UNIVERSITE DES ANTILLES ET DE LA GUYANE

Faculté de Sciences Exactes et Naturelles

Année 2014

N° attribué par la bibliothèque

/ / / / / / / / / /

THESE

Pour obtenir le grade de docteur de l'Université des Antilles et de la Guyane,
Discipline : Microbiologie/Bioinformatique

Présentée par **David COUVIN**

Mise au point, développement et gestion d'une base de données mondiale des génotypes circulants de bacilles tuberculeux : méthodes moléculaires et outils Web pour cartographier, comprendre et maîtriser l'épidémie.

Sous la direction du Dr. Nalin RASTOGI

Soutenue le *02 décembre 2014* à l'Institut Pasteur de la Guadeloupe

Jury :

Rapporteur :	Professeur Miguel VIVEIROS
Rapporteur :	Docteur Igor MOKROUSOV
Examinateur :	Professeur Olivier GROS
Examinateur :	Professeur Raymond CESAIRE
Examinateur :	Docteur Antoine TALARMIN
Directeur de Thèse :	Professeur Nalin RASTOGI

A la mémoire de Mamie Dalie Aya et Papi Jo :
« On reconnait un arbre à ses fruits »

«Etre libre, ce n'est pas seulement se débarrasser de ses chaînes ;
c'est vivre d'une façon qui respecte et renforce la liberté des autres.»
- Nelson Mandela

Remerciements

Je tiens à remercier toutes les personnes qui m'ont apportées leur aide durant ces années de travail de recherche à l'Institut Pasteur de la Guadeloupe (IPG). Je voudrais tout d'abord remercier mon directeur de thèse, le Dr. Nalin RASTOGI, qui m'a permis de réaliser ce travail dans de bonnes conditions et dans la bonne humeur! Merci d'avoir fait naitre en moi cette envie d'écrire. Je remercie aussi le Directeur de l'IPG, le Dr. Antoine TALARMIN, ainsi que les autres membres de la direction dont Monsieur SIGISCAR et Madame DESIREE pour m'avoir permis de mener à bien ce travail.

Je remercie les chercheurs et techniciens de l'Unité Tuberculose et Mycobactéries pour leur soutien tout au long des années que j'ai passé au laboratoire : Dr. Thierry ZOZIO, Dr. Julie MILLET, Mlle Elisabeth STREIT, Mr. Michel PIOCHE, Mr. Frédéric PRUDENTE, Mme Mylène BERCHEL, ainsi que les bioinformaticiens ayant séjourné à l'Unité: Dr. Véronique HILL, Mr. Thomas BURGUIERE, Mr. Christophe DEMAY.

Je remercie également tous les chercheurs, étudiants et stagiaires de passage au laboratoire, notamment Dr. Adolfo BARRETO, Mlle Audrey DAVID, Mr. Iñaki CERVERA-MARZAL, Mlle Zahra LAMIN, Mlle Anne-Gaël BOMER, Mlle Franziska SCHUSTER, Mr. Nadir MEZIDI, Mme Naira DEKHIL, et tous les autres présents à l'Institut avec lesquels j'ai pu partager de bons repas, des biscuits, des « bokits », des « agoulous », des « oréos » et bien plus encore ; notamment Mlle Mirna MOUSSA, Mlle Syndia SADIKALAY, Mlle Daniella GOINDIN, Dr. Mélanie FALORD, Mlle Laura WEIL, Dr. Stéphanie GUYOMARD, Mr. Tony TANT, Mr. Reggie LANOIR, Mr. Thibault COUSIN, Mr. Fabrice GABRIEL, Mlle Christelle ZIMBAN, Mlle Christèle ALDINI, Mr. Thomas SIMONCINI, Mlle Célia DUCAT, et Mlle Angélique MARECHAUX, ainsi que la bande à Quentin, Clément, et Baptiste.

Je remercie aussi les enseignants-chercheurs de l'UAG, spécialement Dr. Sébastien REGIS, Dr. Manuel CLERGUE, Pr. Martine COLLARD, Dr. Jean-Luc HENRY, Dr. Enguerran GRANDCHAMP, ainsi que tous les autres enseignants, pour m'avoir soutenu durant ces années. Je remercie également le Pr. Kristin BENNETT et son équipe pour leur aide et leur collaboration dans de nombreux projets. Je tiens aussi à remercier Mr. Robby ROMNEY, et Mme Anna ROY pour leur aide durant ce travail de thèse.

Enfin, un grand merci à tous mes amis et à toute ma famille, ma mère et mon père, mes sœurs Lolo et Sassou, mes frères Samomo, Teddy et Manou, mes cousines et mes cousins, mes oncles et tantes, et ma Mamie Jo. Je remercie aussi Nounou pour tout.

Toute ma reconnaissance s'adresse également aux membres du jury qui m'ont fait l'honneur de juger ce travail.

RESUME

La tuberculose (TB), maladie infectieuse contagieuse, est causée par une mycobactérie appartenant au complexe *Mycobacterium tuberculosis* (MTBC). Malgré de nombreuses campagnes de vaccination par le BCG (Bacilles de Calmette et Guérin) et les traitements antituberculeux, on observe une réémergence de la maladie depuis les années 80. Ce bouleversement s'explique par la coïnfection avec le virus du VIH/SIDA, la désorganisation des systèmes de santé et par l'apparition des bactéries résistantes aux principaux antibiotiques antituberculeux. Dans ce contexte une meilleure connaissance des mouvements et de l'évolution des clones circulants du bacille tuberculeux permettrait de détecter plus rapidement et de façon plus pertinente l'émergence d'épidémies. Le contrôle et la surveillance de la maladie représentent des moyens essentiels pour lutter contre l'expansion mondiale de la tuberculose. C'est ainsi que l'unité de la Tuberculose et des Mycobactéries de l'Institut Pasteur de la Guadeloupe (IPG) a mis en place une base de données de profils génotypiques pour l'étude de l'épidémiologie globale de la tuberculose.

Dans le cadre de cette thèse, nous décrirons les méthodes de bioinformatique qui ont été utilisées pour la gestion et l'exploitation de la 6ème version (SITVIT2) de cette base de données des génotypes pour mieux comprendre l'évolution et la dissémination mondiale du bacille tuberculeux. Cette base intègre aussi plusieurs marqueurs moléculaires tels que les MIRU-VNTRs (« mycobacterial interspersed repetitive units-variable number of tandem repeats ») et le Spoligotype43 servant à mieux décrire et identifier les familles de profils génotypiques du MTBC. La banque de données SITVIT2 ne cesse d'évoluer, et elle contient actuellement des informations épidémiologiques sur 111 635 isolats cliniques. Cette base de données est consultable en ligne à l'adresse suivante : http://www.pasteur-guadeloupe.fr:8081/SITVIT2/. Nous avons également réalisé une base de données nommée SITVITBovis qui est dédiée à la tuberculose bovine. Cette base de données sera accessible à l'adresse : http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis/.

Parallèlement au développement et à la gestion des bases de données, ce travail de thèse s'est appuyé sur plusieurs études épidémiologiques et phylogéographiques mettant en corrélation les données disponibles sur la résistance aux médicaments ou les caractéristiques démographiques (dont le sexe, l'âge, le statut VIH, et l'origine du patient). Notre initiative de recherche permet ainsi d'améliorer la caractérisation phylogénétique détaillée des lignées du MTBC, ainsi que l'épidémiologie des clones en circulation, afin d'élaborer une cartographie géographique des isolats cliniques prédominants pour les bacilles tuberculeux principalement impliqués dans la maladie, à l'échelle nationale, régionale et mondiale. La superposition ultérieure de ces cartes avec des données sociopolitiques, économiques et démographiques obtenues auprès de Systèmes d'information géographique (SIG) ont permis de dresser un portrait précis des disparités actuelles par pays ou sous-région. Cette thèse représente une importante collaboration avec plusieurs équipes de chercheurs travaillant également dans le domaine de l'épidémiologie moléculaire de la tuberculose.

L'objectif à long terme de ce travail, est d'optimiser la structure de la base, et de pérenniser l'enrichissement de celle-ci (notamment par l'automatisation et la facilité de la saisie de données). Une représentation optimale et une meilleure accessibilité des données de la base conforteraient les efforts faits pour le contrôle et la surveillance de la TB dans monde.

ABSTRACT

Tuberculosis (TB) is a contagious infectious disease caused by mycobacteria belonging to the *Mycobacterium tuberculosis* complex (MTBC). Despite numerous campaigns of vaccination by the BCG (Bacillus Calmette-Guérin) vaccine, and TB treatments, there was a resurgence of the disease since the 80s. This disruption is due to co-infection with HIV/AIDS, disorganization of health systems and the emergence of bacteria resistant to anti-TB drugs. In this context, a better understanding of the movements and changes of circulating clones of tubercle bacilli would detect faster and more relevant emerging epidemics. The control and monitoring of the disease are essential means to fight against the worldwide spread of tuberculosis. Thus the TB and Mycobacteria unit of the Pasteur Institute of Guadeloupe (IPG) has developed a database of genotypic profiles for the study of global epidemiology of tuberculosis.

In this thesis, we discuss the Bioinformatical methods that have been used for the management and development of the 6th version (SITVIT2) of this database of genotypes to better understand the evolution and global dissemination of the tubercle bacillus. The database also includes several molecular markers such as MIRU-VNTRs (mycobacterial interspersed repetitive units- variable number of tandem repeats) and Spoligotype43 allowing a better description and identification of the families of genotypic profiles of Mycobacterium tuberculosis complex. SITVIT2 database is constantly evolving, and it currently contains epidemiological information on 111,635 clinical isolates. This database is available online at the following address: http://www.pasteur-guadeloupe.fr:8081/SITVIT2/. We also developed a database named "SITVITBovis", which is dedicated to bovine tuberculosis. This database will be accessible at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis/.

Alongside the development and management of databases, this work was based on several epidemiological and phylogeographic studies correlating data on drug resistance or demographic characteristics (including sex, age, HIV status, and the origin of the patient). Our research initiative is thus focused to further improve in depth phylogenetic characterization of MTBC lineages, as well as the epidemiological analysis of circulating clones to generate evidence-based geographical mapping of predominant clinical isolates of tubercle bacilli causing the bulk of the disease both at country and regional level. Further superposition of these maps with socio-political, economical and demographical available through Geographic Information Systems (GIS) allows to have a precise view of prevailing disparities at the level of country or sub-region. This thesis represents an important collaboration with several researchers teams also working in the field of molecular epidemiology of tuberculosis.

The long-term goal of this work is to further optimize the database structure, and sustain enriching it (including automation and ease of data entry). Optimum performance and accessibility of data in the database would reinforce efforts to control and surveillance of TB in the world.

AVANT-PROPOS

A travers l'histoire évolutive et la phylogéographie du complexe *Mycobacterium tuberculosis*, nous avons cherché à explorer la dissémination de la tuberculose à travers les migrations et les corrélations pouvant exister entre différents peuples et civilisations. Cette découverte du «méta-monde» de *M. tuberculosis* lié à l'histoire et la géographie de l'humanité permet d'effectuer des voyages virtuels enrichissants car la tuberculose qui sévit ne fait que refléter le devenir et l'évolution des hommes et des femmes d'aujourd'hui. Parce qu'une image vaut mille mots ("*a picture is worth a thousand words*"), la cartographie des génotypes circulants de bacilles tuberculeux développée au cours de cette thèse permet de revisiter quelques pans de l'histoire, une histoire qui s'assemble et se reconstruit au gré des peuples qui vivent ensemble.

TABLE DES MATIERES

INTRODUCTION..... 10

1. Introduction générale.....10

2. Les mycobactéries et le complexe Mycobacterium tuberculosis (MTBC).....11

3. Une longue histoire : de la Phtisie à la tuberculose 16

4. Différentes formes cliniques de la tuberculose 18

4.1 Tuberculose pulmonaire.....18

4.2 Tuberculose extra-pulmonaire.....19

5. Focus sur la tuberculose bovine 20

6. Epidémiologie et phylogéographie mondiales de la tuberculose.....22

6.1 Epidémiologie mondiale de la tuberculose.....22

6.2 Quelques données démographiques et socio-économiques.....23

6.3 Phylogéographie mondiale de la tuberculose.....24

6.4 Méthodes de génotypage et classification des souches bactériennes.....27

7. Diagnostic, transmission, traitement et moyens de lutte.....29

7.1 Diagnostic et transmission 29

7.2 Les antibiotiques antituberculeux et la résistance 30

7.3 Le vaccin BCG.....32

7.4 La stratégie Stop TB et le programme DOTS 33

8. Migrations de populations.....33

9. L'utilité des Bases de Données pour la recherche sur la tuberculose.....36

10. SITVIT : Une Base de Données en constante évolution 36

11. La Bioinformatique d'aujourd'hui.....37

12. Objectifs de ce travail 39

MATERIELS ET METHODES..... 40

1. Méthodes de génotypage et techniques basées sur la PCR.....40

1.1 L'IS6110-RFLP 40

1.2 Le spoligotypage ou « spacer oligo nucleotide typing » (spoligotyping).....41

1.3 MIRU-VNTR typing 42

2. Drug Susceptibility Typing (DST) 46

3. Tests, Outils, et Lois de Biostatistiques.....46

4. Index de Discrimination de Hunter-Gaston (HGDI) et Taux de transmission récente 48

5. Théorie des Graphes et Algorithmes de Reconstruction Phylogénétique 49

6. Bases de données, Logiciels de Phylogénie et outils Bionformatiques 50

6.1 Les bases de données et outils disponibles pour l'épidémiologie moléculaire de la TB....50
6.3 Logiciels de reconstruction phylogénétique et outils de bioinformatique....54
7. Réseaux bayésiens....55
8. Outils de programmation Web et Technologie Java....57
8.1 Petit historique des technologies web ayant marqué le siècle dernier....57
8.2 L'interface Web pour consulter la base de données....57
9. MySQL pour la Gestion de notre Base de Données Relationnelle....58
10. Langage Perl pour le traitement interne des données....59
11. Collection et interprétation des données dans SITVIT2....61
RESULTATS ET DISCUSSIONS....63
1. Chapitre 1 – Les Bases de Données et les Outils Bioinformatiques....63
1.1 Résumé du chapitre....63
1.2 Nécessité de développer une approche fondée sur l'utilisation concomitante de l'épidémiologie moléculaire, les bases de données et les outils bioinformatiques....64
1.3 Les bases de données successives développées au cours de ce travail: SITVITWEB et SITVIT2....65
1.3.1 Présentation de SITVITWEB (Article 1)....66
1.3.2 Présentation de SITVIT2 (Article 2 en préparation + Article 2bis)....68
1.3.3 Un exemple de l'utilisation de SITVIT2 centré sur la dissémination globale de la lignée Beijing – une urgence mondiale (Article 3)....87
1.3.4 Développement de la base SITVITBovis dédiée à la tuberculose bovine (Article 4 en préparation)....93
1.3.5 SITVIT-KBBN : une ressource basée sur les réseaux bayésiens (Article 5)....100
2. Chapitre 2 – Utilisation de typage moléculaire et outils Web pour cartographier et comprendre l'épidémie au niveau mondial....102
2.1 Résumé du chapitre....102
2.2 Epidémiologie de la tuberculose en Amérique latine....103
2.2.1 Etude épidémiologique au Mexique (Article 6)....105
2.2.2 Etude épidémiologique en Colombie (Article 7)....107
2.2.3 Etudes épidémiologiques au Pérou (Article 8 + travail en cours)....110
2.2.4 Etudes épidémiologiques au Brésil (Article 9 + travail en cours)....117
2.3 Epidémiologie de la tuberculose en Afrique de l'Ouest et de l'Est....123
2.3.1 Etude épidémiologique en Guinée-Bissau (Article 10)....125
2.3.2 Etude épidémiologique en Ethiopie (Article 11)....127
2.4 Epidémiologie de la tuberculose en Asie de l'Ouest....131

2.4.1 Etude épidémiologique en Irak « après guerre » (Article 12) 131

2.4.2 Etude épidémiologique au Sultanat d'Oman (travail réalisé avec la collaboration de Dr. JPN Singh) 134

2.5 Etude épidémiologique de la tuberculose dans une population pédiatrique de l'Inde (Article 13) 137

2.6 Etude épidémiologique de la tuberculose en Malaisie (travail en cours) 139

2.7 Epidémiologie de la tuberculose dans la région de Sichuan en Chine (Article 14) 142

2.8 Epidémiologie des formes susceptible et multi-résistante de la tuberculose en Bulgarie (travail réalisé avec la collaboration de Dr. Stefan Panaiotov) 144

2.9 Epidémiologie de la tuberculose en Scandinavie 148

2.9.1 Etude épidémiologique en Suède (Article 15) 148

2.9.2 Etude épidémiologique en Finlande (Article 16) 150

2.10 Epidémiologie de la tuberculose dans la région Rhône-Alpes, France (travail en cours) 154

2.11 Epidémiologie de la tuberculose dans la région de la Vénétie en Italie (Article 17) 158

2.12 Epidémiologie de la tuberculose multi-résistante en Espagne (travail réalisé en collaboration avec Dr. Sofia Samper) 160

CONCLUSION ET PERSPECTIVES 165

BIBLIOGRAPHIE 170

LISTE DES PUBLICATIONS (REALISEES OU EN COURS) 181

ANNEXES 183

INTRODUCTION

1. Introduction générale

Aujourd'hui encore, la tuberculose, maladie en réémergence, demeure un fléau majeur qui menace notre monde. Malgré de nombreux efforts prophylactiques et thérapeutiques plus ou moins efficaces, ainsi qu'une baisse notable de son incidence dans certains pays industrialisés, la tuberculose continue à menacer aussi bien les pays en voie de développement que les régions développées. Selon le Rapport sur la tuberculose dans le monde 2012 publié par l'Organisation Mondiale de la Santé (OMS ; http://www.who.int/tb/publications/global_report/fr/), les chiffres fournis font apparaître:

- une baisse continue du nombre de personnes malades de la tuberculose, cependant on note toujours un nombre impressionnant au niveau mondial; en effet, 8,7 millions de nouveaux cas sont venus s'ajouter en 2011 à la longue liste des patients déjà touchés par cette maladie.
- un nombre estimatif de 1,4 million de décès dus à la tuberculose, dont 500 000 femmes;
- une diminution des taux d'incidence et de mortalité dans les six régions de l'OMS, bien que les régions africaine et européenne ne soient pas encore en voie d'atteindre l'objectif consistant à réduire de moitié la mortalité par rapport à 1990 d'ici 2015;
- des progrès toujours très lents dans la riposte à la tuberculose multirésistante (MDR), avec seulement un patient sur cinq diagnostiqué dans le monde ;
- un grave déficit de financement des activités de soins et de lutte contre la tuberculose. Entre 2013 et 2015, ces activités nécessiteront jusqu'à 8 milliards de dollars (US $) par an dans les pays à revenu faible ou intermédiaire.
- la co-infection TB-VIH et l'apparition de tuberculose MDR/XDR (la tuberculose XDR se définissant comme une tuberculose MDR résistant à une fluoroquinolone ainsi qu'à au moins un des trois antituberculeux de seconde intention injectables – capréomycine, kanamycine et amikacine) compliquent encore la gestion de la maladie et augmentent considérablement le taux de mortalité dû à la tuberculose chez les patients immunodéprimés, ce qui constitue un problème de santé majeur au niveau mondial.

En ce qui concerne la Caraïbe et l'Amérique latine, la situation épidémiologique de la tuberculose est très disparate d'un territoire à l'autre avec des taux d'incidence rapportés par l'OMS variant de moins de 2 à plus de 200 cas de tuberculose pour 100 000 habitants. Il en est de même pour les départements français d'Amérique (DFA – Guadeloupe, Martinique, Guyane française) pour lesquels des taux d'incidence relativement faibles, inférieurs à 10 cas pour 100 000 habitants, ont été rapportés pour les deux territoires insulaires, Guadeloupe et Martinique tandis qu'en Guyane française l'incidence atteint 25 cas avec culture positive pour 100 000 habitants (52 cas pour 219 000 habitants en 2010; source IP Guadeloupe), plaçant la Guyane en tête des départements français pour l'incidence de tuberculose.

La tuberculose est fortement liée aux populations qui la transmettent et la propagent – une observation qui permet de supposer que les dynamiques de transmissions différentes d'un territoire à l'autre seraient à l'origine de la variabilité génotypiques des souches circulantes. Dans ce contexte, nos travaux antérieurs ont souligné que la spécificité phylogéographique des clones de bacilles tuberculeux est étroitement liée à l'origine géographiques et l'ethnicité des patients. Au cours de cette thèse, nous décrivons les méthodes moléculaires et outils Web permettant efficacement de cartographier, comprendre et maîtriser l'épidémie grâce à une approche basée sur la mise au point, le développement et la gestion d'une base de données mondiale des génotypes circulants de bacilles tuberculeux. Ces outils sont particulièrement utiles pour la surveillance et l'étude des filières de transmission de la tuberculose au niveau local, régional et mondial.

2. Les mycobactéries et le complexe *Mycobacterium tuberculosis* (MTBC)

Les bactéries sont des organismes vivants unicellulaires et procaryotes présents dans tous les milieux. Parmi les différents types de bactéries existantes, *Mycobacterium tuberculosis* fait partie de l'embranchement des *Actinobacteria* (**Figure 1**).

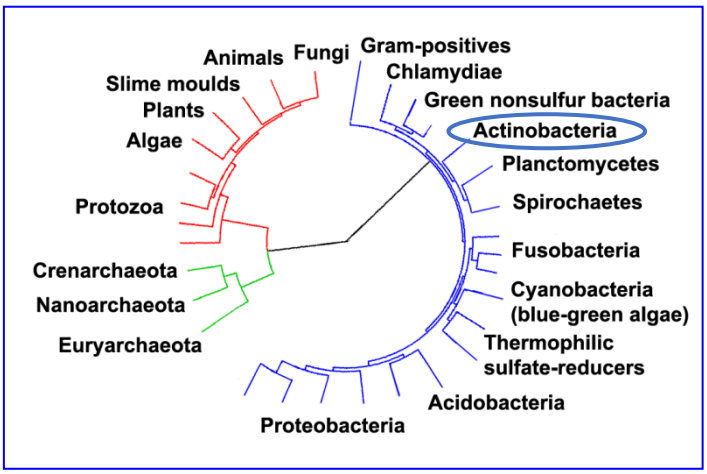


Figure 1 : Arbre phylogénétique, basé sur un génome complètement séquencé (**Ciccarelli FD 2006**). Sur la figure, les eucaryotes sont en rouge, les procaryotes différenciés par les *Archaea* en vert, et les *Eubacteria* (vraies bactéries) sont en bleu.

Les mycobactéries (**Figure 2**) appartiennent au genre *Mycobacterium*, de la famille des *Mycobacteriaceae*, de l'ordre des Actinomycétales et de la classe des Actinobactéries (**Shinnick et Good, 1994**). Il s'agit de bacilles droits ou légèrement incurvés de 1 à 10 µm de long sur 0,2 à 0,6 µm de large, immobiles, ne formant pas de spores, conidies, hyphes aériens ou capsules. Les mycobactéries sont définies comme des BAAR (bacilles acido-alcoolo-résistants), généralement non observables après coloration de Gram, on les caractérise notamment de bactéries « Gram neutre » (**Rastogi et coll., 2001**). La technique de référence Ziehl-Neelsen permet de colorer et d'identifier les mycobactéries, en utilisant les propriétés acido-alcoolo-résistantes de leur paroi (**David et coll., 1989**). Les mycobactéries sont omniprésentes dans l'environnement mais elles peuvent devenir néfastes pour l'homme dans certaines conditions comme l'immunodépression, la présence de lésions, l'âge avancé ou

encore la présence de maladie(s) préexistante(s). La classification des espèces figurant à l'intérieur du genre *Mycobacterium* a beaucoup évolué, notamment grâce à l'apparition des techniques de biologie moléculaire. La classification de Runyon définit 4 groupes parmi les mycobactéries sur la base de caractères culturaux (temps de croissance sur milieux solides) et morphologiques (pigmentation des colonies). La pratique courante distingue les mycobactéries tuberculeuses, pathogènes obligatoires dont le réservoir est l'homme ou certains mammifères (*Mycobacterium tuberculosis* Complex) et les mycobactéries non tuberculeuses encore dites « atypiques », présentes dans l'environnement, qui ne sont pas des pathogènes obligatoires. Parmi ces dernières figurent *M. avium, M. intracellulare, M. marinum, M. kansasii, M. xenopi, M. simiae, M. fortuitum, M. smegmatis, M. ulcerans, M. scrofulaceum* …

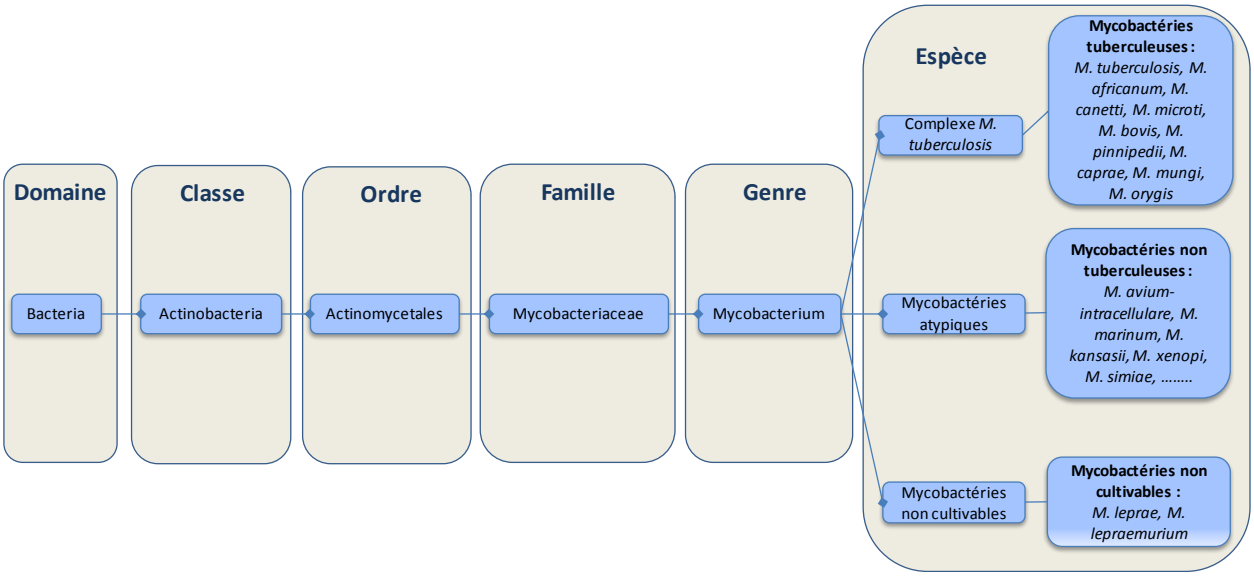


Figure 2 : Représentation schématique des mycobactéries.

On peut également classer les mycobactéries selon leur vitesse de croissance : croissance rapide versus croissance lente – et un arbre phylogénétique 16S permet aisément de les discriminer (**Figure 3**). Il faut cependant noter l'existence d'espèces non cultivables sur les milieux de culture comme *Mycobacterium leprae*, agent étiologique de la lèpre humaine.

Mycobacterium tuberculosis (ou bacille de Koch) est l'agent infectieux responsable de la tuberculose humaine. *M. tuberculosis* fait partie de tout un complexe, le complexe *Mycobacterium tuberculosis* (MTBC). Il ressemble à un fin bâtonnet mesurant environ 2 à 4 µm de long et environ 0,4 µm de large (**Figure 4**). La croissance de cette bactérie est lente, avec un temps de doublement pouvant varier entre 12h et 24h (**Harshey et Ramakrishnan, 1977**). L'ADN de *M. tuberculosis* présente plus de 99,9 % d'identité avec celui des autres membres du MTBC composé notamment de *M. bovis* – l'agent de la tuberculose bovine ; *M. bovis* BCG, la souche vaccinale dérivée de *M. bovis* ; M. africanum – un pathogène de l'homme d'origine africaine ; *M. canettii* – un bacille tuberculeux atypique avec des colonies « lisses » fréquemment observé chez les patients originaires de Djibouti ; *M. microti*, responsable de la tuberculose chez certains rongeurs (***voir description détaillée ci-après***).

Le génome de la souche de référence : *Mycobacterium tuberculosis* H37Rv, a été complètement séquencé en 1998 (**Cole et coll., 1998**). Il est formé d'un chromosome circulaire de 4 411 529 paires de bases à haut pourcentage en G/C (65,6%), codant 3924

gènes. Les travaux génétiques et génomiques sur *M. tuberculosis* et sur d'autres mycobactéries ont permis de recueillir d'innombrables informations sur l'organisation génétique des mycobactéries. Ces données génétiques ont, de plus, apporté des éléments clés dans la compréhension de la physiopathologie de la bactérie, pouvant aboutir, sur un plus long terme, à l'identification de cibles pour la mise au point de nouveaux traitements ou vaccins. Grâce à leur cellule riche en lipide et à la grande proportion de guanine et de cytosine dans leur ADN, les mycobactéries présentent de grandes capacités de conservation/préservation (**Leão et Portaels, 2007**).

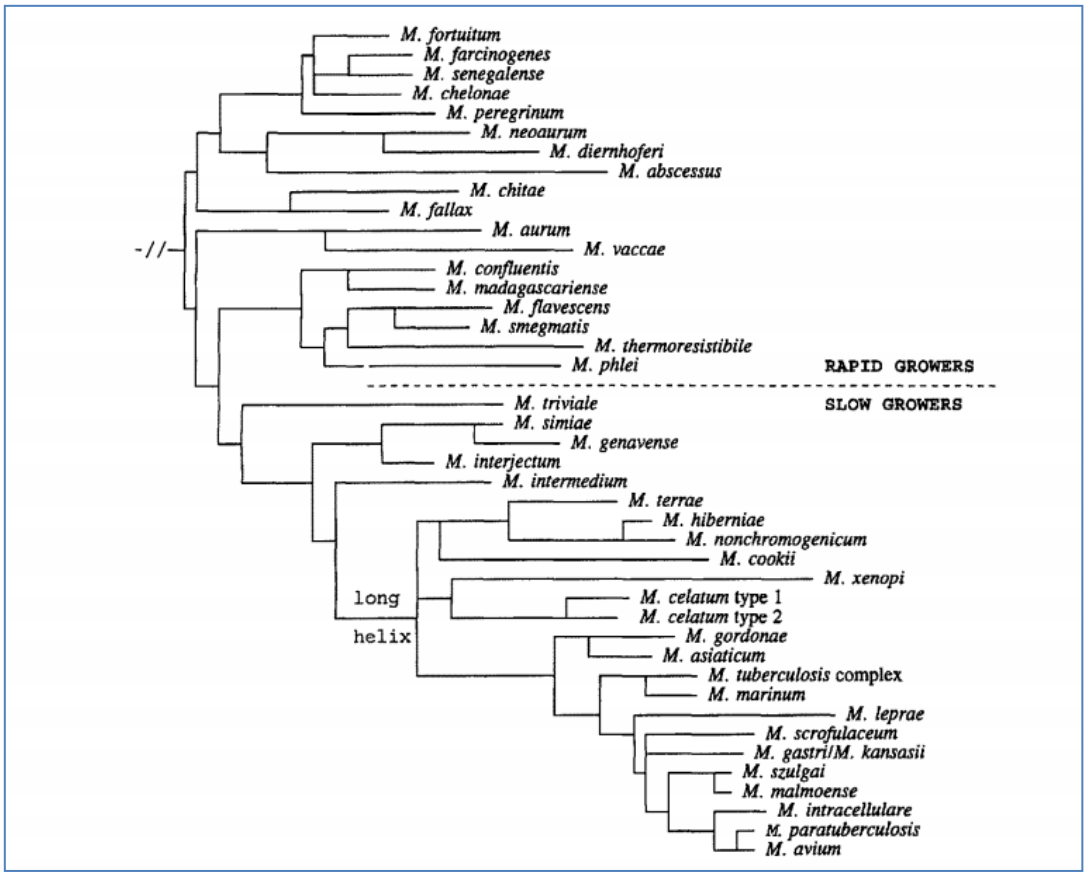


Figure 3 : Arbre phylogénétique 16S permettant de discriminer les mycobactéries à croissance rapide et lente (tiré de **Pitulle et coll., 1992**, **Rogal et coll., 1990** et **Kirschner et coll., 1993**).

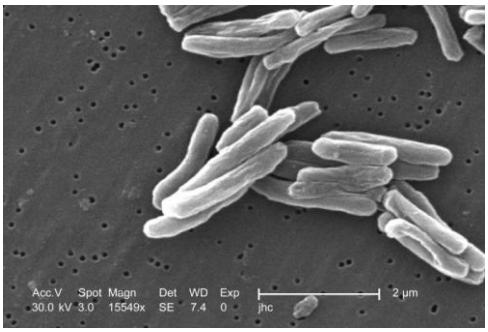


Figure 4 : *Mycobacterium tuberculosis* grossi 15549 fois (Source CDC).

L'analyse du génome de *M. tuberculosis* a également permis d'identifier plusieurs nouvelles familles de gènes précédemment inconnues, comme par exemple les gènes codant

pour les protéines PE et PPE, qui occupent 10 % du génome. L'analyse du génome a aussi indiqué que 14 gènes *esx* codant pour les protéines de la famille ESAT-6 (certains membres de la famille ESAT-6 étant impliqués dans le pouvoir pathogène de *M. tuberculosis*) sont présents dans 11 régions différentes (**Cole et coll., 1998**). Lors des études de génomique comparative employant différentes techniques d'hybridation, plusieurs régions de différence (RD1-RD14), codant pour environ 140 protéines, ont été trouvées absentes chez *M. bovis* BCG par rapport à la souche virulente *M. tuberculosis* H37Rv, ce qui permet d'étudier les différences génétiques potentiellement impliquées dans la virulence de certains membres du complexe. Ainsi, des études concernant la présence ou l'absence des régions RD chez un grand nombre de souches de MTBC ont permis de définir des relations phylogénétiques entre les différents membres du complexe (**Figure 5**) et de proposer un nouveau schéma de l'évolution des bacilles tuberculeux, remettant en question l'hypothèse précédente selon laquelle *M. bovis* serait l'ancêtre de *M. tuberculosis* (**Brosch et coll., 2002** ; **van Ingen et coll., 2012**).

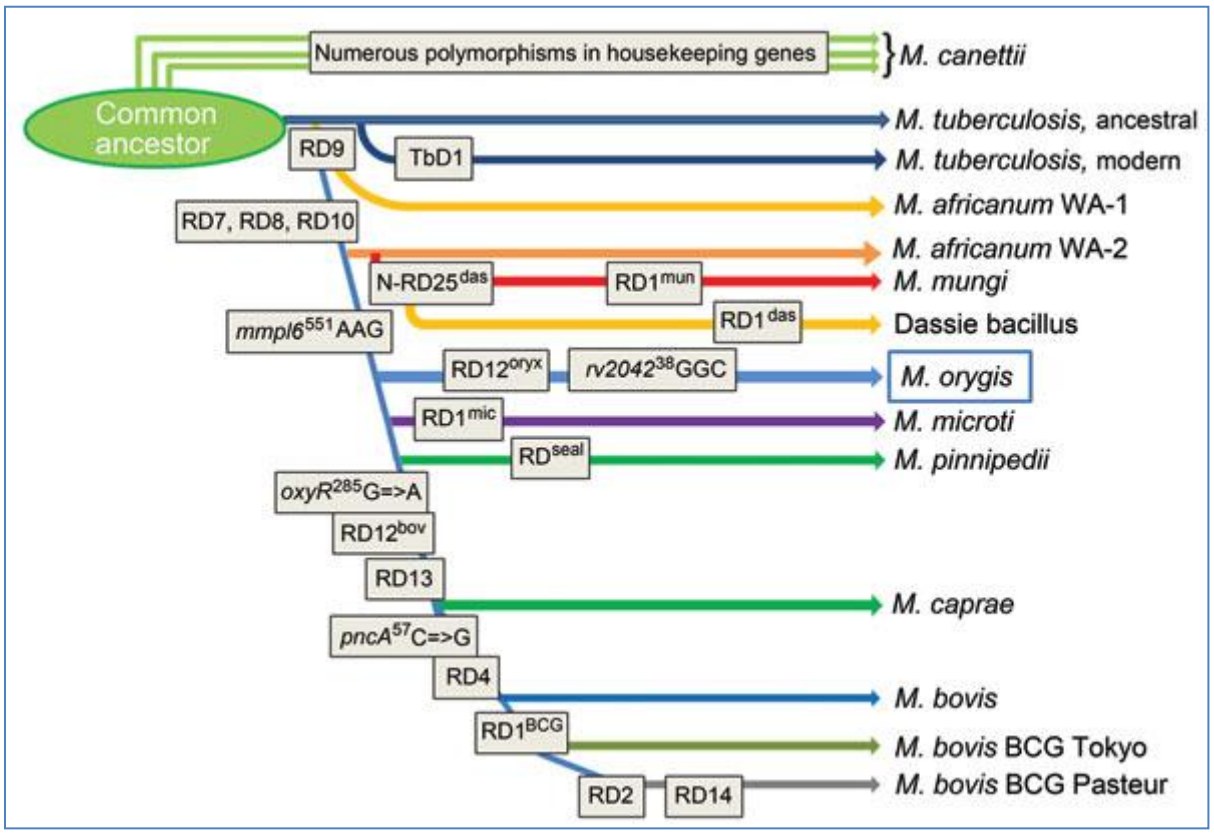


Figure 5 : Illustration schématique des relations phylogénétiques entre les membres du complexe *M. tuberculosis* selon **van Ingen et coll.,** (**2012**), basée sur l'étude initiale de **Brosh et al.** (**2002**).

En conclusion, le complexe *Mycobacterium tuberculosis* (MTBC) est écologiquement très vaste, et composé de plusieurs sous-espèces incluant :

- ***Mycobacterium tuberculosis*** qui est l'agent pathogène de la tuberculose humaine. On le retrouve plus rarement chez les bovidés après un contact avec un homme infecté (**Serrano et coll., 1986**).

- ***Mycobacterium africanum*** qui touche aussi l'homme principalement (**Castets et coll., 1968**), mais peut aussi infecter d'autres animaux (**Gudan et coll., 2008** ; **Rahim et coll., 2007**).
- ***Mycobacterium canettii*** qui a été décrit avec précision par D van Soolingen (**van Soolingen et coll., 1997**), est particulièrement associé à la Corne de l'Afrique (essentiellement Djibouti) (**Koeck et coll., 2011** ; **Fabre et coll., 2010** ; **Brosch et coll., 2002** ; **Pfyffer et coll., 1998**), et correspondrait à l'espèce progénitrice à partir de laquelle *M. tuberculosis* a émergé (**Fabre et coll., 2004** ; **Gutierrez et coll., 2005** ; **Supply et coll., 2013** ; **Blouin et coll., 2014**). Son réservoir naturel, sa spécificité d'hôtes, ainsi que son mode de transmission demeurent encore méconnus (**Miltgen et coll., 2002**).
- ***Mycobacterium microti*** qui touche principalement les rongeurs (**Wells et coll., 1937**), est non-pathogène pour l'homme (**Hart et coll., 1977**). Il fut nommé plus tard *M. tuberculosis var. muris*, faute de pouvoir être distingué alors de *M. tuberculosis*. Un vaccin atténué de cette souche fut utilisé en Tchécoslovaquie de 1951 à 1969 tandis que des essais étaient conduits en Grande Bretagne de 1950 à 1952 avec des formes non atténuées (http://fr.wikipedia.org/wiki/Mycobacterium_microti). Un demi-million d'enfants furent vaccinés par ce vaccin (**Sula et coll., 1976**).
- ***Mycobacterium bovis*** qui est l'agent causal de la tuberculose bovine, peut aussi toucher l'homme (suite à la consommation de produits laitiers non-pasteurisés par exemple), et d'autres animaux sauvages ou domestiques (**Thoen et coll., 1984** ; **Chartier et coll., 1991** ; **Thorel et coll., 1998**). Il est probable que la tuberculose bovine n'aurait jamais atteint des proportions épidémiques chez l'humain si elle n'était pas transmise par les aliments (**Francis J, 1958**).
- ***Mycobacterium pinnipedii*** (ou bacille du phoque) qui est responsable de la tuberculose chez les pinnipèdes, peut aussi rarement infecter l'homme (**Cousins et coll., 2003**).
- ***Mycobacterium caprae*** qui est principalement isolé chez les caprins (**Aranaz et coll., 1999**), peut concerner les humains dans de rares cas (**Kubica et coll., 2003**).
- ***Mycobacterium mungi*** qui est principalement retrouvé chez les mangoustes bagués (« banded Mongooses »), présente un spectre d'hôte et des modalités de transmission à ce jour inconnus (**Alexander et coll., 2010**).
- ***Mycobacterium orygis*** (ou oryx) qui a été récemment décrit chez des membres de la famille des bovidés (oryx, gazelles, antilopes, cerfs, …) (**van Ingen et coll., 2012**). Un cas de transmission de ce microorganisme d'un humain à une vache a déjà été documenté (**Dawson et coll., 2012**).
- **Dassie bacillus** (ou bacille Dassie) qui est semblable à M. microti, et est responsable de la tuberculose chez les rats des rochers (Petromus typicus) (**Mostowy et coll., 2004** ; **Parsons et coll., 2008**).

3. Une longue histoire : de la Phtisie à la tuberculose

L'histoire de la tuberculose est à la fois riche et complexe, elle se faufile à travers les époques depuis la préhistoire. Cette maladie, qui a inspirée de nombreuses œuvres artistiques au cours du temps, continue à jouer un rôle majeur dans nos sociétés actuelles. Depuis l'isolement et la caractérisation d'un ancien ADN de *M. tuberculosis* datant de 17 000 ans avant J. C. chez une espèce de bison disparue, nous savons que la TB était présente en Amérique vers la fin du pléistocène (**Rothschild et coll., 2001**). La tuberculose est donc une maladie très ancienne. Dans les restes humains, l'ADN a permis de retracer la présence de la tuberculose dans les momies égyptiennes, avec la caractérisation de *M. tuberculosis* et *M. africanum* (**Zink et coll., 2003**). L'origine de la tuberculose humaine serait associée à la transition démographique du néolithique il y a environ 11000 ans. En effet, le développement de la domestication des animaux a augmenté la probabilité de transfert zoonotique de nouveaux agents pathogènes pour l'homme tandis que les innovations agricoles ont entrainé l'augmentation des densités de population qui a contribué à maintenir le cycle infectieux de la tuberculose (**Wolfe et coll., 2007**). Cependant, on continue à se poser des questions sur l'origine de la tuberculose humaine. Une hypothèse en ferait (i) une zoonose provenant de la tuberculose bovine (i.e. l'humain aurait été infecté par l'animal) (**Stead, 1997** ; **Wolfe et coll., 2007**) ; dans une autre hypothèse au contraire, (ii) la tuberculose bovine résulterait d'une zoonose inverse (i.e. la propagation de la tuberculose humaine aux bovins) (**Gibbons, 2008** ; **Smith et coll., 2009**). Quoiqu'il en soit, la tuberculose affiche des caractéristiques d'une pathologie très ancienne : le phénomène de latence puis la réactivation de la maladie est caractéristique d'une entité très ancienne (**Barry et coll., 2009**).

La maladie était largement répandue en Egypte et à Rome (**Zink et coll., 2003** ; **Donoghue et coll., 2004**), et bien que faiblement répandue, elle était aussi présente sur le continent Américain avant l'arrivée de Christophe Colomb (**Salo et coll., 1994** ; **Konomi et coll., 2002** ; **Sotomayor et coll., 2004**). Cependant, on continue de chercher à savoir si la tuberculose aurait des origines indépendantes dans les 2 hémisphères (Nord et Sud), ou si elle aurait été apportée dans les Amériques par les Européens (du moins sous sa forme récente, lignées « Euro-américaines »).

Au cours du temps (depuis 4000-2000 avant J-C), la tuberculose a existé sous différentes appellations (**Tableau 1**) (surnommée phtisie par **Hippocrates** vers 460 avant J-C), touchant des hommes de différentes origines ou classes sociales, et concernant aussi d'autres mammifères. Bien qu'**Aristote** (384-322 avant J-C) considérât que la maladie était contagieuse, la plupart des auteurs grecs pensaient que la maladie était héréditaire, et en partie le résultat de la faiblesse mentale ou morale d'un individu. Plus tard, **Clarissimus Galen** (131-201 AD) décrivit la tuberculose comme une maladie due à la malnutrition (**Pease, 1940**). L'héliothérapie fut conseillée vers le 5ème siècle par **Caelius Aurelianus**. Les physiciens romains recommandèrent aux malades de se baigner dans l'urine humaine, de manger du foie de loup, et de boire du sang d'éléphant. Au Moyen-âge, on pensait que le fait de toucher les souverains d'Angleterre et de France avait le pouvoir de soigner les personnes atteintes du « King's Evil » ou scrofula (scrophula ou struma) – le gonflement des ganglions lymphatiques, souvent associés à la tuberculose.

L'épidémie de tuberculose en Europe, connue plus tard sous le nom de « Great White Plague » (grande peste blanche), a probablement commencé au début du 17ème siècle, pour persister jusqu'au 19ème (**Leão et Portaels, 2007**). Cette période coïncide également avec les expéditions Européennes vers les Amériques et l'Afrique (avec le commerce d'esclaves).

Certaines études ont aussi démontré la présence de la maladie en Asie, et durant le 19ème siècle, un pic d'incidence fut observé en Chine et en Inde.

Tableau 1 : Les différentes désignations de la tuberculose (**Leão et Portaels, 2007**).

Noms	Forme Clinique
Phtisie	Tuberculose (nom Grec)
Maladie du poumon	Tuberculose
Consomption	Tuberculose
Lupus vulgaris	Tuberculose de la peau
Maladie mésentérique	Tuberculose des ganglions lymphatiques abdominaux
Mal de Pott	Tuberculose de la colonne vertébrale
Scrofule	Tuberculose des ganglions lymphatiques du cou
Ecrouelles (King's evil)	Tuberculose des ganglions lymphatiques du cou
Peste blanche (White plague)	Tuberculose pulmonaire
"White swelling"	Tuberculose des os
Tuberculose miliaire	Tuberculose disséminée

Gaspard Laurent Bayle (1774-1816) a prouvé que les bacilles tuberculeux sont la cause de la maladie, et non une conséquence de celle-ci. Le nom « tuberculose » a vu le jour suite à cette découverte de Bayle. Le terme « tuberculose » a donc été utilisé par de nombreux experts en médecine, pour définir la maladie causée par ces fameux « tubercules ». Cependant, c'est un peu plus tard que le germe causant la tuberculose sera formellement identifié. Le 24 Mars 1882, **Robert Koch** rendit publique sa découverte de l'agent pathogène de la tuberculose (TB), *Mycobacterium tuberculosis*, découverte qui lui value le prix Nobel de Physiologie ou Médecine en 1905. L'introduction des sanatoriums (**Figure 6**) fut un événement majeur de la fin du 19ème siècle pour le traitement antituberculeux. Ces établissements proposaient aux patients une nourriture équilibrée, une cure d'air frais, de lumière et de soleil.

Figure 6 : Le sanatorium de Waverly Hills, situé à Louisville dans l'état du Kentucky (USA) qui fonctionna de 1910 à 1961.

4. Différentes formes cliniques de la tuberculose

Bien que la tuberculose soit une maladie qui concerne généralement les poumons, le bacille peut se disséminer dans l'organisme et causer d'autres formes cliniques en se localisant dans d'autres parties du corps. Il existe notamment des formes de tuberculoses ganglionnaires, ostéo-articulaires, uro-génitales ou cutanées. Les formes miliaires et méningées de la tuberculose sont les plus létales.

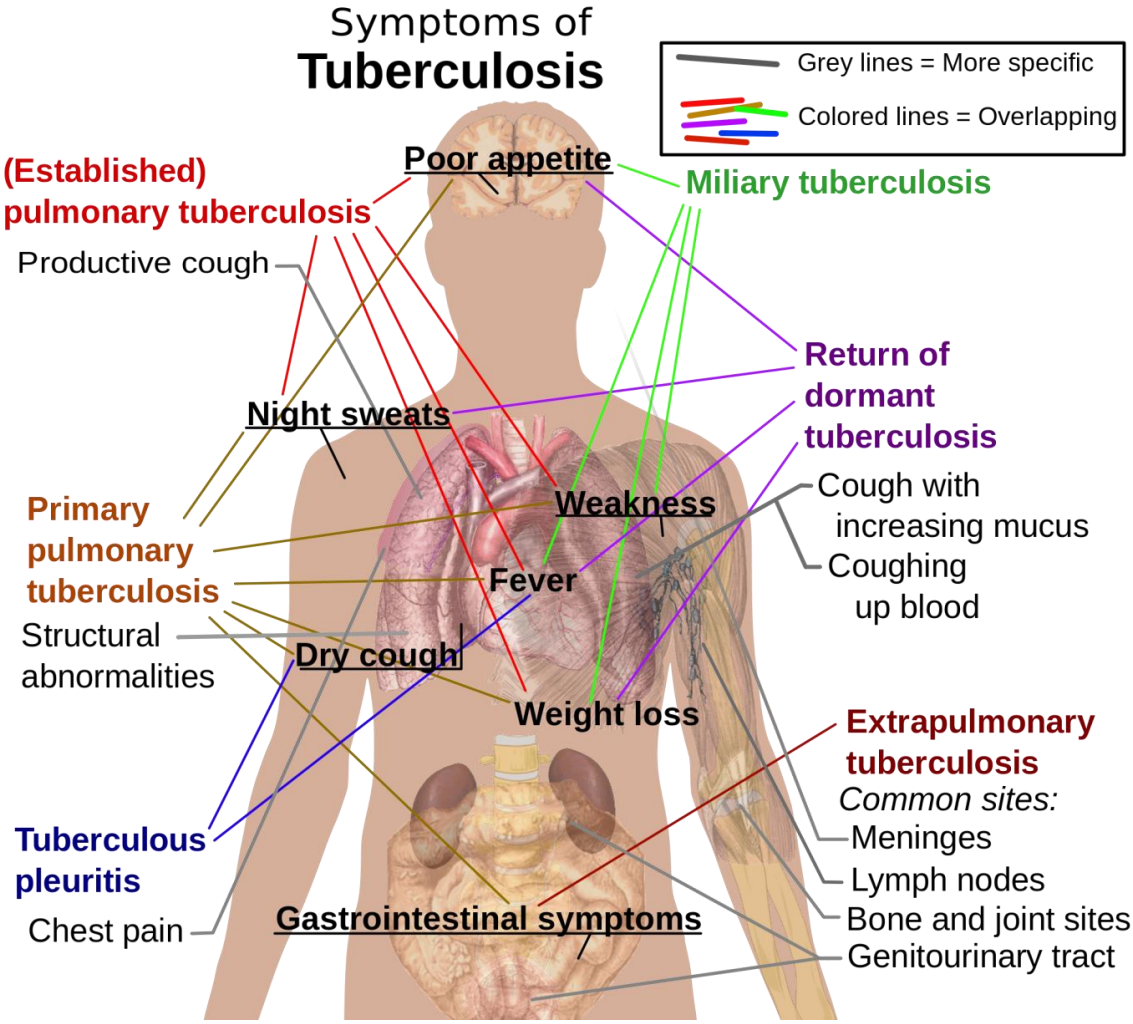


Figure 7 : Les symptômes dus à la tuberculose (**eMedicineHealth**. Author: George Schiffman, MD, FCCP. Last Editorial Review: 1/15/2009).

4.1 Tuberculose pulmonaire

La tuberculose pulmonaire (TBP) est la forme la plus fréquente de la maladie. Elle se caractérise dans sa forme classique par une fièvre prolongée, une toux persistante de plus de 2 semaines, pouvant être accompagnée d'hémoptysie (rejet de sang), un amaigrissement de 5 à 10 kg (survenant en quelques mois), un essoufflement, des douleurs thoraciques, et des sueurs nocturnes.

La radiographie du thorax permet de diagnostiquer une atteinte classique des apex (extrémité supérieure des poumons), avec parfois des cavernes (cavités dans le tissu pulmonaire,

contenant de l'air et des bacilles tuberculeux). Ces cavernes restent présentes lors de la guérison, se calcifient et produisent des séquelles radiologiques appelées « tâches au poumon ».

4.2 Tuberculose extra-pulmonaire

Mis à part le poumon, la tuberculose peut aussi concerner d'autres localisations pour lesquelles on parlera de tuberculose extra-pulmonaire. Dans la plupart des cas, elle est le résultat d'une dissémination hématogène. Le bacille de Koch peut se loger dans différents organes, et provoquer des symptômes aussi graves que ceux de la tuberculose pulmonaire. On notera les formes suivantes :

(i) La tuberculose urogénitale (ou génito-urinaire) est la tuberculose extra-pulmonaire la plus fréquente. Cette tuberculose peut faire partie d'une infection disséminée ou être localisée uniquement au tractus urogénital. Elle peut atteindre chez l'homme, des organes comme le rein, se propager dans les uretères puis la vessie, et peut concerner aussi les canaux déférents, la prostate, les épididymes et les testicules. Cependant, cette forme de tuberculose atteint plutôt l'appareil urinaire chez la femme (**Eichbaum et Rubin, 2002** ; **Frieden et coll., 2003** ; **Gokce et coll., 2002**).

(ii) La tuberculose ganglionnaire concerne essentiellement les enfants et les adolescents. Dans ce cas de tuberculose, les bacilles présents dans les poumons vont passer dans la circulation sanguine ou dans le système lymphatique. La bactérie va se localiser dans les ganglions lymphatiques dans lesquels elle provoquera une inflammation. Dans la majorité des cas, cette tuberculose attaque les ganglions cervicaux (**Emeline Ferard**, www.maxisciences.com, avril 2013).

(iii) La tuberculose intestinale atteint principalement l'intestin grêle et le colon. Le bacille de Koch emprunte dans ce cas, la voie sanguine ou la voie digestive (par une éventuelle ingestion de la bactérie). La majorité des cas de tuberculose intestinale, quasiment disparue en Europe, se retrouve en Afrique (**Emeline Ferard**, www.maxisciences.com, avril 2013).

(iv) La tuberculose ostéo-articulaire (ou osseuse) se situe particulièrement chez les patients des pays en développement d'Afrique et d'Asie, ainsi que chez les personnes immunodéprimées. Elle s'attaque aux os du malade, et touche dans la plupart des cas, la colonne vertébrale, la hanche ou les genoux. Si elle n'est pas traitée, elle peut causer la destruction des articulations (**Emeline Ferard**, www.maxisciences.com, avril 2013).

(v) La tuberculose miliaire représente 2% des atteintes tuberculeuses, et survient plus spécialement chez les personnes âgées, les enfants de moins de deux ans, et les personnes atteintes de VIH/SIDA. C'est une forme disséminée de la maladie qui peut apparaitre suite à une tuberculose mal soignée et atteindre les organes de l'abdomen, le squelette osseux, l'appareil génital, ou les glandes surrénales. La tuberculose miliaire est très grave (plus sévère que la tuberculose pulmonaire

classique), et peut conduire à une insuffisance respiratoire aigüe (**Emeline Ferard**, www.maxisciences.com, avril 2013).

La tuberculose bovine a notamment souvent été associée à des formes de tuberculose extra-pulmonaire (**Lari et coll., 2009** ; **Fallico et coll., 2014**).

5. Focus sur la tuberculose bovine

La tuberculose bovine (BTB) est une des formes de tuberculose à laquelle nous avons attaché une importance particulière au cours de ce travail de thèse, puisque une interface web entière a été développée en se focalisant sur ce type unique de tuberculose (interface SITVITBovis, qui sera vue plus en détail dans la section « Résultats et Discussion »). BTB est globalement moins meurtrière que la tuberculose classique (due à *M. tuberculosis*).

Le commerce de produits agricoles ou animaliers (**Figure 8A**), ainsi que la consommation de produits laitiers (**Figure 8B**) mal pasteurisés, peuvent favoriser la transmission de cette tuberculose bovine. A l'échelle nationale et mondiale, beaucoup d'efforts sont faits pour mieux prévenir et contrôler la BTB (http://www.oie.int/fileadmin/home/fr/media_center/docs/pdf/disease_cards/bovine-tb-fr.pdf).

Jusqu'aux années 1920, date d'apparition des mesures de contrôle dans les pays développés, la tuberculose bovine était l'une des maladies majeures des animaux domestiques à l'échelle mondiale. Aujourd'hui, cette forme de tuberculose que l'on retrouve essentiellement chez les bovins et les animaux sauvages, est une importante zoonose (maladie des animaux qui peut également toucher l'homme).

La BTB est une maladie listée dans le *Code sanitaire pour les animaux terrestres* de l'Organisation mondiale de la santé animale (OIE) et elle doit être déclarée à l'OIE (conformément au Code sanitaire pour les animaux terrestres de l'OIE ; http://www.oie.int/fr/).

Des isolements ont été obtenus à partir de buffles, bisons, ovins, caprins, équidés, camélidés, porcs, sangliers, cerfs, antilopes, chiens, chats, renards, visons, blaireaux, furets, rats, primates, lamas, koudous, élans, tapirs, wapiti, éléphants, sitatungas, oryx, addax, rhinocéros, opossums, écureuils, loutres, phoques, lièvres, taupes, ragondins, coyotes et plusieurs félins prédateurs dont les lions, les tigres, les léopards et les lynx.

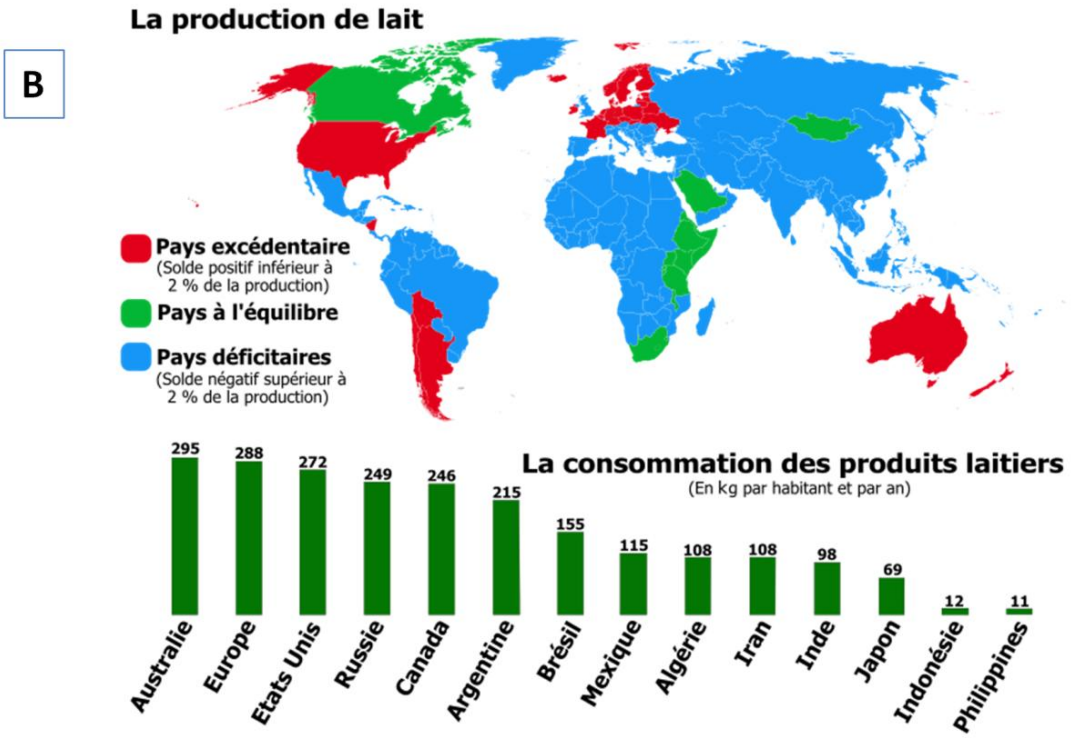


Figure 8 : (**A**) Exportation mondiale des produits agricoles et des aliments en 2010 (http://cartographie.sciences-po.fr/fr/exports-agricultural-and-food-products-world-2010). (**B**) Production et consommation des produits laitiers (http://fr.wikipedia.org/wiki/Produit_laitier).

6. Epidémiologie et phylogéographie mondiales de la tuberculose

6.1 Epidémiologie mondiale de la tuberculose

Bien que l'incidence mondiale de la tuberculose ait tendance à décroitre, cette maladie représente un fléau menaçant pour de nombreuses populations. L'épidémie globale de tuberculose semble être liée à la pauvreté, la promiscuité, les conditions de vie précaires, ou d'autres aspects socio-démographiques et socio-économiques difficiles (**Farmer et coll., 2006** ; **Keshavjee et coll., 2008**).

La **Figure 9** schématise de manière très brève, quelques facteurs de risques pouvant interagir entre eux, ainsi qu'avec une ou plusieurs populations. L'exposition à ces facteurs peut favorablement déclencher une maladie chez un individu, et des études épidémiologiques (telles que des études cas-témoins ou études de cohortes) permettraient de quantifier ou de mesurer l'impact d'un ou de plusieurs de ces facteurs sur la survenue d'une maladie dans une population donnée. Les caractéristiques socio-économiques et démographiques jouent un rôle indubitable dans la survenue potentielle de maladies, et l'étude de ces différents paramètres corrélés aux analyses moléculaires apporte une vue d'ensemble sur la caractérisation précise d'un événement.

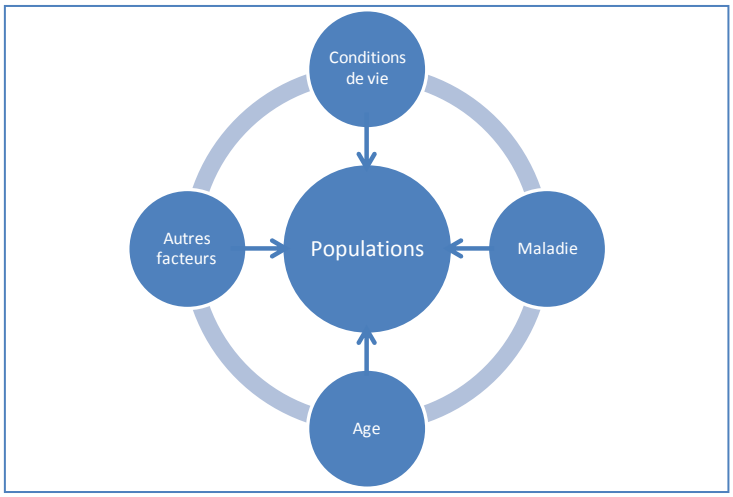


Figure 9 : Schéma représentant des facteurs de risques pouvant influer sur des populations données.

En 2008, l'incidence mondiale de la tuberculose fut estimée à 139 cas pour 100 000 habitants. Les pays pauvres sont clairement les plus touchés, et ils représentent 95% des nouveaux cas dans le monde.

La tuberculose est une maladie infectieuse liée à la pauvreté, la malnutrition, la surpopulation, et l'immunosuppression. Certains paramètres spécifiques comme les conditions difficiles de travail qui existaient durant la période de la révolution industrielle auraient facilité la transmission de la tuberculose (**Orcau et coll., 2011**).

L'OMS (ou WHO « World Health Organization »), a fixé des objectifs mondiaux (stratégie « Stop TB » et programme DOTS) afin d'inverser la tendance concernant l'incidence de la tuberculose, et de réduire de 50% le taux de mortalité (par rapport à 1990) (**WHO Report, 2013**).

En 2012, la tuberculose a conduit à un total de 8,6 millions de nouveaux cas estimés, et 1.3 million de morts (incluant 320000 morts parmi les patients co-infectés par le VIH/SIDA), et son taux de mortalité a régressé de 45% depuis 1990.

La tuberculose est répandue dans le monde entier, et même si certaines régions du monde sont faiblement touchées, la vigilance concernant cette constante menace, doit être maintenue. L'incidence de la tuberculose est très importante en Afrique, en Asie du Sud-Est et du Sud, en Amérique Latine, mais aussi dans les pays d'Europe de l'Est et la Russie, ainsi que dans la région caribéenne avec notamment Haïti et la République Dominicaine (**Figure 10**) (**WHO Report, 2013**).

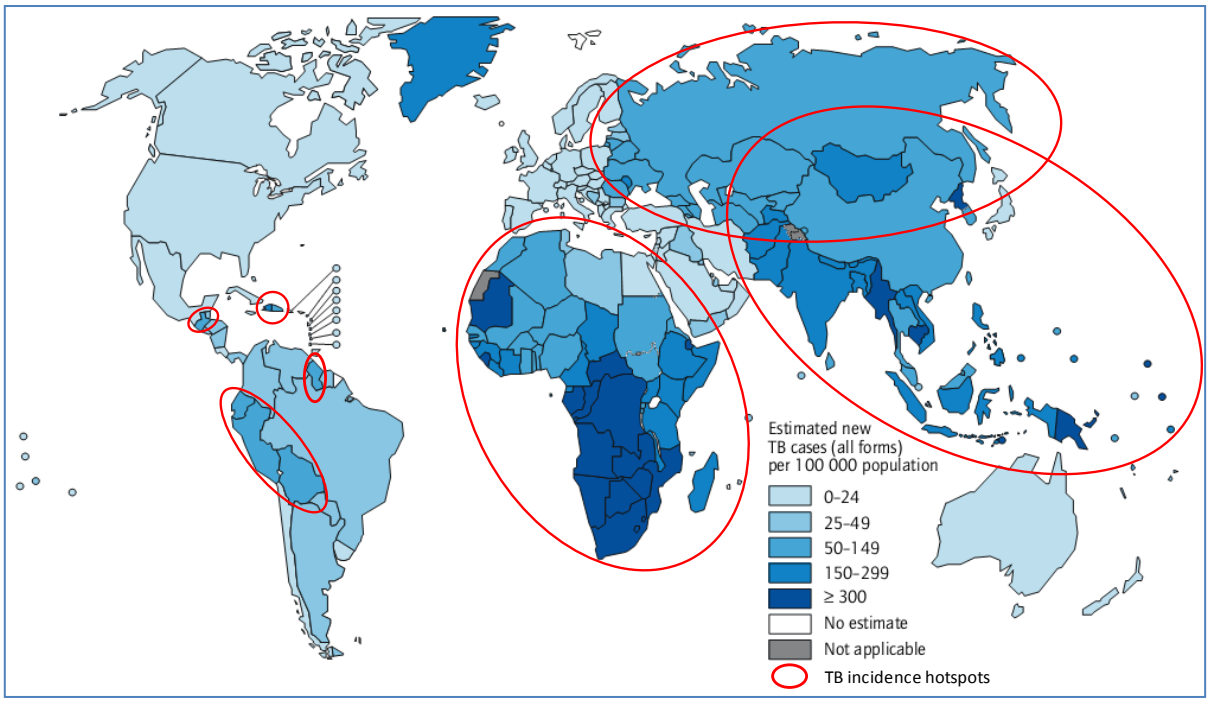


Figure 10 : Taux d'incidence estimé de la tuberculose en 2011 (**WHO Report, 2012**).

6.2 Quelques données démographiques et socio-économiques

Etudier la tuberculose reviendrait en quelque sorte à étudier conjointement les facteurs de risques associés à la maladie. En considérant des caractéristiques telles que l'âge, le sexe, le statut VIH, la densité de population, et le produit intérieur brut (PIB) des pays (entre autres), nous pouvons chercher à comprendre les possibles corrélations ou combinaisons pouvant s'ajouter au tableau de chasse de la tuberculose.

Cette analyse multidisciplinaire permet de renforcer les stratégies de lutte antituberculeuse, en proposant un ensemble d'approches pouvant s'imbriquer entre elles, et mesurer les différentes implications et corrélations possibles entre la maladie et les divers facteurs.

L'évolution des situations socio-économiques et démographiques des pays du monde, est un indice qui peut apporter une meilleure compréhension épidémiologique de la transmission/dissémination de la tuberculose dans ces pays.

Un lien peut certainement être établi entre la distribution globale de *M. tuberculosis* et plusieurs autres facteurs de risques tels que la moyenne d'âge globale selon les pays (http://en.wikipedia.org/wiki/Population_pyramid), la distribution des PIB de chaque pays (http://fr.wikipedia.org/wiki/Liste_des_pays_par_PIB_nominal), la densité de population (http://fr.wikipedia.org/wiki/Liste_des_pays_par_densit%C3%A9_de_population), et à la situation globale du virus VIH/SIDA (http://www.who.int/gho/hiv/en/).

Ainsi, les caractères socio-économiques et démographiques peuvent entrer en compte et jouer un rôle majeur dans l'étude épidémiologique de la tuberculose.

6.3 Phylogéographie mondiale de la tuberculose

La phylogéographie est l'étude des principes et processus qui gouvernent la distribution des lignées généalogiques ou phylogénétiques (**Avise et coll., 1987** ; **Avise et coll., 1998** ; **Kidd et Ritchie, 2006**). Elle étudie en particulier les phénomènes de spéciation conduisant à la distribution et à la structuration des populations.

Les principales méthodes de typage et d'épidémiologie moléculaire (qui seront abordées dans la partie suivante), nous ont permis d'identifier et de mettre en évidence les particularités des souches de tuberculose circulants dans le monde, et participent de ce fait, à la lutte anti-tuberculose.

Le complexe *Mycobacterium tuberculosis* (MTBC) constitue un groupe génétique homogène. Au sein des bases de données de l'Institut Pasteur de la Guadeloupe, les lignées phylogénétiques de MTBC sont assignées à l'aide du profil de spoligotypage des souches (Cf. Partie « Matériels et Méthodes »).

Les paragraphes suivants correspondent à une brève description des principales lignées phylogénétiques :

La lignée **AFRI** (*M. africanum*) est principalement retrouvée dans la région de l'Afrique de l'Ouest. Selon plusieurs études, la prévalence de cette lignée a tendance à décroitre avec le temps (**Groenheit et coll., 2011** ; **Koro Koro et coll., 2013**).

La lignée **BOVIS** (*M. bovis*) concerne généralement les espèces animales telles que les bovins. Bien que l'incidence globale de cette lignée soit peu élevée chez les humains, elle est tout de même répandue dans les 4 coins du globe (particulièrement en Europe) (**Smith et coll., 2011**).

La famille **Manu** est une famille récemment dénommée, qui pourrait être un clone ancestral du principal groupe génétique 1 (PGG1) (**Brudey et coll., 2006**). Elle est notamment présente en Inde, et en Egypte (**Helal et coll., 2009**). Cependant, la lignée « Manu » pourrait également provenir d'une infection mixte de souches de différentes familles (**Lazzarini et coll., 2012**).

La lignée **East African-Indian** (EAI), qui a été renommée « Indo-Oceanic » (**Gagneux et coll., 2006**), est fréquente dans les régions d'Afrique de l'Est, du Sud et du Sud-est asiatique (**Kremer et coll., 1999**). Elle a d'abord été décrite en Guinée-Bissau (**Källenius et coll., 1999**). La famille EAI est subdivisée en plusieurs sous-lignées (voir Annexes).

La lignée **Beijing** est aussi un membre du groupe PGG1, elle est caractérisée par une signature de spoligotypage spécifique (absence des espaceurs 1-34, et présence des espaceurs 35-43) (**van Soolingen et coll., 1995**). Cette lignée qui a été fortement associée aux souches de tuberculose multi-résistantes aux antimicrobiens (**Plikaytis et coll., 1994** ; **Bifani et coll., 2002**), est principalement originaire de l'Asie de l'Est (**Mokrousov et coll., 2005**), et présente une tendance à se répandre de manière hétérogène dans le reste du monde (en se basant notamment sur différents marqueurs génétiques tels que les MIRU-VNTRs) (**Iwamoto et coll., 2007** ; **Allix-Béguec et coll., 2014**).

La lignée **Central-Asian** (CAS) est une lignée particulièrement retrouvée dans le sous-continent indien (**Bhanu et coll., 2002** ; **Filliol et coll., 2003**). L'étude des différents marqueurs moléculaires (spoligotypes et MIRU-VNTRs) suggère que cette lignée pourrait être l'ancêtre de la famille Beijing (**Sola et coll., 2003**). Elle est subdivisée en plusieurs sous-lignées dont CAS1-Delhi, CAS1-Kili, et CAS2 (voir Annexes).

La lignée **Haarlem** a été décrite en 1999 au Pays-Bas (**Kremer et coll., 1999**), elle est principalement retrouvée dans les pays du Nord de l'Europe. Elle est aussi présente dans la Caraïbe, et en Afrique centrale où elle est supposée avoir été introduite durant la colonisation européenne (**Filliol et coll., 2003**).

La lignée **Ural** a été particulièrement bien décrite dans une étude récente (**Mokrousov, 2012**), et est prédominante dans les régions qui bordent la chaine de montages de l'Oural (la Russie, les pays de l'ex URSS), mais aussi en Iran.

La lignée **Latin American and Mediterranean** (LAM), comme son nom l'indique, concerne plutôt les pays d'Amérique latine, les pays bordant la Méditerranée, mais elle est aussi fréquente dans les autres pays d'Afrique (**Brudey et coll., 2006** ; **Demay et coll., 2012**).

La lignée **Cameroon** (précédemment nommée LAM10-CAM), ne faisant plus partie de la lignée LAM, est une lignée endémique au Cameroun, mais elle est aussi présente dans d'autres pays d'Afrique Centrale et de l'Ouest, ainsi que dans la Caraïbe (**Koro Koro et coll., 2013** ; **Niobe-Eyangoh et coll., 2004**).

La lignée **Turkey** (précédemment appelée LAM7-TUR), était aussi jadis considérée comme une LAM. Cette lignée est spécifique de la Turquie, mais on la retrouve également dans d'autres pays (notamment les pays voisins de la Turquie comme la Bulgarie) (**Zozio et coll., 2005** ; **Abadia et coll., 2010** ; **Kisa et coll., 2012**).

La lignée **X** semble être une famille liée aux pays anglo-saxons (**Brudey et coll., 2006** ; **Demay et coll., 2012**). On la retrouve surtout en Afrique du Sud, aux USA, et dans la Caraïbe.

La lignée **S**, bien que sa proportion soit faible dans nos bases de données, est prédominante en Europe du Sud (particulièrement en Sicile et en Sardaigne), et on la retrouve aussi en Afrique du Nord et du Sud (**Warren et coll., 2002** ; **Sola et coll., 2001**).

Les isolats de la famille **T** font partie d'un groupe de lignées mal définies (« ill defined »). Ce groupe est néanmoins inclus dans le groupe des lignées dites « Euro-américaines » (LAM, X, Haarlem, S, et T) (**Gagneux et coll., 2006**). Il serait intéressant de se poser la question de l'origine exacte de cette famille T mal définie.

La distribution globale des lignées phylogénétiques majeures de MTBC contenues dans notre base de données SITVIT2 est représentée dans la **Figure 11**.

Les outils et méthodes algorithmiques et de reconstruction phylogénétique permettant d'établir des liens entre les diverses familles de souches de tuberculose seront plus amplement abordés dans la partie « Matériels et Méthodes ».

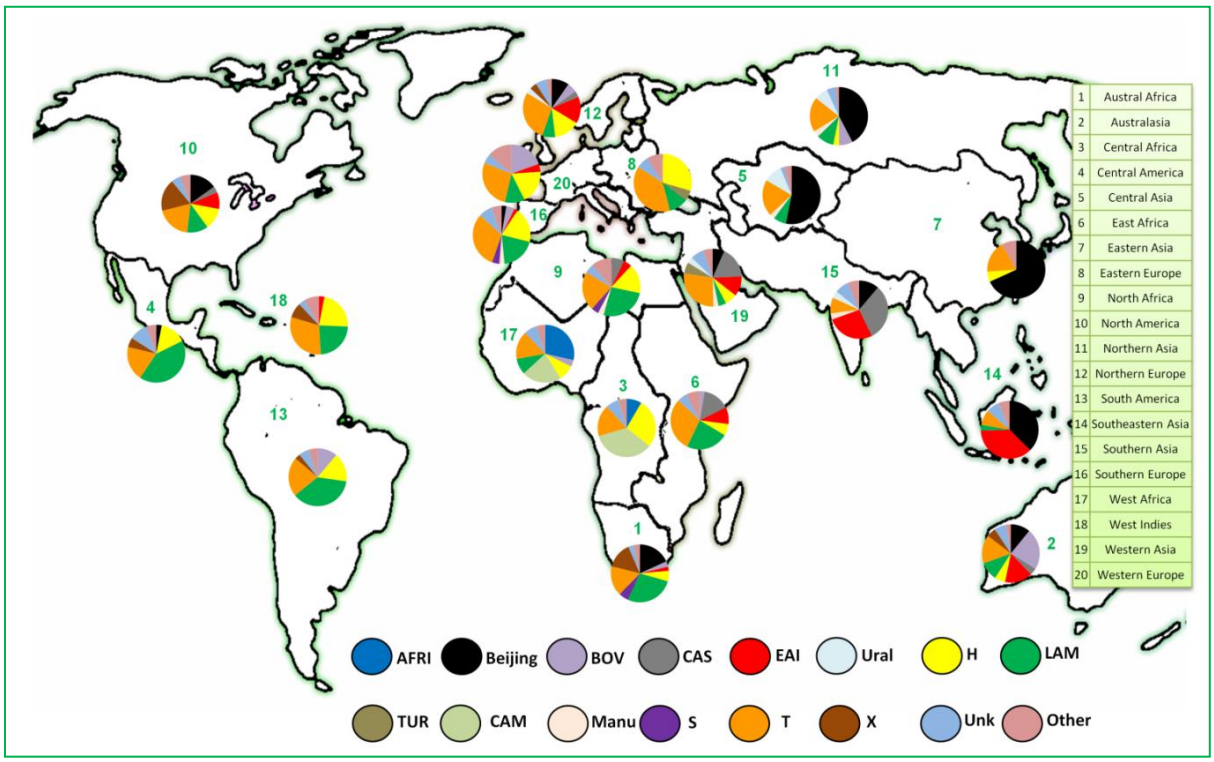


Figure 11 : Distribution géographique des grandes familles génotypiques de *M. tuberculosis*.

6.4 Méthodes de génotypage et classification des souches bactériennes

L'évolution des entités biologiques et notamment microbiologiques est une question qui continue à alimenter la recherche de principes et de modèles qui expliqueraient la variation des espèces. Les grandes théories de l'évolution telles que la « sélection naturelle » de Charles Robert Darwin (1809-1882) (**Darwin, 1859**), ou la « Philosophie zoologique » de Jean-Baptiste de Lamarck (1744-1829) (**Lamarck, 1809**), ont contribué à des avancées scientifiques notables, et ont révolutionné de manière générale la biologie. De plus, les travaux de Louis Pasteur (1822-1895) reconnus notamment dans le domaine de la microbiologie, ont participé à l'essor de cette discipline à l'échelle mondiale.

Le complexe *M. tuberculosis* représente un groupe génétique fortement homogène, et plusieurs membres de MTC partagent en moyenne plus de 99.7% d'identité nucléotidique (**Kato-Maeda et coll., 2001**). Pour différencier les souches du MTBC, la réaction en chaine par polymérase ou « polymerase chain reaction » (PCR) est utilisée. C'est une technique de biologie moléculaire permettant de dupliquer l'ADN qui a révolutionné le domaine de la recherche. Cette technologie a fait l'objet d'un prix Nobel de Chimie entre 1993 et 1994, et représente l'une des découvertes les plus importantes du XXème siècle (**The PCR Revolution, 2010**). Elle est notamment très utilisée dans les laboratoires de microbiologie, et particulièrement pour le diagnostic de maladies infectieuses.

Les méthodes actuelles de génotypage moléculaire, basées sur la PCR (**Jagielski et coll., 2014**), et les innovations menées dans le domaine de l'épidémiologie moléculaire de la tuberculose, ont permis et permettent une meilleure classification des isolats cliniques de bacilles tuberculeux (**García de Viedma et coll., 2011**).

En dehors de leur utilisation en épidémiologie (**García de Viedma et coll., 2011**), les méthodes de typage moléculaire sont également utiles pour les études évolutives (**Rastogi et Sola, 2007**). Il s'agit essentiellement des méthodes suivantes (Cf. Partie « Matériels et Méthodes »), qui sont plus ou moins utilisées pour les études épidémiologiques et phylogénétiques :

- le typage par IS6110-RFLP (« Restricted Fragment Length Polymorphism »), qui a longtemps été considérée comme une technique de référence (« gold standard ») pour le typage de *M. tuberculosis* (**van Embden et coll., 1993**);
- le spoligotypage (**Kamerbeek et coll., 1997**) ;
- le typage MIRU-VNTRs (**Supply et coll., 2006**) ;
- les « Large Sequence Polymorphisms » (LSP) (**Gagneux et coll., 2006**) ;
- les Régions de différence (RD) ;
- les « Single Nucleotide Polymorphisms » (SNP) (**Gutacker et coll., 2006**).

Le polymorphisme d'un seul nucléotide ou SNP (« Single Nucleotide Polymorphism ») est une technique qui a pu être mise en œuvre grâce à la génomique comparative. Il existe différents types de SNPs : (i) les SNPs régulateurs influants sur la synthèse protéique ; (ii) les SNPs synonymes, pour lesquels plusieurs formes d'un même SNP mènent à la même séquence polypeptidique ; (iii) les SNPs non synonymes, pour lesquels les séquences produites diffèrent. Nous pouvons aussi noter que les SNPs non synonymes sont impliqués dans la résistance aux médicaments antituberculeux. Cette résistance est presque toujours associée à des mutations ponctuelles (mutations non synonymes, petites délétions et

duplications), dans des loci chromosomiques spécifiques (**Jagielski et coll., 2014**). De plus, les SNPs permettent d'identifier des groupes de clusters phylogénétiquement distincts (**Fillol et coll., 2006**), et leur utilité pour la compréhension de la TB a été démontrée. Aussi, le besoin de la réalisation d'une base de données basée sur les SNPs s'est déjà fait ressentir (**Stucki et Gagneux, 2013**).

L'efficacité des RD/LSPs dans l'identification et l'analyse phylogénétique des membres du complexe *M. tuberculosis* est avérée. La distinction entre les LSPs qui se produisent une fois par rapport à celles qui se produisent de façon répétée dans une région génomique peut fournir des indications sur les rôles biologiques des LSPs et identifier des marqueurs phylogénétiques utiles (**Alland et coll., 2007**). Les RD/LSPs ont aussi permis de mettre en évidence 6 lignées phylogéographiques importantes : (i) Indo-Oceanic, (ii) East-Asian, (iii) East-African-Indian, (iv) Euro-American, (v) West-African 1, (vi) West-African 2 (**Figure 12**) (**Gagneux et coll., 2006**).

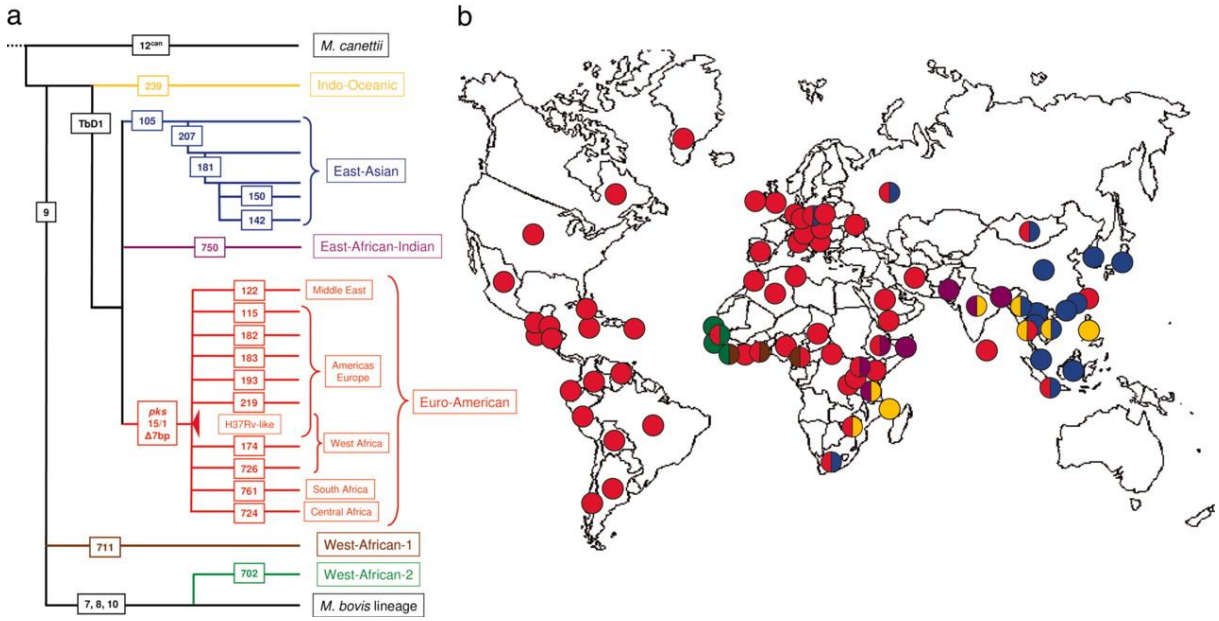


Figure 12 : Distribution géographique de MTBC. (a) arbre phylogénétique, et (b) cartographie basée sur les RDs/LSPs. Les couleurs indiquent les différentes lignées phylogéographiques (**Gagneux et coll., 2006**).

La distinction entre trois groupes génétiques de *M. tuberculosis* a été réalisée sur la base de deux polymorphismes se produisant à haute fréquence dans les gènes codants la catalase-peroxydase et la sous-unité « A » de la gyrase, conduisant à la classification de trois principaux groupes génétiques (PGG); les bactéries du groupe 1 étant ancestrales à celles des groupes 2 et 3 (**Sreevatsan et coll., 1997**).

En utilisant une collection mondiale de souches de MTBC et 212 SNPs, Filliol et coll. ont identifié six principales ramifications, les groupes de clusters SNP phylogénétiquement distincts (SCGs), et cinq sous-groupes (**Filliol et coll., 2006**). Les SCGs étaient fortement associés à l'origine géographique des échantillons de *M. tuberculosis* et le lieu de naissance des hôtes humains.

En parallèle, à l'aide de puces à ADN pour identifier de manière compréhensive les LSPs, une association stable entre les souches de MTBC et leurs hôtes humains a également été observée (**Hirsh et coll., 2004**) (**Tableau 2**).

Tableau 2 : Comparaison de différentes nomenclatures de lignées de *M. tuberculosis*.

Spoligotyping-based (Filliol 2003)	PGG	SCG (Filliol 2006)	SNP-based (Gutacker 2006)	LSP (Gagneux 2006)
East-African-Indian (EAI)	PGG1	SCG 1	sSNP-I	Indo-Oceanic (Lineage 1)
Beijing	PGG1	SCG 2	sSNP-II	East-Asian (Lineage 2)
Central-Asian (CAS)	PGG1	SCG 3a	sSNP-IIA	East-African-Indian (Lineage 3)
Haarlem	PGG2	SCG 3b	sSNP-III	Euro-American (Lineage 4)
X1	PGG2	SCG 3c	sSNP-IV	Euro-American (Lineage 4)
X1,X2,X3	PGG2	SCG 4	sSNP-V	Euro-American (Lineage 4)
LAM	PGG2	SCG 5	sSNP-VI	Euro-American (Lineage 4)
T (Miscellaneous)	PGG2-3	SCG 6	sSNP-VII sSNP-VIII	Euro-American (Lineage 4)
Bovis	PGG1	SCG 7	(MTBC)	(MTBC)
M. africanum	PGG1	NA	NA	West-African 1 (Lineage 5)
M. africanum	PGG1	NA	NA	West-African 2 (Lineage 6)

Aujourd'hui, de nouvelles approches sont en phase de révolutionner le domaine de l'épidémiologie moléculaire. Les nouvelles générations de séquençages (NGS) et le séquençage complet du génome (« Whole Genome Sequencing », WGS), sont des techniques de pointe qui commencent à devenir indispensables pour l'analyse moléculaire et épidémiologique de la tuberculose (**van Soolingen, 2014**).

7. Diagnostic, transmission, traitement et moyens de lutte

7.1 Diagnostic et transmission

L'agent pathogène *M. tuberculosis* est à l'origine soit d'une maladie évolutive (« tuberculose maladie ») ou d'une infection latente (« tuberculose infection »). Un tiers de la population mondiale serait infectée par une tuberculose latente. La transmission se fait par voie aérienne par inhalation de gouttelettes de salive contenant le bacille pathogène. Seules les personnes atteintes d'une « tuberculose maladie » sont contagieuses (par leur toux).

Le test cutané à la tuberculine, ou test de Mantoux (intradermoréaction), peut permettre de dépister la tuberculose en révélant la primo-infection (première contamination d'un individu par une mycobactérie du complexe *M. tuberculosis*). Il consiste à injecter 0,1 mL de solution de tuberculine (à 10 Unité Internationales) dans le derme. Cependant depuis la généralisation du BCG, ce test a beaucoup moins de valeur pour dépister la tuberculose latente. Suspecté sur la clinique et la radiographie pulmonaire, un cas de tuberculose doit être confirmé par des examens bactériologiques.

L'examen direct (ED) (qui consiste en un frottis) d'expectoration constitue l'examen le plus rapide en terme de rendu décisionnel au médecin soignant. Cet examen direct est l'examen privilégié par l'OMS dans la lutte contre la tuberculose car il nécessite peu de moyens et peut être réalisé même dans les régions reculées. Cependant sa sensibilité n'est pas excellente et dépend largement de l'opérateur. Dans les pays développés, cet examen direct est systématiquement accompagné d'une mise en culture qui est l'examen de référence. La culture permet d'identifier précisément la bactérie mais aussi d'effectuer un antibiogramme qui vise à tester la sensibilité des souches bactériennes vis-à-vis de plusieurs antibiotiques.

Depuis quelques années, des méthodes moléculaires permettent de faire un diagnostic de tuberculose sur les prélèvements primaires. Toutes ne sont pas très fiables.

7.2 Les antibiotiques antituberculeux et la résistance

Le traitement de la « tuberculose maladie » peut être relativement long et coûteux. L'OMS a notamment défini des schémas thérapeutiques normalisés adaptés en fonction du type de tuberculose rencontré mais aussi en fonction des systèmes de santé des pays qui auront à les appliquer et de leurs moyens financiers disponibles. Ces schémas normalisés présentent l'avantage de (i) réduire les risques d'apparition d'une pharmaco-résistance, (ii) réduire les coûts et faciliter l'estimation des besoins en médicaments, (iii) favoriser un approvisionnement régulier en médicaments, et (iv) simplifier la formation du personnel.

Les antibiotiques antituberculeux (ou simplement antituberculeux) sont sélectionnés en fonction des cas. Pour un patient qui n'a jamais été traité antérieurement, il convient de lui prescrire des antituberculeux de première ligne tels que : isoniazide (INH), rifampicine (RIF), pyrazinamide (PZA), et ethambutol (ETH). Pour éviter le développement de souches résistantes aux antibiotiques il est obligatoire d'utiliser une multi-thérapie (utilisation de plusieurs antibiotiques) pour élargir le spectre d'activité et renforcer l'effet bactéricide, ainsi que pour limiter l'émergence de résistances aux antibiotiques. Les antituberculeux de seconde ligne sont les suivants : les fluoroquinolones et des antibiotiques injectables de la famille des aminosides (capreomycine, kanamicine, amikacine).

La résistance aux antibiotiques antituberculeux représente actuellement un problème majeur qui vient déstabiliser les efforts de lutte anti-TB. Bien que l'incidence globale de la tuberculose soit en baisse, les souches multi-résistantes (MDR) et extrêmement résistantes (XDR) ont tendance à augmenter de manière inquiétante. Le **Tableau 3** présente de manière succincte les différents antibiotiques antituberculeux de première ou de seconde ligne, leurs concentrations minimales inhibitrices (CMI), et les mécanismes de résistance.

Tableau 3 : Concentrations minimales inhibitrices (« minimum inhibitory concentration ») et mécanismes de résistance des antituberculeux de première et seconde ligne (**Zhang et Yew, 2009**).

Drug (year of discovery)	MIC µg/ml	Gene(s) involved in resistance	Gene function	Role	Mechanism of action	Mutation frequency %
Isoniazid (1952)	0.02–0.2	*kat*G *inh*A	Catalase-peroxidase Enoyl ACP reductase	Pro-drug conversion Drug target	Inhibition of mycolic acid biosynthesis and other multiple effects	50–95 8–43
Rifampicin (1966)	0.05–1	*rpo*B	β subunit of RNA polymerase	Drug target	Inhibition of RNA synthesis	95
Pyrazinamide (1952)	16–50 (pH 5.5)	*pnc*A	Nicotinamidase/pyrazinamidase	Pro-drug conversion	Depletion of membrane energy	72–97
Ethambutol (1961)	1–5	*emb*B	Arabinosyl transferase	Drug target	Inhibition of arabinogalactan synthesis	47–65
Streptomycin (1944)	2–8	*rps*L *rrs* *gid*B	S12 ribosomal protein 16S rRNA rRNA methyltransferase (G527 in 530 loop)	Drug target Drug target Drug target	Inhibition of protein synthesis	52–59 8–21 ?
Amikacin/kanamycin (1957)	2–4	*rrs*	16S rRNA 16S rRNA	Drug target	Inhibition of protein synthesis	76
Capreomycin (1960)		*tly*A	2′-O-methyltransferase			
Quinolones (1963)	0.5–2.5	*gyr*A *gyr*B	DNA gyrase subunit A DNA gyrase subunit B	Drug target	Inhibition of DNA gyrase	75–94
Ethionamide (1956)	2.5–10	*eta*A/*eth*A	Flavin monooxygenase	Prodrug conversion	Inhibition of mycolic acid synthesis	37
		*inh*A		Drug target		56
PAS (1946)	1–8	*thy*A	Thymidylate synthase	Drug activation?	Inhibition of folic acid and iron metabolism?	36

MIC = minimum inhibitory concentration; ACP = acyl carrier protein; PAS = para-aminosalicylic acid.

Il existe plusieurs phénotypes de résistance : les souches bactériennes peuvent être soient (i) sensibles aux antituberculeux ; (ii) résistantes à un ou plusieurs antituberculeux n'incluant pas simultanément l'INH et la RIF ; (iii) multi-résistantes ou MDR incluant simultanément la résistance à l'INH et la RIF, avec ou non la résistance à d'autres antibiotiques ; (iv) extrêmement résistantes ou XDR définis comme MDR avec la résistance additionnelle à un antibiotique de la familles des fluoroquinolones, et au moins un des antibiotiques antituberculeux injectables (capréomycine, kanamycine, amikacine). La recherche des mécanismes qui conduisent à la résistance aux antibiotiques (**Figure 13**), est un sujet de grande importance qui n'a pas fini de susciter beaucoup d'intérêt.

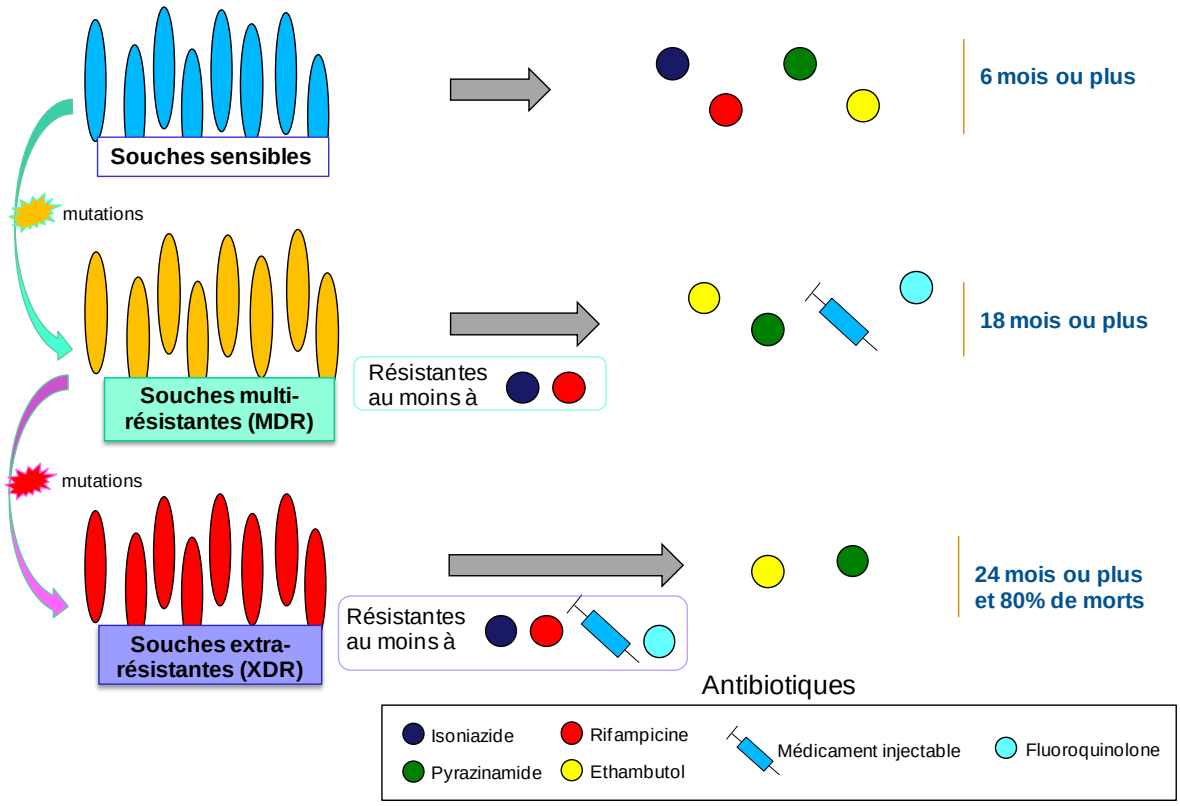


Figure 13 : Schéma simplifié des mécanismes et traitements de la tuberculose résistante aux antibiotiques.

De nouvelles méthodes de traitement sont donc urgemment souhaitées pour venir à bout de cette forme particulière de tuberculose. Une étude récente innovante a notamment montré la possibilité de traiter la tuberculose résistante avec des plantes (**Madikizela et coll., 2014**).

7.3 Le vaccin BCG

En 1886, un premier vaccin fut expérimenté par Vittorio Cavagnis (**Gazette hebdomadaire de médecine et de chirurgie. - 1887**) tandis qu'à cette même époque Robert Koch tenta vainement de développer un sérum curatif basé sur la tuberculine (**Gradmann, 2001**). En 1902, Behring essaya un vaccin contre la tuberculose bovine à partir d'un bacille d'origine humaine atténué : le bovovaccin. Behring proposa également, la tuberculase (http://www.bium.univ-paris5.fr/histmed/asclepiades/pdf/sbalcerowiak.pdf**)**, qui ne connut pas un réel succès. Toujours dans le domaine vétérinaire, Koch essaya le tauruman. Pour mémoire, il faut aussi citer le sérum de Marmorek (1904) (**Guy Grenier, 2002**), le sérum de Maragliano, et les sérums de Richet et Héricourt.

C'est en 1921, qu'Albert Calmette et Camille Guérin de l'Institut Pasteur de Lille essayent avec succès le premier vaccin contre la tuberculose sur lequel ils travaillaient depuis 1908 (conçu à la base pour être un vaccin vétérinaire). Baptisé BCG (Bacille de Calmette et Guérin ou Bilié de Calmette et Guérin), ce vaccin issu d'une souche vivante atténuée de

Mycobacterium bovis deviendra obligatoire en France en 1950 (http://fr.wikipedia.org/wiki/Tuberculose).

De nos jours, l'efficacité de ce vaccin est remise en cause, et un nouveau vaccin largement applicable et efficace est désespérément souhaité pour le contrôle de la tuberculose (**Kaufmann et coll., 2014**).

7.4 La stratégie Stop TB et le programme DOTS

La stratégie Stop TB (ou Stratégie Halte à la Tuberbulose) est une stratégie mise en place par l'Organisation Mondiale de la Santé (OMS), visant à obtenir une réduction spectaculaire des cas de tuberculose d'ici à 2015. Cette stratégie prône également la mise au point de nouveaux outils efficaces pour une meilleure prévention et un meilleur traitement de la tuberculose, ainsi que la disparition à long terme de cette maladie (http://www.who.int/tb/strategy/stop_tb_strategy/fr/).

La démarche DOTS se situe au cœur de la stratégie Halte à la tuberculose. Cette stratégie tente de renforcer la lutte antituberculeuse à l'aide de cinq éléments principaux : (i) Volonté politique accompagnée d'une augmentation et d'une pérennisation des financements ; (ii) Dépistage des cas par des examens bactériologiques de qualité assurée ; (iii) Traitement normalisé avec surveillance et soutien des patients ; (iv) Système efficace d'approvisionnement en médicaments et de gestion pharmaceutique ; (v) Système de suivi et d'évaluation et mesure des effets (http://www.who.int/tb/dots/fr/).

8. Migrations de populations

L'époque dans laquelle nous vivons est particulièrement marquée par un essor significatif de la migration des populations du globe. De 1950 à 2012, le nombre de déplacements pour le loisir et le tourisme a littéralement explosé, passant de 25 millions à plus de 1 milliard (http://www2.unwto.org/). De plus, on observe beaucoup de déplacements pour raisons économiques de populations venant des pays en voie de développement immigrant vers les pays plus industrialisés.

Le monde est donc en mouvement constant, les populations migrent pour le tourisme, le travail, ou pour fuir la pauvreté (http://www.unfpa.org/pds/migration.html). Nous pouvons prendre l'exemple du bassin Méditerranéen avec le drame récent de Lampedusa (http://fr.wikipedia.org/wiki/Naufrage_du_3_octobre_2013_%C3%A0_Lampedusa), qui témoigne bien de la situation critique à laquelle doivent faire face les autorités publiques, politiques et les services de santé. On peut également prendre l'exemple de la Caraïbe (**Figure 14**) qui représente une véritable plaque tournante économique et de trafics illicites.

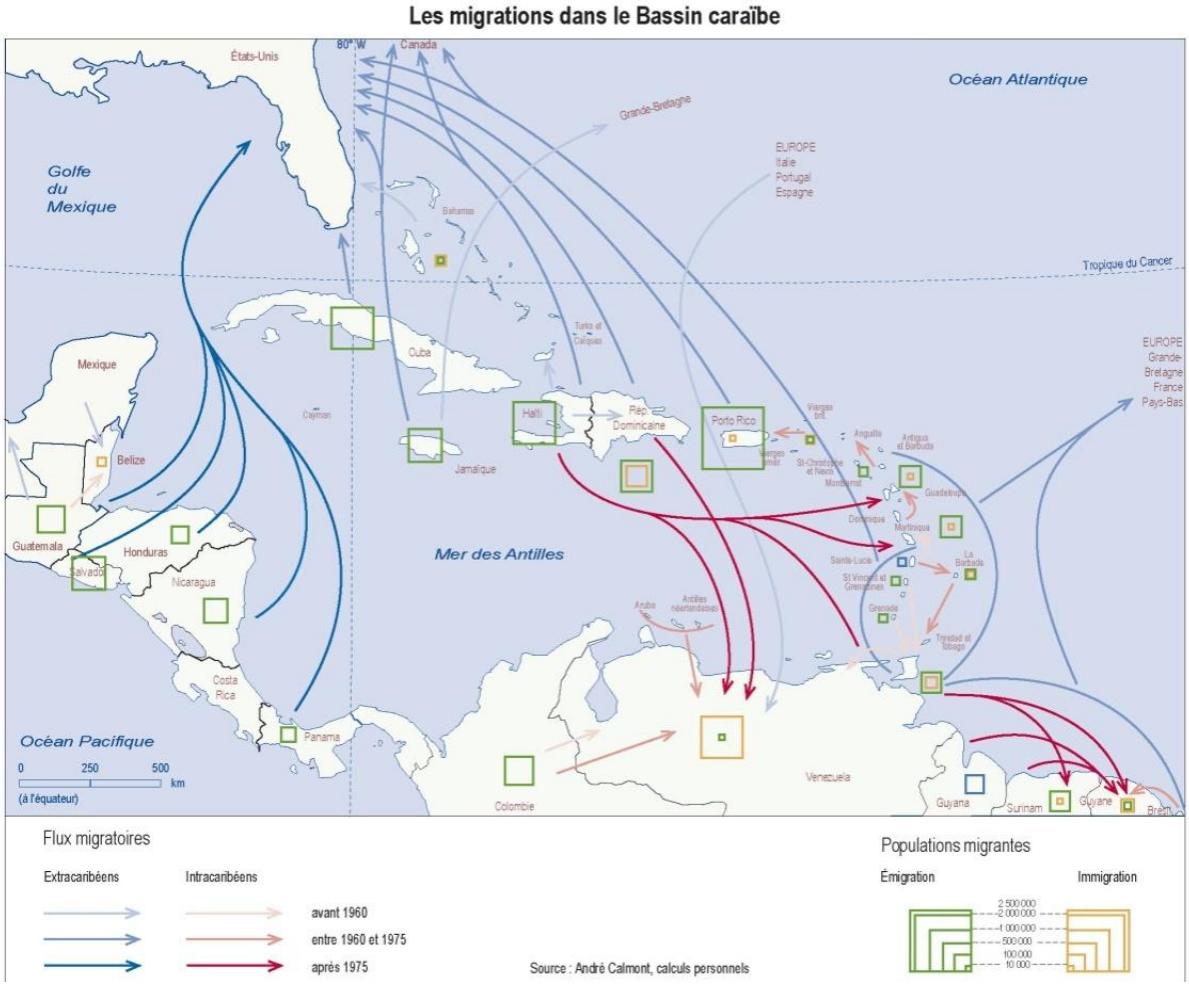


Figure 14 : migrations de populations dans le bassin Caribéen (http://www1.univ-ag.fr/aihp-geode/page5/page17/page40/page59/page59.html).

Cet espace caribéen serait aussi propice à la dispersion de la tuberculose, qui est très diversifiée en fonction des îles ou régions (**Millet et coll., 2014**).

Lorsque nous jetons un coup d'œil sur les migrations historiques qui se sont déroulées au cours du temps (https://genographic.nationalgeographic.com/human-journey/ ; http://en.wikipedia.org/wiki/Human_migration), nous pouvons tisser des liens étroits entre les différentes explorations, colonisations, ou conquêtes avec la dispersion globale de la tuberculose (**Hirsh et coll., 2004** ; **Gagneux et coll., 2006** ; **Wirth et coll., 2008** ; **Hill et coll., 2012**). A l'heure actuelle, les cas de tuberculose observés dans les pays « riches » sont majoritairement dus aux populations migrantes qui sont le plus souvent infectées par les génotypes prédominants dans leur pays d'origine (**Vanhomwegen et coll., 2011** ; **Fallico et coll., 2014**). Il est difficile de répertorier de manière précise toutes les migrations et tous les mouvements de populations qui se sont déroulés au cours du temps. Nous pouvons évoquer les diverses explorations, la route de la soie, la découverte du nouveau monde, la progression des empires, les croisades, le commerce triangulaire, les colonisations et décolonisations, et les différentes guerres locales ou mondiales, qui pourraient avoir un rôle potentiel dans la dissémination de la tuberculose au niveau mondial. **La Figure 15** témoigne des migrations actuelles des populations mondiales qui cherchent à fuir la pauvreté, ou faire du tourisme.

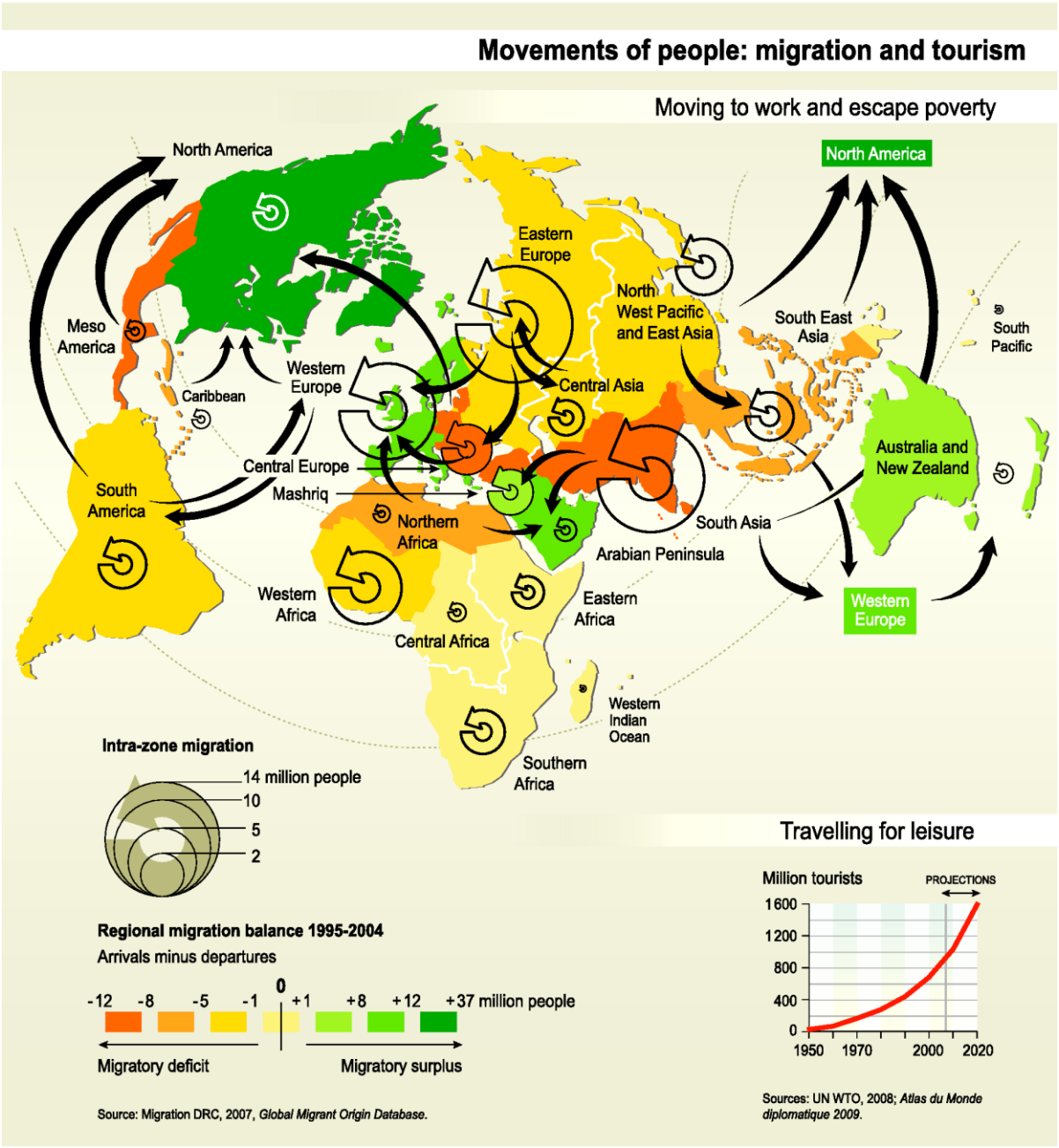


Figure 15 : Les migrations mondiales (http://www2.unwto.org/).

9. L'utilité des Bases de Données pour la recherche sur la tuberculose

De nos jours, les bases de données représentent des outils essentiels pour le contrôle et la surveillance des maladies ou d'autres systèmes. Ces bases permettent le stockage de grandes quantités d'informations, de manière structurée, facilitant le traitement des données. Les bases de données s'inscrivent aussi dans le panel des outils contribuant à la réalisation des Objectifs du Millénaire pour le développement (OMD) visant à réduire la pauvreté et la maladie au niveau mondial (https://www.un.org/fr/millenniumgoals/).

Les bases de données sont similaires aux patrimoines historiques ou aux monuments que nous devons constamment maintenir et affermir, dans cette nouvelle ère du « Big Data », où l'information stockée peut atteindre une taille de plus en plus volumineuse (à l'ordre du Téra octet ou du Péta octet). De plus, les bases de données en ligne sont devenues d'importantes interfaces de publication de données biologiques (**Howe et coll., 2008**).

Ces collections de données, de plus en plus développées dans la science et le domaine de la santé publique, permettent de stocker et d'analyser les informations dont le nombre est en augmentation rapide.

Nous pouvons très brièvement citer les bases de données, dédiées à la recherche sur la tuberculose, qui seront décrites plus en détail dans la partie « Matériels et Méthodes » : SpolDB4 (**Brudey et coll., 2006**), SITVITWEB (**Demay et coll., 2012**), SpolTools (**Reyes et coll., 2008** ; **Tang et coll., 2008**), MIRU-VNTRplus (**Allix-Béguec et coll., 2008** ; **Weniger et coll., 2010**), TB GIMS (**CDC MMWR, 2010**), Mbovis.org (**Smith et Upton, 2012**), MycoDB.es (**Rodriguez-Campos et coll., 2012**), TB-lineage (**Shabbeer et coll., 2012**), tbvar (**Joshi et coll., 2013**), InTB (**Soares et coll., 2013**), PolyTB (**Coll et coll., 2014**), et « Genome-wide *Mycobacterium tuberculosis* variation » (GMTV) (**Chernyaeva et coll., 2014**).

10. SITVIT : Une Base de Données en constante évolution

La première base de données a été lancée, il y a plus de quinze ans à l'Institut Pasteur de la Guadeloupe (IPG), lorsqu'un stagiaire nommé Jérôme Maïsetti a pris l'initiative d'entrer les « motifs » de spoligotypes disponibles à partir de nos propres isolats des Caraïbes (n=218 souches), dans une feuille de calcul Excel, et de les mettre en commun avec les données publiées des isolats provenant d'autres pays (n=392 souches).

Une fois les spoligotypes triés, nous avons réalisé que l'on pouvait non seulement définir des motifs prédominants, mais aussi retracer l'origine des souches et leurs mouvements potentiels. C'est ainsi que la base de données de 610 spoligotypes provisoirement nommée SpolDB1 est née, et a conduit à la première description de 69 spoligotypes majeurs afin de mieux comprendre l'origine de la tuberculose et sa transmission (**Sola et coll., 1999**). Le développement de SpolDB1 a été suivi par le lancement de SpolDB2 contenant des données sur 3319 isolats (**Sola et coll., 2001**), puis SpolDB3 avec 13008 isolats regroupés en 813 « shared-types » (contenant 11708 isolats) et 1300 profils orphelins (**Filliol et coll., 2002** ; **Filliol et coll., 2003**). Plus récemment, le développement de la quatrième

version SpolDB4 (n=39295 isolats) en 2006 (**Brudey et coll., 2006**), et de SITVITWEB (n=62582 isolats) en 2012 (**Demay et coll., 2012**), a permis d'avoir un meilleur aperçu phylogéographique de la circulation des lignées génotypiques de MTBC dans le monde entier. La dernière version mise à jour de notre base de données nommée SITVIT2 (n=111635 isolats), contient des informations de génotypage sur environ deux fois plus de souches que dans la version précédente. Il convient de souligner que les deux dernières versions SITVITWEB et SITVIT2 sont multi-marqueurs, contenant à la fois des données de spoligotypage et des MIRU-VNTRs (« mycobacterial interspersed repetitive units-variable number of tandem repeats »). Dans ces versions récentes, une interface Web permet à l'utilisateur de rechercher des souches selon des critères tels que l'année d'isolement, le pays d'isolement de la souche, le pays d'origine du patient, ou le nom du contributeur; ainsi que les recherches combinées supplémentaires dans SITVIT2, permettant aussi d'obtenir les données sur la résistance aux médicaments, les caractéristiques démographiques et d'autres informations épidémiologiques. La nouvelle base qui est en cours de publication ne cessera pas de s'enrichir de nouvelles données de génotypage de MTBC, grâce à une interface consacrée aux membres de l'unité Tuberculose et Mycobactéries de l'IPG, provisoirement appelée « SITVITAdmin ».

11. La Bioinformatique d'aujourd'hui

La bioinformatique est un champ de recherche multidisciplinaire pouvant être associée à des domaines aussi variés que la biologie, la médecine, l'informatique, les mathématiques, la chimie, ou la physique. Le terme « bioinformatique » est apparu en 1970 dans une publication de Paulien Hogeweg et Ben Hesper (université d'Utrecht, Pays-Bas) (**Hesper et Hogeweg, 1970**), en référence à l'étude des processus d'information dans les systèmes biotiques (**Attwood et coll., 2011**).

La biologie et/ou la microbiologie sont des domaines complexes en constante évolution. De nouvelles méthodes et techniques apparaissent jour après jour, pour analyser, interpréter et modéliser les données biologiques. La bioinformatique fait donc le lien entre les 2 domaines (biologie et informatique) en essayant de faciliter l'interprétation et le traitement des informations, en proposant des algorithmes plus ou moins complexes permettant par exemple de mieux comprendre les mécanismes génétiques favorisant le cancer colorectal, de réaliser la comparaison ou l'alignement des séquences d'ADN, de construire des arbres phylogénétiques, ou encore de modéliser l'interaction des gènes ou protéines (**Figure 16**).

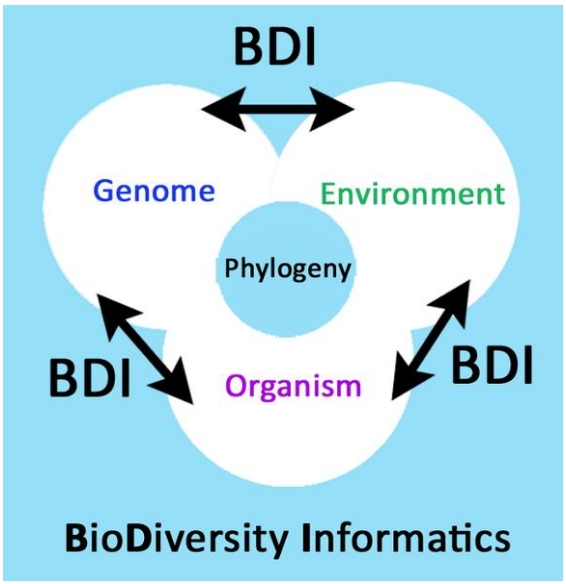


Figure 16 : Schéma représentant les transformations de la biologie évolutive dues à l'augmentation des données disponibles sur la variation dans les génomes, les organismes et l'environnement (**Losos et coll., 2013**).

« La biologie évolutive est dopée par les progrès de l'informatique qui offrent un accès à un nombre croissant de données sur la diversité et les variations des gènes, des génomes des organismes et de l'environnement. Tout cela peut être relié à l'arbre de vie (phylogénie), des populations à des clades entiers, via de nouveaux protocoles et réseaux dans le domaine de l'informatique de la biodiversité » (**Losos et coll., 2013**).

Nous verrons donc au cours de ce travail, comment la bioinformatique nous a permis d'assembler, en partie, les pièces du « puzzle gigantesque » de *Mycobacterium tuberculosis*.

Les méthodes de bio-statistique sont très utilisées dans les domaines de recherche médicale et d'épidémiologie. Ces techniques permettent de représenter des données de manière quantitative ou qualitative, en donnant une certaine valeur ou un certain poids à l'analyse ou la comparaison des données. De nombreux outils et tests statistiques sont utilisés dans ce travail de thèse afin de montrer les différences significatives et les liens pouvant exister entre plusieurs variables (facteurs de risque potentiels) en relation avec la tuberculose.

12. Objectifs de ce travail

M. tuberculosis est un pathogène largement répandue dans le monde entier. Malgré une baisse notable de son incidence dans de nombreux pays, la tuberculose est toujours un problème majeur de santé publique. Des réponses concertées et coordonnées sont encore nécessaires pour contrôler la maladie de manière globale.

A travers ce travail, nous essaierons de répondre à quelques questions concernant l'épidémiologie moléculaire et la surveillance mondiale de la tuberculose. Cette thèse s'appuie essentiellement sur deux thématiques principales :

- Le développement, l'implémentation, et le maintien des bases de données SITVIT et outils Web pour mieux caractériser, visualiser, contrôler, et faciliter l'interprétation des isolats cliniques de MTBC. Les bases de données créées sont basées sur le système de gestion de base de données relationnelle (SGBDR) MySQL.
- Les études épidémiologiques et phylogéographiques utilisant diverses caractéristiques démographiques ou cliniques (telle que la résistance aux antituberculeux). Ces études ont pour but d'établir une cartographie mondiale ou régionale des clones circulants de *M. tuberculosis*.

Les données génotypiques qui ont été enregistrées, au fil du temps, dans les bases de données de l'Institut Pasteur de la Guadeloupe, permettent de dresser un portrait précis de l'épidémiologie mondiale de la tuberculose. Cette thèse apporte une amélioration structurelle des bases de données, en mettant en place de nouvelles architectures favorisant la gestion des données. L'intégration de nouvelles variables spécifiques aux données épidémiologiques et démographiques (sexe, âge, statut VIH, origine) des patients donne un meilleur aperçu et facilite l'interprétation du schéma relationnel existant pour la base de données SITVIT2. Cette initiative se révèle très utile pour la surveillance épidémiologique de la tuberculose et l'étude des chaînes de transmission de la maladie au niveau local, régional et international. De plus, l'utilisation des nombreux outils Web en conjonction avec les études épidémiologiques diverses ou les données contenues dans SITVIT2, permet de visualiser et cartographier les isolats cliniques de *M. tuberculosis* en fonction de plusieurs caractéristiques cliniques (résistance aux antibiotiques) ou démographiques.

Les résultats obtenus dans ce travail de thèse mettent en relief différents aspects de l'épidémie mondiale de la tuberculose. Les bases de données et les diverses cartographies adaptées aux analyses épidémiologiques nous donnent une vue d'ensemble de MTBC dans le but de mieux comprendre et surveiller sa propagation mondiale.

MATERIELS ET METHODES

1. Méthodes de génotypage et techniques basées sur la PCR

La caractérisation des clones circulant de *Mycobacterium tuberculosis* est essentielle pour comprendre les modalités de fonctionnement et de transmission de la bactérie. L'identification de l'espèce par les examens bactériologiques ne suffit pas pour les recherches épidémiologiques auxquelles nous nous intéressons, il faut identifier les souches d'un foyer de contamination. Les méthodes de typage basées sur la PCR (« Polymerase Chain Reaction ») représentent une révolution pour l'épidémiologie moléculaire de la tuberculose. Les bactéries appartenant au complexe *M. tuberculosis* ont un génome bien conservé mais certaines régions sont polymorphiques, et elles représentent les cibles des marqueurs moléculaires. Le typage par IS6110-RFLP (« Restricted Fragment Length Polymorphism ») (**van Embden et coll., 1993**) a été remplacé par d'autres techniques plus facile à mettre en œuvre telles que le spoligotypage (**Kamerbeek et coll., 1997**) et le typage MIRU-VNTR (**Supply et coll., 2006**). Il existe aussi d'autres méthodes comme les « Large Sequence Polymorphisms » (LSP) (**Gagneux et coll., 2006**), les Régions de différence (RD), et les « Single Nucleotide Polymorphisms » (SNP) (**Gutacker et coll., 2006** ; **Filliol et coll., 2006**).

Le laboratoire de recherche sur la tuberculose et les mycobactéries de l'Institut Pasteur de la Guadeloupe utilise principalement 2 méthodes de génotypage : le spoligotypage ou spoligotyping (« spacer oligo nucleotide typing »), et le génotypage MIRU-VNTR (« Mycobacterial Interspersed Repetitive Units – Variable Number Tandem Repeats »). Voyons de manière un peu plus détaillée, l'utilité de ces marqueurs moléculaires et de ces méthodes de génotypage.

1.1 L'IS6110-RFLP

IS6110 est un élément d'insertion retrouvé exclusivement dans le complexe *M. tuberculosis* (**van Embden et coll., 1993**). Cet élément a un unique site de restriction pour l'enzyme *Pvu*II. L'IS6110-RFLP est une méthode de référence jadis considérée comme un « gold standard », cependant, elle fut remplacée par la suite par d'autres méthodes plus rapides, moins laborieuses et moins couteuses. La technique IS6110 est fréquemment utilisée pour discriminer les souches au sein des clusters principaux (**Figure 17 et 18**).

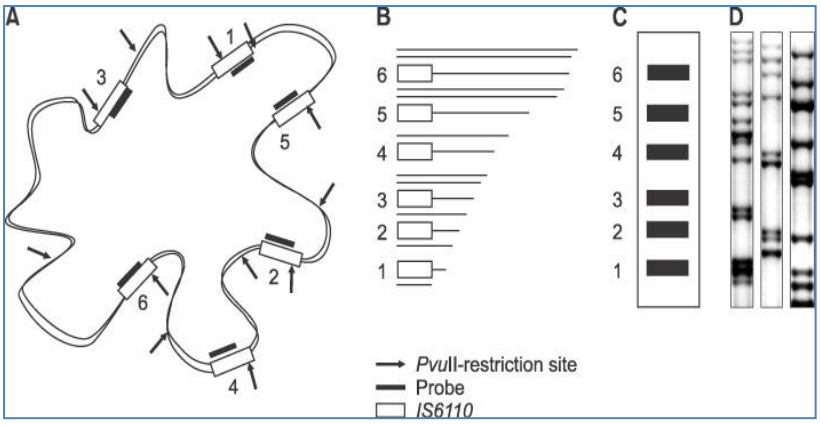


Figure 17 : Principes et procédures de l'IS6110-RFLP (**Jang et coll., 2011**). (**A**) Obtention de IS6110-RFLP par digestion enzymatique (*Pvu*II) ; (**B**) Fragments d'ADN séparés sur gel d'agarose, et transféré sur membrane de nylon ; (**C**) Hybridation avec un fragment de séquence d'IS6110 ; (**D**) image réelle d'un motif d'IS6110-RFLP.

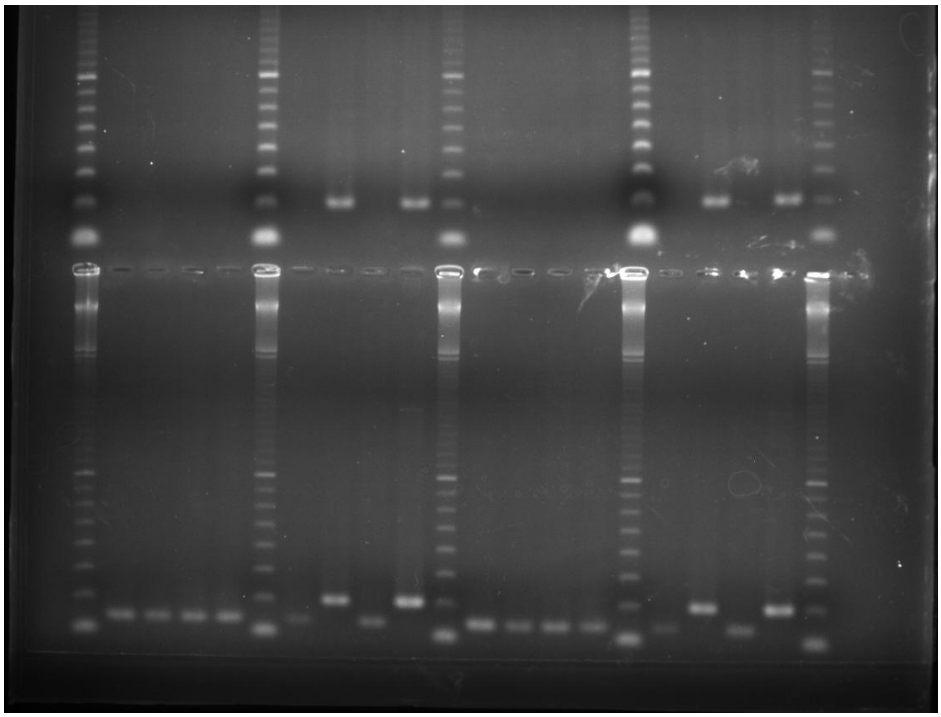


Figure 18 : Photo de gel de souche de la famille/lignée LAM, correspondant au typage IS6110.

1.2 Le spoligotypage ou « spacer oligo nucleotide typing » (spoligotyping)

Le spoligotypage est une méthode basée sur une région d'ADN polymorphique appelée locus DR (« Direct Repeat ») (**Kamerbeek et coll., 1997**). Cette région est présente dans le génome des bactéries du complexe *M. tuberculosis*, et est composée d'un nombre variable de répétitions directes (DVRs). Le spoligotypage est une méthode, basée sur la PCR, très fréquemment utilisée en épidémiologie moléculaire et en phylogéographie du MTBC. Initialement identifié dans la souche vaccinale de *M. bovis* BCG (**Hermans et coll., 1991**), le locus DR appartient à la famille de séquences d'ADN répétées de type CRISPR (« clustered regularly interspaced short palindromic repeat »), et contient des séquences répétées parfaites identiques de 36 pb entrecoupées par des espaceurs uniques de 34-41 pb. Les DVRs présentent un important polymorphisme parmi les isolats cliniques de *M. tuberculosis*. Le spoligotypage43 révèle la présence ou l'absence de 43 espaceurs pour une souche donnée (**Figure 19**). Le motif binaire généré par cette méthode a déjà montré de grandes capacités à véhiculer des informations phylogénétiques importantes (**Sola et coll., 2001**). Le **Tableau 4** indique comment passer du code binaire au code octal.

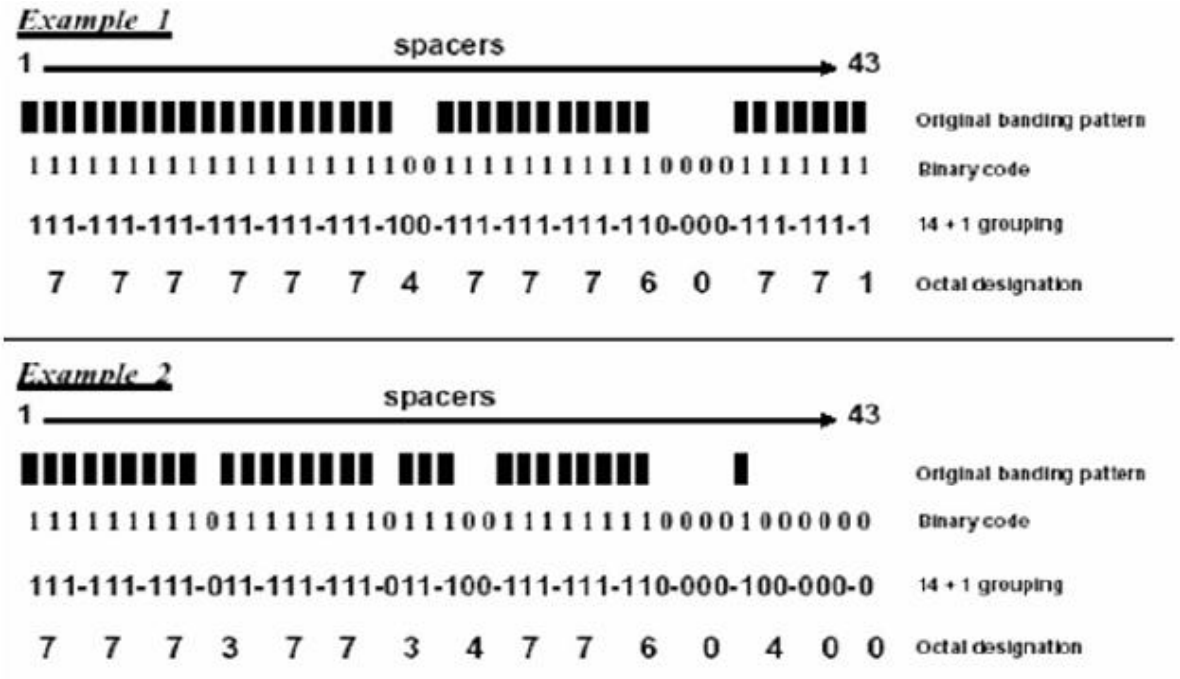


Figure 19 : Exemple de spoligotypage, et passage du code binaire au code octal (http://www.cdc.gov/tb/).

Tableau 4 : Equivalence entre le code binaire et le code octal.

Valeur binaire	Valeur octale
000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

1.3 MIRU-VNTR typing

Le typage MIRU-VNTRs (« mycobacterial interspersed repetitive units – variable number of tandem repeats ») est une technique basée sur la présence de courtes séquences d'ADN répétées dans le génome des mycobactéries, et fut développée par Philip Supply (**Supply et coll., 2001** ; **Supply et coll., 2006**). Les mini-satellites MIRU-VNTR sont constitués de marqueurs indépendants tels que (i) le classique 12-loci MIRU (ou MIRU12), (ii) le 15-loci MIRU (MIRU15), (iii) le format complet 24-loci MIRU (MIRU24), et (iv) les 5-loci « exact tandem repeats » (ETRs, ou VNTR) (**Frothingham et coll., 1998**).

Au sein de la base de données SITVIT, les MIRU-VNTRs sont identifiés par des codes alphanumérique, les lettres A, B, C, etc….., représentent successivement 10, 11, 12, etc……. nombres de copie (**Tableau 5**).

Tableau 5 : Exemple de codage alphanumérique d'un motif MIRU12.

Locus (MIRU12)	2	4	10	16	20	23	24	26	27	31	39	40
Nombre de répétitions	2	2	11	5	2	5	2	4	2	4	3	10
Codage	2	2	B	5	2	5	2	4	2	4	3	A

La **Figure 20** schématise les différents marqueurs utilisant les loci MIRU-VNTRs. Certains loci sont plus ou moins discriminants en fonction de la famille phylogénétique d'une souche donnée de MTBC.

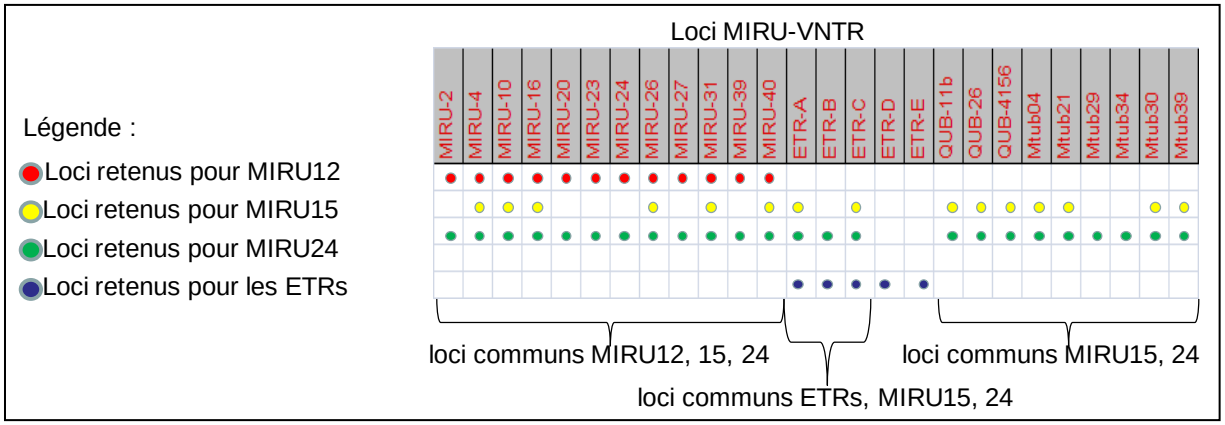


Figure 20 : Schéma représentant les loci impliqués pour chaque marqueur MIRU-VNTRs.

Le **Tableau 6** indique la correspondance entre les différentes désignations MIRU-VNTRs et les loci VNTR associés. Ces loci spécifiques sont présents dans le génome de la souche de référence de *M. tuberculosis* H37Rv (**Figure 21 et 22**).

La méthode de typage basée sur les MIRU-VNTRs est de plus en plus utilisée dans les études épidémiologiques se consacrant à caractériser les souches de MTBC. Cette technique d'identification est plus discriminante que le spoligotypage, les MIRU-VNTRs ayant une horloge moléculaire plus rapide que celle des spoligotypes. Cependant, il peut y avoir homoplasie entre ces marqueurs (similitude entre 2 motifs provenant de chemins évolutifs différents), bien que le spoligotypage serait plus enclin à l'homoplasie que le typage MIRU-VNTRs (**Comas et coll., 2009**).

Spoligotypage et typage MIRU-VNTRs sont les nouveaux « gold standard » dans le typage moléculaire de *M. tuberculosis*, et le 24-loci MIRU (MIRU24) est la méthode la plus adaptée, particulièrement pour mener des études de qualité. Mais pour réaliser une étude vraiment robuste, il est conseillé d'utiliser conjointement d'autres marqueurs génétiques tels que les IS6110, les SNPs, les RD/LSPs.

Tableau 6 : Correspondance entre les désignations MIRU-VNTRs et les loci VNTR.

Désignation MIRU-VNTRs	Loci VNTR
MIRU-2	VNTR-0154
MIRU-4	VNTR-0580
MIRU-10	VNTR-0960
MIRU-16	VNTR-1644
MIRU-20	VNTR-2059
MIRU-23	VNTR-2531
MIRU-24	VNTR-2687
MIRU-26	VNTR-2996
MIRU-27	VNTR-3007c
MIRU-31	VNTR-3192
MIRU-39	VNTR-4348
MIRU-40	VNTR-0802
ETR-A	VNTR-2165
ETR-B	VNTR-2461
ETR-C	VNTR-0577
QUB-11b	VNTR-2163
QUB-26	VNTR-4052
QUB-4156	VNTR-4156
Mtub04	VNTR-0424
Mtub21	VNTR-1955
Mtub29	VNTR-2347
Mtub30	VNTR-2401
Mtub34	VNTR-3171
Mtub39	VNTR-3690

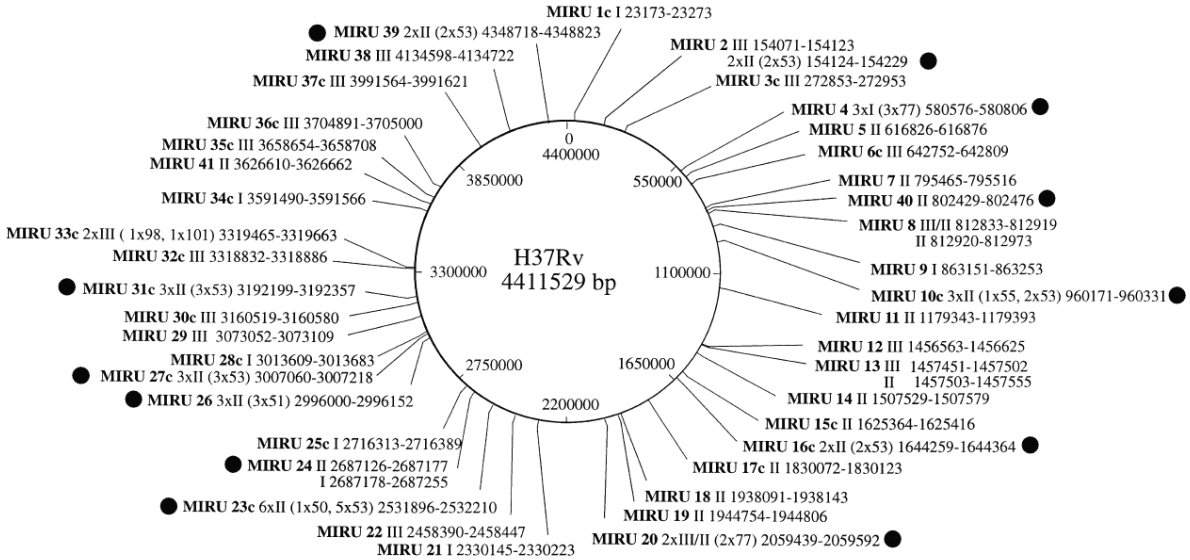


Figure 21 : Position des 41 loci MIRU sur le génome de *M. tuberculosis* H37Rv (**Supply et coll., 2001**). Le nom des loci apparait en gras, le caractère « c » est affecté aux MIRUs présentant une orientation inverse à celle définie par Cole et al. (1998). Les caractères romains indiquent le type du MIRU (type I, II, ou III). La position de chaque locus est aussi indiquée (loci VNTR représentés par des nombres), et les points noirs révèlent les MIRUs du format 12.

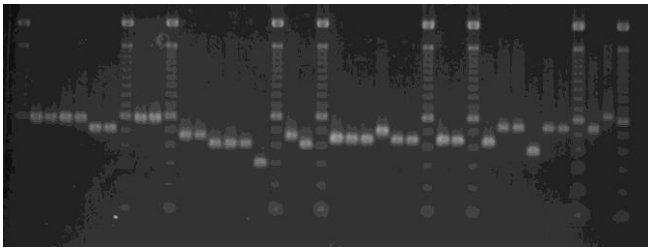


Figure 22 : Photo de gel (agarose) correspondant au typage MIRU-VNTR.

La **Figure 23** schématise le chromosome hypotétique d'un isolat de MTBC avec les éléments cibles des différentes méthodes de typage.

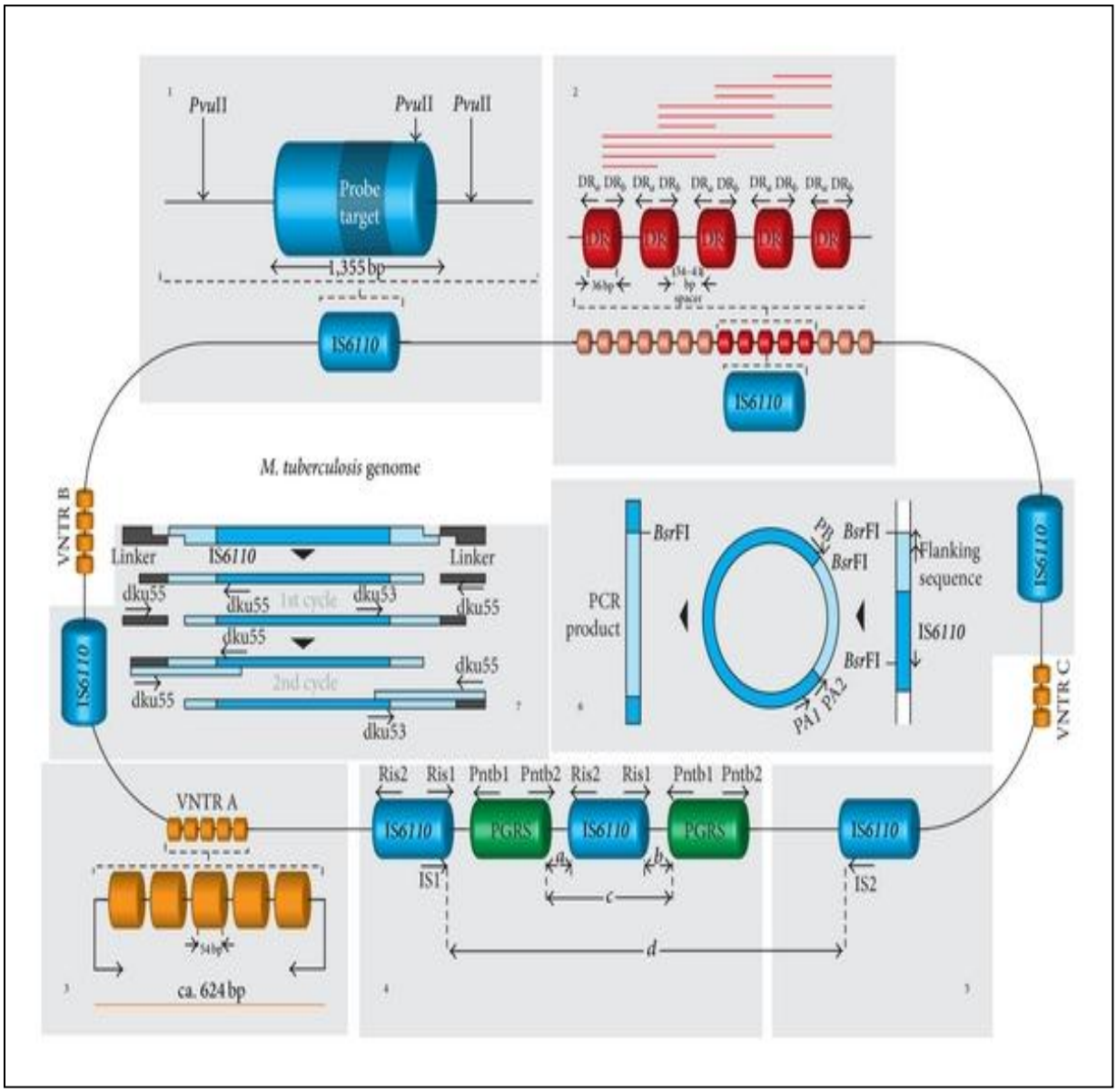


Figure 23 : Représentation schématique des éléments répétitifs utilisés comme cibles pour différentes méthodes de génotypage du complexe M. tuberculosis : (1) IS6110-RFLP (2) Spoligotypage (3) les minisatellites MIRU-VNTRs (4/5) DRE-PCR (6) PCR inverse semi nichée (7) Ligation Mediated-PCR (**Jagielski et coll., 2014**).

2. Drug Susceptibility Typing (DST)

La résistance aux anti-tuberculeux représente un challenge mondial auquel doit faire face les services de santé publique. Un diagnostique correcte des formes de résistances des souches, est donc primordial pour un meilleur traitement de la maladie.

On dénote plusieurs formes de résistances aux anti-tuberculeux, et une codification bien précise a été mise en œuvre dans la base de données SITVIT2 afin de bien identifier ces résistances :

- Le Code 1 désigne les souches « Pansusceptibles », c'est-à-dire non-résistantes aux antibiotiques.

- Le Code 2 représente les souches de tuberculose multi-résistantes ou MDR (résistance simultanée à l'isoniazide et la rifampicine, avec ou sans résistance aux autres antibiotiques).

- Le Code 3 indique une résistance à au moins un anti-tuberculeux (n'importe lequel), mais n'incluant pas simultanément l'isoniazide (INH) et la rifampicine (RIF).

- Le Code 4 est utilisé pour les souches dites ultra-résistantes ou XDR, correspondant aux souches MDR avec une résistance additionnelle à un fluoroquinolone, et une résistance à au moins un des médicaments injectables (capréomycine, kanamycine et amikacine).

3. Tests, Outils, et Lois de Biostatistiques

Les biostatistiques sont un moyen efficace d'estimer la validité ou la qualité d'un travail de recherche scientifique. Afin de mener à bien nos études épidémiologiques, différents tests et méthodes statistiques ont été utilisés par l'intermédiaire de logiciels reconnus. Les tests les plus communément appliqués à travers ce travail de thèse, ont été les suivants :

- Le test du Chi-carré (Chi^2 ou Khi^2) a été proposé par le statisticien Karl Pearson en 1900 (**Stigler, 2008**). Cette technique statistique permet de tester la différence ou la dépendance entre plusieurs variables, et de dire si cette différence est statistiquement significative ou non. D'autres principes statistiques tels que l'hypothèse nulle (H_0) ou le risque alpha (α) permettent de déterminer la significativité de la différence entre variables (http://fr.wikipedia.org/wiki/Test_du_%CF%87%C2%B2).
- Le test exact de Fisher, qui doit son nom à son inventeur Ronald Fisher, est un test similaire au test de Chi^2, et contrairement à ce dernier, le « Fisher's exact test » est non-paramétrique, et peut être appliqué avec des nombres d'échantillon peu élevés.
- Le test t de Student permet de comparer des groupes de variables quantitatives comme par exemple la distribution des âges des patients dans 2 populations distinctes.

On peut privilégier l'utilisation d'un test par rapport à un autre, en fonction du type de variable (variable quantitative ou variable qualitative) ou en fonction de la taille des

échantillons étudiés. Le **Tableau 7** récapitule les principaux tests statistiques, en fonction des variables et/ou groupes de populations étudiées.

Tableau 7 : Les différents tests statistiques, paramétriques ou non, en fonction des types de variables.

		Variable continue		Variable binaire	
		Test paramétrique	**Test non paramétrique**	**Grands échantillons**	**Petits échantillons**
2 groupes	Indépendants	Test *t* de Student	Test de Mann-Whitney	Test de Chi²	Test exact de Fisher
	Appariés	Test *t* de Student apparié	Test de Wilcoxon	Test de McNemar	Test exact de McNemar
Plus de 2 groupes	Indépendants	Analyse de variance	Test de Kruskal-Wallis	Test de Chi²	Test exact de Fisher
			Test de Cuzick (tendance)	Test de tendance	
1 groupe 2 variables continues		Coefficient de corrélation de Pearson	Coefficient de corrélation de Spearman		

La valeur P qui peut être calculée à la suite des différents tests statistiques, est la probabilité que l'hypothèse nulle H_0 (i.e. il n'y a pas de grande différence entre les variables) soit vraie. Le seuil de cette valeur P est généralement fixé à 5% ou 0,05, et lorsque notre valeur P est inférieure à 0,05 on peut rejeter l'hypothèse nulle (i.e. il y existe réellement une différence significative entre les variables). On peut également considérer une différence comme étant marginalement significative lorsque, par exemple, la valeur P est comprise entre 0,05 et 0,08 (ou d'autres intervalles de probabilités). Les valeurs P sont très utilisées dans les études statistiques ou épidémiologiques, cependant, beaucoup d'experts se sont interrogés sur la précision ou l'efficacité de ce paramètre statistique (**Alves et Yu, 2014**).

L'Odds Ratio (OR), aussi connu sous le nom de rapport des chances ou rapport des cotes est une mesure statistique très utilisée en épidémiologie, pour déterminer comme la valeur P, la dépendance entre les variables étudiées.

Les Intervalles de Confiance (IC) permettent de définir une marge d'erreur ou d'estimer une imprécision, en calculant une « fourchette » à partir de ce qui est observé sur un échantillon (c'est sur ce même principe que reposent les sondages d'opinion) (**Huguier et Flahault, 2000**). En général, on détermine un intervalle de confiance à 95%, ce qui signifie que la fourchette en question sera calculée afin de se donner 95 chances sur 100 pour que la « réalité » se situe entre les limites de la fourchette (on peut aussi se fixer des limites plus ou moins exigeantes) (**Huguier et Flahault, 2000**).

Parallèlement à ces analyses uni-variées, nous avons utilisé des analyses multi-variées pour bien mettre en évidence la dépendance et l'interaction des variables entre elles.

La régression logistique permet d'analyser la valeur explicative propre de chacune des covariables étudiées. Dans cette méthode statistique, la variable expliquée doit être qualitative à deux classes (i.e. binaire). Cependant, les covariables étudiées peuvent être quantitatives ou qualitatives (**Huguier et Flahault, 2000**). La régression logistique est donc un outil privilégié pour les analyses multidimensionnelles (ou multi-variées). Il existe plusieurs autres outils et méthodes statistiques tels que le modèle de Cox ou encore la régression multiple, qui sont

bien détaillés dans le livre de Michel Huguier et Antoine Flahault (**Huguier et Flahault, 2000**).

Les calculs de moyenne, médiane, ou écart type, nous ont également permis de tirer des conclusions significatives par rapport aux différentes études épidémiologiques. Les logiciels de statistique comme STATA12, MegaStat12, et R (version 2.14.1), ont été utilisés pour appliquer les différents tests statistiques à nos études épidémiologiques.

4. Index de Discrimination de Hunter-Gaston (HGDI) et Taux de transmission récente

L'**HGDI** (Index de Discrimination de Hunter-Gaston) a été souvent utilisé à travers l'outil V-DICE (disponible au lien : http://www.hpa-bioinformatics.org.uk/cgi-bin/DICI/DICI.pl), pour comparer les différents marqueurs de génotypage (MIRU-VNTRs ou Spoligotype) et déterminer les motifs les plus discriminants (**Hunter et Gaston, 1988**). Cet index est l'application du plus ancien « Simpson index of diversity » décrit en 1948. L'HGDI est décrit par l'équation suivante :

$$HGDI = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{S} n_j(n_j - 1)$$

Soit N le nombre total de souches, S le nombre total de motifs et n_j le nombre total de souches pour le $j^{ème}$ motif.

Le **taux de transmission récente** est une technique qui permet de détecter les cas de contamination et de transmission récente. Pour calculer ce taux, on émet l'hypothèse que tous les patients porteurs d'une même souche bactérienne ont été contaminés potentiellement par une même source. La notion de « grappe » en épidémiologie moléculaire est définie comme au moins 2 isolats cliniques présentant des profils/patterns génétiques strictement identiques. Il y a 2 méthodes de calcul pour estimer les cas en « grappe » :

- La méthode « n » tient compte de tous les cas en grappe d'une étude donnée, et est utilisée pour estimer le nombre de patients qui appartiennent à une chaine de transmission active (**Alland et coll., 1994**).
- La méthode « n-1 » admet qu'un cas par grappe soit indexé (cas index) sans être comptabilisé dans les cas en grappe (**Small et Behr, 1997**). On utilise cette méthode pour estimer les cas de transmission active dans les cas de réactivation de la maladie. Le taux de transmission récente se définit par la soustraction du *nombre de grappes* au *nombre de cas en grappe*, le tout divisé par le *nombre total de cas* (**Murray et Alland, 2002**).

5. Théorie des Graphes et Algorithmes de Reconstruction Phylogénétique

L'utilisation de certains principes de la théorie des graphes nous ont permis de mener à bien des études épidémiologiques pointues. Divers outils et logiciels ont rendus possible la construction de représentations schématiques facilitant l'interprétation de données parfois très denses. Les algorithmes suivant ont été le plus souvent utilisés pour une visualisation optimale de l'évolution des souches de *M. tuberculosis* :

- L'UPGMA (« Unweighted Pair Group Method with Arithmetic mean ») est un algorithme très utilisé pour la construction d'arbres phylogénétiques. L'UPGMA permet de transformer une matrice de distances (matrice répertoriant une distance arithmétique pouvant exister entre différents organismes) en un arbre enraciné.
- Le « Neighbor-Joining » est une méthode d'analyse de cluster (ou groupe) qui est aussi utilisée pour la construction d'arbres phylogénétiques. Cette méthode tient compte des différences de vitesse d'évolution entre les différentes branches de l'arbre phylogénétique. Cette technique fournit un arbre non polarisé (non enraciné) (**Saitou et Nei, 1987**).
- Le Maximum de Parcimonie (ou parcimonie de Wagner) est une méthode phylogénétique permettant de construire des arbres de classification hiérarchique après enracinement. Cette méthode part du principe que l'évolution est parcimonieuse puisque qu'elle requiert un minimum de changements évolutifs. L'arbre optimal serait celui qui minimise le nombre global de mutations, et qui privilégierait les branches les plus courtes. La parcimonie de Wagner repose sur 3 étapes : (i) la recherche de tous les cladogrammes possibles, (ii) la mesure de la longueur de chaque arbre et (iii) enfin la sélection de ceux présentant la plus petite longueur.
- Le « Minimum Spanning Tree » (MST, ou arbre couvrant de poids minimal) est un graphe non-dirigé et connecté (i.e. il existe un lien entre chaque nœud du graphe, aucun noeud n'est isolé). Le MST est un graphe reliant tous ses noeuds de façon à ce que le poids global de l'arbre (la distance entre les noeuds) soit le plus minimal possible. Au cours de ce travail de thèse, le MST a souvent été utilisé, à travers les logiciels BioNumerics et MLVA Compare (logiciels qui seront décrit au paragraphe 6.3), pour mettre en évidence les relations évolutionnaires existantes entre les souches du complexe *Mycobacterium tuberculosis*.
- Le « Hierarchical Layout » est une méthode de repésentation hiérarchique des noeuds d'un graphe, favorisant la lecture en empêchant le chevauchement des arêtes. Ce type de visualisation linéaire peut être notamment généré par l'application SpolTools (Cf. partie 6.1).
- L'arbre basé sur l'algorithme de Fruchterman-Reingold, utilisant les « Force-directed algorithms » (algorithmes de dessin basés sur les forces), est un graphe dirigé qui n'est pas nécessairement connecté. Cet arbre peut également être généré grâce à SpolTools. L'idée d'un algorithme basé sur les forces est de considérer une certaine force s'appliquant entre chaque paire de nœuds. L'algorithme de Fruchterman-Reingold est un algorithme permettant une représentation graphique rafinée, cet algorithme est

souvent utilisé pour représenter des graphes en 2D ou en 3D. Cependant, la limitation de cette technique serait que lorsque le nombre de nœuds est important, les arêtes délimitant ces nœuds peuvent se chevaucher.

Dans les études épidémiologiques qui vont être présentées dans la partie Résultats et Discussion, les 2 dernières méthodes « Hierarchical Layout » et « Fruchterman-Reingold » seront désignées en tant que « Spoligoforests ».

6. Bases de données, Logiciels de Phylogénie et outils Bionformatiques

6.1 Les bases de données et outils disponibles pour l'épidémiologie moléculaire de la TB

La tuberculose est un problème de santé publique majeur qui continue à susciter beaucoup d'intérêts (**WHO Report 2013**). Bien que l'incidence globale de la maladie a tendance à diminuer, nous devons toujours rester vigilant et garder un œil sur cette menace. Durant ces dernières années, nous avons pu assister à la création et au développement de nombreuses bases de données et outils web dédiés à l'épidémiologie moléculaire et l'évolution de la tuberculose. En voici quelques exemples :

(i) SpolTools est un outil Web de visualisation conçu pour manipuler et analyser des données de spoligotypage de MTBC (**Reyes et coll., 2008**; **Tang et coll., 2008**). Cet outil contient également un répertoire en ligne des isolats spoligotypés collectés dans la littérature publiée (i.e. 30 jeux de données contenant 1179 motifs de spoligotypage correspondant à 6278 isolats cliniques). SpolTools permet de dessiner des arbres Spoligoforest (utlisant les algorithmes de « Fruchterman-Reingold » et le « Hierarchical Layout ») qui illustrent les relations évolutives et les liens de parenté entre spoligotypes dans une étude donnée. Cette application est aussi utilisée pour le double typage CRISPR de *Corynebacterium diphtheriae* (**Mokrousov, 2009**). SpolTools est disponible à l'adresse suivante : http://www.emi.unsw.edu.au/spolTools/.

(ii) MIRU-VNTR*plus* est un ensemble d'outils Web dédié à l'analyse des données de typage moléculaire sur la tuberculose, en particulier les formats 12-, 15- et 24-loci MIRU-VNTRs (**Allix-Béguec et coll., 2008**; **Weniger et coll., 2010**). Cette collection d'outils permet la recherche de souches similaires, la création d'arbres phylogénétiques (MST, UPGMA, Neighbor-Joining), et la cartographie des informations géographiques. En outre, la base de données de MIRU-VNTR*plus* fournit également des résultats détaillés (l'origine géographique, les profils de sensibilité aux médicaments, lignées génétiques, les motifs de spoligotypage, les profils SNP et LSP, et les empreintes digitales d'IS6110-RFLP) sur une collection de 186 souches de référence bien caractérisées. MIRU-VNTR*plus* est disponible à l'adresse : http://www.miru-vntrplus.org/.

(iii) TB Genotyping Information Management System (TB GIMS) est un système sécurisé basé sur le Web, conçu pour améliorer l'accès et la diffusion de l'information de génotypage à l'échelle nationale États-Unienne (**MMWR, 2010**). Il stocke et gère les données de génotypage des patients atteints de tuberculose aux États-Unis. Ce système permet aux utilisateurs autorisés d'enregistrer et de suivre les isolats de MTBC pour et à partir de laboratoires de génotypage spécifiques (sous contrat). L'outil TB GIMS informe immédiatement les laboratoires et les programmes travaillant sur la TB des résultats de génotypage obtenus et des mises à jour effectuées ; il corrèle les données des isolats avec les données de surveillance obtenues au niveau des patients ; il génère des rapports sur les grappes de génotypes, notamment sur la répartition des génotypes à l'échelle nationale ; et il fournit des cartes des grappes de génotypes à l'échelle du pays, des états ou des comtés. Cette base de données n'est pas accessible au public.

(iv) Mbovis.org est une base de données contenant plus de 1400 motifs spoligotypes appartenant aux lignées phylogénétiques de MTBC ayant comme particularité la région RD9 supprimée : *M. africanum*, *M. bovis* (antilopes), *M. microti*, *M. pinnipedii*, *M. caprae* et *M. bovis* (**Smith et Upton, 2012**). Cette base de données est disponible à l'adresse suivante : http://www.mbovis.org/.

(v) MycoDB.es est une base de données espagnole dédiée à la tuberculose animale (**Rodriguez-Campos et coll., 2012**), qui a été créée comme un outil épidémiologique au niveau national (focalisé sur l'Espagne). Elle contient 401 motifs de spoligotypes différents correspondant à 17273 isolats appartenant à *M. bovis*, *M. caprae* et *M. tuberculosis*. MycoDB.es contient aussi une quantité limitée de données MIRU-VNTRs. Malheureusement, cette base de données est limitée à l'agence espagnole de la santé animale - Centro de Vigilencia Sanitaria Veterinaria (VISAVET) : http://www.vigilanciasanitaria.es/mycodb/.

(vi) TB-Lineage est un outil en ligne pour la classification phylogénétique et l'analyse des génotypes de MTBC, utilisant les spoligotypes et éventuellement le locus MIRU 24 (**Shabbeer et coll., 2012**). Il a été développé et testé à l'aide des données de génotypage fournies par les CDC (« Centers for Disease Control and Prevention », centres épidémiologiques) d'Atlanta sur 37066 isolats cliniques correspondant à 3198 motifs de spoligotypage et 5430 motifs MIRU-VNTRs. Cependant, si les données du locus MIRU 24 ne sont pas disponibles, le système utilise les prédictions faites par un classificateur bayésien naïf basé seulement sur les données de spoligotypage. La précision de la classification automatique en utilisant les deux marqueurs spoligotype et MIRU24 est supérieure à 99%. En utilisant le spoligotypage seul, la précision de classification est supérieure à 95%. TB-Lineage est disponible en ligne à l'adresse : http://tbinsight.cs.rpi.edu/run_tb_lineage.html. Cette application web fournit également un outil permettant de générer des spoligoforests afin de visualiser la

diversité génétique et la parenté des génotypes de MTBC ainsi que leurs lignées phylogénétiques. SITVIT-KBBN est un sous-projet faisant partie de TB-Lineage qui a permis la prédiction des lignées des spoligotypes de la base de données SITVIT2, en utilisant les réseaux bayésiens (**Aminian et coll., 2014**)

(vii) tbvar est une base de données utilisant un pipeline de calcul systématique qui permet d'annoter des variants fonctionnels potentiels et des variants associées à la résistance aux médicaments à partir des données cliniques de re-séquençage du complexe *M. tuberculosis* (**Joshi et coll., 2013**). À cette fin, les auteurs de ce travail ont re-analysé l'ensemble des données de re-séquençage correspondant à plus de 450 isolats de MTBC disponibles dans le domaine public afin de créer une carte de Variome (l'ensemble des variations génétiques) comprenant plus de 29000 variations de nucléotide unique. Cette base de données peut être intérogée en cherchant directement les variants génétiques (par exemple, 1417019, 3037367, 4222628, etc) ; les gènes (par exemple, *katG*, *pncA*, *gyrA*, etc) ; les « RvIDs » (par exemple, Rv1059, Rv1069c, Rv3693, etc); ou les plages de positions sur le génome (10000-15000; 30000-35000; 80000-85000, etc). tbvar est disponible à l'adresse : http://genome.igib.res.in/tbvar/.

(viii) InTB est une interface Web pour le stockage et l'analyse des données cliniques, socio-démographiques et de typage moléculaire sur la tuberculose (**Soares et coll., 2013**). Elle permet d'insérer et de télécharger des données de génotypage standards en collaboration avec un large éventail de variables cliniques et socio-démographiques qui sont utilisées pour caractériser la maladie. L'interface permet également de classer les nouveaux isolats dans un ensemble bien caractérisé sur la base de références internes, de représenter les données de différentes façons (courbes, graphes), et de générer des arbres phylogénétiques basés sur des données filtrées combinant l'information moléculaire et clinique/socio-démographiques. Construit grâce à des logiciels open source, le code source complet de l'application ainsi que les divers « packages » prêts à l'emploi sont disponibles au lien : http://www.evocell.org/inTB.

(ix) PolyTB est aussi une ressource en ligne, qui permet de rechercher les polymorphismes pouvant exister à l'intérieur de gènes d'intérêt, ainsi que d'examiner la diversité génomique des souches de MTBC (**Coll et coll., 2014**). Dans cette base de données, les auteurs ont traité les données brutes de séquençage (plus de 1500 isolats cliniques correspondant à 8 études) afin d'établir un catalogue de SNPs (n=74039), de petits « indels » ou insertions/délétions dans une séquence nucléotidique (n=4810), et des variantes structurelles plus importantes (n=800). Le code source de PolyTB est librement accessible en ligne à l'adresse : http://pathogenseq.lshtm.ac.uk/polytb.

(x) Genome-wide Mycobacterium tuberculosis variation (GMTV) est une base de données qui répertorie les variations génomiques des souches de *M. tuberculosis*

collectées à travers la Russie. La base de données GMTV (**Chernyaeva et coll., 2014**), qui comprend 1 084 génomes et plus de 69 000 SNPs ou variants Indel, peut être interrogée en fonction de l'information de variation du génome, ainsi que les associations présumées avec la résistance aux médicaments, l'origine géographique, les étapes et résultats cliniques. Cette base de données est accessible en ligne (http://mtb.dobzhanskycenter.org).

(xi) SpolDB4 et SITVITWEB sont des bases de données de génotypage de MTBC développées à l'Institut Pasteur de la Guadeloupe (IPG) (**Brudey et coll., 2006** ; **Demay et coll., 2012**). La plus récente version SITVITWEB est une base de données multi-marqueurs contenant des informations sur 62582 isolats cliniques correspondant à 153 pays d'origine des patients et 105 pays d'isolement des souches. De plus, SITVITWEB contient des informations de spoligotypage sur 58180 isolats cliniques, ainsi que 8161 isolats ayant l'information de typage 12-loci MIRU-VNTRs, et 4626 isolats cliniques avec l'information de typage 5-locus « Exact Tandem Repeats » (ETRs). SITVITWEB est librement accessible en ligne à l'adresse suivante : http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE.

(xii) La base de données SITVIT2 est aussi développée à l'IPG. SITVIT2 contient 111635 isolats cliniques répartis dans 169 pays d'origine des patients et 131 pays d'isolement. Comme ses précédentes versions, elle contient des informations de génotypage de MTBC sur les données de spoligotypage (n=103856 isolats cliniques), les données de typage 12-loci MIRU-VNTRs (n=19311 isolats cliniques), et les données de typage 5-locus ETRs (n=10294 isolats cliniques). Cependant, contrairement aux versions précédentes, SITVIT2 inclut également les données de typage sur les 15-loci MIRU-VNTRs (n=4702 isolats cliniques), et les 24-loci MIRU-VNTRs (n=1672 isolats cliniques). SITVIT2 sera bientôt disponible en ligne à l'adresse : http://www.pasteur-guadeloupe.fr:8081/SITVIT2.

(xiii) SITVITBovis est une base de données extraite de SITVIT2. Elle est dédiée à la tuberculose bovine. Un total de 25741 isolats cliniques ont été recueillis à partir de 60 pays d'origine des patients et 75 pays d'isolement. L'ensemble des isolats (n=25741) contenait l'information de spoligotypage. Les données de typage 12-loci MIRU-VNTRs concernaient 344 isolats cliniques, les 5-locus ETRs étaient disponibles pour 415 isolats, et les 15-loci MIRU-VNTRs étaient disponibles pour 104 isolats cliniques. SITVITBovis sera bientôt disponible à l'adresse : http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis.

6.2 Autres outils ou ressources d'instituts et agences de santé collectant des données sur la TB

D'autres organismes comme l'OMS, ou l'institut de veille sanitaire (InVS), entre autres, participent activement à la collecte et à l'analyse de données sur la tuberculose. Les liens suivants sont accessible en ligne :

- Institut de veille sanitaire : http://www.invs.sante.fr/display/?doc=surveillance/tuberculose/BDD_tuberculose/index_BK.htm.
- "Center for Disease Control and Prevention" (CDC) : http://www.cdc.gov/tb/.
- Organisation Mondiale de la Santé (OMS) : http://www.who.int/tb/en/.
- TBalliance : http://www.tballiance.org/.
- "International Union Against Tuberculosis and Lung Disease" (IUATLD) : http://www.theunion.org/.

6.3 Logiciels de reconstruction phylogénétique et outils de bioinformatique

De nombreux outils et logiciels de bioinformatique et de phylogénie ont été employés pour visualiser, analyser, et calculer les relations évolutionnaires existantes entre les souches de MTBC pour des études épidémiologiques bien spécifiques. Ces outils performants ont apporté des données intéressantes aux analyses conduites. Quelques uns de ces outils et logiciels sont listés ci-après:

(a) Le logiciel BioNumerics (versions 3.5 et 6.6 ; Applied Maths, Sint-Martens-Latem, Belgium ; http://www.applied-maths.com/bionumerics) a été utilisé pour réaliser les arbres phylogénétiques (MSTs, UPGMA,…) avec les données de génotypage (spoligotypage, MIRU-VNTRs), mais aussi en fonction des différents paramètres socio-démographiques, épidémiologiques, ou des données de résistance aux anti-tuberculeux (avec la version 6.6 du logiciel). En plus des outils d'analyse phylogénétique, BioNumerics 6.6 propose aussi des outils statistiques de visualisation et de cartographie des données. La version 7.1 du logiciel (version d'évaluation obtenue pour une période limitée) nous a permis d'analyser les isolats de MTBC de manière simplifée, avec un meilleur traitement des profils MIRU-VNTRs, grâce à des modules/plugins facilitant l'intégration de ces données. Cette version d'essai contenait aussi des outils permettant un meilleur rendu visuel et graphiques des cartographies basées sur Google Map.

(b) MLVA Compare (développé par Ridom et GenoScreen) nous a aussi permis de réaliser l'étude phylogénétique des isolats cliniques de MTBC. Ce logiciel facilite notamment l'interprétation et l'analyse des données de typage MIRU-VNTRs. De plus, le logiciel rend possible l'assignation des lignées phylogénétiques en association avec l'outil Web MIRU-VNTR*plus*.

(c) L'outil Web SpolTools a été utilisé pour dessiner des arbres Spoligoforests à partir des données de spoligotypage. Cet outil Web a permis de générer des arbres basés sur les algorithmes de Fruchterman-Reingold et le « Hierarchical Layout ». Les spoligoforests dessinés à partir de cet outil sont des arbres utilisant le langage de création de graphe DOT, ce qui rend les arbres transportables et modifiables par d'autres logiciels (notamment le logiciel GraphViz).

(d) Le logiciel GraphViz (http://www.graphviz.org) (**Ellson et coll., 2002**) a donc été utilisé pour modifier les spoligoforests créés à partir de SpolTools, et pour colorier les graphes en fonction des lignées phylogénétiques.

(e) L'outil Web V-DICE (http://www.hpa-bioinformatics.org.uk/cgi-bin/DICI/DICI.pl) a été utilisé pour calculer les HGDI en fonction des différents marqueurs génotypiques (Spoligotypes, MIRU-VNTRs). Grâce à cet outil, le pouvoir discriminant peut être calculé à partir des séquences séparées par des loci ou considérées dans leur globalité.

(f) L'application WebLogo (version 2.8.2) disponible au lien : http://weblogo.berkeley.edu/ (**Schneider et Stephens, 1990** ; **Crooks et coll., 2004**), a été utilisée pour évaluer et visualiser la diversité allélique entre les différents motifs d'un groupe donné appartenant au même marqueur moléculaire (spoligotypes, ou MIRU-VNTRs). La méthode de représentation basée sur les données de spoligotypage a déjà été nommée « Spoligologos » (**Driscoll et coll., 2002**).

(g) L'algorithme de clustering OKM (« Overlaping k-means ») (**Guillaume Cleuziou, Université d'Orléans**) est une extension des k-moyennes pour la recherche de classes recouvrantes. Cet algorithme a été appliqué pour mettre en évidence et chercher à mieux comprendre la désignation « Ambigüe » de certaines lignées phylogénétiques basées sur les spoligotypes.

7. Réseaux bayésiens

Au cours de ce travail de thèse, les réseaux bayésiens ont été utilisés pour proposer une meilleure prédiction des lignées de MTBC à l'aide des données de spoligotypage et des MIRU-VNTRs.

Les réseaux bayésiens permettent de modéliser des connaissances sous forme de relations probabilistes. Ces relations sont représentées sous la forme d'un graphe (**Figure 24**). Les réseaux bayésiens peuvent être notamment utilisés dans les domaines de la biologie et de la microbiologie pour la reconstruction et la reconnaissance de séquence de nucléotides dans la molécule d'ADN. Les réseaux bayésiens peuvent également être efficaces dans de nombreux

autres domaines tels que la physique, la thermo dynamique, ou la robotique. L'apprentissage d'un réseau bayésien est constitué de deux tâches :

- l'apprentissage de la structure
- l'apprentissage des tables de probabilités conditionnelles.

Pour l'apprentissage des tables de probabilités conditionnelles (**Tableau 8**), on dispose de l'algorithme Espérance-Maximisation ou « Expectation-Maximisation » (EM), qui est performant. Par contre l'apprentissage de la structure est un problème combinatoire NP-diffcile (**Chickering et coll., 2004**).

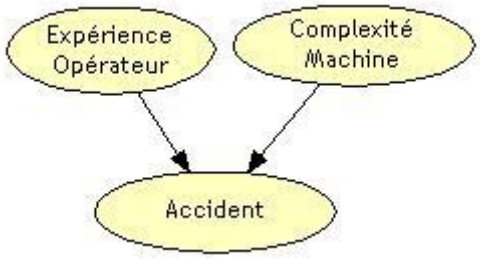


Figure 24 : graphe schématisant un réseau bayésien.

Tableau 8 : Table de probabilité d'accident en fonction des déterminants spécifiques.

Complexité	Faible			Moyenne			Elevée		
Expérience	Faible	Moyenne	Elevée	Faible	Moyenne	Elevée	Faible	Moyenne	Elevée
Accident	1.0	0.5	0.1	1.5	1.0	0.6	2.0	1.6	1.1
Pas d'accident	99.0	99.5	99.9	98.5	99.0	99.4	98.0	98.4	98.9

Les réseaux bayésiens sont des outils de modélisation et de raisonnement très pratiques lorsque les situations sont incertaines ou les données incomplètes (face à la grande quantité de données, il y a une incapacité de l'expert à en extraire les informations pertinentes).

Un réseau bayésien B=(G,V) peut être modélisé à l'aide d'un graphe acyclique dirigé G = (X,E), dont les sommets représentent un ensemble de variables aléatoires $X=\{X_1,\ldots\ldots,X_n\}$, et $V_i= (P(X_i/X_{Pa}(X_i)))$ est la matrice des probabilités conditionnelles du noeud i connaissant l'état de ses parents $Pa(X_i)$ dans G.

Un réseau bayésien B représente une distribution de probabilité sur X dont la loi jointe est :

$$\mathcal{P}(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} \mathcal{P}(X_i/X_Pa(X_i))$$

La **Figure 25** (le réseau ASIA) illustre un autre exemple de réseau bayésien servant à évaluer la propagation de signes médicaux en fonction de certains facteurs. Par exemple, on pourra rechercher avec cette méthode, la probabilité d'attraper la tuberculose en visitant l'Asie (**Lauritzen et Spiegelhalter, 1988**).

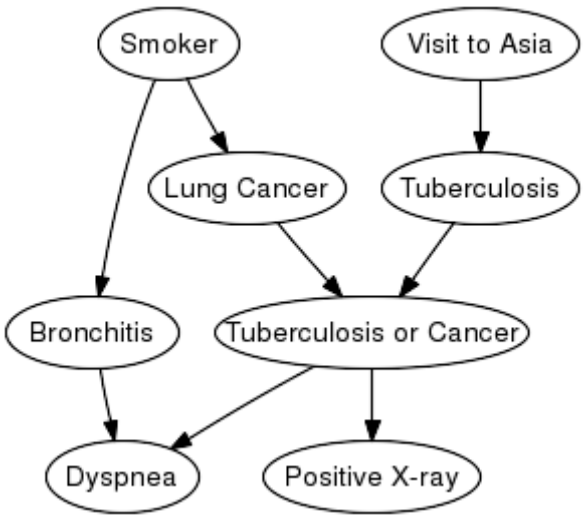


Figure 25 : le réseau ASIA (**Lauritzen et Spiegelhalter, 1988**).

8. Outils de programmation Web et Technologie Java

8.1 Petit historique des technologies web ayant marqué le siècle dernier

- En 1970 naissait le protocole TCP/IP pour uniformiser le transit des informations sur internet.
- En 1984, le protocole http (Hypertext Transfer Protocol) a permis de mettre en réseaux les pages internet (textes + liens).
- En 1990, le langage HTML (HyperText Markup Language) a vu le jour pour mieux structurer les pages web.
- En 1994 naissait le consortium W3C (http://www.w3.org/).
- En 1995 apparaissait Javascript, un langage de programmation de scripts, sans lequel les pages web d'aujourd'hui n'auraient pas la même valeur.
- Et pour finir, 1996 fut l'année d'invention de la CSS (« Cascading Style Sheets » ou feuilles de style en cascade), cet outil permet de donner du style aux sites web.

8.2 L'interface Web pour consulter la base de données

La plateforme de développement Eclipse ® (https://www.eclipse.org/) a permis de réaliser le développement web en implémentant des classes du langage de programmation Java, et des langages Web tels que HTML, JSP (« Java Server Pages »), JavaScript et Ajax. Ces différentes technologies nous ont permis de développer des pages web dynamiques en respectant le modèle MVC (Modèle Vue Contrôleur). De plus, les outils de programmation d'interface Google API et DHTMLX ont permis de réaliser beaucoup de représentations graphiques et des tableaux dynamiques dans l'application web. Aussi, le module java JXL a favorisé la lecture, l'écriture et le traitement des fichiers Excel, à partir de l'application.

Pour déployer notre application Web à travers le protocole Internet HTTP, nous avons utilisé le Server Apache Tomcat 6 ® (http://tomcat.apache.org/) qui est spécialisé dans les applications Web contenant des pages JSP.

Les classes Java Servlets, développées dans le langage JSP permettent de jouer un rôle de contrôleur, acteur principal du MVC (**Figure 26**), et d'établir un lien entre les pages Web et la base de données. La connexion à la base de données se fait à l'aide du pattern Java « singleton » pour s'assurer qu'une seule instance de notre connexion à la base sera mise en œuvre pendant toute la durée de notre application.

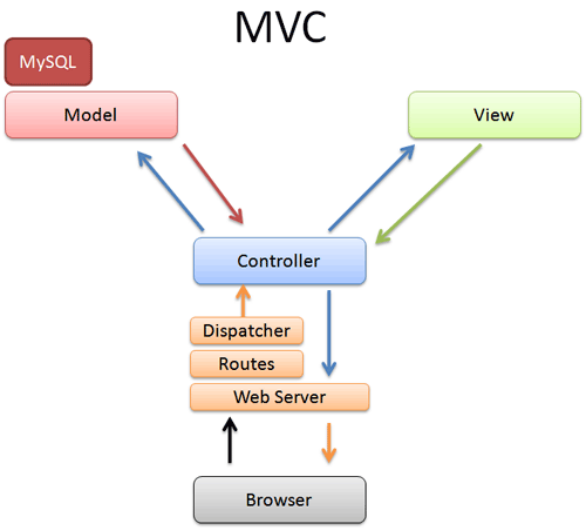


Figure 26 : Modèle MVC (Source : blog.bluepyth.fr).

9. MySQL pour la Gestion de notre Base de Données Relationnelle

Le Système de Gestion de Base de Données Relationnelle (SGBDR) MySQL a été utilisé pour le développement, la gestion et le maintien de la base de données.

Les logiciels MySQL Administrator ® et MySQL Query Browser ® ont été nécessaires pour la consultation, la modification, la mise à jour, et l'insertion/suppresion de données. Ces outils ont aussi servis à réaliser des tests, et à sauvegarder la base.

Le logiciel SQL Maestro for MySQL ® est spécialisé pour la conception des « triggers » ou déclencheurs SQL (processus ou instruction SQL se déclenchant à partir d'une certaine action sur la base). Un compilateur intégré au logiciel permet de compiler le code et de vérifer à tout moment la validité des triggers. Le schéma relationnel de la base de données n'a pas cesser d'évoluer au cours du temps, et ce dernier sera peut être amené à évoluer à nouveau dans le futur (**Figure 27**).

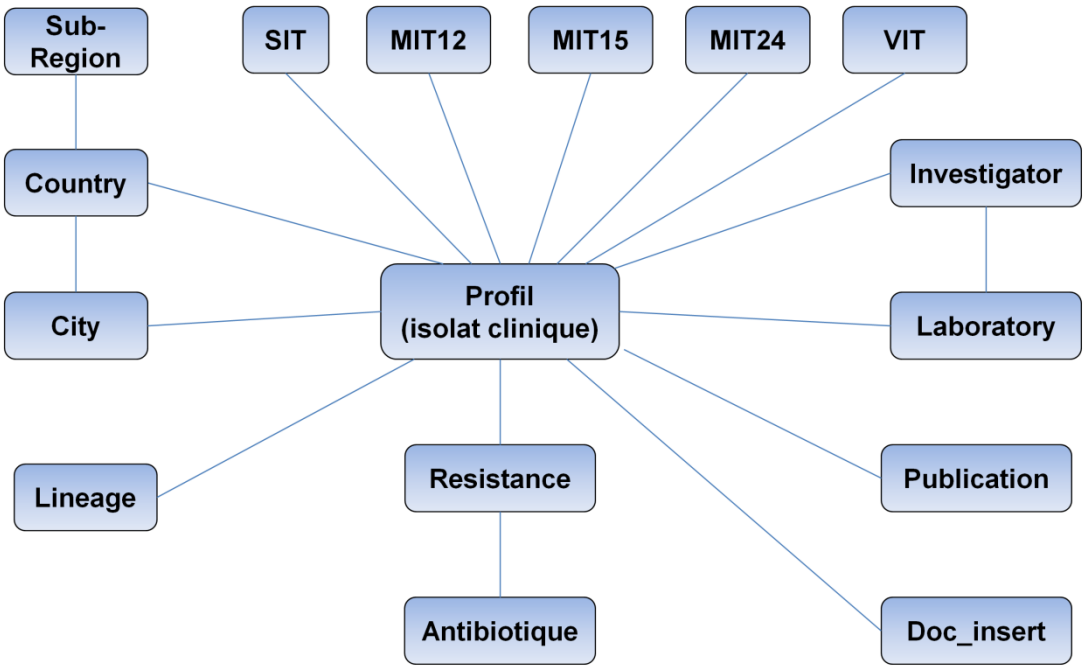


Figure 27 : Schéma relationnel ou Modèle Conceptuel de Données (MCD) très simplifié de la récente version de la base de données.

10. Langage Perl pour le traitement interne des données

Le langage Perl a été utilisé pour la migration et le traitement des données. C'est un langage puissant et utile dans de nombreuses situations : administration système, manipulation de textes (mail, logs, linguistique, génétique), programmation réseau (CGI, mod_perl, etc), bases de données, interfaces graphiques etc.

Perl est un langage interprété, polyvalent, et particulièrement adapté au traitement et à la manipulation de fichiers et des données, notamment du fait de l'intégration des expressions régulières dans la syntaxe même du langage. Il permet de naviguer aisément à travers différents fichiers et différentes bases de données.

Dans un même script, on peut accéder (en lecture et/ou en écriture) simultanément, à des fichiers et bases de données sans passer par d'autres classes ou algorithmes complexes, mais en utilisant simplement des packages.

Le logiciel ActivePerl (http://www.activestate.com/activeperl), utilisé pour pouvoir programmer en langage Perl sous un Système d'Exploitation (SE) Windows, inclut généralement tous les packages (bibliothèques) nécessaires pour faire tourner les programmes, et d'autres, plus spécifques sont téléchargeables. Par exemple, pour réaliser le script de traitement et de migration des données, nous avons téléchargé deux packages spécifiques permettant d'écrire et de lire dans un fichier Excel :

- « Spreadsheet : :WriteExcel » (pour écrire dans un Fichier Excel)
- « Spreadsheet : :ParseExcel » (pour lire un Fichier Excel)

Le développement du code Perl a été réalisé à l'aide de l'éditeur de texte générique Notepad++ (codé en langage C++), qui intègre la coloration syntaxique de code source pour de nombreux langages, dont le langage Perl. La **Figure 28** schématise les traitements principaux qui ont été réalisés à l'aide de Perl et d'autres outils pour la création et le développement de la base de données SITVIT2.

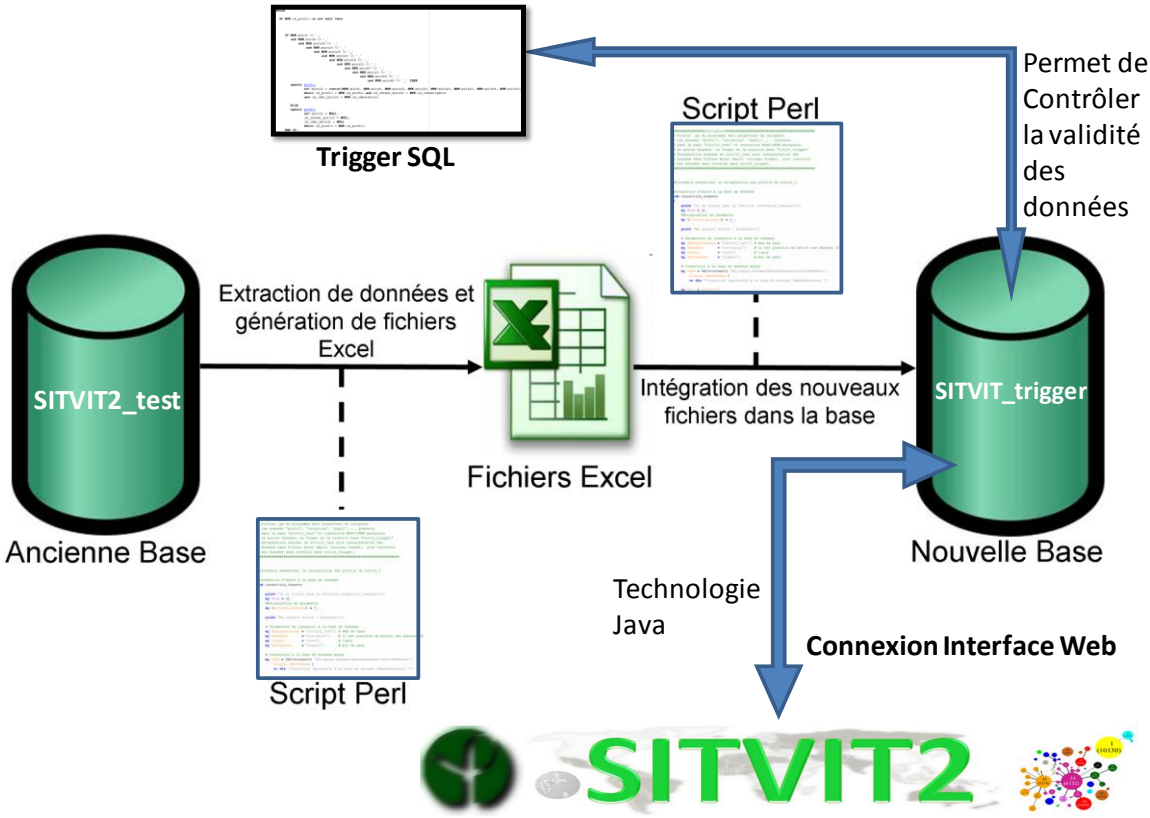


Figure 28 : Schématisation des phases de traitement des données jusqu'à la connexion avec l'interface Web.

11. Collection et interprétation des données dans SITVIT2

Pour chaque version de la base de données, il existe un manuel d'utilisation et d'instruction permettant aux contributeurs de prendre connaissance des modalités d'entrée des données dans la base. Les données de patients qui sont enregistrées dans SITVIT, sont désidentifiées et obtiennent un code spécifique nommé « isoNumber » qui permet un meilleur suivi des données en cas d'échanges inter-laboratoires. Il existe différentes versions de fichiers d'import de données. Ces différentes versions sont accessibles en ligne, en se référant aux interfaces Web correspondantes.

La procédure d'entrée des données dans SITVIT2 est schématisée dans la **Figure 29**, et se fait à partir d'un ficher Excel de la manière suivante :

- L'auteur doit identifier ses souches à l'aide d'un numéro de souche (qui serait identifiable par le laboratoire dans laquel les souches ont été isolées). Ce numéro de souche devra être entré dans la colonne « Strain » du fichier Excel.
- Lorsque le spoligotypage a été réalisé, l'auteur peut fournir la valeur octale de ce marqueur (exemple : 000000000003771) qui sera rentré dans la colonne « Spoligotype43 » du fichier.
- Si l'auteur a appliqué le typage MIRU-VNTRs ou les ETRs, l'ordre des différents loci doit être respecté, i.e :
    - Pour les 5-locus ETRs (colonne « VNTR »), l'ordre est : ETR-A, B, C, D, et E.
    - Pour les 12-loci MIRU (colonne « MIRU12 »), l'ordre est : MIRU 2, 4, 10, 16, 20, 23, 24, 26, 27, 31, 39 et 40.
    - Pour les 15-loci MIRU (colonne « MIRU15 »), l'ordre est : MIRU 4, 10, 16, 26, 31, et 40, ETR-A, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub30, et Mtub39.
    - Pour les 24-loci MIRU (colonne « MIRU24 »), l'ordre des 12 premiers loci est le même que pour les MIRU12, suivi par : ETR-A, ETR-B, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub29, Mtub30, Mtub34, et Mtub39.
- Les pays d'isolement ou d'origine des patients doivent respecter le code iso-alpha-2 (http://fr.wikipedia.org/wiki/ISO_3166-1_alpha-2) avant d'être rentré dans les colonnes correspondantes du fichier Excel.
- Le nombre d'isolats doit aussi être indiqué dans le fichier. Généralement ce nombre est égal à 1.
- La ville d'isolement des souches doit être identifiée par son identifiant geonames-ID correspondant (http://www.geonames.org/).
- Un code laboratoire (exemple « 01 ») est aussi entré dans le fichier afin de différencier les laboratoires distincts pouvant se trouver dans un même pays d'isolement.
- Le type de résistance aux anti-tuberculeux (Code 0, 1, 2, 3, ou 4) doit être indiqué dans la colonne correspondante. Noter que le code 0 indique que l'information de résistance est inconnue.
- Le sexe, l'âge, et le statut de sérologie VIH du patient, peuvent également être entrés dans le fichier.

- La colonne « Remark » peut être utilisée pour ajouter un commentaire ou une autre information importante.
- L'année d'isolement de la souche doit être entrée. Dans le cas où l'année serait inconnue, on entre la chaîne « 0000 ».
- Le nom précis de l'espèce appartenant au complexe *Mycobacterium tuberculosis* peut être aussi spécifié dans le fichier (exemple : *M. tuberculosis*, *M. bovis*, *M. mungi*,….).

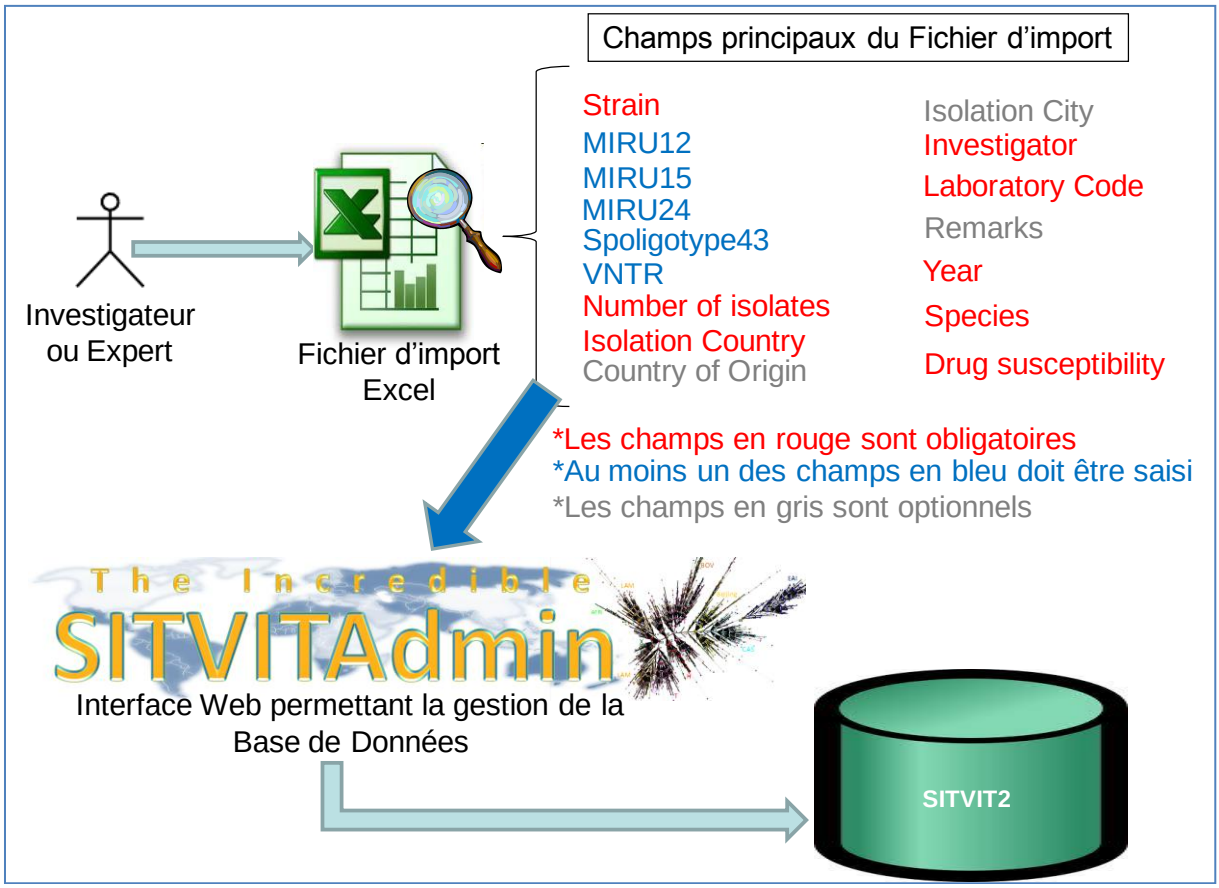


Figure 29 : Schématisation de la procédure d'entrée des données dans SITVIT2.

RESULTATS ET DISCUSSIONS

1. Chapitre 1 – Les Bases de Données et les Outils Bioinformatiques

1.1 Résumé du chapitre

Dans ce chapitre, nous présenterons les bases de données de génotypage du complexe *Mycobacterium tuberculosis* élaborées à l'IPG, dans le cadre d'une initiative concertée visant à lutter contre l'épidémie mondiale de la tuberculose. De la toute première version au format Excel en 1999 (SpolDB1 ; n = 610 isolats cliniques) à la quatrième au format MySQL en 2006 (SpolDB4 ; n = 39 295 isolats cliniques), ces bases de données ont brossé un premier tableau de la phylogéographie des lignées génotypiques de MTBC en circulation, en utilisant le typage oligonucléotidique des espaceurs (spoligotypage) qui permet d'étudier le polymorphisme du locus DR (Direct Repeat). Les deux dernières versions multimarqueurs sous MySQL constituent respectivement la 5e version de SITVITWEB, publiée en 2012 (n = 62582 isolats cliniques), qui repose sur le spoligotypage et le typage MIRU-VNTR 12 loci, et la 6e version, dénommée SITVIT2 publiée en 2014 (n = 111 635 isolats cliniques) dans le cadre de ma thèse et qui contient des données de spoligotypage et de MIRU-VNTR 12, 15 ou 24 loci. Sur cette dernière version, une interface en ligne permet à l'utilisateur de rechercher des souches dans la base de données au moyen de critères tels que l'année, le pays d'isolement, le pays d'origine ou le nom du chercheur. Cette interface permet en outre d'effectuer des recherches mixtes dans SITVIT2, permettant d'obtenir les données de génotypage de certaines souches, conjointement avec leur répartition géographique, ainsi que les données disponibles sur la résistance aux antibiotiques ou les caractéristiques démographiques et épidémiologiques. Notre initiative de recherche a ainsi pour but d'améliorer la caractérisation phylogénétique détaillée des lignées du MTBC, ainsi que l'épidémiologie des clones en circulation, afin d'élaborer une cartographie géographique factuelle des isolats cliniques prédominants pour les bacilles tuberculeux principalement impliqués dans la maladie, à l'échelle nationale et régionale. La superposition ultérieure de ces cartes avec des données sociopolitiques, économiques et démographiques obtenues auprès de Systèmes d'information géographique (SIG) dressera un portrait précis des disparités actuelles par sous-région, selon la classification des Nations Unies. Il est important de comprendre le détail de ces disparités et faiblesses pour que les décideurs politiques et les autorités de santé publique puissent prendre les mesures qui s'imposent pour mieux surveiller, comprendre et contrôler l'épidémie mondiale de la tuberculose.

Nous aborderons dans ce chapitre d'abord les 2 bases de données successives développées au cours de ce travail: SITVITWEB (**Article 1**), SITVIT2 (**Article 2 en préparation** + **Article 2bis**), ainsi qu'un exemple précis de l'utilisation de SITVIT2 centré sur la dissémination globale de la lignée Beijing (**Article 3**), puis le développement de la base SITVITBovis dédiée à la tuberculose bovine (**Article 4 en préparation**), et finalement de SITVIT-KBBN : un outil web basé sur les réseaux bayésiens (**Article 5**).

1.2 Nécessité de développer une approche fondée sur l'utilisation concomitante de l'épidémiologie moléculaire, les bases de données et les outils bioinformatiques

De nos jours, les bases de données sont de plus en plus utilisées pour mener à bien les études épidémiologiques en tout genre. La collecte d'informations, et leur stockage dans des bases de données sont des méthodes très utiles permettant le datamining et l'interprétation des données microbiologiques, mais cette collection de données qui doit toujours être maintenue et pérennisée, doit aussi être considérée avec précaution.

Près de vingt ans après que la tuberculose (TB) ait été déclarée « urgence de santé publique mondiale » par l'Organisation mondiale de la santé (OMS), et malgré les nets progrès réalisés pour atteindre les objectifs internationaux de 2015, fixés dans le cadre des objectifs du Millénaire pour le développement, la TB demeure la deuxième maladie infectieuse la plus meurtrière au monde après le VIH/SIDA (rapport de l'OMS, **WHO 2013**). Ainsi, la TB demeure un problème économique et sanitaire grave, non seulement dans les pays en développement, mais aussi au sein des nations développées, en raison des co-infections TB/VIH et de l'émergence de souches multirésistantes aux médicaments (Multi-Drug Resistant, MDR) et, plus récemment, de souches extrêmement résistantes aux médicaments (XDR), complexifiant la gestion de la maladie et augmentant considérablement la mortalité par TB chez les patients immunodéprimés.

De plus, l'intensification des déplacements et des migrations de population à laquelle on assiste depuis plusieurs décennies, que ce soit pour des motifs touristiques ou professionnels, soulève un nouveau problème dans les pays où la TB était en déclin : la modification des scénarios socio-épidémiologiques en raison d'une immigration massive en provenance de pays où la TB est fortement endémique (**García de Viedma et coll., 2011**). Parmi les questions qui se posent immanquablement, on retiendra notamment : la comparaison entre le rôle de la transmission récente et celui de la réactivation/importation dans les cas de TB d'origine étrangère ; les répercussions d'une éventuelle importation de souches de *M. tuberculosis* non caractérisées à ce jour ; et la transmission croisée entre cas de différentes nationalités. Il est donc important de comprendre comment se transmettent les bacilles tuberculeux, de savoir quels sont les clones impliqués dans les épidémies et/ou les cas résistant à l'action des médicaments, d'identifier de nouveaux clones susceptibles de voir le jour dans un milieu donné ou d'être en voie d'extinction, d'identifier les sous-populations les plus à risque de contracter l'infection et les facteurs de risque associés, et de pouvoir interpréter ces résultats à la lumière de l'évolution du MTBC. Il est évident qu'il serait difficile, aujourd'hui, de répondre à ces questions sans l'aide de l'épidémiologie moléculaire.

On considérait, il y a encore peu de temps, que toutes les souches de MTBC adaptées à l'Homme étaient pratiquement identiques ; la question de la variabilité génétique individuelle entre espèces du MTBC n'a donc pas retenu une grande attention et la plupart des travaux réalisés en la matière se sont focalisés sur des organismes précis. L'arrivée des techniques moléculaires et leur généralisation dans les études des populations de bacilles tuberculeux ont fait évoluer les idées et les technologies. Bien que le MTBC constitue un groupe remarquablement homogène sur le plan génétique, avec des signes tangibles d'évolution clonale, des études récentes ont démontré que la diversité génétique entre les différents clones était nettement plus importante qu'on ne le pensait, ce qui pourrait avoir un impact sur leurs propriétés pathobiologiques. Une étude de référence menée sur une collection mondiale de

souches de MTBC, à partir des données de séquence d'ADN de sept paires de mégabases, a ainsi révélé une diversité génétique importante (**Hershberg et coll., 2008**). Les auteurs de cette étude pensent qu'une grande partie de cette diversité génétique provient d'une dérive génétique qui pourrait être liée à des évènements démographiques et migratoires humains, avec des conséquences fonctionnelles comme l'émergence et la propagation d'une TB résistante aux médicaments.

Il va sans dire que le diagnostic et la confirmation bactériologique précoces de la TB ainsi que la détermination de la résistance aux médicaments constituent des étapes clés dans la lutte contre l'épidémie de TB, et que les laboratoires de référence constituent des structures de premier plan en matière de diagnostic et de lutte antituberculeuse à l'échelle mondiale. Dans un tel contexte, nous avons mis au point une approche fondée sur l'utilisation concomitante de l'épidémiologie moléculaire, des bases de données et des outils bioinformatiques pour des études aux niveaux local, régional et mondial. Pour la surveillance de la TB à l'échelle mondiale, nos données regroupent non seulement nos propres données, mais aussi celles recueillies auprès de différents laboratoires participants implantés dans le monde entier. Ce chapitre rappellera brièvement les étapes suivies pour constituer ces bases de données et les outils en ligne mis au point pour permettre de créer des cartes de répartition géographique mondiale montrant la fréquence des génotypes de TB dans le monde à différentes échelles géographiques. Nous citerons également nos études, publiées ou en cours, qui utilisent ces données et nous présenterons les perspectives qui se dessinent aujourd'hui.

1.3 Les bases de données successives développées au cours de ce travail: SITVITWEB et SITVIT2

À ce jour, la dernière version de notre base de données, SITVIT2 (n = 111 635 isolats), contient les données de génotypage d'environ deux fois plus de souches que la version précédente ; elle sera soumise pour publication en 2014. Il convient de souligner que ces deux versions récentes sont des bases de données multimarqueurs au format MySQL qui contiennent des données de spoligotypage et de typage MIRU-VNTR, limitées aux MIRU 12 loci dans SITVITWEB ; et des données de MIRU-VNTR 12, 15 et 24 loci dans SITVIT2. Sur ces dernières versions, une interface en ligne permet à l'utilisateur de rechercher des souches dans la base de données au moyen de critères tels que l'année, le pays d'isolement, le pays d'origine ou le nom du chercheur ; cette interface permet en outre d'effectuer des recherches mixtes dans SITVIT2, permettant d'obtenir les données de génotypage de certaines souches, conjointement avec leur répartition géographique, ainsi que les données disponibles sur la résistance aux antibiotiques ou les caractéristiques démographiques et épidémiologiques. Grâce à l'élaboration de ces bases de données successives, l'IPG a considérablement amélioré sa connaissance du génotypage et de la phylogénie/phylogéographie de la TB dans le monde. Mon travail de thèse concerne directement les bases de données SITVITWEB et SITVIT2 décrites ci-dessous.

*1.3.1 Présentation de SITVITWEB (**Article 1**)*

La base de données SITVITWEB a été publiée en 2012. C'est une base de données multi-marqueurs intégrant des profils génotypiques de Spoligotypage, 12-loci MIRUs, et 5-loci ETRs (ou VNTR). Elle contient des informations de génotypage sur 62582 isolats cliniques (correspondant à 153 pays d'origine des patients, et isolés dans 105 pays d'isolement). Un total de 7105 motifs spoligotypes a été répertorié, correspondant à 58180 isolats. La ressource de données SITVITWEB contient également un total de 2379 profils MIRU12 (à partir de 8161 isolats), et 458 profils de 5-loci ETRs (à partir de 4626 isolats cliniques). La base de données est publiquement disponible au lien suivant : http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE.

La **Figure 30A** présente la distribution des lignées phylogénétiques définies dans SITVITWEB au niveau mondial (un aperçu de capture d'écran de l'interface SITVITWEB est visible dans la **Figure 30B**). On peut notamment observer qu'en Amérique du Nord, près de toutes les lignées (LAM, Beijing, T, X, Haarlem) sont présentent dans les mêmes proportions (représentant environ 20% des souches) alors que les lignées CAS, AFRI et S sont représentées en dessous de 4% dans cette même macro-région. Dans la région des Caraïbes, les lignées de proportions proches Haarlem (29,2%), LAM (27,7%) et T (27,7%), et la lignée X (8,2%) représentent les lignées phylogénétiques majoritaires. La lignée LAM est très répandue en Amérique du Sud avec une proportion de 49,3%, suivie de la famille T (26,7%) et de la famille Haarlem (15,7%). En Europe, T est la famille la plus représentée, avec une proportion dépassant 30% dans les régions du Nord, du Sud, et de l'Ouest de l'Europe. La lignée Haarlem est répandue en Europe de l'Est et est la deuxième famille la plus représentée en Europe. La lignée EAI est répandue en Europe du Nord (24,9%). Cette forte présence correspond certainement à des cas importés de la tuberculose. Enfin, la famille S est surtout visible en Europe du Sud (5,8%). Comme observé en Amérique du Sud, la lignée LAM est également la plus répandue dans toutes les sous- régions de l'Afrique (à l'exception de l'Ouest de l'Afrique où la lignée AFRI est celle qui prédomine avec 37% des souches, suivie par l'Afrique Centrale avec 7,3% de souches AFRI). La lignée LAM est suivie par le groupe T, qui est présent dans toutes les macro-régions ou sous- régions avec à peu près la même proportion (environ 20%).

La lignée Beijing qui est plutôt répandue dans les régions d'Asie de l'Est, est aussi visible dans 3 sous-régions africaines (au Nord, à l'Est, et au Sud), mais elle est très répandue en Afrique australe avec plus de 20% de souches. La lignée CAS est visible en Afrique du Nord et de l'Est à 11,8% et 7,2% respectivement. Les isolats de la famille S sont bien visibles en Afrique du Nord (8%) et en Afrique du Sud (5,8%). Enfin, en ne considérant que le continent Africain, la famille X dépasse le seuil de 4% uniquement en Afrique du Sud (17,2%), tandis que la lignée EAI dépasse 4% en Afrique du Nord (6,8%) et en Afrique de l'Est (8,2%).

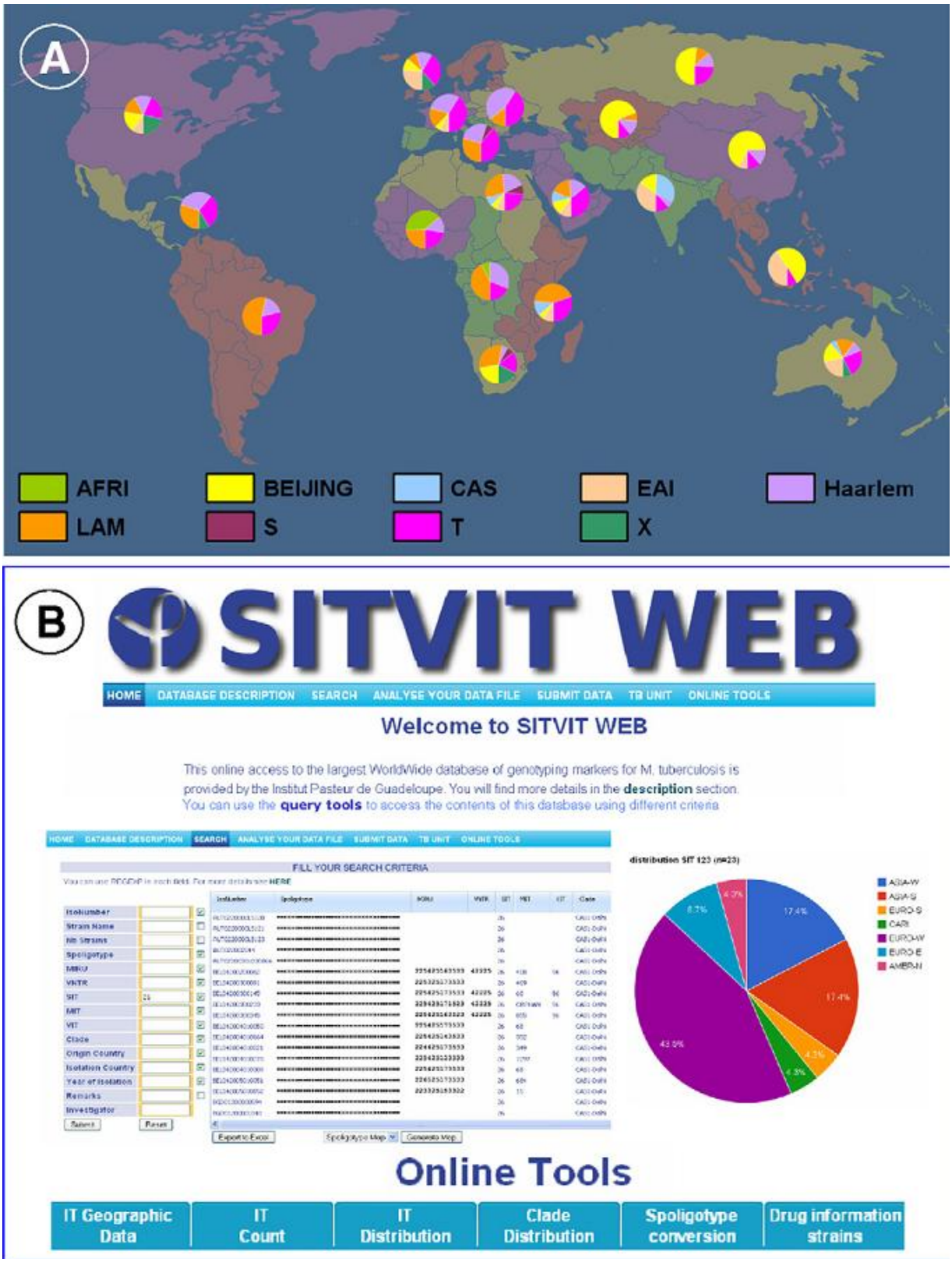


Figure 30 : (**A**) Distribution géographiques des lignées de MTBC basées sur le spoligotypage dans SITVITWEB (seuls les pourcentages supérieurs à 4% sont pris en compte pour la réalisation des camemberts); (**B**) Aperçu (capture d'écran) de l'interface SITVITWEB.

Lorsque l'on considère la famille Beijing, bien qu'elle soit prédominante en Asie de l'Est, elle est aussi très répandue en Asie du Nord (la Russie), en Asie Centrale, et en Asie du Sud-est dans des proportions approchant les 50%. EAI est la famille la plus fréquente dans le

Sud- Est et en Asie du Sud (représentant respectivement 37,6% et 3 % des isolats). La lignée CAS est présente dans l'Asie du Sud, et au Moyen-Orient (Asie de l'Ouest) à 30,1% et 9,6% respectivement. En ce qui concerne le groupe de lignées PGG2/3 (groupe de lignées dites « modernes »), la famille T est la plus fréquente en Asie de l'Ouest (représentant 35,4% des isolats), alors que les familles T et LAM sont retrouvées dans presque le même pourcentage (environ 20%) en Australasie. Les souches de la famille EAI et Beijing sont aussi bien visibles dans cette région du globe.

Concernant les outils web, nous avons utilisé la technologie Java (pages JSP, Javascript, Google API, Ajax…), pour une meilleure visualisation des souches contenues dans SITVITWEB. Cette interface met également à disposition de l'utilisateur, de nombreuses fonctionnalités favorisant le géo-référencement au niveau des sous-régions et des pays des isolats.

*1.3.2 Présentation de SITVIT2 (**Article 2 en préparation** + **Article 2bis**)*

Les fonctionnalités du futur site Internet SITVIT2 seront améliorées par rapport à la version actuelle de SITVITWEB, tant sur le plan numérique qu'au niveau de l'interface de recherche ; la désignation des lignées demeurera quant à elle pratiquement inchangée, à quelques exceptions près. À l'heure où nous écrivons ces lignes, SITVIT2 contient au total 111 635 isolats cliniques de MTBC prélevés chez des patients originaires de 169 pays (1032 villes/131 pays d'isolement). Pour la collecte des données, nous avons enrichi la base de données avec les résultats de génotypage obtenus à l'IPG ou transmis par différents laboratoires coinvestigateurs et partenaires, ou avec des données extraites d'études publiées (**Demay et coll., 2012**). Le site Internet a été développé en langage JSP (Java Server Pages) et est hébergé sur un serveur d'applications Apache Tomcat gratuit (http://tomcat.apache.org). La technologie Java a été utilisée comme précédemment pour SITVITWEB. Comme sur les versions précédentes, la description des caractères génétiques des isolats cliniques consultable dans SITVIT2 repose sur une clé d'identification unique (IsoNumber) qui reprend les informations relatives au pays d'isolement, le code du laboratoire, l'année d'isolement, un code pour la résistance aux médicaments (0 à 4) et un numéro d'isolat unique attribué par le laboratoire ou l'hôpital participant. Conformément aux principes éthiques relatifs au traitement électronique des données, cette procédure permet d'attribuer un numéro anonyme que seule la personne ayant fourni les données (le laboratoire de microbiologie qui a transmis les données), et aucun autre utilisateur, peut décoder pour remonter jusqu'aux informations relatives au patient. SITVIT2 utilise le système de marquage automatique de SpolDB4 et de SITVITWEB, qui attribue un numéro SIT à chaque spoligotype présent dans 2 souches ou plus de la base de données, et un numéro MIT à chaque profil MIRU présent dans 2 souches ou plus. Les numéros MIT relatifs aux formats MIRU-VNTR 12, 15 et 24 loci sont respectivement dénommés 12-MIT, 15-MIT et 24-MIT, tandis que ceux réservés aux 5-loci ETRs sont dénommés VIT. Le terme « orphelin » fait quant à lui référence aux profils décrits pour un seul isolat qui ne correspond à aucun des profils enregistrés dans la banque de la base de données SITVIT2.

Parmi les fonctionnalités du site Internet, on peut notamment citer plusieurs outils de recherche pour la conversion des formats de spoligotype (de binaire en octal et vice versa), l'envoi et l'analyse des données de spoligotypage et de MIRU sous différents formats, ainsi que la recherche par critères pour les marqueurs, seuls ou combinés, l'année et le pays

d'isolement/d'origine, le nom du chercheur, des cartes de répartition géographique et les informations connexes sur la démographie, l'épidémiologie et la résistance aux antituberculeux. Il est donc possible d'effectuer des recherches individuelles ou groupées en téléchargeant un fichier Excel correctement formaté (modèle disponible sur le site). Pour toutes les recherches, l'utilisateur pourrait s'attendre à obtenir un rapport détaillé sur les marqueurs (numéros SIT et MIT), les lignées phylogénétiques et les informations connexes disponibles dans la base de données sous un format anonymisé. Parmi les améliorations prévues, il sera possible de visualiser les correspondances de nomenclature des profils MIT dans SITVIT2, d'après la nomenclature de MIRU-VNTRplus (on notera qu'une comparaison des lignées entre les deux bases de données n'a pas révélé de différences significatives; résultats non présentés).

Les informations concernant le sexe des patients sont disponibles pour 18 944 isolats cliniques, leur âge est connu pour 16 968 personnes, et la sérologie VIH est connue pour 2673 patients. Nous avons remarqué que les lignées LAM, X, et BOV étaient présentes majoritairement chez les patients VIH-positifs (valeur $P < 0,001$). Grâce à l'information sur l'année d'isolement disponible pour de nombreuses souches, nous avons pu corréler divers paramètres épidémiologiques (résistance aux antituberculeux, lignées phylogénétiques) avec le temps pour décrire l'évolution probable de la tuberculose en fonction de ces paramètres. L'information sur la résistance aux antibiotiques, disponible pour 19 606 isolats, a conduit à la conclusion d'une association des lignées tels que CAS et Beijing avec les souches pharmaco-résistantes (résistantes à un ou plusieurs antibiotiques, MDR ou XDR ; avec une valeur $P < 0,001$).

Pour autant que nous puissions en juger, SITVIT2 est la plus grande base de données multi-marqueurs internationale de génotypage accessible au public. Un aperçu de l'interface affichant la distribution mondiale des lignées est donné (**Figure 31A**). Cette base est librement consultable à l'adresse suivante : http://www.pasteur-guadeloupe.fr:8081/SITVIT2. En effet, la taille de cette base de données va continuer à croître dans l'avenir au fur et à mesure des soumissions de données. Le **Tableau 9** présente une brève comparaison des distributions phylogénétiques contenues dans les 2 bases de données SITVITWEB vs. SITVIT2. Nous constatons notamment que la distribution de certaines lignées ne varie pas considérablement entre les deux bases de données. Les proportions étaient semblables pour les familles : Beijing (9,84% vs 9,72% respectivement) CAS (3,69% vs 3,91% respectivement), le Cameroun ou LAM10-CAM (1,04% vs 0,98% respectivement), la Turquie ou LAM7-TUR (0,59% vs 0,53% respectivement), et Manu (1,08% vs 0,95% respectivement). Par contre, nous pouvons remarquer la progression de LAM7-TUR (famille Turkey) dans les pays européens de l'Est, représentant environ 6% dans SITVIT2 et moins de 2% dans SITVITWEB. On peut également remarquer la différence de proportion de la lignée AFRI, en ce qui concerne l'Afrique de l'Ouest. Cette lignée qui représente environ 37% des souches dans SITVITWEB, ne représente qu'environ 29% des isolats dans SITVIT2. Nous pouvons enfin désigner l'augmentation de la lignée CAS en Asie occidentale et en Afrique de l'Est, représentant respectivement dans SITVITWEB vs SITVIT2, autour de 12% contre 18% (Asie occidentale), et environ 10% contre 15% (Afrique de l'Est) (**Figure 31B**).

Tableau 9 : Comparaison des distributions de lignées majeures de MTBC dans SITVITWEB et SITVIT2.

Major Lineages	SITVITWEB (n=62582)		SITVIT2 (n=111635)	
	Nb	%	Nb	%
Beijing	6159	9.84	10850	9.72
AFRI	695	1.11	965	0.86
BOV	6486	10.36	25741	23.06
CANETTII	12	0.02	12	0.01
CAS	2480	3.96	4362	3.91
EAI	4674	7.47	6617	5.93
Haarlem	6425	10.27	9572	8.57
Ural	633	1.01	1008	0.90
LAM	8042	12.85	12245	10.97
Cameroon (previously LAM10-CAM)	650	1.04	1095	0.98
Turkey (previously LAM7-TUR)	370	0.59	593	0.53
Manu	675	1.08	1064	0.95
MICROTI	29	0.05	29	0.03
PINI	152	0.24	159	0.14
S	1151	1.84	1606	1.44
T	12038	19.24	17947	16.08
X	4088	6.53	4683	4.19

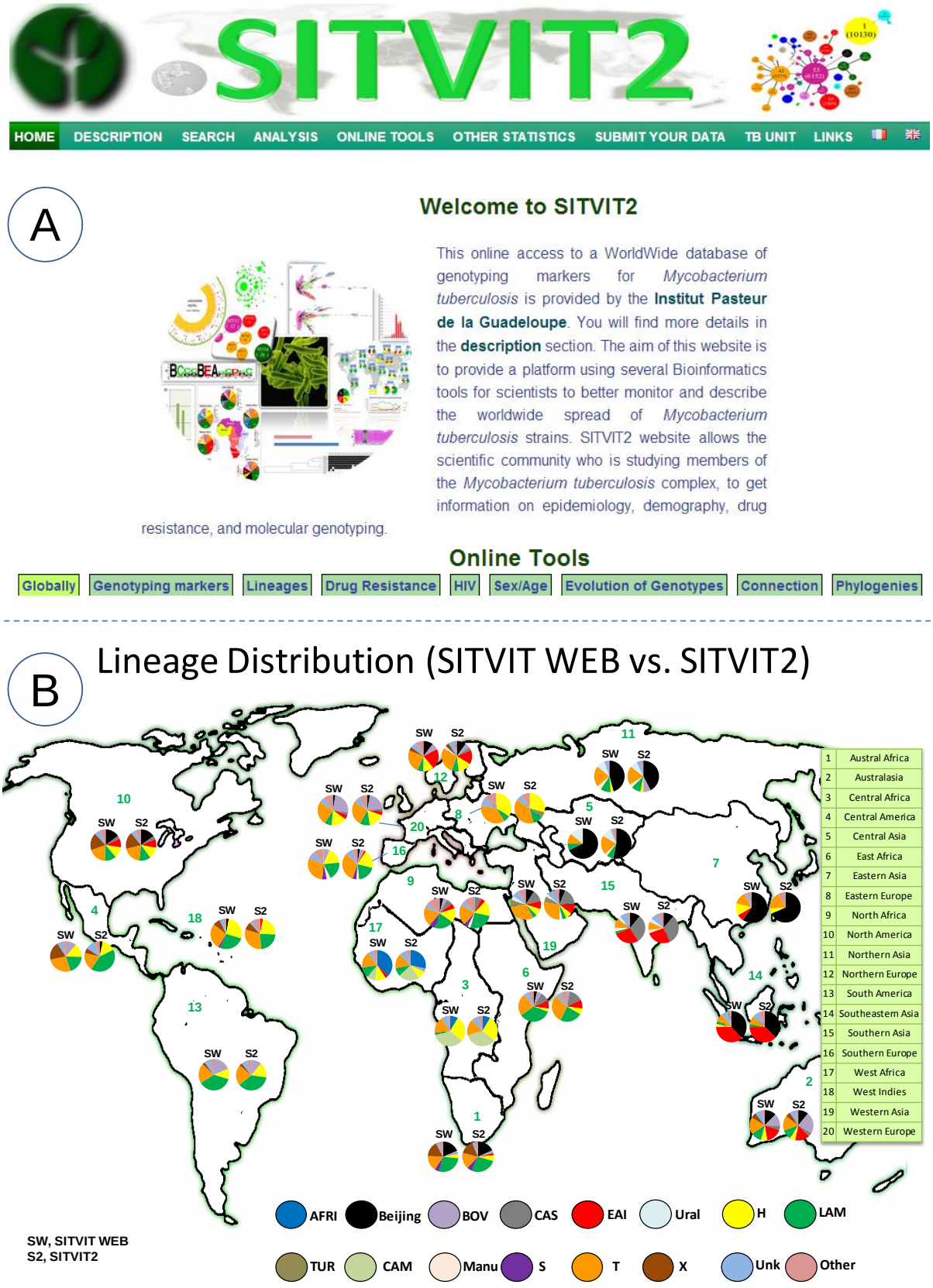


Figure 31 : (**A**) Imprime écran de la nouvelle interface Web de SITVIT2 ; (**B**) Distribution et comparaison des lignées phylogénétiques de SITVITWEB par rapport à SITVIT2 (les lignées représentant plus de 2% des isolats sont considérées).

Lien entre lignée phylogénétique et sérologie VIH/SIDA chez les patients

Parmi les 111 635 isolats disponibles dans SITVIT2, l'information sérologique concernant le VIH était connue pour 2 673 isolats de patients. 607 patients (22,71%) étaient VIH-positifs et 2066 patients (77,29%) étaient VIH-négatifs.

Les patients VIH-positifs ont été trouvés principalement en Russie, en Amérique du Sud et en Afrique de l'Est (à proportion de 67,74%, 40,23% et 33,25% respectivement). Entre 16,67% et 18,89% des patients étaient VIH-positifs dans les Caraïbes, les populations d'Europe du Sud et d'Afrique de l'Ouest. Tandis que la proportion de patients VIH-positifs représentait moins de 5% dans le sous-continent indien (Asie du Sud) et en Asie centrale. La moyenne d'âge entre les 2 groupes de patients VIH-négatifs et VIH-positifs, était similaire, soit respectivement 34,36 ans (avec un écart type de 15,62) chez les séronégatifs et 33,88 ans (avec un écart type de 8,69) chez les séropositifs. Néanmoins, l'âge moyen des patients variait beaucoup en fonction de la sous-région d'isolement (**Figure 32**). Notons aussi que, en ce qui concerne notre base de données, les patients VIH-positifs entre 0-20 ans étaient visibles dans la population péruvienne, espagnole, ainsi qu'au Mozambique et au Nigeria. Toutefois, la différence de sex-ratio était significative entre les patients VIH-négatifs (1279/733 = 1,74 sex-ratio mâle/femelle) et les patients VIH-positifs (438/155 = 2,83 ratio mâle/femelle), avec une valeur de P < 0,0001 (Odds Ratio = 1,62; intervalle de confiance à 95% (1,32 ; 1,99)) (**Figure 32**).

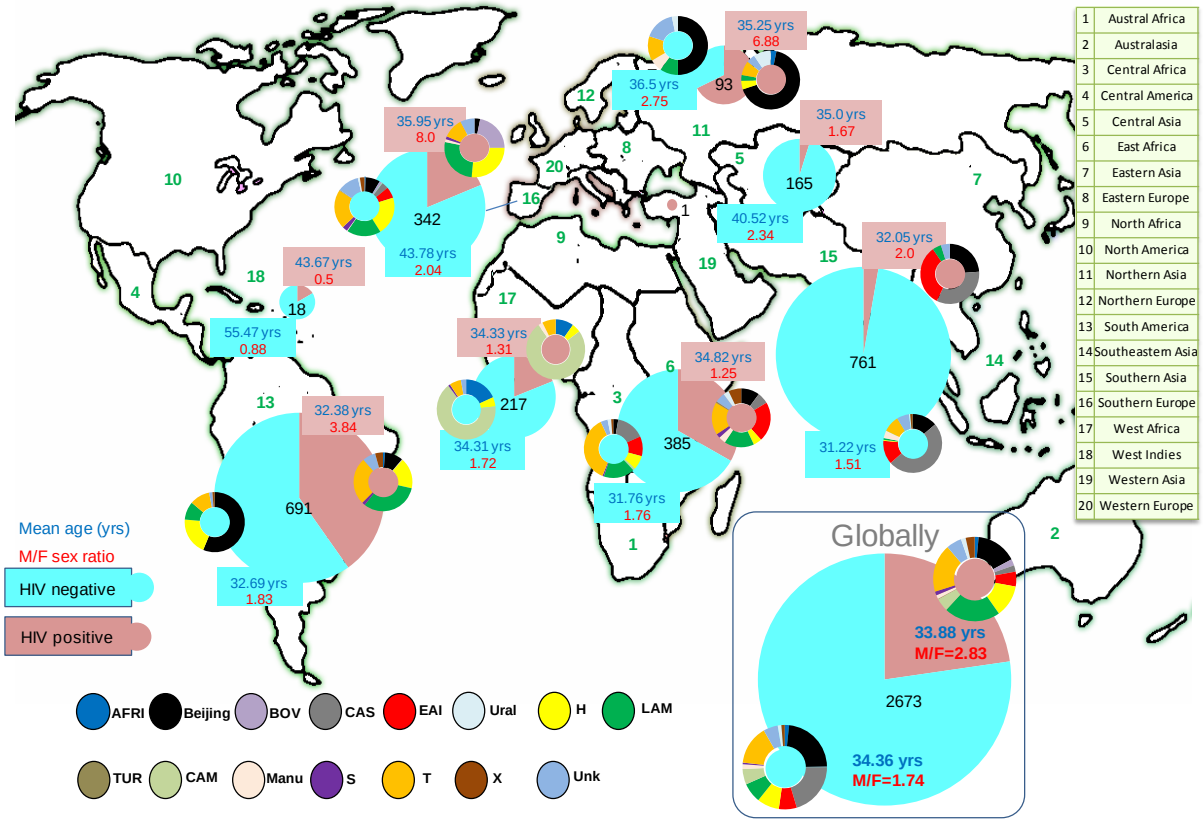


Figure 32 : Distribution globale des patients VIH-positifs et –négatifs dans SITVIT2, en fonction des lignées phylogénétiques. Le nombre présent à l'intérieur du camembert indique le nombre de souches avec l'information de sérologie VIH.

Répartition de la résistance aux antituberculeux selon les pays et régions et les lignées génotypiques potentiellement impliquées

Parmi un total de 16 464 isolats (extraits de SITVIT2 en Mars 2013), un total de 5640 isolats étaient résistants à au moins un antibiotique. Parmi eux, 2105 (37,32% des isolats résistants) l'étaient à un ou plusieurs antibiotiques, mais n'impliquant pas l'isoniazide (INH) et la rifampicine (RIF) en même temps ; 3353 (59,45%) isolats étaient multi-résistants (MDR), et 182 (3,23%) étaient ultra-résistants (XDR).

Les souches MDR ont été trouvées principalement au Kazakhstan (n = 59 ou 80,82%), en Inde (n = 558 ou 62,42%), en Russie (n = 151 ou 56,77%), au Mexique (n = 181 ou 47,01%), en République dominicaine (n = 37 ou 43.02%), en Chine (n = 182 ou 42,42%), au Venezuela (n = 55 ou 41,98%), en Turquie (n = 147 ou 34,11%), en Pologne (n = 39 ou 33,33%), au Kirghizistan (n = 16 ou 30,19%), en Espagne (n = 356 ou 29,67%), et en Guyane/Guyana (n = 18 ou 20,0%). En outre, la totalité des isolats enregistrés à partir d'Afrique du Sud (n = 785), d'Éthiopie (n = 74) et du Panama (n = 37) étaient MDR dans SITVIT2 (**Figure 33**).

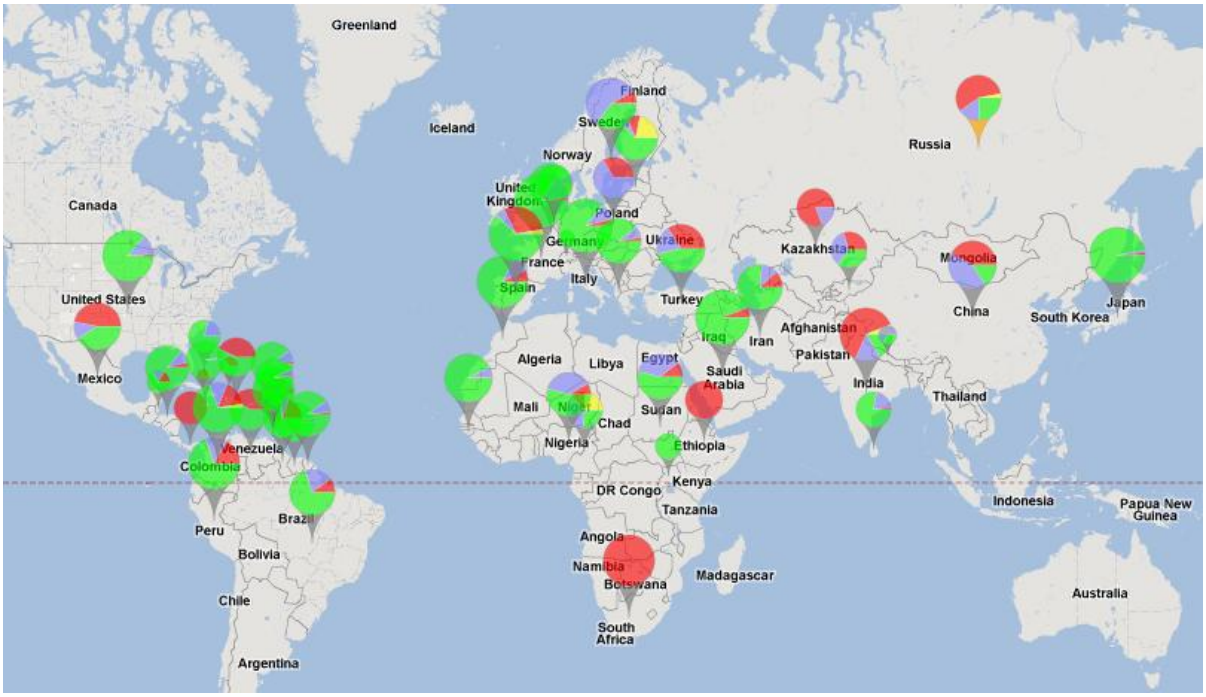


Figure 33 : Répartition par pays de la résistance aux antibiotiques : souches pan-susceptibles (en vert), souches résistantes à au moins un médicament n'incluant pas simultanément l'INH et la RIF (en bleu), souches multi-résistantes ou MDR (en rouge), souches ultra-résistantes ou XDR (en jaune).

Les lignées phylogénétiques de MTBC basées sur le spoligotypage ont été utilisées pour souligner certaines spécificités pouvant exister entre les souches face à la forme de résistance aux antituberculeux (**Figure 34**). Globalement, les lignées CAS, Beijing, Turkey (anciennement LAM7-TUR), Bovis et Manu étaient plus associées à la tuberculose résistante aux antituberculeux (résistance à un ou plusieurs médicaments, MDR ou XDR) ; tandis que AFRI, Haarlem, T, et la lignée « Ural » récemment définie (**Mokrousov, 2012**) sont globalement plus associées à la tuberculose pansusceptible (chi-carré = 320,43, valeur P <

0,0001). Les lignées restantes X, S, EAI, LAM, et le Cameroun (anciennement LAM10-CAM) étaient représentées en proportions à peu près similaires dans les deux groupes de souches pan-susceptibles et résistantes aux médicaments (**Figure 34**). Enfin, lorsque nous nous concentrons au niveau des sous-régions, nous avons remarqué qu'une même lignée peut être différemment associée aux souches résistantes en fonction de la région géographique d'isolement. Nous pouvons noter par exemple les spécificités suivantes :

(i) Les lignées LAM et Haarlem qui sont généralement liées aux souches MTBC pansusceptibles dans la plupart des régions, sont plus souvent associées à des souches résistantes dans la Caraïbe;

(ii) Comme nous l'avons précédemment montré dans le travail dédié aux souches de la lignée Beijing, l'association de cette dernière lignée avec la résistance aux antibiotiques diffère selon la région dans laquelle on se trouve;

(iii) Contrairement à la lignée AFRI, la lignée Cameroun était plus souvent associée à des souches résistantes en Afrique occidentale et centrale;

(iv) La lignée EAI est significativement associée à la résistance en Afrique du Nord, mais pas en Asie de l'Ouest et du Sud, dans les Caraïbes, en Amérique du Sud, et en Europe occidentale;

(v) La lignée CAS est significativement associée à la tuberculose résistante en Asie du Sud, en Afrique du Nord, et en Afrique de l'Est (particulièrement l'Ethiopie), mais pas au Moyen-Orient et en Europe du Nord;

(vi) Bien que les souches de la famille S constituent un très faible pourcentage dans notre étude, on remarque une association statistiquement significative à la résistance dans la Caraïbe et en Amérique central.

Cependant, il peut exister un biais potentiel dans les analyses comparatives basées sur la distribution mondiale des génotypes par rapport à : (i) la sérologie VIH, (ii) la résistance aux antituberculeux, et (iii) la distribution des groupes d'âge et du sexe des patients. De plus, concernant la résistance aux antituberculeux, nous ne pouvons pas exclure la possibilité d'émergence de sous-clones MDR au sein de la population généralement susceptible. Nous pouvons prendre l'exemple de l'émergence récente de souches MDR Ural en Europe de l'Est, tandis que l'avis général sur ce génotype est qu'il est peu résistant et moins transmissible.

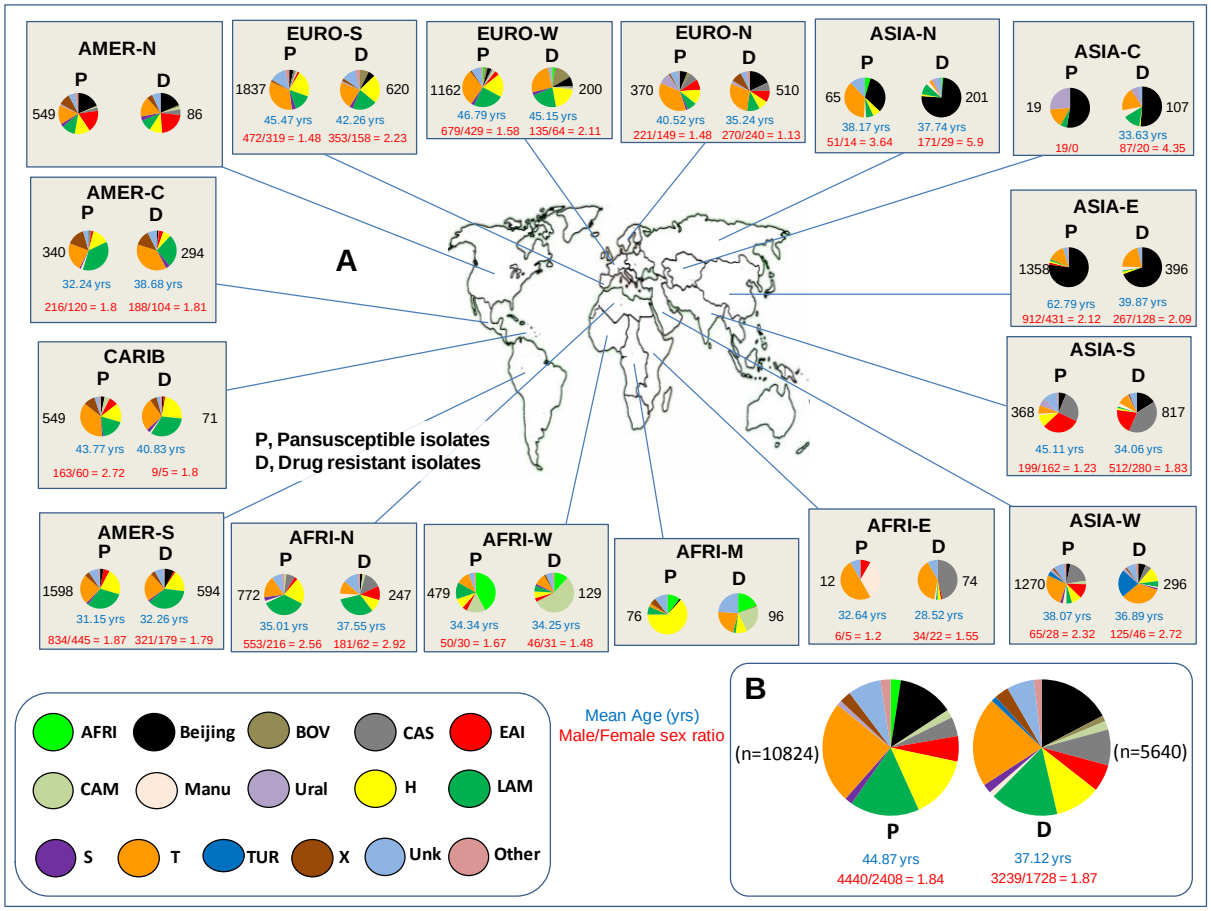


Figure 34 : Distribution des lignée phylogénétiques en fonction des caractéristiques démographiques des patients (âge et sexe) et la résistance aux antituberculeux dans les différentes régions du monde (**A**) et globalement (**B**).

Représentation schématique de la migration des souches résistantes aux antibiotiques

Ayant des données sur les principales lignées phylogénétiques et l'origine géographiques des patients, nous sommes partis d'hypothèse que l'on peut tracer un mouvement potentiel de migration des souches résistantes aux antibiotiques. Par exemple, la présence des cas de multirésistance chez les patients originaires d'Afrique de l'Est et d'Asie du Sud-Est observée en Europe du Nord ne fait que refléter l'apport des souches importées de pays endémiques ayant une large proportion de souches résistantes. Ainsi, nous nous sommes intéressé à des cas où l'origine des patients était connue (n=12 243 isolats, dont 4 035 résistants à un ou plusieurs antibiotiques). La **Figure 35** met en évidence les principales sous-régions d'origine des patients ayant une forte proportion d'isolats résistants et se propageant dans d'autres régions du globe. On peut remarquer qu'une forte proportion de patients porteurs de souches résistantes par rapport aux souches pansusceptible provient de l'Afrique de l'Est, d'Asie du Sud, d'Amérique du Sud, d'Europe de l'Est, d'Afrique du Nord, d'Asie de l'Ouest, d'Afrique de l'Ouest, d'Asie du Sud; en direction des pays d'Europe (Nord, Ouest et Sud) et Amérique du Nord. Par conséquent, notre base de données accessible en ligne constitue un outil intéressant pour avoir un 1er aperçu des tendances d'évolution et diffusion de la résistance aux antituberculeux dans le temps et dans l'espace.

Cependant, il est important de souligner que les données représentées dans le schéma ci-dessous (**Figure 35**) ne sont pas complètes. Pour une raison de visibilité, nous avons considéré les régions d'origine pour lesquelles le nombre de souches résistantes était supérieur à 20. Nous pouvons notamment évoquer les migrations provenant de la Russie (Asie du Nord) vers les pays d'Europe de l'Ouest et les USA.

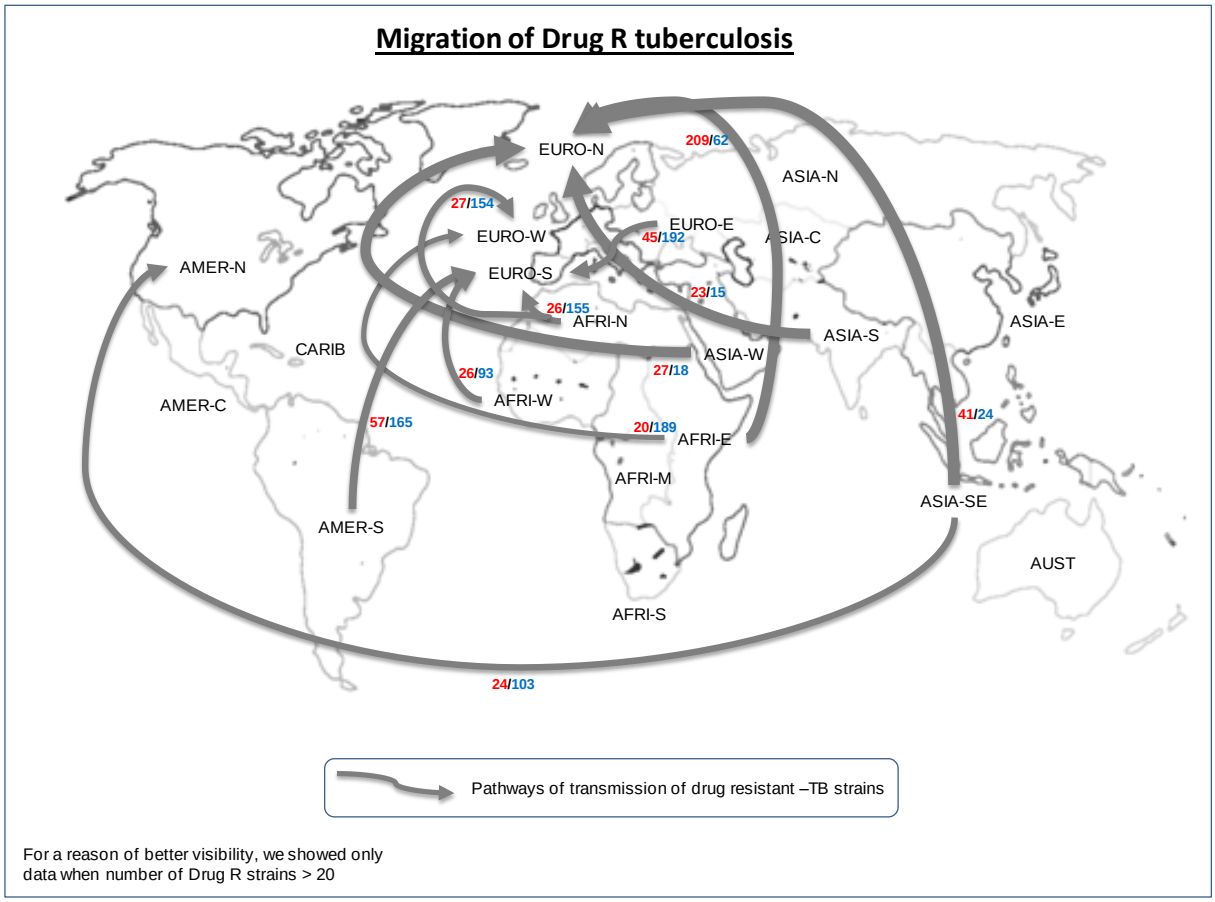


Figure 35 : Carte schématisant les potentielles et principales migrations des souches de tuberculose résistante aux antituberculeux.

Age et sexe des patients tuberculeux dans SITVIT2

Certains clivages ont été observés lorsque nous avons comparé les groupes d'âge et les sex-ratios des patients (**Figure 36**). Les personnes les plus âgées ont été retrouvées en Asie de l'Est (particulièrement au Japon) et dans les pays européens, par opposition aux pays africains où les patients tuberculeux étaient relativement jeunes. Cette observation est en accord avec les données recueillies au niveau mondial (http://en.wikipedia.org/wiki/Population_pyramid), soulignant que les patients de l'hémisphère Sud sont en moyenne plus jeunes que ceux de l'hémisphère Nord. En considérant SITVIT2, on constate que la tranche d'âge des 0-20 ans a été principalement enregistrée chez les patients provenant d'Amérique latine (Amérique centrale et du Sud) et du sous-continent indien (atteignant près de 20% des patients). De plus, la tranche d'âge des 21-40 ans était prédominante (60%) en Afrique (de l'Ouest, du Nord et Afrique de l'Est).

En corrélation avec ce qui a été vu au niveau macro-géographique, nous avons remarqué que certaines lignées/sous-lignées étaient plus visibles chez les patients plus jeunes (entre 0-20 ans) comme : BOV_1 (22,22%), CAS1-Delhi (18,53%), CAS1-Kili (20,00%), LAM 3 (16,58%), LAM6 (15,32%), la famille Turkey ou LAM7-TUR (14.02%), T2-Uganda (17,65%), T3-ETH (18,92%), T5-Madrid2 (24,24%), T5-RUS1 (14,04%), Ural-1 (14,88%), X1 (15,70%), et X3 (15,87%).

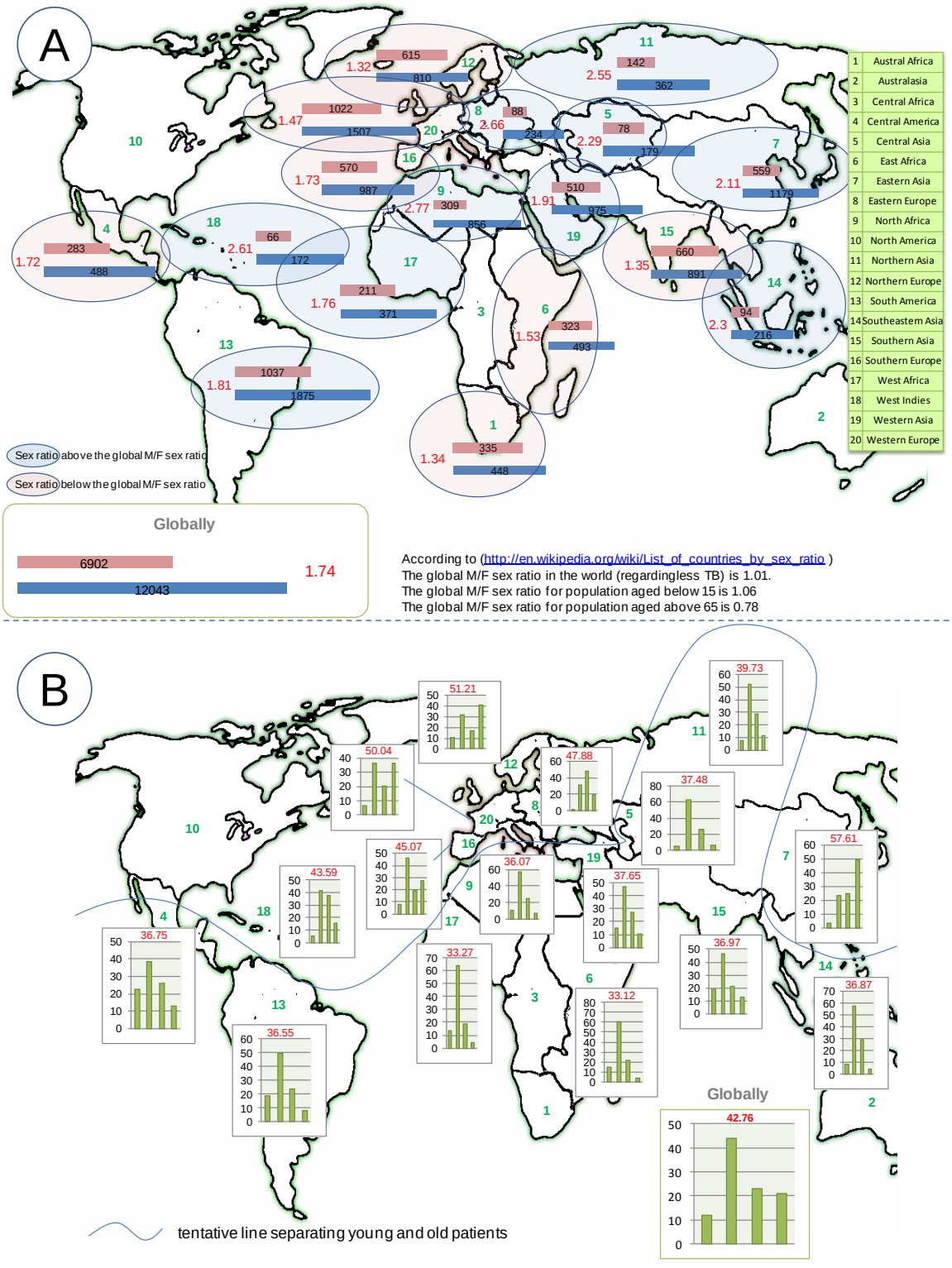


Figure 36 : (**A**) Distribution globale des sex-ratios ; et (**B**) Distribution des groupes d'âge par sous-régions. Les tranches d'âge représentées de gauche à droite sont : 0-20 ans, 21-40 ans, 41-60 ans, et >60 ans.

Impact de la migration sur l'épidémie mondiale de la tuberculose

A l'image du monde dans lequel nous vivons, qui ne fait qu'influer sur l'épidémie mondiale de la tuberculose à travers ses bouleversements et changements, nous assistons à des modifications géopolitiques, socioéconomiques et démographiques majeures qui ne cessent de redessiner les différentes facettes épidémiologiques de *M. tuberculosis*. Ces variations sont perceptibles grâce à des outils moléculaires et des bases de données. Il est évident que les cas de tuberculose dus à des souches importées (provenant de patients étrangers) sont aujourd'hui devenus un problème considérable dans les pays industrialisés. Ainsi les souches d'importation remplacent progressivement les souches autochtones dans l'Europe de l'Ouest et du Nord, et en Amérique du Nord ; tandis que les souches modernes (Euro-Américaines) remplacent les souches ancestrales (*M. africanum*) dans l'Afrique de l'Ouest.

En effet, les patients étrangers sont généralement porteurs de souches spécifiques majoritairement présentes dans leurs pays d'origine. La caractérisation des souches tuberculeuses est donc modifiée en fonction des patients d'origine étrangère impliqués dans une étude donnée. La **Figure 37** indique, lorsque l'information d'origine de la souche est connue, la distribution des lignées phylogénétiques au niveau des sous-régions en fonction de l'origine des patients (autochtones vs. étrangers). On remarque notamment la présence des souches de lignées EAI et CAS (généralement prédominantes respectivement en Asie du Sud-Est et en Asie du Sud) chez les patients étrangers d'Europe du Nord de l'Ouest et du Sud. On peut aussi noter la forte présence de souches de lignée BOVIS chez les patients autochtones d'Europe du Nord, de l'Ouest et du Sud, ainsi que chez les patients d'Amérique du Sud et d'Australasie. Ces données sur la distribution détaillée des souches des patients autochtones versus étrangers sont disponibles en ligne (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/).

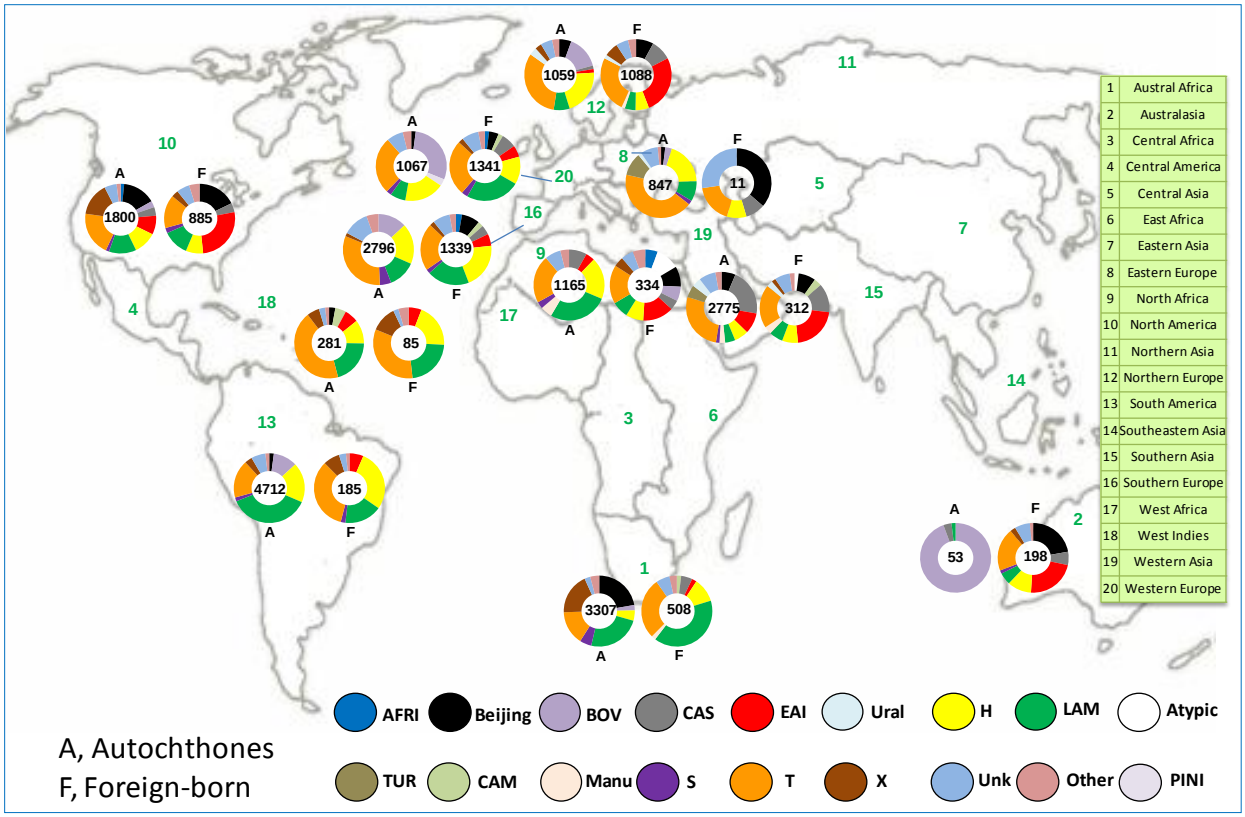


Figure 37 : Distribution des lignées phylogénétiques en fonction de l'origine des patients (autochtones vs. étrangers). Le nombre figurant à l'intérieur des anneaux représente le nombre de souches impliqué dans chaque groupe de patients.

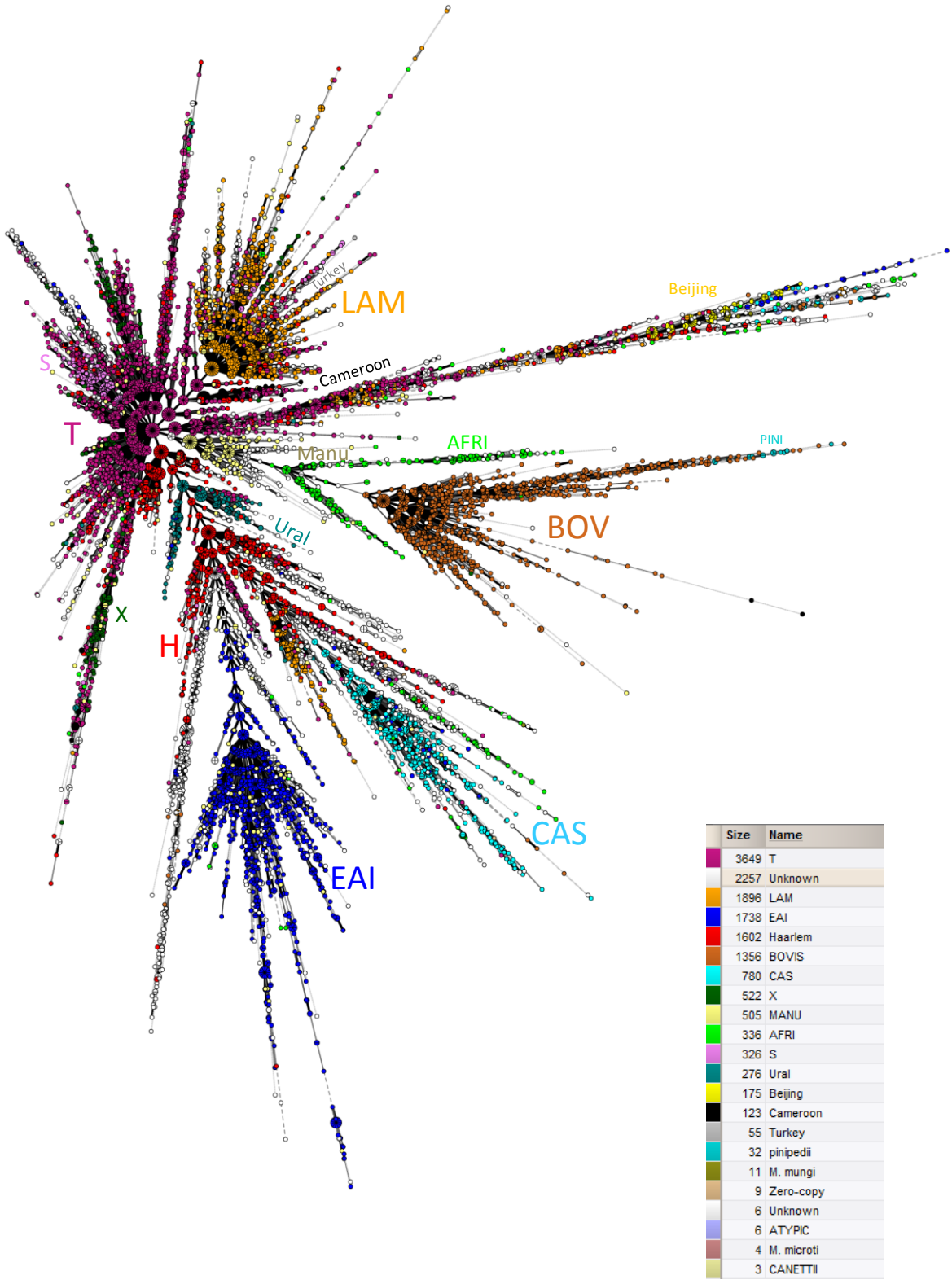
Génotypage et phylogénie globale

Parmi les 111 635 isolats cliniques enregistrés dans SITVIT2, nous pouvons dénombrer : 103 856 isolats pour lesquels le marqueur de spoligotypage était disponible ; 10 294 isolats pour lesquels le marqueur 5-locus ETRs était disponible ; 19 311 isolats pour le marqueur 12-loci MIRU-VNTRs, 4 702 isolats pour le marqueur 15-loci MIRU ; et 1 672 isolats pour les 24-loci MIRU-VNTRs. Le « Minimum Spanning Tree » (MST) de la **Figure 38** a été construit sur la base des spoligotypes, en fonction de tous les isolats disponibles pour ce marqueur (n=103 856).

Le MST de la **Figure 38** présente tous les profils spoligotypes contenus dans SITVIT2, en fonction de leur lignée phylogénétique (couleur des nœuds), et de leur dispersion dans les différentes sous-régions du monde (séparations à l'intérieur des nœuds). Nous remarquons que les lignées majeures sont plus ou moins bien délimitées dans les grappes de spoligotypes. Les motifs appartenant aux lignées « Euro-américaines » se trouvent principalement dans les parties supérieure et centrale de l'arbre (T, S, LAM, Haarlem), alors que certaines grappes excentrées de l'arbre concernent plutôt des isolats de lignées ancestrales (AFRI, BOVIS, CAS, EAI), les groupes de spoligotypes des lignées AFRI et BOV étant particulièrement rapprochés. Nous remarquons aussi que les lignées inconnues (« Unknown ») se trouvent le plus souvent à proximité des lignées « Euro-américaines » (H, LAM, T), cependant, nous pouvons également noter un nombre important de lignées inconnues proches de la famille Manu. Les spoligotypes appartenant à la lignée Cameroon (anciennement nommée LAM10-CAM) sont plus proches des génotypes du groupe T.

Le MST de la **Figure 39** a été construit à partir des motifs 12-MIRU-VNTR complets (i.e. ayant une valeur précise pour chaque locus) présents dans la base de données (n=17919 isolats). Comme pour le MST basé sur le spoligotypage (**Figure 38**), les différents nœuds du présent MST (**Figure 39**) sont subdivisés en fonction du nombre de sous-régions correspondant à un motif MIRU donné. Malgré le fait que les lignées de notre base soient basées sur le spoligotypage, ce MST permet de bien distinguer certaines lignées ancestrales PGG1 telles que Beijing CAS et EAI qui sont regroupées dans la partie droite de l'arbre. Les motifs MIRU12 correspondant à la lignée BOV sont plutôt visibles dans l'une des extrémités supérieures de l'arbre, non loin de certains motifs correspondant à la lignée AFRI. Cependant, les motifs 12-MIRU appartenant aux lignées dites « Euro-américaines » sont plutôt dispersés dans l'arbre, et nous remarquons que certains motifs peuvent être subdivisés parmi plusieurs lignées différentes. Les nombres indiqués sur l'arbre (**Figure 39**) représentent quelques 12-MITs qui sont généralement bien visibles (correspondant à un nombre de souches non-négligeable) dans notre base de données SITVIT2. Enfin, cet arbre MST a permis une relativement bonne différenciation du groupe génétique ancestral (PGG1) par rapport au groupe plus récent (PGG2/3), et il met également en évidence quelques 12-MITs bien spécifiques à certaines lignées phylogénétiques, en particulier celles qui ont été récemment désignées comme « central nodes » à l'aide des profils 12-loci MIRU-VNTRs (**Hill et coll., 2012**). D'autres MSTs ont été construits à partir des marqueurs de typage 15 et 24-loci MIRU-VNTRs (**Annexes 1** et **2**). Le **Tableau 10** indique brièvement les 12-MITs majoritaires dans notre base de données (avec un nombre d'isolats ≥ 50), ainsi que leur association avec les lignées désignées par spoligotypage dans SITVIT2 d'une part, et les lignées obtenues à l'aide de l'outil MIRUVNTR*plus* d'autre part. Mises à part quelques rares divergences (surlignées en gris dans le tableau), les lignées attribuées par SITVIT2 correspondent bien avec celles attribuées par MIRUVNTR*plus* dans cet exemple.

Nous avons développé un algorithme permettant de mettre en évidence les génotypes rares (spoligotypes ou MIRU-VNTRs). A travers cet algorithme, nous nous sommes intéressés aux profils génotypiques contenant un nombre minimal d'isolats cliniques (présents au moins en deux exemplaires dans la base de données) et se répartissant sur un nombre limité de sous-régions. Le pseudo-code de l'algorithme ainsi que l'interface utilisateur permettant de l'utiliser dans l'application web de SITVIT2 sont visibles en **Annexes 3A** et **3B**



Size	Name
3649	T
2257	Unknown
1896	LAM
1738	EAI
1602	Haarlem
1356	BOVIS
780	CAS
522	X
505	MANU
336	AFRI
326	S
276	Ural
175	Beijing
123	Cameroon
55	Turkey
32	pinipedii
11	M. mungi
9	Zero-copy
6	Unknown
6	ATYPIC
4	M. microti
3	CANETTII

Figure 38 : MST construit à l'aide de tous les spoligotypes de SITVIT2 (n=103 856). Les séparations entre les nœuds représentent le nombre de sous-régions différentes partagées par un profil spoligotype donné. Veuillez noter que le nombre indiquant la taille de chaque lignée phylogénétique doit être considéré en fonction des différentes sous-régions et des profils distincts de spoligotypage. Ce nombre ne représente pas le nombre total de souches contenues pour une lignée donnée. Les liens entre chaque nœud indiquent la distance rapprochée (i. e. peu de différences) ou éloignée entre les profils spoligotypes (lien représenté par une ligne noire continue épaisse, une ligne continue grise, une ligne en « tirets », ou une ligne pointillée).

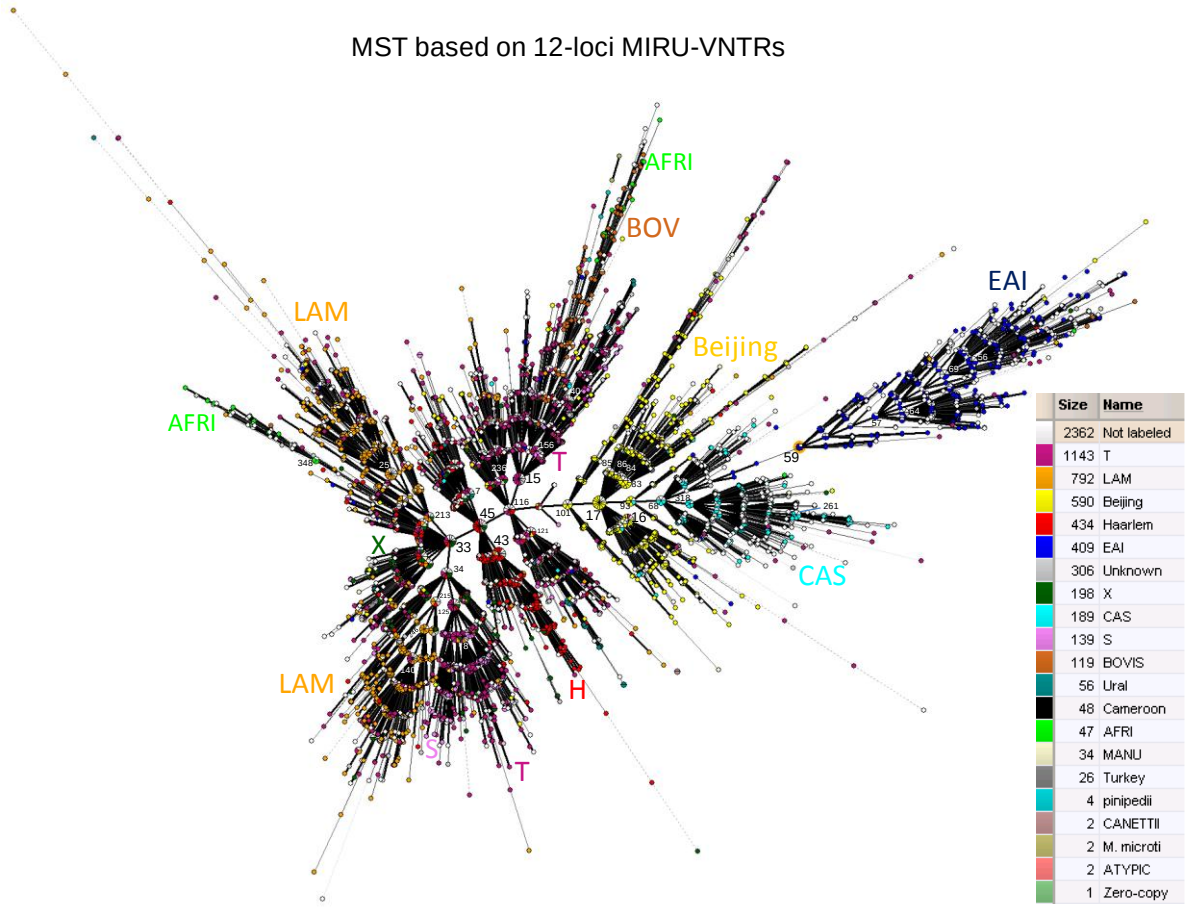


Figure 39 : MST basé sur les motifs 12-loci MIRU-VNTRs complets contenus dans SITVIT2 (n=17919 isolats). Les séparations entre les nœuds représentent le nombre de sous-régions différentes partagées par un profil MIRU12 donné. Les nombres présents à côté des nœuds représentent les 12-MITs. La signification des liens séparant les nœuds est la même que celle de l'arbre précédent.

Le **Tableau 11** représente une mise à jour des règles binaires des lignées basées sur le spoligotypage précédemment décrites dans l'étude correspondant à l'**Article 1** (i.e. SITVITWEB), et indique les SITs représentatifs de ces lignées. Ces règles de description des spoligotypes principaux de SITVIT2 (repésentant un total de 71 lignées) ont été mises en œuvre en se basant sur l'intégralité des motifs de spoligotypage ayant une lignée bien définie dans notre base de données. L'étude de la fréquence globale, basée sur des représentations schématiques des spoligotypes (« SpoligoLogos ») (**Driscoll et coll., 2002**), de la présence ou de l'absence d'espaceur en fonction des 71 différentes sous-lignées, nous à permis d'émettre de nouvelles hypothèses concernant les espaceurs généralement présents ou absents. Les carrés noirs et rouges (remplis et/ou vides) désignant les 43 espaceurs des spoligotypes ont la même signification que dans les règles précédemment décrites. Cependant, les carrés bleus (vides ou remplis) désignent les espaceurs qui sont plus souvent absents ou présents en fonction des lignées/sous-lignées. Ces nouvelles règles pourraient faciliter l'interprétation des spoligotypes et leur assignation à une lignée/sous-lignée spécifique. Aussi, l'**Annexe 4** montre les associations entre les SITs représentatifs et les profils 12 et 15-loci MIRU majeurs. Dans cette figure, un lien est aussi fait avec la moyenne d'âge, le sexe, et le statut VIH des patients en fonction des profils SITs représentatifs. La liste complète des SITs existant dans SITVIT2, accompagnés des lignées qui leurs sont attribuées par une mise à jour des règles expertes se basant sur la description binaire des spoligotypes, sera disponible en ligne dans l'application web.

Nous devons cependant admettre qu'il existe encore des incertitudes incontournables d'interprétation de profils spoligotypes présentant de longues délétions (comme par exemple le SIT254) (**Mokrousov et coll., 2014**).

En conclusion, SITVIT2 est une large collection de données épidémiologiques, démographiques et génétiques basée sur le complexe *Mycobacterium tuberculosis*. Malgré certaines limites, cette base de données contribue de manière significative à mieux comprendre et contrôler l'évolution mondiale de la tuberculose. Les différents MSTs construits témoignent de la grande diversité phylogénétique de MTBC, et de la classification plus ou moins complexe des souches en fonction des méthodes de typage utilisées.

Tableau 10 : Représentation des 12-MITs prédominants (n≥50 isolats) dans SITVIT2.

12-MIT	MIRU-VNTRs	Nb in DB	Predominant Spoligotyping based Lineages in SITVIT2 (%)	Lineage in MIRU-VNTRplus
17	223325173533	479	Beijing (97.09)	Beijing
45	225325153323	447	H3 (44.6), H1 (37.77)	Haarlem
16	223325153533	331	Beijing (97.61)	Beijing
33	224325153323	272	X2 (54.37), T1 (17.5), H3 (12.5)	?
68	225425173533	226	CAS1-Delhi (52.94), CAS (32.35)	?
8	223125153324	221	T1 (74.36)	?
42	225313153323	217	H3 (63.76), H1 (13.42), H2 (11.41)	Haarlem
43	225323153323	207	H1 (52.24), H3 (28.36), H2 (5.97)	Haarlem
34	224325153324	173	X1 (32.26), X3 (26.88), X2 (15.05), T1 (11.83)	?
40	225125113322	155	T1 (65.71), T3-ETH (17.14), T2 (7.14)	TUR
56	254326223432	154	EAI2-Manila (69.15), EAI2-nonthaburi (14.89)	EAI
25	224226153321	147	LAM9 (54.33), LAM5 (23.62), LAM4 (5.51)	LAM
213	224326153323	144	LAM3 (86.52)	LAM
190	124326153220	106	LAM9 (95.19)	LAM
144	224325153314	103	X1 (50.0), T1 (31.82)	?
163	224126152321	100	LAM11-ZWE (62.5), LAM9 (15.63)	LAM
246	124326153324	96	LAM4 (86.96), LAM9 (6.52)	LAM
157	223425153322	95	T1 (70.0), T3 (8.0), H3 (8.0)	?
261	227425113434	95	CAS1-Kili (90.0), CAS (10.0)	?
104	222325173543	88	Beijing (100.0)	Beijing
224	223326153321	86	T1 (85.71), T2 (5.71)	LAM
128	223226153321	85	LAM1 (67.44), LAM9 (18.6)	LAM
271	226425153533	84	CAS1-Delhi (86.36), EAI5 (9.09)	?
46	225325153324	83	H1 (46.0), H3 (22.0), X3 (14.0), T1 (8.0)	Haarlem
7	222325153323	82	Ural-2 (22.0), T1 (20.0), H3 (18.0), H1 (16.0)	?
69	254326223434	82	EAI3-IND (100.0)	EAI
318	226425173533	81	CAS1-Delhi (76.92), CAS2 (7.69), CAS (7.69)	Delhi/CAS
117	224325143324	80	T4-CEU1 (47.06), X2 (17.65), T1 (11.76), X1 (9.8)	?
15	223325153322	79	T1 (72.73), T2 (7.27)	?
310	215125113322	78	Tukey (57.97), T1 (21.74), T3-OSA (18.84)	TUR
152	225325153322	77	H3 (45.16), H1 (32.26), T1 (6.45)	Haarlem
194	214125113322	77	Turkey (92.11)	TUR
633	224315153324	76	T2 (89.19)	Cameroon
156	223125153322	74	T1 (44.0), T5 (32.0), T3 (8.0), T2 (8.0)	?
374	226425173423	73	CAS1-Delhi (83.33)	Delhi/CAS
83	223325163533	72	Beijing (100.0)	Beijing
32	224325153322	71	X1 (38.1), X2 (16.67), T1 (9.52), X3 (7.14)	Haarlem
12	223315153323	69	Cameroon (89.19)	Cameroon
135	221325173533	68	Beijing (100.0)	Beijing
140	124326153224	67	LAM9 (20.93), T1 (18.6), T5-RUS1 (9.3), T-tuscany (6.98), LAM6 (6.98)	LAM
30	224325143323	63	X2 (30.77), T1 (15.38), LAM (11.54), H1 (11.54), LAM5 (7.69)	?
1211	225425183324	61	H3 (93.44), H1 (6.56)	?
116	223325153323	60	T1 (31.43), X2 (20.0),H3 (14.29), X1 (11.43), H1 (5.71), T2 (5.71)	?
64	254326223513	57	EAI1-SOM (59.26)	EAI
26	224226163321	54	LAM2 (51.85), LAM1 (27.78), LAM9 (11.11)	LAM
112	223325143324	53	X1 (36.85), T3 (21.05), T1 (18.42), T4 (5.26), T2 (5.26)	?
237	223126152321	53	LAM11-ZWE (36.36), LAM4 (36.36), LAM9 (27.27)	LAM
396	223125153323	53	T1 (52.17), LAM3 (43.48)	?
456	224315153321	53	Cameroon (66.67), T2 (33.33)	Cameroon
212	233325153324	51	S (82.5), LAM3 (5.0)	S
35	224325153325	50	H3 (27.78), Ural-2 (22.22), T1 (11.11), X1 (11.11), X3 (8.33), LAM9 (5.56), T2 (5.56)	X
231	223125143324	50	T1 (71.43), T (14.29), H2 (7.14)	?

Tableau 11 : Mise à jour des règles de description des lignées/sous-lignées basées sur le spoligotypage.

Typical SIT	Lineage	New Binary Spoligo Rules															Octal Code	Total (Clade)
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
181	AFRI_1	■□■	■■■	□□□	■■■	■■■	■■■	■■■	■■■	■□■	■■■	■■■	■■■	■□□	■□■	■	7707777777777671	570
331	AFRI_2	■□■	■□■	■□□	□□□	■□■	■□■	■□□	□□□	■□■	■■■	■■■	■■□	□□□	■□■	■	7740776077777071	199
438	AFRI_3	■□■	■□■	■□□	□□□	■□■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■□■	■	7740777777777071	69
	AFRI	■■■	■■■	■□□	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	■□■	■	7747777777777671	127
2669	ATYPIC	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□	0000000000000000	110
1	Beijing	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□■■	■■■	■	0000000000003771	10850
482	BOV_1	■□□	■■■	■□□	■■■	■□■	□■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□□	□	6767737777777600	17427
683	BOV_2	■■□	■■□	■□□	□□□	■■■	□■■	■□■	■□■	■□■	■■■	■■■	■■■	■□□	□□□	□	6640737777777600	3258
479	BOV_3	■□■	■□□	□□□	□□□	□□■	□■■	■□■	■□■	■□■	■■□	■□■	■□■	■□□	□□□	□	6400137777777600	164
647	BOV_4-CAPRAE	□■□	□□□	□□□	□□□	□□□	□■■	■■■	■■■	■■■	□■■	■■■	■■■	■■□	□□□	□	2000037777377600	1735
	BOV	■■□	■■■	■■□	■■■	■■■	■■■	■■■	■■■	■□■	■■■	■■■	■■■	■□□	□□□	□	6767777777777600	3157
592	CANETTII	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□■	□□□	□□■	□□□	□□□	□	0000000001011000	12
26	CAS1-Delhi	■■■	□□□	□■■	■■■	■■■	■■■	■■■	■□□	□□□	□□□	□□□	□■■	■■■	■■■	■	7037777400037771	3220
21	CAS1-Kili	■■■	□□□	□■■	□■■	■■■	■■■	■□□	□□□	□□□	□□□	□□□	□■■	■■■	■■■	■	7033774000017771	444
288	CAS2	■■■	□□□	□■■	■■■	■■■	■■■	■■■	■□□	□□□	□□□	□□□	■■■	■■■	■■■	■	7003777400037771	168
	CAS	■■■	□□□	□■■	■■■	■■■	■■■	■■■	■□□	□□□	□□□	□□□	□■■	■■■	■■■	■	7037777400037771	530
48	EAI1-SOM	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□■	□■■	■■■	□■■	■	7777777774137731	1004
745	EAI1-MYS	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□■	□■■	□□■	■■■	■	7777777774133331	42
19	EAI2-Manila	■■□	■■■	■■■	■■■	■■■	■■■	■□□	■■■	■■■	■□□	□□■	□■■	■■■	■■■	■	6777774774137771	1147
89	EAI2-nonthaburi	■■□	■■■	■□□	□□□	□□□	□□□	□□□	□□□	□■■	■□□	□□■	□■■	■■■	■■■	■	6740000034137771	112
	EAI2	■■□	■■■	■■■	■■■	■■■	■■■	■□□	■■■	■■■	■□□	□□■	□■■	■■■	■■■	■	6777774774137771	25
11	EAI3-IND	■□□	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□■	□■■	□□□	■■■	■	4777777774137071	982
139	EAI4-VNM	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	■□□	□□■	□■■	■■■	■■■	■	7777777744137771	426
236	EAI5	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□■	□■■	■■■	■■■	■	7777777774137771	1831
591	EAI6-BGD1	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□■	■■■	■□□	□□■	□■■	■■■	■■■	■	7777777574137771	679
1898	EAI7-BGD2	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□□	□□□	□□□	□■■	■■■	■	7777777700037771	104
109	EAI8-MDG	■□□	■■■	■■■	■■■	■■■	■■□	□■■	■■■	■■■	■□□	□□■	□■■	■■■	■■■	■	4777773774137771	275
47	H1	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□□	□■□	□□□	■■■	■■■	■	7777777774020771	3001
2	H2	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	■□□	□□□	□■□	□□□	■■■	■■■	■	0000000044020771	413
50	H3	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	□■□	□□□	■■■	■■■	■	7777777777207771	4287
777	Ural-1	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□■□	□□□	■■■	■■■	■	7777777774207771	682
127	Ural-2	■□■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□■□	□□□	■■■	■■■	■	5777777774207771	326
20	LAM1	■■□	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	6777776077607771	1135
17	LAM2	■■□	■■■	■■■	■■■	□■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	6777376077607771	915
33	LAM3	■■■	■■■	■■□	□□■	■■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	7761776077607771	2009
60	LAM4	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	□■■	■	7777776077607731	700
93	LAM5	■■■	■■■	■■■	■■■	□■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	7777376077607771	629
64	LAM6	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■□■	■■□	□□□	■■■	■■■	■	7777776075607771	635
290	LAM8	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	7777776067607771	66
42	LAM9	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	7777776077607771	4668
59	LAM11-ZWE	■■■	■■■	■■■	■■■	■■■	■■■	■■■	□□□	■■■	□□□	■■□	□□□	■■■	■■■	■	7777776060607771	1057
209	LAM12-Madrid1	■■■	■■■	■□□	□□□	□□■	■■■	■■□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	7760176077607771	49
41	Turkey	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□□□	■□□	■■■	■■□	□□□	■■■	■■■	■	7777774047607771	593
61	Cameroon	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□□	□■■	■■■	■■□	□□□	■■■	■■■	■	7777777437607771	1095
523	Manu_ancestor	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■	7777777777737771	45
100	Manu1	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□■	■■■	■■■	■	7777777777737771	209
54	Manu2	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□■■	■■■	■■■	■	7777777777637771	584
1378	Manu3	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	□□□	■■■	■■■	■	7777777777707771	221
539	microti	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	■■□	□□□	□	0000000000000600	29
593	PINI1	□□□	■■■	■□□	□□□	□□□	□□□	□□□	□■■	■■■	■■■	■■■	■■■	■■□	□□□	□	0740000377777600	33
637	PINI2	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□□□	■■■	■■■	■■■	■■■	■■□	□□□	□	0000000077777600	119
	PINI	□□□	■■■	■□□	□□□	□□□	□□□	□□□	□■■	■■■	■■■	■■■	■■■	■■□	□□□	□	0740000377777600	7
34	S	■■■	■■■	■■□	□■■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7763777777607771	1606
451	T-H37Rv	■■■	■■■	■■■	■■■	■■■	■■■	■□□	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777774777607771	180
1737	T-tuscany	■■■	■■■	■■■	■■■	■□□	□□□	□□□	□□■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777600177607771	63
53	T1	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777777777607771	11846
280	T1-RUS2	■■■	■□■	□□□	□□□	□□□	□□□	■□■	■■■	■■■	■□■	■■□	□□□	■■■	■■■	■	7700007777607771	118
52	T2	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■□	□■■	■	7777777777607731	1849
135	T2-uganda	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	□■■	□	7777777777607730	148
37	T3	■■■	■■■	■■■	■■■	□■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777737777607771	894
149	T3-ETH	■■■	■■■	■■■	□□□	□□□	□□□	□□■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7770003777607771	505
627	T3-OSA	■■■	■□□	□□■	■■■	□■■	■■■	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7417377777607771	64
40	T4	■■■	■■■	■■■	■■■	■■■	■■■	□■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777773777607771	265
39	T4-CEU1	■■■	■■■	■■■	■■■	■■■	■■■	□■■	■□□	■■■	■■■	■■□	□□□	■□□	■■■	■	7777773477607471	238
44	T5	■■■	■■■	■■■	■■■	■■■	■■■	■■■	■□■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777777577607771	478
58	T5-Madrid2	■■■	■■■	■■■	■■■	■■■	■■■	■□■	■□■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777775577607771	204
254	T5-RUS1	■■■	■■■	■■■	■■■	■■□	□□□	□□□	□□□	■■■	■■■	■■□	□□□	■■■	■■■	■	7777600077607771	233
119	X1	■■■	■■■	■■■	■■■	■■■	■■□	■□■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7777767777607771	1733
137	X2	■■■	■■■	■■■	■■■	■■■	■■□	■■■	■■■	■■■	■■■	■■□	□□□	■■□	□□□	■	7777767777607601	1579
92	X3	■■■	□□□	□□□	□□□	■■■	■■□	■■■	■■■	■■■	■■■	■■□	□□□	■■■	■■■	■	7000767777607771	1354
405	Zero-copy	■■■	■■■	■■■	■■■	■■■	■■■	□□□	□□□	□□□	□□□	□□□	□□□	□□□	□■■	■	7777770000000011	22

*1.3.3 Un exemple de l'utilisation de SITVIT2 centré sur la dissémination globale de la lignée Beijing – une urgence mondiale (**Article 3**)*

Des efforts supplémentaires sont souhaités pour lutter de manière concertée et coordonnée contre la tuberculose (TB). Les outils moléculaires et les outils Web sont indispensables pour contrôler la propagation mondiale de la tuberculose, surveiller le phénomène de résistance aux antituberculeux, et déterminer les populations à risque. De plus, ces outils sont nécessaires pour identifier les clones du complexe *Mycobacterium tuberculosis* (MTBC) circulants. Dans cette étude, nous avons pris l'exemple spécifique de la lignée Beijing qui est une lignée reconnue pour son association avec l'émergence mondiale des tuberculoses multi-résistantes (MDR) et ultra-résistantes (XDR). Un total de 10 850 isolats appartenant à la lignée Beijing (à partir de 169 pays d'isolement et 92 pays d'origine du patient), a été extrait de SITVIT2 et analysé dans cette étude. L'information concernant le spoligotypage des isolats, était disponible pour l'intégralité des données (n = 10 850), les minisatellites MIRU-VNTRs étaient disponibles pour 2896 isolats, et l'information de résistance aux médicaments a été cartographiée au niveau macro-géographique et à l'échelle des pays pour 2846 souches de la famille Beijing.

Par rapport aux autres lignées (non-Beijing) du complexe *M. tuberculosis* (**Figure 40A**), la proportion de souches de la famille Beijing était plus élevée en Asie de l'Est (n = 3519/5216, ou 67,47%), suivie de l'Asie-Centrale (n = 170/320, ou 53,13%), l'Asie du Nord (n = 744/1924, ou 38,67%), l'Asie du Sud-est (n = 1048/3059, ou 34,26%), l'Afrique australe (n = 823/4824, ou 17,06%), l'Amérique du Nord (n = 2179/14557, ou 14,97%) et l'Asie du Sud (n = 532/4883, ou 10,89%). Les sous-régions restantes montraient une proportion de souches Beijing inférieure à 10%. Bien que le sex-ratio mâle/femelle global était de 1,96, la répartition des sex-ratios mâle/femelle dans les sous-régions (**Figure 40B**) connaissait des variations importantes et statistiquement significatives (**Tableau 12**); ce ratio était le plus élevé en Russie avec une valeur de 4,93. Cependant, nous avons observé des sex-ratios s'étalant de 0,8 à 1,13 en Amérique Centrale, dans les Caraïbes, en Afrique de l'Est et en Europe du Nord (valeur P < 0,0001).

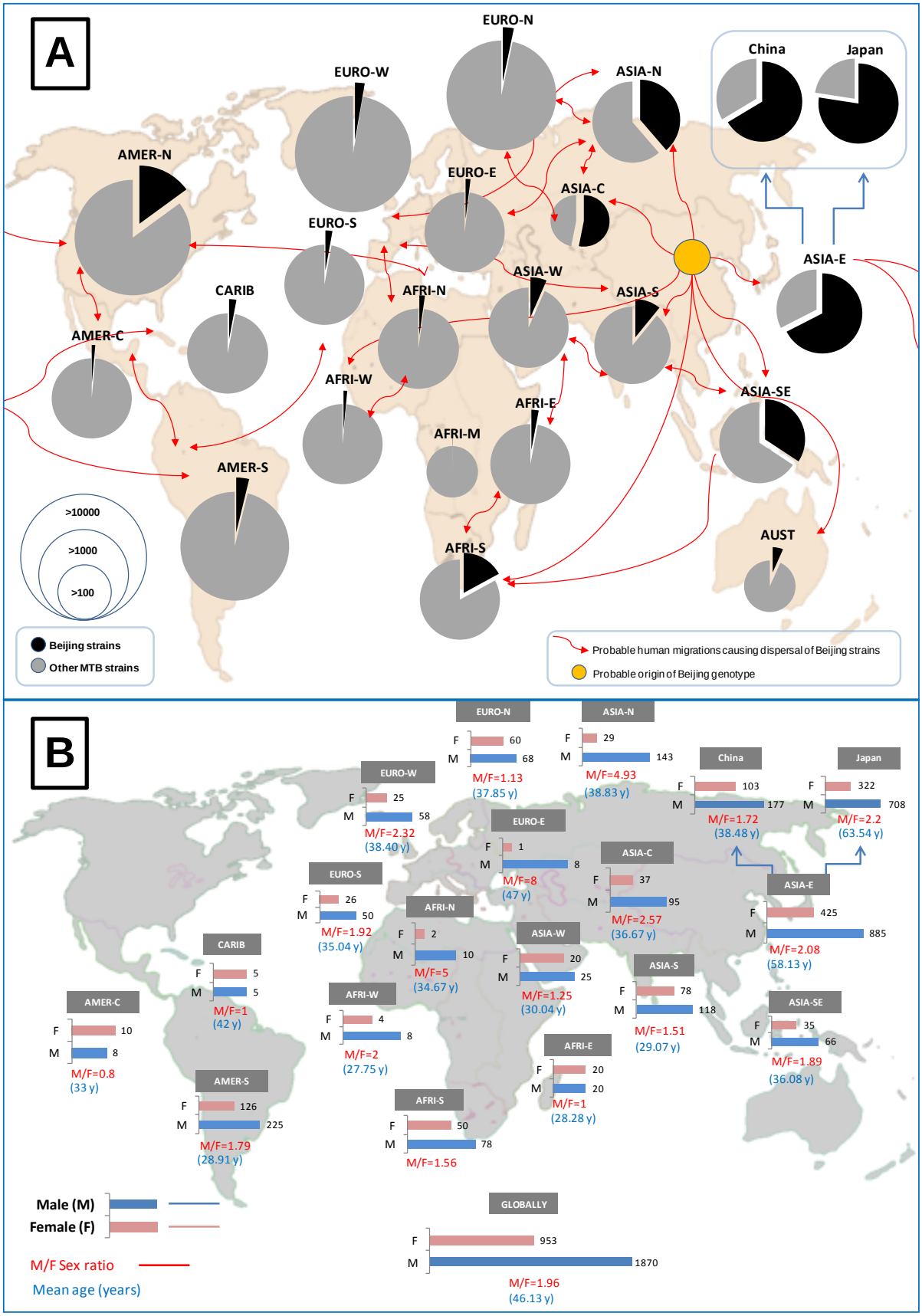


Figure 40 : (**A**) Distribution phylogéographique des souches de la famille Beijing comparée aux autres souches (non-Beijing) de MTBC. (**B**) Distribution des sex-ratios mâle/femelle et des moyennes d'âge au niveau des macro-régions.

Tableau 12 : Matrice de corrélation indiquant les différences statistiques (valeurs P) des sex-ratios entre les différentes sous-régions étudiées.

	AMER-C	ASIA-C	AFRI-E	ASIA-E	AFRI-N	ASIA-N	EURO-N	AMER-S	ASIA-SE	ASIA-S	EURO-S	AFRI-W	CARIB	ASIA-W	EURO-W	AFRI-S
AMER-C	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_
ASIA-C	0.018*	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_
AFRI-E	0.70	0.0097*	_	_	_	_	_	_	_	_	_	_	_	_	_	_
ASIA-E	0.038*	0.30	0.02*	_	_	_	_	_	_	_	_	_	_	_	_	_
AFRI-N	0.04*	0.32	0.04*	0.36	_	_	_	_	_	_	_	_	_	_	_	_
ASIA-N	0.0001*	0.019*	<0.0001*	<0.0001*	0.67	_	_	_	_	_	_	_	_	_	_	_
EURO-N	0.49	0.0017*	0.73	<0.001*	0.044*	<0.0001*	_	_	_	_	_	_	_	_	_	_
AMER-S	0.092	0.10	0.081^	0.22	0.14	<0.0001*	0.029*	_	_	_	_	_	_	_	_	_
ASIA-SE	0.092	0.28	0.092	0.65	0.18	0.0008*	0.062^	0.82	_	_	_	_	_	_	_	_
ASIA-S	0.19	0.028*	0.23	0.042*	0.094	<0.0001*	0.21	0.37	0.39	_	_	_	_	_	_	_
EURO-S	0.094	0.35	0.098	0.75	0.19	0.0024*	0.077^	0.78	0.95	0.40	_	_	_	_	_	_
AFRI-W	0.23	0.77	0.31	1	0.32	0.23	0.28	0.56	0.60	0.45	0.61	_	_	_	_	_
CARIB	0.75	0.13	0.64	0.31	0.11	0.021*	0.55	0.28	0.26	0.37	0.33	0.36	_	_	_	_
ASIA-W	0.43	0.042*	0.61	0.092	0.074^	0.0001*	0.78	0.26	0.26	0.57	0.26	0.36	0.51	_	_	_
EURO-W	0.04*	0.74	0.032*	0.66	0.28	0.015*	0.015*	0.32	0.51	0.13	0.58	0.72	0.18	0.11	_	_
AFRI-S	0.18	0.059^	0.22	0.13	0.11	<0.0001*	0.21	0.52	0.49	0.90	0.49	0.48	0.36	0.53	0.19	_
EURO-E	0.033*	0.25	0.035*	0.29	0.61	0.54	0.036*	0.11	0.14	0.079^	0.15	0.26	0.091	0.062^	0.21	0.088

Non significant differences (p-value between>0.085)

Marginally significant differences (^p-value between 0.05 and 0.085)

Significant differences (*p-value between<0.05)

L'information concernant la résistance aux médicaments était disponible pour 2 846 souches (c'est à dire 26,23% des isolats de la famille Beijing) avec la répartition suivante : les souches pansusceptibles, n = 1713 (60,19%) ; les souches résistantes à au moins un médicament antituberculeux, n = 320 (11,24%) ; les souches MDR, n = 736 (25,86%) ; et les souches XDR, n = 77 (2,71%). La répartition mondiale de la résistance aux médicaments entre les souches de la famille Beijing et les souches non-Beijing, dans les différentes sous-régions (la Chine et le Japon sont également indiqués), est illustrée à la **Figure 41A**. Tandis que l'évolution de la résistance aux médicaments au fil du temps pour les souches de famille Beijing, pendant une période de 14 ans (1998-2011), est résumée dans la **Figure 41B**. A travers ces données, nous avons observé une progression de la proportion de souches MDR et XDR au niveau mondial depuis 2003, et une diminution relative de souches pansusceptibles (**Figure 41B**). La tuberculose pansusceptible concerne principalement des patients âgés > 60 ans (44%) alors que la tuberculose résistante (résistance à au moins un antibiotique, MDR et XDR) concerne préférentiellement les patients âgés entre 21-40 ans (représentant 52 à 58%). Bien que la proportion de souches résistantes aux médicaments était significativement plus élevée pour les souches de la famille Beijing comparée aux souches non-Beijing à l'échelle mondiale, d'importantes variations dans la répartition de la résistance aux médicaments ont été observées (**Figure 41A**). La multirésistance aux médicaments était significativement liée à la lignée Beijing en Russie, en Asie du Sud, en Asie du Sud-est, et dans les pays européens, mais pas en Amérique, en Asie occidentale, en Chine et au Japon (malgré le fait que la proportion de souches résistantes aux médicaments était significativement plus élevée en Chine qu'au Japon).

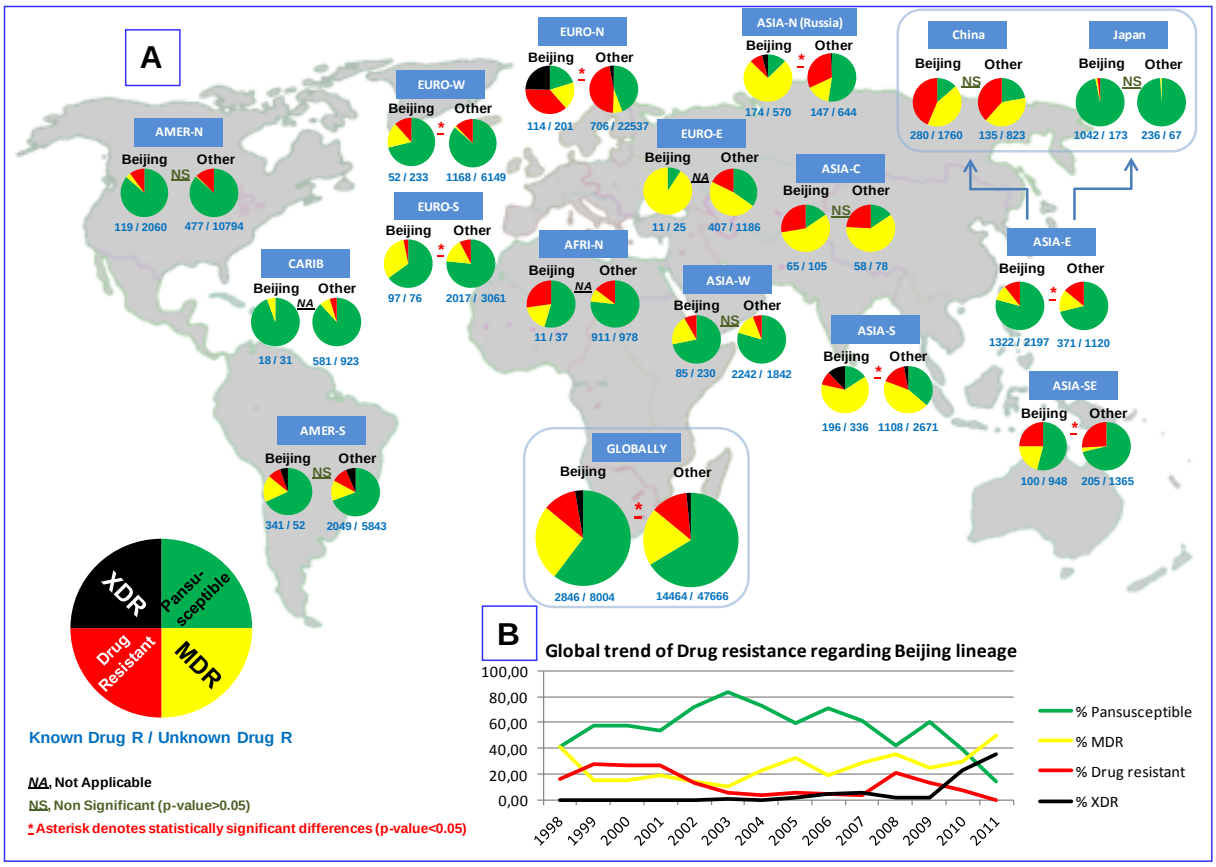


Figure 41 : (**A**) Répartition mondiale de la résistance aux antituberculeux entre les familles Beijing et non-Beijing ; et (**B**) évolution de des formes de résistance parmi la lignée Beijing, entre 1998 et 2011.

La diversité mondiale et la spécificité géographique des souches de la famille Beijing pourraient être spécifiquement visualisés grâce aux WebLogos (http://weblogo.berkeley.edu/) appliqués aux données de motifs 12-loci MIRU-VNTR disponibles (n = 1 827 souches ; **Figure 42A**). Comme le montre cette figure, la diversité allélique vue en conjonction avec le pouvoir discriminant (HGDI), met clairement en évidence certaines spécificités géographiques ainsi que la parenté des profils de 12-loci MIRU en fonction des souches dominantes de la famille Beijing. Par exemple, on peut noter la similitude des WebLogos entre la Russie, le Kirghizistan, l'Asie du Sud, et de l'Europe de l'Est; où l'on retrouve une proportion élevée de motifs MIT 16 (modèle MIRU12 : 223325153533) mais avec un HGDI faible (de l'ordre de 0,428 à 0,674 au Kirghizistan en Asie du Sud ; **Figure 42A**). D'autre part, l'HGDI était supérieur à 0,8 dans toutes les autres régions, avec une plus haute valeur de 0,997 en Chine. À l'exception de la Chine, de l'Amérique du Sud, des Caraïbes, et de la France, qui ont tous été caractérisées par des WebLogos bien spécifiques, toutes les autres régions (notamment l'Amérique du Nord, l'Europe du Nord, le Japon, l'Afrique du Sud, et l'Europe du Sud), ont été caractérisées par la prédominance du modèle MIRU12 223325173533 (MIT 17). Les pouvoirs discriminants élevés, en collaboration avec le nombre important de souches orphelines, observés en Chine (HGDI : 0,997) et au Japon (HGDI : 0,945) corroborent le fait que les souches de la lignée Beijing ont été historiquement présentes dans l'Est de l'Asie et ont divergé au fil du temps, par opposition à leur diffusion récente en Russie, qui a probablement eu lieu au cours du 20ème siècle (**Mokrousov, 2008**).

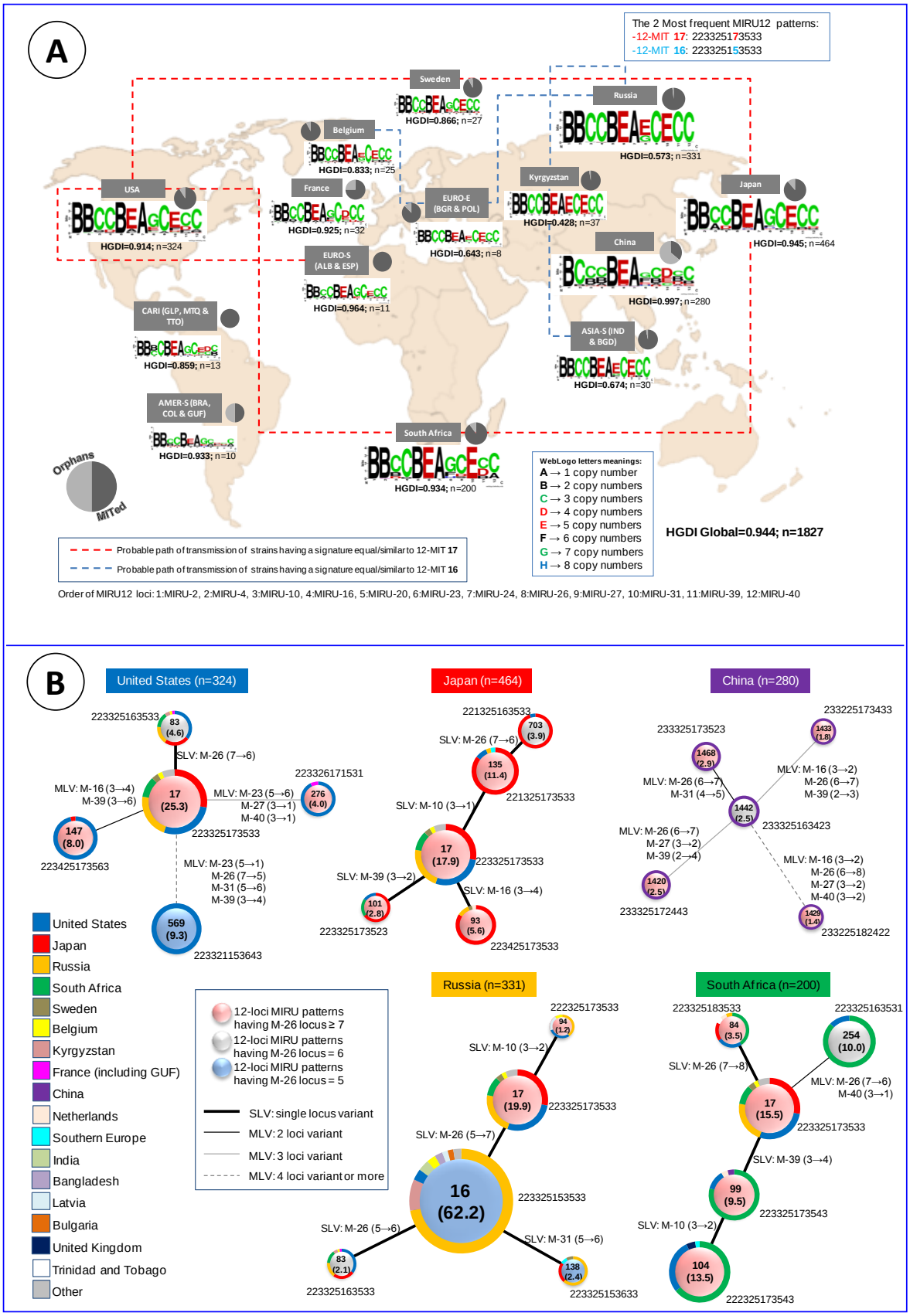


Figure 42 : (**A**) Diversité mondiale des patterns WebLogos basés sur les MIRU12. (**B**) Relations évolutives des MIRU12 prédominants dans les pays ayant plus de 100 isolats. Les MITs sont inscrits à l'intérieur des nœuds, et la couleur de nœud varie en fonction du locus MIRU26 (rose pour 7 copies, gris pour 6 copies, et bleu pour 5 copies).

Les **Figure 42B et 43** représentent des MSTs construits à l'aide de l'information de typage MIRU-VNTRs (12-, 15-, et 24-loci) en fonction du pays d'isolement des isolats. Les MSTs de la **Figure 42B** basés sur les MIRU12, ont été dessinés pour mettre en évidence les 5 motifs MIRU les plus prédominants dans les pays ayant plus de 100 isolats disponibles avec l'information de typage MIRU12 : Etats-Unis (n=324), Chine (n=280), Japon (n=464), Russie (n=331), et Afrique du Sud (n=200). La **Figure 43** met en évidence la géo-spécificité des motifs MIRU-VNTRs (12, 15, et 24-loci). On peut notamment remarquer une grande diversité au niveau des motifs (12-, 15-, ou 24-loci) isolés en Chine.

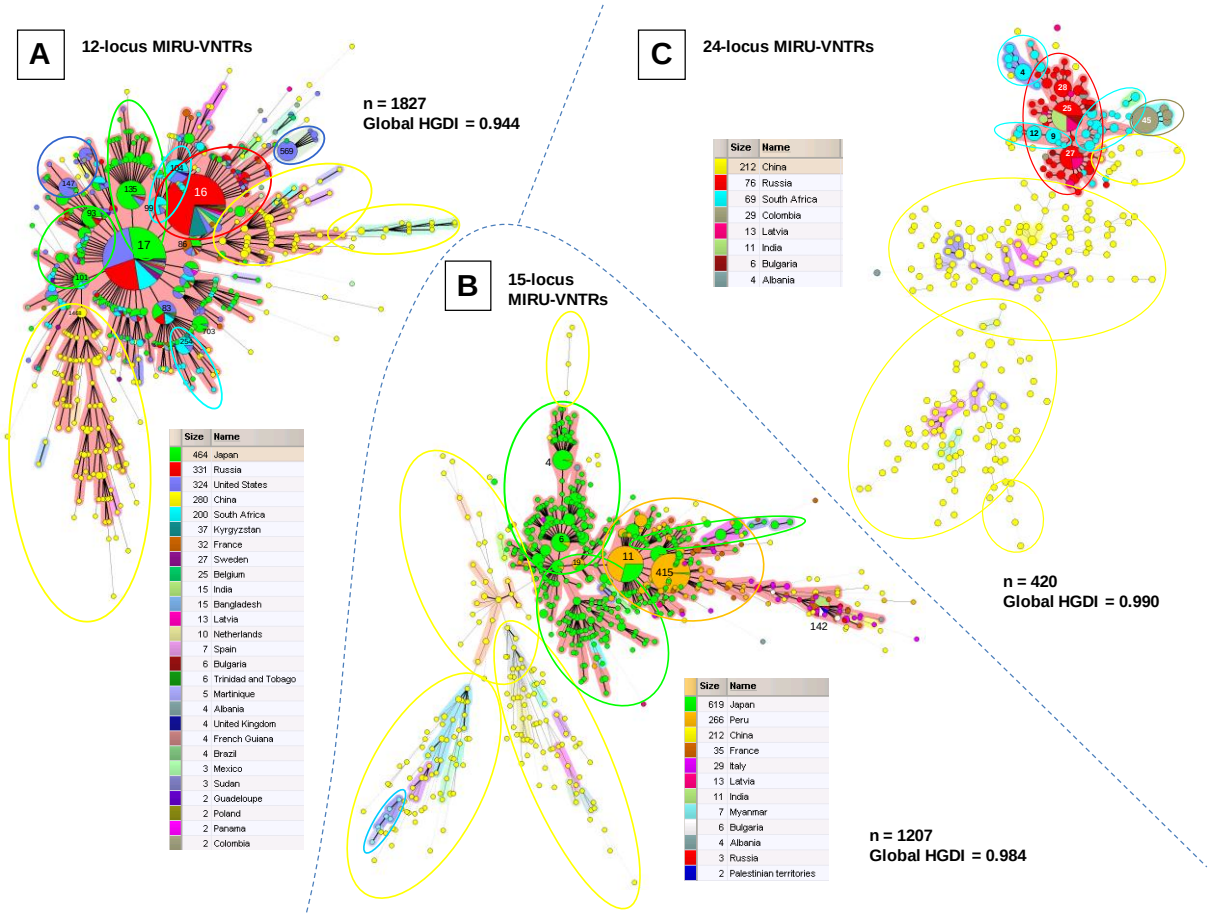


Figure 43 : MSTs illustrant les relations évolutives entre les profils 12-loci (**A**), 15-loci (**B**), et 24-loci MIRU-VNTRs (**C**), en fonction du pays d'isolement des souches (représenté par une couleur différente pour chaque pays). La taille des nœuds est proportionnelle au nombre de souches. La signification des liens est similaire à celle des arbres MSTs précédents

Pour conclure, notre étude qui se focalise sur les souches de la lignée Beijing de MTBC, a permis de bien mettre en évidence les clivages et spécificités géographiques existants au niveau démographique, épidémiologique, ou encore au niveau de la résistance aux antituberculeux. Ces résultats justifient bien l'utilité des outils moléculaires et Web pour la surveillance de la tuberculose au niveau mondial.

*1.3.4 Développement de la base SITVITBovis dédiée à la tuberculose bovine (**Article 4 en préparation**)*

Par ailleurs, j'ai également développé une base de données pour la tuberculose bovine ; nommée SITVITBovis. Nous sommes donc certains que cette base trouvera elle aussi sa place auprès de la communauté scientifique en tant qu'outil utile non seulement pour la génétique moléculaire des populations, la démographie historique et la modélisation épidémiologique de la tuberculose bovine, mais aussi pour les analyses génétiques fondamentales.

La tuberculose bovine (BTB) est un risque majeur dans les pays industrialisés dans lesquels l'élevage des animaux et le commerce des produits agricoles sont des pratiques courantes. Cependant la BTB est aussi importante dans les pays en voie de développement. SITVITBovis représente un extrait de la base de données SITVIT2. Cette portion de base de données est uniquement consacrée à l'épidémiologie mondiale de la tuberculose bovine. SITVITBovis contient des informations de génotypage sur 25741 isolats cliniques correspondant à 60 pays d'origine des patients et/ou souches animales, et 75 pays d'isolement. Les marqueurs moléculaires impliqués dans cette base sont : les spoligotypes, les MIRU12, les MIRU15, et les 5-loci ETRs (ou VNTR). La base de données SITVITBovis permet d'identifier les profils spoligotypes de tuberculose bovine, en associant notre méthode d'identification classique (attribution d'un SIT ou « Spoligo International Type »), à la méthode d'identification (nomenclature) de la base de données Mbovis.org (http://www.mbovis.org/) qui permet d'attribuer des numéros SB aux spoligotypes étudiés. Cette interface dédiée à la tuberculose bovine apporte de nombreuses fonctionnalités permettant de géo-référencer les souches, mais aussi de déterminer les spécificités des différents hôtes humains ou animaux de la maladie. Cette application est disponible en ligne à l'adresse suivante : http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis. Grâce aux outils de programmation et langages SQL et Java, nous avons pu créer une interface/application Web répondant à de nombreuses fonctionnalités, et permettant la consultation en ligne de données épidémiologique sur la tuberculose bovine (**Figure 44**).

Figure 44 : Capture d'écran de l'interface Web de SITVITBovis.

Diversité des lignées bovines dans le monde

L'application web SITVITBovis nous donne donc un aperçu sur la distribution géographique de différentes sous-lignées phylogénétiques de *M. bovis* (BOV_1, BOV_2, BOV_3, BOV_4-CAPRAE). Elle permet également la distinction entre les isolats humains et les isolats animaux afin d'aider à une meilleur compréhension des filières de transmission, et des facteurs responsables de l'expansion de la tuberculose bovine au niveau mondial.

Dans notre étude, la sous-lignée BOV_1 est globalement majoritaire, représentant 67.7% des isolats cliniques de BTB, suivie par BOV_2 (12.7%), BOV ou BOV_Like (12.3%), BOV_4-CAPRAE (6.7%), et BOV_3 (0.6% des isolats).

(i) BOV_1 est notamment prédominante de manière significative (nombre de souches ≥25) en Asie du Nord (Russie), en Europe de l'Ouest et du Sud, sur tout le continent Africain (Nord, Centre, Ouest, Sud, et Est), au Moyen-Orient (Asie de l'Ouest), en Asie du Sud, en Amérique du Nord et du Sud, avec des proportions qui oscillent entre 43% et 99% (**Figure 45**).

(ii) La lignée BOV_2 est très majoritaire en Europe du Nord (84% des isolats cliniques), en Asie de l'Est (68%), en Australasie (54%), en Amérique du Sud (35%), en Amérique du Nord (25%), en Amérique Centrale (9%), et en Afrique australe (8%) (**Figure 45**). Ce qui pourrait suggérer une relativement forte relation de commerce ou d'échange entre ces différentes régions.

(iii) La lignée BOV_3 (lignée minoritaire) est très peu visible dans notre étude. Elle est cependant bien présente en Asie de l'Est (24%), en Amérique du Nord (12%), et en Amérique du Sud (4%).

(iv) La lignée BOV_4-CAPRAE est majoritairement présente en Europe de l'Est (représentant 83% des souches). Cette lignée représente 8% des isolats cliniques en Europe de l'Ouest et du Sud, et 6% des isolats en Afrique du Nord.

(v) La lignée BOV (désignée comme BOV_Like dans les versions précédentes de nos bases de données) est prédominante en Amérique Centrale (avec 59% des isolats cliniques). Cette lignée représente aussi 46% des souches bovines en Afrique de l'Est. Les souches « BOV » sont aussi répandues en Afrique du Nord, de l'Ouest, et Central, ainsi qu'en Amérique du Nord et du Sud, en Europe de l'Ouest, et du Sud, au Moyen-Orient, et en Asie du Sud, avec des proportions allant de 12% à 23% (**Figure 45**).

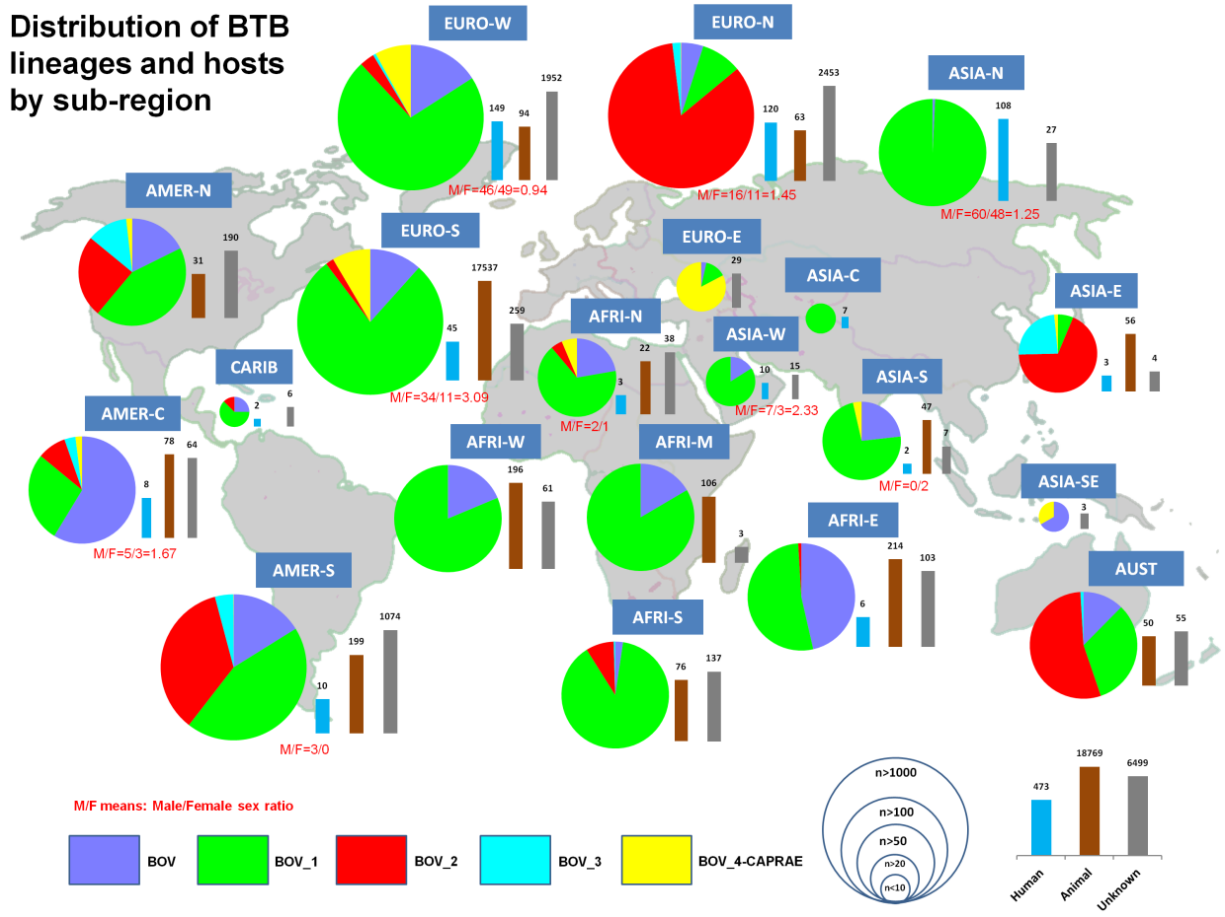


Figure 45 : Distribution phylogéographique mondiale des souches bovines et de leurs hôtes.

Distinction entre hôtes humains et animaux

Cette étude nous a permis de faire la distinction entre les souches humaines (n=473) et animales (majoritairement des bovins ; n=18769). La majorité des souches isolées d'animaux ont été retrouvées en Europe du Sud (majoritairement en Espagne). Alors que la majorité des souches BTB chez l'homme sont isolées en Europe de l'Ouest (n=149) et du Nord (n=120), suivi de la Russie (n=108) (**Figure 45**).

Nous avons aussi pu dessiner un arbre phylogénétique de type « Minimum spanning tree » (MST) permettant de mettre en valeur les profils spoligotypiques impliqués chez les hôtes humains, bovins, et les autres animaux sauvages (**Figure 46**). On peut remarquer que la majorité des souches provenant de l'homme appartiennent au profil SIT482/BOV_1 qui contient ¼ de souches humaines et ¾ de souches animales. D'autres profils spoligotypiques comme les SIT663/BOV_1 et SIT820/BOV contiennent aussi des souches humaines, mais en proportions moindres. D'autres spoligotypes isolés minoritaires et/ou orphelins concernaient également des souches humaines (**Figure 46**). Certains motifs majoritaires comme le SIT696/BOV_1, SIT695/BOV, SIT645/BOV4-CAPRAE et SIT3714/BOV4-CAPRAE, sont présents uniquement chez les souches isolées à partir des bovins. Environ 10% des souches du motif spoligotype SIT683/BOV_2 appartenaient aux isolats provenant des autres animaux sauvages.

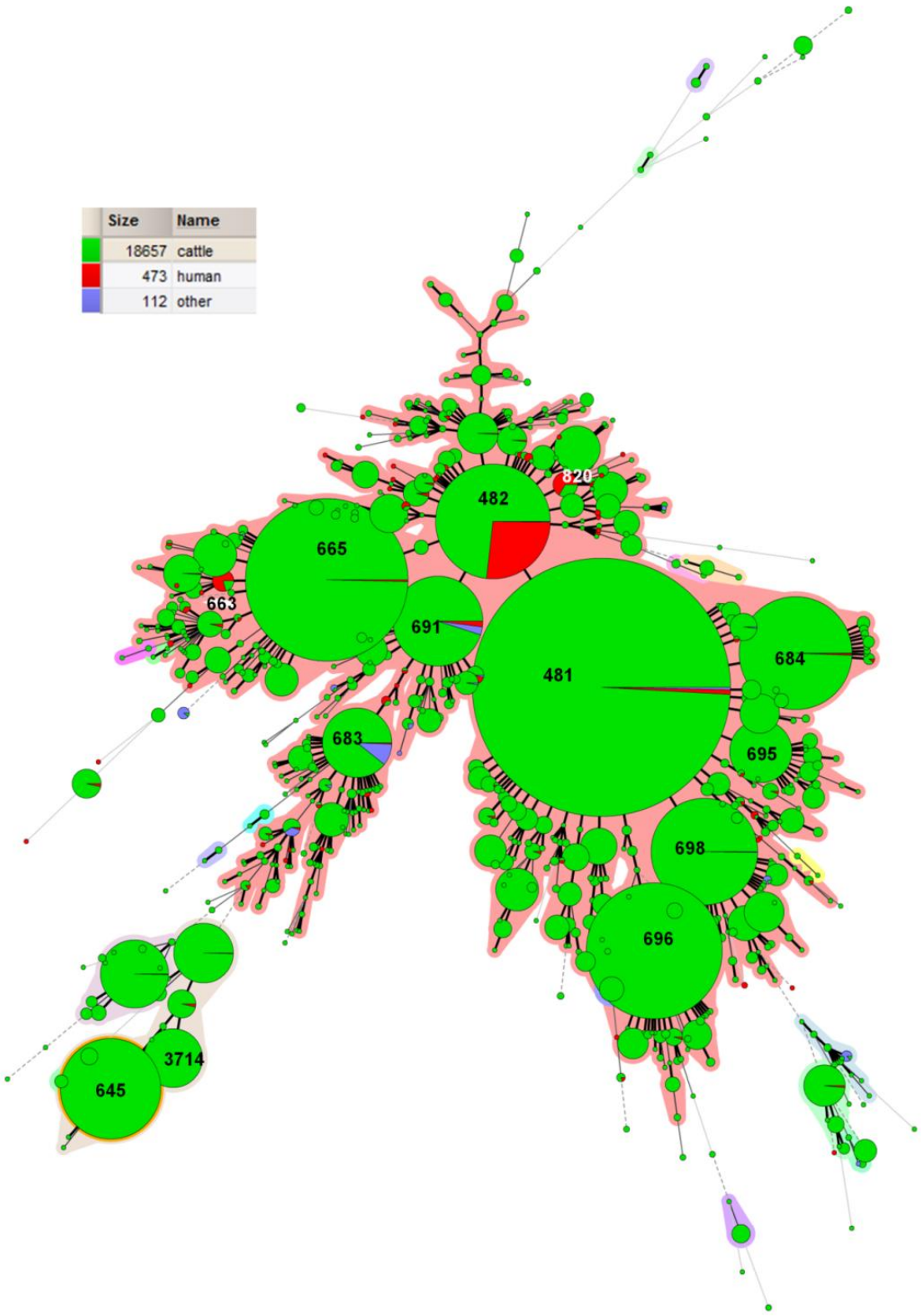


Figure 46 : MST construit avec le logiciel BioNumerics sur la base des spoligotypes, et en fonction des différents types d'hôtes : humain (en rouge), bovin (en vert), et autre animal sauvage (en bleu). La taille des nœuds (profils spoligotypes) est proportionnelle au nombre de souches partagées par un profil donné. Le numéro à l'intérieur des nœuds indique le « Spoligotype International Type » (SIT).

Distinction des profils spoligotypiques prédominants chez les animaux autres que les bovins

Le précédent MST basé sur toutes les souches ayant un hôte identifié, nous a déjà indiqué que le profil SIT683/BOV_2 contenait la plupart des isolats provenant d'animaux autres que les bovins (« wildlife animals », n=112). On se rend compte dans ce nouveau MST (**Figure 47**) que ce profil SIT683 est particulièrement prédominant chez les cochons. Les buffles (« Buffalo ») sont très visibles dans le profil SIT691/BOV_1 qui partage également une grande quantité de souches provenant de cochons. Le profil SIT481/BOV_1 est partagé à la fois par des lions, des buffles, des chèvres et des léopards, et est également retrouvé chez une hyène et un koudou (antilope). Une majorité de wapitis (« elk ») partage le profil SIT3858/BOV_1 avec un cerf de Virginie (« white-tailed deer »). Ces 2 espèces sont répandues sur le continent Américain (particulièrement dans le Nord). Ces différentes observations suggèrent une réelle transmission de la maladie entre différents animaux. Il existerait aussi une préférence géographique et ou environnementale/climatique, qui favoriserait la transmission.

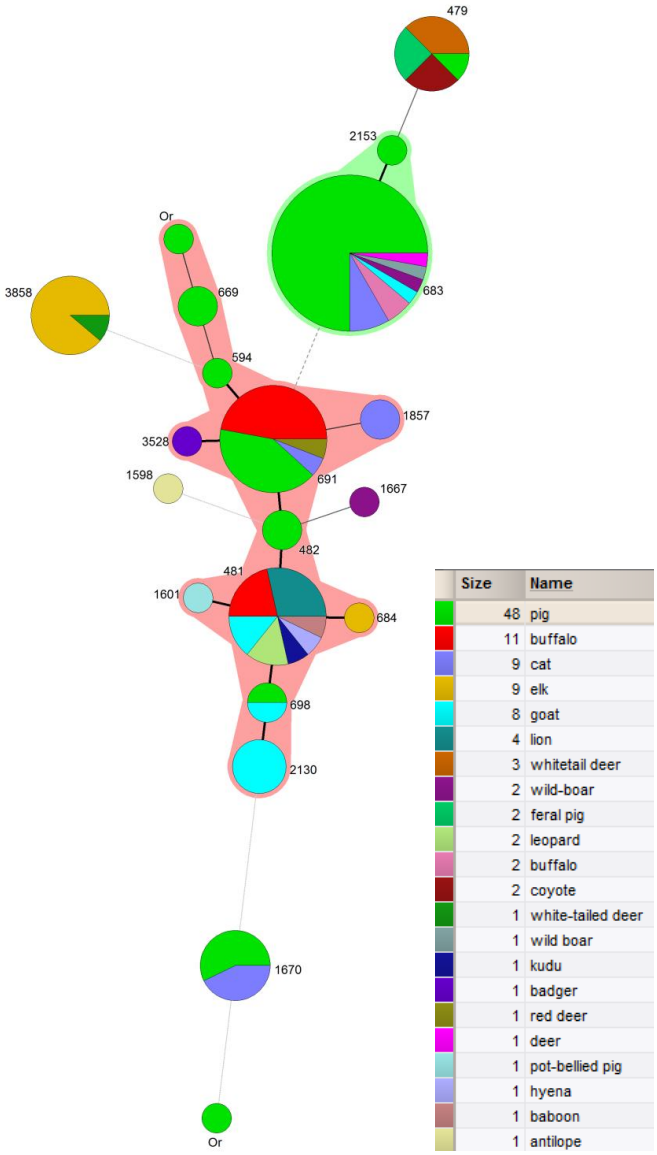


Figure 47 : MST construit sur la base des spoligotypes, en fonction de tous les animaux différents des bovins (animaux sauvages, n=112). Chaque nœud possède un numéro SIT. La taille des nœuds est proportionnelle au nombre de souches.

Tuberculose bovine et résistance aux antituberculeux

La souche *M. bovis* est naturellement résistante à la pyrazinamide dans la plupart des cas. Dans notre étude, l'information de résistance était connue pour 108 isolats cliniques. Parmi ces isolats : (i) 33 (30.56%) isolats étaient Pansusceptibles (c'est-à-dire non-résistants) ; (ii) seulement 2 souches (1.85%) étaient multirésistantes (MDR ; INH+RIF) ; (iii) 46 souches (42.59%) étaient résistantes à au moins 1 antibiotique (généralement la pyrazinamide ou PZA) ; (iv) 27 souches (25%) étaient ultrarésistantes (XDR).

Distribution des groupes d'âge des patients en Europe

La comparaison des données entre SITVITBovis (souches humaines) et SITVIT2 montre que la distribution des groupes d'âge des patients touchés par la tuberculose bovine (*M. bovis*) est différente comparée à celle des patients touchés par les pathogènes de *M. tuberculosis*. Lorsque l'on compare la distribution des groupes d'âge 0-20 ans, 21-40 ans, 41-60 ans, et >60 ans, chez les patients infectées par les souches appartenant aux sous-lignées BOV_1, H3, LAM9, T1, Beijing, CAS1-Delhi, et EAI5 (**Tableau 13**), on constate que les souches BOV_1 sont plus répandues aussi bien chez les patients jeunes (0-20 ans) que chez les patients âgés (>60 ans).

Tableau 13 : Comparaison de la distribution des groupes d'âge chez les patients porteur des souches de la sous-lignée BOV_1 par rapport aux autres sous-lignées majeures de la base de données SITVIT2.

Lineage	Number of strains with Age information	%0-20 yrs	%21-40 yrs	%41-60 yrs	%>60 yrs
Beijing	2640	9,51	39,92	22,20	28,37
BOV_1	**162**	**22,22**	**25,31**	**15,43**	**37,04**
CAS1-Delhi	1085	18,53	50,05	22,30	9,12
EAI5	393	12,21	51,15	26,46	10,18
H3	1369	11,91	37,84	22,50	27,76
LAM9	1333	11,78	46,59	25,58	16,05
T1	2641	11,09	39,57	22,95	26,39

En considérant uniquement les souches bovines isolées en Europe (au Nord, au Sud, et à l'Ouest), nous constatons une disparité dans la distribution des groupes d'âge. Pour tous les groupes d'âge (0-20 ans, 21-40 ans, 41-60 ans, et >60 ans), on observe que la proportion des patients d'Europe de l'Ouest (EURO-W) est plus ou moins similaire, se situant autour de 20% (entre 18 et 24%) sauf pour le groupe d'âge >60 ans qui représente environ 35% des patients d'EURO-W (**Figure 48**). Cependant, on observe une majorité de patients appartenant au groupe des 21-40 ans (51% des patients) pour l'Europe du Sud (EURO-S). Enfin, nous pouvons noter que la majorité des patients provenant de l'Europe du Nord (EURO-N) appartient au groupe d'âge >60 ans (81% des cas). Ces résultats semblent suggérer que la tuberculose bovine est un problème important qui touche particulièrement la population active en Europe du Sud, tandis qu'il s'agirait d'une réactivation chez les patients âgées en Europe du Nord.

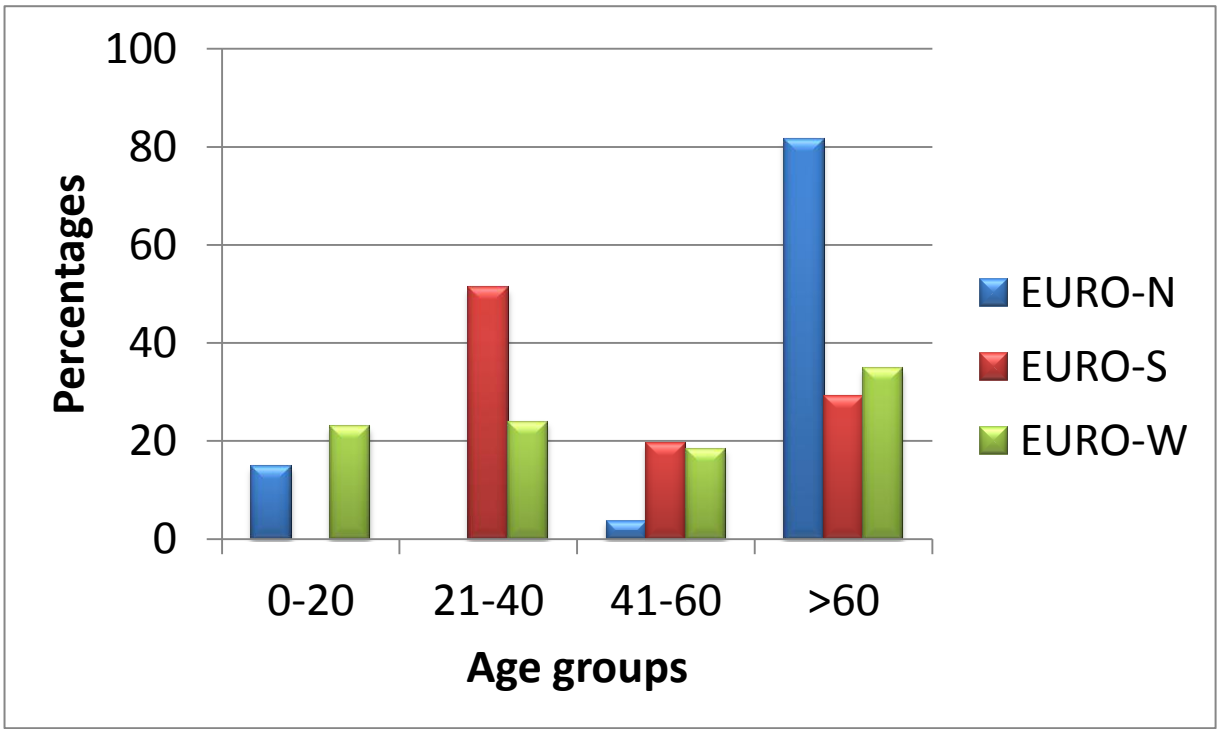


Figure 48 : Distribution des groupes d'âge concernés par la tuberculose bovine en Europe.

Quelle serait la tendance globale d'évolution de la tuberculose bovine en fonction des différents hôtes ?

Une surveillance accrue de la tuberculose bovine est nécessaire pour une meilleure compréhension de la maladie. Malgré une baisse notable de la tuberculose au niveau mondial, on aperçoit une augmentation particulièrement épisodique de la tuberculose bovine chez l'hôte humain. D'un point de vue global, dans la base de données SITVITBovis, la proportion de la BTB chez l'homme représentait 5.18%, 6.12%, 10.47%, 27.81%, et 35.82%, en 1994, 1998, 2001, 2006, et 2009 respectivement (**Figure 49**). Cependant n'ayant pas l'exhaustivité des données recueillies, nous ne pouvons pas tirer de conclusions épidémiologiques dans l'immédiat.

SITVITBovis étant un extrait de SITVIT2, tout changement réalisé dans ce dernier sera automatiquement refléter dans la base de données dédiée à la tuberculose bovine. De plus, notre approche est un peu limitée car SITVITBovis se base sur les 5-loci ETRs, 12, et 15-loci MIRU-VNTRs classiques, alors qu'il existe d'autres loci VNTR de typage pour *M. bovis*. Nous envisageons donc que l'ajout d'autres schémas de loci VNTR dans nos bases de données permettrait de mieux identifier la tuberculose bovine.

En résumé, cette base de données apporte un aperçu global de la distribution et de l'évolution des souches appartenant à *M. bovis* en fonction de divers paramètres épidémiologiques et démographiques. SITVITBovis représente ainsi une ressource importante pour la surveillance de la tuberculose bovine.

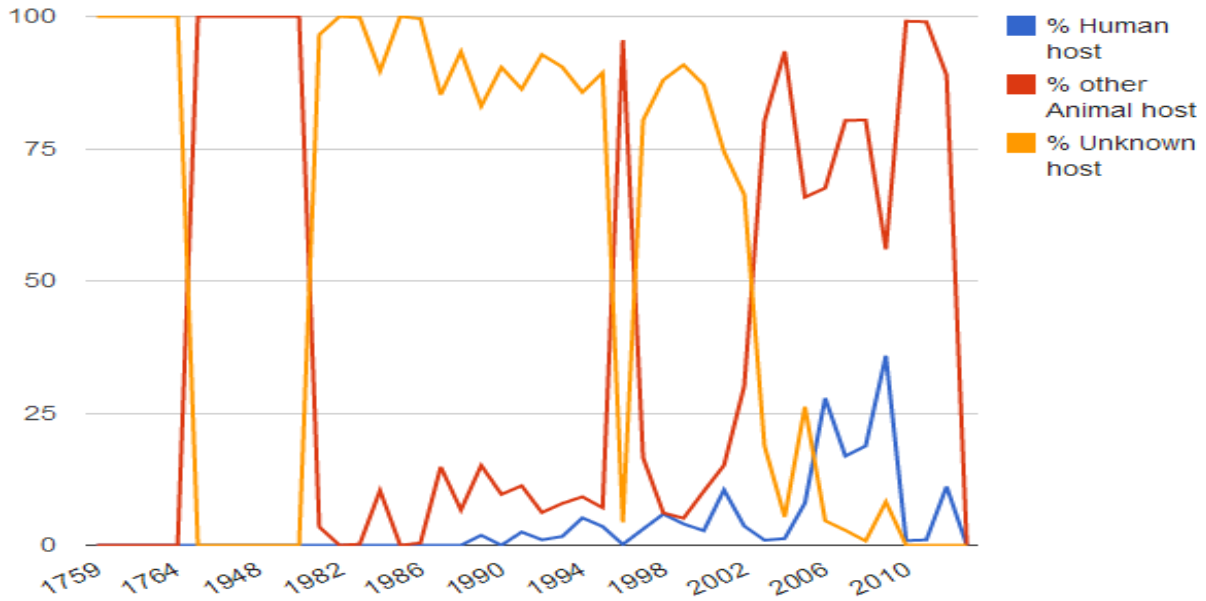


Figure 49 : Evolution temporelle globale de la tuberculose bovine en fonction des différents hôtes (http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis/stata.jsp).

*1.3.5 SITVIT-KBBN : une ressource basée sur les réseaux bayésiens (**Article 5**)*

Une nouvelle approche utilisant les réseaux bayésiens a été développée pour la classification phylogénétique des bactéries du complexe *Mycobacterium tuberculosis* (MTBC). Cet algorithme permet de prédire et d'identifier les lignées phylogénétiques à partir de nos règles de classification SITVIT2. L'application KBBN (« knowledge-based Bayesian network ») est donc basée sur des règles expertes concernant 69 lignées/sous-lignées majeures de la base de données SITVIT2. L'approche a été validée en utilisant deux bancs d'essai obtenus à partir de deux grandes bases de données de génotypage (notre base SITVIT2 et celle du CDC d'Atlanta) pour prédire les lignées/sous-lignées phylogénétiques de MTBC. La prédiction des lignées a été réalisée en utilisant le marqueur génétique de spoligotypage qui est plus souvent disponible lors de l'enregistrement des souches des patients dans les bases de données. Les résultats ont montré que l'utilisation des réseaux bayésiens dans la prédiction des lignées peut considérablement augmenter la précision de la classification lorsque les données relatives aux patients sont insuffisantes. L'outil SITVIT-KBBN est disponible en ligne: http://tbinsight.cs.rpi.edu/run_tb_lineage.html [*Il s'agit d'un travail de collaboration avec les chercheurs du Rensselaer Polytechnic Institute, Troy NY, USA, coordonnée par le Dr. Kristin Bennett*].

Un score (F-Mesure) permet de déterminer avec quelle précision les lignées phylogénétiques ont été prédites dans l'outil KBBN (**Tableau 14**). La F-Mesure a été choisie car elle capture de manière efficace la performance de notre modèle sur les ensembles de données multi-classées asymétriques étudiées. Le modèle développé est très précis car il permet de classifier correctement 94,3% de tous les spoligotypes contenu dans notre base de données (avec une valeur F de 0,93 pour les 69 lignées distinctes). La lignée T, qui est connue pour être mal définie, a souvent été confondue avec une grande variété de lignées différentes dont LAM6, T1-RUS, T3-ETH, T3-OSA, et AFRI.

Tableau 14 : F-Mesure appliquée et entrainée sur 10 633 isolats cliniques de SITVIT2.

Clade	*F*-measure	Clade	*F*-measure	Clade	*F*-measure
AFRI	0.800	H	0.736	PINI	0.750
AFRI_1	0.944	H1	0.924	PINI1	1.000
AFRI_2	0.908	H2	0.875	PINI2	0.667
AFRI_3	0.966	H3	0.915	S	0.976
Beijing	1.000	H3-Ural-1	0.873	T	0.926
BOV	0.948	H37Rv	0.958	T1-RUS2	0.778
BOV_1	0.993	H4-Ural-2	0.933	T2	0.953
BOV_2	1.000	LAM	0.947	T2-Uganda	0.991
BOV_3	0.644	LAM1	0.977	T3	0.964
BOV_4-Caprae	0.891	LAM11-ZWE	0.954	T3-ETH	0.65
Cameroon	0.929	LAM12-Madrid1	0.947	T3-OSA	0.626
CANETTI	1.000	LAM2	0.991	T4	0.988
CAS	0.937	LAM3	0.988	T4-CEU1	1.000
CAS1-Delhi	0.961	LAM4	0.970	T5	0.984
CAS1-Kili	0.973	LAM5	0.978	T5-Madrid2	1.000
CAS2	0.921	LAM6	0.856	T5-RUS1	0.949
EAI	0.982	LAM8	1.000	T-Tuscany	1.000
EAI1-SOM	0.986	Manu_ancestor	1.000	Turkey	0.928
EAI2-Manila	0.984	Manu1	0.991	X1	0.989
EAI2-Nonthaburi	1.000	Manu2	1.000	X2	0.963
EAI3-IND	0.963	Manu3	1.000	X3	0.995
EAI4-VNM	1.000	Microti	0.750	ZERO	0.800
EAI6-BGD1	0.989				
EAI7-BGD2	1.000			AVERAGE	**0.930**
EAI8-MDG	1.000				

Cependant, quelques différences de précisions de prédictions des lignées ont été observées entre les bases de données testées (SITVIT et la base du CDC). Le **Tableau 15** compare les résultats du KBBN, du BN (réseau bayésien), des règles seules (règles de classification des lignées) et du SVM (« support vector machine » non linéaire et linéaire) sur les deux bancs d'essai. Les règles de classification utilisées seules ont une précision globale très pauvre, cependant elles conduisent à améliorer de manière statistiquement significative la précision du BN sur les deux jeux de données. Les résultats indiquent une certaine cohérence entre la précision du SVM (linéaire et non linéaire) et du KBBN. Néanmoins, la méthode KBBN qui permet de générer de manière aléatoire des données observables (http://en.wikipedia.org/wiki/Generative_model), a un avantage sur la méthode SVM, notamment sur la classification prédictive des lignées phylogénétiques. En résumé, SITVIT-KBBN basée sur les réseaux bayésiens facilite grandement la classification et l'interprétation des isolats MTBC en adéquation avec les règles d'identification mises en place par les experts

et représente un outil prometteur de prédictions des lignées phylogénétiques de MTBC basées sur le spoligotypage et/ou 12 loci MIRU-VNTR.

Tableau 15 : F-mesure moyenne des méthodes KBBN, BN, Règles seule et SVM (non linéaire et linéaire) sur deux bancs d'essai : SITVIT-CV (un échantillon de SITVIT2) et CDC-Sublineage (un échantillon de la base du CDC).

Dataset	Model				
	KBBN	BN	Rules-only	SVM nonlinear	SVM linear
SITVIT-CV	0.945	**0.771**	**0.345**	**0.903**	**0.914**
CDC-Sublineage	0.981	**0.934**	**0.312**	0.994	0.993

2. Chapitre 2 – Utilisation de typage moléculaire et outils Web pour cartographier et comprendre l'épidémie au niveau mondial

2.1 Résumé du chapitre

Dans ce chapitre nous décrirons les nombreuses collaborations avec les différentes équipes du monde concernant les aspects épidémiologiques et phylogénétiques de la tuberculose. Ces collaborations ont été rendus possibles notamment grâce aux bases de données et aux outils web développés. Cependant, ils ont en même temps permis d'enrichir nos bases tout au long de ce travail. Le travail minutieux de cartographie a permis de répondre aux questions : Qui, Quand, et Où – si cher à des épidémiologistes. Comme l'épidémiologie n'est pas seulement l'étude de la répartition, nous avons également cherché à savoir les déterminants démographiques et cliniques des populations concernées. Cependant nous sommes conscients de fait qu'il s'agit souvent d'études isolées avec les biais inhérents de sélection (et d'informations disponibles car pas toujours représentatives), il s'agit tout de même d'une contribution importante qui souligne les particularités propres aux pays et régions du monde sur la survenue de la tuberculose maladie dans une population donnée.

Pour ne pas alourdir la lecture, les diverses études épidémiologiques sont résumées ici en une balade descriptive qui commence dans l'Amérique latine (le Mexique, la Colombie, le Pérou et le Brésil) ; se poursuit en Afrique (la Guinée Bissau à l'Ouest et l'Ethiopie à l'Est) et au Moyen Orient/Asie de l'Ouest (Sultanat d'Oman et l'Irak « après-guerre »); continue avec la visite de l'Asie du Sud/Sud-Est/Est (l'Inde, la Malaisie, la Chine); et se termine avec l'Europe avec des pays aussi variés que la Bulgarie, la Finlande, la Suède, la France, l'Espagne et l'Italie (**Figure 50**). Ainsi, à travers cette thèse, nous décrivons le « méta-monde » merveilleux et complexe des bacilles tuberculeux grâce à ces études collaboratives.

Figure 50 : Utilisation de typage moléculaires et outils Web pour cartographier et comprendre l'épidémie de la tuberculose au niveau mondial – chaque point dénote une étude collaborative effectuée dans le cadre de cette thèse.

2.2 Epidémiologie de la tuberculose en Amérique latine

« On ne discute pas pour avoir raison, mais pour comprendre » Borges (Jorge Luis).

« L'Amérique latine est généralement définie comme la partie de l'Amérique où l'on parle des langues dérivées du latin (ou langues romanes). Bien que le français soit une langue romane, elle n'est que rarement et très partiellement retenue pour définir l'Amérique latine. (…) D'autre part, l'examen attentif de l'histoire et de la culture de l'Amérique latine montre que cette dernière ne constitue pas un ensemble culturellement homogène : l'expression « Amérique latine » efface le passé précolombien de la région, pour ne s'attacher qu'à la langue des colonisateurs. La définition de l'Amérique latine est cohérente avec le concept du « partage du monde » signé en 1494 par l'Espagne et le Portugal lors du Traité de Tordesillas. » (http://fr.wikipedia.org/wiki/Am%C3%A9rique_latine).

Bien que les lignées LAM, Haarlem et T prédominent dans cette région du globe, nous avons remarqué de nombreuses spécificités en fonction des différents pays d'Amérique latine. Nous notons que la lignée LAM est très majoritaire au Venezuela (représentant près de 69% des isolats cliniques), au Honduras (53% des isolats), au Paraguay (50%), en République Dominicaine (42%), au Brésil et en Colombie (40% chacun). Les proportions de la lignée LAM sont moins élevées dans les autres pays de cette zone (**Figure 51**). En considérant ces pays d'Amérique latine, nous constatons que la lignée EAI est très visible au Suriname où elle prédomine largement (36% des isolats cliniques), au Guyana (19% des isolats), au Mexique (12% des souches), et en Guyane Française (11% des souches). Nous remarquons aussi que la famille Beijing est répandue à Cuba (11% des souches), au Panama (6% des isolats) et au Pérou (5%). Bien que les souches bovines (lignée Bovis) récoltées ne proviennent pas nécessairement d'hôtes humains (certaines souches étant collectées à partir d'autres animaux), elles représentent un risque potentiel en Argentine (53%), au Mexique (13%) ainsi qu'au Brésil (11% des isolats). Nous notons également la forte présence de la lignée X au Mexique (15% des isolats), en Haïti et en Guadeloupe (10% des isolats), mais aussi dans d'autres îles

des Caraïbes « latines », excepté Cuba où cette lignée est pratiquement invisible. Cette absence de lignée X à Cuba peut être due au passé historique et aux relations paradoxales qui existent entre ce pays et les Etats-Unis (http://fr.wikipedia.org/wiki/Relations_entre_Cuba_et_les_%C3%89tats-Unis).

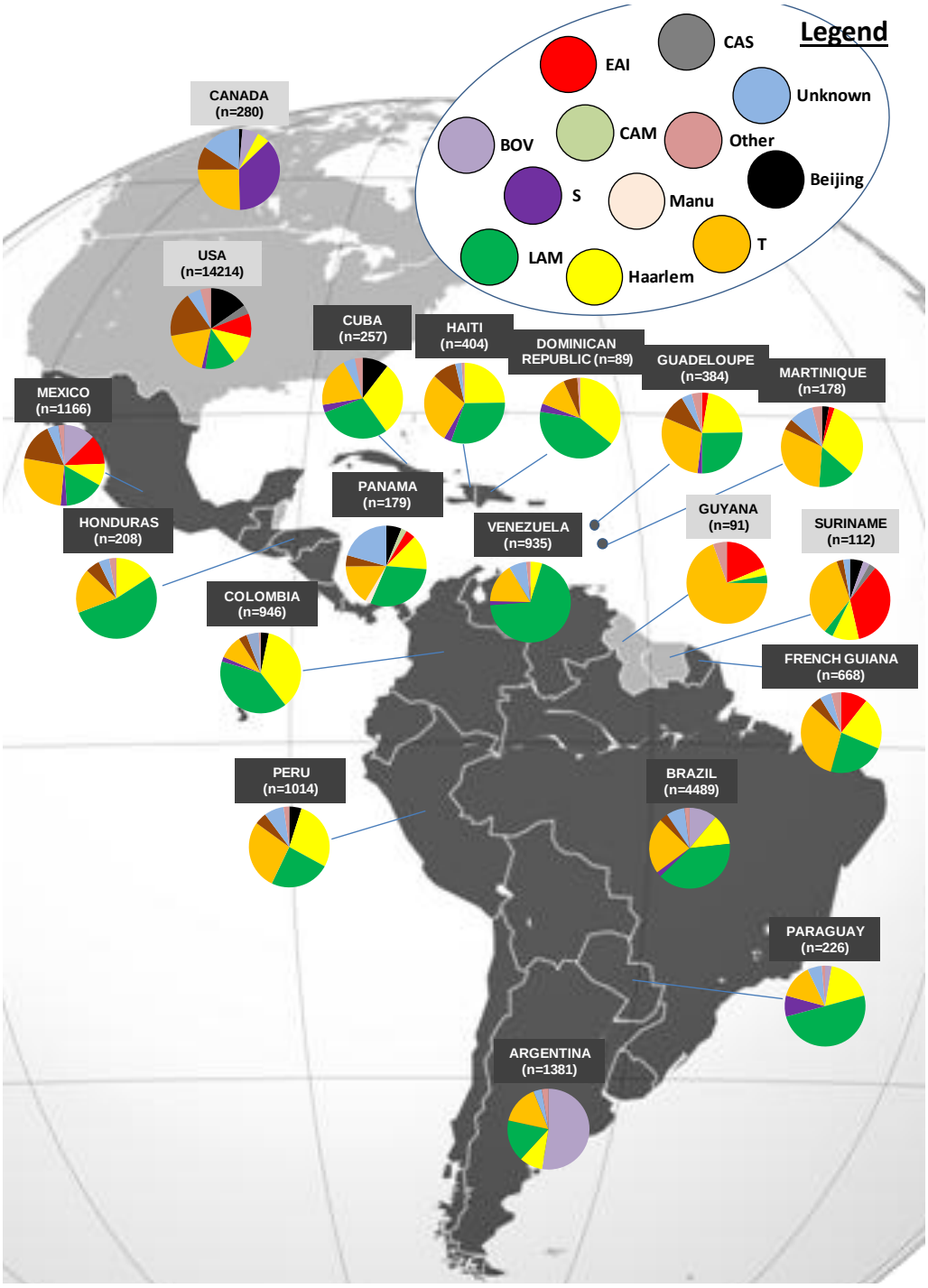


Figure 51 : Distribution des lignées phylogénétiques majeures en Amérique latine (pays indiqués en gris foncé) selon la base de données SITVIT2.

Dans notre redécouverte de l'Amérique latine, nous avons plus spécialement visité le Mexique (**Article 6**), la Colombie (**Article 7**), le Pérou (**Article 8**), et le Brésil (**Article 9**).

*2.2.1 Etude épidémiologique au Mexique (**Article 6**)*

« Pendant cette période, l'Espagne s'est enrichie grâce à l'exploitation des mines d'argent mexicaines, alors les plus riches du monde, dont le produit transite via Anvers, première place financière mondiale, pour permettre d'importer des biens de l'Inde, où les marchands sont friands d'argent-métal. Les espagnols implantent aussi la culture de la canne à sucre et du café, alors que sur le plan humain, la population amérindienne chuta de 80%, à cause des épidémies et des travaux forcés. On estime qu'avant l'arrivée des Espagnols, le Mexique central comptait 25 millions d'habitants. Il en restait un million vers 1650 » (http://fr.wikipedia.org/wiki/Mexique).

Présentation de l'étude

Dans notre étude (**Article 6**), nous nous sommes intéressés de manière spécifique à la diversité génétique des souches multi-résistantes (MDR) de MTBC. La tuberculose est un problème de santé publique important au Mexique. Cependant, nous disposons de peu d'informations sur la diversité génétique des souches de MTBC qui circulent dans le pays. Dans ce travail, nous avons analysé 109 isolats MDR recueillies dans 23 états différents du Mexique en 2003. Ces isolats ont été rétrospectivement caractérisés par spoligotypage et typage 12-loci MIRU-VNTR. Les résultats de cette étude suggèrent que la majorité des cas n'étaient pas liés au niveau épidémiologique.

Résultats principaux

Les 109 isolats récoltés au cours de cette étude ont été testés avec la méthode DST (« Drug Susceptibility Testing ») afin de confirmer leur état de résistance. Tous les 109 isolats étaient résistants au moins à l'isoniazide (INH) et la rifampicine (RIF). En plus du test de résistance à l'INH et la RIF, le test de résistance à la streptomycine (STR), l'éthambutole (ETH), et la pyrazinamide (PZA), a aussi été effectué. Par ailleurs, 32/109 isolats (29.4%) étaient uniquement résistants à l'INH et la RIF ; 31/109 isolats (28.4%) étaient simultanément résistants à 3 antibiotiques ; 24/109 isolats (22%) étaient simultanément résistants à 4 antibiotiques ; et 24/109 isolats (20.2%) étaient simultanément résistants à 5 antituberculeux.

Dans ce travail, le profil SIT53 (de la famille T1) représentait le motif prédominant pour 34/109 (31.2%) isolats. Les lignées majeures représentées dans l'étude étaient les lignées T et LAM qui représentaient respectivement environ 60% et 11% des souches. Tandis que d'autres lignées comme les familles Beijing, EAI, et Haarlem, étaient minoritaires dans notre étude.

Nous avons notamment construit un MST, en utilisant la version 6.6 de BioNumerics (http://www.applied-maths.com/bionumerics), qui illustre bien l'évolution entre les souches des motifs de la famille Beijing typés en utilisant les 12-loci MIRU de notre étude (n=3 souches de la famille Beijing) par rapport à 327 isolats des USA (**Figure 52**). Ce MST montre l'unicité et la particularité des motifs basés sur le typage 12-loci MIRU-VNTRs de la famille Beijing au Mexique par rapport aux souches collectées aux USA. Nous pouvons remarquer que le profil identifié « or_MEX2 » est particulièrement bien éloigné des autres profils.

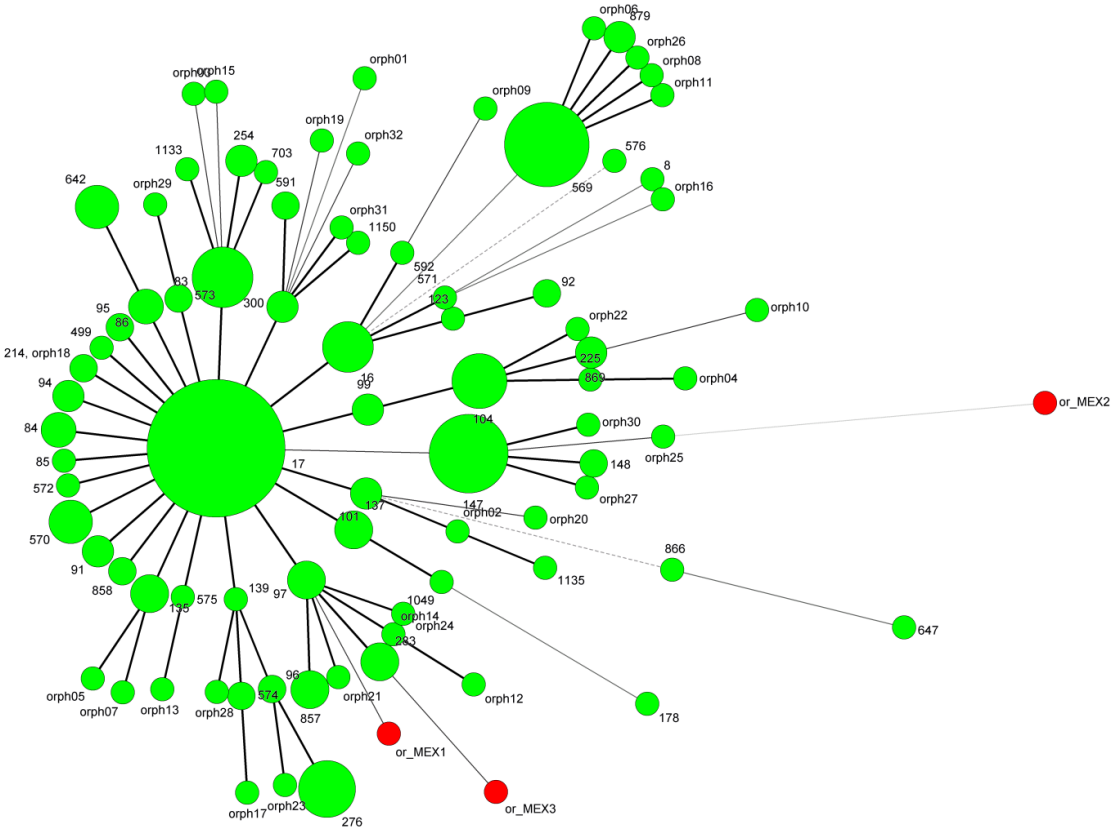


Figure 52 : MST illustrant la diversité des profils 12-loci MIRU du Mexique (souches rouges) par rapport aux profils des Etats-Unis (souches vertes). La taille des nœuds est proportionnelle au nombre d'isolats. Lorsque la distance évolutionnaire entre 2 profils est grande, les arêtes sont en pointillés, alors que lorsque la distance entre 2 motifs est petite, les arêtes sont représentées en trait épais.

Enfin, cette étude nous a montré que le motif spoligotype SIT53/T1 était majoritaire parmi ces 109 souches multi-résistantes, ce qui pourrait une forte association de ce profil avec la tuberculose MDR. En outre, les régions du sud des États-Unis sont également concernées par la circulation des profils SIT53. Cette situation pourrait s'expliquer par l'importante population d'origine mexicaine résidant dans cette région (**Soini et coll., 2000**). Néanmoins, d'autres études ont montré que le SIT53 a été plus rarement distribué dans les états se situant au long de la frontière américano-mexicaine malgré la présence d'un nombre important de personnes d'origine hispanique dans ces états (**Chiang et coll., 2009**).

Conclusion

En résumé, il s'agit de l'étude moléculaire la plus complète de la tuberculose multirésistante réalisée au Mexique. La caractérisation moléculaire menée, améliore notre compréhension de la transmission de la tuberculose dans ce pays, bien que l'utilisation d'une technique plus discriminante telle que le typage 24-loci MIRU-VNTR (MIRU24) serait plus adaptée pour mieux caractériser les isolats du complexe *M. tuberculosis* circulants.

*2.2.2 Etude épidémiologique en Colombie (**Article 7**)*

« La Colombie est ethniquement diverse. L'interaction entre les descendants des premiers habitants indigènes, les colons espagnols, les populations africaines déportées dans le pays comme esclaves et l'immigration du 20ème siècle venue d'Europe et du Moyen-Orient, ont produit un patrimoine culturel varié, également influencé par la grande variété géographique du pays. La majorité des centres urbains sont situés dans les hauts plateaux de la cordillère des Andes, mais le territoire colombien englobe également la forêt amazonienne, les Llanos et les côtes Caraïbes et Pacifique. Écologiquement, la Colombie est l'un des 17 pays mégadivers du monde (http://fr.wikipedia.org/wiki/Pays_m%C3%A9gadivers), et est considéré comme le plus mégadivers par kilomètre carré. » (http://fr.wikipedia.org/wiki/Colombie).

Présentation de l'étude

Dans cette étude (**Article 7**), nous nous sommes employés à caractériser les souches de MTBC qui circulent en Colombie, le 3ème pays le plus peuplé d'Amérique latine. De nombreuses méthodes de génotypage ont été utilisées pour mener à bien ce travail : le spoligotypage, l'IS6110-RFLP, et le typage 24-loci MIRU-VNTR. Un total de 414 isolats cliniques ont été isolés à partir de patients adultes dans 3 différentes villes de Colombie (Medellin, Cali, et Popayan) entre les années 2005 et 2008. Ce travail a permis de décrire de manière très précise les souches de MTBC responsables de la transmission de la tuberculose en Colombie.

Résultats principaux

Les lignées LAM et Haarlem étaient prédominantes dans cette étude, et nous pouvons noter l'importance particulière des profils spoligotypes SIT42/LAM9 et SIT62/H1 qui représentaient tous deux 53.3% des isolats. Ce travail nous a aussi permis de remarquer qu'il existait certaines associations entre les données de génotypage, et d'autres caractéristiques tels que la multi-résistance aux antituberculeux et la séropositivité due au virus VIH/SIDA. Nous avons notamment pu mettre en relief l'association du SIT45/H1 avec les souches MDR en comparant avec les SITs prédominants (>2%) de l'étude (valeur $P < 0.001$). De plus, une différence significative a été observée entre les SITs prédominants et la sérologie VIH positive (avec une valeur $P = 0.044$). La proportion de patients VIH-positifs s'est révélée élevée parmi les isolats appartenant aux motifs SIT53/T1 (2/13 VIH-positifs) et SIT727/H1 (2/19 VIH-positifs). La différence entre les SITs prédominants et les 3 plus grands nombres de copie (bandes 8, 9, et 11) d'IS6110-RFLP était aussi significative (valeur $P = 0.011$). Cependant, aucune différence significative n'a été observée lorsque l'on a comparé les sex-ratios parmi tous les SITs prédominants (valeur $P > 0.5$) (**Figure 53**).

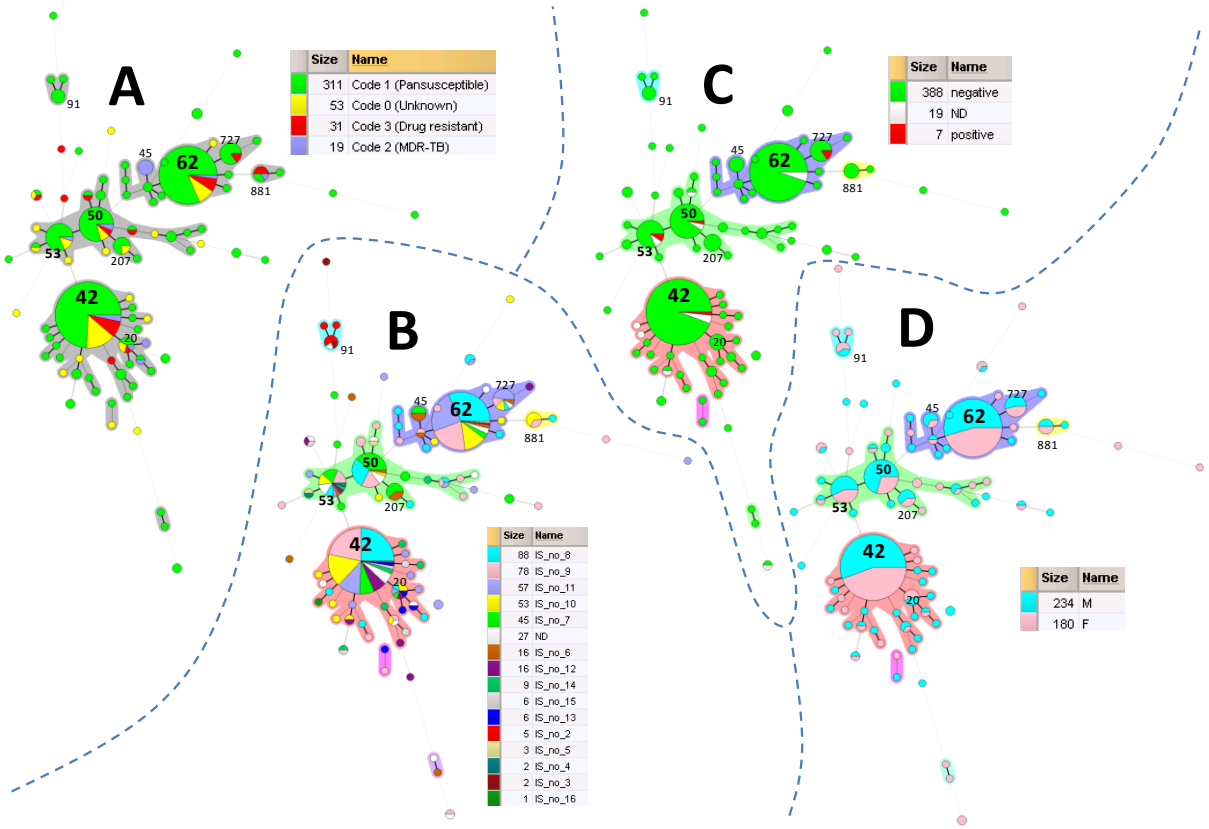


Figure 53 : MSTs illustrant la diversité des souches « spoligotypées » en fonction de plusieurs paramètres : (**A**) La résistance aux antituberculeux ; (**B**) L'IS6110-RFLP ; (**C**) La sérologie VIH ; (**D**) La différence de sex-ratio. Comme pour les MSTs précédents, la taille des nœuds indiquent le nombre de souches, et l'épaisseur et la couleur des arêtes indiquent la distance entre les nœuds.

Une cartographie globale des souches tuberculeuses colombiennes a permis de mettre l'accent sur la diversité phylogénétique et phylogéographique existant au sein même du pays, en considérant les différentes villes/régions étudiées (Medellin, Cali, et Popayan). Nos données ont montré que la lignée LAM représentait 39,6% des isolats de Medellín (Etat libre d'Antioquia), 39,1% des isolats de Cali (Etat de Valle) et 24,0% des isolats de plusieurs villes dans l'état de Cauca (**Figure 54A**). La lignée Haarlem comprenait 48,7% du groupe d'isolats de Medellín et 39,0% du groupe d'isolats de Cali, mais elle n'était pas représentée parmi les isolats de Cauca. En revanche, la lignée T « mal définie » représentait 40,0% des isolats de Cauca, incluant la ville de Popayan (la ville principale), alors que la proportion d'isolats de cette lignée était bien moins élevée à Medellín (6,4%) et à Cali (11,0%). Aucun isolat appartenant à la famille Beijing n'a été identifié dans cette étude. Les données de génotypage disponibles pour deux autres villes de Colombie (Buenaventura et Bogotá) dans la base de données SITVIT2, ont montré une prédominance des isolats appartenant aux lignées LAM et Haarlem. D'autres lignées telles que T, X et S sont moins représentées dans ces deux villes. Cependant, contrairement au résultat de notre étude, la lignée Beijing était fréquente dans la ville de Buenaventura (représentant près de 14% des isolats de cette ville, selon la base de données SITVIT2).

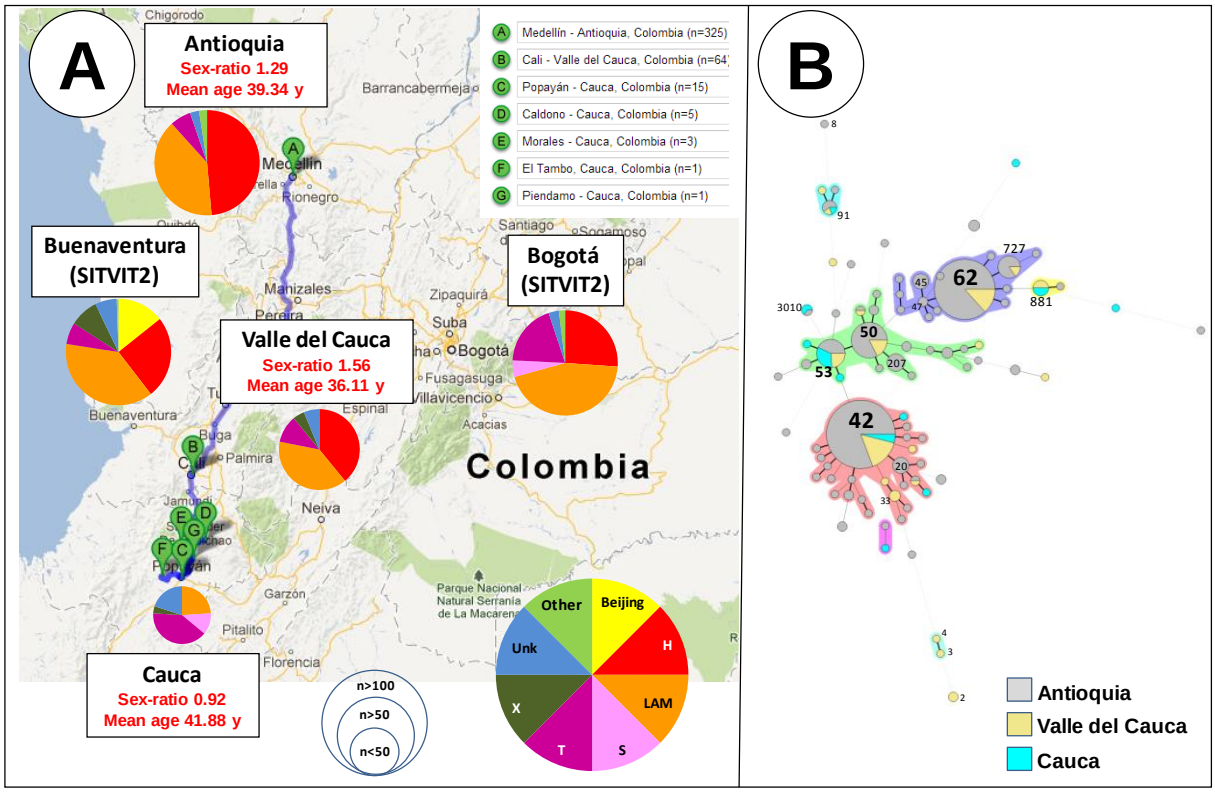


Figure 54 : (**A**) Distribution phylogéographique des souches de MTBC observées dans cette étude, et (**B**) MST basé sur le spoligotypage selon la ville/région d'origine des patients.

Le MST de la **Figure 54B** a révélé une relation évolutionnaire étroite entre les principaux spoligotypes trouvés dans cette étude, soit le SIT42/LAM et le SIT62/Haarlem (principalement retrouvés dans les états d'Antioquia et de Valle). En revanche, les isolats de l'état de Cauca ont été plus associés avec la lignée T « mal définie ». En effet, le SIT53/T1 a été reporté en plus grande proportion dans l'état de Cauca que dans d'autres régions du pays. De plus, nous avons trouvé une différence significative en comparant la distribution des lignées d'Antioquia, de Valle, et de Cauca dans cette étude (P < 0,001).

Conclusion

Pour finir, cette étude a révélé une dominance des lignées LAM et Haarlem du complexe MTBC dans plusieurs régions de la Colombie, en particulier dans les deux grands milieux urbains (Medellín et de Cali). Les motifs spoligotypes dominants étaient le SIT 42/LAM et le SIT62/Haarlem. La proportion de la lignée Haarlem était plus élevée en Colombie par rapport aux pays voisins (**Tableau 16**), ce qui suggère des conditions particulières de coévolution de cette lignée avec la population humaine.

Tableau 16 : Distribution des proportions des SITs prédominants dans cette étude, par rapport aux pays voisins : Venezuela (n=935), Brésil (n=4556), et Pérou (n=1296). L'Astérix (*) indique une différence significative.

						Distribution in neighboring countries					
SIT	Spoligotype Description	Octal Code	This study n/t (%)	COLOMBIA without study n/t (%)	p-value	VENEZUELA n/t (%)	p-value	BRAZIL n/t (%)	p-value	PERU n/t (%)	p-value
20	■■□■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	677777607760771	7/414 (1.69)	6/432 (1.39)	0.7212	42/935 (4.49)	0.0112*	147/4556 (3.23)	0.0843	5/1296 (0.39)	0.0056*
42	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	777777607760771	124/414 (29.95)	93/432 (21.53)	0.005*	111/935 (11.87)	<0.0001*	401/4556 (8.80)	<0.0001*	73/1296 (5.63)	<0.0001*
45	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□■□□□□■■■■■■■■	777777764020771	6/414 (1.45)	0/432 (0.00)	0.013*	0/935 (0.00)	<0.001*	9/4556 (0.20)	<0.0001*	0/1296 (0.00)	<0.0002*
50	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777777777720771	34/414 (8.21)	14/432 (3.24)	0.0018*	13/935 (1.39)	<0.0001*	239/4556 (5.25)	0.0112*	160/1296 (12.35)	0.0210*
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777777777760771	21/414 (5.07)	28/432 (6.48)	0.3805	49/935 (5.24)	0.8978	322/4556 (7.07)	0.1252	118/1296 (9.10)	0.009*
62	■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	777777774020731	97/414 (23.43)	68/432 (15.74)	0.0048*	5/935 (0.53)	<0.0001*	1/4556 (0.02)	<0.0001*	1/1296 (0.08)	<0.0001*
91	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■■	700036777760771	5/414 (1.21)	7/432 (1.62)	0.6119	2/935 (0.21)	0.0314*	4/4556 (0.09)	<0.001*	25/1296 (1.93)	0.3305
207	■■■■■□■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■■	767777777720771	8/414 (1.93)	0/432 (0.00)	0.0032*	0/935 (0.00)	<0.0001*	2/4556 (0.04)	<0.0001*	1/1296 (0.08)	<0.0001*
727	■■■■■■■■■■■■□■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	777737774020731	13/414 (3.14)	10/432 (2.31)	0.4606	1/935 (0.11)	<0.0001*	0/4556 (0.00)	<0.0001*	0/1296 (0.00)	<0.0001*
881	■■■■■■■■■□□■■■■■■■■■■■■■□□□□□□□□□□□□■■■□■■■	776377770000731	6/414 (1.45)	10/432 (2.31)	0.3555	1/935 (0.11)	0.0042*	0/4556 (0.00)	<0.0001*	0/1296 (0.00)	<0.0002*

*2.2.3 Etudes épidémiologiques au Pérou (**Article 8 + travail en cours**)*

« Après avoir connu une période d'hyperinflation au cours des années 1980, la monnaie péruvienne connaît, quant à elle, une période de stabilité par rapport au dollar et aux monnaies européennes. Le taux d'inflation moyen s'est stabilisé autour de 3%, et reste depuis plusieurs années dans les limites de la fourchette fixée par la Banque centrale du Pérou (entre 1% et 3%). Le taux de dollarisation de l'économie reste toutefois élevé, s'établissant à 60% des crédits en 2006 au secteur privé contre 82% en 2000. Parmi les grandes orientations financières, figure un axe majeur : la restructuration de la dette publique tant extérieure qu'intérieure. Elle a permis de voir passer la part de la dette interne en 5 ans de 22% à 29%, traduisant la confiance des marchés dans les obligations d'État. La dette publique globale, à la fin septembre 2007, atteint 31% du PIB. En termes de compétitivité, le Pérou est considéré aujourd'hui comme la première économie d'Amérique latine. Selon l'IMD Business School en Suisse (http://www.imd.org/), le Pérou est en 35ème place au classement de la compétitivité mondiale (2008). » (http://fr.wikipedia.org/wiki/P%C3%A9rou).

Présentation de la première étude

Dans notre première étude (**Article 8**), nous avons mené une analyse épidémiologique basée sur le spoligotypage de 794 souches de bacilles tuberculeux de patients (une souche par patient) provenant de la ville de Lima (capitale du Pérou) entre 1999 et 2005. Ce travail nous a notamment permis de réaliser des analyses phylogénétiques, et d'explorer les relations phylogénétiques pouvant exister entre l'infection VIH/SIDA et la résistance aux antituberculeux.

Résultats principaux

Dans notre analyse, la moyenne d'âge des patients était de 31.9 ans, 63% d'entre eux étaient de sexe masculin, et 30.4% d'entre eux étaient VIH-positifs. La mono-résistance à la RIF concernait 4.7% des souches, alors que la mono-résistance à l'INH touchait 8.7%, et les souches MDR représentaient 17.3% dans l'étude. Un total de 498 souches (62.7% de tous les isolats) formait des clusters (profil présent au moins en 2 exemplaires) à partir de 9 profils

spoligotypes prédominants dans notre étude : SIT50/H3 (16.4% des isolats), SIT53/T1 (12.3% des isolats), SIT33/LAM3 (8.3%), SIT42/LAM9 (7.4%), SIT1/Beijing (5.5%), SIT47/H1 (3.9%), SIT222/Unknown (3.0%), SIT1355/LAM (3.0%), et SIT92/X3 (2.8%). Par ailleurs, un Spoligoforest basé sur l'algorithme de Fruchterman-Reingold a été construit avec l'application SpolTools (http://www.emi.unsw.edu.au/spolTools) (**Reyes et coll., 2008** ; **Tang et coll., 2008**) pour mettre en évidence les liens probables de parenté qui peuvent exister entre les motifs spoligotypes (évoluant généralement par perte d'un ou plusieurs espaceurs), et permettre l'identification des spoligotypes orphelins (présents en un seul exemplaire) dans notre étude (**Figure 55**).

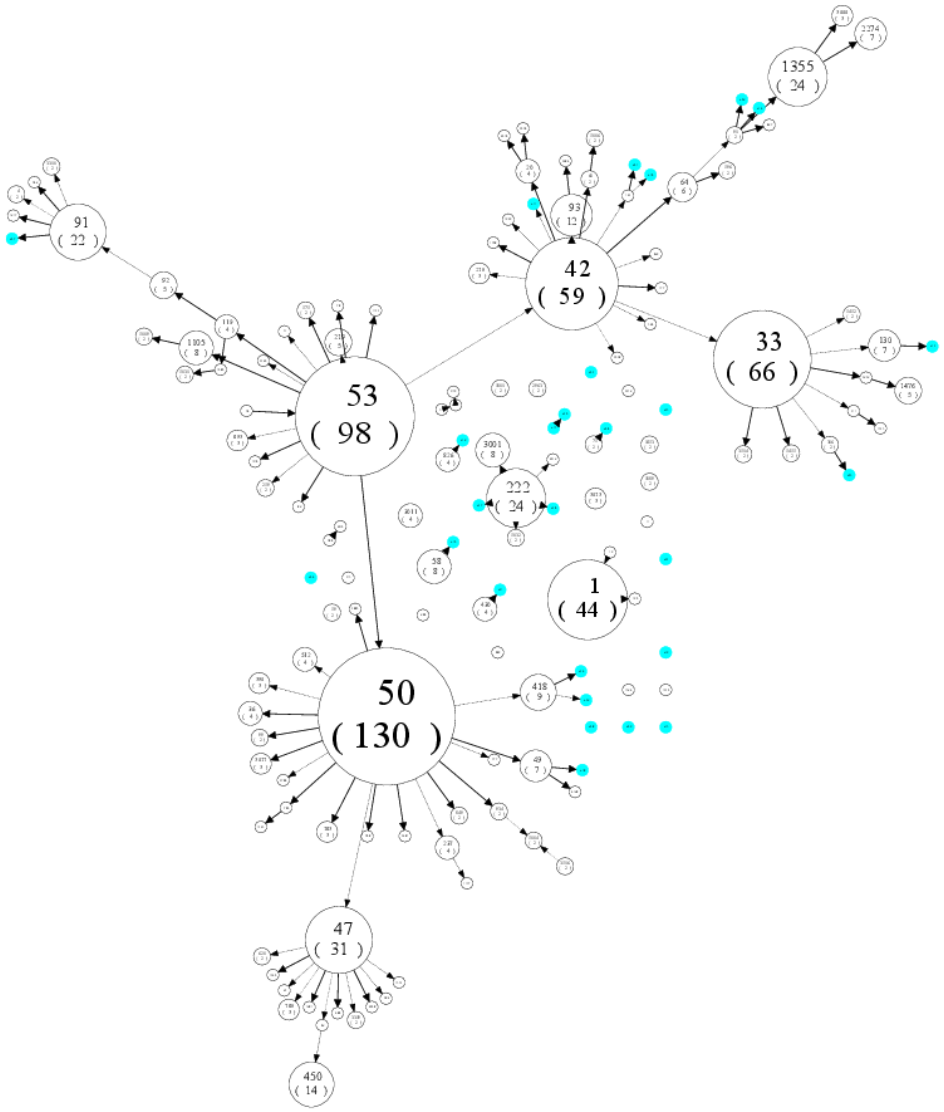


Figure 55 : Spoligoforest basé sur l'algorithme de Fruchterman-Reingold basé sur tous les spoligotypes de l'étude (n=794). Les flèches des arêtes pointent vers les spoligotypes descendants (ceux qui ont perdu des espaceurs par rapport à leurs antécédants). La taille des nœuds est proportionnelle au nombre d'isolats. Les lignes noires (épaisses) relient les motifs spoligotypes qui sont très semblables (le poids maximum de ressemblance étant de 1,0 symbolisé par la perte d'un seul espaceur), tandis que les lignes séparées par des tirets représentent les liens de poids compris entre 0,5 et 1, et les lignes pointillées un poids de moins de 0,5. Les souches orphelines sont indiquées en cyan.

Un arbre MST a été construit avec BioNumerics (version 3.5) pour montrer les relations phylogénétiques existantes entre les spoligotypes de notre étude. Cet arbre permet d'identifier les familles phylogénétiques qui prédominent, notamment le groupe PGG2/3 contenant les lignées Haarlem, LAM, T, X, et S, qui représentent 83.8% des isolats cliniques. Les lignées Haarlem (représentant 228/794, soit 28.7% des isolats), LAM (représentant 225/794, soit 28.3% des isolats), et T (161/794, ou 20.3%) étaient particulièrement visibles dans notre étude (**Figure 56**).

Figure 56 : MST (réalisé avec BioNumerics version 3.5) représentant la diversité phylogénétique des spoligotypes de notre étude (n=794 isolats). La couleur des nœuds indique le nombre de souches (plus la couleur est foncée et plus il y a de souches). La signification des liens est pareille aux arbres MSTs précédents.

Le **Tableau 17** montre l'association des spoligotypes majeurs avec la sérologie VIH/AIDS et la résistance aux antibiotiques (MDR vs. non-MDR). Parmi les patients co-infectés avec le VIH, le spoligotype SIT42/LAM9 était très associé avec la tuberculose MDR (multi-résistante), en effet, 87.5% des patients VIH-positifs ayant le profil SIT42 avaient aussi la tuberculose MDR. De plus, 60.0% des souches des patients VIH-positifs ayant le profil SIT53/T1 étaient également MDR.

Tableau 17 : Fréquence des SITs dominants par rapport au status VIH des patients et à la résistance aux antituberculeux (MDR vs. non-MDR).

	HIV negative		HIV positive		SIT total number of isolates[1] (% MDR[2])
SIT (of clade)	MDR	Non-MDR	MDR	Non-MDR	
50 (H3)	7	96	3	22	128 (7.8%)
53 (T1)	8	49	24	16	97 (33.0%)
33 (LAM3)	1	53	0	11	65 (1.54%)
42 (LAM9)	4	15	35	5	59 (66.1%)
1 (Beijing)	3	32	1	8	44 (9.1%)
47 (H1)	2	24	0	5	31 (6.5%)
22 (unknown)	2	16	2	4	24 (16.7%)
1355 (LAM)	1	12	4	6	23 (21.7%)
91 (X3)	1	14	1	6	22 (9.1%)
others	24	183	12	70	289 (12.5%)
total	53	494	82	153	782 (17.3%)

[1]12 of 794 isolates lacked either MDR or HIV status data, leaving 782 here as the denominator.
[2]Percentage of SIT-specific isolates with MDR.
doi:10.1371/journal.pone.0065873.t004

Conclusion

En résumé, nous avons présenté la distribution des souches de *M. tuberculosis* isolées à Lima, au Pérou. Cette étude a mis en évidence une proportion importante de nouveaux spoligotypes, suggérant une épidémie prolongée, et clonale, de la tuberculose MDR chez les patients VIH-positifs en milieu hospitalier.

Présentation de la seconde étude

Lors d'une autre étude menée au Pérou (**Article en préparation**), nous nous sommes concentrés sur l'analyse des souches ultra-résistantes (XDR) circulant entre 2007 et 2009. Un total de 142 souches a été typé à l'aide du Spoligotypage et du typage 15-loci MIRU. Cette étude a notamment mis en relief le fait que, contrairement aux autres pays d'Amérique latine où la lignée LAM est généralement la plus dominante, les lignées Haarlem (43.7% des isolats) et T (27.5% des isolats) étaient prédominantes. La lignée LAM représentait 16.2% des souches, et la lignée Beijing représentait un taux record de 9.2% des souches, soulignant l'histoire particulière du Pérou et ses relations avec les populations d'Asie de l'Est (**Iwamoto et coll., 2012**). Dans notre étude, 90.9% des souches provenant de patients préalablement traités. 62.7% des isolats étaient résistants à la fois à la capréomycine (CAP) et la kanamycine (KAN), 15.5% des souches étaient résistants uniquement à la CAP, et 21.8% étaient résistants uniquement à la KAN. De plus, les SIT219/T1 et SIT50/H3 étaient les plus majoritaires (n=21 isolats pour chaque motif) dans notre étude.

Résultats principaux

L'analyse basée sur les 15-loci MIRU-VNTRs (MIRU15) a montré, comme prévu, une plus grande diversité des profils MIRU par rapport au spoligotypes. En effet, seulement 12 génotypes MIRU15 correspondant à 49 isolats ont été identifiés. L'arbre phylogénétique (MST) basé sur la combinaison spoligotypes-15-loci MIRUs (**Figure 57**) témoigne de la diversité des souches de cette étude qui représente une première analyse des souches XDR circulants au Pérou. Nous remarquons que les lignées Haarlem et T sont très diversifiées. Les couples SITs/15-MITs suivants étaient prédominants dans l'étude : SIT 3001 / 15-MIT 242

(n=6 isolats), SIT 219 / 15-MIT 236 (n=5 isolats), SIT 47 / 15-MIT 234 (n=4 isolats), SIT 50 / 15-MIT 240 (n=3 isolats), et SIT 1 / 15-MIT 235 (n=3 isolats). Notons que, mis à part les Etats-Unis d'Amérique, les SITs 3001/H3 et 219/T1 n'ont été isolés qu'au Pérou.

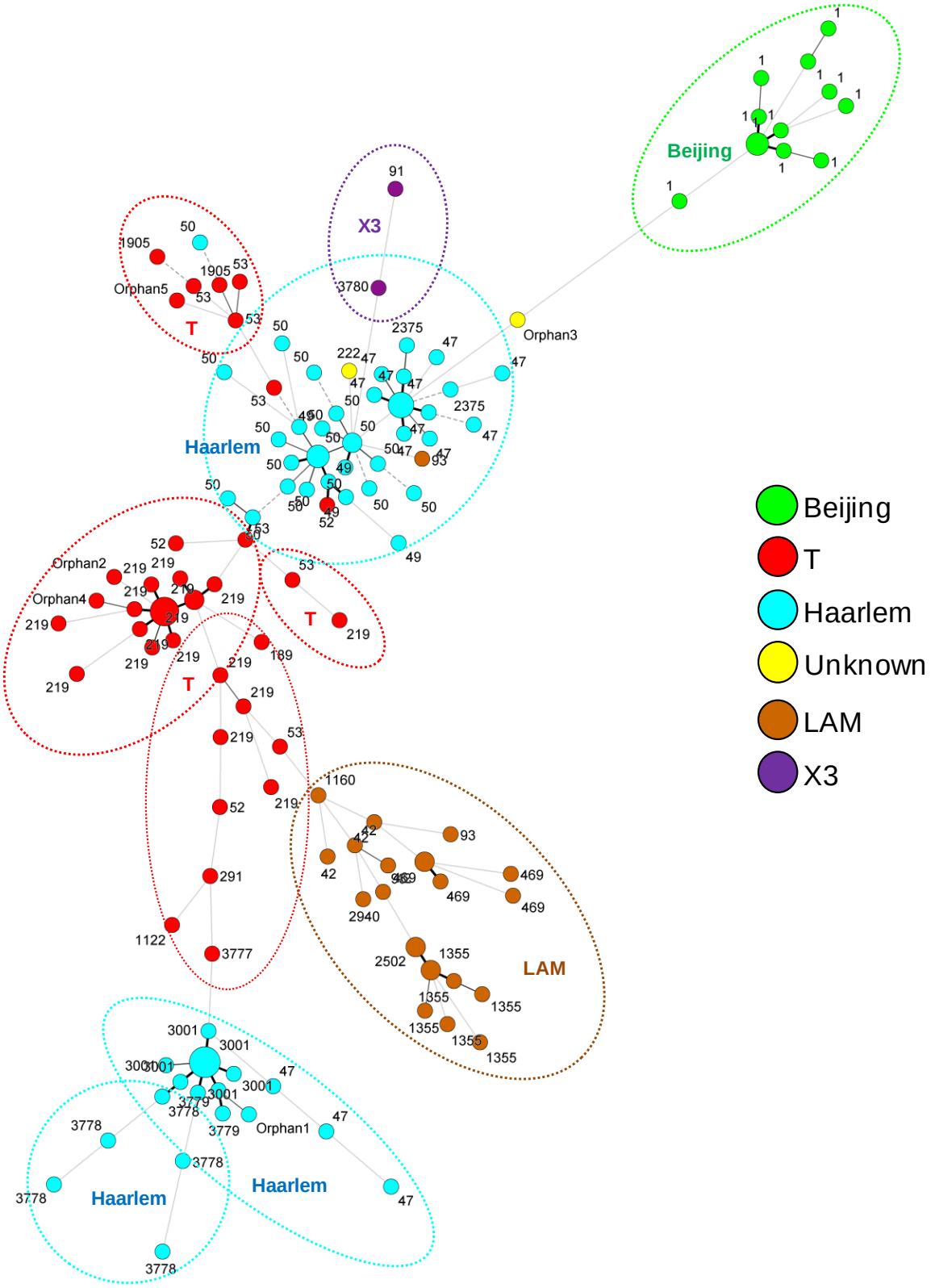


Figure 57 : MST composite basé sur le typage MIRU15 et le spoligotypage des 142 souches de l'étude. Les numéros indiqués représentent les SITs identifiés, les mêmes numéros SITs (répartis dans différents nœuds) apparaissent avec des profils MIRU15 différents.

En outre, une analyse comparative basée sur les 15-loci MIRUs nous a permis de souligner la particularité des souches péruviennes de la famille Beijing par rapports aux souches Beijing provenant du Japon (n=603), de la Chine (n=209), ou encore de la France (n=35).

Un MST fut construit à partir de ces données (**Figure 58A**), et une représentation graphique utilisant WebLogo (http://weblogo.berkeley.edu) permis de visualiser la diversité des souches de la lignée Beijing en comparant avec les MIRU15 des autres pays étudiés (**Figure 58B**). Le **Tableau 18** basé sur les données recueillies pour l'analyse WebLogo, permet de connaître la moyenne des nombres de copies pour chaque locus MIRU15 étudié. A travers le MST (**Figure 58A**), on peut noter que la majorité des souches péruviennes sont concentrés ensemble, non loin du profil 15-MIT 11. Aussi, les profils per006 et 15-MIT 234 sont phylogénétiquement proches des souches japonaises de la lignée Beijing, tandis que le profil per009 en haut du MST est proche des motifs MIRU15 chinois.

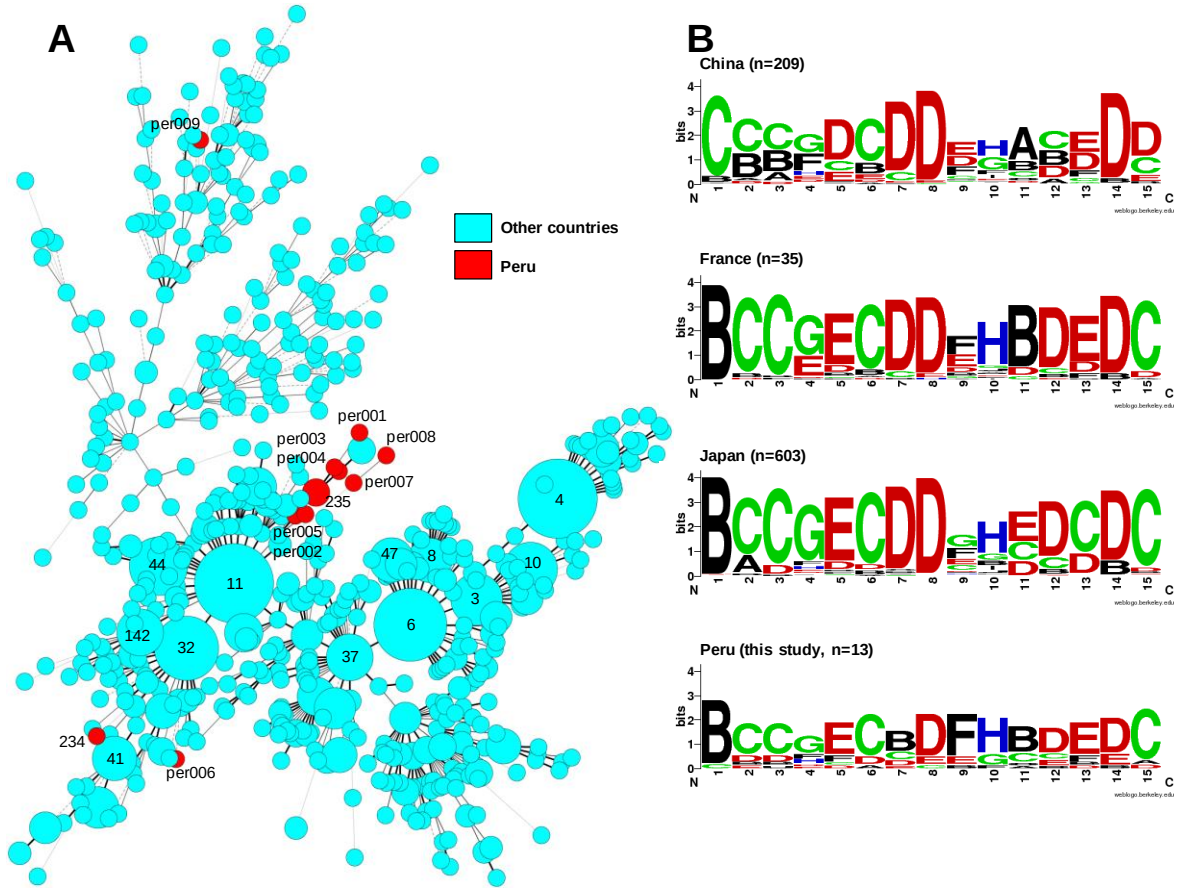


Figure 58 : Analyse basée sur le typage MIRU15 des souches de la famille Beijing : (**A**) MST représentant la diversité génétique des souches péruviennes (n=13), représentées en rouge par rapport aux souches provenant des autres pays enregistrés dans SITVIT2 (n=863) représentées en cyan; les nombres à l'intérieur des nœuds indiquent les 15-MITs ; (**B**) WebLogo comparant les différents loci des motifs MIRU15 de Chine (n=209), de France (n=35), et du Japon (n=603), avec les isolats de notre étude (n=13). L'ordre des loci MIRU15 (axe des abscisses sur les logos) est le celui utilisé classiquement dans cette thèse (cf. Matériels et Méthodes). L'axe des ordonnées représente la variabilité des nombres de copies pour un loci MIRU15 donné. Les lettres symbolisent les nombres de copies, avec le code « label/nombre de copies » : A/1, B/2, C/3, D/4, E/5, F/6, G/7, H/8, I/9, J/10, K/11, L/12, M/13, N/14, O/15, P/16, Q/0, U/Unknown.

Tableau 18 : Moyenne des nombres de copies de chaque locus MIRU15 pour les pays étudiés, et en fonction de notre étude.

Locus MIRU15	Chine	France	Japon	Pérou
1_MIRU-4	2.9	2	2	2.1
2_MIRU-10	2.6	3	2.5	3.2
3_MIRU-16	2.4	3.1	3.1	3.1
4_MIRU-26	6.3	6.3	6.9	6.6
5_MIRU-31	4	4.7	4.8	4.9
6_MIRU-40	3	2.9	3	3
7_ETR-A	3.9	3.9	3.8	2.8
8_ETR-C	4	4.2	4	4.1
9_QUB-11b	4.9	5.1	5.6	5.5
10_QUB-26	7.1	7.6	7.4	7.6
11_QUB-4156	1.6	2.3	4.2	2
12_Mtub04	2.8	3.8	3.7	3.8
13_Mtub21	4.7	4.8	3.3	4.8
14_Mtub30	3.9	3.8	3.6	4.1
15_Mtub39	3.7	3	3.1	2.9

Conclusion

Pour conclure, ce travail présente pour la première fois au Pérou, la caractérisation génétique et les relations évolutives des souches XDR du complexe *Mycobacterium tuberculosis*. Cette étude a permis de mettre en évidence la proportion importante des lignées Haarlem et T. En outre, il n'y avait pas de preuve d'une lignée prédominante associé à la condition d'ultra-résistance (XDR). Une surveillance et un contrôle régulier des souches de MTBC circulant au Pérou sont nécessaires pour mieux prévenir et limiter l'expansion de la maladie.

*2.2.4 Etudes épidémiologiques au Brésil (**Article 9 + travail en cours**)*

« Le Brésil est le pays le plus vaste et le plus peuplé d'Amérique latine avec une superficie de 8 511 965 km². Colonie prospère de la couronne portugaise durant plusieurs siècles, le pays est aujourd'hui le seul héritage du Portugal en terre américaine. Le Brésil possède des caractéristiques communes avec la plupart des autres pays d'Amérique du Sud : c'est un pays en majorité catholique, la densité humaine moyenne est assez faible et la population est fortement urbanisée. Ce pays présente de forts contrastes géographiques et sociologiques. Alors qu'une grande partie du territoire est couverte par la forêt amazonienne pratiquement vide d'êtres humains, la côte sud-est abrite les mégapoles de São Paulo et Rio de Janeiro. Les inégalités économiques sont parmi les plus élevées du monde. Cependant, le Brésil fait partie des grandes puissances émergentes aux côtés de la Chine, de l'Inde, de l'Afrique du Sud et de la Russie. » (http://fr.wikipedia.org/wiki/Br%C3%A9sil).

Présentation de la première étude

Une première étude (**Article 9**) a été menée dans la région de São Paulo qui est l'état le plus peuplé du Brésil. Cet état est également celui qui compte le plus grand nombre de cas de tuberculose chaque année (plus de 18500 cas par an), avec une incidence de 40.5 cas pour 100 000 habitants en 2010. Notre étude constitue le premier aperçu des génotypes de bacilles tuberculeux circulant à São Paulo. Nous avons étudié un total de 740 isolats de MTBC, résistants et susceptibles aux antituberculeux, recueillis entre 2006 et 2008.

Résultats principaux

La lignée LAM était la plus prédominante dans l'étude, représentant 331/740 (44.7%) des isolats, suivi par la lignée Haarlem (n=153/740, 20.7%), et la lignée T (n=148/740, 20%). Les SITs prédominants dans notre étude étaient les suivants : SIT50/H3 (n=84), SIT53/T1 (n=62), SIT42/LAM9 (n=48), SIT17/LAM2 (n=46), SIT33/LAM3 (n=36), SIT64/LAM6 (n=31), SIT47/H1 (n=24), SIT60/LAM4 (n=23), SIT51/T1 (n=21), SIT20/LAM1 (n=20), et SIT317/X2 (n=13). Le **Tableau 19** montre qu'il existe une corrélation entre l'état de résistance aux médicaments des spoligotypes regroupés en clusters et le sexe et l'âge des patients. L'âge moyen des patients variait de 31 ans (SIT 4 / lignée inconnue) à 46 ans (SIT 137 / X2). Le sex-ratio homme / femme observé variait de 1.2 (SIT 1905 / T1) à 12.0 (SIT 137 / X2).

Tableau 19 : Distribution des SITs majeurs en fonction de la résistance aux antibiotiques, du sexe, et de l'âge des patients. Abréviations: I, isoniazide; R, rifampicine; SM, streptomycine.

SIT (lineage) – No. of isolates	Drug resistant patterns, No. (%)						Gender, No. (%)		Age (years)	
	MDR	Monoresistance			Other-resistance	Susceptible	Male	Female	Mean	Median
		I	R	S						
1 (Beijing) – 7	3 (42.8)	1 (14.3)	0	1 (14.3)	0	2 (28.6)	5 (71.4)	2 (28.6)	41.0 (25–58)	37.0
177 (LAM9) – 9	4 (44.4)	1 (11.1)	0	1 (11.1)	0	3 (33.3)	7 (77.8)	2 (22.2)	38.0 (26–50)	38.0
1905 (T1) – 11	3 (27.3)	7 (63.6)	0	0	0	1 (9.1)	6 (54.5)	5 (45.5)	38.0 (17–54)	51.0
4 (U) – 12	7 (58.3)	4 (33.3)	0	0	0	1 (8.3)	8 (66.7)	4 (33.3)	31.0 (20–51)	29.0
137 (X2) – 13	6 (46.1)	4 (30.8)	0	2 (15.4)	0	1 (7.7)	12 (92.3)	1 (7.7)	46.0 (20–71)	48.0
51 (T1) – 21	1 (4.8)	3 (14.3)	0	12 (57.1)	1 (4.8)	4 (19.0)	16 (76.2)	5 (23.8)	40.0 (25–66)	41.0
60 (LAM4) – 23	15 (65.2)	5 (21.7)	0	0	0	3 (13.0)	13 (56.5)	10 (43.5)	33.0 (10–62)	31.0
47 (H1) – 24[a]	4 (16.7)	6 (25.0)	2 (8.,3)	2 (8.3)	3 (12.5)	6 (25.0)	18 (75.0)	6 (25.0)	34.3 (20–54)	33.0
20 (LAM1) – 26	6 (23.1)	4 (15.4)	2 (7.7)	3 (11.5)	3 (11.5)	8 (30.8)	18 (69.2)	8 (30.8)	39.8 (21–69)	40.0
64 (LAM6) – 31	9 (29.0)	11 (35.5)	0	1 (3.2)	1 (3.2)	9 (29.0)	20 (64.5)	11 (34.5)	38.3 (16–74)	37.0
33 (LAM3) – 36	11(30.5)	4 (11.1)	3 (8.3)	7 (19.4)	2 (5.5)	9 (25.0)	20 (55.6)	16 (44.4)	39.7 (17–70)	40.0
17 (LAM2) – 46	18 (39.1)	14 (30.4)	0	3 (6.5)	1 (2.2)	10 (21.7)	36 (78.3)	10 (21.7)	36.5 (15–76)	33.0
42 (LAM9) – 48	13 (27.1)	7 (14.6)	2 (4.2)	5 (10.4)	6 (12.5)	15 (31.2)	32 (66.7)	16 (33.3)	36.5 (12–63)	35.5
53 (T1) – 62[a]	10 (16.1)	10 (16.1)	2 (3.2)	19 (30.6)	3 (4.8)	17 (27.4)	44 (71.0)	18 (29.0)	35.0 (15–66)	34.0
50 (H3) – 84[b]	31 (36.9)	16 (19.0)	2 (2.4)	14 (16.7)	5 (5.9)	15 (17.8)	65 (77.4)	19 (29.6)	36.3 (16–64)	34.0

On remarque aussi dans ce dernier tableau, que la proportion d'isolats MDR dépassait les 40% pour les SITs suivant : SIT1/Beijing, SIT4/Unknown, SIT60/LAM4, SIT177/LAM9, et SIT137/X2.

Conclusion

Enfin, dans cette première étude concernant les isolats de MTBC circulant à São Paulo, les lignées relativement anciennes du groupe PGG1 (Beijing, CAS, EAI, PINI) étaient rarement présents, alors que les lignées récentes du groupe PGG2 (LAM, T, S, X, Haarlem) prédominaient dans cette étude. De plus, ce travail a permis de démontrer que les profils SITs les plus répandus dans l'état de São Paulo incluent aussi bien des isolats résistants que des isolats sensibles aux antibiotiques.

Présentation de la seconde étude

Dans une seconde étude brésilienne (**Article en préparation**), nous avons enquêté sur les isolats de MTBC présents dans l'état de « Mato Grosso do Sul » (MS) dont les villes principales sont Dourados (186 357 habitants) et Corumbá (101 089 habitants). Nous avons analysé 149 isolats à partir de patients tuberculeux indigènes (85/149, 57%) et non-indigènes (64/149, 43%), provenant de 18 villes différentes de MS, entre 2002 et 2007. Les isolats ont été génotypés en utilisant les méthodes suivantes : le spoligotypage, le typage 12-loci MIRU (MIRU12), et l'analyse RD^{Rio} Large Sequence Polymorphism (LSP). La lignée LAM était prédominante dans cette étude (n=98/149, 65.8% des isolats). Parmi ces LAM, la sous-lignée LAM9 était la plus fréquente (77/98, 78.6% des isolats de la famille LAM).

Résultats principaux

Les patients de sexe masculin (n=108/149) prédominaient dans cette étude, représentant 72.5% des souches. Les patients étaient âgés de 2 à 93 ans, cependant, une majorité d'entre eux se trouvaient dans la tranche d'âges 20-60 ans (n=76/149, 51%). La différence entre les patients indigènes et non-indigènes était statistiquement significative en ce qui concerne les groupes d'âge 25-34 et 55-64 ans (valeur P = 0.0083, Odds Ratio = 5.341; 95% intervalle de confiance (1.32; 26.74)) (**Figure 59**).

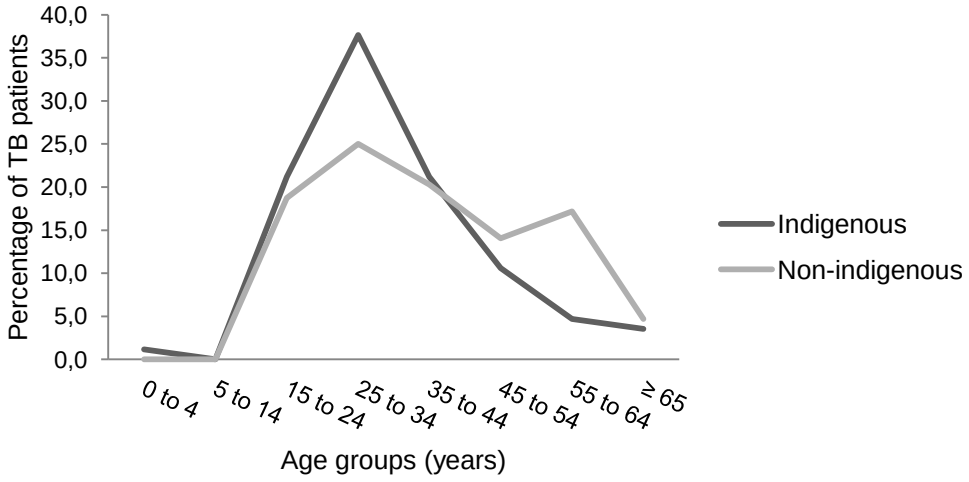


Figure 59 : Distribution des groupes d'âge entre les patients indigènes et non-indigènes.

Les profils spoligotypiques majeurs sont visibles dans le Spoligoforest de la **Figure 60**, construit en tant que « hierarchical layout ». Nous pouvons dénoter la dominance des profils SIT42/LAM9, SIT177/LAM9, SIT53/T1, et SIT20/LAM1.

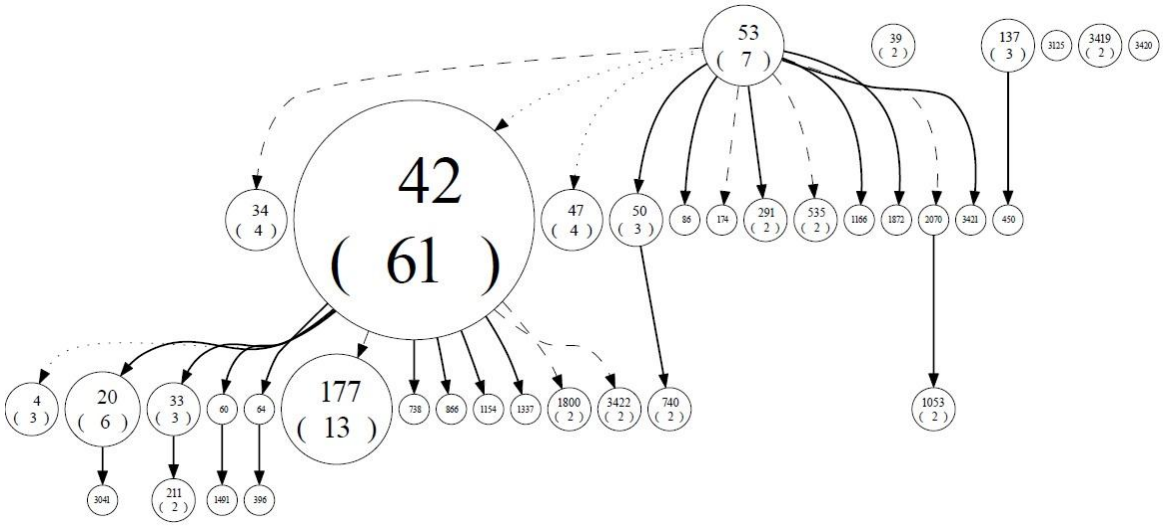


Figure 60 : Spoligoforest basé sur les 149 spoligotypes de l'étude, et construit en tant que « hierarchical layout ». La signification des différents liens et la taille des nœuds est similaire au précédent spoligoforest.

Notre étude a également révélé des différences notables entre indigènes et non-indigènes en ce qui concerne la distribution des lignées. En effet, la proportion des lignées LAM et S était significativement plus élevée chez la population indigène, alors que la proportion de la lignée Haarlem était plus élevée dans la population non-indigène (valeur P = 0.048). La différence de proportion des génotypes SIT20/LAM1 et SIT47/H1 est notamment statistiquement significative entre les deux populations mentionnées (valeur P = 0.033) (voir **Tableau 20** et **Figure 61**).

Tableau 20 : Distribution des lignées phylogénétiques majeures, et différence des sex-ratios entre les populations indigènes et non-indigènes.

Parameters	Indigenous patients (n=85)	Non indigenous patients (n=64)	Global dataset (n=149)	p-value
Total no. of isolates	85 (57%)	64 (43%)	149	
Sex				
Male	63 (74.1%)	45 (70.3%)	108 (72.5%)	0.61
Female	22 (25.9%)	19 (29.7%)	41 (27.5%)	
Sex ratio M/F	2.9	2.4	2.6	
Clades				
Haarlem	3 (3.5%)	7 (10.9%)	10 (6.7%)	
LAM	60 (70.6%)	38 (59.4%)	98 (65.8%)	
S	4 (4.7%)	0 (0.0%)	4 (2.7%)	**0.048**
T	13 (15.3%)	10 (15.6%)	23 (15.4%)	
Unknown/Others	3 (3.5%)	8 (12.5%)	11 (7.4%)	
X	2 (2.4%)	1 (1.6%)	3 (2.0%)	

Le MST suivant (**Figure 61**) indique quels profils spoligotypes et/ou MIRU12, sont impliqués chez les patients indigènes ou non-indigènes, et de sexe masculin ou féminin. Ce MST montre notamment les 5 couples SITs/12-MITs (indiqués en bleu dans la **Figure 61C**) les plus prédominants (SIT42/MIT25, SIT177/MIT25, SIT42/MIT1532, SIT42/MIT1530 and SIT42/MIT1535). On remarque aussi une grande dispersion dans la répartition des génotypes entre les différents groupes ethniques et le sexe des patients. Nous avons pu mettre en évidence 4 profils spoligotypes retrouvés uniquement chez les patients non-indigènes, il s'agit des SITs : le SIT211/LAM3, le SIT535/T1, le SIT740/H3, et le SIT3419/H3. Pareillement, 5 autres génotypes étaient retrouvés uniquement chez les patients indigènes : le SIT20/LAM1, le SIT34/S, le SIT291/T1, le SIT1053/T1, et le SIT1800/LAM9.

Conclusion

En résumé, bien que le nombre de souches étudiées soit peu élevé, cette étude reflète bien la diversité génotypique des isolats de *Mycobacterium tuberculosis* entre des populations distinctes dans l'état de Mato Grosso do Sul, au Brésil. Cependant, une meilleure connaissance des génotypes de M. tuberculosis circulants dans cette région, permettrait de renforcer les programmes nationaux de contrôle de la tuberculose.

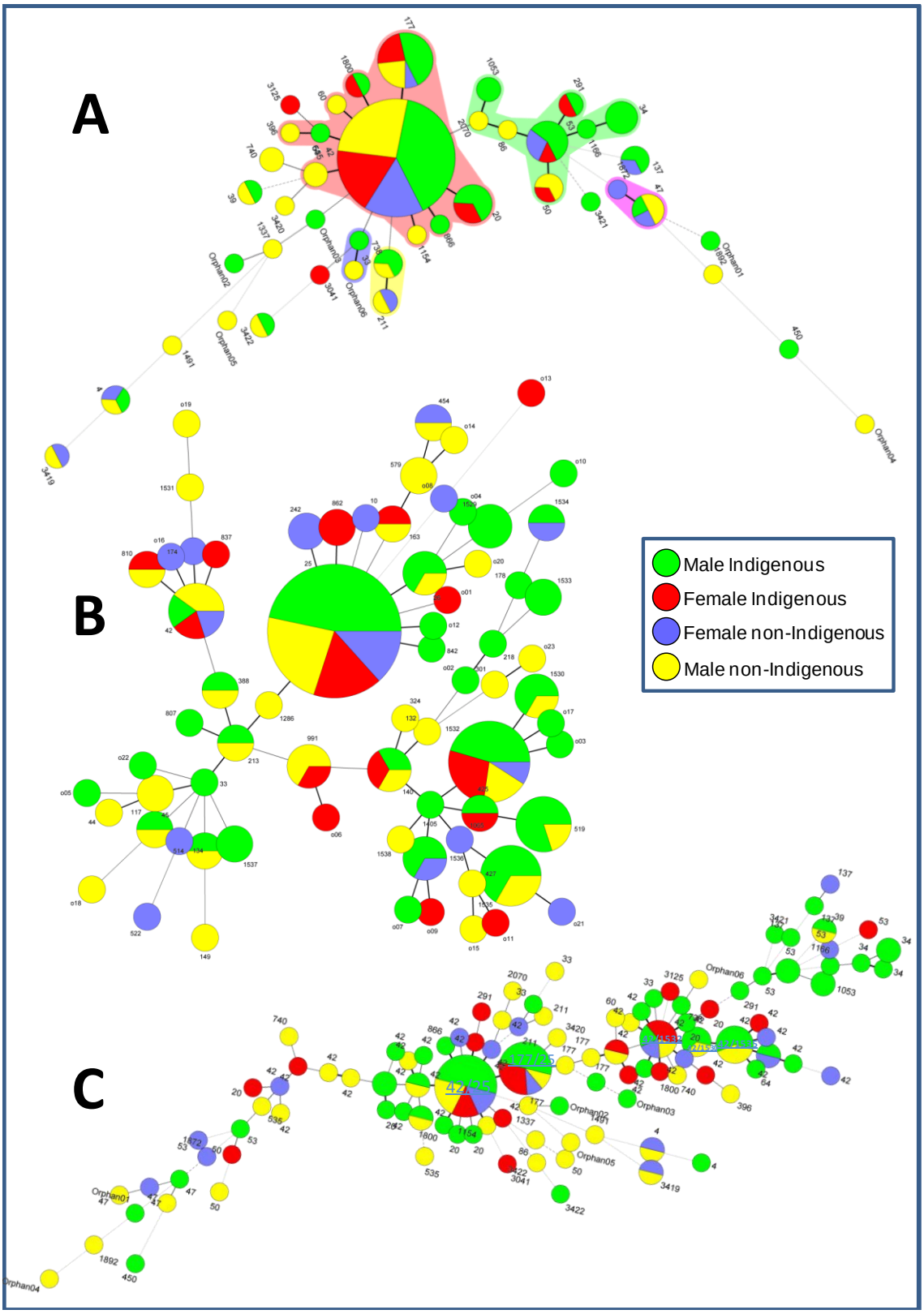


Figure 61 : MSTs montrant la répartition des différentes ethnies (indigènes et non-indigènes) au sein de chaque génotype spoligotype et/ou 12-loci MIRU en fonction du sexe (masculin ou féminin) des patients : (**A**) MST basé sur le spoligotypage ; (**B**) MST basé sur le typage 12-loci MIRU ; (**C**) MST combiné basé à la fois sur le spoligotypage et le typage MIRU12.

2.3 Epidémiologie de la tuberculose en Afrique de l'Ouest et de l'Est

« De longues lianes ne suffisent pas pour construire une case, il en faut aussi des courtes pour parfaire les angles. » Proverbe Africain.

« L'Afrique subsaharienne reste la région du monde où l'on enregistre le plus fort taux de mortalité infantile (121 pour mille contre 7 pour mille dans les pays développés). Mais un rapport conjoint de l'Unicef, l'OMS et la Banque mondiale souligne que des progrès notables ont été accomplis. Ce taux a en effet diminué entre 1990 et 2010 dans la plupart des États concernés, à l'exception de la Somalie, du Zimbabwe et du Cameroun. Quatre pays - Madagascar, le Liberia, l'Érythrée et le Malawi - sont même en passe d'atteindre les objectifs du millénaire pour le développement (OMD). Ce plan, adopté en l'an 2000 par les Nations unies, a pour but d'éradiquer la pauvreté dans le monde d'ici à 2015. Parmi les moyens d'y parvenir : réduire la mortalité infantile de deux tiers. » (**GEO N°403 de septembre 2012 p.18**).

Dans notre base de données SITVIT2, les données génotypiques d'isolats de *M. tuberculosis* sont nombreuses (**Tableau 21**). Le continent Africain montre une diversité génétique particulière dans la répartition des lignées phylogénétiques. Cette diversité génotypique observable témoigne de la complexité historique de ce continent.

A travers ce voyage épidémiologique, nous avons d'abord fait escale en Guinée-Bissau (**Article 10**), puis nous nous sommes affairés à redécouvrir l'Ethiopie, l'un des berceaux de l'humanité (**Article 11**).

Tableau 21 : Inventaire des souches collectées en Afrique et répertoriées dans SITVIT2.

Country of Isolation	nb of strains	%	total in region
BDI (Burundi)	10	0,07	
COM (Comoros)	12	0,09	
ETH (Ethiopia)	1507	11,17	
KEN (Kenya)	80	0,59	
MDG (Madagascar)	640	4,75	
MOZ (Mozambique)	511	3,79	**AFRI-E** n=4138 (30,68%)
MUS (Mauritius)	23	0,17	
MWI (Malawi)	150	1,11	
REU (Reunion)	17	0,13	
TZA (Tanzania)	410	3,04	
UGA (Uganda)	42	0,31	
ZMB (Zambia)	462	3,43	
ZWE (Zimbabwe)	274	2,03	
CAF (Central African Republic)	151	1,12	
CMR (Cameroon)	566	4,20	**AFRI-M** n=808 (5,99%)
COD (Democratic Republic of the Congo)	2	0,01	
TCD (Chad)	89	0,66	
DZA (Algeria)	151	1,12	
EGY (Egypt)	230	1,71	
LBY (Libya)	58	0,43	**AFRI-N** n=2145 (15,90%)
MAR (Morocco)	899	6,67	
SDN (Sudan)	284	2,11	
TUN (Tunisia)	523	3,88	
BWA (Botswana)	5	0,04	
NAM (Namibia)	81	0,60	**AFRI-S** n=4580 (33,96%)
SWZ (Swaziland)	1	0,01	
ZAF (South Africa)	4493	33,31	
BEN (Benin)	4	0,03	
BFA (Burkina Faso)	1	0,01	
CIV (Ivory Coast)	50	0,37	
GIN (Guinea)	4	0,03	
GMB (Gambia)	376	2,79	**AFRI-W** n=1816 (13,46%)
GNB (Guinea-Bissau)	423	3,14	
MLI (Mali)	20	0,15	
MRT (Mauritania)	2	0,01	
NGA (Nigeria)	837	6,21	
SEN (Senegal)	92	0,68	
SLE (Sierra Leone)	7	0,05	

*2.3.1 Etude épidémiologique en Guinée-Bissau (**Article 10**)*

« La Guinée-Bissau est le 3ème producteur de noix de cajou d'Afrique, et le 6ème mondial, avec sa production de 120 000 tonnes par an lui rapportant 60 millions de dollars. Bien que le pays soit entouré d'états francophones (Sénégal, Guinée), la langue officielle de la Guinée-Bissau est le portugais. Du fait de sa pauvreté et de sa désorganisation économique, la Guinée-Bissau est une proie facile pour les trafiquants de drogue de l'Amérique du Sud qui l'utilisent comme passerelle pour atteindre l'Union européenne. » (http://fr.wikipedia.org/wiki/Guin%C3%A9e-Bissau).

Présentation de l'étude

L'étude épidémiologique menée en Guinée-Bissau (**Article 10**) a consisté à analyser 414 souches de MTBC entre 1989 et 2008 par la méthode de spoligotypage. Les souches de Guinée-Bissau ont ensuite été comparées aux souches contenues dans la base de données SITVIT2. Notre étude a permis d'identifier les lignées majoritaires circulant en Guinée-Bissau. Ainsi, nous avons pu mettre en évidence les lignées suivantes : AFRI (n = 195, 47,10%), LAM (n = 75, 18.12%), T (n = 53, 12.8%), Haarlem (n = 37, 8.85%), EAI (n = 25, 6.04%), « Unknown » (n = 12, 2.87%), Beijing (n = 7, 1.68%) , clade X (n = 4, 0.96%), Manu (n = 4, 0.97%), CAS (n = 2, 0.48%). Deux souches de la lignée LAM isolées en 2007 appartenaient à la famille Cameroun (SIT61).

Résultats principaux

La **Figure 62** indique la distribution globale des lignées phylogénétiques du continent Africain, ainsi que par sous-régions. Cette représentation réalisée à partir des données inclues dans SITVIT2 en 2011 nous donne un petit aperçu de l'évolution survenue dans les différentes régions d'Afrique. Nous remarquerons que la distribution des lignées phylogénétiques varie d'une sous-région à l'autre, montrant la spécificité de chaque région.

Notre étude a souligné la prédominance particulière de la lignée AFRI en Guinée-Bissau, pays qui recueillerait le plus fort taux d'isolats appartenant à la lignée AFRI, lorsque l'on compare ces données avec SITVIT2. La présente analyse a aussi permis de mettre en évidence le caractère phylogénétique singulier de la famille de souches « Guinea-Bissau ». Cette famille de souches a été classifiée en tant que membre de la sous-lignée spoligotypique AFRI_1 (selon les règles de classification basées sur le spoligotypage). Les isolats de cette famille contiennent un motif de RFLP distinct avec un faible nombre de séquences d'insertion IS6110, et ne présentent pas les régions de différence RD7, RD8, RD9 et RD10, RD701 et RD702.

Ce travail a également permis de révéler la présence de profiles de spoligotypes endémiques à la Guinée-Bissau. Ainsi, les 2 profils spoligotypes SIT 181 et SIT 187 étaient majoritaires en Guinée-Bissau, et représentaient respectivement 36.92% et 52.87% des isolats cliniques de la base de données SITVIT2 (**Tableau 22**).

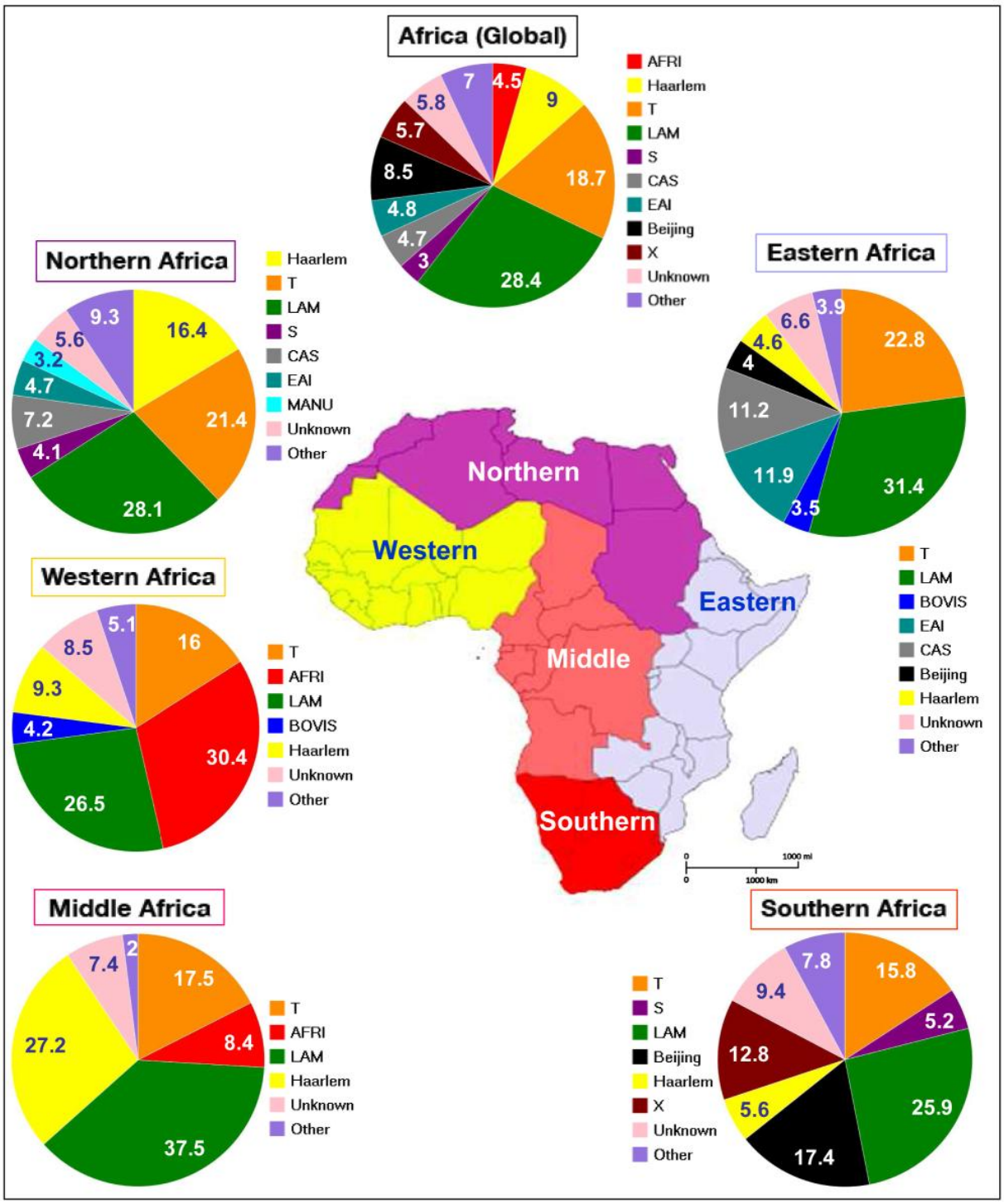


Figure 62 : Distribution géographique des lignées génotypiques basées sur le spoligotypage de *Mycobacterium tuberculosis* dans différentes sous-régions de l'Afrique (n = 11 956 isolats cliniques) dans la base de données SITVIT2.

Tableau 22 : Descriptions des SITs prédominants (représentant au moins 10 isolats) dans l'étude.

SIT (Octal Number) Clade Spoligotype Description	Total (%) in GNB	% in GNB as compared to SITVIT2	Distribution in Regions with ≥ 5% of a given SIT	Distribution in Countries with ≥ 5% of a given SIT
181 (770777777777671) AFRI_1	96(23.19)	36.92	AFRI-W 74.23, AMER-N 8.08, EURO-W 7.69	GNB 37.30, GMB 32.31, USA 0.08
187 (770777777767671) AFRI_1	46(11.11)	52.87	AFRI-W 74.71, AMER-N 12.64, AFRI-N 5.75	GNB 54.02, GMB 18.39, USA 12.64
42 (777777607760771) LAM 9	36(8.7)	1.31	AMER-S 28.38, AMER-N 15.49, EURO-S 12.16, AFRI-N 10.48, EURO-W 6.68	USA 14.50, BRA 9.82, MAR 8.62, COL 7.49, ITA 6.39
47 (777777774020771) H1	18(4.35)	1.47	AMER-N 20.95, EURO-W 20.13, EURO-S 13.77, AMER-S 10.43, EURO-E 8.39	USA 19.40, AUT 10.27, ITA 7.25, BRA 5.87
50 (777777777720771) H3	12(2.9)	0.43	AMER-N 21.85, AMER-S 15.90, EURO-W 15.36, EURO-S 12.63, EURO-E 6.46, AFRI-S 5.17	USA 21.31, AUT 7.39, ESP 6.60, ITA 5.27
20 (677777607760771) LAM 1	11(2.66)	1.57	AMER-S 24.39, AMER-N 24.39, AFRI-S 12.69, EURO-S 11.26, EURO-W 8.13, CARI 6.13	USA 22.68, BRA 14.12, NAM 8.84, PRT 6.99, VEN 5.99
53 (777777777760771) T1	10(2.42))	0.21	AMER-N 19.40, AMER-S 14.26, EURO-W 12.64, EURO-S 9.88, ASIA-W 8.57, AFRI-S 6.41	USA 17.04, ZAF 6.27, ITA 5.06
527 (677777777403771) EAI5	9(2.17)	100.00	AFRI-W 100.00	GNB 100.00

Conclusion

En conclusion, la famille de souches « Guinée-Bissau », fait partie de la lignée AFRI (*M. africanum*) prédominante en Afrique de l'Ouest. Cette famille présente une spécificité phylogéographique extrêmement élevé pour l'Afrique de l'Ouest, avec pour épicentre de sa distribution, la Guinée-Bissau. Cependant, les tendances temporelles indiquent cependant que cette famille de souches est en déclin dans la plupart des régions d'Afrique, y compris la Guinée-Bissau (valeur P = 0.048).

*2.3.2 Etude épidémiologique en Ethiopie (**Article 11**)*

« L'Éthiopie, aujourd'hui constitutionnellement laïque, est un pays où de nombreuses croyances coexistent. Après l'Arménie, c'est la deuxième plus ancienne nation chrétienne au monde. S'y trouvent aujourd'hui des orthodoxes monophysites, des catholiques et des protestants. Par ailleurs, un tiers de ses habitants sont musulmans et des minorités religieuses comme les Falachas ou des animistes y vivent aussi. Sur le plan international, l'Éthiopie est signataire de la Déclaration des Nations unies dès 1942 et devient l'un des 51 États membres fondateurs de l'ONU. Addis-Abeba est aujourd'hui le siège de la Commission économique pour l'Afrique (CEA) et de l'Union africaine. » (http://fr.wikipedia.org/wiki/%C3%89thiopie).

Présentation de l'étude

Cette étude (**Article 11**) visait à caractériser les isolats cliniques de MTBC circulant dans une zone à prédominance pastorale (dédiée à l'élevage) en Ethiopie, un pays où la tuberculose est la deuxième cause de mortalité. La culture des échantillons de crachats,

prélevés sur un total de 325 cas suspects de tuberculose pulmonaire, a été faite pour isoler les souches de MTBC. Le spoligotypage a été utilisé pour caractériser 105 isolats de culture positive, et le résultat a été comparé à la base de données internationale SITVIT2. Quarante-quatre motifs spoligotypes distincts ont été observés, correspondant à 35 « Spoligotype International Types » (SITs) contenant 96 isolats cliniques, et 9 motifs orphelins. 27 SITs contenant 83 isolats correspondent à un SIT préexistant dans la base de données, alors que 8 SITs (n=13 isolats) ont été nouvellement créés. Un total de 19 SITs contenant 80 isolats ont été regroupés en clusters dans cette étude (le clustering global étant de 76,19%).

Résultats principaux

Le **Tableau 23** est un récapitulatif des motifs spoligotypes prédominants dans notre étude, avec leurs distributions par pays. Parmi ces motifs prédominants, les SITs suivants étaient les plus importants : SIT 149 de la lignée T3-ETH, SIT 37 de la lignée T3, SIT 53 de la famille T1, et SIT 54 de la lignée Manu2. Ces derniers SITs représentaient à eux seuls, 36.2% des isolats cliniques de notre étude.

Tableau 23 : Représentation des SITs prédominants (contenant au moins 3 souches) dans notre étude.

SIT (Lineage) Octal Number Spoligotype Description	Nb in study (%)	% in study vs. database	Distribution in countries with >=3% of a given SITs
37 (T3) 777737777760771	10 (9.52)	2.04	ETH 17.72, USA 8.15, SWE 5.5, SAU 5.3, CHN 5.3, FXX 4.48, ITA 4.28, ZAF 3.46, DNK 3.46, FIN 3.06
149 (T3-ETH) 777000377760771	10 (9.52)	2.07	ETH 58.06, DNK 11.78, SAU 6.61, USA 4.75, SWE 4.55, NLD 3.51
53 (T1) 777777777760771	9 (8.57)	0.14	USA 13.25, FXX 7.91, ITA 5.36, BRA 5.15, ZAF 4.88, TUR 3.49, AUT 3.44, CHN 3.1
54 (Manu2) 777777777763771	9 (8.57)	3.56	CHN 14.63, IND 13.83, EGY 13.44, USA 7.91, SAU 7.91, ZAF 6.32, IRQ 5.53, RUS 3.95, ETH 3.56
52 (T2) 777777777760731	6 (5.71)	0.65	CHN 11.82, USA 11.61, SWE 9.76, FXX 9.0, BEL 4.88, ITA 3.47, JPN 3.25, ETH 3.25
25 (CAS1-Delhi) 703777740003171	4 (3.81)	0.64	ETH 23.83, SAU 15.14, SDN 11.11, USA 8.37, IND 7.89, OMN 6.76, IRQ 4.67
26 (CAS1-Delhi) 703777740003771	4 (3.81)	0.26	IND 30.19, USA 14.47, PAK 9.51, OMN 7.2, BGD 5.48, SAU 5.15, IRN 4.1, ITA 3.17, ETH 3.1
119 (X1) 777776777760771	3 (2.86)	0.28	USA 58.23, ZAF 13.22, MEX 9.61
121 (H3) 777777775720771	3 (2.86)	5.77	TWN 36.54, ETH 19.23, FXX 11.54, CZE 7.69, USA 5.77, SAU 5.77
523 (Manu_ancestor) 777777777777771	3 (2.86)	6.25	USA 10.42, SAU 8.33, JPN 8.33, CHN 8.33, NGA 6.25, MYS 6.25, KOR 6.25, ETH 6.25, ZAF 4.17, VNM 4.17, TWN 4.17, THA 4.17, BEL 4.17
3897 (Manu2) 777347777763671	3 (2.86)	75	ETH 75.0, USA 25.0

Les trois lignées dominantes T, CAS, et Manu représentaient 76,19% des isolats cliniques. Le SIT149/T3-ETH était l'un des deux motifs les plus représentés avec le SIT37/T3. Le spoligoforest (**Figure 63**) construit grâce à l'algorithme de Fruchterman-Reingold disponible dans l'outil « SpolTools », met bien en évidence les motifs spoligotypes présents dans notre étude.

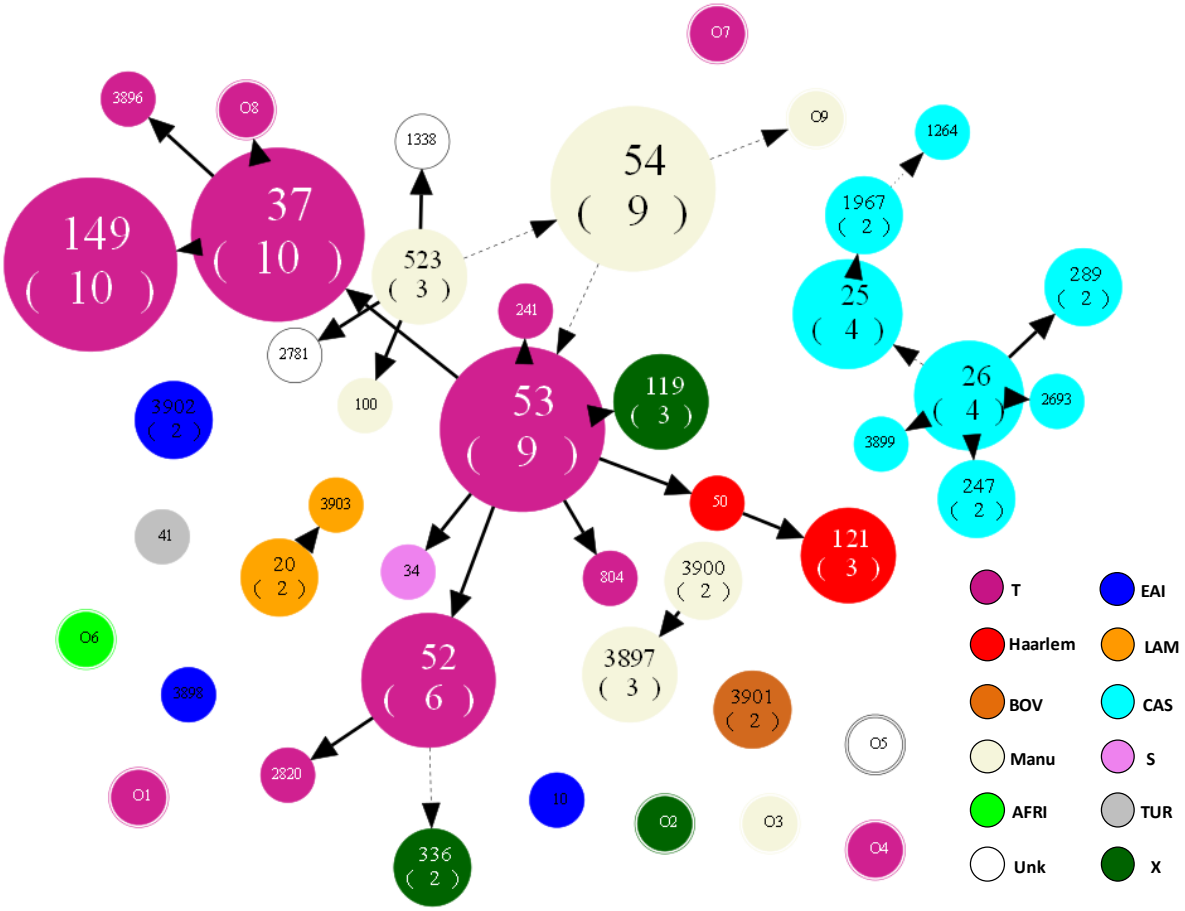


Figure 63 : Arbre Spoligoforest dessiné à l'aide de l'algorithme de Fruchterman-Reingold de l'outil SpolTools (http://www.emi.unsw.edu.au/spolTools). Les couleurs indiquent les différentes lignées phylogénétiques, et la taille des nœuds est proportionnelle au nombre d'isolats.

Les MSTs de la **Figure 64** représente les motifs spoligotypes retrouvés dans cette étude, en fonction de divers paramètres : les lignées phylogénétiques, l'ethnicité des patients, le statut VIH/SIDA, l'occupation (éleveurs ou non-éleveurs), et la résidence (urbaine ou rurale).

On remarque notamment une association significative des souches des patients VIH-positifs parmi les souches de la lignée Manu (avec une valeur P = 0.019 lorsque l'on compare avec les souches de la lignée CAS). De plus, un nombre statistiquement significatif de patients séropositifs appartenait à l'ethnie Amhara (valeur P = 0,009). De plus, un nombre important de patients atteints du VIH étaient des non-éleveurs (valeur P = 0,012). Les patients appartenant à l'ethnie Amhara comptaient plus de résidents urbains tandis que les patients de l'ethnie Afar étaient beaucoup plus ruraux (valeur P <0,001). Les patients de l'ethnie Amhara étaient plutôt des non-éleveurs tandis que les patients d'origine ethnique Afar étaient

principalement des éleveurs (p <0,001); ce résultat coïncide aussi avec le fait que sérologie VIH positive ait été trouvée de manière plus signifiante chez les patients de l'ethnie Amhara (essentiellement non-éleveurs).

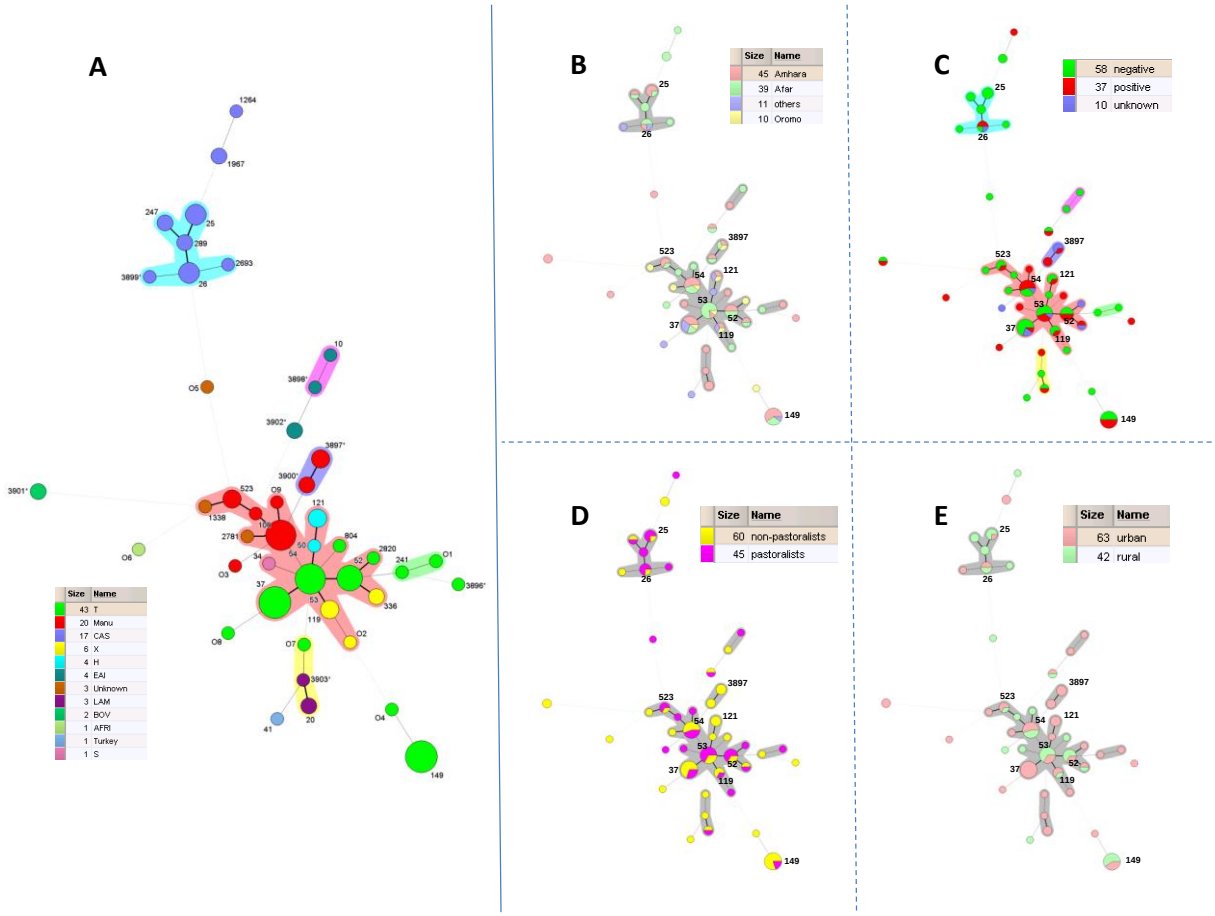


Figure 64 : MSTs construits sur la base des données de spoligotypage et en fonction de divers paramètres : (A) les lignées phylogénétiques ; (B) l'ethnicité ; (C) le statut VIH ; (D) l'occupation (éleveurs ou non-éleveurs) ; (E) la résidence (rurale ou urbaine).

Conclusion

En résumé, contrairement aux études précédentes, nous avons montré que les souches de la lignée Manu constituent non seulement une lignée dominante, mais sont également associés à l'infection au virus VIH dans la région d'Afar, en Ethiopie. Le niveau élevé de regroupement suggère la présence d'une transmission récente qui devrait être étudiée plus en détail en utilisant des marqueurs de génotypage supplémentaires.

2.4 Epidémiologie de la tuberculose en Asie de l'Ouest

« Cette roue sur laquelle nous tournons est pareille à une lanterne magique. Le soleil est la lampe; le monde, l'écran; Nous sommes les images qui passent. » Omar Khayyam.

« Par certains côtés, l'ambiguïté du terme « Moyen-Orient » est un avantage, puisqu'il peut être utilisé dans des cultures et des circonstances politiques différentes. Cette ambiguïté gêne certains géographes, qui, cependant, ont essayé de populariser « Asie du Sud-ouest » comme alternative (« Southwest Asia »), bien que cela n'ait eu que peu de succès. D'autres ont fait leur apparition comme « Asie de l'Ouest » (West Asia), lequel est devenu le terme d'usage en Inde, à la fois par le gouvernement et les médias. » (http://fr.wikipedia.org/wiki/Moyen-Orient).

Dans cette partie, nous visiterons 2 pays d'Asie de l'Ouest, l'Irak (**Article 12**) et le Sultanat d'Oman (**article en préparation**).

*2.4.1 Etude épidémiologique en Irak « après guerre » (**Article 12**)*

« L'Irak couvre une grande partie de la Mésopotamie, berceau de grandes civilisations parmi les plus anciennes. C'est sur les berges du Tigre, passant par Bagdad, que l'écriture est née, il y a 5 000 ans. Aux époques achéménide, parthe et sassanide, le territoire de l'Irak (l'empire sémite de Babylone) est intégré à l'Empire perse, formant, peu avant sa conquête par les Arabes (au IIIème siècle par la tribu des Banu Lakhm) et son islamisation, la province sassanide du Khvarvaran. » (http://fr.wikipedia.org/wiki/Irak).

Présentation de l'étude

Notre objectif était de mettre en évidence et caractériser les souches de *M. tuberculosis* circulant à Bagdad (**Article 12**), ainsi que dans d'autres villes de l'Irak. Un total de 270 souches isolées entre 2010 et 2011 à partir de patients atteints de tuberculose ont été analysées par spoligotypage alors que le typage MIRU-VNTR a été réalisé pour un sous-échantillon de souches (n=110). L'analyse a révélé que 94,1% des isolats appartenaient à des clades de génotype connu : la lignée CAS (39,6%), la lignée T « mal définie » (29,6%), la lignée Manu (7,4%), la lignée Haarlem (7%), la famille Ural (4,1%), la lignée LAM (3,3%), la lignée X (0,7%), la lignée Turkey (précédemment nommée LAM7-TUR ; 0,7%), la lignée EAI (0,7%), la lignée S (0,7%), et la lignée « Unknown » (5,9%).

Résultats principaux

Concernant les informations obtenues grâce à un arbre MST (**Figure 65**), les résultats obtenus, sur les 270 isolats de MTBC, ont montré que 2 profils spécifiques, SIT1144/T1 et SIT309/CAS1-Delhi, étaient majoritaires dans cette étude (représentant 6,3 % des isolats pour chaque profil). Les relations d'évolution existant entre les isolats iraquiens, visibles sur l'arbre (**Figure 65**), montrent clairement que l'épidémie de tuberculose observée en Iraq est essentiellement due à 2 groupes de souches de MTBC appartenant aux lignées T et CAS.

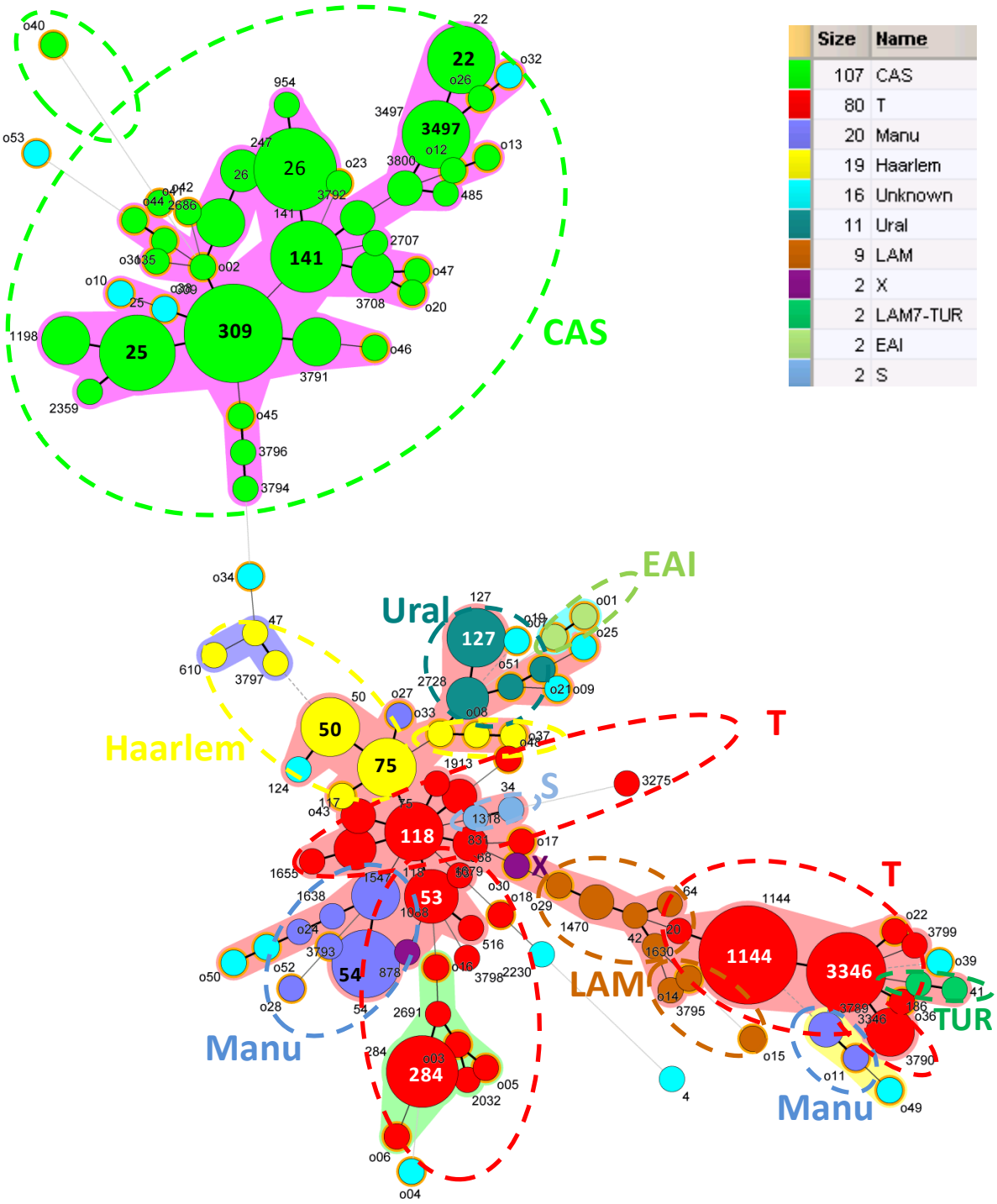


Figure 65 : MST représentant les profils de spoligotypage présents dans notre étude, ainsi que les lignées phylogénétiques correspondantes.

Comme le montre la **Figure 66** (géoréférencement réalisé avec une API Google), il existe des différences significatives entre la ville de Bagdad et les autres villes iraquiennes, en matière d'informations démographiques et de données de résistance aux médicaments. En effet, avec un rapport de masculinité hommes/femmes de 1,49 à Bagdad, contre 3,04 dans les autres gouvernorats, la proportion de patients féminins est nettement supérieure dans la ville de Bagdad (valeur p = 0,009 ; *Odds Ratio* = 0,49 et IC 95 % (0,28 ; 0,86)). Le rapport entre les cas traités pour la première fois et les cas retraités varie considérablement d'un groupe à

l'autre, la proportion de patients retraités étant plus élevée dans les autres gouvernorats iraquiens (valeur p = 0,007 ; OR = 2,1, IC 95 % (1,19 ; 3,62)). Enfin, le taux de TB MDR est supérieur dans les gouvernorats autres que Bagdad (p > 0,12 ; différence non statistiquement significative).

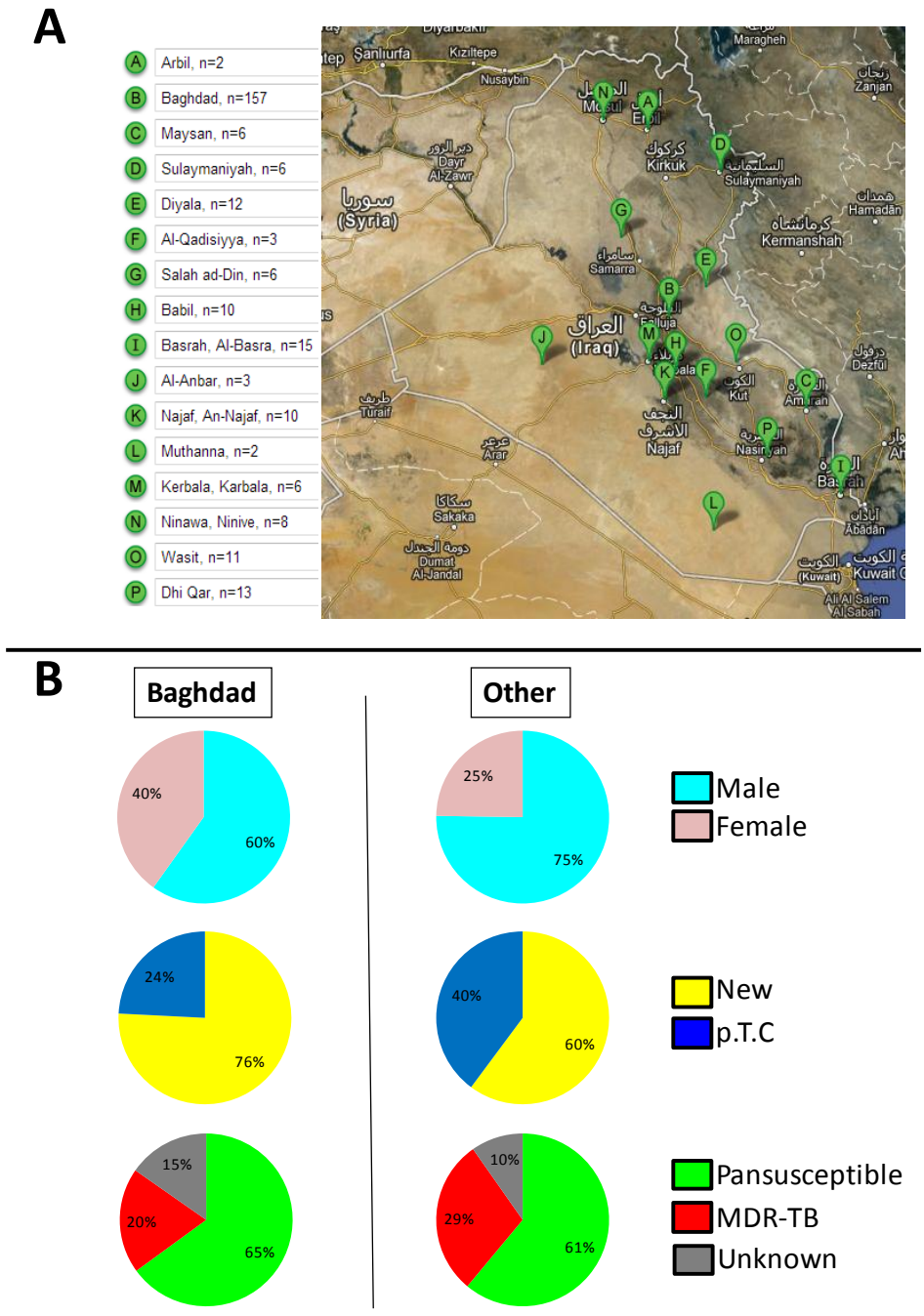


Figure 66 : (**A**) Capture d'écran de Google Maps montrant la répartition des souches de MTBC en fonction de la ville d'isolement, à Bagdad et dans d'autres villes d'Iraq (n = 270 isolats), et (**B**) comparaison des données en fonction de l'origine des patients et du rapport de masculinité, de la phase de traitement ou de la résistance aux médicaments à Bagdad et dans les autres gouvernorats.

Conclusion

Pour conclure, notre étude représente une première initiative de description épidémiologiques et évolutionnaire des isolats de *M. tuberculosis* en Irak. Cette étude a permis particulièrement de mettre en évidence la prédominance du profil SIT1144/T1. De plus, en utilisant les 24-loci MIRU-VNTRs, nous avons relevé un fort taux de transmission récente (RTI) de 33.6%.

2.4.2 Etude épidémiologique au Sultanat d'Oman (travail réalisé avec la collaboration de Dr. JPN Singh)

« Le Sultanat d'Oman compte environ 2,8 millions d'habitants (recensement de 2010). Plus de 0,8 millions d'habitants sont d'origine étrangère, ce qui représente environ 1/3 de la population totale du pays. La concentration de travailleurs immigrés du pays est l'une des plus hautes au monde. La société Omanaise est principalement une jeune société avec une estimation de 41.2% de la population âgée de moins de 15 ans. L'Oman partage une histoire particulière avec la Tanzanie. » (http://en.wikipedia.org/wiki/Demographics_of_Oman)

Présentation de l'étude

Notre étude a reposé sur la caractérisation, par la méthode de spoligotypage, de 786 isolats de MTBC, incluant 316 isolats publiés provenant d'une étude précédente (**Al-Maniri et coll., 2010**). La comparaison des données publiées par rapport aux données non-publiées a permis de mettre en évidence l'évolution moléculaire du paysage génotypique dans le Sultanat d'Oman.

Résultats principaux

Les résultats ont montré que les lignées majoritaires étaient les lignées CAS et EAI avec la présence de motifs spoligotypes très diversifiés comme le montre le spoligoforest basé sur l'algorithme de Fruchterman-Reingold (**Figure 67**). A travers ce spoligoforest, on remarque que le profil SIT26/CAS1-Delhi est le plus prédominant (n=109 isolats). Les profils suivants sont eux aussi visibles dans l'étude : SIT25/CAS1-Delhi (n=42), SIT11/EAI3-IND (n=33), SIT1/Beijing (n=31), SIT8/EAI5 (n=29), SIT3804/EAI1-SOM (n=26). Ce dernier SIT3804 a été nouvellement créé, et est retrouvé uniquement en Oman, ce qui pourrait suggérer l'endémicité de ce profil dans ce pays. La description de ce motif spoligotype est similaire au profil classique SIT48/EAI1-SOM (avec l'absence de l'espaceur 40), cependant, nous pouvons également noter l'absence de l'espaceur 4 et des espaceurs 10-17 chez ce nouveau profil SIT3804. Aussi, nous avons remarqué la présence non négligeable du profil SIT59/LAM11-ZWE (n=11) qui s'expliquerait par l'histoire partagée entre l'Oman et la Tanzanie (pays voisin de la Zambie et du Zimbabwe où le profil SIT59 est majoritaire).

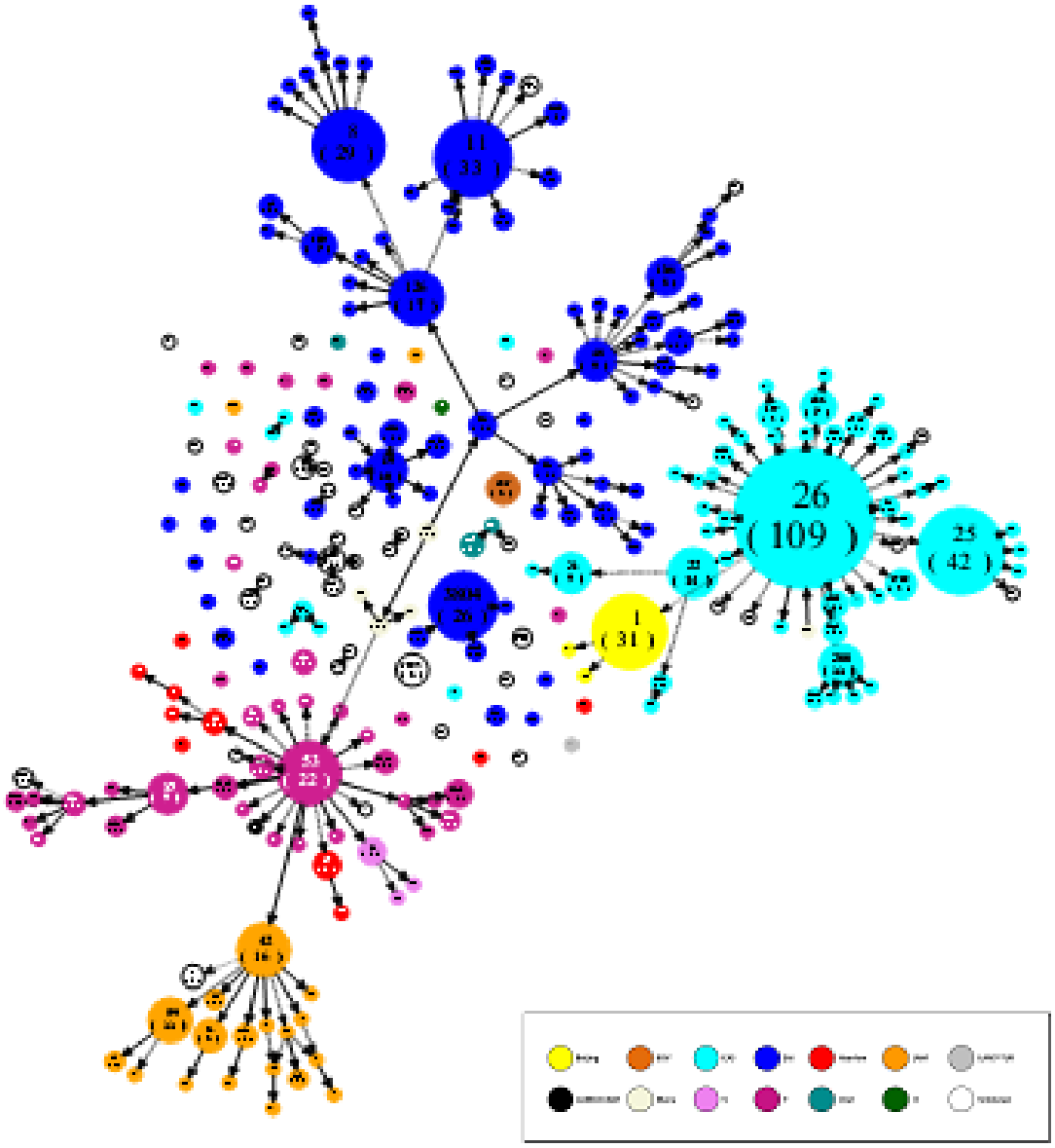


Figure 67 : Spoligoforest basé sur l'algorithme de Fruchterman-Reingold. Les couleurs représentent les différentes lignées phylogénétiques, notamment CAS (cyan), EAI (bleu), T (violet foncé), LAM (orange). Une meilleure qualité d'image est disponible sur demande.

Des différences significatives ont été observées lors de la comparaison des isolats provenant des patients d'origine étrangère par rapport aux Omanais, en fonction des données précédemment publiées (souches isolées entre 2005-2007) et des données non-publiées (souches isolées entre 2008-2010) (**Figure 68A**). La **Figure 68B** montre un arbre MST basé sur les profils de spoligotypage sur lequel on peut bien mettre en évidence le fait que certains motifs spoligotypes sont plus visibles parmi les patients immigrés (c'est le cas des motifs SIT11/EAI3-IND et SIT288/CAS2), alors que d'autres motifs sont plus majoritaires chez les patients Omanais (c'est le cas des SIT25/CAS1-Delhi, SIT8/EAI5, et SIT3804/EAI1-SOM). De plus, la lignée CAS était particulièrement plus répandue chez les patients Omanais, tandis que la lignée EAI était plus visible chez les patients immigrés (**Figure 68C**), malgré une certaine diversité de répartition des lignées en fonction des différentes régions du Sultanat d'Oman.

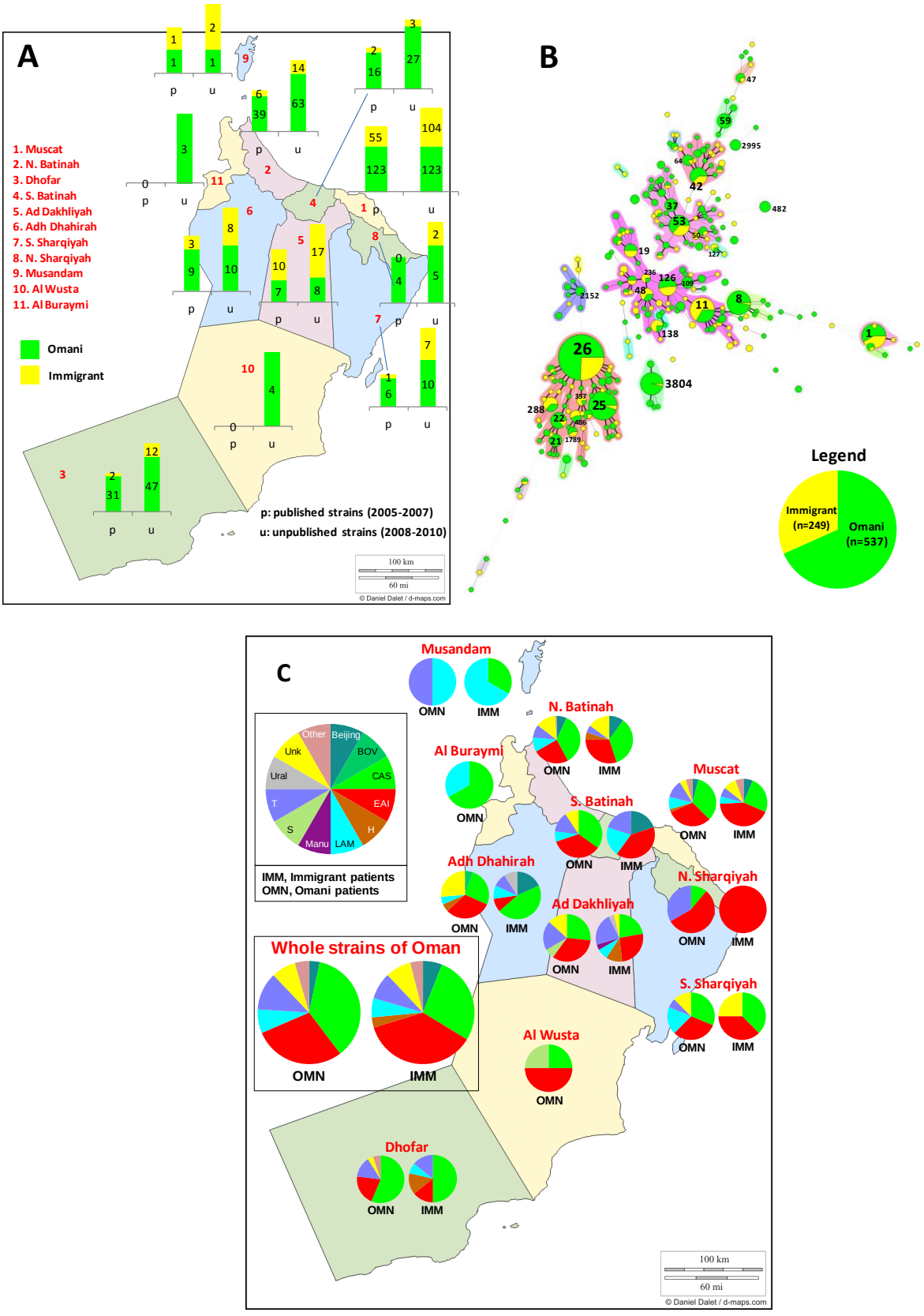


Figure 68 : Distribution des patients en fonction de leur origine, et des différentes régions du Sultanat d'Oman : (**A**) Carte géographique répertoriant les immigrants par rapport aux Omanais ; (**B**) MST représentant les profils spoligotypes des patients en fonction de l'origine ; (**C**) Distribution des lignées en fonction de l'origine des patients et de la région du pays.

Conclusion

En conclusion, la caractérisation des souches de *M. tuberculosis* des patients du Sultanat d'Oman a révélé à la fois une forme de similarité avec la présence de profils couramment détectés dans cette région ou les régions voisines, mais aussi une forme de spécificité des isolats avec notamment la présence du profil SIT3804. Des analyses supplémentaires, telles que les typages MIRU-VNTRs, seraient très bénéfiques pour mieux identifier et comparer les isolats circulant dans ce pays où le taux d'immigration est particulièrement élevé.

2.5 Etude épidémiologique de la tuberculose dans une population pédiatrique de l'Inde (**Article 13**)

« Sur la terre tout a une fonction, chaque maladie une herbe pour la guérir, chaque personne une mission. » Proverbe Indien.

« L'Inde est le deuxième pays le plus peuplé du monde. Ce pays est un foyer de civilisations parmi les plus anciens du monde, la civilisation de la vallée de l'Indus s'y est développée dès 5000 av. J.C. Le sous-continent indien a abrité de vastes empires et est présent sur les routes commerciales dès l'antiquité. » (http://fr.wikipedia.org/wiki/Inde)

Présentation de l'étude

Notre étude (**Article 13**) s'est focalisée sur une population de patients pédiatriques dans la ville de New Delhi. Un total de 111 patients pédiatriques a été analysé. La majorité des patients (n= 97/111 ou 8,38%) appartenait au groupe d'âge 11-15 ans, la moyenne d'âge des patients était de 13,4 ans, et le sex-ratio homme/femme était de 29/82 (0,35). Le test de résistance aux antituberculeux a révélé que, parmi les 97 patients âgés de 11 à 15 ans, 31/97 (32%) souches étaient pan-susceptibles, 66/97 (68%) étaient résistantes à au moins un antibiotique ; dans ce même groupe 55/97 (56.7%) étaient multi-résistantes (MDR).

Résultats principaux

Le spoligotypage, illustré par un MST et deux Spoligoforests en **Figure 69**, a permis de mettre clairement en évidence la prédominance des motifs suivants : le SIT1/Beijing (28/97 soit 28.9%), suivi par le SIT26/CAS1-Delhi (27/97 soit 27.8%), le SIT53/T1 (6/97 soit 6.2%), le SIT25/CAS1-Delhi (3/97 soit 3.1%), et le SIT250/Beijing (2/97 soit 2.1%). Dans cette étude, les lignées phylogénétiques prédominantes étaient : CAS (n=41 ou 45.56%), Beijing (n=30 ou 33.33%), et EAI (n=5 ou 5.56%).

L'implication des lignées Beijing et CAS1-Delhi chez les patients pédiatriques, suggère que ces 2 lignées jouent un rôle majeur dans la transmission active de la maladie. Cette étude a également révélé une forte prédominance du motif SIT1/Beijing par rapport à d'autres études menées en Inde. On remarque aussi un net contraste lorsque l'on compare la distribution des SITs prédominants avec les autres isolats recueillis en Inde, ainsi que les isolats enregistrés dans les pays voisins tels que le Bangladesh, le Pakistan, le Népal, le Sri Lanka, la Chine, le Myanmar, et l'Arabie Saoudite (**Tableau 24**)

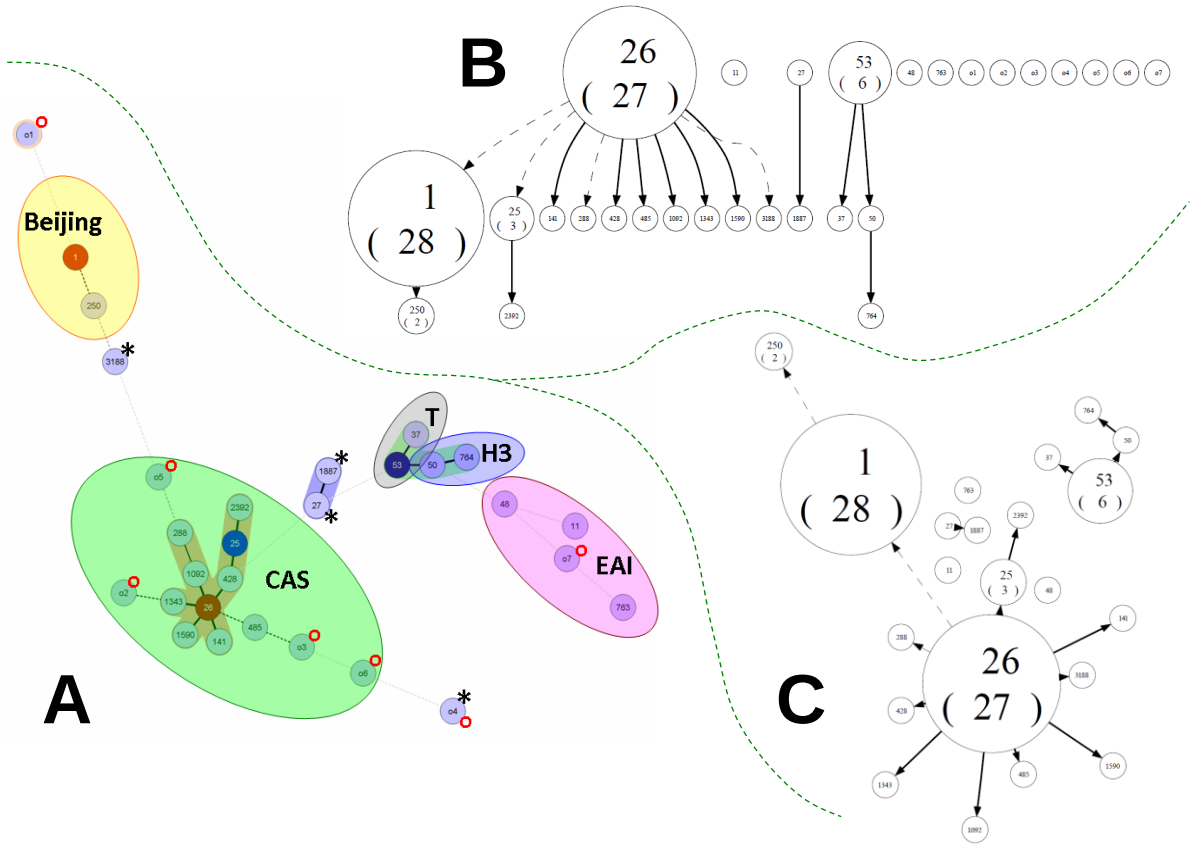


Figure 69 : Evolution des motifs spoligotypes représentée par un MST (**A**), un « Hierarchical layout » (**B**), et un arbre basé sur l'algorithme de Fruchterman-Reingold (**C**). Dans le MST, la lettre « o » (en rouge) désigne un motif orphelin, et les astérix « * » désignent les motifs ayant une lignée phylogénétique inconnue.

Tableau 24 : Distribution et comparaison des clusters spoligotypes majeurs de notre étude, dans les pays voisins de l'Inde.

Octal Code (SIT) Lineage	India			Distribution in neighboring countries (SITVIT2)						
	A	B	C							
	This study N/t (%)	New Delhi SITVIT2, N/t (%)	All India SITVIT2, N/t (%)	Bangladesh N/t (%)	Pakistan N/t (%)	Sri Lanka N/t (%)	Nepal N/t (%)	China N/t (%)	Myanmar N/t (%)	Saudi-Arabia N/t (%)
000000000003771 (SIT1) Beijing	28/90 (31.11)	209/2148 (9.73)	262/2647 (9.90)	181/1130 (16.02)	23/385 (2.97)	14/100 (14.0)	8/49 (16.33)	1885/3072 (61.36)	98/310 (31.61)	76/1818 (4.18)
703777740003771 (SIT26) CAS1-Delhi	27/90 (30.0)	420/2148 (19.55)	457/2647 (17.26)	83/1130 (7.35)	144/385 (37.40)	0/100 (0)	11/49 (22.45)	4/3072 (0.13)	5/310 (1.61)	78/1818 (4.29)
777777777760771 (SIT53) T1	6/90 (6.67)	78/2148 (3.63)	92/2647 (3.48)	38/1130 (3.36)	11/385 (2.86)	2/100 (2.0)	2/49 (4.08)	194/3072 (6.31)	3/310 (0.97)	157/1818 (8.64)
703777740003171 (SIT25) CAS1-Delhi	3/90 (3.33)	44/2148 (2.05)	49/2647 (1.85)	4/1130 (0.35)	6/385 (0.78)	0/100 (0)	1/89 (1.12)	1/3072 (0.03)	0/310 (0)	94/1818 (5.17)
000000000000371 (SIT250) Beijing	2/90 (2.22)	0/2148 (0)	3/2647 (0.11)	1/1130 (0.09)	0/385 (0)	0/100 (0)	0/49 (0)	8/3072 (0.26)	0/310 (0)	0/1818 (0)

Conclusion

En conclusion, notre étude témoigne d'une importante association des lignées Beijing et CAS avec la tuberculose résistante aux antibiotiques. De plus, ce travail réalisé au sein d'une population de patients pédiatriques, indique une transmission active de la maladie, et une progression particulière de la famille Beijing dans la zone étudiée. Pour prévenir la dissémination de la maladie, une grande attention des services de santé devrait être portée sur les enfants de Delhi.

2.6 Etude épidémiologique de la tuberculose en Malaisie (**travail en cours**)

« La bonne semence, même si elle tombe dans la mer, deviendra une île. » Proverbe Malaisien.

« La Malaisie est un pays d'Asie du Sud-est. Au sens strict du terme, les Malais sont les populations qui parlent la langue malaise et qui habitent le littoral oriental de l'île de Sumatra, les îles Riau, la péninsule Malaise et le littoral de l'île de Bornéo. » (http://fr.wikipedia.org/wiki/Malaisie)

Serions-nous en présence d'une nouvelle sous-lignée appartenant au groupe EAI en Malaisie ?

Présentation de l'étude

Dans cette étude, nous avons analysé 220 isolats cliniques du complexe *Mycobacterium tuberculosis*. Ces souches ont été collectées entre 2009 et 2013 dans les villes Malaisiennes de Kelantan (n=184) et de Kuala Lumpur (n=36). La méthode du spoligotypage a été mise en œuvre pour analyser ces souches Malaisiennes. La moyenne d'âge des patients était de 49,2 ans, 62,7% de ces derniers étaient de sexe masculin, et 93,2% des patients étaient d'origine Malaysienne. Les lignées majeures dans cette étude étaient les lignées EAI (n=124 ; 56%) et Beijing (n=63 ; 29%). Parmi les génotypes de la famille EAI, nous avons bien mis en évidence les isolats appartenant au motif SIT 745/EAI1, qui est un profil très spécifique de la Malaisie.

Résultats principaux

La **Figure 70** représente 2 arbres phylogénétiques (Spoligoforest et MST) illustrant l'évolution « spoligotypique » des souches de la Malaisie. On remarque bien que les profils SITs prédominants sont les suivant : le SIT 1 (Beijing) représentant 25.5% des isolats, suivi par le SIT 745 (EAI1-SOM) représentant 15.0% des isolats, le SIT 591 (EAI6-BGD1) avec 5.9% des isolats, le SIT 256 (EAI5, 5.5%), le SIT 236 (EAI5, 4.6%), le SIT 19 (EAI2-Manila, 4.1%), le SIT 89 (EAI2-nonthaburi, 2.3%), et le SIT 50 (H3, 1.4%).

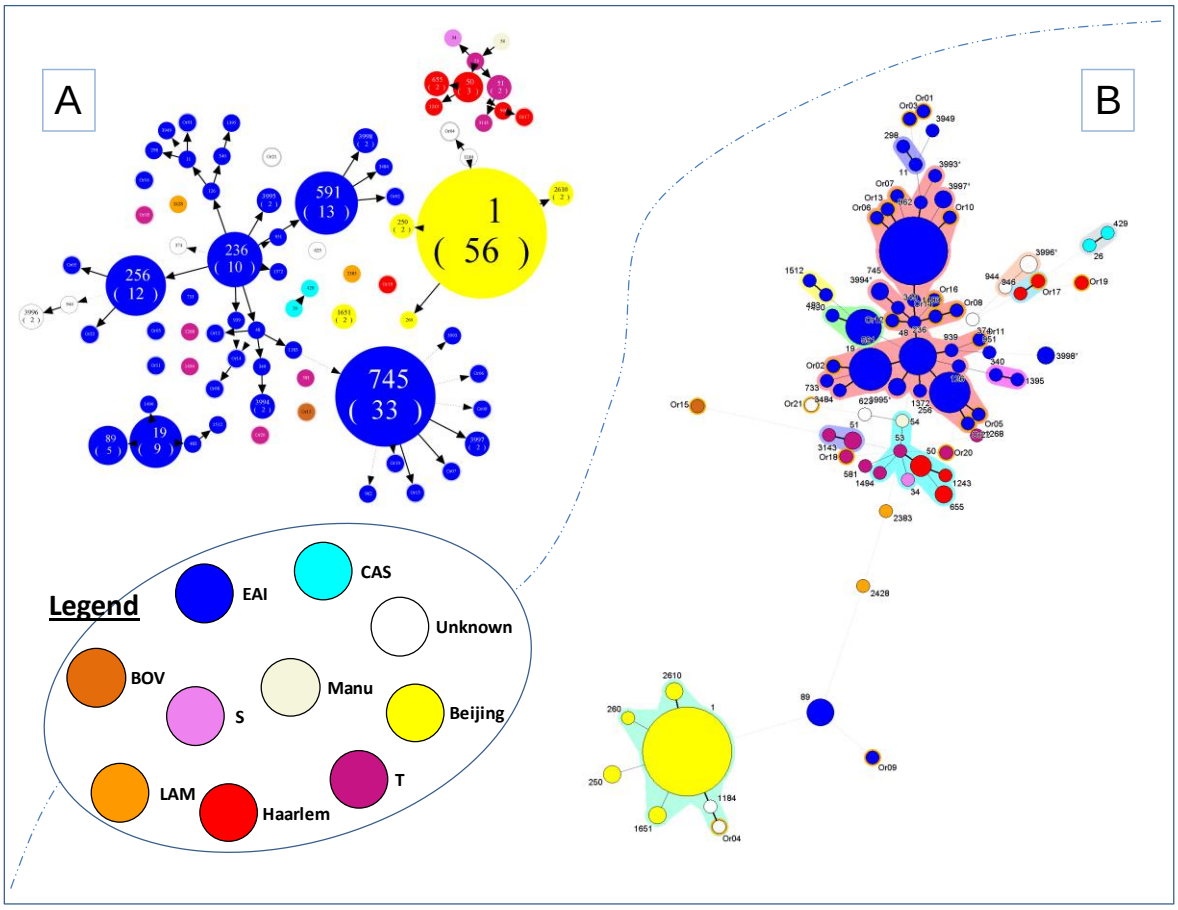


Figure 70 : Spoligoforest basé sur l'algorithme de Fruchterman-Reingold (**A**) ; et MST (**B**) mettant en évidence les relations d'évolution des spoligotypes des isolats de MTBC.

D'un point de vue plus global, nous avons également comparé les souches de notre étude avec les isolats de Malaisie précédemment enregistrés dans SITVIT2 (n=749), ainsi que les isolats de quelques pays voisins : l'Indonésie (n=422), la Myanmar (n=310), les Philippines (n=155), le Singapour (n=333), la Thaïlande (n=307), le Vietnam (n=783), la Chine (n=3072), le Japon (n=1569), La Corée du Sud (n=152), la Taiwan (n=423), l'Inde (n=2647), le Bangladesh (n=1090), le Pakistan (n= 385), le Népal (n=49), et le Sri-Lanka (n=100). La **Figure 71** illustre bien la diversité des lignées phylogénétiques qui prédominent dans les pays voisins de la Malaisie. On remarque notamment que la zone Asie du Sud-est est marquée par la forte présence des lignées EAI et Beijing.

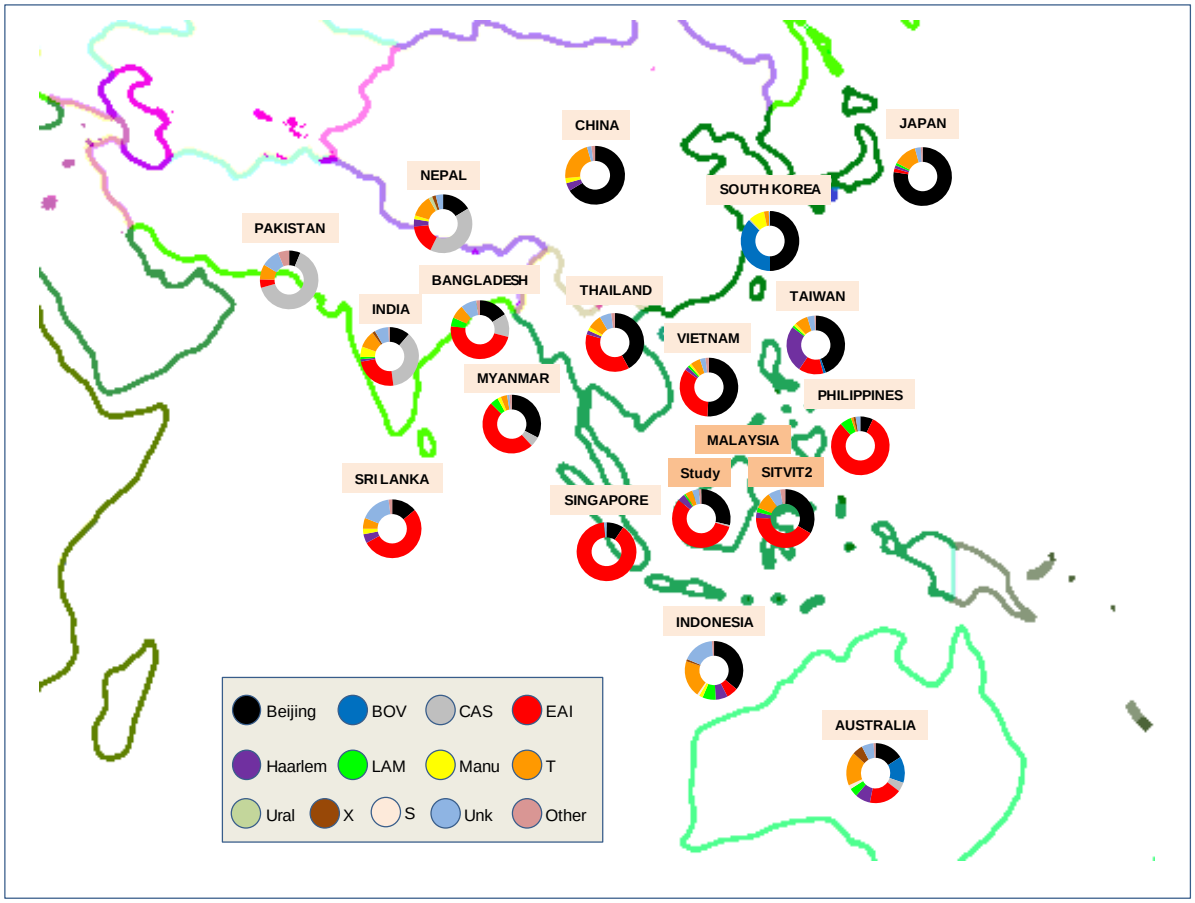


Figure 71 : Cartographie de la distribution des lignées phylogénétiques majeures dans les pays voisins de la Malaisie (données extraites de la base de données SITVIT2), et comparaison des lignées de notre étude avec les données précédemment enregistrées dans SITVIT2.

Conclusion

Pour finir, notre étude a permis de mettre en évidence l'éventuelle possibilité de la présence d'une lignée spécifique à la Malaisie, qui pourrait être provisoirement nommée « EAI1-MYS ». Le profil SIT 745 serait le spoligotype représentatif de cette nouvelle lignée (**Figure 72**). Cependant, d'autres analyses sont certainement nécessaires pour valider cette observation. Une étude mettant en corrélation les génotypes de spoligotypage et de typage MIRU-VNTRs viendrait apporter davantage de précision sur la possibilité de l'existence de cette nouvelle famille de MTBC.

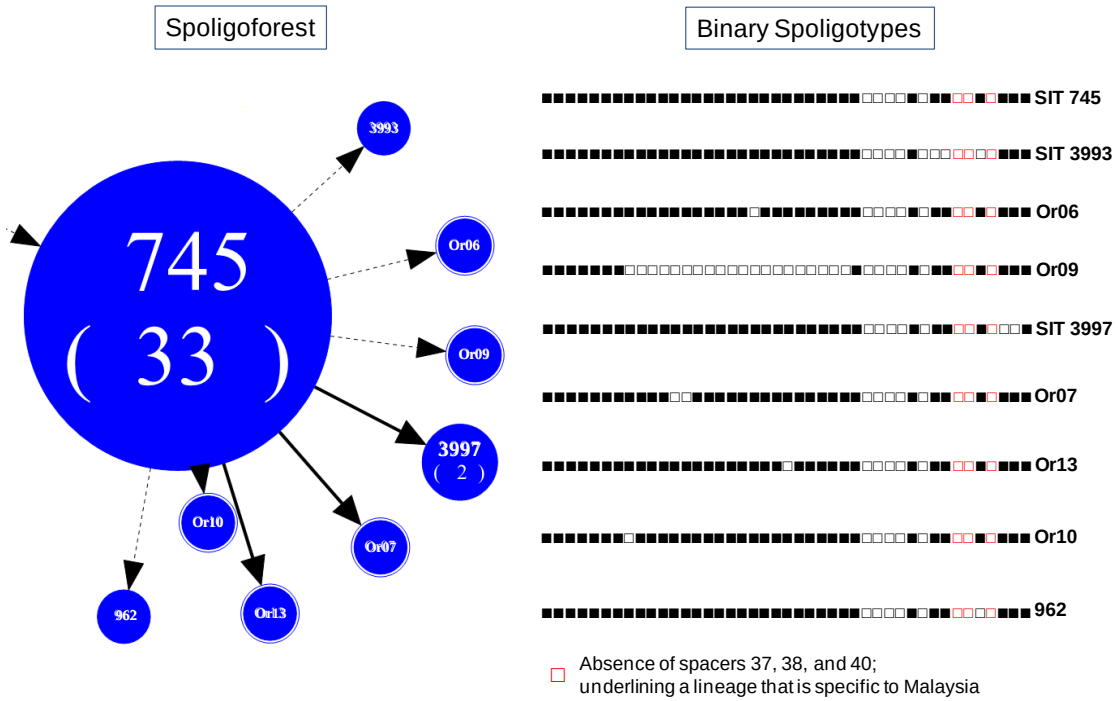


Figure 72 : Partie zoomée de l'arbre spoligoforest de la **Figure 70**, montrant le SIT 745 et ses descendants, ainsi que les profils spoligotypes de ceux-ci qui sont remarquables par l'absence spécifique des espaceurs 37, 38, et 40, qui soulignerait l'existence d'une nouvelle lignée provisoirement désignée «EAI1-MYS ».

2.7 Epidémiologie de la tuberculose dans la région de Sichuan en Chine (**Article 14**)

« Ce n'est pas le but de la promenade qui est important mais les petits pas qui y mènent. » Proverbe Chinois.

« La Chine, longtemps une civilisation très avancée, a été et reste le foyer de nombreuses innovations dans les domaines des sciences et des arts. Elle est à l'origine de nombreuses inventions majeures telles la boussole, le papier, le billet de banque, les pâtes alimentaires ou la poudre à canon. La civilisation chinoise a fortement imprégné toute l'Asie de l'Est, notamment aux niveaux religieux (confucianisme et taoïsme) et linguistique. » (http://fr.wikipedia.org/wiki/Chine)

Présentation de l'étude et résultats principaux

Nous avons analysé 306 isolats de *M. tuberculosis* provenant de la province de Sichuan en Chine (**Article 14**). Cette région est particulièrement marquée par un taux important de tuberculose résistante ou multi-résistante. Notre étude était basée sur les méthodes de typage 24-loci MIRU-VNTRs et spoligotypage avec des pouvoirs discriminants (HGDI) respectifs de 0,999 pour les MIRUs et de 0,595 pour les spoligotypes. Nous avons également observé une grande diversité allèlique parmi les différents loci des profils 24-MIRU-VNTRs tant au niveau des souches de la lignée Beijing que des non-Beijing (**Tableau 25**).

Tableau 25 : Diversité allèlique des 24-loci MIRU-VNTRs de l'étude.

VNTR locus	VNTR alias	Number of alleles	Range of repeats	Allelic diversity (*h*) for		
				Beijing isolates (*n* = 212)	Non-Beijing isolates (*n* = 94)	All isolates (*n* = 306)
154	MIRU02	3	1–3	0	0.052	0.016
580	MIRU04	7	1–6,8	0.179	0.629	0.356
960	MIRU10	6	0–5	0.588	0.582	0.628
1644	MIRU16	6	0–5	0.662	0.67	0.669
2059	MIRU20	3	1–3	0.023	0.092	0.048
2531	MIRU23	7	1,3–8	0.178	0.169	0.178
2687	MIRU24	1	1	0	0	0
2996	MIRU26	10	1–10	0.72	0.728	0.79
3007	MIRU27	5	0–4	0.596	0.648	0.622
3192	MIRU31	6	1–6	0.561	0.669	0.671
4348	MIRU39	7	0–6	0.728	0.682	0.749
802	MIRU40	5	1–5	0.529	0.712	0.63
2165	ETR-A	5	1–5	0.293	0.662	0.503
2461	ETR-B	3	1–3	0.023	0.472	0.226
577	ETR-C	5	2–6	0.104	0.189	0.128
424	Mtub04	8	0–6,8	0.725	0.694	0.75
1955	Mtub21	9	1–9	0.709	0.719	0.805
2347	Mtub29	5	2–6	0.069	0.093	0.079
2401	Mtub30	5	2–6	0.138	0.245	0.437
3171	Mtub34	4	1–4	0.111	0.151	0.127
3690	Mtub39	6	1–6	0.608	0.776	0.691
2163b	QUB-11b	14	1–10,19–22	0.775	0.837	0.821
4052	QUB-26	11	1–11	0.745	0.814	0.783
4156	QUB-4156	8	0–7	0.611	0.489	0.581

Comme on aurait pu s'y attendre, le motif spoligotype prédominant dans notre étude était le SIT1/Beijing (n=194 ; 63,4% des isolats), suivi par le SIT53/T1 (n=11 ; 3.7%), le SIT334/T1 (n=9 ; 3%), et le SIT52/T2 (n=8 ; 2,7%). La lignée Beijing était particulièrement associée à la forme de tuberculose multi-résistante (**Tableau 26**).

Tableau 26 : Distribution de la tuberculose résistante aux antituberculeux en fonction des lignées phylogénétiques.

Family	No. (%) of isolates with drug-resistant patterns					
	MDR	Non-MDR	STR	INH	RIF	EMB
Beijing	104 (33.97)	61 (19.94)	21 (6.86)	8 (2.61)	16 (5.23)	2 (0.65)
T	32 (10.46)	15 (4.9)	4 (1.31)	2 (0.65)	3 (0.98)	1 (0.33)
Unknown	9 (2.94)	1 (0.33)	0	0	0	0
Orphan	6 (1.96)	1 (0.33)	2 (0.65)	0	0	0
Manu2	4 (1.31)	6 (1.96)	0	0	0	0
H3	1 (0.33)	2 (0.65)	2 (0.65)	0	0	1 (0.33)
LAM9	1 (0.33)	0	1 (0.33)	0	0	0
Total	157 (51.3)	86 (28.1)	30 (9.8)	10 (3.27)	19 (6.21)	4 (1.31)

La tuberculose MDR a été plus fréquemment retrouvée chez les patients appartenant au groupe d'âge 21-40 ans (**Tableau 27**).

Tableau 27 : Distribution de l'âge et du sexe des patients en fonction de la résistance aux médicaments.

Parameters	Number (%) of total drug-resistant isolates investigated (this study)	Number (%) of MDR isolates investigated (this study)	Number (%) of total drug-resistant isolates (n = 708)	Number (%) of total isolates (n = 2723)
Age				
⩽20	27 (8.82)	14 (4.58)	62 (8.76)	278 (10.21)
21–40	172 (56.21)	91 (29.74)	370 (52.26)	1402 (51.49)
41–60	83 (27.12)	44 (14.38)	197 (27.82)	711 (26.11)
⩾61	24 (7.84)	9 (2.94)	79 (11.16)	332 (12.19)
Gender				
Male	204 (66.67)	110 (35.95)	483 (68.22)	1884 (69.19)
Female	102 (33.33)	47 (15.36)	225 (31.78)	839 (30.81)

Conclusion

Pour conclure, notre étude a permis de bien mettre en relief la diversité des isolats de *M. tuberculosis* résistants aux antituberculeux dans la région de Sichuan, en utilisant la double technique spoligotypage et typage 24-loci MIRU-VNTRs. Bien que la lignée Beijing était prédominante dans ce travail, d'autres génotypes tels que la famille T, Manu, Haarlem, et LAM, étaient aussi visibles. La présente étude a aussi montré que les 24-loci MIRU-VNTRs sont très utiles pour discriminer les souches de tuberculose dans cette région du monde. Enfin, nos résultats suggèrent que l'épidémie de tuberculose résistante de la région de Sichuan serait plutôt due à une résistance acquise au lieu d'une transmission de la tuberculose résistante.

2.8 Epidémiologie des formes susceptible et multi-résistante de la tuberculose en Bulgarie (travail réalisé avec la collaboration de Dr. Stefan Panaiotov)

« Il n'y a point de forêt sans arbres tordus. » Proverbe Bulgare.

« La Bulgarie est un État successeur de la Grande Bulgarie et le plus ancien pays slave encore en existence. Elle était un empire puissant au Moyen Age, dans la période de 681 à 1018 (Premier Empire) et de 1185 à 1396 (Second Empire). Après 1371, à cause d'une crise politique, le Second Empire se morcelle en quelques petites principautés, qui tombent sous les attaques ottomanes à la fin du siècle.» (http://fr.wikipedia.org/wiki/Bulgarie)

Présentation de l'étude

Dans ce travail, nous avons analysé 333 souches appartenant au complexe *M. tuberculosis*. Le spoligotypage a été utilisé pour discriminer la majorité des souches, alors que les typages 5-locus ETRs, 12-, 15-, et 24-loci MIRU-VNTRs ont été appliqués à un nombre réduits d'isolats cliniques. Contrairement à ce que l'on avait pu observer dans des études précédentes (**Figure 73**), le profil spoligotype SIT41/Turkey était le plus prédominant dans notre étude avec 86 isolats. Le SIT125 qui était désigné comme un membre de la sous-lignée LAM3 dans SITVITWEB, a été renommé « T2/LAM3 (controversial signature) » dans SITVIT2 suite à une discussion avec des experts. De plus, notre étude contenait 175 souches de tuberculose MDR et 157 souches pan-susceptibles (l'information était indisponible pour un unique isolat).

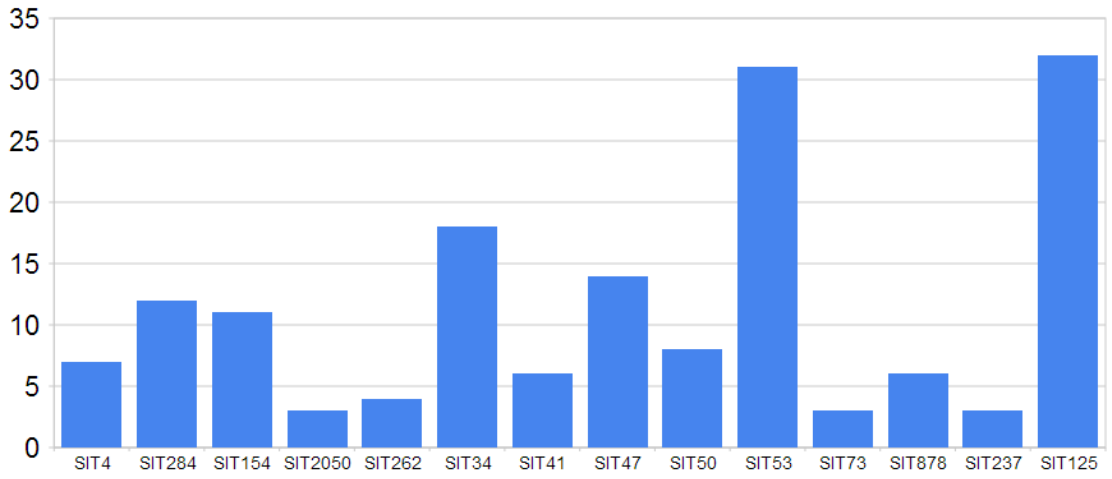


Figure 73 : Distribution des SITs (nombre d'isolats) prédominants en Bulgarie, selon SITVITWEB.

Résultats principaux

Les lignées prédominantes dans notre étude étaient LAM7-TUR (redéfinie « Turkey »), T1, et « T2/LAM3 (controversial signature) ». Le tableau suivant (**Tableau 28**) représente la distribution des profils SITs prédominants dans l'étude, en fonction de l'information de résistance (Pan-susceptible vs. MDR). On peut noter que la proportion de souches de tuberculose MDR est très élevée parmi les isolats appartenant aux SITs 41 et 284 (avec une valeur P<0,0001).

Tableau 28 : Distribution des SITs prédominants de notre étude, en fonction de la résistance aux antituberculeux.

SIT (Lineage) (number of strains)	Spoligotype Description	Octal Number	Number of Pansusceptible strains	Number of MDR strains
41 (LAM7-TUR) (n=86)	■■■■■■■■■■■■■■■■■■■■□□□□□■□□■■■■■■□□□□■■■■■■■	777777404760771	3	83
53 (T1) (n=43)	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777777777760771	21	22
125 (T2/LAM3 (controversial signature)) (n=33)	□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■□□□□■■■□■■■	000000007760731	25	8
284 (T1) (n=25)	□□□□■■■■■■■□□■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	037637777760771	9	16
1122 (T1) (n=15)	■■■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	767777777760771	15	0
154 (T1) (n=14)	■■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	757777777760771	7	7
47 (H1) (n=10)	■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	777777774020771	9	1
4 (Unknown) (n=9)	□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■□□□□■■■■■■■	000000007760771	7	2
34 (S) (n=8)	■■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	776377777760771	7	1
50 (H3) (n=7)	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	777777777720771	5	2

La **Figure 74** représente un MST basé sur tous les spoligotypes de notre étude. Alors que la **Figure 75** illustre des MSTs combinés qui mettent en corrélation le spoligotypage et les méthodes de typage 12/15/24-loci MIRU-VNTRs. Le MST basé sur les spoligotypes (**Figure 74**) ne distingue pas bien la famille « Turkey » (précédemment LAM7-TUR) des profils de la lignée LAM9.

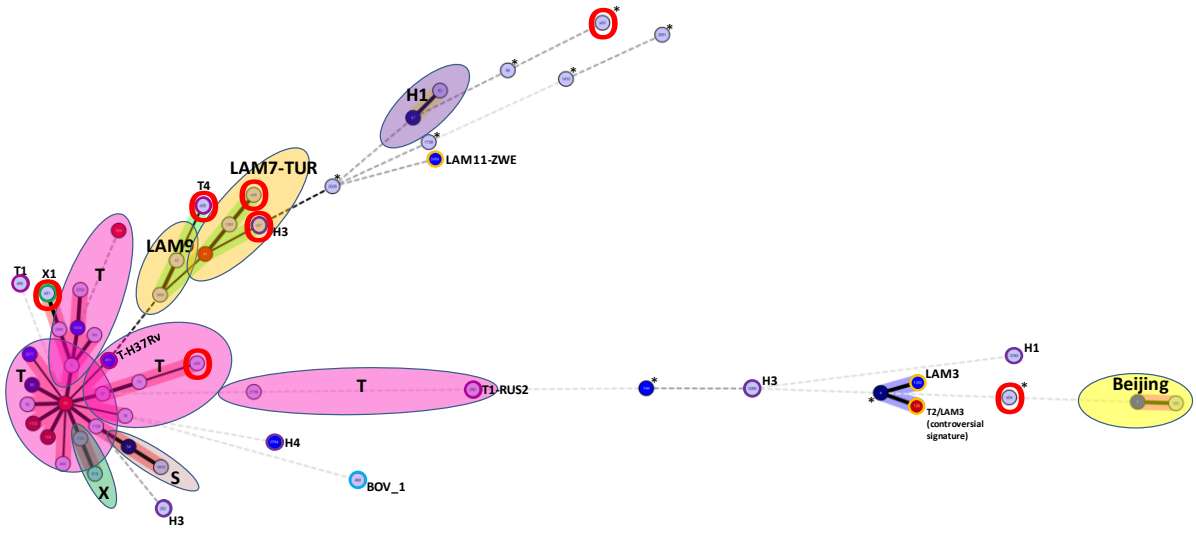


Figure 74 : MST basé sur les données de spoligotypage des isolats (n=332). La lettre « o » en rouge indique les profils orphelins, et l'étoile « * » indique les lignées non-identifiées (inconnues).

La figure basée sur l'analyse combinée des spoligotypes et des MIRU-VNTRs (**Figure 75**) montre différents degrés de discrimination des souches en fonction des différents marqueurs MIRU-VNTRs 12, 15, ou 24 loci, et nous dénombrons respectivement 46, 50, et 56 nœuds dans ces MSTs. On remarque que l'arbre combiné avec les 12-loci MIRU-VNTRs discrimine moins bien les souches appartenant au profil SIT41/Turkey que les arbres combinés avec les 15 et 24-loci MIRU-VNTRs. De plus, les profils 12-MIT 194 (214125113322), 15-MIT 203 (141132341B34343), et 24-MIT 115

(2141251133223241B3434433) étaient les plus associés au SIT41/Turkey dans cette étude (profils représentés en rouge).

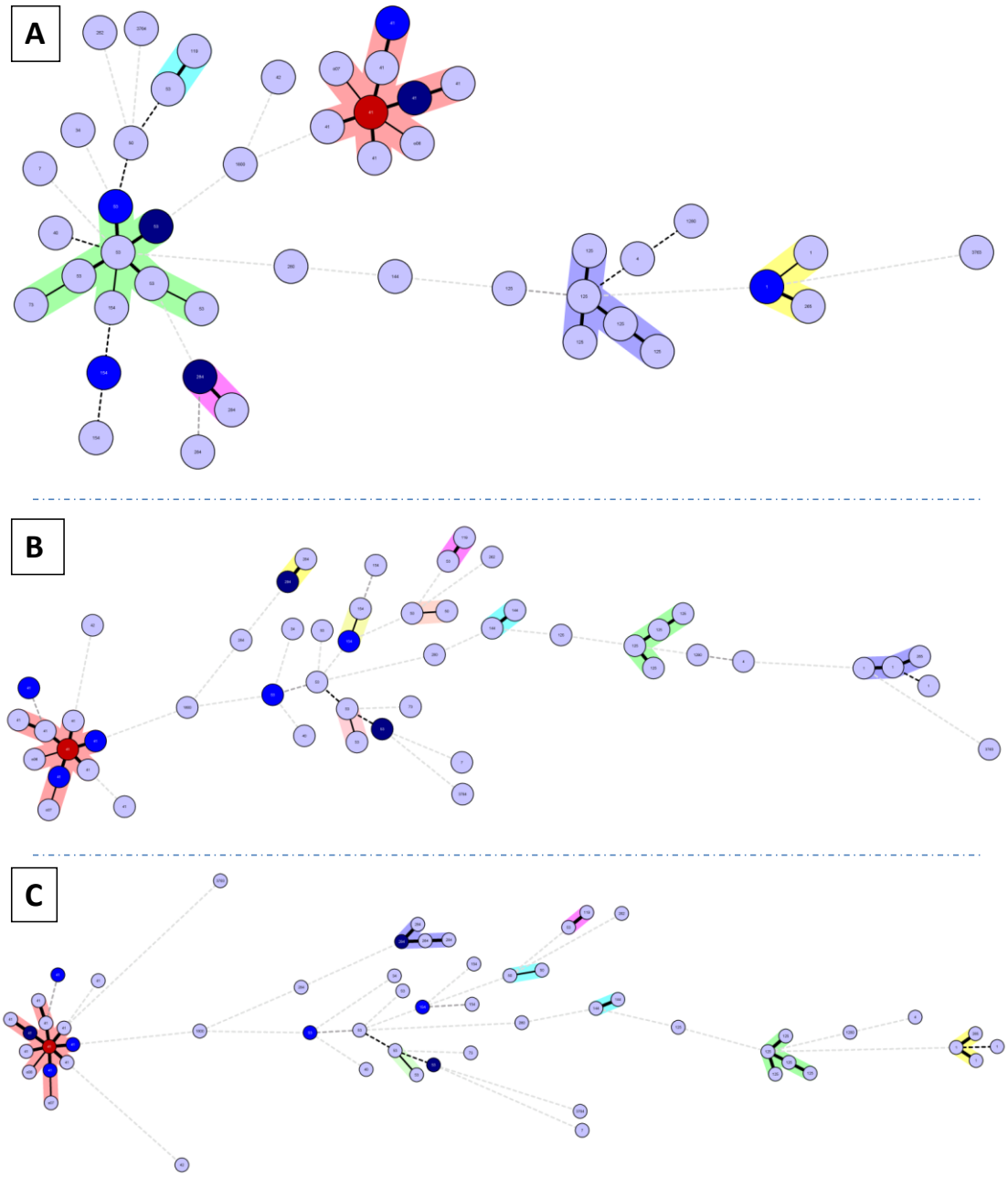


Figure 75 : MSTs combinés avec les données de spoligotypage et de typage MIRU-VNTRs : (**A**) Combinaison avec les 12-loci MIRU-VNTRs ; (**B**) Combinaison avec les 15-loci MIRU-VNTRs ; (C) Combinaison avec les 24-loci MIRU-VNTRs. Le numéro apparaissant à l'intérieur des nœuds des arbres correspond aux SITs associés aux profils MIRU-VNTRs. La couleur des nœuds est relative au nombre de souches.

Conclusion

Conclusion, notre étude fait ressortir une progression potentielle du spoligotype SIT41/Turkey en Europe de l'Est. Les isolats appartenant au SIT41 et au SIT284 étaient fortement associés à la tuberculose MDR. Ces observations suggèrent une probable transmission de la tuberculose MDR provenant de profils prédominant en Asie de l'Ouest (essentiellement la Turquie) et se dirigeant vers la Bulgarie.

2.9 Epidémiologie de la tuberculose en Scandinavie

« La corde à lier les pensées n'est pas encore tressée. » Proverbe Scandinave.

« La Scandinavie est en effet la terre d'origine d'une partie des peuples germaniques et des Vikings. La Scandinavie géographique incluait la Norvège, la Suède et des régions de Finlande, mais la Scandinavie politique incluait également le Danemark. Politiquement la Suède et la Norvège ont été unies dans une union sous un monarque et la Finlande a constitué une partie de la Suède. » (http://fr.wikipedia.org/wiki/Scandinavie).

Dans cette partie, nous verrons quelques aspects des paysages génotypiques de MTBC présents en Scandinavie. Nous nous intéresserons particulièrement à la Suède (**Article 15**) et à la Finlande (**Article 16**).

*2.9.1 Etude épidémiologique en Suède (**Article 15**)*

« La Suède parvint à conserver sa neutralité pendant la Première et la Seconde Guerres mondiales, à l'exception notable du soutien logistique et militaire apporté à la Finlande lors de la tentative d'invasion soviétique de 1939-1940. À la suite de l'échec de l'instauration d'une union de défense scandinave, le pays persista dans sa politique de neutralité au cours de la guerre froide et n'est jusqu'aujourd'hui membre d'aucun traité d'alliance militaire. Elle adhéra toutefois à l'Union européenne en 1995. » (http://fr.wikipedia.org/wiki/Su%C3%A8de)

Présentation de l'étude

Dans ce travail (**Article 15**), nous avons fait l'état des lieux des souches du complexe *M. tuberculosis* potentiellement impliquées dans l'épidémie de la tuberculose Suédoise du siècle dernier. L'étude s'est donc focalisée sur les patients nés avant 1945 (patients âgés). Un total de 409 isolats a été analysé par spoligotypage et typage IS6110-RFLP. Nous avons remarqué que les lignées « Euro-américaines » appartenant au groupe PGG2/3 (T, Haarlem, LAM, X, S) étaient plus répandues chez ces patients âgés (n=363/409 ou 88,8%), alors que les lignées anciennes du groupe PGG1 (CAS, Beijing, EAI, Manu, BOV) ne représentaient que 5,9% des isolats.

Résultats principaux

Les lignées Haarlem (41,1%) et T (37,7%) étaient les plus prédominantes dans notre étude. Nous avons dessiné 2 MSTs distincts (**Figure 76**), l'un contenant uniquement les motifs spoligotypes « SITés », et l'autre incluant les spoligotypes orphelins. Parmi les 45 souches orphelines, 22 appartenaient à la famille T, 17 appartenaient à la lignée Haarlem, 2 à la lignée BOV, une souche appartenait à la famille LAM, et 3 isolats étaient de lignée inconnue. Ces résultats indiquent que les souches isolées à partir des patients âgés sont moins clustérisées que les souches provenant des patients plus jeunes qui étaient plutôt concernés par les souches de tuberculose ancestrale appartenant au groupe PGG1.

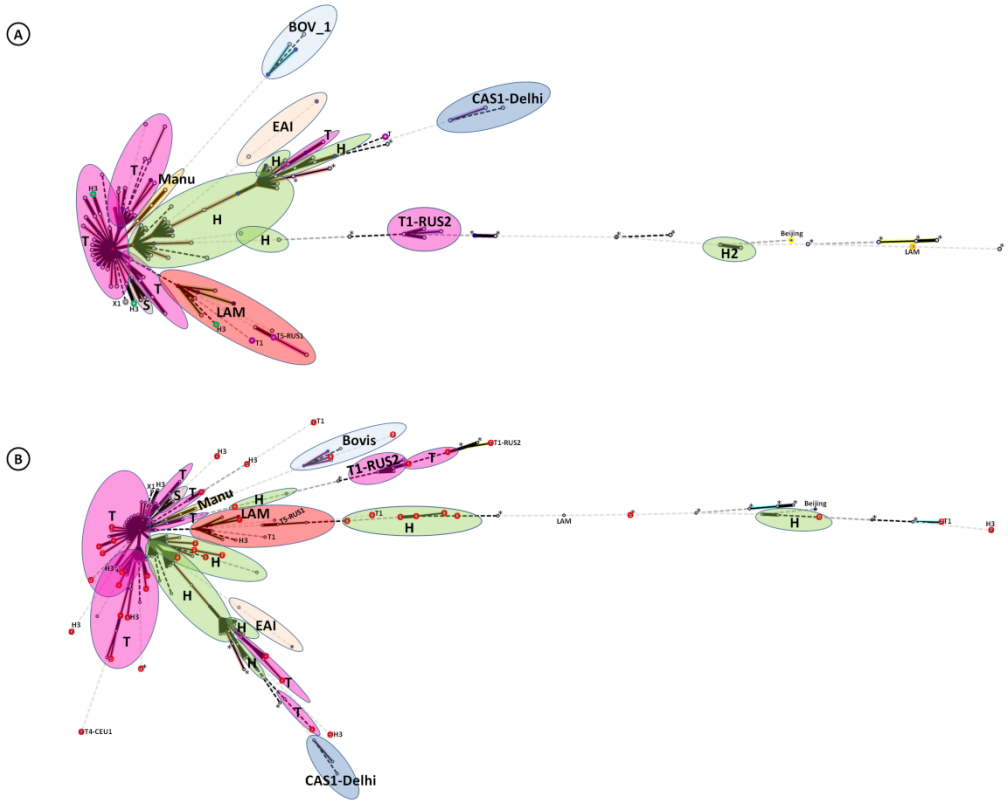


Figure 76 : MST basé sur les spoligotypes : (**A**) SITs seuls, (**B**) Toutes les souches incluant les spoligotypes orphelins. La lettre « o » en rouge indique les profils orphelins, et l'étoile « * » indique les lignées non-identifiées (inconnues).

De plus, presque toutes les souches orphelines appartenant au groupe de tuberculose récente PGG2/3 se trouvaient dans des positions terminales éloignées de leurs familles génotypiques respectives sur l'arbre MST (**Figure 76B**). Et ce, contrairement aux nœuds centraux qui ont été essentiellement définis par les familles de spoligotypes tels que SIT50/H3 et SIT53/T1. Cette observation suggère que la plupart des souches orphelines qui existaient en Suède il ya un siècle, lorsque la tuberculose était hautement endémique, ont probablement subi une extinction parce qu'elles ont évolué en tant que membres terminaux au sein de ces familles qui ont eu du mal à proliférer.

Conclusion

En conclusion, notre étude a montré que les lignées récentes appartenant au groupe PGG2/3 étaient prédominantes et généralement associées aux patients âgés suédois. Cependant, les lignées anciennes du groupe PGG1 étaient associées aux patients jeunes qui sont habituellement des patients issus de l'immigration.

*2.9.2 Etude épidémiologique en Finlande (**Article 16**)*

« Longtemps sous domination du voisin suédois, sous le règne duquel fut fondée la première capitale, Turku, la Finlande demeure marquée par les nombreux conflits qui l'ont opposée durant toute son histoire récente à la Russie, dont elle n'a obtenu son indépendance de fait qu'au terme d'une guerre civile en 1917, de deux nouveaux conflits pendant la Seconde Guerre mondiale, et enfin d'une longue tutelle indirecte connue sous le nom de finlandisation durant la Guerre froide. » (http://fr.wikipedia.org/wiki/Finlande)

Présentation de l'étude

Dans cet article (**Article 16**), nous avons analysé 1048 souches de *M. tuberculosis* isolées entre 2008 et 2011. Contrairement à l'étude précédente menée en Suède, l'immigration (bien qu'existante) est un problème moins important en Finlande. Parmi les 1048 souches analysées à l'aide du spoligotypage et du typage 15-loci MIRU-VNTRs, 733 (69,9%) provenaient de patients suédois, 264 (25,2%) provenaient de patients nés à l'étranger, et 51 (4,9%) provenaient de patients d'origine inconnue. Ces analyses ont permis une meilleure discrimination des souches de tuberculose circulant en Finlande.

Résultats principaux

Une étude comparative avec les pays voisins de la Finlande a été réalisée en dessinant une cartographie phylogénétique des souches de *M. tuberculosis* présentes dans la Scandinavie (**Figure 77**). Les lignées majoritaires de notre étude étaient les lignées T (n=346 ; 33%) et Haarlem (n=237 ; 22,6%). La distribution des lignées contenues dans notre étude, est similaire à celle qui a été recueillie précédemment (basée sur les données de SITVIT2) en Finlande. Cependant, les lignées CAS, EAI, et Beijing sont plus visibles dans notre étude. De plus, la lignée « Ural » est particulièrement retrouvée en Finlande, ce qui n'est pas le cas en Suède et en Norvège.

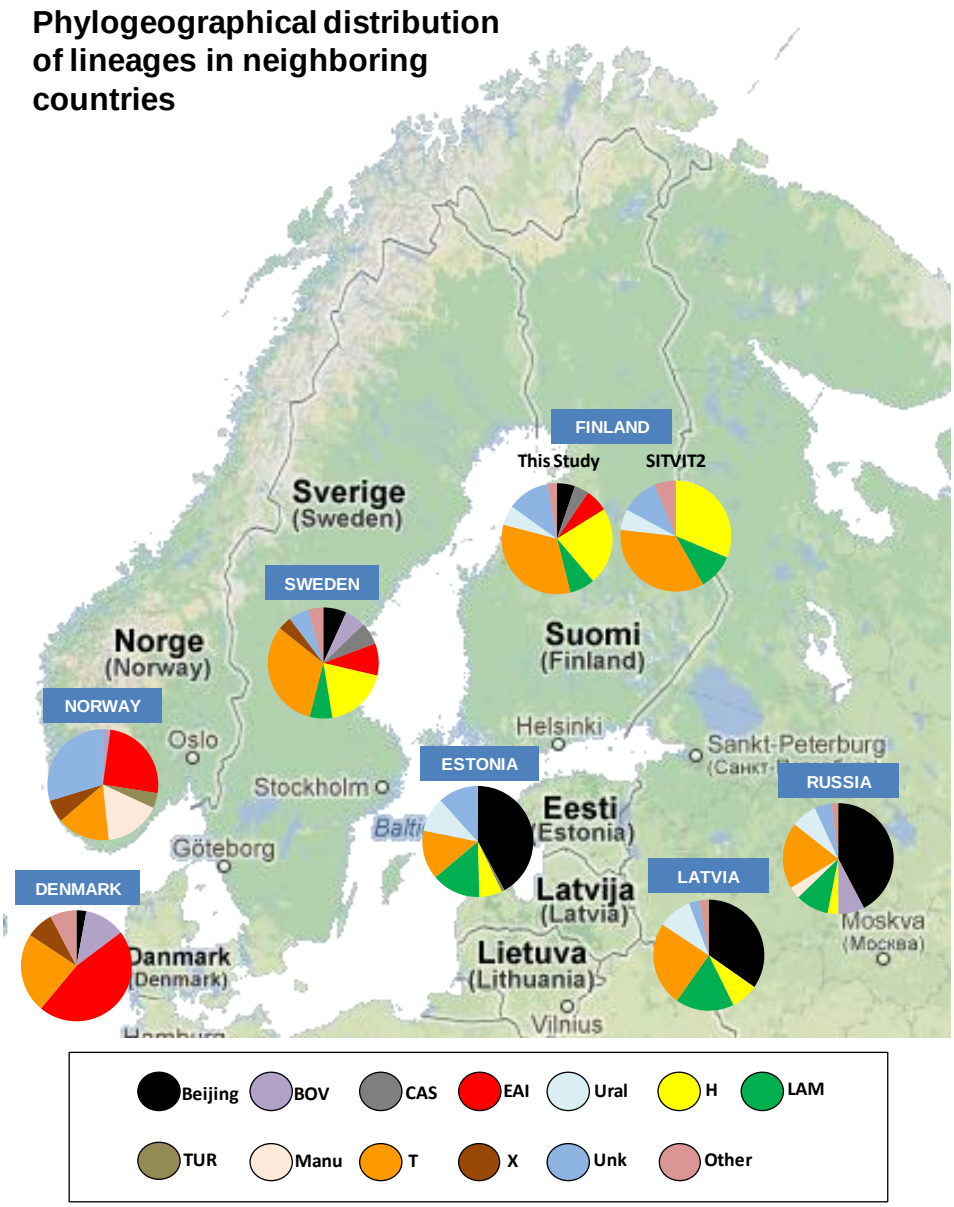


Figure 77 : Distribution phylogéographique des lignées de *M. tuberculosis* en Finlande et dans ses pays voisins.

Un arbre MST basé à la fois sur les données de spoligotypage et de typage 15-loci MIRU-VNTRs a été dessiné pour mettre en évidence la diversité des isolats de Finlande (**Figure 78**). Le cluster spoligotype-15MIRU le plus important appartenait au profil SIT42/LAM9.

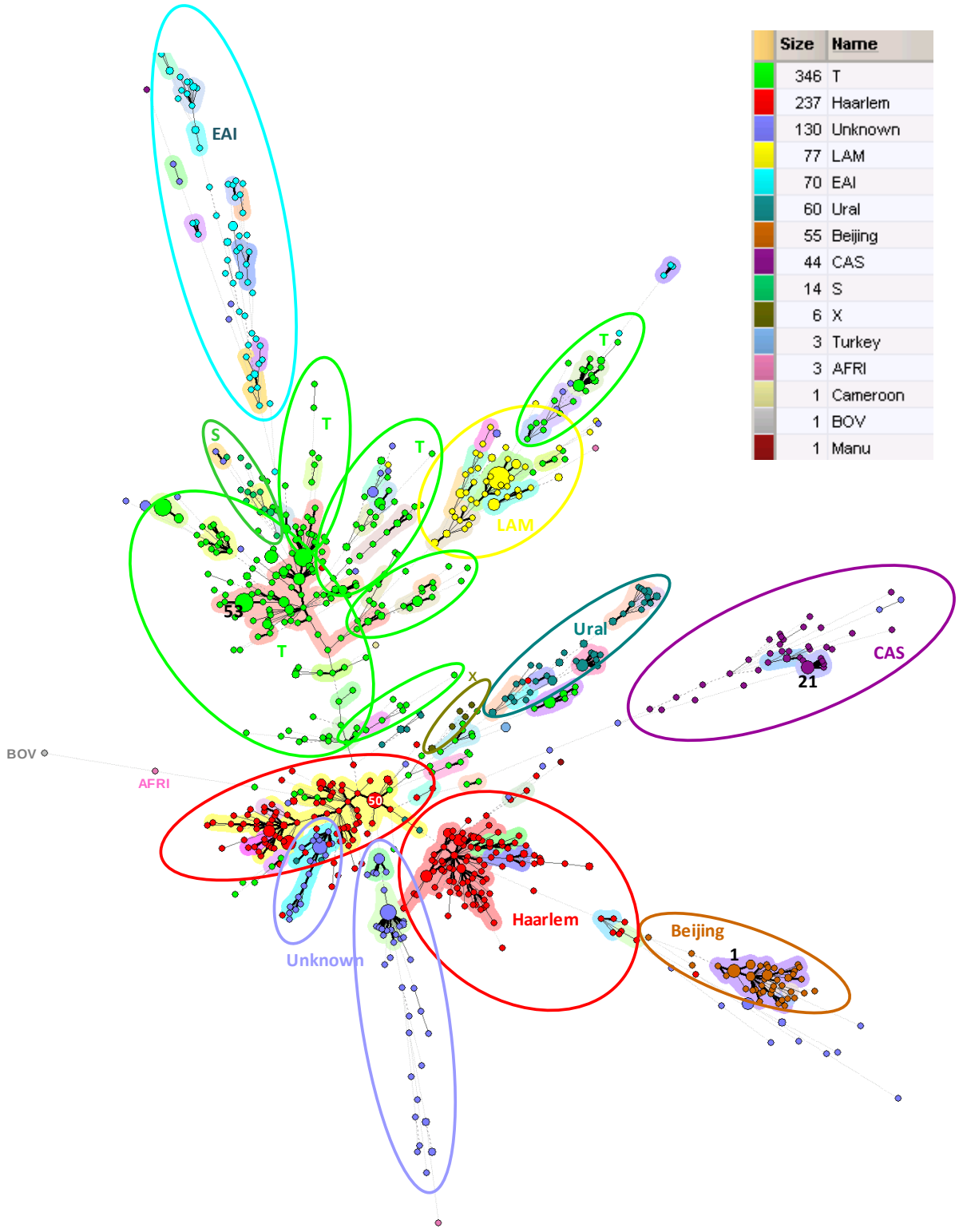


Figure 78 : MST basé sur la combinaison du spoligotypage et du typage 15-loci MIRU-VNTRs.

Cette étude nous a également permis, de montré que certains profils « F-Types » étaient exclusivement liés aux patients Finlandais. Ces « F-types » concernaient 68 isolats (**Figure 79**).

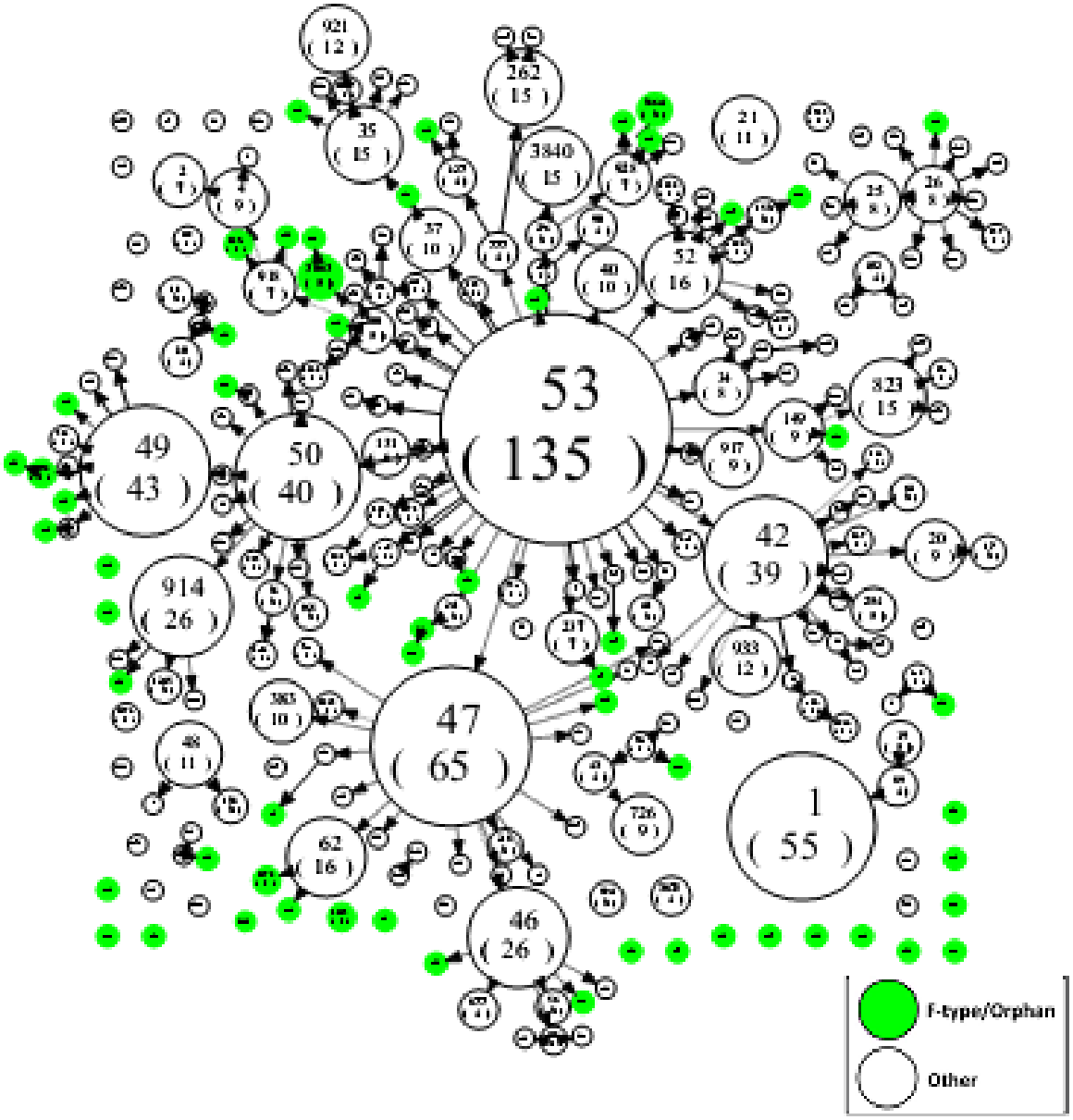


Figure 79 : Arbre spoligoforest basé sur l'algorithme de Fruchterman-Reingold, représentant les spoligotypes spécifiques à la Finlande « F-types » (coloriés en vert dans la figure). Une meilleure qualité d'image est disponible sur demande.

Conclusion

Pour conclure, nous pouvons dire qu'une grande partie des souches de Finlande a été isolée à partir de patients Finlandais plus âgés (>60 ans). Plusieurs spoligotypes non identifiés précédemment ont été révélés dans notre étude. De plus, nous avons aussi détecté des profils bien spécifiques à la Finlande (« F-types »). Finalement, la situation de la tuberculose en Finlande est bien distincte des situations rencontrées dans d'autres pays à basse incidence de tuberculose.

2.10 Epidémiologie de la tuberculose dans la région Rhône-Alpes, France (**travail en cours**)

« A chaque problème, une solution. » Proverbe Français.

« La France occupe en 2013, le cinquième rang mondial pour le produit intérieur brut. Son économie est de type capitaliste avec une intervention étatique assez forte. Elle figure parmi les premiers mondiaux dans les secteurs de l'agroalimentaire, de l'aéronautique, de l'automobile, des produits de luxe, du tourisme et du nucléaire. » (http://fr.wikipedia.org/wiki/France)

Présentation de l'étude

Nous rapportons ici les résultats d'une étude sur la diversité génotypique des souches du complexe *Mycobacterium tuberculosis* (MTBC) isolées dans la région Rhône Alpes de 2000 à 2010, dans le cadre d'une démarche qualité et épidémiologique menée par l'observatoire Rhône-Alpes des mycobactéries, différents centres hospitaliers, et laboratoires. Cette étude concernait 2257 souches (1 isolat par patient) de *M. tuberculosis* n=2226 ; *M. bovis* n=95 ; et *M. africanum* n=36. L'ensemble des souches a été typé par spoligotypage (n=2257) et une partie par 12/15-loci MIRU-VNTRs (n=1698 isolats, incluant 974 ayant un profil complet).

Résultats principaux

Les lignées génotypiques par spoligotypage dans notre étude sont : T (n=762, 33.76%) ; Haarlem (n=514, 22.77%) ; LAM (n=413, 18.30%, dont n=53, 2.35% LAM10-CAM) ; BOV (n=95, 4.21%) ; Beijing (n=84, 3.72%) ; EAI (n=66, 2.92%) ; S (n=63, 2.79%) ; AFRI (n=36, 1.60%) ; CAS (n=30, 1.33%) ; X (n=9, 0.40%) ; la famille Turquie (précédemment classifiée comme LAM7-TUR ; n=7, 0.31%). Les motifs les plus prédominants correspondent aux SIT53/T1 (n=346, 15,3%) > SIT50/H3 (n=166, 7,35%) > SIT42/LAM9 (n=125, 5.5%) > SIT1/Beijing (n=72, 3,2%) > SIT47/H1 (n=71, 3,1%). Ainsi, la lignée Euro-Américaine qui regroupe les souches d'évolution récente (T, LAM, Haarlem, X, S) regroupe 1768 ou 78,33% des souches (**Figure 80**). Le MST de la **Figure 80** permet de bien distinguer les spoligotypes appartenant aux lignées ancestrales PGG1 (EAI, AFRI, BOV, CAS, Beijing) de ceux appartenant aux lignées modernes PGG2/3 (T, LAM, Haarlem, X, S). Nous pouvons également apercevoir que la majorité des souches orphelines occupent une position terminale dans l'arbre et concernent plutôt les profils du groupe PGG2/3. Un spoligoforest a aussi été construit à partir des SITs prédominants et en fonction de l'origine des patients (**Annexe 5**). Nous notons que l'information sur l'origine atteste bien le fait que les patients d'origine étrangère sont généralement infectés par les génotypes prédominants dans leur pays d'origine.

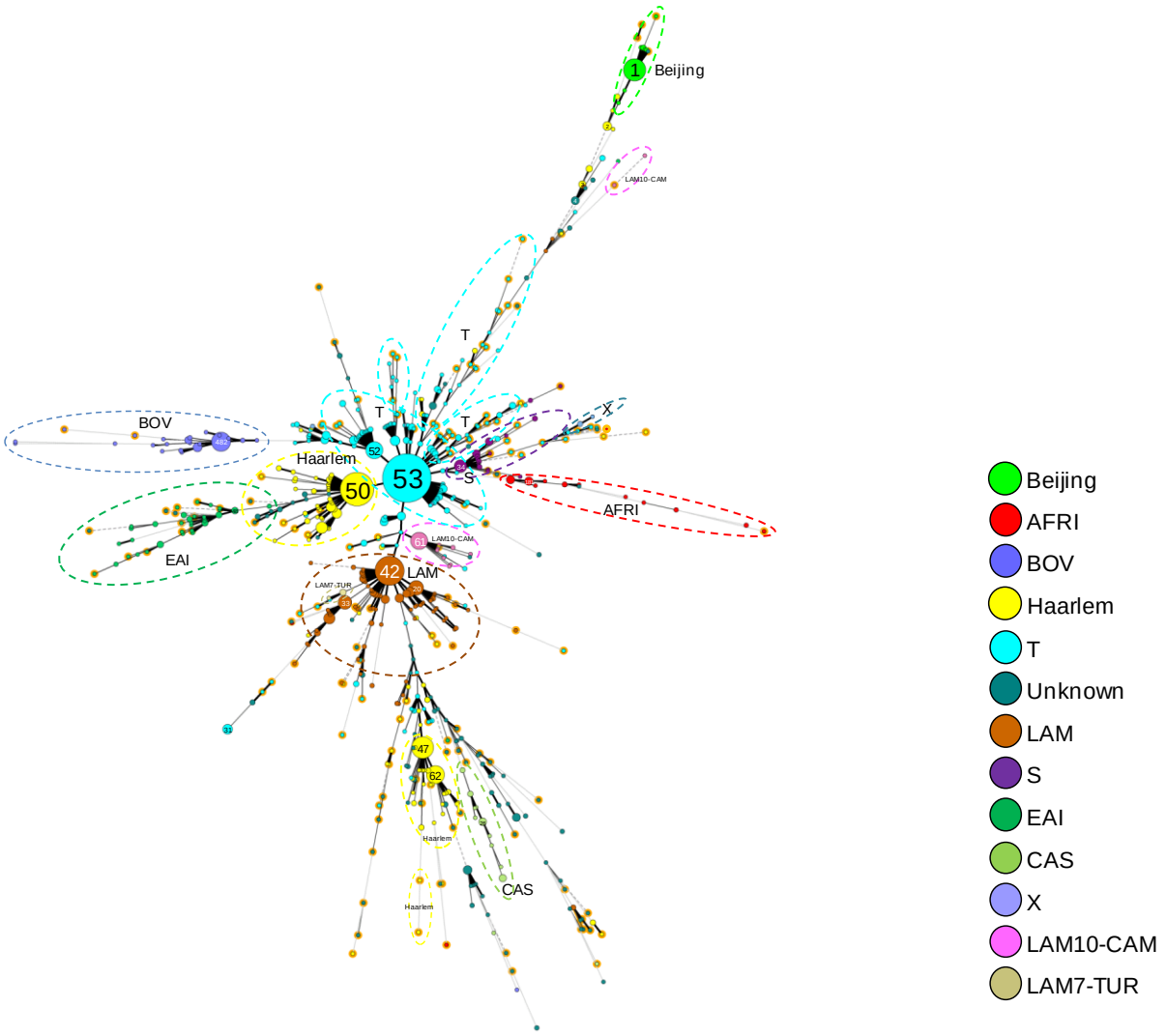


Figure 80 : MST représentant tous les motifs de spoligotypage (n=2257). La couleur de chaque nœud indique la lignée à laquelle appartient le profil. Les spoligotypes orphelins sont entourés en orange. La taille des nœuds est proportionnelle au nombre de souches et la signification des liens est similaire aux MSTs précédents.

L'origine est connue pour 927 patients : 376 (40,6%) français et 551 (59,4%) étrangers (**Figure 81**). Les patients d'origine française sont relativement plus âgés (moyenne d'âge de 58,42 ans) que les patients d'origine étrangère (moyenne d'âge de 42,38 ans). L'Afrique représente à elle seule 70% des cas chez les patients d'origine étrangère, dont 39% du Maghreb et 30,1% d'Afrique subsaharienne. Dans cet échantillon, AFRI, EAI, LAM, Beijing, S, et CAS sont majoritairement retrouvés chez les patients d'origine étrangère tandis que les familles Haarlem, T et BOV se retrouvent majoritairement chez les patients d'origine française (**Tableau 29**). Bien que cela ne soit pas très significatif, notre étude nous a également permis de mettre en évidence la présence de motifs rares principalement trouvés / isolés à partir de patients d'origine française. Nous pouvons par exemple noter la présence des SITs suivants : le SIT 488/H1 (n=3 isolats provenant de patients français), et le SIT 713/T1 (n=3).

A

	AFRI-M (%)	AFRI-E (%)	AFRI-N (%)	AFRI-W (%)	AMER-C (%)	CARI (%)	AMER-N (%)	AMER-S (%)	ASIA-E (%)	ASIA-N (%)	ASIA-S (%)	ASIA-SE (%)	ASIA-W (%)	EURO-E (%)	EURO-S (%)	EURO-W (%)	Unknown (%)
2000-2003 n=412	34 (37.36)	5 (20.00)	101 (46.33)	30 (50.85)	0 (0.00)	0 (0.00)	1 (100)	2 (66.67)	1 (16.67)	0 (0.00)	5 (45.45)	10 (31.25)	7 (28.00)	9 (24.32)	13 (50.00)	186 (48.95)	8 (0.60)
2004-2007 n=896	41 (45.05)	15 (60.00)	85 (38.99)	22 (37.29)	0 (0.00)	0 (0.00)	0 (0.00)	1 (33.33)	5 (83.33)	2 (50.00)	4 (36.36)	17 (53.13)	13 (52.00)	20 (54.05)	12 (46.15)	155 (40.79)	504 (37.70)
2008-2010 n=949	16 (17.58)	5 (20.00)	32 (14.68)	7 (11.86)	1 (100)	1 (100)	0 (0.00)	0 (0.00)	0 (0.00)	2 (50.00)	2 (18.18)	5 (15.63)	5 (20.00)	8 (21.62)	1 (3.85)	39 (10.26)	825 (61.71)
	91	25	218	59	1	1	1	3	6	4	11	32	25	37	26	380	1337

B

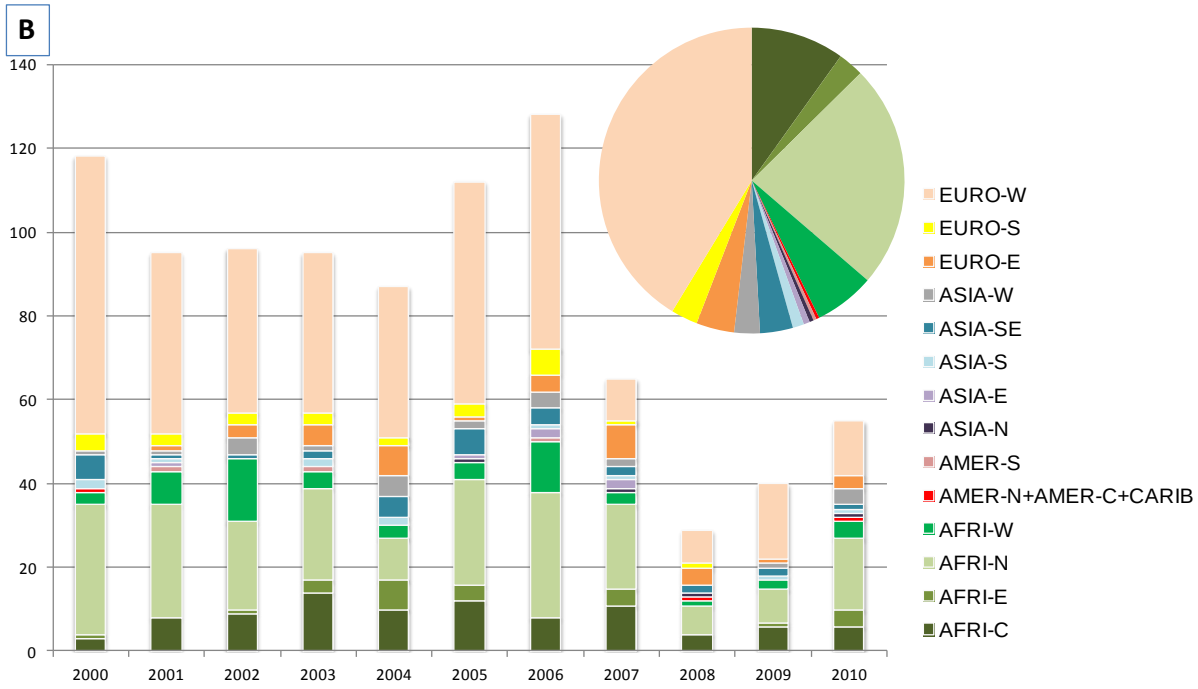


Figure 81 : Représentation du nombre de souches en fonction de la région d'origine des patients : Tableau de répartition par groupe d'années (**A**) ; et Histogramme de la répartition par année (**B**).

Tableau 29 : Distribution des lignées phylogénétiques entre les patients Français et les patients nés à l'étranger.

Phylogenetic lineages	French	Foreign-born	French + Foreign-born	French/ Total %	Foreign-born/ Total %
T	146	165	311	46,9	53,1
LAM	45	97	142	31.7	68.3
Cameroon / LAM10-CAM	5	20	25	20.0	80.0
Turkey / LAM7-TUR	0	1	1	0.0	100.0
Haarlem and Ural	106	109	215	49,3	50,7
Unknown	31	34	65	47,7	52,3
EAI	3	32	35	8,6	91,4
Beijing	9	27	36	25	75
S	10	22	32	31,3	68,8
Africanum	2	20	22	9,1	90,9
CAS	0	13	13	0	100
Bovis	19	10	29	65,5	34,5
X	0	1	1	0	100
Total	376	551	927	40,6	59,4

La **Figure 82A** corrobore bien le fait que la plupart des patients d'origine étrangère sont infectés par les génotypes prédominants dans leur pays d'origine. Alors que la **Figure 82B** montre l'évolution des lignées de M. tuberculosis appartenant aux groupe PGG1 (AFRI, CAS, EAI, Beijing, BOV) dans notre étude. On peut remarquer une certaine tendance de décroissance de la proportion de souches de la famille *M. africanum* (ou AFRI) au fil du temps.

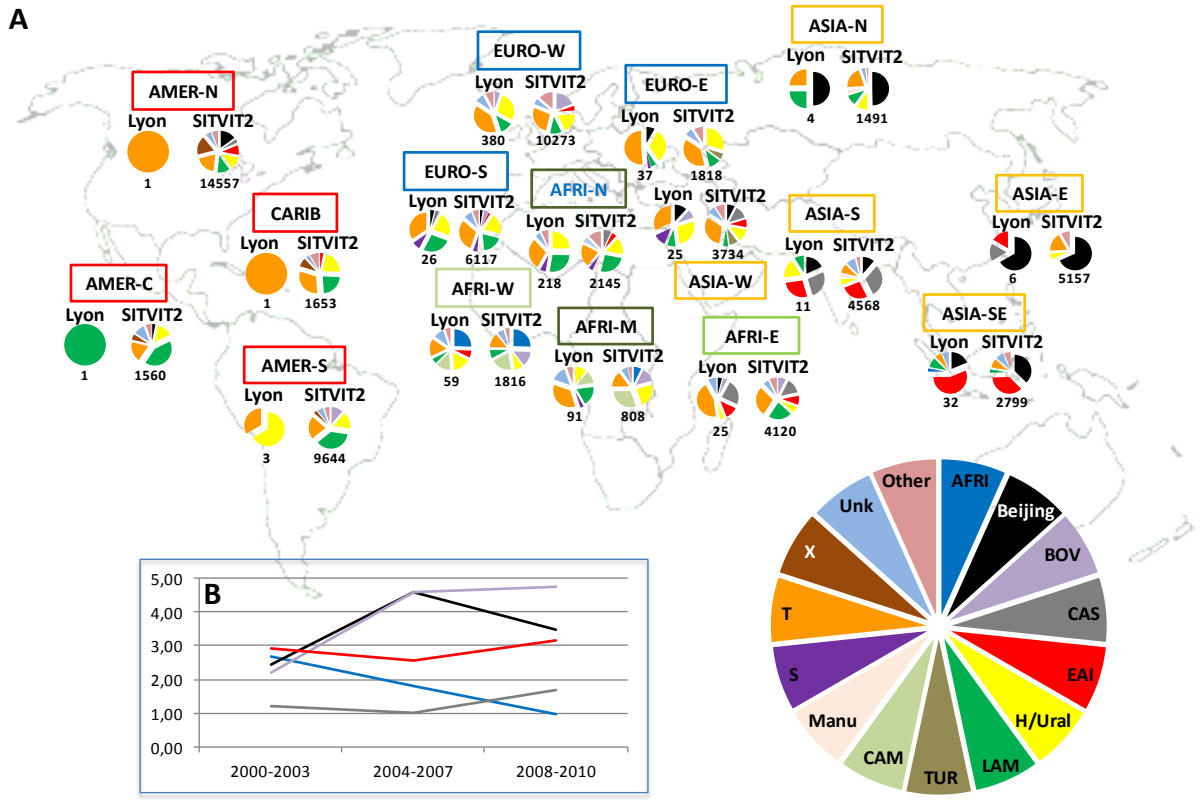


Figure 82 : Distribution phylogéographique des principales lignées de *M. tuberculosis* : (**A**) comparaison entre les données de l'étude et la base de données SITVIT2 ; (**B**) tendance globale d'évolution des souches PGG1 en fonction du groupe d'années dans la présente étude.

Les analyses de la résistance aux antituberculeux, la localisation de la maladie (pulmonaire ou extra-pulmonaire), le sexe et l'âge des patients, ainsi que l'étude phylogénétique et statistique de l'ensemble des données ont permis de tirer des conclusions importantes quant aux facteurs de risque de développer la maladie. Ainsi, les souches résistantes y compris les formes MDR sont couramment associées avec la tranche d'âge 21-40 ans ($p<0,001$) ; la tuberculose extra-pulmonaire est plus répandue chez les femmes tandis que la forme pulmonaire prédomine chez les hommes ($p<0,001$) ; les lignées BOV et CAS sont particulièrement prédominantes chez les patients atteints de tuberculose extra-pulmonaire ($p<0,001$).

Conclusion

Pour conclure, notre étude nous a permis de mettre en relief les différentes facettes de la transmission de *M. tuberculosis* chez les patients Français et nés à l'étranger. Nous avons

pu mettre en évidence la particularité et la tendance des souches de tuberculose dans la région Rhône-Alpes. Notre étude souligne également la nécessité de prendre des mesures plus efficaces pour contrôler les transmissions récentes de tuberculose ainsi que les réactivations de tuberculose latente (LTBI), particulièrement chez les patients d'origine étrangère.

2.11 Epidémiologie de la tuberculose dans la région de la Vénétie en Italie (**Article 17**)

« Quand il y a plusieurs cuisiniers, la soupe est trop salée. » Proverbe Italien.

« L'apport de l'Italie à la civilisation occidentale est immense : elle est notamment le berceau de la civilisation étrusque, de la Grande-Grèce, de l'Empire romain, du Saint-Siège, des Républiques maritimes, de l'Humanisme et de la Renaissance. » (http://fr.wikipedia.org/wiki/Italie)

Présentation de l'étude

Dans cette étude, nous avons analysé 705 souches du complexe MTB isolées entre 2006 et 2009 dans la région de la Vénétie, située au Nord-ouest de l'Italie. Le spoligotypage a été réalisé pour l'intégralité des souches (n=705), alors que le typage 15-loci MIRU-VNTRs a permis de mieux identifier certaines souches appartenant à des motifs/lignées phylogénétiques particuliers (SIT1/Beijing, SIT19/EAI2-Manila, SIT26/CAS1-Delhi, SIT34/S, SIT61/LAM10-CAM, et SIT181/AFRI_1 ; n=83).

Résultats principaux

Parmi tous les patients, 410 (58.16%) étaient d'origine étrangère et 295 (41.84%) étaient Italiens. La plupart des patients étrangers provenaient de l'Europe de l'Est (32,1%), de l'Afrique du Nord (17,1%), de l'Asie du Sud (14,4%), et de l'Afrique de l'Ouest (14,2%). Les patients âgés (>70 ans) prédominaient chez les autochtones (46,4% des cas), alors que les patients plus jeunes (21-40 ans) prédominaient chez les étrangers ($p<0,0001$). Les lignées T et Haarlem étaient majoritaires dans notre étude, représentant respectivement 27,9% et 25,1% des isolats. Notre étude a révélé une proportion particulièrement grande de souches ancestrale PGG1 (17,87% des isolats). On peut noter que la lignée CAS était répartie de manière semblable chez les patients Italiens (3,73%) et les patients nés à l'étranger (5.85%). Les lignées CAS, BOV et EAI étaient plus associées aux formes de tuberculose extra-pulmonaire (**Tableau 30**). De plus, cette étude a montré pour la première fois une association du SIT 4 (Unknown) avec des souches de tuberculose extra-pulmonaire. A travers cette étude, nous avons également pu mettre en corrélation l'évolution des profils spoligotypes en fonction de divers paramètres démographiques et/ou cliniques tels que le sexe et l'origine des patients, ainsi que la résistance aux antituberculeux.

Tableau 30 : Distribution phylogénétique en fonction de la localisation de la maladie.

Lineage	Extra-pulmonary TB (%)	Extra-pulmonary or Pulmonary TB (%)	Pulmonary TB (%)	Total
AFRI	5 (33.33)	1 (6.67)	9 (60.00)	15
BOVIS	9 (56.25)	0 (0.00)	7 (43.75)	16
Beijing	5 (16.13)	1 (3.23)	25 (80.65)	31
CAS	13 (37.14)	6 (17.14)	16 (45.71)	35
Cameroon	1 (6.25)	1 (6.25)	14 (87.50)	16
EAI	11 (44.00)	3 (12.00)	11 (44.00)	25
Haarlem	27 (16.77)	13 (8.07)	121 (75.16)	161
LAM	22 (25.29)	4 (4.60)	61 (70.11)	87
Manu	1 (25.00)	0 (0.00)	3 (75.00)	4
S	7 (21.21)	1 (3.03)	25 (75.76)	33
T	35 (17.77)	4 (2.03)	158 (80.20)	197
Turkey	0 (0.00)	1 (33.33)	2 (66.67)	3
Unknown	17 (26.98)	3 (4.76)	43 (68.25)	63
Ural	0 (0.00)	0 (0.00)	16 (100.00)	16
X	0 (0.00)	1 (33.33)	2 (66.67)	3
Total	**153 (21.70)**	**39 (5.53)**	**513 (72.77)**	**705**

La discrimination de certain génotypes par le typage 15-loci MIRU-VNTRs, a permis d'identifier des clusters bien spécifiques associés à des patients ayant le plus souvent la même origine géographique (**Figure 83**).

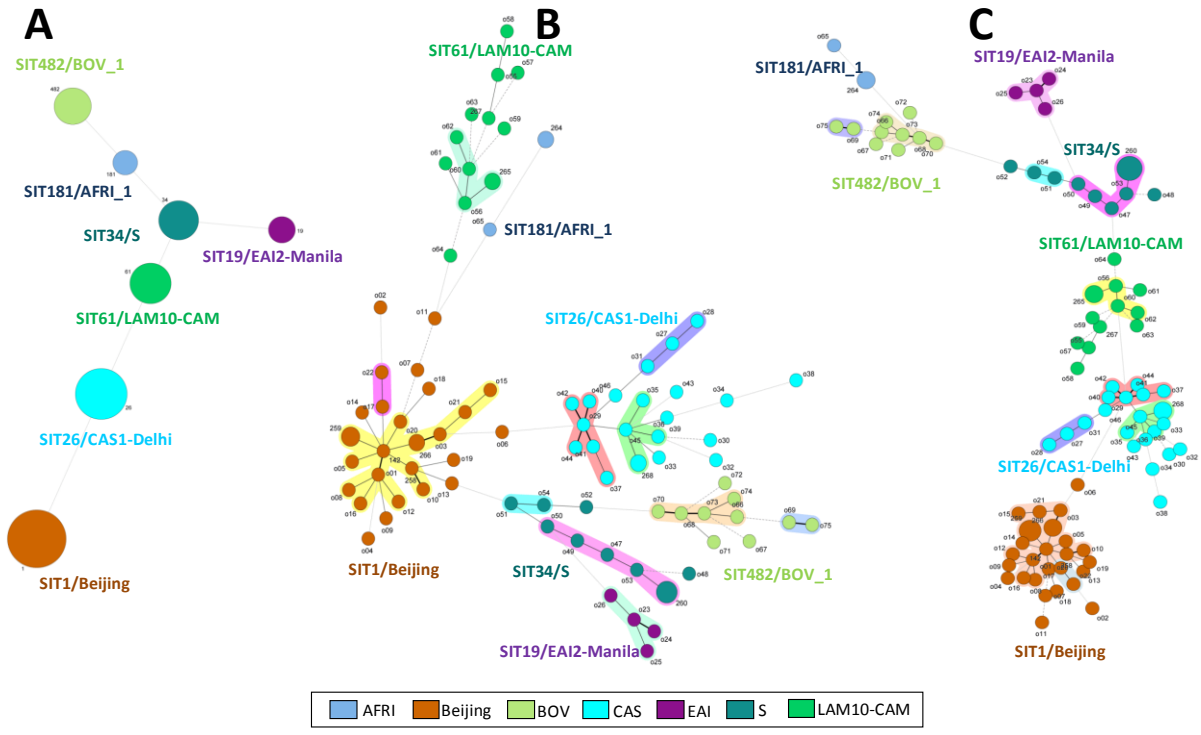


Figure 83 : MST représentant le typage des génotypes particuliers de spoligotypage (**A**), par la méthode 15-loci MIRU seule (**B**), ou par la combinaison du spoligotypage et du typage 15-loci MIRU (**C**).

Conclusion

En conclusion, notre analyse a montré une forte proportion de réactivation de la tuberculose latente chez les patients âgés Italiens. Notre étude suggère également une transmission active de la tuberculose chez les patients Italiens et d'origine étrangère. Pour finir, une surveillance et un contrôle accru de la tuberculose est plus que nécessaire pour combattre l'épidémie en Italie.

2.12 Epidémiologie de la tuberculose multi-résistante en Espagne (travail réalisé en collaboration avec Dr. Sofia Samper)

« Bon arbre, porte bonne ombre. » Proverbe Espagnole.

« L'Espagne est bordée par la mer Méditerranée au sud et à l'est, à l'exception du territoire britannique de Gibraltar et le détroit du même nom qui sépare le continent européen de l'Afrique. Au nord, les Pyrénées constituent une frontière naturelle avec la France, l'Andorre, et le golfe de Gascogne.» (http://fr.wikipedia.org/wiki/Espagne)

Présentation de l'étude

Notre étude a été menée en étroite collaboration avec le « *Spanish MDR-TB Surveillance Network* ». Nous avons analysé 366 isolats de *M. tuberculosis* multi-résistants (MDR) à travers différentes régions d'Espagne. Le spoligotypage a été utilisé sur la totalité des souches de l'étude, alors que le typage 12-loci MIRU-VNTRs a été appliqué à 106 isolats cliniques, dans le but de décrire et d'évaluer la dissémination de la tuberculose MDR en Espagne entre 1998 et 2005.

Résultats principaux

Les lignées LAM (n=94, 25.68%) et Haarlem (n=91, 24.86%) étaient les lignées les plus dominantes dans notre étude, suivies par les lignées T (n=69, 18.85%), Bovis (n=27, 7.38%), et Beijing, (n=21, 5.74%). La **Figure 84** présente une distribution phylogéographique des lignées phylogénétiques de notre étude, ainsi que les lignées de certains pays/régions avoisinants. Cette même figure montre également un MST qui permet d'identifier les relations génotypiques des isolats en fonction de régions espagnoles importantes (telles que Madrid, Galicia, Andalusia, Canaries et Valencia).

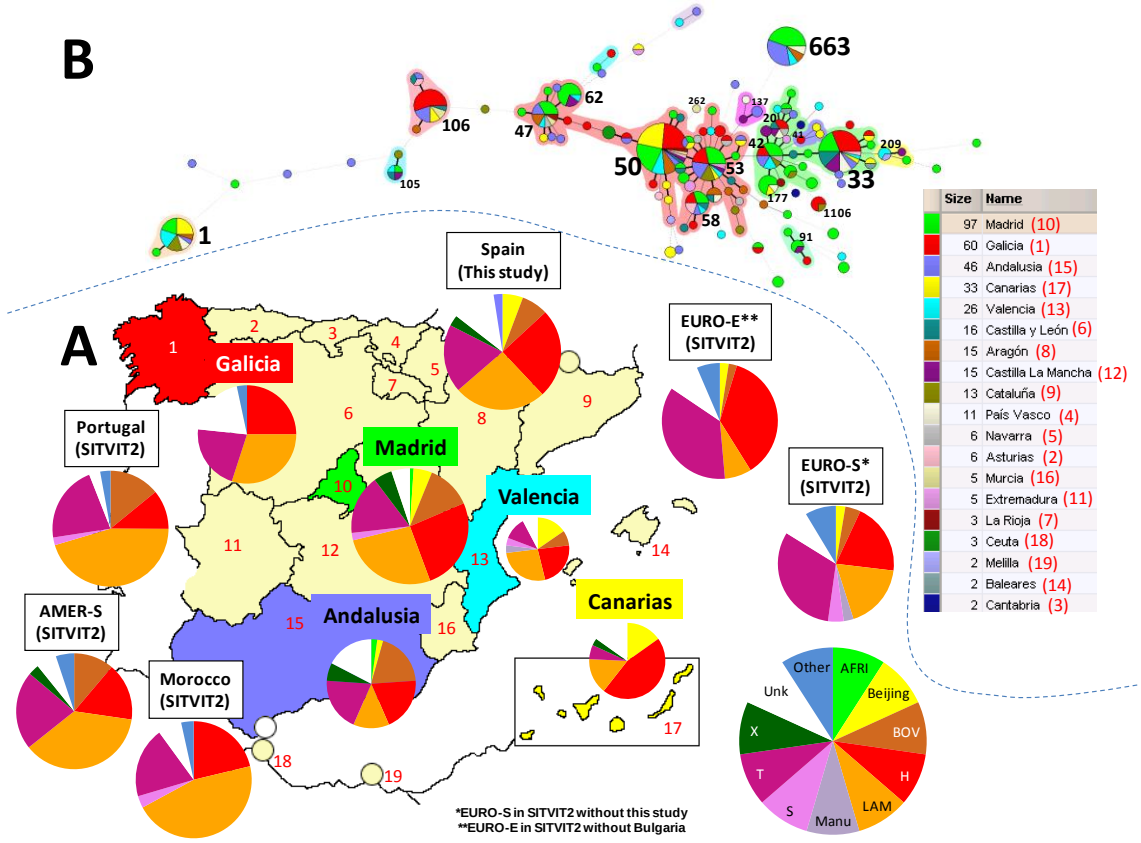


Figure 84 : Distribution phylogéographique des lignées des régions espagnoles et pays/régions avoisinants (**A**) ; et MST basé sur les spoligotypes de l'étude répartis dans les différentes régions de l'Espagne (**B**).

La présente étude souligne une distribution particulière des motifs de spoligotypage. La présence des SIT/lignées suivants est très remarquable : SIT50/H3 (n=47), SIT663/BOV_1 (n=27), SIT33/LAM3 (n=32), SIT53/T1 (n=21), SIT1/Beijing (n=20), et SIT106/Unknown (n=20) (**Figure 85**). Nous avons aussi dénoté la présence non négligeable du SIT58/T5-Madrid2 (n=10). Le SIT106 (précédemment identifié en tant que « LAM-like ») devrait faire l'objet d'investigations supplémentaires pour que nous puissions obtenir une classification plus claire. En effet, ce profil SIT106 est visiblement proche des profils appartenant à la lignée Haarlem lorsque l'on se fie au MST, mais il est aussi plus proche des spoligotype de type LAM lorsque l'on regarde le Spoligoforest.

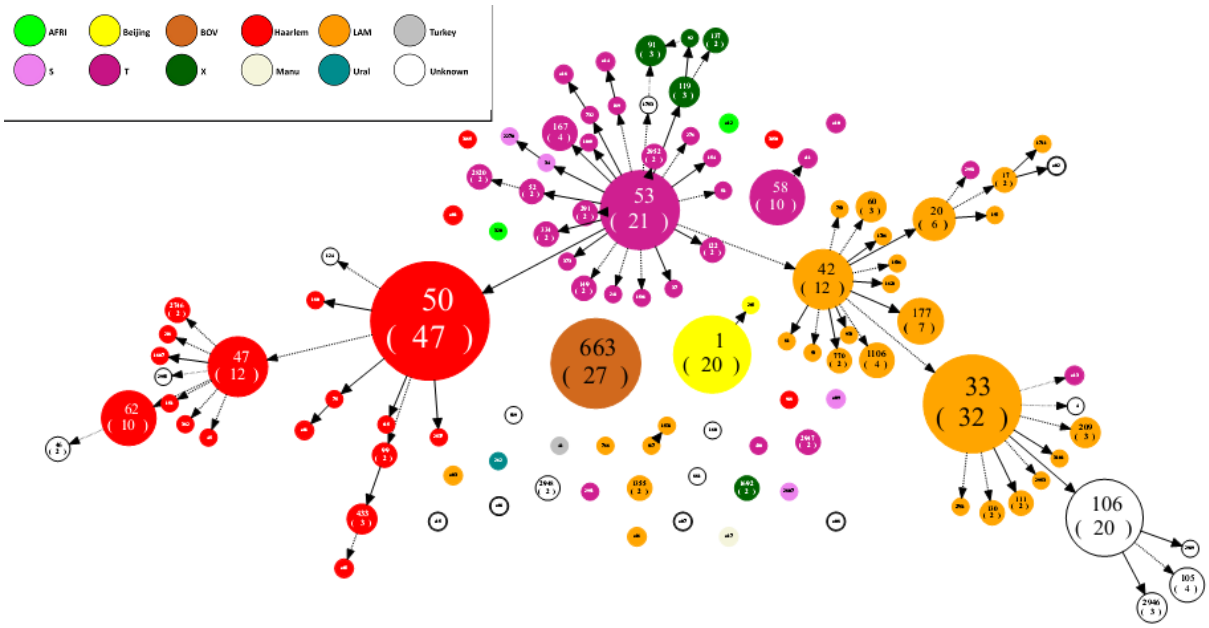


Figure 85 : Spoligoforest utilisant l'algorithme de Fruchterman-Reingold, illustrant les différentes proportions des motifs spoligotypes de notre étude.

Le typage 12-loci MIRU-VNTRs appliqué à 106 isolats, nous a permis de mieux discriminer certaines souches. Les MSTs de la **Figure 86** présentent la diversité des 106 isolats en utilisant le spoligotypage seul (**Figure 86A**), le MIRU typage seul (**Figure 86B**), et la combinaison des deux méthodes de typage MIRU et spoligotype (**Figure 86C**). Les positions des profils correspondant au SIT106/Unknown, dans les différents MSTs, témoignent bien de la complexité de la classification phylogénétique précise de ce profil qui était le plus souvent lié au profil 12-MIT 251 (224326133323) dans l'étude. Nous remarquons aussi que les isolats appartenant aux lignées T et LAM sont subdivisés et forment des groupes distincts par rapport aux autres isolats dans l'arbre basé sur les 12-loci MIRUs (**Figure 86B**). Nous notons également que les isolats appartenant à la lignée Haarlem sont mieux regroupés dans ce même arbre.

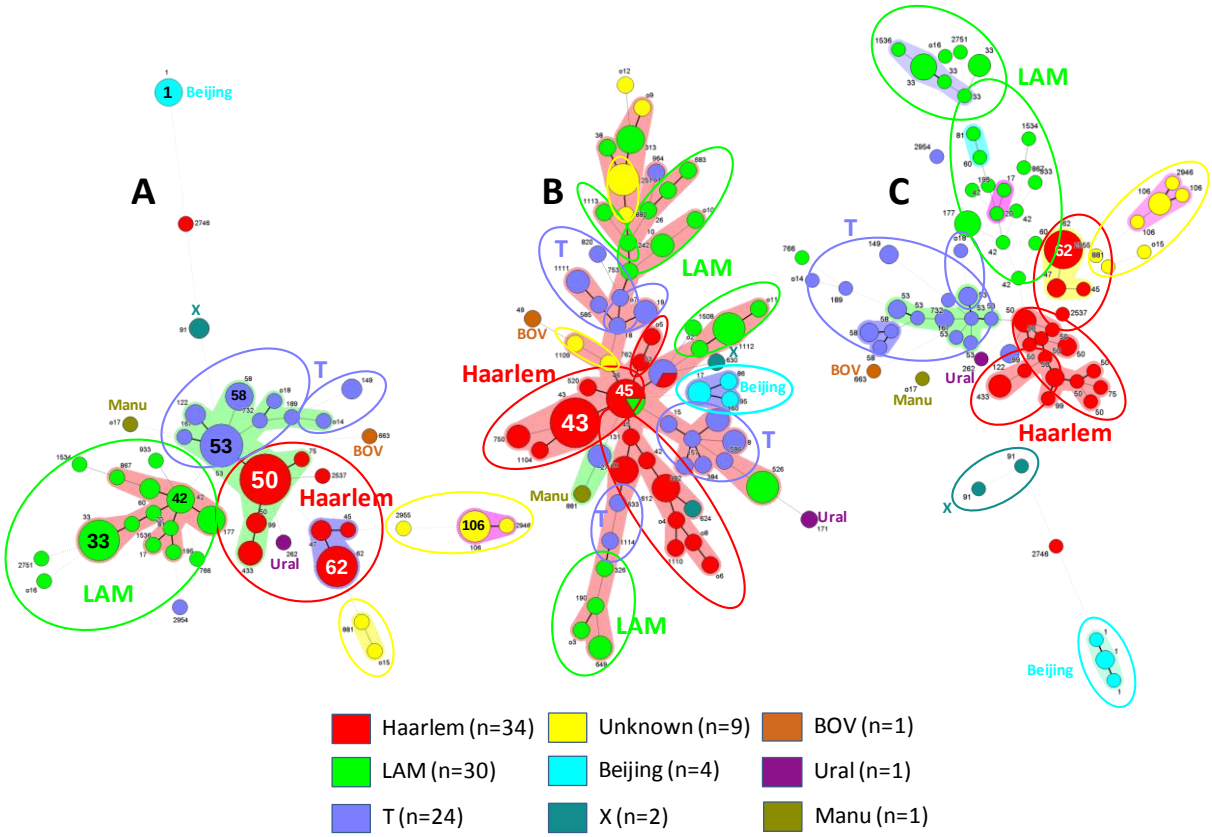


Figure 86 : MSTs basés sur les 106 isolats de l'étude contenant l'information 12-loci MIRU : (**A**) MST construit seulement à l'aide du spoligotypage ; (**B**) MST construit seulement à l'aide du typage 12-loci MIRU-VNTRs ; (**C**) MST construit en combinant spoligotypage et typage 12-loci MIRU-VNTRs.

En ce qui concerne l'ensemble de l'étude, les souches appartenant au SIT663/BOV_1 étaient toutes des souches ultra-résistantes (XDR). Aucune différence significative n'a été observée dans la répartition des génotypes en fonction du sexe des patients (**Figure 87A**). Le SIT663/BOV_1 était significativement associée aux souches XDR (**Figure 87B**). Les souches isolées à partir de patients VIH-positifs ont été notamment visibles chez les patients partageant les SIT663/BOV_1 et SIT20/LAM1 (p-value <0,009 par rapport au SIT50/H3 le plus prédominant; **Figure 87C**). On peut remarquer que le SIT106/Unknown et le SIT663/BOV_1 ont été trouvés exclusivement chez les patients Espagnols, alors que les patients partageant le SIT177/LAM9 étaient tous originaires de l'Afrique de l'Ouest.

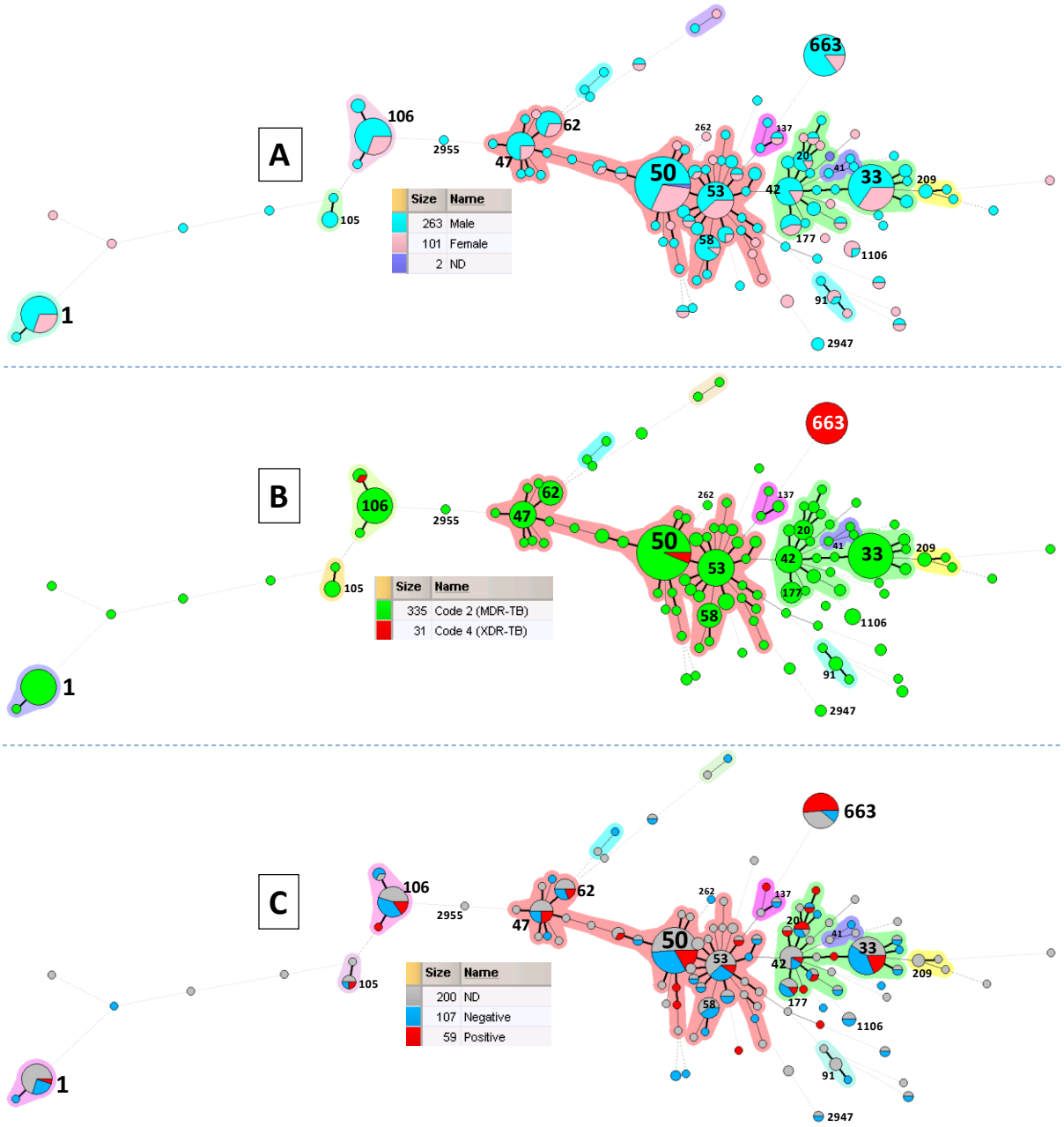


Figure 87 : MSTs basés sur le spoligotypage de l'étude, en fonction de différents paramètres : (**A**) le sexe des patients ; (**B**) la résistance aux antituberculeux (MDR vs. XDR) ; (**C**) la sérologie VIH/SIDA.

Conclusion

En résumé, notre étude a permis de faire un bref état des lieux de la tuberculose multi-résistante circulant en Espagne. Cependant, le pouvoir discriminant du typage 12-loci MIRU-VNTRs ou du spoligotypage n'est pas suffisant pour détecter de manière précise les filières épidémiologiques de transmission. Mais ces méthodes de typages permettent de décrire globalement les lignées phylogénétiques qui circulent en Espagne.

CONCLUSION ET PERSPECTIVES

Ce voyage épidémiologique, qui est le fruit d'intenses collaborations avec différentes équipes de chercheurs, nous a permis de mettre en exergue divers paysages génotypiques et phylogéographiques de MTBC à travers le monde. La base de données représente un « joyau » qui doit être pérennisé, et pris en compte lors d'analyses épidémiologiques diverses de MTBC. Malgré quelques limites concernant les méthodes de génotypage utilisées et la structure principale de l'application, cette ressource très riche nous donne un aperçu global et plus localisé des tendances d'évolution du complexe *M. tuberculosis*. Cette thèse représente une cartographie globale, mettant à contribution de nombreux outils Web et/ou bioinformatiques, favorisant la compréhension et la surveillance de la tuberculose au niveau mondial. Au cours de l'histoire des bases de données SpolDB/SITVIT, nous avons pu être témoins de la conception d'outils clés qui permettent de réaliser des études phylogénétiques variées, de décrire et identifier les isolats cliniques, et de déterminer de futurs tendances et profils génétiques impliqués dans la transmission mondiale de la tuberculose. Un algorithme permettant de mettre en évidence les génotypes rares ou confinés dans un nombre précis de régions géographiques, a été développé et est intégré dans la nouvelle interface de la base de données SITVIT2.

Le développement des bases de données de génotypage de bacilles tuberculeux, ainsi que les nombreuses études épidémiologiques sur lesquelles sont appliquées ces bases, permettent de tirer des conclusions importantes sur certains aspects de l'épidémiologie mondiale de la TB, et peuvent contribuer à diverses stratégies de lutte antituberculeuse. La mise au point des bases de données SITVIT et le développement des outils Web associés apportent une vue globale sur la situation mondiale de la tuberculose, en mettant en évidence différents clivages par rapport aux zones géographiques, à la virulence des génotypes, aux profils de résistance aux antituberculeux, aux données démographiques des patients (notamment le sexe, l'âge, et l'origine), ou encore les coinfections TB/VIH. Une meilleure caractérisation génotypique des isolats circulants de MTBC et les diverses études épidémiologiques qui en découlent ont permis non seulement d'enrichir les bases de données mais ont également contribué à apporter des éléments pertinents pour le renforcement des programmes de lutte antituberculeuse à l'échelle des pays et des régions étudiés. De plus, ces données permettent d'avoir une meilleure compréhension de l'épidémie ainsi qu'un aperçu des aspects évolutifs concernant le MTBC nécessaires pour proposer une meilleure prise en charge des patients.

Les études épidémiologiques apportent des réponses adaptées et ciblées en fonction des pays concernés. Ainsi, les données de génotypage accompagnées des données démographiques et épidémiologiques (résistance aux antituberculeux, l'âge, le sexe, le statut VIH, maladie pulmonaire vs. extra-pulmonaire, etc..), permettent de tirer des conclusions significatives sur des associations probables entre les génotypes et les différents paramètres. Nous pouvons citer notre étude sur la base de données des isolats MDR d'Espagne ainsi que d'autres études réalisées en Amérique latine (Mexique, Colombie, Pérou, Brésil), en Afrique (Guinée Bissau, Ethiopie) et en Asie (Irak, Oman, Inde, Malaisie, Chine) qui apportent des informations pertinentes non seulement pour les programmes de lutte antituberculeuses localement, mais aussi sur la diversité génétique et la géo-spécificité des isolats MTBC. On peut citer pêle-mêle: (i) l'exemple de la famille « Guinée Bissau » pour ce pays, et la description d'une lignée provisoirement désignée « EAI1-MYS » spécifique à la Malaisie concernant la géo-spécificité des isolats, (ii) la transmission active de la tuberculose MDR

chez les jeunes enfants à New Delhi, Inde ; et (iii) finalement différentes études menées en Europe (Bulgarie, Finlande, Suède, France, Italie, Espagne) qui mettent en évidence à la fois la réactivation de la maladie chez les patients âgés autochtones versus les cas importés de tuberculose active le plus souvent retrouvés chez les patients d'origine étrangère. De plus, lorsque l'on prend l'exemple des études concernant la résistance aux antituberculeux et les coinfections TB/VIH, on peut déduire que certaines lignées phylogénétiques sont d'avantage retrouvées chez les patients VIH+ (par exemple, la lignée LAM en Amérique du Sud et la lignée « Cameroon » en Afrique de l'Ouest), ou majoritaires parmi les souches résistantes aux antituberculeux (par exemple, la lignée EAI en Afrique du Nord) (**Figures 32 et 34**).

L'étude des lignées phylogénétiques est importante pour avoir une vue d'ensemble de l'épidémie de la TB qui sévit actuellement et s'inscrit comme un sujet prioritaire pour le développement de nouveaux outils permettant de mieux comprendre, contrôler et appréhender la tuberculose (**López et coll., 2003** ; **Ford et coll., 2013**). En effet, les lignées phylogénétiques spécifiques de MTBC peuvent jouer un rôle majeur dans la virulence, la résistance aux antibiotiques, et la susceptibilité à l'infection par rapport à l'âge ou au sexe des patients. De nombreuses études sont actuellement en cours pour tirer tout le bénéfice de cette nouvelle compréhension de l'épidémiologie globale de MTBC. On peut citer entre autres la conception de kits de diagnostic puissants tels que le « Multiplex Ligation-dependent Probe Amplification » (MLPA) qui permet à la fois la détection de marqueurs de résistance aux antituberculeux et la discrimination des souches en lignées génotypiques distinctes (**Bergval et coll., 2012**) ; le développement de nouveaux outils WEB tels que PolyTB (**Coll et coll., 2014a**) et GMTV database (**Chernyaeva et coll., 2014**) ; ainsi que le nouveau « SNP barcode » pour le typage de MTBC (**Coll et coll., 2014b**). Dans les années à venir, ces outils seront indispensables pour étudier les associations entre les paramètres bactériologiques, démographiques, cliniques et épidémiologiques versus les variations génomiques observées chez MTBC (y compris SNPs).

Malgré le biais possible de recrutement des isolats pour différentes régions géographiques, la version la plus récente des bases de données (n=111635 isolats de 169 pays) donne un aperçu intéressant de la phylogénie mondiale de MTBC, particulièrement la façon dont les souches s'organisent, évoluent, et se distinguent des unes et des autres dans la population mondiale – basée sur le spoligotypage (**Figure 38**, **Tableau 11**), ou MIRU-VNTRs (**Figure 39**, **Tableau 10**, **Annexes 1 et 2**).

Au cours de ce mémoire, nous avons aussi abordé le thème de la migration des populations. Les données disponibles dans la collection SITVIT mettent bien en évidence les potentiels flux migratoires des souches bactériennes provenant généralement de patients originaires de pays en voie de développement qui se dirigent le plus souvent vers les pays industrialisés. Ainsi, la cartographie schématisant les flux de migration de tuberculose résistante donne une idée sur l'origine géographique potentielle des souches résistantes (**Figure 35**), soulignant les filières de transmission et de la propagation de la tuberculose y compris les cas d'antibiorésistance. De plus, les études spatio-temporelles nous apportent des indices significatifs sur l'évolution mondiale de bacilles tuberculeux dont on peut voir un exemple sur: http://www.pasteur-guadeloupe.fr:8081/SITVIT2/stata.jsp. Une telle approche facilite notamment le développement des modèles prédictifs en ciblant les pays/régions qui représentent les « hotspots » en se référant à l'évolution temporelle des bactéries résistantes et aux facteurs socio-économiques associés. Ainsi, l'étude menée sur la lignée Beijing reconnue pour sa forte association avec l'émergence de tuberculose MDR et XDR (**Article 3**) constitue un exemple intéressant car elle souligne les clivages importants concernant la résistance aux

antituberculeux des isolats ainsi que l'âge, le sexe et l'origine géographique des patients; permettant de faire des prédictions quant aux risques de propagation de la maladie. En résumé, à travers les données récoltées et les outils développés dans cette thèse, nous avons pu :

(i) mettre au point et améliorer les bases de données de génotypage de MTBC,
(ii) réaliser une étude phylogénétique globale de MTBC en se basant sur les profils de spoligotypage et de typage MIRU-VNTRs,
(iii) mettre en évidence des potentielles associations entre certaines lignées phylogénétiques et divers caractéristiques géographiques, démographiques, cliniques et épidémiologiques,
(iv) estimer certaines filières probables de transmission de la tuberculose, notamment pour les formes résistantes de la maladie,
(v) identifier des profils spécifiques à des régions géographiques précises,
(vi) mettre en relief certains cas de transmissions actives et de réactivations de la maladie (tuberculose latente)
(vii) contribuer aux programmes de surveillance et de lutte antituberculeuse dans diverses régions du monde, favorisant ainsi une meilleure gestion des patients,
(viii) contribuer à une meilleure compréhension biologique et épidémiologique de la tuberculose.

Pour conclure, ce travail de thèse s'est beaucoup appuyé sur l'utilité des bases de données et outils web dans la surveillance et le contrôle de la tuberculose au niveau régionale et mondial. Les différentes études épidémiologiques ont été menées dans des pays variés à faible, moyenne, ou haute incidence de tuberculose. Grâce à la cartographie mondiale des génotypes de MTBC recueillis à travers de nombreuses régions du monde, nous souhaitons favoriser la sélection de souches bactériennes représentatives de la diversité mondiale de MTBC ; un tel panel serait très important pour l'élaboration et la conception de kits de diagnostic, de nouveaux vaccins, et le développement des nouveaux antituberculeux.

PERENNISATION ET ENRICHISSEMENT DES BASES DE DONNEES SITVIT

Cette base de données sera bientôt enrichie de nouveaux paramètres et de nouvelles méthodes d'analyse telle que le « Whole Genome Sequencing » (WGS). SITVIT pourrait aussi inclure d'autres loci MIRU-VNTRs en dehors des 24-loci classiques, à l'exemple du consensus récent concernant les loci hypervariables pour la famille « Beijing » (**Allix-Béguec et coll., 2014**). D'autres efforts devront être réalisés pour maintenir cette base et la rendre plus accessible et autonome. De nouveaux outils logiciels et matériels pourraient également enrichir et rendre plus performantes nos bases de données.

Dans le but de mieux cerner l'évolution des isolats cliniques de MTBC, nous prévoyons de développer de nouveaux outils de visualisation et de classification qui permettront d'établir des phylogénies plus robustes à partir des génotypes obtenus par spoligotypage et/ou typage MIRU-VNTRs.

Afin de mieux comprendre et maitriser l'épidémie, nous nous sommes également penché sur des méthodes de classification basées sur les réseaux bayésiens et une version optimisée de l'algorithme de k-moyennes (ou « K-means » en anglais) :

(i) Une application Web (en développement) provisoirement nommée « TbBayesNet », initiée par Dr. Véronique Hill, permettra de prédire en fonction des motifs de spoligotypage ou de typage 12-loci MIRU-VNTRs, les lignées phylogénétiques probables, les régions probables où le motif en question est visible, ainsi que les loci (MIRU) ou espaceurs (Spoligotypes) manquants, en utilisant des probabilités et des scores statistiques.

(ii) Un programme informatique basé sur une optimisation des k-moyennes sera mis en œuvre pour mieux détecter et identifier les souches de *M. tuberculosis* auxquelles nous avions (dans la version SITVITWEB) attribué des désignations de lignées « ambigües » (**Tableau 31**). Cette optimisation des k-moyennes fera partie d'un travail à venir.

Tableau 31 : Sous-Lignées ambigües dans SITVITWEB, et leurs interprétations provisoires dans SITVIT2.

Sous lignée génotypique ambigüe	Nouvelle notation de la sous lignée
AMBIGOUS :LAM1 LAM4	LAM
AMBIGOUS :EAI6-BGD1 EAI1-SOM	EAI
AMBIGOUS :T3 T2	T
AMBIGOUS :T4 T2	T
AMBIGOUS :LAM5 LAM4	LAM
AMBIGOUS :T4 T3	T
AMBIGOUS :T2-uganda H37Rv	T
AMBIGOUS :T4 T5	T
AMBIGOUS :T3 T5	T
AMBIGOUS :T4 T3 T2	T
AMBIGOUS :X1 T5	X1
AMBIGOUS :T2 T5	T
AMBIGOUS :T3 X1	X1

Nous avons également débuté un projet nommé « MIRUSkyScrapper », avec une équipe de chercheurs du Rensselaer Polytechnic Institute (RPI) aux USA, pour mieux visualiser les différents loci MIRU-VNTRs en fonction des lignées phylogénétiques ou des régions géographiques.

Dans un futur proche, nous réaliserons une étude concernant la cartographie mondiale des isolats appartenant aux lignées EAI et CAS afin de mettre en évidence plusieurs caractéristiques de l'évolution de ces profils génétiques et tenter de déchiffrer les facteurs probablement liés à la dispersion de ces lignées. Les lignées CAS et EAI sont généralement prédominante dans les régions bordant l'océan indien et le pacifique. La **Figure 88** illustre quelques échanges commerciaux (« route de la soie ») qui ont eu lieu il y a des milliers d'années. Cette figure suggère que la transmission de la tuberculose (qui serait une maladie aussi vieille que l'humanité) est étroitement liée à la migration humaine (comme le montre déjà plusieurs études). Nous pouvons donc analyser en profondeur l'histoire de certaines migrations et commerces humains pour mieux comprendre la diversité de la tuberculose dans des régions spécifiques, et pendant des périodes précises.

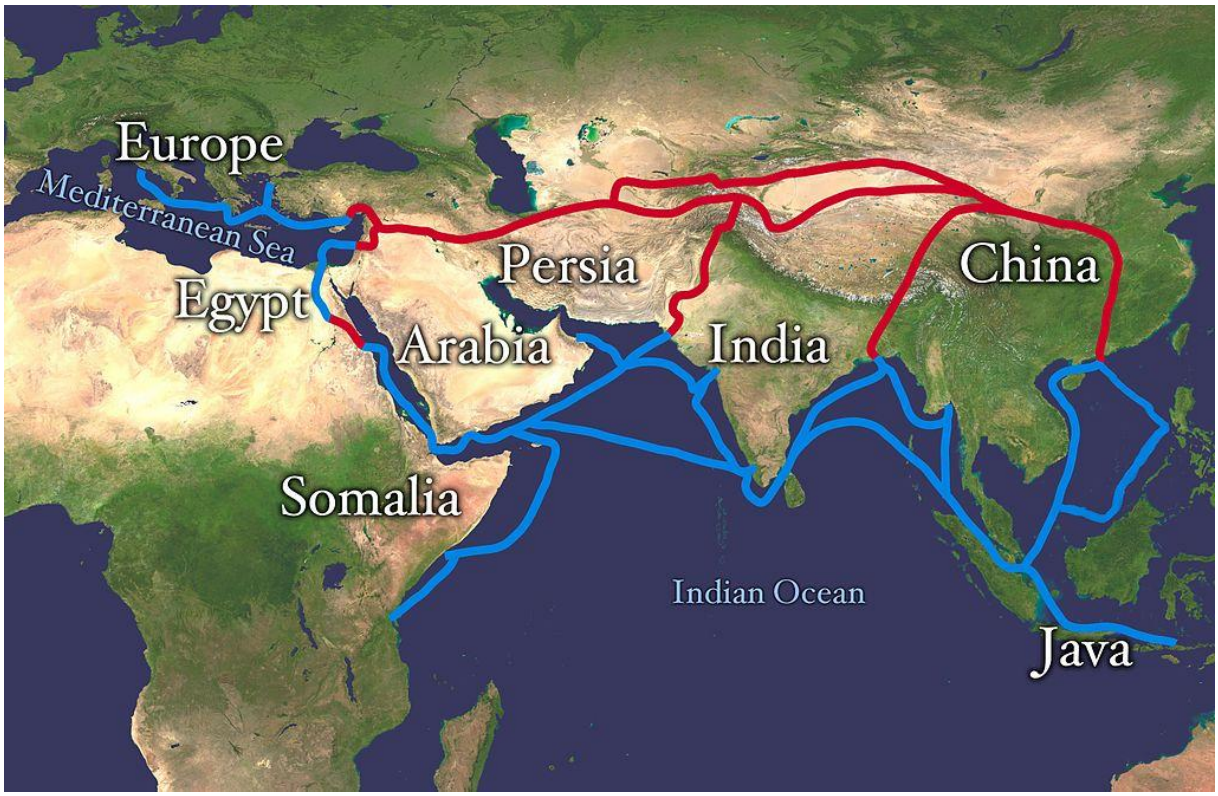


Figure 88 : Représentation géographique de « la route de la soie » (http://en.wikipedia.org/wiki/Silk_Road).

Avec l'avancée des nouvelles technologies de l'information et de la communication, et la progression croissante des données biologiques, de nouveaux challenges et impératifs ont vu le jour. Les manières d'interpréter, de stocker, d'analyser les données sont nombreuses et variées ; mais les façons les plus optimales de réaliser ces différentes tâches sont encore à développer ou inventer. Bien que différents outils bioinformatiques soient déjà disponibles pour réaliser un nombre conséquent d'action, il existe encore des besoins considérables correspondant à différents cas de figure et différents domaines ou sous-domaines d'étude concernant l'épidémiologie moléculaire de la tuberculose. En même temps, la nécessité de centraliser les divers outils et l'accès aux données, est un sujet d'actualité, qui faciliterait le traitement et la gestion des informations. Une nouvelle ère approche, et au moment même où nous entendons parler de « Big Data » ou de « Cloud », de futurs concepts plus complexes sont certainement en développement. Le futur des bases de données SITVIT est une thématique qui nous tient à cœur. Nous envisageons donc un système de gestion et de maintenance plus autonome pour les nouvelles versions à venir. L'interface « SITVITAdmin » facilite d'ores et déjà la gestion de la base de données, ainsi que l'enregistrement et le traitement des informations épidémiologiques et génotypiques des membres du complexe *M. tuberculosis*.

BIBLIOGRAPHIE

« Gazette hebdomadaire de médecine et de chirurgie. - 1887. - série 2, tome 24 » [archive], sur web2.bium.univ-paris5.fr (consulté le 15 juillet 2010).

Abadia E, Zhang J, dos Vultos T, Ritacco V, Kremer K, Aktas E, Matsumoto T, Refregier G, van Soolingen D, Gicquel B, Sola C. (2010). Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infect Genet Evol. 2010; 107: 1066–1074.

Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, Gey van Pittius NC. Novel *Mycobacterium tuberculosis* complex pathogen, M. mungi. Emerg Infect Dis. 2010 Aug;16(8):1296-9.

Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, Drucker E, Bloom BR. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med. 1994 Jun 16;330(24):1710- 6.

Alland D, Lacher DW, Hazbón MH, Motiwala AS, Qi W, Fleischmann RD, Whittam TS. Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. J Clin Microbiol. 2007; 45(1):39-46.

Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol. 2008; 46: 2692-2699.

Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, Campos-Herrero I, Mokrousov I, Niemann S, Kontsevaya I, Rastogi N, Samper S, Sng LH, Warren RM, Supply P. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. J Clin Microbiol. 2014; 52(1):164-72.

Alves G, Yu Y-K (2014) Accuracy Evaluation of the Unified P-Value from Combining Correlated P-Values. PLoS ONE 9(3): e91225. doi:10.1371/journal.pone.0091225

Aminian M, Couvin D, Shabbeer A, Hadley K, Vandenberg S, Rastogi N, Bennett KP. Predicting *Mycobacterium tuberculosis* complex clades using knowledge-based Bayesian networks. Biomed Res Int. 2014;2014:398484.

Aranaz A, Liébana E, Gómez- Mampaso E, Galán JC, Cousins D, Ortega A, Blázquez J, Baquero F, Mateos A, Súarez G, Domínguez L. Mycobacterium tuberculosis subsp. caprae subsp. nov.: a taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain. Int J Syst Bacteriol. 1999; 49 Pt 3:1263 - 73.

Attwood T.K., Gisel A., Eriksson N-E. and Bongcam-Rudloff E., « Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective » [archive], sur Bioinformatics - Trends and Methodologies, InTech, 2011 (http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/concepts-historical-milestones-and-the-central-place-of-bioinformatics-in-modern-biology-a-european-).

Avise JC, Arnold J, Ball Jr. RM, Bermingham, E, Lamb T, Neigel, JE, Reeb CA, and Saunders J. 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. Ann. Rev. Ecol. System 18: 489-522.

Avise JC, Walker D, Johns GC. Speciation durations and Pleistocene effects on vertebrate phylogeography. Proc Biol Sci. 1998 Sep 22;265(1407):1707-12.

Barry CE 3rd, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ, Young D. 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nat Rev Microbiol. 7: 845-855.

Bergval I, Sengstake S, Brankova N, Levterova V, Abadía E, Tadumaze N, Bablishvili N, Akhalaia M, Tuin K, Schuitema A, Panaiotov S, Bachiyska E, Kantardjiev T, de Zwaan R, Schürch A, van Soolingen D, van 't Hoog A, Cobelens F, Aspindzelashvili R, Sola C, Klatser P, Anthony R. Combined species identification, genotyping, and drug resistance detection of *Mycobacterium tuberculosis* cultures by MLPA on a bead-based array. PLoS One. 2012; 7(8):e43240.

Bhanu NV, van Soolingen D, van Embden JD, Dar L, Pandey RM, Seth P. Predominace of a novel *Mycobacterium tuberculosis* genotype in the Delhi region of India. Tuberculosis (Edinb). 2002;82(2-3):105-12.

Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol 2002 Jan; 10:45-52.

Blouin Y, Cazajous G, Dehan C, Soler C, Vong R, Hassan MO, Hauck Y, Boulais C, Andriamanantena D, Martinaud C, Martin É, Pourcel C, Vergnaud G. Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, Djibouti. Emerg Infect Dis. 2014 Jan;20(1):21-8.

Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci USA. 99: 3684-3689.

Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C. 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6: 23.

Castets M, Boisvert H, Grumbach F, Brunel M, and Rist N. Tuberculosis bacilli of the African type: preliminary note. Rev. Tuberc. Pneumol. 1968, 32:179- 184.

CDC. 2010. Launch of TB Genotyping Information Management System (TB GIMS). Morbidity and Mortality Weekly Report (MMWR) March 19, 59(10); 300.

Charles Darwin, 1859. De l'origine des espèce. Titre anglais original : "The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life".

Chartier F, Chartier C, Thorel MF, Crespeau F. [A new case of *Mycobacterium bovis* pulmonary tuberculosis in the dromedary (Camelus dromedarius) in Mauritania]. Rev Elev Med Vet Pays Trop. 1991;44(1):43- 7.

Chernyaeva EN, Shulgina MV, Rotkevich MS, Dobrynin PV, Simonov SA, Shitikov EA, Ischenko DS, Karpova IY, Kostryukova ES, Ilina EN, Govorun VM, Zhuravlev VY, Manicheva OA, Yablonsky PK, Isaeva YD, Nosova EY, Mokrousov IV, Vyazovaya AA, Narvskaya OV, Lapidus AL, O'Brien SJ. Genome-wide *Mycobacterium tuberculosis* variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. BMC Genomics. 2014 Apr 25;15:308.

Chiang CY, Lee JJ, Yu MC, Bai KJ, Lin TP, Luh KT. Tuberculosis-related deaths without treatment. Int J Tuberc Lung Dis. 2009 Dec;13(12):1563-5.

Ciccarelli FD (2006). "Toward automatic reconstruction of a highly resolved tree of life."; Science 311(5765): 1283-7.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature; 396(6707):190.

Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, Viveiros M, Portugal I, Drobniewski F, Gagneux S, Glynn JR, Pain A, Parkhill J, McNerney R, Martin N, Clark TG. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis (Edinb). 2014a; 94(3):346-54.

Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun. 2014b; 5:4812.

Comas I, Homolka S, Niemann S, Gagneux S. 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One. 4: e7815.

Cousins DV, Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, Duignan P, Murray A, Dupont C, Ahmed N, Collins DM, Butler WR, Dawson D, Rodríguez D, Loureiro J, Romano MI, Alito A, Zumarraga M, Bernardelli A. Tuberculosis in seals caused by a novel member of the *Mycobacterium tuberculosis* complex: Mycobacterium pinnipedii sp. nov. Int J Syst Evol Microbiol. 2003 Sep;53(Pt 5):1305- 14.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14: 1188-1190.

David H, Lévy- Frébault V, Thorel MF (1989). Méthodes de laboratoire pour mycobactériologie clinique. p. 1 - 87. Institut Pasteur, Paris, France.

Dawson KL, Bell A, Kawakami RP, Coley K, Yates G, Collins DM. Transmission of Mycobacterium orygis (*M. tuberculosis* complex species) from a tuberculosis patient to a dairy cow in New Zealand. J Clin Microbiol. 2012 Sep;50(9):3136-8.

Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N (2012). SITVITWEB - a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol; 12(4):755-66.

DM Chickering, D Heckerman, and C Meek. Large sample learning of Bayesian networks is NP-hard. The Journal of Machine Learning Research, 5 :1287 1330, 2004.

Donoghue HD, Spigelman M, Greenblatt CL, Lev-Maor G, Bar-Gal GK, Matheson C, Vernon K, Nerlich AG, Zink AR. Tuberculosis: from prehistory to Robert Koch, as revealed by ancient DNA. Lancet Infect Dis. 2004 Sep;4(9):584-92.

Driscoll JR, Bifani PJ, Mathema B, McGarry MA, Zickas GM, Kreiswirth BN, Taber HW. 2002. Spoligologos: a bioinformatic approach to displaying and analyzing *Mycobacterium tuberculosis* data. Emerg Infect Dis. 8(11): 1306-9.

Eichbaum Q., Rubin E.J. : Tuberculosis. Advances in labatory diagnosis and drug susceptibility testing. Am. J. Clin. Pathol., 2002 ; 118 : S3-17.

Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. 2002. Graphviz – Open Source Graph Drawing Tools. In: Mutzel P, Jünger M, Leipert S (Editors), Heidelberg: Springer-Verlag Berlin. pp. 483-484.

Fabre M, Koeck JL, Le Fleche P, Simon F, Herve V, Vergnaud G, et al. High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of

"Mycobacterium canettii" strains indicates that the M. tuberculosis complex is a recently emerged clone of "M. canettii". J Clin Microbiol. 2004; 42:3248–55.

Fabre, Michel; Hauck, Yolande; Soler, Charles; Koeck, Jean-Louis; Van Ingen, Jakko; Van Soolingen, Dick; Vergnaud, Gilles; Pourcel, Christine (2010). "Molecular characteristics of *'Mycobacterium canettii*' the smooth *Mycobacterium tuberculosis* bacilli". Infection, Genetics and Evolution 10 (8): 1165–73.

Fallico L, Couvin D, Peracchi M, Pascarella M, Franchin E, Lavezzo E, Rassu M, Manganelli R, Rastogi N, Palù G. Four year longitudinal study of *Mycobacterium tuberculosis* complex isolates in a region of North-Eastern Italy. Infect Genet Evol. 2014; 26:58-64.

Farmer PE, Nizeye B, Stulac S, Keshavjee S. Structural violence and clinical medicine. PLoS Med. 2006 Oct;3(10):e449.

Filliol I, Driscoll JR, Van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Anh DD, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Källenius G, Kassa-Kelembho E, Koivula T, Ly HM, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, De Waard JH, Sola C, Rastogi N. 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg Infect Dis. 2002 Nov; 8(11): 1347-9.

Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Kassa-Kelembho E, Ho ML, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, de Waard JH, Sola C, Rastogi N. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J Clin Microbiol. 2003 May;41(5):1963-70.

Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, Alland D. 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol. 2006; 188: 759-772.

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet. 2013; 45(7):784-90.

Francis, J. 1958. Tuberculosis in animals and man. Cassell and Company, London, UK, 357 pp.

Frieden T.R., Sterling R., Munsiff S., Watt C.J., Dye C. : Tuberculosis. The Lancet, 2003 ; 362 : 887-899.

Frothingham R, Meeker-O'Connell WA. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 1998; 144:1189-96.

Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA. 103(8): 2869-73.

García de Viedma D, Mokrousov I, Rastogi N. 2011. Innovations in the molecular epidemiology of tuberculosis. Enferm Infecc Microbiol Clin. 29 (Suppl 1): 8-13.

Gibbons A. American Association of Physical Anthropologists meeting. Tuberculosis jumped from humans to cows, not vice versa. Science. 2008 May 2;320(5876):608.

Gokce G., Kilicarslan H., Ayan S., Tas F., Akar R., Kaya K., Gultekin E.Y : Genito urinary tuberculosis : a review of 174 cases. Scand. J. Infect. Dis., 2002 ; 34 : 338-340.

Gradmann C. (2001) Robert Koch and the pressures of scientific research: tuberculosis and tuberculin. Med Hist. Jan 2001; 45(1): 1–32.

Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, Riccardi F, Couvin D, Hill V, Rastogi N, Koivula T, Källenius G. 2011. The Guinea-Bissau family of *Mycobacterium tuberculosis* complex revisited. PLoS One. 6(4): e18601.

Gudan A, Artukovi c B, Cvetnic Z, Spici c S, Beck A, Hohsteter M, Nagli c T, Bata I, Grabarevic Z. Disseminated tuberculosis in hyrax (Procavia capensis) caused by *Mycobacterium africanum*. J Zoo Wildl Med. 2008 Sep;39(3):386- 91.

Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J Infect Dis. 2006; 193: 121-128.

Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog. 2005;1:e5.

Guy Grenier. 100 ans de médecine francophone: histoire de l'Association des médecins de langue française du Canada. Éditions MultiMondes, 2002. Lien : http://multim.com//titre/?ID=129.

Harshey RM, Ramakrishnan T (1977). Rate of ribonucleic acid chain growth in Mycobacterium tuberculosis H37Rv. J Bacteriol; 129(2):616-22.

Hart PD, Sutherland I. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. Br Med J. 1977 Jul 30;2(6082):293- 5.

Helal ZH, Ashour MS, Eissa SA, Abd-Elatef G, Zozio T, Babapoor S, Rastogi N, Khan MI. Unexpectedly high proportion of ancestral Manu genotype *Mycobacterium tuberculosis* strains cultured from tuberculosis patients in Egypt. J Clin Microbiol. 2009 Sep;47(9):2794-801.

Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect Immun. 59: 2695-2705.

Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol. 2008, 6(12): e311.

Hesper B et Hogeweg P, « Bioinformatica: een werkconcept », Kameleon, vol. 1, no 6,? 1970, p. 28–29

Hill V, Zozio T, Sadikalay S, Viegas S, Streit E, Kallenius G, Rastogi N. MLVA based classification of *Mycobacterium tuberculosis* complex lineages for a robust phylogeographic snapshot of its worldwide molecular diversity. PLoS One. 2012;7(9):e41991.

Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc Natl Acad Sci USA. 2004; 101: 4871-4876.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. (2008) Big data: The future of biocuration. Nature. 455(7209): 47-50.

Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 1988;26:2465-2466.

Iwamoto T, Grandjean L, Arikawa K, Nakanishi N, Caviedes L, Coronel J, Sheen P, Wada T, Taype CA, Shaw MA, Moore DA, Gilman RH. Genetic diversity and transmission characteristics of Beijing family strains of *Mycobacterium tuberculosis* in Peru. PLoS One 2012;7(11):e49651.

Iwamoto T, Yoshida S, Suzuki K, Tomita M, Fujiyama R, Tanaka N, Kawakami Y, Ito M. Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on *Mycobacterium tuberculosis* strains predominated by the Beijing family. FEMS Microbiol Lett. 2007 May;270(1):67-74.

Jagielski T, van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. 2014. Current Methods in the Molecular Typing of *Mycobacterium tuberculosis* and Other Mycobacteria. Biomed Res Int. 2014 (Article ID 645802): 1-21.

Jean-Baptiste de Lamarck, 1809. Philosophie zoologique.

Joshi KR, Dhiman H, Scaria V. (2013) tbvar: a comprehensive genome variation resource for *Mycobacterium tuberculosis*. Database Vol. 2013: article ID bat083.

Källenius G, Koivula T, Ghebremichael S, Hoffner SE, Norberg R, Svensson E, Dias F, Marklund BI, Svenson SB. Evolution and clonal traits of *Mycobacterium tuberculosis* complex in Guinea-Bissau. J Clin Microbiol. 1999 Dec;37(12):3872-8.

Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol. 35: 907-914.

Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11: 547-554.

Kaufmann SH, Lange C, Rao M, Balaji KN, Lotze M, Schito M, Zumla AI, Maeurer M. Progress in tuberculosis vaccine development and host-directed therapies--a state of the art review. Lancet Respir Med. 2014 Apr;2(4):301-20.

Keshavjee S, Gelmanova IY, Pasechnikov AD, Mishustin SP, Andreev YG, Yedilbayev A, Furin JJ, Mukherjee JS, Rich ML, Nardell EA, Farmer PE, Kim JY, Shin SS. Treating multidrug-resistant tuberculosis in Tomsk, Russia: developing programs that address the linkage between poverty and disease. Ann N Y Acad Sci. 2008;1136:1-11.

Kidd DM et Ritchie MG, « Phylogeographic information systems : putting the geography into phylogeography », Journal of biogeography, vol. 33, no 11,? 2006, p. 1851-1865.

Kirschner P, Meier A, Böttger EC. Genotypic identification and detection of Mycobacteria: facing novel and uncultured pathogens. In: Persing DH, Tenover F, White TJ, Smith TF (ed): Diagnostic molecular microbiology: principles and applications. American Society for Microbiology, Washington DC, 1993, p. 173-190.

Kisa O, Tarhan G, Gunal S, Albay A, Durmaz R, Saribas Z, Zozio T, Alp A, Ceyhan I, Tombak A, Rastogi N. (2012). Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey. PLoS One. 2012; 7(1): e30331.

Koeck, J.-L.; Fabre, M.; Simon, F.; Daffé, M.; Garnotel, É.; Matan, A. B.; Gérôme, P.; Bernatas, J.-J. et al. (2011). "Clinical characteristics of the smooth tubercle bacilli *'Mycobacterium canettii*' infection suggest the existence of an environmental reservoir". Clinical Microbiology and Infection 17 (7): 1013–9.

Konomi N, Lebwohl E, Mowbray K, Tattersall I, Zhang D. Detection of mycobacterial DNA in Andean mummies. J Clin Microbiol. 2002 Dec;40(12):4738-40.

Koro Koro F, Kamdem Simo Y, Piam FF, Noeske J, Gutierrez C, Kuaban C, Eyangoh SI. 2013. Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping. J Clin Microbiol. 51: 299-302.

Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PW, Martín C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, Musser JM, van Embden JD. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J Clin Microbiol. 1999 Aug;37(8):2607-18.

Kubica T, Rüsch- Gerdes S, Niemann S. *Mycobacterium bovis* subsp. *caprae* caused one-third of human *M. bovis*- associated tuberculosis cases reported in Germany between 1999 and 2001. J Clin Microbiol. 2003 Jul;41(7):3070- 7.

Lari N, Rindi L, Cristofani R, Rastogi N, Tortoli E, Garzelli C. Association of *Mycobacterium tuberculosis* complex isolates of BOVIS and Central Asian (CAS) genotypic lineages with extrapulmonary disease. Clin Microbiol Infect. 2009 Jun;15(6):538-43.

Lazzarini LC, Rosenfeld J, Huard RC, Hill V, Lapa e Silva JR, DeSalle R, Rastogi N, Ho JL. *Mycobacterium tuberculosis* spoligotypes that may derive from mixed strain infections are revealed by a novel computational approach. Infect Genet Evol. 2012 Jun;12(4):798-806.

Leão S, Portaels F. 2007. Chapter 1 – History. In Tuberculosis 2007: from basic science to patient care, Edited by Palomino JC, Leao S, Ritacco V. 2007, 25-52, Amedeo Online Textbooks; http://pdf.flyingpublisher.com/tuberculosis2007.pdf

López B, Aguilar D, Orozco H, Burger M, Espitia C, Ritacco V, Barrera L, Kremer K, Hernandez-Pando R, Huygen K, van Soolingen D. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. Clin Exp Immunol. 2003; 133(1):30-7.

Losos JB, Arnold SJ, Bejerano G, Brodie ED 3rd, Hibbett D, Hoekstra HE, Mindell DP, Monteiro A, Moritz C, Orr HA, Petrov DA, Renner SS, Ricklefs RE, Soltis PS, Turner TL. Evolutionary biology for the 21st century. PLoS Biol. 2013;11(1):e1001466.

M. Huguier, A. Flahaut. Biostatistiques au quotidien. Paris : Elsevier (2000), 204 p. ISBN 2-84299-174-5.

Madikizela B, Ndhlala AR, Finnie JF, Van Staden J. Antimycobacterial, anti-inflammatory and genotoxicity evaluation of plants used for the treatment of tuberculosis and related symptoms in South Africa. J Ethnopharmacol. 2014 Apr 28;153(2):386-91.

Millet J, Baboolal S, Streit E, Akpaka PE, Rastogi N. A first assessment of *Mycobacterium tuberculosis* genetic diversity and drug-resistance patterns in twelve Caribbean territories. Biomed Res Int. 2014;2014:718496.

Miltgen, Jean; Morillon, Marc; Koeck, Jean-Louis; Varnerot, Anne; Briant, Jean-François; Nguyen, Gilbert; Verrot, Denis; Bonnet, Daniel; Vincent, Véronique (2002). "Two Cases of Pulmonary Tuberculosis Caused by *Mycobacterium tuberculosis* subsp. canetti". Emerging Infectious Diseases 8 (11): 1350–2.

Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, Narvskaya O. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res 2005; 15:1357-1364.

Mokrousov I. 2009. *Corynebacterium diphtheriae*: genome diversity, population structure and genotyping perspectives. Infect Genet Evol. 9(1):1-15.

Mokrousov I. 2012. The quiet and controversial: Ural family of *Mycobacterium tuberculosis*. Infect Genet Evol. 12: 619-629.

Mokrousov I. 2008. Genetic geography of *Mycobacterium tuberculosis* Beijing genotype: a multifacet mirror of human history? Infect Genet Evol. 8:777-785.

Mokrousov I, Jiao WW, Wan K, Shen A. 2014. Stranger in a strange land: Ibero-American strain of *Mycobacterium tuberculosis* in Tibet, China. Infect Genet Evol. 26:323-6.

Mostowy S, Cousins D, Behr MA. Genomic interrogation of the dassie bacillus reveals it as a unique RD1 mutant within the *Mycobacterium tuberculosis* complex. J Bacteriol. 2004 Jan;186(1):104-9.

Murray M, et Alland D. 2002. Methodological problems in the molecular epidemiology of tuberculosis. Am. J. Epidemiol. 155:565 - 71.

Niobe-Eyangoh SN, Kuaban C, Sorlin P, Thonnon J, Vincent V, Gutierrez MC. Molecular characteristics of strains of the cameroon family, the major group of *Mycobacterium tuberculosis* in a country with a high prevalence of tuberculosis. J Clin Microbiol. 2004 Nov;42(11):5029-35.

Orcau À, Caylà JA, Martínez JA. Present epidemiology of tuberculosis. Prevention and control programs. Enferm Infecc Microbiol Clin. 2011 Mar;29 Suppl 1:2-7.

Parsons S, Smith SG, Martins Q, Horsnell WG, Gous TA, Streicher EM, Warren RM, van Helden PD, Gey van Pittius NC. Pulmonary infection due to the dassie bacillus (*Mycobacterium tuberculosis* complex sp.) in a free-living dassie (rock hyrax-Procavia capensis) from South Africa. Tuberculosis (Edinb). 2008 Jan;88(1):80-3.

Pease AS. Some remarks on the diagnosis and treatment of tuberculosis in antiquity. Isis 1940; 31: 380-93.

Pfyffer, GE, Auckenthaler R, van Embden JDA, van Soolingen D. Mycobacterium canetti, the smooth variant of *M. tuberculosis*, isolated from a Swiss patient exposed in africa. 1998, Emerg. Infect. Dis. 4:631-634

Pitulle C, Dorsch M, Kazda J, Wolters J, Stackebrandt E (1992). Phylogeny of rapidly growing members of the genus Mycobacterium. International Journal of Systematic Bacteriology, 42: 337-343.

Plikaytis BB, Marden JL, Crawford JT, Woodley CL, Butler WR, Shinnick TM. Multiplex PCR assay specific for the multidrug-resistant strain W of *Mycobacterium tuberculosis*. J Clin Microbiol. 1994 Jun;32(6):1542-6.

Rahim Z, Möllers M, te Koppele- Vije A, de Beer J, Zaman K, Matin MA, Kamal M, Raquib R, van Soolingen D, Baqi MA, Heilmann FG, van der Zanden AG. Characterization of *Mycobacterium africanum* subtype I among cows in a dairy farm in Bangladesh using spoligotyping. Southeast Asian J Trop Med Public Health. 2007 Jul;38(4):706- 13.

Rastogi N, Legrand E, Sola C (2001). The mycobacteria: an introduction to nomenclature and pathogenesis. Rev Sci Tech; 20(1):21-54.

Rastogi N, Sola C. 2007. Chapter 2 – Molecular evolution of the *Mycobacterium tuberculosis* complex. In Tuberculosis 2007: from basic science to patient care, Edited by Palomino JC, Leao S, Ritacco V. 2007, 53-91, Amedeo Online Textbooks; http://pdf.flyingpublisher.com/tuberculosis2007.pdf

Reyes JF, Francis AR, Tanaka MM. 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics. 9: 496.

Rodriguez-Campos S, González S, de Juan L, Romero B, Bezos J, Casal C, Álvarez J, Fernández-de-Mera IG, Castellanos E, Mateos A, Sáez-Llorente JL, Domínguez L, Aranaz A; Spanish Network on Surveillance Monitoring of Animal Tuberculosis. 2012. A database for animal tuberculosis (mycoDB.es) within the context of the Spanish national program for eradication of bovine tuberculosis. Infect Genet Evol. 12(4): 877-82.

Rogall T, Wolters J, Flohr T, Böttger EC. Towards a phylogeny and definition of species at the molecular level within the genus Mycobacterium. International Journal of Systematic Bacteriology 1990, 40: 323-330.

Rothschild BM, Martin LD, Lev G, Bercovier H, Bar-Gal GK, Greenblatt C, Donoghue H, Spigelman M, Brittain D. 2001. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. Clin Infect Dis. 33: 305-311.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems.Journal of the Royal Statistical Society. Series B (Methodological), 50(2) :p p. 157 224, 1988.

Saitou, N., and Nei, M. (1987), "The neighbor-joining method: a new method for reconstructing phylogenetic trees", Mol. Biol. Evol., 4(4):406-425.

Salo WL, Aufderheide AC, Buikstra J, Holcomb TA. Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. Proc Natl Acad Sci U S A. 1994 Mar 15;91(6):2091-4.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097-6100.

Serrano JE, Garrido Contreras AA, and Vizcaino LL. Considérations sur la tuberculose chez les animaux de compagnie. 1986, Méd. Mal. Infect. 10:556 - 7

Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP. 2012. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. Infect Genet Evol. 12: 789-797.

Shinnick TM, Good RC (1994). Mycobacterial taxonomy. Eur J Clin Microbiol Infect Dis; 13(11):884-901.

Small, P. and Behr, M. Molecular fingerprinting of *Mycobacterium tuberculosis*: how can it help the clinician? Clinical Infectious Diseases. 1997 25 : 806- 10

Smith NH, Berg S, Dale J, Allen A, Rodriguez S, Romero B, Matos F, Ghebremichael S, Karoui C, Donati C, Machado Ada C, Mucavele C, Kazwala RR, Hilty M, Cadmus S, Ngandolo BN, Habtamu M, Oloya J, Muller A, Milian-Suazo F, Andrievskaia O, Projahn M, Barandiarán S, Macías A, Müller B, Zanini MS, Ikuta CY, Rodriguez CA, Pinheiro SR, Figueroa A, Cho SN, Mosavari N, Chuang PC, Jou R, Zinsstag J, van Soolingen D, Costello E, Aseffa A, Proaño-Perez F, Portaels F, Rigouts L, Cataldi AA, Collins DM, Boschiroli ML, Hewinson RG, Ferreira Neto JS, Surujballi O, Tadyon K, Botelho A, Zárraga AM, Buller N, Skuce R, Michel A, Aranaz A, Gordon SV, Jeon BY, Källenius G, Niemann S, Boniotti MB, van Helden PD, Harris B, Zumárraga MJ, Kremer K. European 1: a globally important clonal complex of *Mycobacterium bovis*. Infect Genet Evol. 2011 Aug;11(6):1340-51.

Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol. 2009 Jul;7(7):537-44.

Smith NH, Upton P. 2012. Naming spoligotype patterns for the RD9-deleted lineage of the Mycobacterium tuberculosis complex; www.Mbovis.org. Infect Genet Evol. 12 (4), pp. 873–876.

Soares P, Alves RJ, Abecasis AB, Penha-Gonçalves C, Gomes MG, Pereira-Leal JB. (2013) inTB – a data integration platform for molecular and clinical epidemiological analysis of tuberculosis. BMC Bioinformatics. 14: 264.

Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, Musser JM. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J Clin Microbiol. 2000 Feb;38(2):669-76.

Sola C, Devallois A, Horgen L, Maïsetti J, Filliol I, Legrand E, Rastogi N. 1999. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. Emerg Infect Dis. 5(3): 404-14.

Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N. 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect Genet Evol. 3: 125-133.

Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. 2001. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. J Mol Evol. 53: 680-689.

Sotomayor H, Burgos J, Arango M. [Demonstration of tuberculosis by DNA ribotyping of *Mycobacterium tuberculosis* in a Colombian prehispanic mummy]. Biomedica. 2004 Jun;24 Supp 1:18-26. Spanish.

Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci USA. 94: 9869-9874.

Stead WW. The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. Clin Chest Med. 1997 Mar;18(1):65-77.

Stephen M. Stigler. Karl Pearson's Theoretical Errors and the Advances They Inspired. Statist. Sci. Volume 23, Number 2 (2008), 151-285

Stucki D, Gagneux S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. Tuberculosis (Edinb). 2013 Jan;93(1):30-9.

Sula L, Ra dkovský I. Protective effects of M. microti vaccine against tuberculosis. J Hyg Epidemiol Microbiol Immunol. 1976;20(1):1- 6.

Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol. 44: 4498-4510.

Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J Clin Microbiol. 39: 3563-3571.

Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat Genet. 2013;45:172–9.

Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. Bioinformatics 2008; 24:2414-415.

The PCR Revolution: Basic Technologies and Applications. Edited by Stephen A. Bustin PhD

Thoen CO, Karlson AG, Himes EM. Disease in domestic and feral animals: *Mycobacterium tuberculosis* complex. 1984, Kubica GP and Wayne LG (ed.), The mycobacteria: a sourcebook, part B. Marcel Dekker, Inc., New York, p. 1209– 1236.

Thorel MF, Karoui C, Varnerot A, Fleury C, Vincent V. Isolation of *Mycobacterium bovis* from baboons, leopards and a sea - lion. Vet Res. 1998 Mar -Apr;29(2):207- 12.

van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small PM. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol. 31, 406-409.

van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, van Soolingen D. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. Emerg Infect Dis. 2012;18(4):653-5.

van Soolingen D, Hoogenboezem T, de Haas PE, Hermans PW, Koedam MA, Teppema KS, Brennan PJ, Besra GS, Portaels F, Top J, Schouls LM, van Embden JD. A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterization of an exceptional isolate from Africa. Int J Syst Bacteriol. 1997 Oct;47(4):1236- 45.

van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD. (1995). Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. J Clin Microbiol. 33: 3234-3238.

van Soolingen D. Whole-genome sequencing of *Mycobacterium tuberculosis* as an epidemiological marker. Lancet Respir Med. 2014 Apr;2(4):251-2.

Vanhomwegen J, Kwara A, Martin M, Gillani FS, Fontanet A, Mutungi P, Crellin J, Obaro S, Gosciminski M, Carter EJ, Rastogi N. Impact of immigration on the molecular epidemiology of tuberculosis in Rhode Island. J Clin Microbiol. 2011 Mar;49(3):834-44.

Warren RM, Streicher EM, Charalambous S, Churchyard G, van der Spuy GD, Grant AD, van Helden PD, Victor TC. Use of spoligotyping for accurate classification of recurrent tuberculosis. J Clin Microbiol. 2002 Oct;40(10):3851-3.

Wells AQ, and Oxon DM. 1937. Tuberculosis in wild voles. Lancet i:1221.

Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res. 38(Web Server issue): W326–W331.

WHO. 2013. Global tuberculosis report 2013. www.who.int/iris/bitstream/10665/91355/1/978941564656_eng.pdf.

Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, van Soolingen D, Rüsch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 2008 Sep 19;4(9):e1000160.

Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. Nature. 447: 279-283.

Zhang Y, Yew WW. Mechanisms of drug resistance in *Mycobacterium tuberculosis*. Int J Tuberc Lung Dis. 2009 Nov;13(11):1320-30.

Zink AR, Sola C, Reischl U, Grabner W, Rastogi N, Wolf H, Nerlich AG. 2003. Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. J Clin Microbiol. 41: 359-367.

Zozio T, Allix C, Gunal S, Saribas Z, Alp A, Durmaz R, Fauville-Dufaux M, Rastogi N, Sola C. Genotyping of *Mycobacterium tuberculosis* clinical isolates in two cities of Turkey: description of a new family of genotypes that is phylogeographically specific for Asia Minor. BMC Microbiol. 2005 Jul 26;5:44.

Magazine GEO N°403 de septembre 2012 p.18 (https://www.t411.me/t/4686063)

LISTE DES PUBLICATIONS (REALISEES OU EN COURS)

En ordre d'apparition dans la thèse

Article 1
Demay C, Liens B, Burguière T, Hill V, **Couvin D**, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N (2012). SITVITWEB – a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol; 12(4):755-66.

Article 2
David Couvin, Nalin Rastogi (2014). Macro-geographical specificities of the prevailing Tuberculosis epidemic as seen through SITVIT2, an updated version of the *Mycobacterium tuberculosis* genotyping database. (en rédaction)

Article 2bis
Couvin D and Rastogi N. (2014). The establishment of databases on circulating genotypes of *Mycobacterium tuberculosis* complex and web tools for an effective response to better monitor, understand and control the tuberculosis epidemic worldwide. Euroreference Les cahiers de la reference vol.12. https://pro.anses.fr/euroreference/Documents/ER12-RESEAUX-MycobacteriumEN.pdf

Article 3
Couvin D and Rastogi N. (2014) Tuberculosis – a global emergency: tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage. Tuberculosis. (Sous-presse)

Article 4
David Couvin, Iñaki Cervera-Marzal, Audrey David, Nalin Rastogi. SITVITBovis – a publicly available database and mapping tool to get a global overview of animal and human cases caused by *Mycobacterium bovis*. (en rédaction)

Article 5
Minoo Aminian, **David Couvin**, Amina Shabbeer, Kane Hadley, Scott Vandenberg, Nalin Rastogi, and Kristin P. Bennett (2014). Predicting *Mycobacterium tuberculosis* Complex Clades Using Knowledge-Based Bayesian Networks. BioMed Research International, vol. 2014, Article ID 398484, 11 pages.

Article 6
Martinez-Guarneros A, Rastogi N, **Couvin D**, Escobar-Gutierrez A, Rossi LM, Vazquez-Chacon CA, Rivera-Gutierrez S, Lozano D, Vergara-Castañeda A, Gonzalez-Y-Merchand JA, Vaughan G (2013). Genetic diversity among multidrug-resistant *Mycobacterium tuberculosis* strains in Mexico. Infect Genet Evol; 14:434-43.

Article 7
Realpe T, Correa N, Rozo JC, Ferro BE, Gomez V, Zapata E, Ribon W, Puerto G, Castro C, Nieto LM, Diaz ML, Rivera O, **Couvin D**, Rastogi N, Arbelaez MP, Robledo J. (2014) Population Structure among *Mycobacterium tuberculosis* Isolates from Pulmonary Tuberculosis Patients in Colombia. PLoS One; 9(4):e93848.

Article 8
Sheen P, **Couvin D**, Grandjean L, Zimic M, Dominguez M, Luna G, Gilman RH, Rastogi N, Moore DA (2013). Genetic diversity of *Mycobacterium tuberculosis* in Peru and exploration of phylogenetic associations with drug resistance. PLoS One; 8(6):e65873.

Article 9
Martins MC, Giampaglia CM, Oliveira RS, Simonsen V, Latrilha FO, Moniz LL, **Couvin D**, Rastogi N, Ferrazoli L (2013). Population structure and circulating genotypes of drug-sensitive and drug-resistant *Mycobacterium tuberculosis* clinical isolates in São Paulo state, Brazil. Infect Genet Evol; 14:39-45.

Article 10
Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, Riccardi F, **Couvin D**, Hill V, Rastogi N, Koivula T, Källenius G (2011). The Guinea-Bissau family of *Mycobacterium tuberculosis* complex revisited. PloS One; 6(4):e18601.

Article 11
Mulugeta Belay, Gobena Ameni, Gunnar Bjune, **David Couvin**, Nalin Rastogi, and Fekadu Abebe (2014). Strain Diversity of *Mycobacterium tuberculosis* Isolates from Pulmonary Tuberculosis Patients in Afar Pastoral Region of Ethiopia. BioMed Research International Volume 2014, Article ID 238532, 12 pages.

Article 12
Ruqaya Mustafa Ali, Alberto Trovato, **David Couvin**, Amina N. Al-Thwani, Emanuele Borroni, Fahim H. Dhaer, Nalin Rastogi, and Daniela M. Cirillo (2014). Molecular Epidemiology and Genotyping of *Mycobacterium tuberculosis* Isolated in Baghdad. BioMed Research International Volume 2014, Article ID 580981, 15 pages.

Article 13
Jyoti Arora, Zeeshan Sidiq, Sangeeta Sharma, Ritu Singhal, Manpreet Bhalla, **David Couvin**, Rohit Sarin, Nalin Rastogi, Vithal Prasad Myneeedu (2014). Phylogenetic associations with drug resistance of *Mycobacterium tuberculosis* isolates in a pediatric population. Int J Tuberc Lung Dis; 18(10):1172-9.

Article 14
Zhao Y, Feng Q, Tang K, Zhang C, Sun H, Luo T, Yang Z, **Couvin D**, Rastogi N, Sun Q (2012). The population structure of drug-resistant *Mycobacterium tuberculosis* clinical isolates from Sichuan in China. Infect Genet Evol; 12(4):718–24.

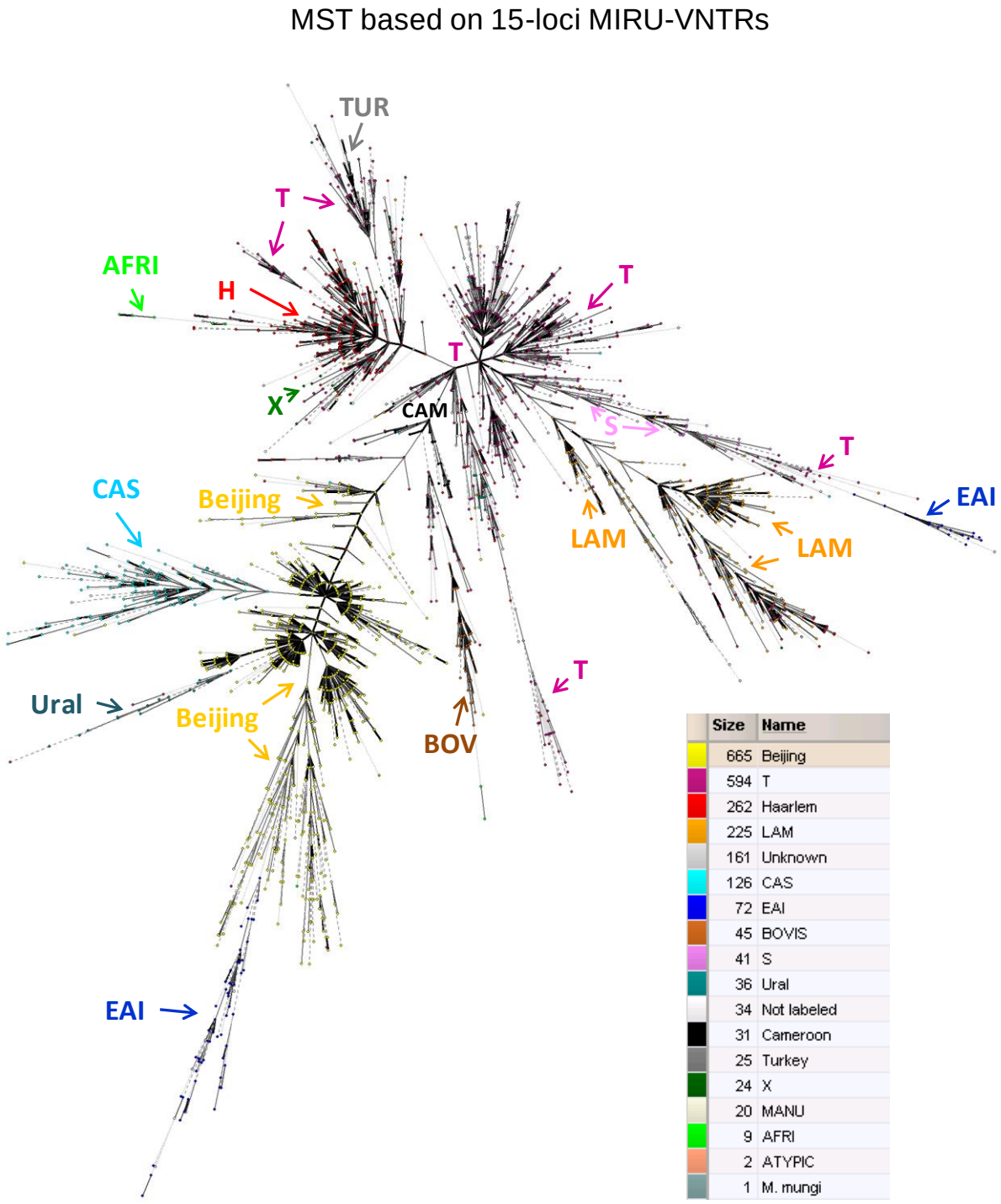
Article 15
Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S, **Couvin D**, Koivula T, Rastogi N, Källenius G (2012). *Mycobacterium tuberculosis* strains potentially involved in the TB epidemic in Sweden a century ago. PLoS One; 7(10):e46848.

Article 16
Smit PW, Haanperä M, Rantala P, **Couvin D**, Lyytikäinen O, Rastogi N, Ruutu P, Soini H (2013). Molecular epidemiology of tuberculosis in Finland, 2008-2011. PLoS One; 8(12):e85027.

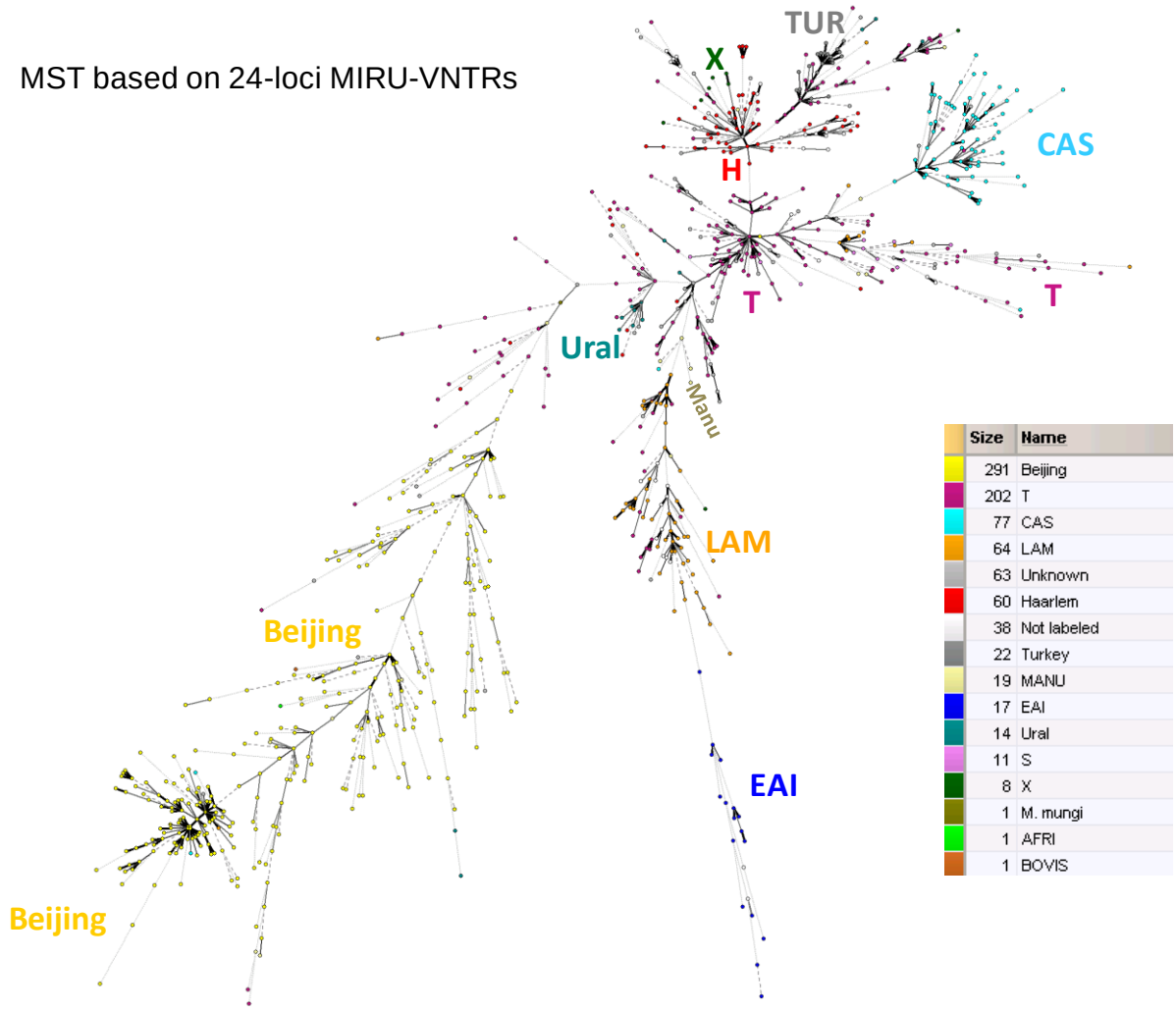
Article 17
Fallico L, **Couvin D**, Peracchi M, Pascarella M, Franchin E, Lavezzo E, Rassu M, Manganelli R, Rastogi N, Palù G. (2014). Four year longitudinal study of *Mycobacterium tuberculosis* complex isolates in a region of North-Eastern Italy. Infect Genet Evol; 26:58-64.

ANNEXES

Annexe 1 : MST construit avec les profils 15-loci MIRU-VNTRs présents et complets (chaque locus a une valeur) dans SITVIT2 (n=3724 isolats cliniques)



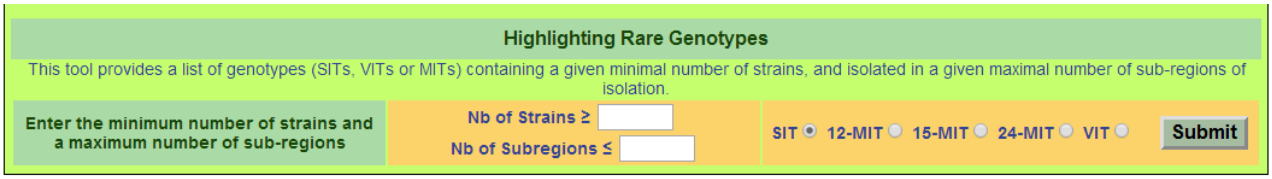
Annexe 2 : MST construit avec les profils 24-loci MIRU-VNTRs présents et complets (chaque locus a une valeur) dans SITVIT2 (n=1463 isolats cliniques).



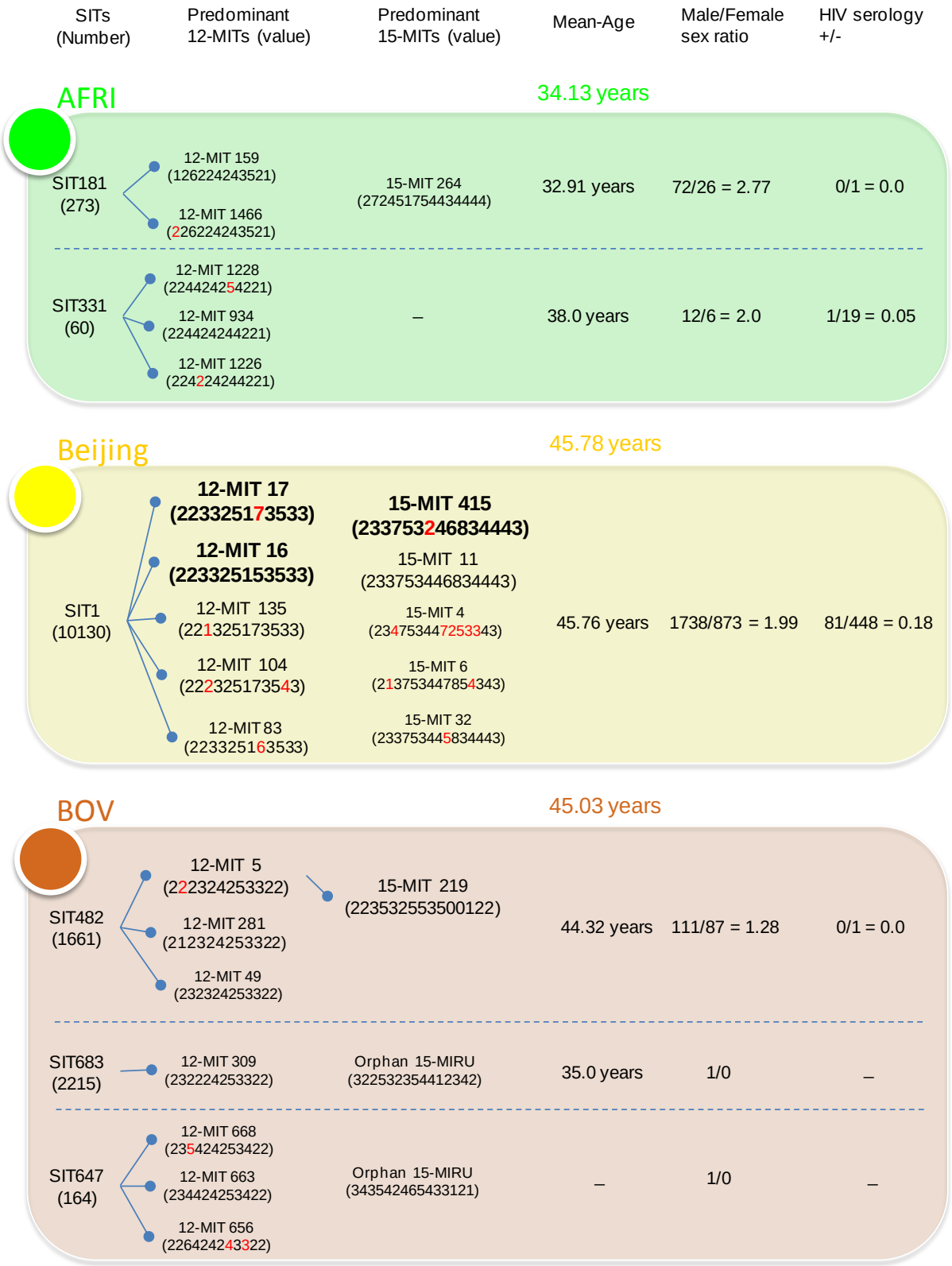
Annexe 3A : Pseudo-code de l'algorithme permmettant la mise en évidence des génotypes rares.

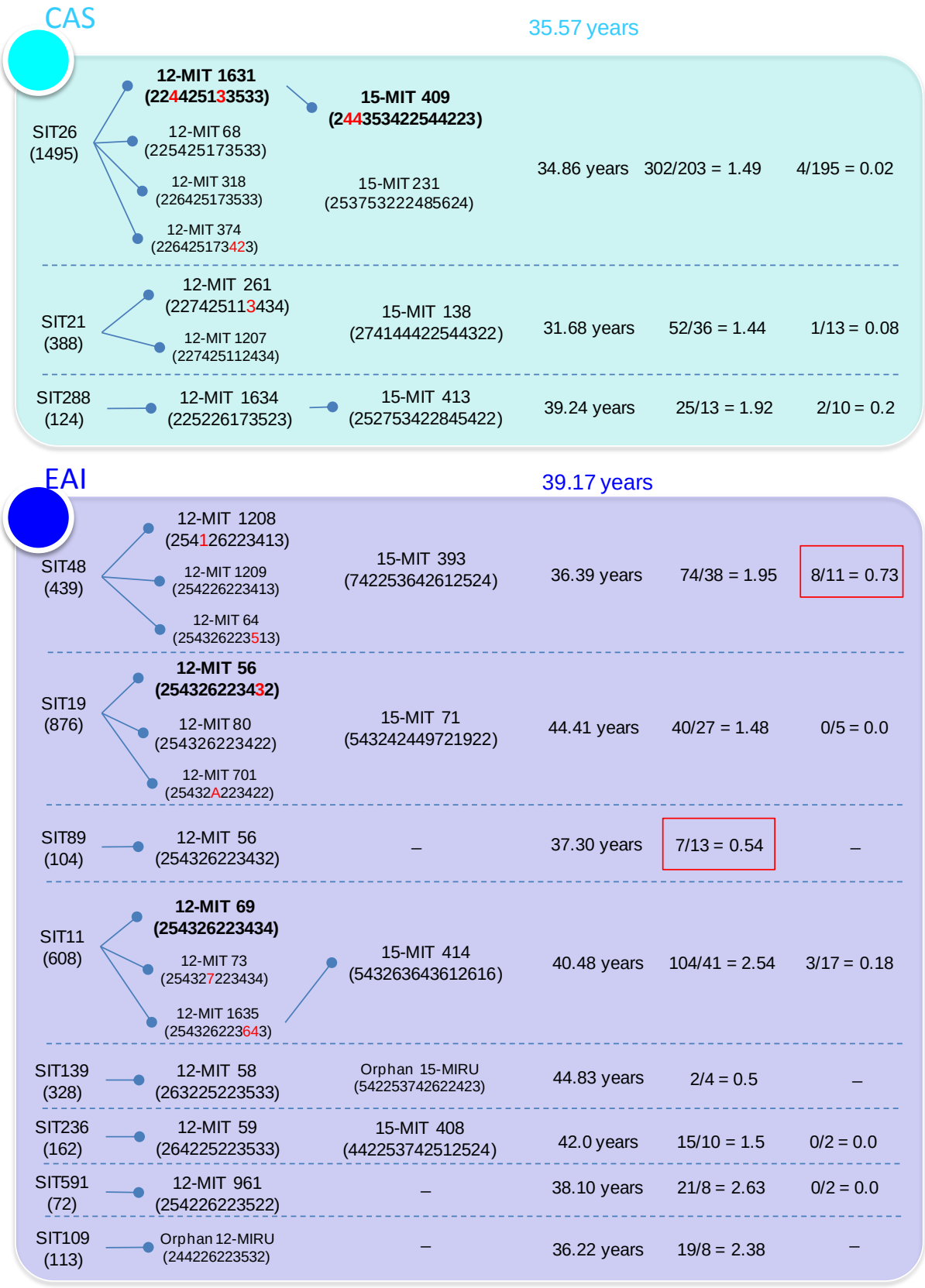
```
Fonction highlighting_rare_genotypes (type_marker, number_strains,
number_regions){

      //La variable type_marker en option de fonction vaut "SIT", "12-MIT",
      //"15-MIT", "24-MIT", ou "VIT"

      //Déclaration de la variable liste_genotypes qui sera retourner à la
      //fin de la fonction ou procédure
      Liste_genotypes = ""

      //Déclaration des variables nb_strains_by_IT et nb_regions_by_IT
      nb_strains_by_IT = 0
      nb_regions_by_IT = 0

      //On ne peut exécuter les instructions que dans le cas où les
      //variables sont bien définies et ont une valeur
      Si (type_marker ≠ null) {

            //On parcours la liste des génotypes en fonction du type_marker
            Pour i de 1 jusqu'à taille_Liste_type_marker {

                  //Récupérer le nombre de souches pour un profil IT donné
                  //dans la liste, en fonction du type de marqueur
                  nb_strains_by_IT = getNbStrainsParIT(IT_genotype_Marker)

                  //Récupérer le nombre de régions contenant un profil IT
                  //donné dans la liste, en fonction du type de marqueur
                  nb_regions_by_IT = getNbRegionsParIT(IT_genotype_Marker)

                  //Sélection des profils concernés en function des
                  //critères (nb strains et nb regions) saisis par
                  //l'utilisateur dans l'interface
                  si(nb_strains_by_IT ≥ number_strains et
                        nb_regions_by_IT ≤ number_regions) {
                        AjouterProfilAListe (Liste_genotypes)
                  //Fin « si »
                  }
            //Fin « Pour »
            }
      //Fin « Si »
      }
      Retourner Liste_genotypes
}//Fin « Fonction »
```

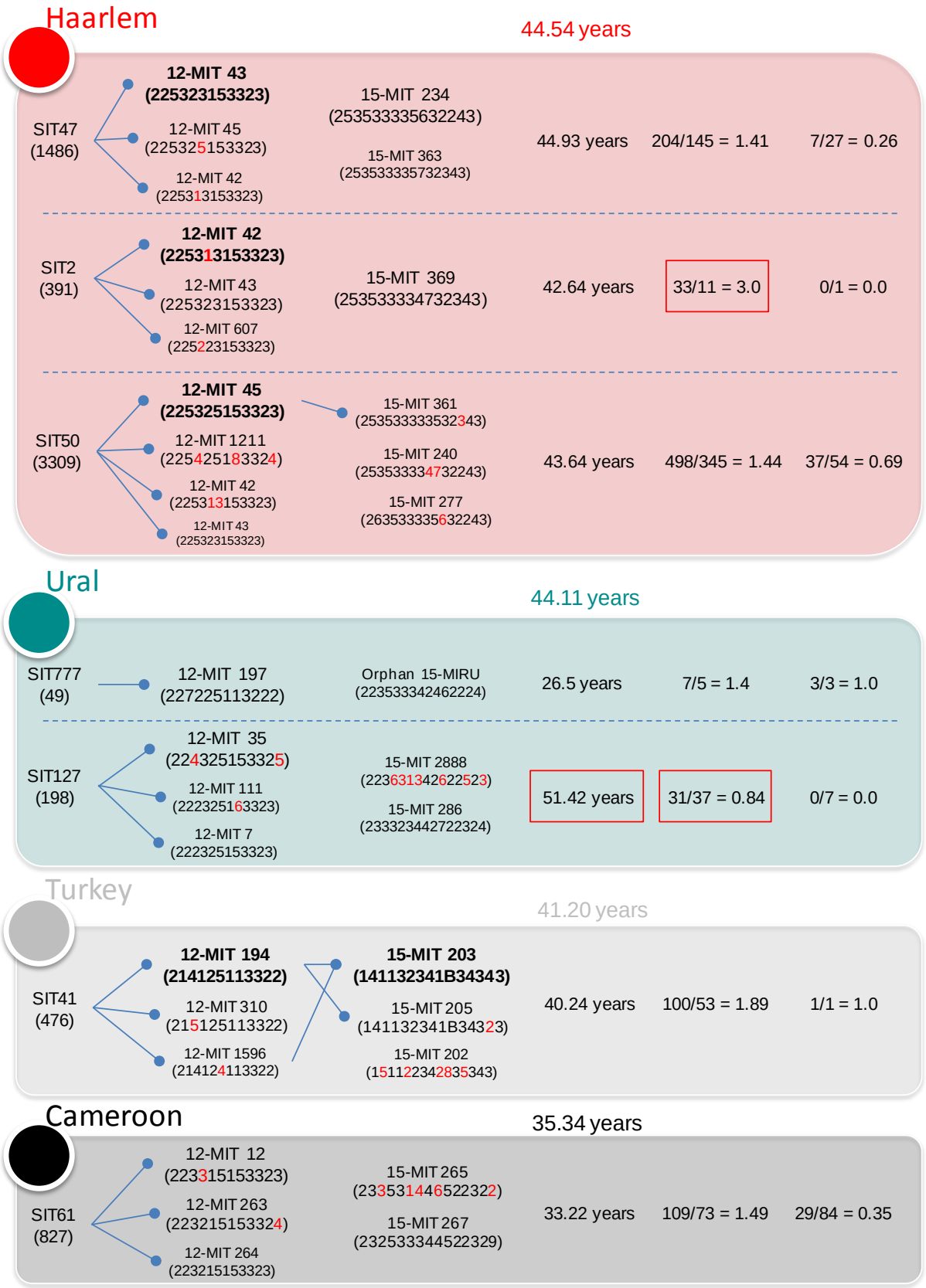
Annexe 3B : Interface web SITVIT2 permettant la mise en évidence des génotypes rares.

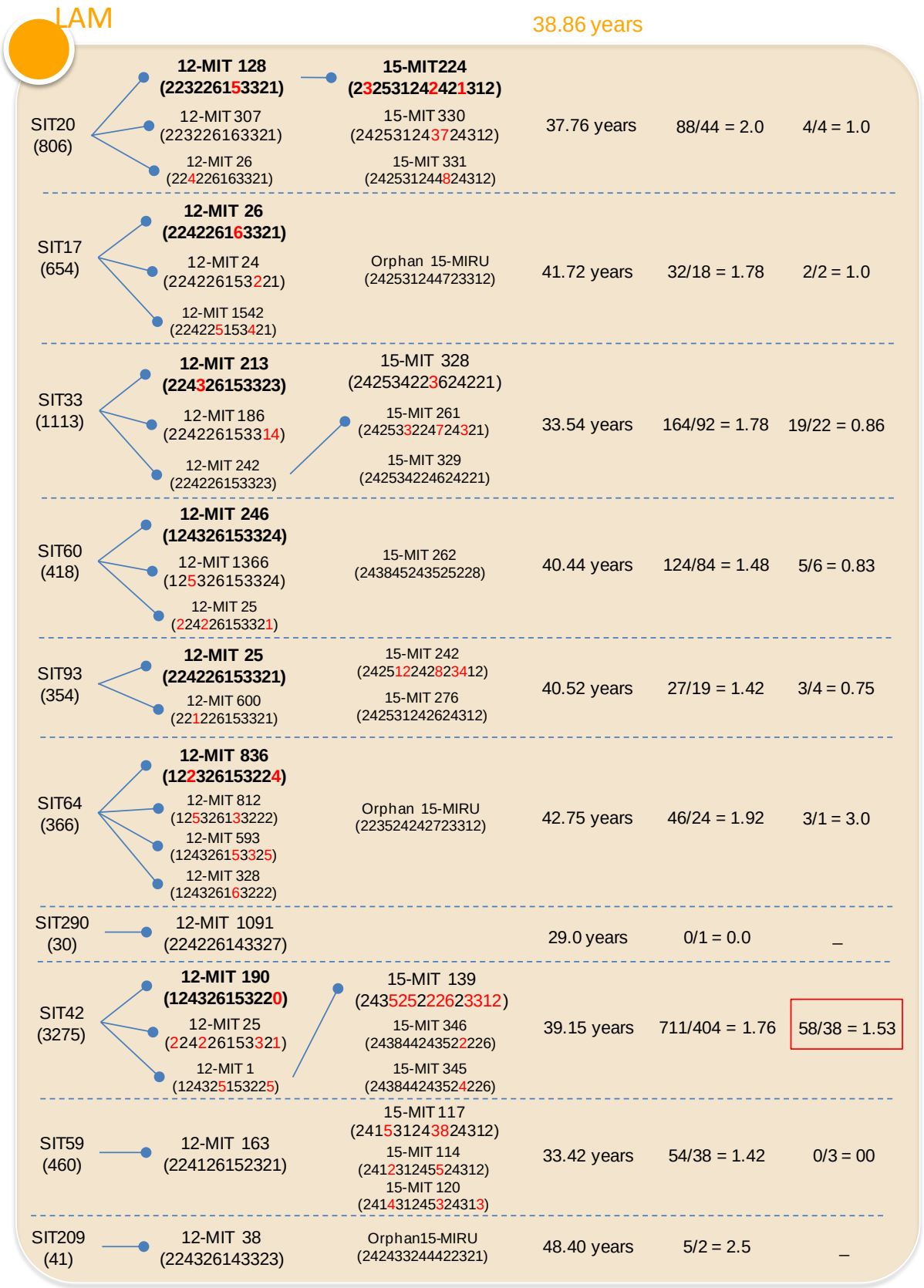


Annexe 4 : Principaux SITs représentant les lignées majeures basées sur le spoligotypage, suivis de leurs associations avec les 12- ou 15-loci MIRU-VNTRs prédominants, ainsi que les données démographiques (sexe, âge, et statut VIH). Les liens (flèches bleues) entre les profils spoligotypes et/ou MIRU sont affichés en fonction de la présence de l'information dans la base de données SITVIT2.




CAS
35.57 years
SIT26 (1495)
12-MIT 1631 (22442513 3533)
12-MIT 68 (225425173533)
12-MIT 318 (226425173533)
12-MIT 374 (2264251734 23)
15-MIT 409 (244 353422544223)
15-MIT 231 (253753222485624)
34.86 years
302/203 = 1.49
4/195 = 0.02
SIT21 (388)
12-MIT 261 (22742511 3434)
12-MIT 1207 (227425112434)
15-MIT 138 (274144422544322)
31.68 years
52/36 = 1.44
1/13 = 0.08
SIT288 (124)
12-MIT 1634 (225226173523)
15-MIT 413 (252753422845422)
39.24 years
25/13 = 1.92
2/10 = 0.2
EAI
39.17 years
SIT48 (439)
12-MIT 1208 (2541 26223413)
12-MIT 1209 (254226223413)
12-MIT 64 (25432622 3513)
15-MIT 393 (742253642612524)
36.39 years
74/38 = 1.95
8/11 = 0.73
SIT19 (876)
12-MIT 56 (25432622343 2)
12-MIT 80 (254326223422)
12-MIT 701 (25432A223422)
15-MIT 71 (543242449721922)
44.41 years
40/27 = 1.48
0/5 = 0.0
SIT89 (104)
12-MIT 56 (254326223432)
–
37.30 years
7/13 = 0.54
–
SIT11 (608)
12-MIT 69 (254326223434)
12-MIT 73 (254327223434)
12-MIT 1635 (25432622364 3)
15-MIT 414 (543263643612616)
40.48 years
104/41 = 2.54
3/17 = 0.18
SIT139 (328)
12-MIT 58 (263225223533)
Orphan 15-MIRU (542253742622423)
44.83 years
2/4 = 0.5
–
SIT236 (162)
12-MIT 59 (264225223533)
15-MIT 408 (442253742512524)
42.0 years
15/10 = 1.5
0/2 = 0.0
SIT591 (72)
12-MIT 961 (254226223522)
–
38.10 years
21/8 = 2.63
0/2 = 0.0
SIT109 (113)
Orphan 12-MIRU (244226223532)
–
36.22 years
19/8 = 2.38
–


Haarlem
44.54 years
SIT47 (1486)
12-MIT 43 (225323153323)
12-MIT 45 (225325153323)
12-MIT 42 (225313153323)
15-MIT 234 (253533335632243)
15-MIT 363 (253533335732343)
44.93 years
204/145 = 1.41
7/27 = 0.26
SIT2 (391)
12-MIT 42 (225313153323)
12-MIT 43 (225323153323)
12-MIT 607 (225223153323)
15-MIT 369 (253533334732343)
42.64 years
33/11 = 3.0
0/1 = 0.0
SIT50 (3309)
12-MIT 45 (225325153323)
12-MIT 1211 (225425183324)
12-MIT 42 (225313153323)
12-MIT 43 (225323153323)
15-MIT 361 (253533333532343)
15-MIT 240 (253533334732243)
15-MIT 277 (263533335632243)
43.64 years
498/345 = 1.44
37/54 = 0.69
Ural
44.11 years
SIT777 (49)
12-MIT 197 (227225113222)
Orphan 15-MIRU (223533342462224)
26.5 years
7/5 = 1.4
3/3 = 1.0
SIT127 (198)
12-MIT 35 (224325153325)
12-MIT 111 (222325163323)
12-MIT 7 (222325153323)
15-MIT 2888 (223631342622523)
15-MIT 286 (233323442722324)
51.42 years
31/37 = 0.84
0/7 = 0.0
Turkey
41.20 years
SIT41 (476)
12-MIT 194 (214125113322)
12-MIT 310 (215125113322)
12-MIT 1596 (214124113322)
15-MIT 203 (141132341B34343)
15-MIT 205 (141132341B34323)
15-MIT 202 (151122342835343)
40.24 years
100/53 = 1.89
1/1 = 1.0
Cameroon
35.34 years
SIT61 (827)
12-MIT 12 (223315153323)
12-MIT 263 (223215153324)
12-MIT 264 (223215153323)
15-MIT 265 (233531446522322)
15-MIT 267 (232533344522329)
33.22 years
109/73 = 1.49
29/84 = 0.35


LAM
38.86 years
SIT20 (806)
12-MIT 128 (232226153321)
12-MIT 307 (223226163321)
12-MIT 26 (224226163321)
15-MIT224 (232531242421312)
15-MIT 330 (242531243724312)
15-MIT 331 (242531244824312)
37.76 years
88/44 = 2.0
4/4 = 1.0
SIT17 (654)
12-MIT 26 (224226163321)
12-MIT 24 (224226153221)
12-MIT 1542 (224225153421)
Orphan 15-MIRU (242531244723312)
41.72 years
32/18 = 1.78
2/2 = 1.0
SIT33 (1113)
12-MIT 213 (224326153323)
12-MIT 186 (224226153314)
12-MIT 242 (224226153323)
15-MIT 328 (242534223624221)
15-MIT 261 (242532247243 21)
15-MIT 329 (242534224624221)
33.54 years
164/92 = 1.78
19/22 = 0.86
SIT60 (418)
12-MIT 246 (124326153324)
12-MIT 1366 (125326153324)
12-MIT 25 (224226153321)
15-MIT 262 (243845243525228)
40.44 years
124/84 = 1.48
5/6 = 0.83
SIT93 (354)
12-MIT 25 (224226153321)
12-MIT 600 (221226153321)
15-MIT 242 (242512242823412)
15-MIT 276 (242531242624312)
40.52 years
27/19 = 1.42
3/4 = 0.75
SIT64 (366)
12-MIT 836 (122326153224)
12-MIT 812 (125326133222)
12-MIT 593 (124326153325)
12-MIT 328 (124326163222)
Orphan 15-MIRU (223524242723312)
42.75 years
46/24 = 1.92
3/1 = 3.0
SIT290 (30)
12-MIT 1091 (224226143327)
29.0 years
0/1 = 0.0
_
SIT42 (3275)
12-MIT 190 (124326153220)
12-MIT 25 (224226153321)
12-MIT 1 (124325153225)
15-MIT 139 (243525252262 3312)
15-MIT 346 (243844243522226)
15-MIT 345 (243844243524226)
39.15 years
711/404 = 1.76
58/38 = 1.53
SIT59 (460)
12-MIT 163 (224126152321)
15-MIT 117 (241531243824312)
15-MIT 114 (241231245524312)
15-MIT 120 (241431245324313)
33.42 years
54/38 = 1.42
0/3 = 00
SIT209 (41)
12-MIT 38 (224326143323)
Orphan15-MIRU (242433244422321)
48.40 years
5/2 = 2.5
_

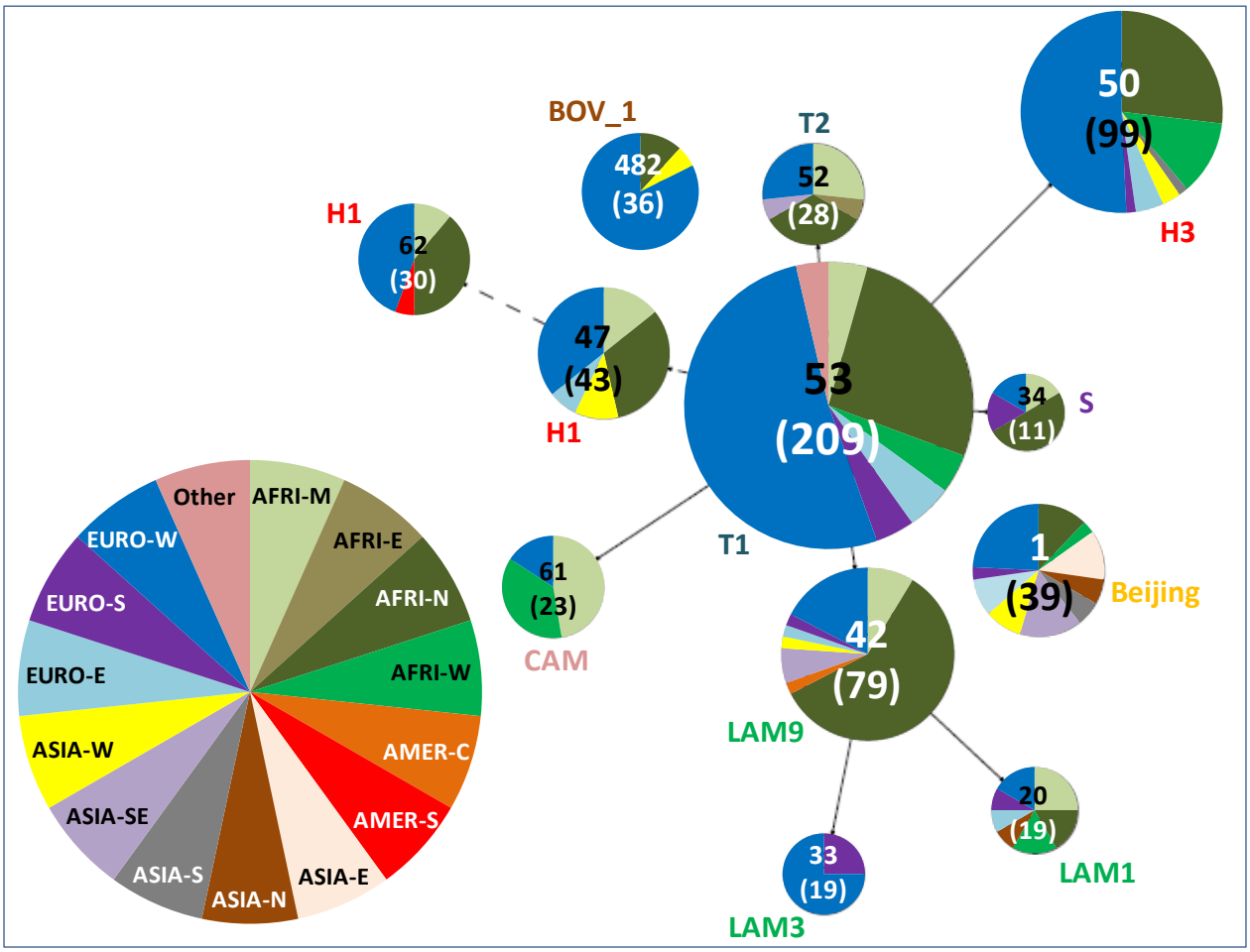
Manu
38.33 years
SIT100 (81)
12-MIT 1146 (223315153322)
12-MIT 1628 (223325143335)
15-MIT 406 (233435333612443)
37.36 years
8/6 = 1.33
1/5 = 0.2
SIT54 (244)
12-MIT 1628 (223325143335)
12-MIT 1138 (223325173323)
15-MIT 406 (233435333612443)
15-MIT 280 (233732225124 23)
37.49 years
57/26 = 2.19
0/15 = 0.0
SIT1378 (39)
–
–
42.5 years
4/2 = 2.0
–
S
46.02 years
SIT34 (836)
12-MIT 212 (233325153324)
12-MIT 1216 (333251532 24)
12-MIT 232 (333225143324)
12-MIT 279 (333251532 22)
15-MIT 260 (333534343143123)
45.41 years
100/64 = 1.56
3/4 = 0.75
X
37.45 years
SIT119 (1077)
12-MIT 34 (224325153324)
12-MIT 112 (223325143324)
12-MIT 23 (224225153324)
15-MIT 208 (353531331833443)
38.93 years
47/32 = 1.47
3/6 = 0.5
SIT137 (1023)
12-MIT 33 (224325153323)
12-MIT 22 (224225153323)
12-MIT 117 (224325143324)
15-MIT 114 (241231245524312)
15-MIT 135 (243533343832445)
37.04 years
24/7 = 3.43
0/5 = 0.0
SIT92 (427)
12-MIT 34 (224325153324)
12-MIT 1224 (225436155324)
12-MIT 46 (225325153324)
Orphan 15-MIRU 243532333834344
36.57 years
57/28 = 2.04
6/8 = 0.75

T

44.79 years

SIT1737 (28)	**12-MIT 140 (124326153224)** 12-MIT 565 (124326163224)		78.0 years	1/0	_
SIT53 (6152)	**12-MIT 8 (223125153324)** 12-MIT 40 (225125113322) 12-MIT 157 (223425153322) 12-MIT 15 (223325153322)	**15-MIT 263 (243533333712343)** 15-MIT 406 (233435333612443) 15-MIT 209 (231535342522226)	43.81 years	1069/635 = 1.68	62/91 = 0.68
SIT280 (70)	12-MIT 210 (222125153324) 12-MIT 8 (223125153324)	15-MIT 215 (431634323523223)	55.23 years	10/5 = 2.0	0/1 = 0.0
SIT52 (901)	**12-MIT 230 (223325153423)** 12-MIT 847 (223325154423) 12-MIT 13 (223325153224) 12-MIT 1100 (222325144323)	15-MIT 161 (223433344822333) 15-MIT 33 (233531342532125)	46.56 years	140/93 = 1.51	1/18 = 0.056
SIT135 (31)	12-MIT 11 (223315143423)	_	_	_	_
SIT37 (467)	**12-MIT 1222 (223415153323)** 12-MIT 112 (223325143324) 12-MIT 157 (223425153322)	15-MIT 181 (231535343522235)	44.08 years	80/48 = 1.67	3/11 = 0.27
SIT149 (465)	**12-MIT 40 (225125113322)**		30.55 years	33/33 = 1.0	4/26 = 0.15
SIT627 (39)	**12-MIT 310 (215125113322)**	15-MIT 14 (151132342842243)	73.48 years	13/12 = 1.08	_
SIT40 (157)	12-MIT 518 (123315133324) 12-MIT 1145 (223321153314)		41.90 years	21/11 = 1.91	0/3 = 0.0
SIT39 (143)	**12-MIT 117 (224325143324)** 12-MIT 220 (224325143325)		54.54 years	17/7 = 2.43	0/3 = 0.0
SIT44 (210)	**12-MIT 156 (223125153322)** 12-MIT 236 (223326153323)		47.76 years	28/10 = 2.8	0/2 = 0.0
SIT58 (173)	**12-MIT 18 (223326153324)** 12-MIT 339 (223326153325) 12-MIT 19 (223326153326)	Orphan 15-MIRU (233536343522222)	38.06 years	28/6 = 4.67	3/5 = 0.6
SIT254 (174)	**12-MIT 326 (124325153224)** 12-MIT 140 (124326153224) 12-MIT 1411 (124325143223)	15-MIT 140 (243524222623312)	37.61 years	32/17 = 1.88	1/3 = 0.33

Annexe 5 : Arbre Spoligoforest basé sur l'algorithme de Fruchterman-Reingold montrant les SITs prédominants dans l'étude épidémiologique menée en région Rhône-Alpes, et leur répartition en fonction des régions d'origine des patients indiquée dans l'étude.



<u>Annexe 6</u> : ***Slam TB*** : Poème/Essai venant appuyer le désir de lutte et d'éradication de la tuberculose.

Français
La tuberculose reste la "tuberculose" car elle continue à tuer,
Douleur et tristesse dans ces lendemains incertains,
Joie et plaisir dans cette brillante heureuse journée,
Incertitudes et précisions dans ces flux mouvementés,
Le temps passe, va, et vient,
Quel serait son avenir? Nous n'en savons pas plus,
Certains progrès et des catastrophes pour le moment,
Mais nous sommes encore là, nous gardons l'espoir d'un avenir meilleur,
Le temps passe, va, et vient,
Cependant, l'amour reste, demeure, et résiste,
Parce que le temps est un assassin silencieux, il a toujours quelque chose à tuer,
Un jour ou l'autre, nous n'aurons plus besoin d'un autre médicament,
D'une façon ou une autre, nous réussiront dans cette affaire,
Le temps passe, vole, et pleure,
Ça y est, c'est la vie que nous vivons. Est-ce cette vie que nous laissons?
A quel point peut-on comprendre cet infini sans limites?

English
TB is still TB, and it continues to kill time after time,
Pain and sorrow in this uncertain tomorrow,
Joy and pleasure in this happy shinny daytime,
Uncertainty and accuracy in this lively flow,
Time is traveling, coming, and passing,
What would be its future? We don't know,
Some advances and disasters for the time being,
But we are still keeping hoping for a better tomorrow,
Time is traveling, coming, and passing,
However, love is staying, remaining, resisting,
Because time is a silent assassin, it always has something to kill,
One day or another, we will not need another pill,
One way or another, we will succeed in this deal,
Time is passing, flying, and crying,
This is it, this is the life we live. This is the life we leave?
How much could we understand in this limitless infinity?

Kréyol
Tibèkiloz ka rété "Tibèkiloz", épi I ka kontinyé tchouyé,
Doulè épi tristès adan on douvini ensèten,
Tchè kontan épi plézi adan on bèl jouné ki ka kléré
Ensètitid épi détay adan on ambians ajité,
Tan la ka pasé, i kalé, épi i ka vini,
Ka i ké douvinn? Nou pa sav plis ki sa,
Kèk pwogré épi déga adan moman-lasa,
Men, nou toujou la, ka gadé èspwa pou on plibon divni,
Tan-la ka pasé, i kalé, épi i ka vini,
Kanmenmsa, lanmou ka rété, dousiné, é i ka kenbé,

Paskè tan sé on lasasen ki paka fè dézòd, i toujou ni on biten a kyouyé,
On jou ouben lòt, nou péké bizwen on dòt rimèd,
Antoutjan, nou ké réyisi adan zafè-lasa,
Tan-la ka pasé, i ka volé, é i ka pléré,
Mi mi-y, vi-la nou ka viv-la. Èskè sé lavi-lasa nou ka lagé?
Adan ki nivo nou pé konpwann biten san fen ni sès-lasa?

Español
La tuberculosis sigue siendo viva, con ganas de matar,
Dolor y la tristeza de esas mañanas, cuando nada esta cierto,
Felicidad y placer de esos días alegros y soleados,
Incerteza y precisión de esos flujos torrentosos,
El tiempo pasa, viene y luego se va,
¿Cuál sería su futuro? Tampoco sabemos algo de eso,
Algunos progresos y el desastre en este momento,
A pesar de todo, aquí estamos, y con esperanza nos quedamos,
El tiempo pasa, venga y luego se va,
El amor se queda, resistente y vivo,
Porque el tiempo es un asesino silencioso, que siempre tiene algo a destrozar,
Un día u otro, no necesitaremos otra medicina,
De una manera o de otra, logremos en este negocio,
El tiempo pasa, vola y luego llora,
Por fin estamos allí. Eso es la vida que vivimos. ¿Es esta la vida que dejaremos?
¿En qué medida podríamos entender este ilimitado infinito?

Português
A tuberculose continua sendo a tuberculose
E mesmo com o passar do tempo, continua matando
Dor e sofrimento em um amanhã incerto,
Alegria e felicidade neste dia tão brilhante,
Incerteza e exatidão neste fluxo diário,
O tempo é passageiro, ele vem, ele vai
E como será este futuro? Nós não sabemos,
Alguns avanços e alguns desastres acontecem o tempo todo,
Mas continuamos com esperança de um amanhã melhor,
O tempo é passageiro, ele vem, ele vai
No entanto, amor é ficar, permanecer, resistir
Porque o tempo é um assassino silencioso, ele sempre tem algo para matar,
Algum dia, nós não precisaremos mais de outra pílula
Algum dia, nós teremos sucesso nessa batalha
Tempo e está passando, voando e chorando
É assim, essa é a vida que vivemos. Essa é realmente a vida que levamos?
Quanto entendemos deste infinito sem limites?

Annexe 7 : Exemples d'analyses d'évolution spatio-temporelle en fonction de divers paramètres (lignée phylogénétique, âge, sexe, sérologie VIH, résistance aux antituberculeux).

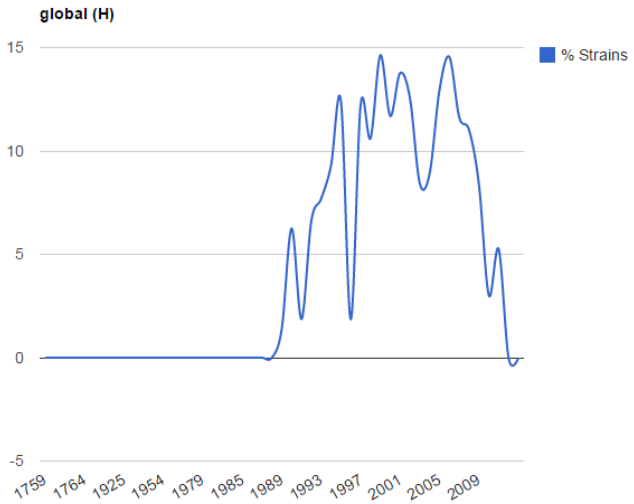


Figure 1 : Evolution globale de la lignée Haarlem (H).

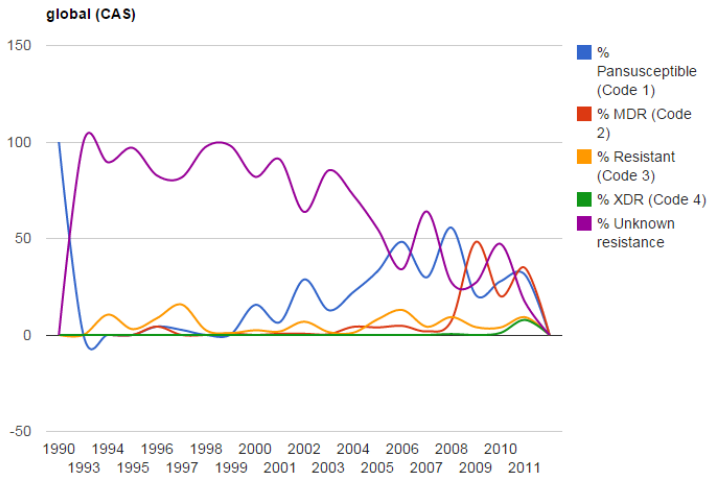


Figure 2 : Evolution globale des souches de lignée CAS en fonction des types de résistance aux antituberculeux.

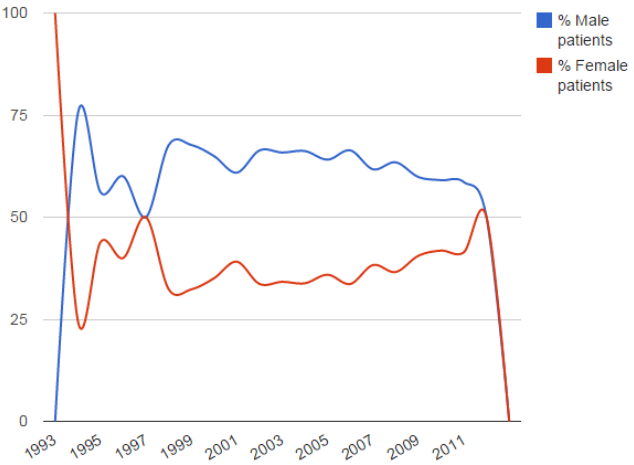


Figure 3 : Evolution globale des souches en fonction du sexe des patients.

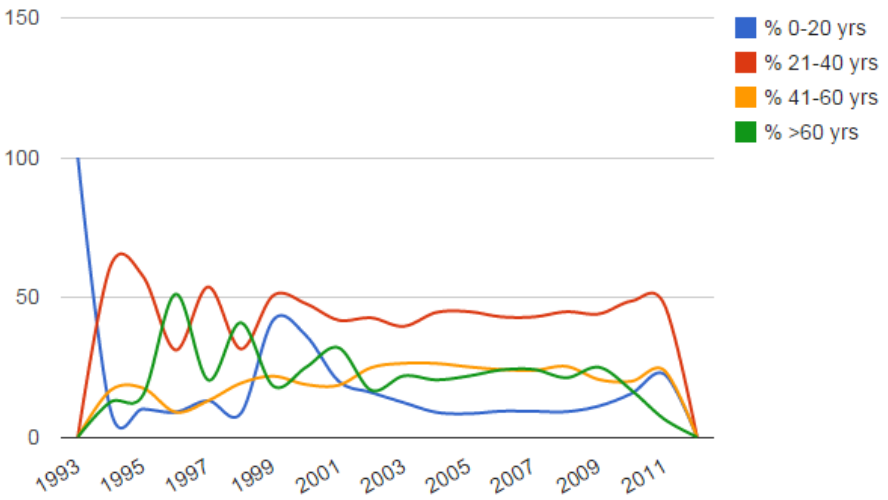


Figure 4 : Evolution globale des souches en fonction des groupes d'âge des patients.

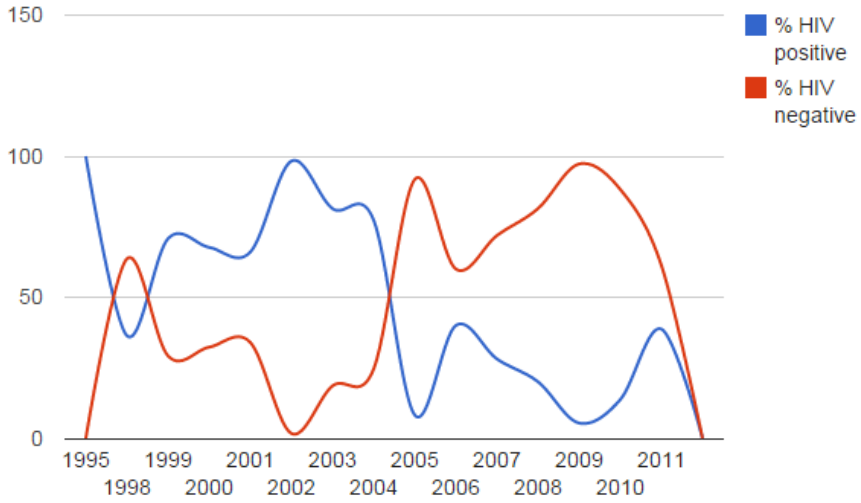


Figure 5 : Evolution globale des souches en fonction de la sérologie VIH/SIDA des patients.

Annexe 8 : Manuel d'utilisation de l'outil Web « SpolSimilaritySearch » (en anglais).

Home

Homepage (**Figure 1**) briefly introduces the SpolSimilaritySearch Web tool which is dedicated to study similarities of all spoligotyping patterns recorded in the SITVIT2 database. The main functionalities of the Website are shown in figures below.

Figure 1: Homepage and brief description of the main functionality.

Select the 43 spacers of your spoligotype

This section is the main functionality of this tool. It allows users to select each of the 43 spoligotypes separately in order to catch thanks to Regular Expresion or REGEXP (http://en.wikipedia.org/wiki/Regular_expression), the more similar patterns (in function of query) contained in SITVIT2 database. A spacer value can be either an absence (□), either a presence (■), or either an undetermined value (-). Note that an undetermined value could be an absence or a presence of spacer.

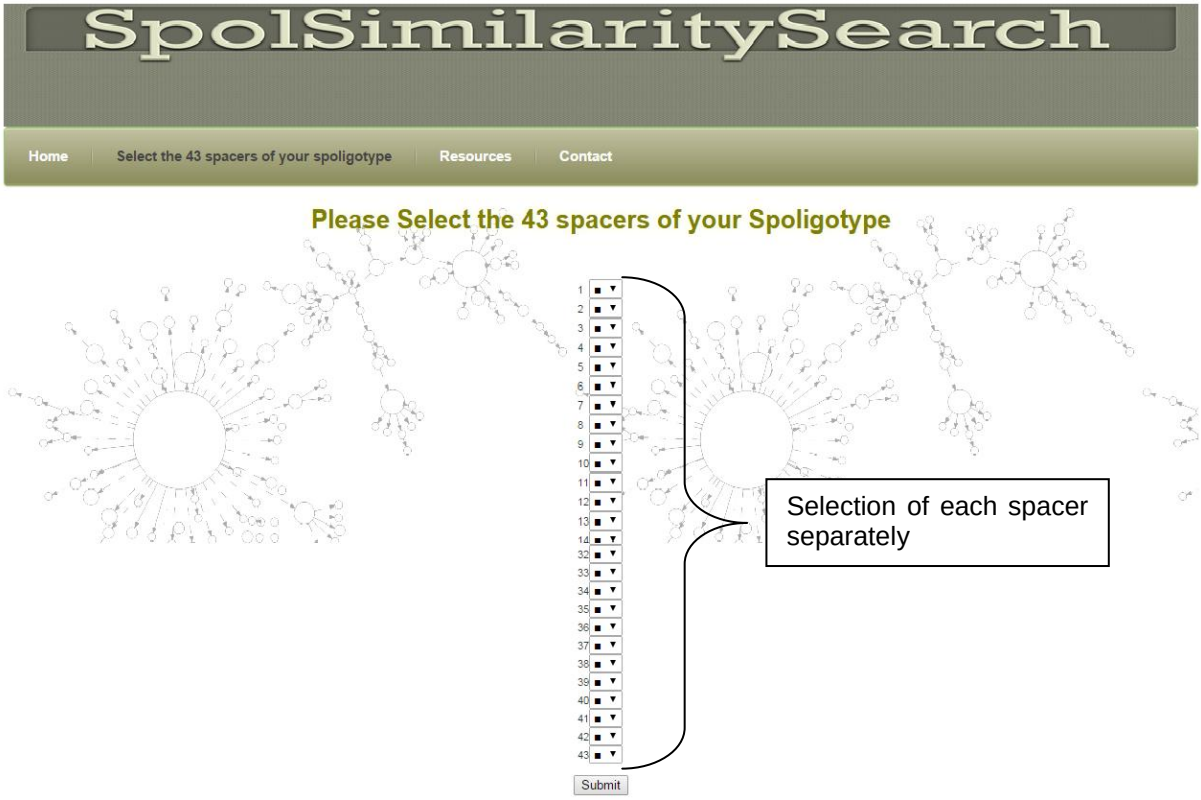


Figure 2: Selection of each spoligotyping spacer for comparison.

After submitting the spoligotype

After submitting the spoligotype profile composed of all the 43 spacers, we obtained the "**Results**" page containing all similar spoligotypes (as compared to SITVIT2 database) in function of the pattern that has been entered. Results also contain the Country Distribution of all similar spoligotypes. Countries were identified by ISO alpha-2 codes (http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2), and distribution was provided in the decreasing order.

For example, when querying for spoligotypes having spacers 33-36 missing and spacers 38, 39, and 40 undetermined (i.e. ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■---■■■), we obtain the following Results (**Figure 3**):

SpolSimilaritySearch

Home | Select the 43 spacers of your Spoligotype | Resources | Contact

Results (Similar spoligotypes)

You have entered the following pattern: ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□...■■■■ ; Regular Expression of Octal form of this spoligotype: 777777777760[0-7]71

Spoligotype (Binary)	Spoligotype (Octal)	Lineage	SIT	Number of Strains	Country Distribution
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777777777760771	T1	53	6152	US=831, FR=496, IT=335, BR=322, ZA=305, TR=218, AT=215, CN=194, MX=179, PE=171, SA=157, ET=154, ES=145, SE=127, BE=119, BG=98, NL=97, MA=96, IN=92, RU=90, GF=82, GB=75, JP=70, PL=66, AR=64, EG=64, DE=56, CO=53, FI=49, VE=49, TN=46, GY=45, AL=43, GR=41, DK=41, ZM=38, GP=37, BD=34, CZ=32, GM=31, LV=30, CI=27, PT=27, ZW=27, HT=27, MG=27, MQ=24, CU=23, YE=22, OM=22, CM=21, SN=21, IR=21, JM=17, AU=17, CA=16, ID=16, HN=16, NG=16, SR=15, NZ=14, MY=14, VN=14, KZ=12, GE=12, MW=12, TH=11, PS=11, PK=11, MZ=11, GW=10, TZ=10, IQ=9, EE=9, BS=8, HU=8, AM=7, NA=7, SD=6, BB=6, PY=6, PA=6, LC=5, KG=5, LY=5, NO=4, DO=4, DZ=4, CF=3, MM=3, LK=2, KE=2, BZ=2, NP=2, TW=2, KM=1, TC=1, AZ=1, RO=1, RE=1, TT=1
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□□■■■■	777777777760471	T1	65	106	BR=41, US=13, HT=12, FR=9, ES=5, AT=4, CM=3, DE=2, BE=2, GP=2, SE=2, CZ=2, MA=1, RU=1, DZ=1, MQ=1, IT=1, KZ=1, GF=1, ZA=1, JP=1
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□■■■■■■	777777777760371	T	240	31	BR=11, US=10, IT=2, DE=2, TW=1, CN=1, CU=1, GF=1, SE=1, AT=1
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■□■■■■	777777777760671	T1	245	20	BR=3, ZM=2, FR=2, IT=2, US=2, EG=1, ZW=1, NL=1, ES=1, VE=1, TN=1, GF=1, AR=1, AT=1
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■	777777777760571	T1	520	20	CZ=5, SA=2, US=2, CN=1, DE=1, PK=1, RU=1, FR=1, TR=1, MQ=1, IT=1, IQ=1, SE=1, AT=1
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□■■■■	777777777760071	T1	787	6	MX=2, US=2, FR=1, AT=1
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■■■■	777777777760171	T1	2366	5	US=2, CN=1, AR=1, LV=1

Number of lines: 7; Total number of strains: 6340

Copyright © 2015, Institut Pasteur de la Guadeloupe. Portions of this page are reproduced from work created and shared by Google and used according to terms described in the Creative Commons 3.0 Attribution License.

Figure 3: Snapshot of 'Results' obtained in function of our selection or query.

Note that in this Result section dots (.) are equivalent to dashes (-) in the Binary form of spoligotype.

Resources

This section provides a set of interesting Web resources or links related to analysis of *Mycobacterium tuberculosis* complex (MTBC).

Contact

This section allows to contact author and supervisor of this tool, and briefly introduces the team working at the lab: "Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de la Guadeloupe".

Annexe 9 : Quelques captures d'écran de l'outil Web « SpolLineages », permettant d'assigner une lignée phylogénétique à des spoligotypes.

Figure 1 : Page d'accueil de l'interface web.

Figure 2 : Page de sélection du ou des profils de spoligotypage.

Figure 3 : Exemple de résultat d'identification d'un profil de spoligotypage similaire à la famille Beijing.

ARTICLES

Article 01

Infection, Genetics and Evolution 12 (2012) 755–766

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology

Christophe Demay [a,1], Benjamin Liens [a,1], Thomas Burguière [a], Véronique Hill [a], David Couvin [a], Julie Millet [a], Igor Mokrousov [b], Christophe Sola [a,2], Thierry Zozio [a], Nalin Rastogi [a,*]

[a] *WHO Supranational TB Reference Laboratory, Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Abymes Cedex, Guadeloupe*
[b] *Laboratory of Molecular Microbiology, St. Petersburg Pasteur Institute, St. Petersburg, Russia*

ARTICLE INFO

Article history:
Available online 17 February 2012

Keywords:
Mycobacterium tuberculosis
Database
Genetic diversity
Epidemiology
Spoligotyping
MIRU–VNTRs

ABSTRACT

Among various genotyping methods to study *Mycobacterium tuberculosis* complex (MTC) genotypic polymorphism, spoligotyping and mycobacterial interspersed repetitive units–variable number of DNA tandem repeats (MIRU–VNTRs) have recently gained international approval as robust, fast, and reproducible typing methods generating data in a portable format. Spoligotyping constituted the backbone of a publicly available database SpolDB4 released in 2006; nonetheless this method possesses a low discriminatory power when used alone and should be ideally used in conjunction with a second typing method such as MIRU–VNTRs for high-resolution epidemiological studies. We hereby describe a publicly available international database named SITVITWEB which incorporates such multimarker data allowing to have a global vision of MTC genetic diversity worldwide based on 62,582 clinical isolates corresponding to 153 countries of patient origin (105 countries of isolation). We report a total of 7105 spoligotype patterns (corresponding to 58,180 clinical isolates) – grouped into 2740 shared-types or spoligotype international types (SIT) containing 53,816 clinical isolates and 4364 orphan patterns. Interestingly, only 7% of the MTC isolates worldwide were orphans whereas more than half of SITed isolates (n = 27,059) were restricted to only 24 most prevalent SITs. The database also contains a total of 2379 MIRU patterns (from 8161 clinical isolates) from 87 countries of patient origin (35 countries of isolation); these were grouped in 847 shared-types or MIRU international types (MIT) containing 6626 isolates and 1533 orphan patterns. Lastly, data on 5-locus exact tandem repeats (ETRs) were available on 4626 isolates from 59 countries of patient origin (22 countries of isolation); a total of 458 different VNTR patterns were observed – split into 245 shared-types or VNTR International Types (VIT) containing 4413 isolates) and 213 orphan patterns. Datamining of SITVITWEB further allowed to update rules defining MTC genotypic lineages as well to have a new insight into MTC population structure and worldwide distribution at country, sub-regional and continental levels. At evolutionary level, the data compiled may be useful to distinguish the occasional convergent evolution of genotypes versus specific evolution of sublineages essentially influenced by adaptation to the host. This database is publicly available at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Although tuberculosis (TB) is a preventable and treatable disease, with an estimated 9.2 million new cases and two million deaths worldwide (WHO Report, 2008), it remains one of the leading infectious diseases worldwide. As a result of inadequate treatment, the proportion of patients with multiple-drug resistant TB (MDR-TB) is constantly increasing, and the extensively drug-resistant TB (XDR-TB) has become a new global threat, particularly for TB/HIV co-infected patients. Lack of compliance during prior treatment appears as the single largest selection factor among patient populations favoring the emergence of MDR-TB, which if not taken seriously and tackled urgently – will contribute to a pool from which higher proportions of XDR-TB bacilli will emerge, making the disease unmanageable. International TB control urgently needs a concerted and coordinated response to monitor and assess global TB spread, drug-resistance surveillance and populations at risk. This is only possible through an integrated approach based on networking of laboratories, TB control programs, and application of new molecular typing techniques to identify, tack and report the circulating and emerging clones of tubercle bacilli worldwide (García de Viedma et al., 2011).

* Corresponding author.
E-mail address: nrastogi@pasteur-guadeloupe.fr (N. Rastogi).
[1] Christophe Demay and Benjamin Liens contributed equally to this work.
[2] Current address: Institut de Génétique et Microbiologie, Orsay, France.

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2012.02.004

Most of the earlier molecular epidemiological studies were performed using IS*6110*-RFLP, considered a gold-standard due to its reproducibility, discriminatory power, and low cost (van Embden et al., 1993). However, this labor-intensive methodology was replaced with alternative PCR-based strategies in the last decade, creating the basis for large-scale, high-throughput genotyping of *Mycobacterium tuberculosis*, well adapted for epidemiological investigations (Rastogi and Sola, 2007). Integration of new genotyping tools – spoligotyping based on the polymorphism of the direct repeat (DR) locus (Kamerbeek et al., 1997), and molecular typing using MIRU–VNTRs minisatellites (Supply et al., 2001, 2006) – has paved the way for a novel approach to understand the distribution of TB and modes of its transmission in the population. The MIRU–VNTR minisatellites constitute a multi-locus marker set since they represent independent markers of a same type (classical 12-locus, discriminatory 15-locus, or full 24-locus formats according to Supply et al. (2001, 2006)). In conjunction with classical epidemiological investigations, these typing methods constitute the new gold-standard for optimal TB surveillance at national, regional, and global levels by "universal genotyping" of patient isolates (Allix-Beguec et al., 2008; Vanhomwegen et al., 2011). At the micro-population level, this strategy allows to determine whether recurrences in a patient are due to reactivation, exogenous reinfection or due to polyclonal infection; and to identify recent transmission versus reactivation, whereas at the macro-population level, it allows to define the global distribution of TB genotypic lineages, to monitor the international spread of high-risk isolates, as well as to explore the evolutionary features of *M. tuberculosis* (García de Viedma et al., 2011).

Publicly available databases such as SpolDB4 (Brudey et al., 2006) and MIRU–VNTRplus (Allix-Beguec et al., 2008) today allow comparing this ever-increasing amount of TB fingerprinting data, and have provided with a new conceptual framework to study global TB epidemiology, and human demographical and TB co-evolutionary history (Shabbeer et al., 2011). The purpose of this paper is to describe one of the biggest international databases SITVITWEB with available genotyping information on 62,582 MTC clinical isolates from 153 countries of patient origin, compiled at Institut Pasteur de la Guadeloupe. It is a significant improvement as compared to the previous version of the SpolDB4 database released 6 years ago (Brudey et al., 2006), which provided only spoligotyping information on 39,295 isolates from 122 countries. By publicly releasing the multimarker SITVITWEB database with appropriate datamining tools, we hope to serve the global research community with a concerted and coordinated response to monitor and assess global TB spread.

2. Material and methods

2.1. Genotyping markers

The genotyping data included in the SITVITWEB were generated using one of the three PCR-based methods that are widely used internationally for *M. tuberculosis* molecular. Each method was applied by investigators using established methodology, namely: (i) spoligotyping which studies variations in the structure of the direct repeat (DR) locus, best translated in a format that reports presence or absence of 43 inter-DR spacers (Kamerbeek et al., 1997); (ii) MIRU–VNTR that describes variations in copy numbers of 40–100 bp minisatellite elements (used in its classical 12-locus format; Supply et al., 2000, 2001); and (iii) 5-locus exact tandem repeats, ETR-A to E (Frothingham and Meeker-O'Connell, 1998). In this paper, the terms Spoligotype International Type (SIT), MIRU International Type (MIT), and VNTR International Type (VIT) designate spoligotypes, 12-locus MIRUs, and 5-locus ETRs respectively, that are shared by two or more patient isolates, as opposed to an "orphan" which designates a pattern reported for a single isolate among the pool of the 62,582 isolates studied worldwide (see below). Note that the suffix "IT" is used throughout this paper to denote clustered isolates in contrast to unique or orphan patterns. Hence SITed, MITed and VITed designations refer to clustered isolates (previously termed as shared-types or STs in SpolDB4 database; Brudey et al., 2006). The term "endemic pattern" designates a pattern found in a specific geographical sub-region(s) as opposed to "ubiquitous" that denotes commonly found *M. tuberculosis* clones without any given geographical specificity. Lastly, terms "clades" and "lineages" are synonymous and define a group of the related genotypes that were defined using slightly modified SpolDB4 rules (Brudey et al., 2006).

2.2. Data collection and SITVITWEB online access tool

The data on a total of 62,582 *M. tuberculosis* complex isolates were either obtained at the Pasteur Institute of Guadeloupe, received from co-investigators and collaborating laboratories, or retrieved from published studies (n = 58,180 spoligotype, n = 4626 5-locus ETRs, n = 8178 12-locus MIRUs), and compiled in an MySQL database (origin of isolates and investigators are available at http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE). SITVITWEB access tool uses a n-tier architecture to provide the best access performance. The main part of this architecture is composed by a MySQL database. In a central part, an Apache Tomcat application server dynamically generates hypertext markup language pages. These pages are the final part of SITVITWEB application, and form the graphical user interface (GUI); the latter is implemented with: Java technology (JavaServer Page, JavaServlet API, Asynchronous JavaScript) and appropriate scripts (dhtmlxGrid, Ajax API, Google Maps API, XML). That's why Javascript activation is recommended at user interface for optimal performance of SITVITWEB. This web application complies with specifications of the World Wide Web Consortium (W3C) and was tested successfully with most of the modern browsers.

2.3. Statistical and bioinformatical analysis

Hunter–Gaston Diversity Index (HGDI), was commonly used to compare typing methods and select the most discriminatory system (Hunter and Gaston, 1988). The definition of HGDI derives from Simpson's diversity index (Simpson, 1949), which varies from 0 (no diversity) to 1 (highest diversity). In the present study, HGDI was used to determine the diversity of each MIRU–VNTR locus in 10 lineages, and is describe by the following equation (where N is the total number of isolates in the sample population for a given locus, S is the total number of distinct repeat unit values identified for the locus and and n_j is the number of isolates having the jth value):

$$\mathrm{HGDI} = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{S} nj(nj-1)$$

Chi-square test for independence also named Pearson's chi-square (χ^2) test is a non-parametric hypothesis testing procedure used to determine whether modalities of two characters are independent. Here, we check if there is a link between geographical places and the distribution of IT or orphans. Lastly, the genetic diversity index H is defined by number of isolates in the different patterns; it was calculated using the following formula: $H = 2n(1 - \sum x_i^2)/(2n-1)$ where n represents the number of individuals and x_i is the frequency of the ith allele for a given marker (Armitage and Berry, 1987).

3. Results and Discussion

3.1. Global overview of genotyping markers

Containing 62,582 clinical isolates from 153 countries of patient origin (105 countries of isolation), and three different genotyping markers (43-spacer Spoligotype format, 12-locus MIRUs and 5-locus ETRs), SITVITWEB constitutes the largest publicly available *M. tuberculosis* complex multiple marker genotyping database. The overall genotyping information on all the clustered isolates by these three respective genotyping methods is summarized in Supplemental Tables S1–S3. These tables provide with detailed listings for each pattern, the total number of isolates corresponding to a given pattern, and their geographical distribution. The clade/subclade designations were also provided for the spoligotyping data according to revised SpolDB4 rules (to be discussed below). We report a total of 7105 spoligotype patterns (corresponding to 58,180 clinical isolates) – grouped into 2740 SITs containing 53,816 clinical isolates and 4364 orphan patterns (Supplemental Table S1). Interestingly, only 7% of the MTC isolates worldwide were orphans whereas more than half of SITed isolates (n = 27,059) were restricted to only 24 most prevalent SITs.

Regarding MIRU–VNTRs, a very useful database for online genotyping analysis of MTC isolates was recently released (Allix-Beguec et al., 2008), and is available online at a freely accessible web-based server (http://www.MIRU-VNTRplus.org). However, it provides information only on a well-characterized set of 186 reference isolates; hence although highly useful for phylogenetical evaluation of a given genotype, this application does not allow to compare genotyping results at a global level. Indeed, the information available is rather aimed to maximize the prediction of lineage for classification purposes than to describe the worldwide diversity of circulating MTC genotypes. In this context, the SITVITWEB application described here represents a 1st integration of worldwide 12-locus MIRU data in any publicly available database (Supplemental Table S2). It contains a total of 2379 MIRU patterns (from 8161 clinical isolates) from 87 countries of patient origin (35 countries of isolation); these were grouped in 847 MITs containing 6626 isolates and 1533 orphan patterns. Last but not least, as an extension to worldwide minisatellite typing data, we also decided to compile available data on 5-locus ETRs that represented the first widely applicable minisatellite typing scheme in TB molecular epidemiology (Frothingham and Meeker-O'Connell, 1998), prior to current use of 12-, 15- and 24-locus MIRU–VNTR typing (Supply et al., 2001, 2006; Mazars et al., 2001). SITVITWEB contains data on 5-locus ETRs on 4626 isolates from 59 countries of patient origin (22 countries of isolation, Supplemental Table S3); it corresponds to 458 individual patterns – 245 VITs containing 4413 isolates and 213 orphan patterns.

3.2. Description of predominant shared-types

As can be seen by the evolution of successive *M. tuberculosis* complex genotyping databases of Institut Pasteur de Guadeloupe, the global genotyping information included in SITVITWEB considerably increased over previous versions of the public databases (from SpolDB1 to SITVITWEB; Supplemental Fig. S1-A). The global genotyping information on 62,589 MTC isolates in function of the distribution of predominant SITs, MITs, and VITs is graphically represented in Fig. 1A to C. It shows the shared-types classified by decreasing number of isolates in each cluster; the top-24 SITs (n = 27,059) and top 63 MITs shown in Fig. 1A and B group 50% of SITed and MITed isolates, while the top 38 VITs shown in Fig. 1C group 80% of VITed strains. Note that the double-headed arrows in these figures show the proportion of isolates included in respective clusters.

3.2.1. Correction of spoligotype data duplications and possible errors

All orphan patterns were visually checked by the expert, and a confirmation was requested from the investigator for doubtful signatures; these essentially concerned patterns that did not follow established rules for any of the assigned lineages. A special attention was also paid to remove all duplications which essentially arose due to the same set of submissions from two different investigators overtime, and could lead to false clusters. As a result, almost one quarter of isolates previously included in SpolDB4 were corrected. In most cases, corrections in orphans resulted in increase of the number of isolates of the most frequent SITs, at the expense of other less-represented SITs. SITVITWEB allows any previous investigator(s) to double check their results as compared to SpolDB4 since all previous strain identification numbers were kept unaltered in the new system (a deleted strain being considered as nonexistent by the system). This problem did not concern the MIT and VIT designations since they were not present in previous versions of the database.

3.2.2. Comparison of genetic diversity of different markers

The genetic diversity of different markers is summarized in Supplemental Fig. S1-A. A comparison of genetic diversity of spoligotyping in SpolDB4 versus SITVITWEB results in a similar diversity index (H = 0.98), showing that the overall diversity did not change in SITVITWEB despite the inclusion of new isolates. This can be explained by the fact that the orphan strain slightly decreased in SITVITWEB and most of the patterns were labeled with existing shared-types. A slightly higher genetic diversity index was observed for 12-locus MIRUs (H = 0.99) as opposed to 5-locus ETRs (H = 0.94).

3.2.3. Comparison of SITed isolates in SITVITWEB versus SpolDB4

Compared to the previous SpolDB4 version, SITVITWEB today reports 801 new SITs which brings the percentage of SITed isolates (n = 53,816) to 93% versus 91% in SpolDB4. Although none of the top-24 SITs was newly created, the description of *M. tuberculosis* diversity has changed both quantitatively (20 SITs accounted for 50% of the SITed isolates in SpolDB4) and qualitatively (the order of the first top-24 SIT has changed and only 18 major SIT are co-shared in previous and new versions). This change is due to increase in both amount and diversity of the newly integrated information as compared to previous SpolDB3 and SpolDB4 versions (Filliol et al., 2003; Brudey et al., 2006); 2 SITs present in SpolDB3 and SpolDB4 top-20 (SIT139 and SIT48), disappeared from current top-24 while six SITs (SIT26, SIT34, SIT17, SIT679, SIT59, SIT92) entered among the top-24 in SITVITWEB. It might be recalled that entry of these six SITs concern preexisting patterns in the database that simply joined the list of 24 most predominant patterns in SITVITWEB. Moreover, only the first three SITs (SIT1, SIT53, SIT50) of the top-24 retained their positions (Fig. 1A). Interestingly, 48.3% (8644/17,891) of newly added SITed isolates are found in the top-24 SITs. Last but not least, the number of orphan isolates in SIVITWEB (n = 4364) decreased to 7.5% of all spoligotyped strains (versus 8.6% in SpolDB4; Supplemental Fig. S1-A) showing that most of the newly spoligotyped isolates belonged to a known SIT rather than an orphan. The observation that almost half of the newly added SITed strains were found in the top-24 SITs (corresponding to SITs already present in SpolDB4) and the slightly decreasing proportion of orphan strains in SIVITWEB compared to SpolDB4 suggests a good coverage of global MTC genotypic diversity in SpolDB4. Nonetheless, both the quality of data and its representativity were improved in SITVITWEB by making corrections to existing data taken from SpolDB4, and also by applying

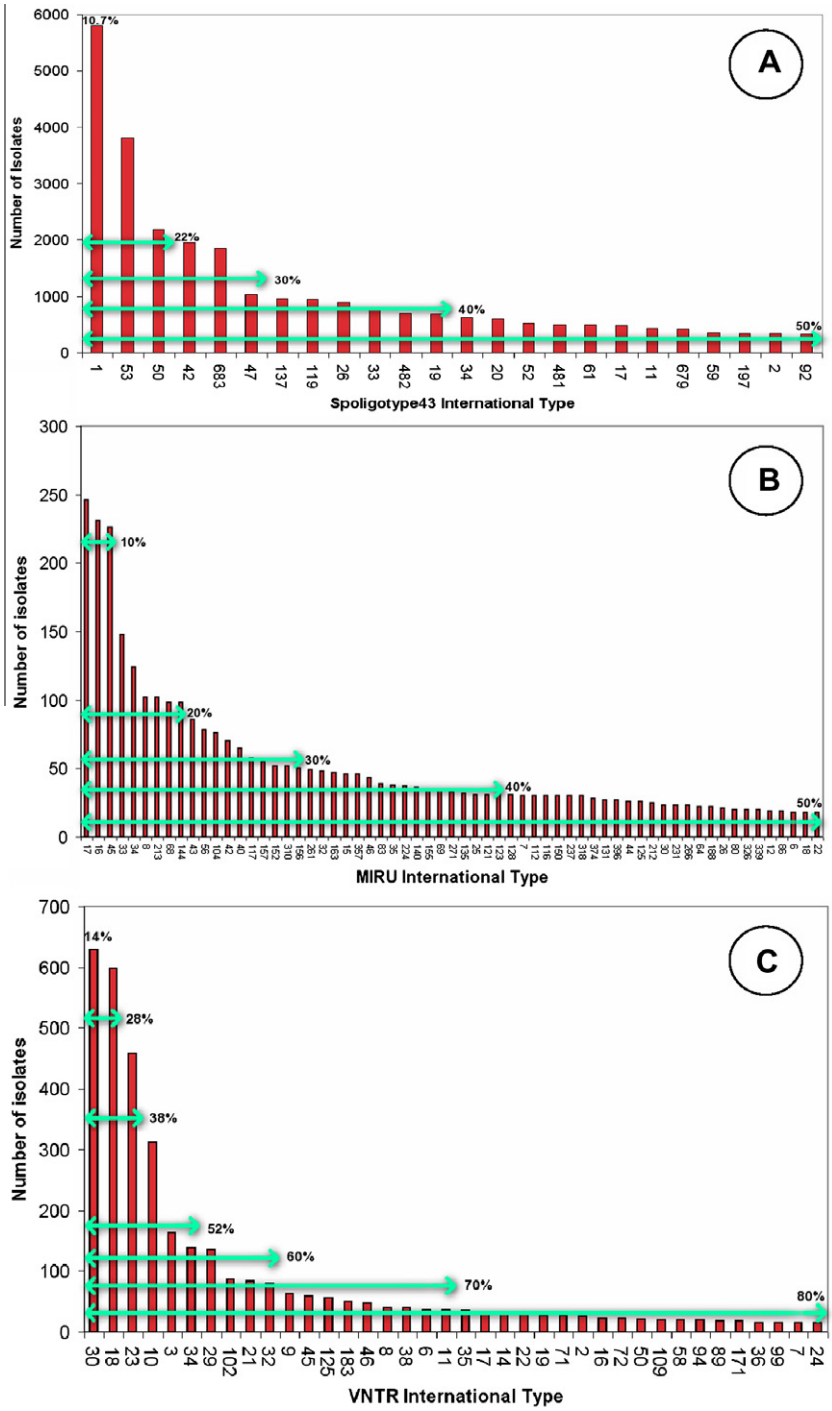


Fig. 1. Distribution of predominant SITs (A), MITs (B) and VITs (C) classified by decreasing number of isolates in the SITVITWEB in each cluster. The top-24 SITs and top 63 MITs shown in Fig. 1A and B group 50% of SITed and MITed strains, while the top 38 VITs shown in Fig. 1C group 80% of VITed strains. Double-headed arrows show the cumulated proportion of isolates included in clusters.

a more stringent procedure for inclusion of new data in SITVITWEB.

3.2.4. Proportion of MITed isolates in SITVITWEB

Regarding the level of clustering, 63 MITs included 50% of the MITed isolates (an average of seven strains per MIT; Fig. 1B). These 63 MIT included 3329 isolates (40.8% of MIRU-typed strains). Three MITs (MIT17, MIT16 and MIT45 were highly predominant and encompassed > 10% of the MITed isolates. With 19% of orphan isolates (1531/8161 strains), MIRU12 marker is marked by the highest orphan rate in SITVITWEB compared to spoligotyping and VNTR markers (Supplemental Fig. S1-A). Regarding 5-locus ETRs, SITVITWEB contains limited information for the time being; nonetheless, >50% of typed isolates were limited to only six VITs (VIT30, VIT18, VIT23, VIT10 VIT3, and VIT34). The most represented shared-types – VIT30, VIT18 included 14% the VNTR-typed isolates each, followed by VIT23 at 10% (Fig. 1C). Lastly, SITVITWEB contained 4285 clinical isolates for which both spoligotyping and MIRU information was available; these represent over 50% of the MIRU-typed strains and >7% of spoligotyped strains. A combination

of the two genotyping markers did not influence a hierarchy of the most represented SITs in the database since almost half (n = 2079) of the isolates with both MIRU and spoligotype information available remained in the top-24 SITs.

3.3. Description and classification of genotypic lineages

3.3.1. Description of spoligotyping-based lineages

As previously for SpolDB4 (Brudey et al., 2006), a spoligotyping based lineage rule in SITVITWEB is defined as a prototype pattern in which the presence/absence of 43 spacers is symbolized by empty or full boxes (Fig. 2). An important feature of SITVITWEB is the transcription and validation of new stringent rules based on 43 binary characters, instead of SpolDB4 rules based on octal codes that assigned a specific octal character to blocks of three spacers. As summarized in Fig. 2, the absence/presence of 43 individual spacers in the new representation is represented by empty/full boxes; black color (spacers whose absence/presence is not essential to define a lineage), red color (spacers whose absence/presence is essential to define a lineage). Since changes in the DR locus are most likely to be neutral that occur via consecutive irreversible deletions of either single or contiguous blocks (van Embden et al., 2000); all stringent spacer absence is considered essential and illustrated by empty red-colored boxes. Note that presence of some spacers, although in a significantly smaller number (e.g. presence of spacer 31 for LAM, X and T lineages as opposed to its absence in Haarlem lineage), was considered essential to develop new rules, and is shown by red-colored full boxes (Fig. 2). Thanks to a subsequent bioinformatical program, an automatic classification procedure was developed in several steps: (i) classification of a pool of isolates, (ii) reclassification of ill-classified or unclassified isolates with redefined expert-based rules to increase classification rate, (iii) blind automated retesting of new pools. Note that all unclassified/ill-classified orphan patterns were ignored during the step 2 in order to keep a sufficiently high stringency level. A summary of the bioinformatical classification approach is provided in Supplemental Fig. S1-B, which compares a new pattern to all sublineages rules followed by its classification in a well-defined sublineage or alternatively as ambiguous.

These new rules were applied to classify each single spoligotype pattern in SITVITWEB to assign a first selection of 68 set of rules. In addition to 62 prototypes of spoligotype lineages/sublineage previously described, we added five "new rules" for definition of variants within existing lineages (AFRI, BOV, CAS, EAI, and PINI), whereas one rule that concerned "all 43 spacers absent" was not given with any genotypic lineage, and is not represented in Fig. 2. These rules allowed to classify almost 3/4 of the SITs corresponding to 95% of spoligotyped isolates (as well as >70% of orphan strains) in the database. A careful expert-based examination using these rules revealed ambiguities for 97 SITs in the previous SpolDB4 version of the database (Supplemental Table S4).

In summary, SITVITWEB classification rules are based on the visual inspection of SpolDB4 rules that were converted from octal to binary interpretations, with five "pseudo-rules" that interpret consensus definitions for five lineages, i.e., AFRI, BOV, CAS, EAI, and PINI. This was done both to avoid sub-lineage names such as CAS1-like, AFRI-like etc., and also as a way to classify many of the ill-classified isolates. Although these rules were translated stringently for a given pattern for automated classification in SITVITWEB, they do not define a prototype per se. The validity of these rules, either for classifying a pool of isolates or for reclassifying ill-classified and/or unclassified isolates should be reinvestigated using other markers in future studies.

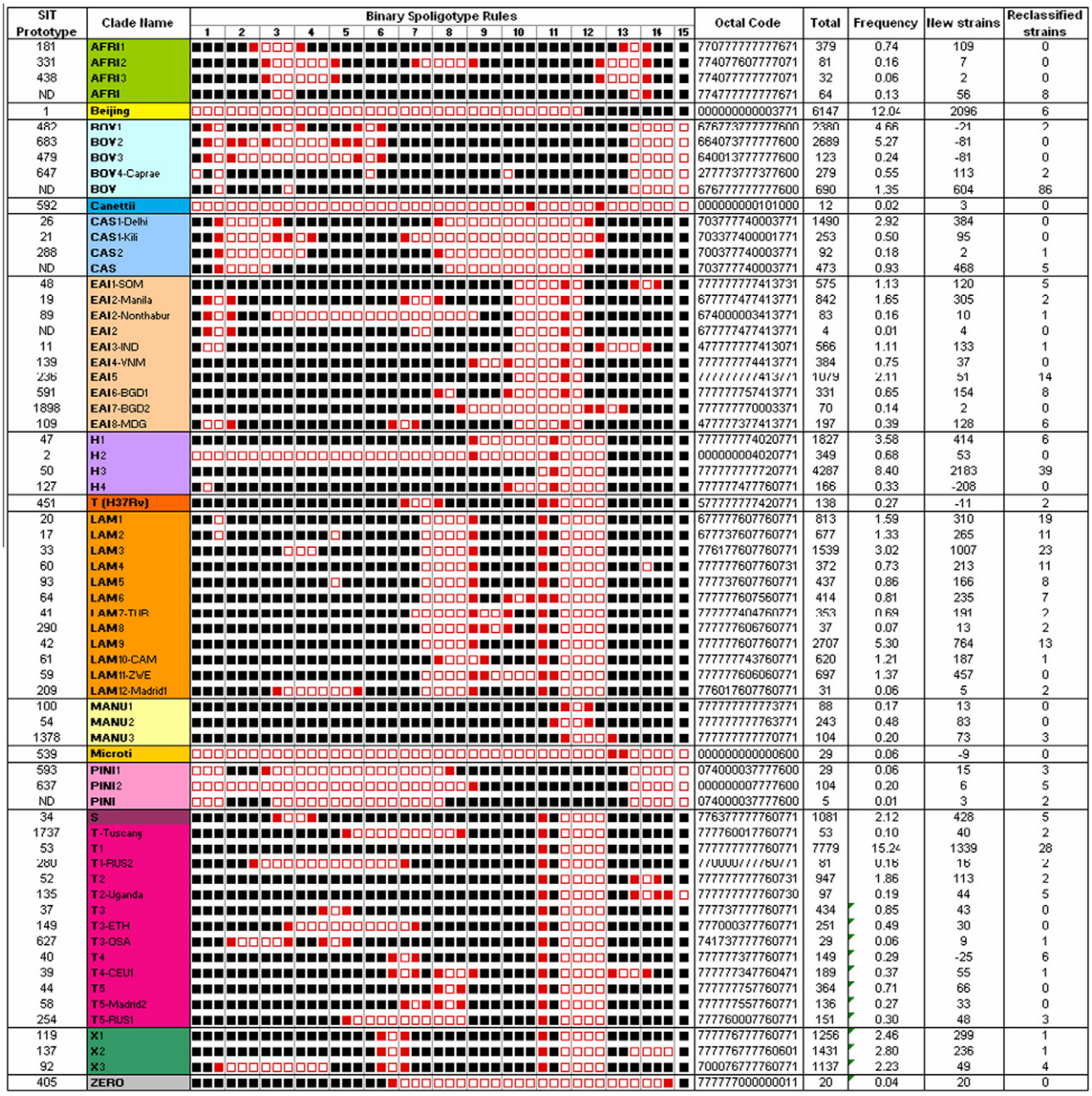
3.3.2. Description of minisatellite-based lineages

Regarding minisatellite-based lineage classification of MTC isolates, a web-based server (http://www.MIRU-VNTRplus.org) provides detailed information on geographical origin, corresponding genetic lineage, IS*6110*-RFLP, 24-locus MIRU–VNTR, spoligotyping, SNPs, and LSPs profiles for a well-characterized set of 186 reference isolates (Allix-Béguec et al., 2008). In this work, the authors have extensively described and tested an algorithm based on best-match analysis followed by tree-based analysis on MIRU–VNTR data (combined or not with spoligotyping data) to describe distribution of isolates with minisatellite data among the various spoligotype families. Hence, although there is no description of minisatellite-based lineages, this approach does allow drawing conclusions based on spoligotype-based classification, which is nonetheless known to be prone to significant limitations (Comas et al., 2009). In the present investigation, we focused on a subset of 4285 isolates for which both 12-locus MIRU and spoligotyping data were available in SITVITWEB. In this dataset, MIRU patterns associated with unclassified/ambiguous spoligotype patterns represented only 3.4% of the MIRU-typed isolates (n = 246 unclassified patterns, and n = 30 ambiguous patterns). The remaining 4009 classified isolates were split in 708 orphan MIRU patterns and 3301 MITed strains; the latter included 3188 (96.6%) SITed isolates and 113 (3.4%) orphan spoligotyped strains. Hence, lineages are more frequently attributed to MIRU patterns when the associated spoligotype pattern is SITed rather than orphan. For example, almost 98% of T, Beijing and LAM lineage isolates could be defined thanks to their association with SITed patterns (their total number represented 58.1% of all classified MITed strains in SITVITWEB).

Since typing based on 5-locus ETRs (ETR-A to E) preceded the 12-MIRU format (Frothingham and Meeker-O'Connell, 1998), we also compiled available information in the same format for 5-ETRs in the SITVITWEB. This information was available on a total of 4626 isolates; 2222 (48%) with spoligotyping data out of which 2098 (or 45.3% of all VNTR-typed strains) were classified in spoligotyping-defined lineages. Among these 2098 duly-classified isolates, 1971 were assigned with VIT designations (VITed) and 127 as orphans. The 1971 VITed isolates included 1883 (95.5%) SITed strains and 88 (4.5%) spoligotyping-based orphans. The T, LAM and H lineages represented 68.6% of the classified VITed isolates, 95.7% of these were defined thanks to the availability of the spoligotyping information. For information purpose, 2404 (52%) ETR-typed isolates were without spoligotyping data (257 VNTR patterns containing 2326 VITed strains and 78 orphans).

3.3.3. Allelic diversity of minisatellites in relation to global sample versus lineages

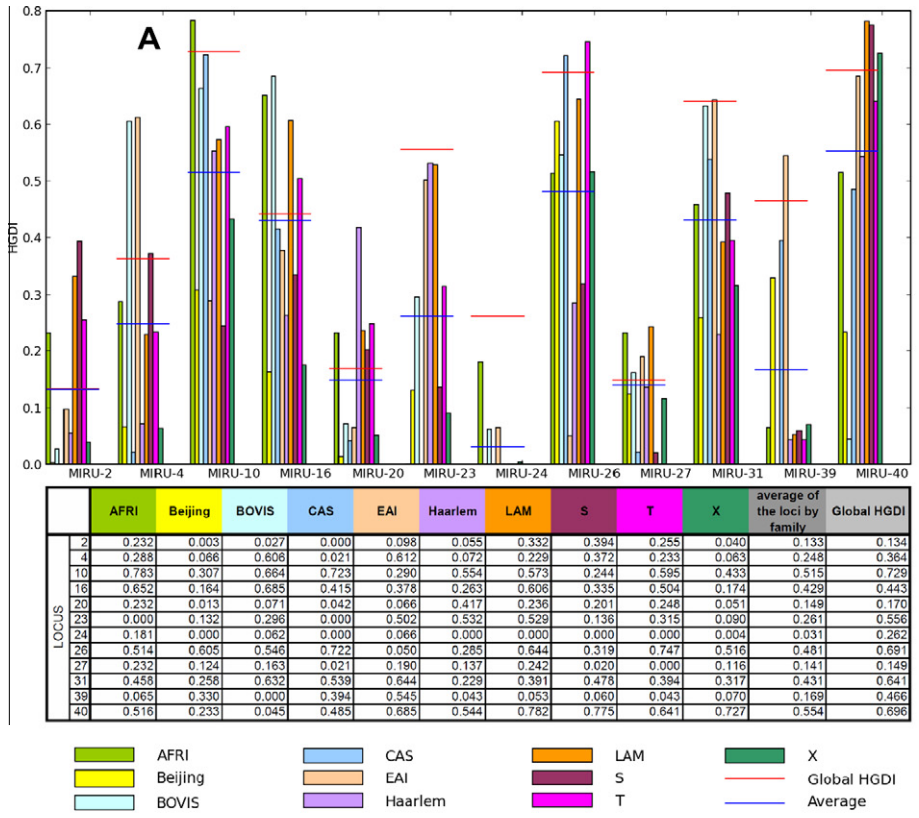
The Hunter–Gaston Diversity Index (HGDI) of minisatellite locus among 10 major spoligotyping-based lineages (Africanum, Beijing, Bovis, CAS, EAI, Haarlem, LAM, S, T, and X) of *M. tuberculosis* complex was calculated both for 12-locus MIRU and 5-locus ETR formats, and is summarized in Fig. 3. Regarding the ill-defined T superfamily that was defined by default in SpolDB4 (Brudey et al., 2006), and whose strains are found scattered among various branches of the Euro-American lineage based on LSPs (Gagneux et al., 2006; Gagneux and Small, 2007) or SNPs (Filliol et al., 2006), we only retained sublineages with known phylogeographical specificity, i.e., T-Tuscany, T1-RUS2 (Russia), T2-Uganda, T3-ETH (Ethiopia), T3-OSA (Osaka), T4-CEU1 (Central Europe), T5-Madrid2. All other T1 to T5 spoligo-prototypes that are not monophyletic, were not retained for calculating allelic diversity. Thus, the sample taken to make these calculations contained 3178 MIRU- and spoligotyped isolates in Fig. 3A (12-locus MIRUs), and 1510 VNTR- and spoligotyped isolates in Fig. 3B (5-locus ETRs). In the graphical representation, the red horizontal lines designate the global HGDI values obtained for each locus versus blue horizontal



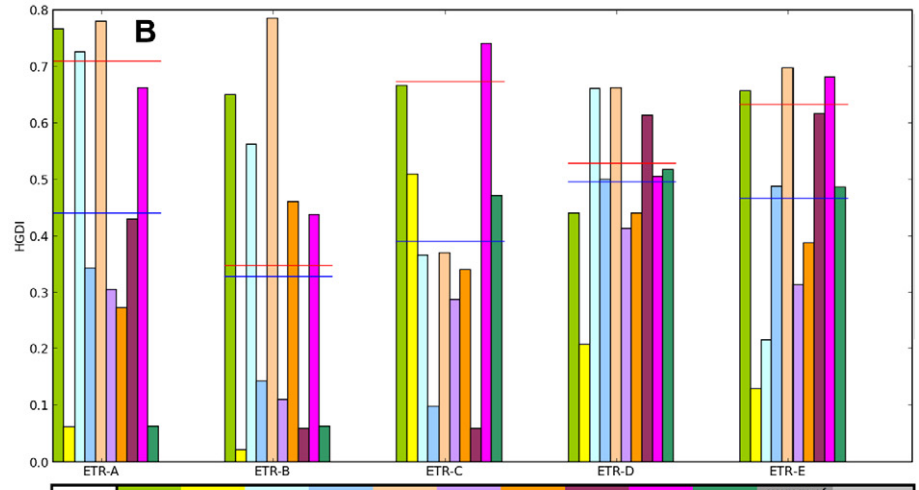
SIT Prototype	Clade Name	Binary Spoligotype Rules 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	Octal Code	Total	Frequency	New strains	Reclassified strains
181	AFRI1		770777777777671	379	0.74	109	0
331	AFRI2		774077607777071	81	0.16	7	0
438	AFRI3		774077777777071	32	0.06	2	0
ND	AFRI		774777777777671	64	0.13	56	8
1	Beijing		000000000003771	6147	12.04	2096	6
482	BOV1		676773777777600	2380	4.66	-21	2
683	BOV2		664073777777600	2689	5.27	-81	0
479	BOV3		640013777777600	123	0.24	-81	0
647	BOV4-Caprae		277773777377600	279	0.55	113	2
ND	BOV		676777777777600	690	1.35	604	86
592	Canettii		000000000101000	12	0.02	3	0
26	CAS1-Delhi		703777740003771	1490	2.92	384	0
21	CAS1-Kili		703377400001771	253	0.50	95	0
288	CAS2		700377740003771	92	0.18	2	1
ND	CAS		703777740003771	473	0.93	468	5
48	EAI1-SOM		777777777413731	575	1.13	120	5
19	EAI2-Manila		677777477413771	842	1.65	305	2
89	EAI2-Nonthabur		674000003413771	83	0.16	10	1
ND	EAI2		677777477413771	4	0.01	4	0
11	EAI3-IND		477777777413071	566	1.11	133	1
139	EAI4-VNM		777777774413771	384	0.75	37	0
236	EAI5		777777774134771	1079	2.11	51	14
591	EAI6-BGD1		777777757413771	331	0.65	154	8
1898	EAI7-BGD2		777777770003371	70	0.14	2	0
109	EAI8-MDG		477777377413771	197	0.39	128	6
47	H1		777777774020771	1827	3.58	414	6
2	H2		000000004020771	349	0.68	53	0
50	H3		777777777720771	4287	8.40	2183	39
127	H4		777777477760771	166	0.33	-208	0
451	T (H37Rv)		577777777420771	138	0.27	-11	2
20	LAM1		677777607760771	813	1.59	310	19
17	LAM2		677737607760771	677	1.33	265	11
33	LAM3		776177607760771	1539	3.02	1007	23
60	LAM4		777777607760731	372	0.73	213	11
93	LAM5		777737607760771	437	0.86	166	8
64	LAM6		777777607560771	414	0.81	235	7
41	LAM7-TUR		777777404760771	353	0.69	191	2
290	LAM8		777777606760771	37	0.07	13	2
42	LAM9		777777607760771	2707	5.30	764	13
61	LAM10-CAM		777777743760771	620	1.21	187	1
59	LAM11-ZWE		777777606060771	697	1.37	457	0
209	LAM12-Madrid1		776017607760771	31	0.06	5	2
100	MANU1		777777777773771	88	0.17	13	0
54	MANU2		777777777763771	243	0.48	83	0
1378	MANU3		777777777770771	104	0.20	73	3
539	Microti		000000000000600	29	0.06	-9	0
593	PINI1		074000037777600	29	0.06	15	3
637	PINI2		000000007777600	104	0.20	6	5
ND	PINI		074000037777600	5	0.01	3	2
34	S		776377777760771	1081	2.12	428	5
1737	T-Tuscany		777760017760771	53	0.10	40	2
53	T1		777777777760771	7779	15.24	1339	28
280	T1-RUS2		770000777760771	81	0.16	16	2
52	T2		777777777760731	947	1.86	113	2
135	T2-Uganda		777777777760730	97	0.19	44	5
37	T3		777737777760771	434	0.85	43	0
149	T3-ETH		777000377760771	251	0.49	30	0
627	T3-OSA		741737777760771	29	0.06	9	1
40	T4		777777377760771	149	0.29	-25	6
39	T4-CEU1		777777347760471	189	0.37	55	1
44	T5		777777757760771	364	0.71	66	0
58	T5-Madrid2		777777557760771	136	0.27	33	0
254	T5-RUS1		777760007760771	151	0.30	48	3
119	X1		777776777760771	1256	2.46	299	1
137	X2		777776777760601	1431	2.80	236	1
92	X3		700076777760771	1137	2.23	49	4
405	ZERO		777777000000011	20	0.04	20	0

Fig. 2. Bioinformatical classification analysis of SITed strains in SITVITWEB. 1st column, typical SIT pattern (prototype) for the lineage/sublineage (ND, not defined since no typical spoligotype pattern could be assigned to these existing patterns, considered pseudo rules for the time-being); 2nd column, sublineage name (with lineages shown in different color schemes that are same as in detailed Supplemental Table S1); 3rd column, binary spoligotype rules (black spacers do not influence sublineage matching, red spacers define lineages/ sublineages), 4th column, octal representation of binary spoligotype patterns; 5th column, total number of a given SIT; 6th column, frequency (%) of a given SIT as compared to the total number of SITed strains; 7th column: number of newly added strains as compared to the previous public database SpolDB4 (note that a negative value indicates strains removed from a lineage because of their misclassification in the previous database); 8th column, number of strains that were reclassified in SITVITWEB using modified classification rules as compared to SpolDB4.

lines that represent the average of HGDI values calculated individually for each of families (note that detailed values for each locus are given in the inserts beneath each graph for clarity). Since families were defined based on spoligotyping data, we note three different kinds of family-specific HGDI for each locus observed. In the 1st case, the family HGDI is under the global HGDI value thus reflecting a reduced variation of a locus in a family compared to the global collection. For example, MIRU-20 locus is less polymorphic in the Beijing clade (0.01) as compared to the global sample (0.17). In the Beijing clade, the number of repeats in MIRU-20 varies from one (0.7% of isolates) to two (99.3%) copies. The second case reflects higher locus diversity at a family level compared to the global sample, e.g., the Haarlem clade shows a higher diversity of MIRU-20 than the global collection (0.42 versus 0.17). Interestingly, similar to the Beijing isolates, the Haarlem strains also had only two alleles of this locus but in different and less unequal proportion: one copy in 29.6%, and two copies in 70.4% of isolates. This marked variation in the biallelic distribution accounts for different HGDI values for MIRU-20 locus in the Beijing and Haarlem isolates (0.01 versus 0.42). On the global scale, one copy in MIRU-20 was present in 9.1% of all isolates, two copies in 90.7%, and three copies in 0.2% of the isolates. For this reason, MIRU-20 is more variable for



LOCUS	AFRI	Beijing	BOVIS	CAS	EAI	Haarlem	LAM	S	T	X	average of the loci by family	Global HGDI
2	0.232	0.003	0.027	0.000	0.098	0.055	0.332	0.394	0.255	0.040	0.133	0.134
4	0.288	0.066	0.606	0.021	0.612	0.072	0.229	0.372	0.233	0.063	0.248	0.364
10	0.783	0.307	0.664	0.723	0.290	0.554	0.573	0.244	0.595	0.433	0.515	0.729
16	0.652	0.164	0.685	0.415	0.378	0.263	0.606	0.335	0.504	0.174	0.429	0.443
20	0.232	0.013	0.071	0.042	0.066	0.417	0.236	0.201	0.248	0.051	0.149	0.170
23	0.000	0.132	0.296	0.000	0.502	0.532	0.529	0.136	0.315	0.090	0.261	0.556
24	0.181	0.000	0.062	0.000	0.066	0.000	0.000	0.000	0.000	0.004	0.031	0.262
26	0.514	0.605	0.546	0.722	0.050	0.285	0.644	0.319	0.747	0.516	0.481	0.691
27	0.232	0.124	0.163	0.021	0.190	0.137	0.242	0.020	0.000	0.116	0.141	0.149
31	0.458	0.258	0.632	0.539	0.644	0.229	0.391	0.478	0.384	0.317	0.431	0.641
39	0.065	0.330	0.000	0.394	0.545	0.043	0.053	0.060	0.043	0.070	0.169	0.466
40	0.516	0.233	0.045	0.485	0.685	0.544	0.782	0.775	0.641	0.727	0.554	0.696



LOCUS	AFRI	Beijing	BOVIS	CAS	EAI	Haarlem	LAM	S	T	X	average of the loci by family	Global HGDI
ETR-A	0.767	0.061	0.725	0.343	0.780	0.305	0.272	0.431	0.662	0.062	0.441	0.709
ETR-B	0.650	0.021	0.562	0.142	0.785	0.109	0.461	0.058	0.438	0.062	0.329	0.347
ETR-C	0.666	0.510	0.366	0.097	0.371	0.287	0.340	0.058	0.740	0.473	0.391	0.674
ETR-D	0.441	0.207	0.660	0.501	0.662	0.413	0.441	0.613	0.506	0.518	0.496	0.529
ETR-E	0.656	0.129	0.216	0.489	0.697	0.314	0.388	0.616	0.680	0.487	0.467	0.632

Fig. 3. Allelic diversity (HGDI) of minisatellite locus among 10 major genotypic lineages (Africanum, Beijing, Bovis, CAS, EAI, Haarlem, LAM, S, T, and X) of *M. tuberculosis* complex in SITVITWEB. (A). 12-locus MIRU format (B). 5-locus ETRs (ETR-A to E). The red horizontal lines designate the global HGDI values obtained for each locus versus blue horizontal lines that represent the average of HGDI values calculated individually for each of families (data extracted from strains on which both minisatellite and spoligotyping results were available; $n = 3178$ strains in A, and $n = 1510$ strains in B). This graph was designed with the 2D plotting library Matplotlib (http://matplotlib.sourceforge.net/). Detailed values for each locus are given in the inserts beneath each graph.

Haarlem and less variable for Beijing isolates compared to the global collection. Lastly, the third case concerns a monomorphic locus in a given family (HGDI = 0), e.g., all 94 strains of CAS family invariably have two copies in MIRU-2 locus, five copies in MIRU-23 locus, and one copy in MIRU-24-locus. Other examples of lineage-specific allelic distribution include four copies in MIRU-23 locus for *Mycobacterium africanum* (AFRI), one copy in MIRU-24-locus for Beijing and LAM lineages, and two copies of MIRU-39 locus for *Mycobacterium bovis* (BOV). All these detailed results are not shown here, but are consultable using the public database link available at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/.

For the majority of locus, the global HGDI is higher than the average HGDI per lineage. Basically, HGDI depends on the number of alleles and the number of isolates sharing these alleles, or,

concisely, on sample richness, evenness, and size. Hence in our study, the HGDI calculation is influenced by sample size which was larger for Beijing and LAM isolates, and small for CAS and AFRI lineages. A closer look at Fig. 3A validated conclusions of previous smaller studies, e.g., the 3 most variable locus in Beijing isolates are MIRU-26 (HGDI = 0.61), MIRU-39 (HGDI = 0.33) and MIRU-10 (HGDI = 0.31), that were previously shown to be the most polymorphic in a much smaller worldwide sample of the Beijing isolates (Mokrousov et al., 2004). On a global scale, the five most discriminatory locus (HGDI > 0.50) are MIRU-10, MIRU-40, MIRU-26, MIRU-31, MIRU-23 and the three least discriminatory locus (HGDI < 0.20) are MIRU-2, MIRU-27, MIRU-20, an observation previously made on a set of 116 *M. tuberculosis* complex clinical isolates from 11 different geographic origins (Sola et al., 2003). In our global MTC sample, on average MIRU-24 was the least polymorphic locus (average HGDI = 0.03) while MIRU-40 was the most polymorphic (average HGDI = 0.55). Lastly, the variability of the VNTRs was also lineage-dependent, e.g., ETR-A had HGDI < 0.065 for Beijing and X lineages as opposed to values > 0.7 for AFRI, BOV, and EAI lineages (for detailed examples, please refer to Fig. 3B). As summarized in the insert beneath the graphical representation, ETR-A (0.71) and ETR-C (0.67) showed higher global HGDI than ETR-D (0.53) > ETR-E (0.63) > ETR-B (0.35). These observations validate previous studies (Supply et al., 2001, 2006; Mazars et al., 2001; Wirth et al., 2008; Cardoso Oelemann et al., 2011), and further corroborate the stable polymorphism of MIRU–VNTRs locus both at lineage and global levels, suggesting their potential use in future assignments of MTC genotypic lineages even when used alone.

3.3.4. Family-specific regrouping of minisatellites

The distribution of the 1018 classified MIRU patterns (n = 3178 isolates) within spoligotype-defined lineages is shown graphically in Supplemental Fig. S2. It is helpful to reiterate that traditional order of locus in the 12-locus MIRU pattern is by increasing locus number, i.e., from MIRU-2 to MIRU-40 (see insert A1 in Supplemental Fig. S2). To better appreciate the distribution of MIRU patterns within lineages and to highlight those co-shared by different lineages, the order of MIRUs in the 12-locus pattern was meaningfully changed as shown in inserts B1 and B2. We changed locus order according to their HGDI values, starting with the least diverse locus MIRU-24 (insert B1, hereby designated as MIRU12*). As shown in insert B2, we then ordered the MIRU12* values from highest to lowest in function of their absolute numerical values. The resulting 1018 MIRU sorted patterns constitute the *X*-axes in the graphical representation shown in Supplemental Fig. S2. As can be seen, we obtained a good separation of ancestral versus modern strains. Note that some patterns are shared by several families like LAM, Haarlem, X and T while others are restricted to one family such as in case of *M. africanum* (AFRI), *M. bovis* (BOV), and EAI. It should be underlined that identical MIRU patterns were found, albeit rarely, in strains from different spoligotype-based families, which may translate three different situations (i) homoplasy or convergent evolution (ii) imprecise classification (iii) typing and/or transcription errors. Regardless, one should also keep in mind the fact that spoligotyping is also significantly prone to homoplasy, and to a higher extent than the MIRU–VNTRs (Comas et al., 2009).

Last but not least, patterns within each family tend to group in blocks at different positions on the *x*-axis due to family-specific regrouping of MIRU-patterns, e.g., the first 125 patterns in Supplemental Fig. S2 are almost always EAI; patterns 130 to 175 include majority of *M. bovis* patterns, etc. Furthermore, the *x*-axis may be divided into three large intervals: (i) subgroup 1, MIRU12* patterns 0–250 include all ancestral MTC lineages (AFRI, EAI, Bovis), (ii) subgroup 2, MIRU12* patterns 300–450 include the majority of CAS and Beijing patterns, (iii) subgroup 3, patterns 525–1025 include the remaining evolutionary modern families (S, T, X, Haarlem, LAM). These findings are highly interesting since they show a definite logic behind our sub groupings and the subdivision based on *katG–gyrA* polymorphism (Sreevatsan et al., 1997), which provided with the first large evolutionary scale subdivision of *M. tuberculosis* complex isolates into three principal genetic groups (PGG). Indeed, the MIRU12* subgroup 1 corresponds to "ancestral" PGG1 strains (TbD1-positive, containing two or more copies of MIRU-24; Ferdinand et al., 2004; Sun et al., 2004), subgroup 2 corresponds to "modern" PGG1 strains (TbD1-negative; containing only one copy of MIRU-24), while the subgroup 3 essentially concerned evolutionary recent PGG2/3 strains, lacking spoligotyping spacers 33–36, that belong to a large LSP-based group called as Euro American Lineage (Gagneux and Small, 2007).

3.4. Geographical distribution of spoligotyping and MIRU data

The worldwide distribution of MTC genotypes and lineages is presented in the SITVITWEB both at the country level (three letter country codes according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3), as well as regional and sub-regional level according to the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm); Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in four sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia was attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia. Supplemental Fig. S3 summarizes the geographical distribution of orphan (0.1% and above) versus shared-types of MTC isolates in various sub-regions upon spoligotyping (A, B, C), and 12-locus MIRU (D, E, F). A and D show the total percentage of orphan strains in a given sub-region, B and E show the percentage of orphans among the total local MTC population for each of the sub-region; while C and F show geographical distribution of SITed and MITed strains for the same sub-regions (note that sub-regions are shown by identical colors throughout).

The distribution of orphan isolates (n = 4364 spoligotypes and 1533 MIRU patterns) is summarized in Supplemental Fig. S3A and D, respectively. Regarding spoligotype orphans, 34.5% come from Europe with almost similar proportions in Western and Southern Europe (15.1%, and 11.8%, respectively). In contrast, Eastern Europe accounts for only 0.9% of all orphan isolates. America is the second region after Europe with 30% orphans with almost similar proportions in Northern and Southern America (14.9% and 13.5%, respectively), followed by Asia (23%) and Africa (11.7%). Regarding orphan MIRU patterns, 54.1% come from Europe, especially from Northern (28.1%) and Western (15.8%) Europe. America represents 22.9% of all orphan isolates (with a prevalence of 17% in North America), followed by Asia (15.3%), Africa and Oceania (<4%).

We also looked at the percentage of orphans among the total local MTC population for each of the sub-regions (Supplemental Fig. S3B and E). In Asia, both Southern and South-Eastern Asia contained a high and equal proportion of spoligotype orphans in their local MTC population (13.9% and 13.3%, respectively). In Europe, contrary to higher rates observed in Southern and Western Europe (13.5% and 8.5%, respectively), Eastern Europe is marked by the lowest orphan rate of all European sub-regions (3.1%), an observation that indirectly supports the high ongoing actively transmitted

TB disease in Eastern Europe. Lastly, the 3 African sub-regions with highest orphan rates (8–10%) are Eastern, Northern and West Africa. South America, the Caribbean and Northern America account respectively for 10.2%, 6% and 4.5% of orphan isolates in their local populations. If one considers the proportion of MIRU orphans among the total local MTC population for each of the sub-regions (Supplemental Fig. S3 E), South-eastern and Southern Asia account for highest proportions (31.1% and 30%, respectively), followed by South America (26.1%), and Northern and Western Europe (21.6% and 21.9%, respectively). On the contrary, both Southern Africa and Northern Asia were characterized by a lower rate of orphans (8.3% and 7.1%, respectively). Interestingly, if the same figures are compared for spoligotype versus MIRU orphans (Supplemental Fig. S3B and E), on average, there were twice more MIRU orphans in each sub-region than spoligotype orphan isolates, which clearly underlines higher discriminatory power of MIRU12 compared to spoligotyping.

Lastly, the geographical distribution of SITed and MITed strains for various sub-regions is summarized in Supplemental Fig. S3C and F. Despite the huge size of SITed strains in SITVITWEB (n = 53,816 present in 22 sub-regions), we are aware of the limitations regarding potential bias generated due to unequal representativity of available data, e.g., 37.4% and 30.3% of all SITed strains were isolated in the American and European regions. The remaining 32.3% of SITed strains very unevenly represented global MTC diversity. Nonetheless Asia and Africa represent 17.5% and 14.1% of SITed isolates, respectively, which still allows to draw interesting conclusions about global MTC genotypic diversity through SITVITWEB. Overall, if one takes in account only sub-regions containing > 1% of SITed strains, these were present in 17 sub-regions and corresponded to 53373 or 99.1% of SITed isolates, and 91.7% of spoligotyped strains. Regarding MITed isolates (n = 6626 present in 16 sub-regions), the European region contained 51.5% of all MITed strains followed by America (19.7%). Northern Europe, Western Europe and Southern Europe account for 23.5%, 13% and 12.4%, respectively, of all MITed strains. These sub-regions count more than 200 distinct patterns. Asia represents 15.6% of MITed isolates (similarly distributed in its sub-regions). African MITed isolates represent 9.8% of all MITed strains (majority from southern Africa). Taking in account only sub-regions containing > 1% of MITed strains, these were present in 14 sub-regions and corresponded to 6584 or 99.3% of MITed isolates (n = 6626), and 91.7% of MIRU-typed strains (n = 8161). Nonetheless, at this point, we prefer not to make over interpretations on country wise endemicity of MITs seeing the small sample size in some regions.

3.5. Worldwide distribution of major M. tuberculosis complex genotypic lineages

The data accumulated in SITVITWEB database and summarized in Supplemental Table S1 and elsewhere in this paper allowed to make conclusions on existing worldwide distribution of the major *M. tuberculosis* lineages based on spoligotyping data (Fig. 4A); note that it neither takes in account the orphans and ill-classified isolates (labeled as ambiguous in SITVITWEB) nor *M. bovis* since it was overrepresented in certain regions, and only percentages superior to 4% are represented in each pie chart. Here is a brief comment on the map – discussing pie charts from left to right: (i) in North America, almost all the lineages (LAM, Beijing, T, X, Haarlem) are found in similar proportions (around 20%; note that CAS, AFRI and S lineages present below 4% are not shown); (ii) In the Caribbean – Haarlem, LAM and T have almost identical proportions (29.2%, 27.7%, 27.7%), with X (8.2%) they are the only ones to exceed the 4%; (iii) LAM is prevalent in South America with 49.3% and it is followed by T (26.7%) and Haarlem (15.7%) clades; (iv) In Europe, T is the most represented family (35%), it is prevalent in Northern, Southern, and Western Europe. Haarlem is prevalent in Eastern Europe and is the second most represented family in Europe (24%). EAI is prevalent in Northern Europe (24.9%), and may relate to imported cases of the disease. Lastly, S exceeds the threshold of 4% only in Southern Europe (5.8%); (v) LAM is the most prevalent clade in all sub-regions of Africa (except for Western Africa where AFRI is the most predominant lineage at 37%, followed by in Middle Africa at 7.3%). It is followed by the ill-defined T group which is present in all sub-regions with roughly the same proportion (~20%). Beijing is found in three African sub-regions (Northern, Eastern, and Southern Africa), although at different rates – at 21.5% it is highly prevalent in Southern Africa. CAS isolates are found in Eastern and Northern Africa at 11.8% and 7.2%, respectively. S isolates are found in Northern Africa (8%) and Southern Africa (5.8%). Finally, X family exceeds the 4% threshold only in Southern Africa (17.2%), while EAI exceeds 4% in 2 sub-regions, Northern (6.8%) and Eastern Africa (8.2%). Regarding PGG1 group of SITed isolates, it may be concluded that: (i) 51.20% of SITed strains from Russia ("Northern Asia") belong to the Beijing family, which is also the most predominant family in Central, Eastern, and South-Eastern Asia (~50%); (ii) EAI is the most frequent family in South-Eastern and Southern Asia (37.6% and 33%, respectively); (iii) CAS is present in Southern and Western Asia at 30.1% and 9.6%. Regarding PGG2/3 group of lineages, T is the most frequent clade in Western Asia (35.4%), while T, and LAM are found in almost the same percentage (~20%) in Australasia (which also contained same proportions of PGG1 strains belonging to EAI, and Beijing).

A similarity between regions at the MTC genotypic lineages may underline long-term historical links and/or more recent but large population movements, as recently shown for Beijing genotype (Mokrousov et al., 2005; Mokrousov, 2008). Thus, gradient of LAM isolates from the South to North America, may speculatively reflect human migration from Latin America to USA in the 20th century, while the presence of the S family (only present in visible proportions in Southern Europe and Northern Africa) may suggest their geographical proximity and historical relationship as a critical factor in shaping local population structures of *M. tuberculosis* in the Mediterranean. Lastly, the Anglo-Saxon ancestry for the X lineage suggested previously (Brudey et al., 2006) was corroborated by its presence in SITVITWEB in Northern Europe, Northern America, Caribbean, South Africa and Australia. In the past, all these areas were a part of the British Empire and received a large and stable human influx during several centuries from the British Isles with an evident founder effect.

In this work, we did not use MIRU typing information to classify MTC isolates, since further studies will be needed to study all the 24-locus individually and search for the link between their allelic diversity and MTC lineages. This is going to be the next 2-year development phase of the database in our laboratory. Nonetheless, the available information on worldwide distribution of MIRU classified patterns is already available to all users thanks to the public database link available at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE; screen shot of the public database is illustrated in Fig. 4B, which allows not only to query the database but also provides various online tools allowing to map the genotypes on a world map. The currently available online tools allow to view IT (shared-type) geographic data, IT count, IT distribution, clade distribution, spoligotype conversion (from binary to octal and vice versa), and available drug-resistance information. The percent distribution of all 12-locus MIRU patterns in various geographical regions is provided in Supplemental Fig. S4A, whereas information on classification of these MIRU patterns split in 10 major spoligotype-based lineages including *M. bovis* is summarized in Supplemental Fig. S4B (note that the latter figure was drawn to facilitate the lineage-specific regrouping of 12-locus MIRU patterns

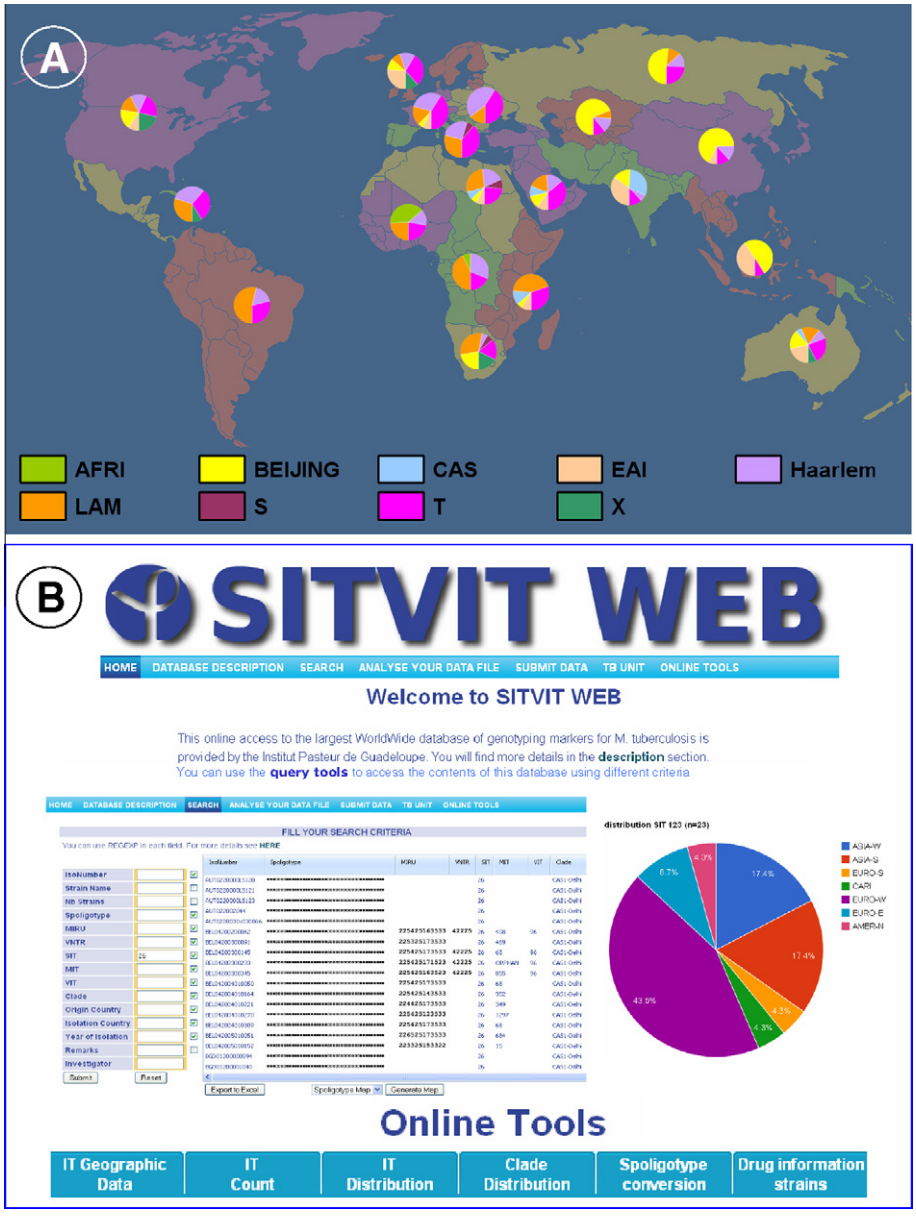


Fig. 4. Geographical distribution of the major *M. tuberculosis* lineages according to spoligotyping data in SITVITWEB (A), and screen shot of the database and online tools available at http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE (B). See text for further details. Note that distribution patterns in A do not take in account orphan or ill-classified MTC strains, nor *M. bovis* (see text for details). Only percentages superior to 4% are shown in pie chart.

in various geographical regions using the same methodology described earlier for Supplemental Fig. S2). It allows to better appreciate the distribution of MIRU patterns within lineages and to highlight those co-distributed in different sub-regions. Briefly, this data concerned 23.4% of spoligotype-classified isolates from Eastern Asia, 21.9% from Northern Asia, 10.4% from Southern Africa. The percentages of the MIRU-typed strains in Western and Eastern Europe, Northern America and the Caribbean lie between 6% and 9%. In all the others sub-regions this percentage is lower than 5%.

The distribution of the lineage-classified MIRU patterns in 13 sub-regions (Supplemental Fig. S4B) shows that Northern America counts more classified patterns ($n = 379$) for 983 isolates, the most represented being patterns belonging to X ($n = 98$), EAI ($n = 76$), and LAM ($n = 72$) lineages. Western Europe is represented by 339 patterns (652 strains), of them 117 patterns in LAM family, 72 in Haarlem and 31 in CAS. Southern Europe is the second European sub-region in term of number of isolates ($n = 163$) and patterns ($n = 82$). In this sub-region, three leading families are LAM, Haarlem, and *Bovis* with respectively 30, 17, 13 patterns. Southern Africa counts 119 classified MIRU patterns (366 strains) while the most represented are patterns are those of LAM ($n = 40$), X ($n = 36$). There is almost exclusively Beijing in Eastern Asia with 65 patterns out of 78 patterns (222 strains) recorded in this sub-region. This is similar to "Northern Asia" (Russia) with 18 Beijing MIRU patterns out of 30 patterns (116 strains) available for this region. In South-eastern Asia only EAI strains were MIRU-typed (43 patterns from 60 strains). In South America there are 142 isolates and 68 patterns, of them *Bovis* counts 22 patterns, LAM 20 patterns, EAI 9 patterns. All the others sub-regions count less than 100 isolates.

3.6. Concluding remarks

Mycobacterium prototuberculosis, the hypothetical progenitor of *M. tuberculosis* was probably a panmyctic species (Gutierrez et al., 2005; Becq et al., 2007; Jang et al., 2008). Contrarily, modern members of the *M. tuberculosis* complex form a strictly clonal

genospecies although few exceptions of this rule are known (McEvoy et al., 2007). A strict clonality implies unidirectional evolution when divergence is generated via intragenomic variation unobscured by lateral gene transfer. Large and phylogenetically relevant subdivisions have been described for *M. tuberculosis* based on different molecular markers in the last decade. Among them, the first attempt based on *katG–gyrA* SNP polymorphism (Sreevatsan et al., 1997) allowed to subdivide *M. tuberculosis* complex isolates into three PGG groupings; PGG1 is considered to evolutionarily older while PGG3 is the youngest which evolved from PGG2. More recently, a novel evolutionary scenario of *M. tuberculosis* complex suggested that ancestral isolates are characterized by the presence of a specific deletion region (TbD1) and modern isolates by TbD1 deletion (Brosch et al., 2002). Use of more discriminatory markers permitted to delineate genetic families with *M. tuberculosis*. One of the initial subdivisions was based on the structure of the DR locus (Kamerbeek et al., 1997) which represents a large polyphyletic family of DNA repeats found in many bacterial species, namely CRISPRs (Jansen et al., 2002; Mojica et al., 2005). Unlike in *Yersinia pestis* (Pourcel et al., 2005), the changes in this locus in *M. tuberculosis* are almost neutral and occur via consecutive irreversible deletions of either single units or contiguous blocks (which occasionally includes IS*6110* mediated disruption and recombination; Beggs et al., 2000; van Embden et al., 2000). Superposition of the above groupings through simultaneous spoligotyping and *katG–gyrA* polymorphism showed that PGG1 includes both ancestral (EAI) and modern (CAS, Beijing) lineages, while PGG2/3 include only modern (Haarlem, LAM, T, and X families) lineages (Gagneux and Small, 2007; reviewed in Rastogi and Sola, 2007). In parallel, SNPs and large sequence polymorphisms (LSPs) allowed inferring more deeply rooted phylogenies (Gutacker et al., 2002; Filliol et al., 2006; Baker et al., 2004; Hirsh et al., 2004), and showed for example that PGG2/3 lineages such as Haarlem, LAM, T, and X were grouped together as "Euro American", a broad LSP-based lineage (Gagneux and Small, 2007).

However, the recent introduction of high-throughput DNA fingerprinting of MTC clinical isolates MIRU–VNTRs (Frothingham and Meeker-O'Connell, 1998; Supply et al., 2001, 2006), allowed to achieve not only a significantly higher discrimination, but also provided with a new insight on MTC evolution (Wirth et al., 2008). The major advantage of the MIRU–VNTR typing as compared to IS*6110* fingerprinting is thanks to its portability (easy digitalization of the generated profiles), easier inter-laboratory exchange, and improved performance of 24-locus typing format (García de Viedma et al., 2011). At present, several databases and online tools are available (El Sahly et al., 2004; Brudey et al., 2006; Allix-Beguec et al., 2008; Mokrousov et al., 2005; Mokrousov, 2008; Grissa et al., 2008a,b; Tang et al., 2008); nonetheless, many are smaller laboratory databases or not publicly available as open access databases. In this sense, the present release of the SITVITWEB makes it the biggest open-access database integrating a large body of information both in terms of number and diversity of entries and markers along with various data-mining tools. It opens new perspectives in future developments aiming to efficiently track the TB epidemic worldwide – both for a better comprehension of current epidemiological situation, but also by opening new research perspectives to identify the smallest number of test isolates and panels which can truly represent the worldwide TB burden based on its exhaustive genotypic diversity. Such panels are a prerequisite to test new drugs to control drug-resistant TB, and also for future vaccine development.

Acknowledgements

The authors are highly grateful to various contributors who very kindly provided genotyping data to successive public databases SpolDB1 to SpolDB4 and the SITVITWEB projects. The full list of contributors is available for each genotyping pattern feeded to the database, and is provided with each interrogation done online. Nonetheless, we would like to mention here the names of those who provided data on 100 or more isolates (note that the names are given in order of the size of data provided from highest to smallest numbers of strains; we regret if any omissions may happen unwillingly): J. R. Driscoll, B. Kreiswirth, T. Victor et al.*, P. M. Hawkey, R. A. Skuce, N. Rastogi et al.**, P. Suffys, S. A. M. Sahal, L. Rigouts, N. Haddad, H. Soini, W. M. Prodinger, A. Gori, M. Fauville-Dufaux et al.***, T. N. Quitugua, U. Singh, C. Garzelli, D. van Soolingen et al.****, V. Katalinic, S. Cheong, N. Morcillo, L. Aristimuno, L. Anh, M. C. Gutierrez, Z. Rahim, T. Koivula, S. Eyangoh, A. Karboul, S. Niemann, F. Stauffer, I. Mokrousov*****, M.J. Zumarraga, N. Kurepina, J. Dale, T. Lillebaeck, R. Jou, M. L. Rossetti, F. Chaves, H. Traore, S. David, D. Garcia de Viedma, I. Parwati, Y. J. Sun, A. Kwara, A. G. Van der Zanden, R. Durmaz, A. Sajduda, R. Hazan, V. Rasolofo-Razanamparany, A. Gibson, B. de Jong, C. Gilpin, M. P. Nicol, J. Maugein, D. Nguyen, R. Diaz, S. Gagneux, T. Iwamoto, C. A. Rodriguez, D. Alland, D. Cousins, V. Ritacco, G. Kallenius, P. Supply, S. Kulkarni, J. Martin, S. Panaiotov, M. H. Féres Saad, F.P. Sweeney, J. Bauer, L. S. Cowan, M. Goyal, M. Kubin, L. Binder, E. Kassa-Kelembho, J. T. Douglas, S. Bonora, O. Dellagostin, T. Tracevska, D. Almeida, O. Ogarkov, E. Aktas, R. Ohata, G. S. Kibiki, P. Farnia, A. Michel, I. Shemyakin, M. Doroudchi, C. E. Salas, O. Kisa, C. Ellermeier, J. De Waard, V. Narbonne, Y. O. Goguet de la Salmonière, M. Ridell, G. Vrioni, F. Portaels. (Asterisks; * T. Victor contributed genotypes jointly with E. Streicher and R. M. Warren; ** N. Rastogi contributed genotypes jointly with his coworkers: A. Devallois, E. Legrand, C. Sola, L. Horgen, I. Filliol, S. Ferdinand, K. Brudey, T. Zozio, J. Millet, V. Hill, and D. Couvin; ***M. Fauville-Dufaux contributed genotypes jointly with C. Allix; ****D. van Soolingen contributed genotypes jointly with K. Kremer. *****I. Mokrousov contributed genotypes jointly with O. Narvskaya. The work done at Nalin Rastogi's laboratory was financed by the International Network of the Pasteur Institutes. Thierry Zozio, Julie Millet, Veronique Hill and David Couvin were awarded Ph.D. fellowships by the European Social Funds through the Regional Council of Guadeloupe. The SITVITWEB database project was partially financed by the Regional Council of Guadeloupe (CR/08-1612: Biodiversité et Risque Infectieux dans les modèles insulaires).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2012.02.004.

References

Allix-Beguec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S., 2008. Evaluation and strategy for use of MIRU–VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J. Clin. Microbiol. 46 (8), 2692–2699.

Armitage, P., Berry, G., 1987. Statistical Methods in Medical Research. Blackwell Scientific Publications, Ltd., Oxford, UK, p. 49–53.

Baker, L., Brown, T., Maiden, M.C., Drobniewski, F., 2004. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerg. Infect. Dis. 10, 1568–1577.

Becq, J., Gutierrez, M.C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O., Deschavanne, P., 2007. Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. Mol. Biol. Evol. 24 (8), 1861–1871.

Beggs, M.L., Eisenach, K.D., Cave, M.D., 2000. Mapping of IS6110 insertion sites in two epidemic strains of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 38, 2923–2928.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D., Cole, S.T., 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc. Natl. Acad. Sci. USA 99, 3684–3689.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong,

S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Gutierrez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ly, H.M., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, M.Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I., Singh, U.B., Somoskovi, A., Skuce, R., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2008. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Cardoso Oelemann, M., Gomes, H.M., Willery, E., Possuelo, L., Batista Lima, K.V., Allix-Béguec, C., Locht, C., Goguet de la Salmonière, Y.O., Gutierrez, M.C., Suffys, P., Supply, P., 2011. The forest behind the tree: phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country. PLoS One 6 (3), e18256.

Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One 4 (11), e7815.

El Sahly, H.M., Wright, J.A., Soini, H., Bui, T.T., Williams-Bouyer, N., Escalante, P., Musser, J.M., Graviss, E.A., 2004. Recurrent tuberculosis in Houston, Texas: a population-based study. Int. J. Tuberc. Lung Dis. 8, 333–340.

Ferdinand, S., Valetudie, G., Sola, C., Rastogi, N., 2004. Data mining of *Mycobacterium tuberculosis* genotyping results using mycobacterial interspersed repetitive units validates clonal structure of spoligotyping-defined families. Res. Microbiol. 155, 647–654.

Filliol, I., Driscoll, J.R., van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valetudie, G., Dang, D.A., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Kassa-Kelembho, E., Ho, M.L., Makristathis, A., Mammina, C., Martin, G., Mostrom, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., de Waard, J.H., Sola, C., Rastogi, N., 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J. Clin. Microbiol. 41, 1963–1970.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., Bobadilla del Valle, M., Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., Leon, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendon, A., Sifuentes-Osornio, J., Ponce de Leon, A., Cave, M.D., Fleischmann, R., Whittam, T.S., Alland, D., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. 188, 759–772.

Frothingham, R., Meeker-O'Connell, W.A., 1998. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 144, 1189–1196.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. USA 103 (8), 2869–2873.

Gagneux, S., Small, P.M., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet. Infect. Dis. 7 (5), 328–337.

García de Viedma, D., Mokrousov, I., Rastogi, N., 2011. Innovations in the molecular epidemiology of tuberculosis. Enferm. Infecc. Microbiol. Clin. 29 (Suppl 1), 8–13.

Grissa, I., Vergnaud, G., Pourcel, C., 2008a. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 1;36 (Web Server issue):W145–148.

Grissa, I., Bouchon, P., Pourcel, C., Vergnaud, G., 2008b. On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. Biochimie 90, 660–668.

Gutacker, M.M., Smoot, J.C., Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N., Musser, J.M., 2002. Genome-Wide Analysis of Synonymous Single Nucleotide Polymorphisms in *Mycobacterium tuberculosis* Complex Organisms. Resolution of genetic relationships among closely related microbial strains. Genetics 162, 1533–1543.

Gutierrez, M.C., Brisse, S., Brosch, R., Fabre, M., Omais, B., Marmiesse, M., Supply, P., Vincent, V., 2005. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog. 1, e5.

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., Small, P.M., 2004. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc. Natl. Acad. Sci. USA 101, 4871–4876.

Hunter, P.R., Gaston, M.A., 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 26, 2465–2466.

Jang, J., Becq, J., Gicquel, B., Deschavanne, P., Neyrolles, O., 2008. Horizontally acquired genomic islands in the tubercle bacilli. Trends Microbiol. 16 (7), 303–308.

Jansen, R., Embden, J.D., Gaastra, W., Schouls, L.M., 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. 43, 1565–1575.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Mazars, E., Lesjean, S., Banuls, A.L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C., Supply, P., 2001. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. Proc. Natl. Acad. Sci. USA 98, 1901–1906.

McEvoy, C.R., Falmer, A.A., Gey van Pittius, N.C., Victor, T., van Helden, P.D., Warren, R.M., 2007. The role of IS6110 in the evolution of *Mycobacterium tuberculosis*. Tuberculosis (Edinb) 87, 393–404.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., Soria, E., 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. 60, 174–182.

Mokrousov, I., 2008. Genetic geography of *Mycobacterium tuberculosis* Beijing genotype: a multifacet mirror of human history? Infect. Genet. Evol. 8, 777–785.

Mokrousov, I., Ly, H.M., Otten, T., Lan, N.N., Vyshnevskyi, B., Hoffner, S., Narvskaya, O., 2005. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res. 15, 1357–1364.

Mokrousov, I., Narvskaya, O., Limeschenko, E., Vyazovaya, A., Otten, T., Vyshnevskiy, B., 2004. Analysis of the allelic diversity of the mycobacterial interspersed repetitive units in *Mycobacterium tuberculosis* strains of the Beijing family: practical implications and evolutionary considerations. J. Clin. Microbiol. 42, 2438–2444.

Pourcel, C., Salvignol, G., Vergnaud, G., 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151, 653–663.

Rastogi, N., Sola, C. 2007. Molecular evolution of the *Mycobacterium tuberculosis* complex, p. 53–91. In Palomino, J. C., Leao, S., Ritacco, V. (eds.), Tuberculosis 2007: From Basic Science to Patient Care. Amedeo Online Textbooks, http://www.tuberculosistextbook.com/index.htm.

Shabbeer A, Ozcaglar C, Yener B, Bennett KP. 2011. Web tools for molecular epidemiology of tuberculosis. Infect Genet Evol. 2011 Aug 28. [Epub ahead of print].

Simpson, E.H., 1949. Measurement of Diversity. Nature 163, 688.

Sola, C., Filliol, I., Legrand, E., Lesjean, S., Locht, C., Supply, P., Rastogi, N., 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect. Genet. Evol. 3, 125–133.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser, J.M., 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. USA 94, 9869–9874.

Sun, Y.J., Lee, A.S., Ng, S.T., Ravindran, S., Kremer, K., Bellamy, R., Wong, S.Y., van Soolingen, D., Supply, P., Paton, N.I., 2004. Characterization of Ancestral *Mycobacterium tuberculosis* by Multiple Genetic Markers and Proposal of Genotyping Strategy. J. Clin. Microbiol. 42, 5058–5064.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rüsch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., Locht, C., 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J. Clin. Microbiol. 39, 3563–3571.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., Locht, C., 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. 36, 762–771.

Tang, C., Reyes, J.F., Luciani, F., Francis, A.R., Tanaka, M.M., 2008. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. Bioinformatics 24, 2414–2415.

van Embden, J.D.A., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., Small, P.M., 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31, 406–409.

van Embden, J.D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B.A., Schouls, L.M., 2000. Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J. Bacteriol. 182, 2393–2401.

Vanhomwegen, J., Kwara, A., Martin, M., Gillani, F.S., Fontanet, A., Mutungi, P., Crellin, J., Obaro, S., Gosciminski, M., Carter, E.J., Rastogi, N., 2011. Impact of immigration on the molecular epidemiology of tuberculosis in Rhode Island. J. Clin. Microbiol. 49, 834–844.

WHO Report 2008. Global tuberculosis control: surveillance, planning, financing. (http://www.who.int/tb/publications/global_report/2008/pdf/fullreport.pdf).

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., Niemann, S., 2008. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 4 (9), e1000160.

Article 02 (en rédaction)

PLoS One

Macro-geographical specificities of the prevailing Tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database

David Couvin[1], Nalin Rastogi[1]*

[1] WHO Supranational TB Reference Laboratory, Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Abymes, Guadeloupe, France

Running title: *SITVIT2 Database*

Key-words: *Mycobacterium tuberculosis*, Database, Spoligotyping, MIRU-VNTRs, Evolution, Epidemiology, Sex, Age, HIV serology, Drug resistance.

---

* Corresponding author. Mailing address: Institut Pasteur de Guadeloupe, Morne Jolivière, BP484, F97183 Abymes Cedex, Guadeloupe. Phone: +590-590-897661. Fax: +590-590-896941. E-mail: nrastogi@pasteur-guadeloupe.fr.

Réseaux

La mise en place de bases de données des génotypes circulants du complexe *Mycobacterium tuberculosis* et des outils web permettant de mieux surveiller, comprendre et contrôler l'épidémie de la tuberculose dans le monde

David Couvin et Nalin Rastogi (nrastogi@pasteur-guadeloupe.fr).
Laboratoire supranational de référence pour la TB de l'OMS, institut Pasteur de la Guadeloupe, Abymes, Guadeloupe, France

Mots-clés : *Mycobacterium tuberculosis*, tuberculose, génotypage, bases de données, spoligotypage, MIRU-VNTR, épidémiologie, phylogénie, résistance aux médicaments.

Résumé

Dans cet article, nous présenterons rapidement plusieurs bases de données de génotypage du complexe *Mycobacterium tuberculosis* (MTBC), élaborées au cours des quinze dernières années à l'Institut Pasteur de la Guadeloupe (IPG), dans le cadre d'une grande initiative concertée visant à lutter contre l'épidémie mondiale de la tuberculose. De la toute première version au format Excel en 1999 (SpolDB1 ; n = 610 isolats cliniques) à la quatrième au format MySQL en 2006 (SpolDB4 ; n = 39 295 isolats cliniques), ces bases de données ont brossé un premier tableau de la phylogéographie des lignées génotypiques de MTBC en circulation, en utilisant le typage oligonucléotidique des espaceurs (spoligotypage) qui permet d'étudier le polymorphisme du locus DR (*Direct Repeat*). Les deux dernières versions multimarqueurs sous MySQL constituent respectivement la 5e version de SITVITWEB, publiée en 2012 (n = 62 582 isolats cliniques), qui repose sur le spoligotypage et le typage MIRU-VNTR (*Mycobacterial Interspersed Repetitive Units/Variable Number of Tandem Repeats*) 12 loci, et la 6e version, dénommée SITVIT2, qui sera publiée en 2014 (n = 111 635 isolats cliniques) et qui contient des données de spoligotypage et de MIRU-VNTR 12, 15 ou 24 loci. Sur ces dernières versions, une interface en ligne permet à l'utilisateur de rechercher des souches dans la base de données au moyen de critères tels que l'année, le pays d'isolement, le pays d'origine ou le nom du chercheur. Cette interface permet en outre d'effectuer des recherches mixtes dans SITVIT2, permettant d'obtenir les données de génotypage de certaines souches, conjointement avec leur répartition géographique, ainsi que les données disponibles sur la résistance aux antibiotiques ou les caractéristiques démographiques et épidémiologiques. Notre initiative de recherche a ainsi pour but d'améliorer la caractérisation phylogénétique détaillée des lignées du MTBC, ainsi que l'épidémiologie des clones en circulation, afin d'élaborer une cartographie géographique factuelle des isolats cliniques prédominants pour les bacilles tuberculeux principalement impliqués dans la maladie, à l'échelle nationale et régionale. La superposition ultérieure de ces cartes avec des données sociopolitiques, économiques et démographiques obtenues auprès de Systèmes d'information géographique (SIG) dressera un portrait précis des disparités actuelles par sous-région, selon la classification des Nations Unies. Il est important de comprendre le détail de ces disparités et faiblesses pour que les décideurs politiques et les autorités de santé publique puissent prendre les mesures qui s'imposent pour mieux surveiller, comprendre et contrôler l'épidémie mondiale de la tuberculose.

Introduction

Près de vingt ans après que la tuberculose (TB) a été déclarée « urgence de santé publique mondiale » par l'Organisation mondiale de la santé (OMS), et malgré les nets progrès réalisés pour atteindre les objectifs internationaux de 2015, fixés dans le cadre des objectifs du Millénaire pour le développement, la TB demeure la deuxième maladie infectieuse la plus meurtrière au monde après le VIH/SIDA (rapport de l'OMS, WHO 2013). Selon les estimations, la TB a été responsable de 8,6 millions de nouveaux cas et de 1,3 million de décès (dont 320 000 chez des patients co-infectés par la TB et le VIH) en 2012. Un examen attentif du rapport de l'OMS révèle que la TB demeure un problème économique et sanitaire grave, non seulement dans les pays en développement, mais aussi au sein des nations développées, en raison des co-infections TB/VIH et de l'émergence de souches multirésistantes aux médicaments (*Multi-Drug Resistant*, MDR) et, plus récemment, de souches extrêmement résistantes aux médicaments (*eXtremely Drug Resistant*, XDR), complexifiant la gestion de la maladie et augmentant considérablement la mortalité par TB chez les patients immunodéprimés.

De plus, l'intensification des déplacements et des migrations de population à laquelle on assiste depuis plusieurs décennies, que ce soit pour des motifs touristiques ou professionnels, soulève un nouveau problème dans les pays où la TB était en déclin : la modification des scénarios socioépidémiologiques en raison d'une immigration massive en provenance de pays où la TB est fortement endémique (García de Viedma *et al.*, 2011). Parmi les questions qui se posent immanquablement, on retiendra notamment : la comparaison entre le rôle de la transmission récente et celui de la réactivation/importation dans les cas de TB d'origine étrangère ; les répercussions d'une éventuelle importation de souches de *M. tuberculosis* non identifiées à ce jour ; et la transmission croisée entre cas de différentes nationalités. Il est donc important de comprendre comment se transmettent les bacilles tuberculeux, de savoir quels sont les clones impliqués dans les épidémies et/ou les cas résistant à l'action des médicaments, d'identifier de nouveaux clones susceptibles de voir le jour dans un milieu donné ou d'être en voie d'extinction, d'identifier les sous-populations les plus à risque de contracter l'infection et les facteurs de risque associés, et de pouvoir interpréter ces résultats à la lumière de l'évolution du complexe *Mycobacterium tuberculosis* (MTBC). Il est évident qu'il serait difficile, aujourd'hui, de répondre à ces questions sans l'aide de l'épidémiologie moléculaire.

On considérait, il y a encore peu de temps, que toutes les souches de MTBC adaptées à l'Homme étaient pratiquement identiques ; la question de la variabilité génétique individuelle

Réseaux

entre espèces du MTBC n'a donc pas retenu une grande attention et la plupart des travaux réalisés en la matière se sont focalisés sur des organismes précis. L'arrivée des techniques moléculaires et leur généralisation dans les études des populations de bacilles tuberculeux ont fait évoluer les idées et les technologies. Bien que le MTBC constitue un groupe remarquablement homogène sur le plan génétique, avec des signes tangibles d'évolution clonale, des études récentes ont démontré que la diversité génétique entre les différents clones était nettement plus importante qu'on ne le pensait, ce qui pourrait avoir un impact sur leurs propriétés pathobiologiques. Une étude de référence menée sur une collection mondiale de souches de MTBC, à partir des données de séquence d'ADN de sept paires de mégabases, a ainsi révélé une diversité génétique importante (Hershberg *et al.,* 2008). Les auteurs de cette étude pensent qu'une grande partie de cette diversité génétique provient d'une dérive génétique qui pourrait être liée à des évènements démographiques et migratoires humains, avec des conséquences fonctionnelles comme l'émergence et la propagation d'une TB résistante aux médicaments.

Bien que la TB soit présente dans le monde entier, son incidence est plus élevée dans certains pays ou certaines régions que dans d'autres : l'Afrique subsaharienne, l'Asie du Sud et du Sud-Est, l'Amérique latine, mais aussi l'Europe de l'Est et la Russie, ainsi que les Antilles (notamment Haïti et la République dominicaine), par exemple. En effet, la TB constitue toujours un problème sanitaire grave dans de nombreux pays des Antilles et d'Amérique latine ; par exemple, l'incidence de la TB à Haïti, qui s'élevait à 330 cas/100 000 habitants dans les années 90, est toujours aussi élevée en 2012, avec 213 cas/100 000 habitants (rapport de l'OMS, 2013). Il va sans dire que le diagnostic et la confirmation bactériologique précoces de la TB ainsi que la détermination de la résistance aux médicaments constituent des étapes clés dans la lutte contre l'épidémie de TB, et que les laboratoires de référence constituent des structures de premier plan en matière de diagnostic et de lutte antituberculeuse à l'échelle mondiale. Dans un tel contexte, l'Institut Pasteur a pris l'initiative de créer le premier laboratoire de référence pour la TB dans la région des Antilles en 1993. Situé à l'Institut Pasteur de la Guadeloupe (IPG), ce laboratoire était initialement destiné à travailler pour les Antilles ; cependant, bien conscients de la portée mondiale de la TB, nous avons dès le départ adopté une approche fondée sur l'utilisation concomitante de techniques bactériologiques et moléculaires.

Au cours des 20 dernières années, nous avons mis au point une approche complète qui intègre le dépistage moléculaire systématique du MTBC et d'autres espèces mycobactériennes, la surveillance de la résistance aux antibiotiques, et l'élaboration de méthodes de diagnostic rapide et de diverses techniques moléculaires utiles pour l'épidémiologique et les études des populations de bacilles tuberculeux aux niveaux local, régional et mondial. Pour la surveillance de la TB à l'échelle mondiale, nous avons créé, depuis 1999, une série de bases de données de génotypage qui regroupent non seulement nos propres données, mais aussi celles recueilles auprès de différents laboratoires participants implantés dans le monde entier. Cette publication rappellera brièvement les étapes suivies pour constituer ces bases de données et les outils en ligne mis au point pour permettre de créer des cartes de répartition géographique mondiale montrant la fréquence des génotypes de TB dans le monde à différentes échelles géographiques. Nous citerons également quelques études, publiées ou en cours, qui utilisent ces données et nous présenterons les perspectives qui se dessinent aujourd'hui.

TB : origine, propagation et co-adaptation avec les hôtes

Depuis l'isolement et la caractérisation d'un ancien ADN de *M. tuberculosis* datant de 17 000 ans avant J. C. chez une espèce de bison disparue, révélant la présence de la TB en Amérique vers la fin du pléistocène (Rothschild *et al.,* 2001), on sait que la TB est une maladie très ancienne. L'étude d'ADN anciens prélevés sur des restes humains a permis de détecter la présence de la TB sur des momies égyptiennes, avec la caractérisation de *M. tuberculosis* et *M. africanum* (Zink *et al.,* 2003). On a alors pensé, par analogie avec d'autres épidémies liées à l'expansion démographique (« crowd diseases"), que l'origine de la TB humaine était liée à la transition démographique néolithique (TDN), qui a commencé il y a environ 11 000 ans. En effet, le développement de la domestication des animaux a accru le risque de transmission zoonotique de nouveaux agents pathogènes à l'Homme, tandis que les innovations agricoles ont favorisé l'accroissement des densités de population, contribuant à entretenir le cycle infectieux (Wolfe *et al.,* 2007). On ne savait cependant pas avec certitude (a) si la TB humaine descendait d'une mycobactérie de ruminant qui avait récemment infecté l'Homme par l'intermédiaire des animaux domestiques ou d'une ancienne mycobactérie humaine qui avait infecté les ruminants sauvages et domestiques ; et (b) si la TB était apparue indépendamment dans les deux hémisphères ou si elle avait été amenée aux Amériques par les Européens. Toutefois, la TB présentait aussi un profil de chronicité, de latence et de réactivation caractéristique d'une maladie pré-TDN (Barry *et al.,* 2009).

C'est grâce à de nouvelles recherches qui ont montré que la TB est probablement aussi vieille que l'humanité (Comas *et al.,* 2013) que l'on a obtenu la réponse à ces questions. En étudiant les différentes variations génétiques au sein du MTBC, des chercheurs sont parvenus à démontrer que la TB s'est certainement propagée dans le monde avec les premiers Hommes modernes ayant quitté l'Afrique. Cette étude a consisté à analyser l'ensemble des génomes d'une collection de 259 souches de MTBC contemporaines provenant du monde entier, puis à comparer la diversité phylogénétique du MTBC à la diversité humaine, déduite de données sur le génome mitochondrial. Les résultats obtenus ont révélé que le MTBC est apparu il y a environ 70 000 ans, qu'il a accompagné les migrations des Hommes anatomiquement modernes hors d'Afrique et qu'il s'est propagé en raison de l'accroissement des densités de population humaine pendant la période néolithique (Comas *et al.,* 2013). Cette origine ancienne de la TB et le fait que la phylogénie génomique du MTBC soit calquée sur celle des génomes mitochondriaux humains ont par ailleurs révélé que la forme pulmonaire de la TB ne s'est pas propagée à l'Homme par l'intermédiaire des animaux domestiques, puisque l'élevage est apparu beaucoup plus tard. Cette étude de référence a également révélé des similitudes frappantes au niveau de l'évolution de l'Homme et du MTBC, suggérant non seulement que l'évolution du MTBC a suivi celle de l'Homme, mais aussi que la diversité du MTBC a directement tiré profit des explosions démographiques humaines. Étant donné que *M. tuberculosis* est un pathogène humain strict, sans réservoir animal ou environnemental connu, il est très probable que les changements démographiques et l'accroissement des densités

Réseaux

de population humaine qui s'opèrent au fil du temps affectent son évolution, comme dans le cas d'un modèle de maladie de groupe « crowd diseases ». En même temps, sa latence et sa chronicité peuvent lui permettre de s'adapter à des densités d'hôtes plus faibles, de survivre, puis de frapper à nouveau lorsque les conditions sont favorables à une infection massive.

Techniques de typage pour l'épidémiologie moléculaire de la TB

Le MTBC est un groupe d'organismes très varié sur le plan écologique ; il comprend *M. tuberculosis, M. africanum* et *M. canettii* (pathogènes humains stricts), M. microti (pathogène des rongeurs) et *M. bovis* (pathogène des bovins qui présente cependant un large spectre d'hôtes), ainsi que *M. pinnipedii* (phoques), *M. caprae* (chèvres), *M. mungi* (mangoustes rayées) et les bacilles de l'oryx (rebaptisé *M. orygis*), du rat des rochers et du chimpanzé, agents étiologiques de la TB chez les espèces animales qui ont donné leur nom (bacilles de « dassie » et « chimpanzee" respectivement). Cependant, le MTBC est un groupe très homogène sur le plan génétique et plusieurs de ses membres partagent, en moyenne, plus de 99,7 % d'identité nucléotidique (Kato-Maeda *et al.,* 2001). Malgré cette remarquable homogénéité génétique, les 20 dernières années ont vu naître de nouvelles techniques moléculaires, largement utilisées aujourd'hui dans les études de génotypage des populations, qui permettent de caractériser les isolats de TB avec précision et de déduire les différentes lignées phylogénétiques qui y sont associées. Des revues détaillées sont disponibles sur l'évolution moléculaire du MTBC (Rastogi et Sola, 2007), les techniques de typage moléculaire actuelles (Jagielski *et al.,* 2014), ainsi que des stratégies et innovations dans le vaste domaine de l'épidémiologie moléculaire de la TB (2) (García de Viedma *et al.,* 2011) ; les lecteurs intéressés sont invités à consulter ces références pour approfondir ces questions.

Bien que la méthode IS6110-RFLP ait longtemps été considérée comme la technique de référence pour le typage de *M. tuberculosis* en raison de sa reproductibilité et de son pouvoir discriminant pour l'épidémiologie moléculaire de la TB (van Embden *et al.,* 1993), cette méthode laborieuse nécessitait de grandes quantités d'ADN et se caractérisait par l'absence de pouvoir discriminant pour le typage des isolats ayant un faible nombre de copies d'IS6110, par exemple dans le sud de l'Inde (Radhakrishnan *et al.,* 2001). De surcroît, l'horloge moléculaire rapide de ce marqueur pour les études d'évolution (de Boer *et al.,* 1999), la complexité des forces qui régissent sa transposition et le risque de convergence génétique (Fang *et al.,* 2001), ainsi que la difficulté à élaborer de grandes bases de données RFLP et la nécessité de disposer de logiciels perfectionnés pour l'analyse des données (Heersma *et al.,* 1998 ; Salamon *et al.,* 1998) ont mis en doute l'intérêt de son utilisation en génétique de l'évolution. Pour les raisons exposées plus haut, la technique IS6110-RFLP a largement fait place, au cours des 10 dernières années, à de nouvelles stratégies de typage par PCR comme le spoligotypage (Kamerbeek *et al.,* 1997) et la technique MIRU-VNTR (*Mycobacterial Interspersed Repetitive Units/Variable Number of Tandem Repeats*) (Supply *et al.,* 2001 ; 2006), jetant ainsi les bases d'un génotypage de *M. tuberculosis* à haut débit et à grande échelle.

S'appuyant sur le polymorphisme du locus DR (*Direct Repeat*), le spoligotypage (de l'anglais SPacer OLIGOnucleotide TYPING, typage oligonucléotidique des espaceurs) fait actuellement partie des techniques de PCR les plus utilisées pour l'épidémiologie moléculaire et l'étude phylogéographique du MTBC. D'abord identifié dans la souche vaccinale M. bovis BCG (Hermans *et al.,* 1991), le locus DR appartient à la famille d'ADN répétitifs CRISPR (*Clustered Regularly Interspaced Short Palindromic Repeats*) et contient plusieurs répétitions parfaites de 36 pb (3 tours d'hélice) identiques, séparées par des espaceurs de 34 à 41 pb uniques. Avec leurs espaceurs non répétitifs, les motifs DR forment plusieurs répétitions directes variables (DVR, Direct Variant Repeats) qui présentent un important polymorphisme entre les isolats cliniques de *M. tuberculosis*. Une étude a précédemment démontré que le profil binaire à 43 caractères obtenu par cette technique véhiculait des informations phylogénétiques déterminantes (Sola *et al.,* 2001). Aujourd'hui, les profils de spoligotypage sont généralement désignés par leur code octal, format descriptif faisant l'objet d'une convention internationale (Dale *et al.,* 2001). En raison de sa simplicité, de son format de résultats binaire et de sa grande reproductibilité, le spoligotypage est très utilisé pour l'épidémiologie moléculaire du MTBC, en tant que technique sur macroréseaux (*macroarrays*), pour analyser la présence ou l'absence de 43 espaceurs prédéfinis (sur 104 espaceurs présents). En effet, la recherche « spoligotyping OR spoligotype » dans PubMed donne 924 articles publiés (interrogation effectuée le 18 mars 2014).

L'une des limites du spoligotypage étant sa tendance à surestimer le regroupement en grappes du MTBC, il a rapidement été proposé de compléter cette technique par un typage reposant sur des minisatellites (MIRU-VNTR), dans le cadre d'une stratégie « à double PCR », en association avec les études épidémiologiques classiques (Sola *et al.,* 2003). Les minisatellites MIRU-VNTR constituent un ensemble de marqueurs multi-loci dans la mesure où ils représentent des marqueurs indépendants d'un même type ; ils sont utilisés aux formats 12 loci classique, 15 loci discriminant ou 24 loci complet (Supply *et al.,* 2001 ; 2006). Cependant, même les MIRU 24 loci ne présentent pas un pouvoir de résolution satisfaisant pour discriminer avec précision des souches de génotype Beijing étroitement apparentées, si bien qu'il a récemment été proposé d'utiliser un autre ensemble de MIRU-VNTR consensus hypervariables 4 loci pour le sous-typage des grappes et complexes clonaux du génotype Beijing (Allix-Béguec *et al.,* 2014).

Techniques de typage moléculaire pour la phylogénie de la TB

Outre leur utilisation en épidémiologie (García de Viedma *et al.,* 2011), les techniques de typage moléculaire sont également utilisées dans les études d'évolution (Rastogi et Sola, 2007). Elles comprennent alors essentiellement deux ensembles de marqueurs : (a) ceux cités plus haut, qui sont largement utilisés dans les études épidémiologiques, mais qui fournissent également des informations phylogénétiques concomitantes – IS6110-RFLP, spoligotypage et MIRU-VNTR (voir ci-après) ; et (b) un ensemble de marqueurs comprenant les polymorphismes de longues séquences (LSP, *Large Sequence Polymorphisms*)/régions de différence (RD) et les polymorphismes mononucléotidiques (SNP, *Single Nucleotide Polymorphisms*), particulièrement utiles pour les études de phylogénie et d'évolution.

L'une des toutes premières études en la matière a eu recours à l'hybridation génomique soustractive pour identifier trois

Réseaux

régions génomiques distinctes, entre des souches virulentes de *M. bovis* et *M. tuberculosis* et la souche avirulente *M. bovis* BCG, respectivement dénommées RD1, RD2 et RD3 (Mahairas *et al.,* 1996). Dans une autre étude, les chercheurs sont parvenus à distinguer trois groupes génétiques de *M. tuberculosis* sur la base de deux polymorphismes apparaissant avec une fréquence élevée dans les gènes codant la catalase-peroxydase et la sous-unité A de la gyrase, aboutissant à une classification en trois groupes génétiques principaux (PGG, *Principal Genetic Groups*), les bactéries du groupe 1 étant ancestrales de celles des groupes 2 et 3 (Sreevatsan *et al.,* 1997). Presque aussitôt, des matrices de chromosomes artificiels bactériens (BAC, *Bacterial Artificial Chromosome*) de la souche H37Rv digérés par une enzyme de restriction ont été utilisées pour révéler la présence de 10 RD entre *M. tuberculosis* et *M. bovis* (RD1 à RD10), dont 7 (RD4 à RD10) étaient délétées chez *M. bovis* (Gordon *et al.,* 1999). Dans une présentation de référence, Brosch *et al.* (2002) ont analysé la répartition de 20 régions variables résultant d'évènements d'insertion-délétion dans le génome des bacilles tuberculeux d'une collection de souches appartenant à toutes les sous-espèces du MTBC, démontrant que, sur la base de la présence ou de l'absence d'une délétion 1 spécifique de *M. tuberculosis* (TbD1, séquence de 2 kb), *M. tuberculosis* pouvait être subdivisé en souches TbD1-positives « anciennes » et en souches TbD1-négatives « modernes » (Brosch *et al.,* 2002). Selon ce nouveau scénario de l'évolution du MTBC, la délétion de RD9 identifie une lignée, représentée par *M. africanum*, *M. microti* et *M. bovis*, qui a divergé de l'ancêtre des souches de *M. tuberculosis* actuelles avant l'apparition de TbD1, découverte qui contredit les hypothèses antérieures selon lesquelles *M. tuberculosis* a évolué à partir d'un ancêtre de *M. bovis* (Brosch *et al.,* 2002). Étant donné qu'aucune de ces régions n'est manquante chez *M. canettii* et d'autres souches ancestrales du MTBC, l'on présume que ce sont des descendants directs des bacilles tuberculeux qui existaient avant la séparation entre la lignée « *M. africanum* – *M. bovis* » et la lignée « *M. tuberculosis* ».

En utilisant une collection mondiale de MTBC et 212 SNP, Filliol *et al.* (2006) ont identifié six groupes de grappes SNP (SCG, *SNP Cluster Groups*) phylogénétiquement différents et à embranchement profond, ainsi que cinq sous-groupes. Ces SCG sont fortement corrélés avec l'origine géographique des échantillons de *M. tuberculosis* et le lieu de naissance des hôtes humains. Les auteurs de l'étude ont proposé un algorithme capable d'identifier deux ensembles minimaux composés de 45 ou de 6 SNP, sur 212 SNP testés, qui peuvent servir à cribler des collections mondiales de MTBC pour étudier l'évolution, la différenciation des souches ou les différences biologiques entre souches. Dans une autre étude, Gutacker *et al.,* (2006) se sont intéressés aux relations génétiques du MTBC en analysant 36 sSNP au sein d'une grande collection de souches isolées de patients inclus dans quatre études des populations aux États-Unis et en Europe, et ils ont affecté cette collection à 1 de 9 grandes grappes génétiques. Une classification similaire a été révélée par analyse d'un panel élargi à d'autres SNPs. Étant donné que les profils de classification des lignées phylogénétiques typées par SNP ont été associés de manière non aléatoire à des profils IS6110, des spoligotypes et des profils MIRU-VNTR, les auteurs pensent que la population du MTBC présente une structure fortement clonale.

Parallèlement, en utilisant des microréseaux d'ADN (*microarrays*) pour identifier avec précision les LSP, des chercheurs ont observé une corrélation stable entre les souches de MTBC et leurs populations d'hôtes humains (Hirsh *et al.,* 2004) ; les analyses phylogénétiques ont non seulement indiqué que les transferts horizontaux de gènes étaient rares au sein du MTBC, mais aussi que les corrélations entre les populations d'hôtes et d'agents pathogènes étaient stables, même dans un milieu urbain cosmopolite (comme San Francisco) et qu'elles dépendaient largement de la composition de la population locale d'immigrés. Les auteurs de cette étude ont conclu que *M. tuberculosis* est organisé en plusieurs grandes populations génétiquement différenciées qui sont à leur tour corrélées directement et de manière stable avec des populations d'hôtes délimitées selon leur lieu d'origine. Une publication postérieure de la même équipe de chercheurs a confirmé cette compatibilité variable entre hôtes et agents pathogènes, la structure de la population mondiale de *M. tuberculosis* étant définie par six lignées phylogéographiques typées par RD/LSP, dont chacune est associée à des populations humaines sympatriques spécifiques : la lignée indo-océanique, la lignée est-asiatique, la lignée est-africaine/indienne, la lignée euro-américaine et deux lignées ouest-africaines (Gagneux *et al.,* 2006).

Tableau 1. Comparaison de la nomenclature des lignées de *M. tuberculosis* fondée sur le spoligotypage avec celles fondées sur les PGG, les groupes de grappes SNP (SCG), le typage SNP et le typage LSP.

Spoligotyping-based (Filliol 2003)	PGG	SCG (Filliol 2006)	SNP-based (Gutacker 2006)	LSP (Gagneux 2006)
East-African-Indian (EAI)	PGG1	SCG 1	sSNP-I	Indo-Oceanic
Beijing	PGG1	SCG 2	sSNP-II	East-Asian
Central-Asian (CAS)	PGG1	SCG 3a	sSNP-IIA	East-African-Indian
Haarlem	PGG2	SCG 3b	sSNP-III	Euro-American
X1	PGG2	SCG 3c	sSNP-IV	Euro-American
X1,X2,X3	PGG2	SCG 4	sSNP-V	Euro-American
LAM	PGG2	SCG 5	sSNP-VI	Euro-American
T (Miscellaneous)	PGG2-3	SCG 6	sSNP-VII sSNP-VIII	Euro-American
Bovis	PGG1	SCG 7	(MTBC)	(MTBC)
M. africanum	PGG1	NA	NA	West-African 1
M. africanum	PGG1	NA	NA	West-African 2

Le **tableau 1** récapitule les correspondances existant entre différentes nomenclatures des lignées génotypiques. Il convient de souligner qu'il faut retenir le nom du marqueur utilisé lorsque l'on désigne une lignée, en particulier la lignée « East-African Indian » (EAI) qui fait référence à deux groupes de *M. tuberculosis* totalement différents selon qu'il s'agit du spoligotypage ou du typage LSP. Une observation intéressante est la bonne congruence qui existe entre le spoligotypage et le typage SNP (Filliol *et al.,* 2006), les spoligotypes EAI et « Beijing » coïncidant respectivement avec les groupes SCG 1 et SCG 2 ; les spoligotypes X et « Central Asian » (CAS) sont également associés à un SCG et/ou sous-groupe donnée. La corrélation avec les SCG n'est pas aussi franche pour les autres lignées. De plus, les différentes lignées identifiées par spoligotypage coïncident bien avec les anciens PGG, si bien

Réseaux

que l'on peut provisoirement classer les souches du MTBC en un groupe TbD1+/PGG1 ancestral (sous-ensemble 1 : *M. africanum* et EAI), un groupe TbD1–/PGG1 moderne (sous-ensemble 2 : « Beijing » et CAS) et un groupe TbD1–/PGG2/3 d'évolution récente (sous-ensemble 3 : « Haarlem », X, S, T et « Latin American and Mediterranean » ou LAM). Néanmoins, les inférences épidémiologiques et phylogénétiques ne sont pas toujours faciles à faire en raison du manque de connaissance des mécanismes mutationnels qui donnent lieu au polymorphisme de ces cibles génomiques. Des études récentes ont démontré que des souches de MTBC non phylogénétiquement apparentées pouvaient parfois présenter le même profil de spoligotypage à la suite d'évènements mutationnels indépendants (Fenner *et al.,* 2011), observation qui corrobore le fait que le spoligotypage est plus enclin à l'homoplasie que les MIRU-VNTR (Comas *et al.,* 2009). En outre, le spoligotypage présente un faible pouvoir discriminant pour les familles associées à l'absence de gros blocs d'espaceurs, par exemple la lignée « Beijing » (Allix-Béguec *et al.,* 2014). Pour toutes ces raisons, nous recommandons d'effectuer une analyse phylogénétique plus fine des clones de MTBC en circulation qui sont les plus significatifs, avec plusieurs marqueurs génétiques, et de comparer les résultats obtenus aux données existantes au niveau mondial – tâche complexe en soi mais réalisable aujourd'hui grâce à des grandes bases de données internationales offrant cette possibilité.

État des lieux des bases de données sur le génotypage de la TB

Notre connaissance de la TB n'a jamais été aussi grande qu'aujourd'hui, mais qu'en est-il de notre capacité à comparer les données produites à toutes les autres données accumulées au fil des ans ? Sommes-nous réellement capables de comparer instantanément les informations génétiques dont nous disposons sur les souches de MTBC en circulation, avec toutes les données démographiques, cliniques, bactériologiques et épidémiologiques enregistrées à des endroits différents ? La nécessité des bases de données ne fait aucun doute dans un tel contexte, et la conception et la constitution de bases de données, pour le contrôle ou la surveillance de la TB ainsi que d'autres maladies infectieuses, vont certainement constituer une étape déterminante pour atteindre les objectifs du Millénaire pour le développement (OMD) fixés par l'OMS pour 2015 (OMS, 2006). En effet, les bases de données permettent de stocker d'énormes quantités d'informations de manière structurée, facilitant le traitement des données et les recherches et simplifiant le processus décisionnel grâce à la fouille de données guidée par les connaissances. Cependant, à l'ère actuelle du *Big Data*, ces bases de données doivent être constamment mises à jour et entretenues, comme du patrimoine ou des monuments historiques, constituant alors un ensemble de données plus volumineux que les bases de données classiques, si bien que des mesures révolutionnaires doivent être prises en matière de gestion, d'analyse et d'accessibilité des données biologiques (Howe *et al.,* 2008). Depuis deux ans, plusieurs bases de données et outils en ligne ont été mis au point dans le domaine de la TB, plus précisément pour étudier l'évolution et l'épidémiologie moléculaire de la TB ; en voici quelques exemples :

- SpolDB4 et SITVITWEB sont des bases de données de génotypage conçues à l'IPG (Brudey *et al.,* 2006 ; Demay *et al.,* 2012). La dernière version de SITVITWEB est une base de données multimarqueurs qui contient les données de génotypage de 62 582 isolats cliniques correspondant à des patients originaires de 153 pays (105 pays d'isolement). Les techniques de typage utilisées sont les suivantes : (a) spoligotypage, n = 7 105 profils de 58 180 isolats cliniques, regroupés en 2 740 types communs ou SIT (*Spoligotype International Types*) (n = 53 816 isolats cliniques) et 4 364 profils orphelins ; (a) MIRU-VNTR 12 loci, n = 2 379 profils de 8 161 isolats cliniques, regroupés en 847 types communs ou MIT (*MIRU International Types*) (n = 6 626 isolats cliniques) et 1 533 profils orphelins ; (c) séquences exactement répétées en tandem (ETR, *Exact Tandem Repeats*) 5 loci, n = 458 profils de 4 626 isolats cliniques, regroupés en 245 types communs ou VIT (*VNTR International Types*) (n = 4 413 isolats cliniques) et 213 profils orphelins. La base de données SITVITWEB peut être consultée librement sur : http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE
- SpolTools propose des applications en ligne et un outil de visualisation permettant de manipuler et d'analyser les données de sérotypage du MTBC (Reyes *et al.,* 2008 ; Tang *et al.,* 2008). Il comprend également une banque en ligne d'isolats spoligotypés, compilée à partir de la littérature publiée (actuellement 30 ensembles de données contenant 1 179 profils de spoligotypage correspondant à 6 278 isolats). Il permet notamment de dessiner des arbres Spoligoforest qui illustrent les relations d'évolution existant entre différents spoligotypes dans un milieu donné. SpolTools est accessible à l'adresse suivante : http://www.emi.unsw.edu.au/spolTools/
- MIRU-VNTRplus est un outil en ligne conçu pour analyser les données de typage moléculaire liées à la TB, en particulier les formats MIRU-VNTR 12, 15 et 24 loci (Allix-Béguec *et al.,* 2008 ; Weniger *et al.,* 2010). Les outils d'exploration de données permettent de rechercher des souches similaires, de créer des arbres phylogénétiques et des arbres de recouvrement minimaux, et de cartographier des données géographiques. La base de données fournit en outre des résultats détaillés (origine géographique, profils de sensibilité aux médicaments, lignées génétiques et profil de spoligotypage, profils SNP et LSP et empreintes IS6110-RFLP), sur une collection de 186 souches de référence bien caractérisées. MIRU-VNTRplus est accessible à l'adresse suivante : http://www.miru-vntrplus.org/
- TB GIMS (*TB Genotyping Information Management System*) est un système en ligne sécurisé conçu pour améliorer l'accès aux données de génotypage ainsi que leur diffusion à l'échelle nationale aux États-Unis (CDC 2010]. Il stocke et gère les données de génotypage relatives aux patients atteints de TB aux États-Unis ; il permet aux utilisateurs habilités d'envoyer des isolats de MTBC aux laboratoires de génotypage participants et de les suivre en ligne ; il informe immédiatement les laboratoires et les programmes travaillant sur la TB des résultats de génotypage obtenus et des mises à jour effectuées ; il corrèle les données des isolats avec les données de surveillance obtenues au niveau des patients ; il génère des rapports sur les grappes de génotypes, notamment sur la répartition des génotypes à l'échelle nationale ; et il fournit des cartes des grappes de génotypes à l'échelle du pays, des états ou des comtés. Malheureusement, cette base de données n'est pas accessible au public.

Réseaux

- Mbovis.org est une base de données de spoligotypage contenant plus de 1 400 profils appartenant aux lignées de MTBC à délétion RD9 suivantes : *M. africanum, M. bovis* (antilope), *M. microti, M. pinnipedii, M. caprae* et *M. bovis* (Smith et Upton, 2012). Cette base de données peut être consultée sur : http://www.mbovis.org/
- MycoDB.es est une base de données espagnole sur la TB animale (Rodriguez-Campos *et al.,* 2012) qui a été créée pour servir d'outil épidémiologique au niveau national (Espagne). Elle contient 401 profils de spoligotypage différents de 17 273 isolats appartenant à *M. bovis, M. caprae* et *M. tuberculosis*, ainsi qu'une quantité limitée de données de MIRU-VNTR. Malheureusement, l'accès à cette base de données est strictement réservé à l'agence espagnole de santé animale (*Centro de Vigilencia Sanitaria Veterinaria, VISAVET*) : http://www.vigilanciasanitaria.es/mycodb/
- TB-Lineage est un outil en ligne qui permet de classer et d'analyser les génotypes du MTBC en grandes lignées, sur la base du spoligotype et éventuellement du profil MIRU 24 loci (Shabbeer *et al.,* 2012). Il a été conçu et testé avec des données de génotypage obtenues auprès des CDC (*Centers for Disease Control and Prevention, centres épidémiologiques*) d'Atlanta, sur 37 066 isolats cliniques correspondant à 3 198 profils de spoligotypage et 5 430 profils de MIRU-VNTR. Cependant, en l'absence de données de MIRU 24 loci, le système utilise les prédictions effectuées avec un classifieur bayésien naïf sur la base des seules données de spoligotypage. La précision de la classification automatique est supérieure à 99 % avec les données de spoligotypage et de MIRU24 ; avec les spoligotypes seuls, elle est supérieure à 95 %. TB-Lineage est disponible librement sur : http://tbinsight.cs.rpi.edu/run_tb_lineage.html. Ce site propose également un outil qui permet de créer des arbres Spoligoforest afin de visualiser la diversité et la corrélation génétiques des génotypes et des lignées associées.
- Tbvar est une base de données interrogeable qui utilise un pipeline de calcul systématique permettant d'annoter les variantes éventuelles, sur le plan fonctionnel et/ou en matière de résistance aux antibiotiques, par rapport aux données de re-séquençage clinique du MTBC (Joshi *et al.,* 2013). Pour cela, les auteurs ont réanalysé des ensembles de données de re-séquençage correspondant à plus de 450 isolats de MTBC disponibles dans le domaine public afin de produire une carte exhaustive des variomes, constituée de plus de 29 000 variations mononucléotidiques. Cette base de données permet de faire des recherches par lieu des variants (1417019, 3037367, 4222628, etc.) ; gènes (*katG, pncA, gyrA,* etc.) ; RvID (Rv1059, Rv1069c, Rv3693, etc.) ; ou plage de positions dans le génome (10000-15000 ; 30000-35000 ; 80000-85000, etc.) ; elle est accessible à l'adresse suivante : http://genome.igib.res.in/tbvar/
- InTB est une interface en ligne qui permet le stockage et l'analyse intégrés d'informations cliniques et sociodémographiques et de données de typage moléculaire sur la TB (Soares *et al.,* 2013). Ce système permet d'introduire et de télécharger des données de génotypage standards, conjointement avec une large gamme de variables cliniques et sociodémographiques utilisées pour caractériser la maladie. Il permet également de classer les nouveaux isolats dans un ensemble d'isolats bien caractérisés sur la base d'étalons internes, de représenter graphiquement plusieurs types de données et de créer des arbres pour des sous-ensembles de données filtrés combinant données moléculaires et informations cliniques/sociodémographiques. Développés au moyen d'un logiciel *open source*, l'intégralité du code source et des packages prêts à l'emploi sont disponibles à l'adresse suivante : http://www.evocell.org/inTB.

Bases de données de génotypage de la TB élaborées à l'IPG

La première base de données a été créée à l'IPG il y a plus de quinze ans, lorsqu'un stagiaire de deuxième cycle nommé Jérôme Maïsetti a pris l'initiative de saisir dans un tableau Excel les profils de spoligotypage dont nous disposions sur nos propres isolats antillais (n = 218 souches), puis de les mettre en commun avec des données publiées (n = 392 isolats) provenant d'autres pays. Une fois les profils triés, nous nous sommes rendu compte que nous pouvions non seulement identifier des profils prédominants, mais également tracer l'origine des souches et leurs éventuels déplacements. Cette base de données de 610 spoligotypes a été provisoirement nommée « SpolDB1 » et a conduit à la toute première description de 69 grands profils de spoligotypage, permettant de mieux comprendre l'origine et la transmission de la TB (Sola *et al.,* 1999). La création de SpolDB1 a été suivie du lancement de SpolDB2, contenant des données sur 3 319 isolats (Sola *et al.,* 2001), puis de SpolDB3, sur 13 008 isolats regroupés en 813 types communs (contenant 11 708 isolats) et 1 300 profils orphelins (Filliol *et al.,* 2002 ; 2003). Plus récemment, l'élaboration de la quatrième version SpolDB4 au format MySQL en 2006, n = 39 295 isolats cliniques (Brudey *et al.,* 2006), et de SITVITWEB en 2012, n = 62 582 isolats cliniques (Demay *et al.,* 2012), a dressé un portrait plus précis de la phylogéographie des lignées génotypiques de MTBC en circulation dans le monde. À ce jour, la dernière version de notre base de données, SITVIT2 (n = 111 635 isolats), contient les données de génotypage d'environ deux fois plus de souches que la version précédente ; elle sera publiée en 2014. Il convient de souligner que ces deux versions récentes sont des bases de données multimarqueurs au format MySQL qui contiennent des données de spoligotypage et de typage MIRU-VNTR, limitées aux MIRU 12 loci dans SITVITWEB ; et des données de MIRU-VNTR 12, 15 et 24 loci dans SITVIT2. Sur ces dernières versions, une interface en ligne permet à l'utilisateur de rechercher des souches dans la base de données au moyen de critères tels que l'année, le pays d'isolement, le pays d'origine ou le nom du chercheur ; cette interface permet en outre d'effectuer des recherches mixtes dans SITVIT2, permettant d'obtenir les données de génotypage de certaines souches, conjointement avec leur répartition géographique, ainsi que les données disponibles sur la résistance aux antibiotiques ou les caractéristiques démographiques et épidémiologiques.
Grâce à l'élaboration de ces bases de données successives, l'IPG a considérablement amélioré sa connaissance du génotypage et de la phylogénie/phylogéographie de la TB dans le monde. Les différentes versions de ces bases de données ont permis la réalisation d'un nombre significatif d'études bilatérales et multilatérales, comme le prouve le très grand nombre de citations des bases de données successives (interrogation effectuée sur Google Scholar le 31 mars 2014) : SpolDB3 (Filliol *et al.,* 2002 ; 2003), 180 + 220, soit un total de 400 citations ; SpolDB4 (Brudey *et al.,* 2006), 661 citations ; et SITVITWEB (Demay *et al.,* 2012), 67 citations, bien que cette base de données soit en ligne depuis seulement un an. Nous sommes donc certains que la prochaine version

Réseaux

SITVIT2 trouvera elle aussi sa place auprès de la communauté scientifique en tant qu'outil utile non seulement pour la génétique moléculaire des populations, la démographie historique et la modélisation épidémiologique de la TB, mais aussi pour les analyses génétiques fondamentales.

Présentation de SITVIT2

Les fonctionnalités du futur site Internet SITVIT2 seront améliorées par rapport à la version actuelle de SITVITWEB, tant sur le plan numérique qu'au niveau de l'interface de recherche ; la désignation des lignées demeurera quant à elle pratiquement inchangée, à quelques exceptions près. À l'heure où nous écrivons ces lignes, SITVIT2 contient au total 111 635 isolats cliniques de MTBC prélevés chez des patients originaires de 169 pays. Pour la collecte des données, nous avons enrichi la base de données avec les résultats de génotypage obtenus à l'IPG ou transmis par différents laboratoires co-investigateurs et partenaires, ou avec des données extraites d'études publiées (Demay *et al.,* 2012). Le site Internet a été développé en langage JSP (*Java Server Pages*) et est hébergé sur un serveur d'applications Apache Tomcat gratuit (http://tomcat.apache.org), dans les locaux de l'IPG. La technologie Java a été utilisée comme décrit précédemment (Demay *et al.,* 2012). Comme sur les versions précédentes, la description des caractères génétiques des isolats cliniques consultable dans SITVIT2 repose sur une clé d'identification unique (IsoNumber) qui reprend les informations relatives au pays d'isolement, le code du laboratoire, l'année d'isolement, un code pour la résistance aux médicaments (0 à 4) et un numéro d'isolat unique attribué par le laboratoire ou l'hôpital participant. Conformément aux principes éthiques relatifs au traitement électronique des données, cette procédure permet d'attribuer un numéro anonyme que seule la personne ayant fourni les données (le laboratoire de microbiologie qui a transmis les données), et aucun autre utilisateur, peut décoder pour remonter jusqu'aux informations relatives au patient. SITVIT2 utilise le système de marquage automatique de SpolDB4 et de SITVITWEB, qui attribue un numéro SIT à chaque spoligotype présent dans 2 souches ou plus de la base de données, et un numéro MIT à chaque profil MIRU présent dans 2 souches ou plus. Les numéros MIT relatifs aux formats MIRU-VNTR 12, 15 et 24 loci sont respectivement dénommés 12-MIT, 15-MIT et 24-MIT, tandis que ceux réservés aux ETR sont dénommés VIT. Le terme « orphelin » fait quant à lui référence aux profils décrits pour un seul isolat qui ne correspond à aucun des profils enregistrés dans la banque de la base de données SITVIT2.
Parmi les fonctionnalités du site Internet, on peut notamment citer plusieurs outils de recherche pour la conversion des formats de spoligotype (de binaire en octal et vice versa), l'envoi et l'analyse des données de spoligotypage et de MIRU sous différents formats, ainsi que la recherche par critères pour les marqueurs, seuls ou combinés, l'année et le pays d'isolement/d'origine, le nom du chercheur, des cartes de répartition géographique et les informations connexes sur la démographie, l'épidémiologie et la résistance aux antituberculeux. Il est donc possible d'effectuer des recherches individuelles ou groupées en téléchargeant un fichier Excel correctement formaté (modèle disponible sur le site). Pour toutes les recherches, l'utilisateur pourrait s'attendre à obtenir un rapport détaillé sur les marqueurs (numéros SIT et MIT), les lignées phylogénétiques et les informations connexes disponibles dans la base de données sous un format anonymisé. Parmi les améliorations prévues, il sera possible de visualiser les correspondances de nomenclature des profils MIT dans SITVIT2, d'après la nomenclature de MIRU-VNTRplus (on notera qu'une comparaison des lignées entre les deux bases de données n'a pas révélé de différences significatives ; résultats non présentés).

Grandes lignées phylogéniques de SITVIT2

Dans SITVIT2, les souches sont classées en grands clades phylogénétiques, selon les signatures décrites précédemment, comprenant plusieurs membres du MTBC (AFRI, *M. africanum* ; BOV, *M. bovis* ; CANETTII, *M. canettii* ; MICROTI, *M. microti* ; PINI, *M. pinnipedii*), ainsi que *M. tuberculosis* stricto sensu, à savoir les clades : « Beijing », CAS (Central Asian), EAI (East-African Indian), « Haarlem/Ural », LAM (Latin-American-Mediterranean), « Cameroon », « Turkey », « Manu », X (lignée à faible de nombre de bandes IS6110) et le clade T, mal défini. Il convient de souligner que certains spoligotypes préalablement classés dans les sous-lignées H3/H4 au sein de la famille « Haarlem » ont récemment été reclassés dans le clade « Ural » (Mokrousov, 2012) ; il s'agit notamment de profils appartenant à la sous-lignée H4 renommés « Ural 2 » et de certains profils préalablement classés dans la sous-lignée H3, mais présentant une autre signature spécifique (présence de l'espaceur 2, absence des espaceurs 29 à 31 et 33 à 36), aujourd'hui renommés « Ural 1 ». De surcroît, deux sous-lignées LAM ont récemment été élevées au niveau de lignée indépendante : LAM10-CAM pour la lignée « Cameroon » (Koro Koro *et al.,* 2013) et LAM7-TUR pour la lignée « Turkey » (Abadia *et al.,* 2010 ; Kisa *et al.,* 2012). Nous avons conservé cette nomenclature pour les lignées génotypiques identifiées par spoligotypage, dans la mesure où elle s'est déjà révélée utile pour des études d'épidémiologie moléculaire locales ou internationales, ainsi que pour suivre l'évolution et la génétique quantitative des bacilles tuberculeux à l'échelle mondiale. On notera que la répartition des isolats cliniques de SITVIT2 est étudiée tant à l'échelle nationale qu'au niveau macrogéographique, par sous-régions, conformément à la classification des Nations Unies (voir http://unstats.un.org/unsd/methods/m49/m49regin.htm) ; régions : AFRI (Afrique), AMER (Amériques), ASIA (Asie), EURO (Europe) et OCE (Océanie), subdivisées en : E (est), M (milieu), C (centre), N (nord), S (sud), SE (sud-est) et W (ouest). Selon ce système de classification, les Antilles (CARIB) appartiennent aux Amériques, tandis que l'Océanie est subdivisée en 4 sous-régions : AUST (Australasie), MEL (Mélanésie), MIC (Micronésie) et POLY (Polynésie). Il convient de signaler que la Russie s'est vue attribuer une nouvelle sous-région (Asie du Nord), au lieu d'être intégrée à l'Europe de l'Est.
Les lecteurs sont invités à se reporter au **tableau 2** pour une comparaison rapide de SITVITWEB et SITVIT2 et des lignées phylogénétiques correspondantes dans les 2 versions, ainsi qu'à la **figure 1** qui illustre l'évolution des souches entre les 2 versions, par sous-région géographique. L'augmentation la plus significative entre les 2 versions concerne l'Europe du Sud et l'Asie de l'Est, suivies de l'Amérique Centrale et du Sud, de l'Afrique du Nord, de l'Est et de l'Ouest, et de l'Europe du Nord et de l'Ouest **(figure 1)**. Quant aux lignées, la proportion du

Réseaux

génotype « Beijing » n'a pas varié de manière significative entre SITVITWEB et SITVIT2 (représentant respectivement 9,84 % contre 9,72 % des isolats au niveau mondial ; tableau 2) ; ce génotype est prédominant en Asie et sa proportion est significative en Amérique du Nord, en Afrique du Sud et en Australasie. Les proportions sont similaires pour les lignées CAS (3,69 % contre 3,91 %), « Cameroon » (anciennement LAM10-CAM : 1,04 % contre 0,98 %), « Turkey » (anciennement LAM7-TUR : 0,59 % contre 0,53 %) et « Manu » (1,08 % contre 0,95 %). Il convient cependant de noter que la plus forte augmentation concerne les souches de *M. bovis* (10,36 % contre 23,06 %), ce qui souligne le potentiel de SITVIT2 pour l'étude de l'épidémiologie de *M. tuberculosis* ainsi que l'étude de la TB bovine.

Tableau 2. Comparaison rapide des bases de données SITVITWEB et SITVIT2 et des grandes lignées phylogénétiques correspondantes du MTBC.

Major Lineages*	SITVITWEB * (n=62582)		SITVIT2 (n=111635)	
	Nb	%	Nb	%
Beijing	6159	9.84	10850	9.72
AFRI	695	1.11	965	0.86
BOV	6486	10.36	25741	23.06
CANETTII	12	0.02	12	0.01
CAS	2480	3.96	4362	3.91
EAI	4674	7.47	6617	5.93
Haarlem/Ural	7058	11.28	10580	9.48
LAM	8042	12.85	12245	10.97
Cameroon (previously LAM10-CAM)	650	1.04	1095	0.98
Turkey (previously LAM7-TUR)	370	0.59	593	0.53
Manu	675	1.08	1064	0.95
MICROTI	29	0.05	29	0.03
PINI	152	0.24	159	0.14
S	1151	1.84	1606	1.44
T	12038	19.24	17947	16.08
X	4088	6.53	4683	4.19

Cartes de répartition mondiale de SITVIT2

La carte de répartition mondiale des grandes lignées de SITVIT2, représentée sur la **figure 2**, illustre les géospécificités phylogéographiques mondiales des isolats de MTBC, donnant une vue d'ensemble de la situation en 2014. Bien que ces géospécificités aient déjà été évoquées dans des études précédentes fondées à la fois sur le spoligotypage (Brudey *et al.,* 2006 ; Demay *et al.,* 2012) et sur le typage LSP (Gagneux *et al.,* 2006), cette carte confirme le profil de répartition et les spécificités actuels grâce aux données conservées dans la base de données SITVIT2. De plus, la reclassification de certaines lignées qui n'apparaissaient pas dans les versions précédentes, comme « Ural », montre que cette lignée est très présente en Russie, en Asie centrale, en Asie du Sud et en Asie de l'Ouest, ainsi qu'en Finlande (Europe du Nord), qui partage une frontière commune et des liens privilégiés avec la Russie, en ayant notamment joué le rôle de zone tampon dans une succession de guerres entre la Russie et la Suède au cours du 18e siècle (http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=ad02). Nous avons par ailleurs confirmé le réétiquetage de LAM7-TUR comme lignée « Turkey » et celui de LAM10-CAM en lignée « Cameroon », grâce à leur répartition phylogéographique dans SITVIT2 **(figure 2)**. Il est important de souligner que la lignée « Turkey » progresse dans les pays d'Europe de l'Est (représentant près de 6 % de toutes les souches de MTBC d'Europe de l'Est en 2014, contre moins de 2 % jusqu'en 2006 ; résultats non présentés). Il convient également de noter la diminution continue de la proportion de la lignée AFRI en Afrique de l'Ouest, qui représentait environ 37 % de toutes les souches de MTBC dans cette sous-région jusqu'en 2006, contre 29 % en 2014. Cette observation corrobore notre précédente hypothèse selon laquelle les souches de *M. africanum* ancestrales d'Afrique de l'Ouest sont peu à peu remplacées par des lignées récentes de MTBC comme « Cameroon » ou d'autres lignées euro-américaines (Groenheit *et al.,* 2011). Enfin et surtout, bien que les proportions globales de CAS n'aient pas beaucoup évolué entre SITVITWEB et SITVIT2 (3,69 % contre 3,91 %), on observe une augmentation en Asie de l'Ouest (12 % contre 18 %) et en Afrique de l'Est (10 % contre 15 %).

Recherche de corrélations fortes entre lignées phylogéniques et caractéristiques démographiques et épidémiologiques dans SITVIT2

Un utilisateur pourra obtenir des données phylogénétiques associées à un certain nombre de paramètres différents tels que l'incidence de la maladie représentée sur des cartes disponibles auprès de l'OMS, la démographie (âge, sex-ratio) ou d'autres caractéristiques dans différentes sous-régions, afin de mettre en évidence les spécificités de chaque pays, région ou population. Si l'on considère que la propagation des clones MDR et XDR de TB en population générale représente une menace de premier plan en matière de lutte antituberculeuse, l'intérêt d'une telle base de données est également de donner une vue d'ensemble de la situation actuelle et de localiser toute émergence éventuelle pour les autorités de santé publique. Ce géoréférencement de nos données, au moyen d'une API Google, peut déjà former un SIG constituant un pendant bactériologique efficace de l'Atlas mondial des maladies infectieuses de l'OMS (http://apps.who.int/globalatlas/), réunissant en une seule plate-forme électronique l'analyse et l'interprétation des données de génotypage et des informations relatives à la démographie, aux conditions socioéconomiques ou aux facteurs environnementaux. Nous pensons qu'une telle cartographie devrait idéalement être associée à des outils et logiciels statistiques et bio-informatiques appropriés afin de mieux décrire le paysage génétique de la TB. Outre les outils

*Les souches sont classées en grands clades phylogénétiques, selon les signatures décrites précédemment (Demay et al., 2012) ; elles comprennent plusieurs membres du MTBC (AFRI, *M. africanum* ; BOV, *M. bovis* ; CANETTII, *M. canettii* ; MICROTI, *M. microti* ; PINI, *M. pinnipedii*), ainsi que des lignées/sous-lignées de M. tuberculosis stricto sensu (les sous-lignées ne sont pas indiquées) : le clade « Beijing », le clade CAS, le clade EAI, les clades « Haarlem/Ural », le clade LAM, les lignées « Cameroon » et « Turkey », la famille « Manu », le clade X à faible nombre de bandes IS6110 et le clade T, mal défini.

Réseaux

Figure 1. (A) Répartition mondiale des isolats de MTBC enregistrés dans SITVITWEB (cercles bleus) et SITVIT2 (cercles verts), le nombre d'isolats par sous-région étant indiqué à l'intérieur de chaque cercle. (B) Évolution progressive d'une série de bases de données (de SpolDB2 à SITVIT2), selon le nombre d'isolats de MTBC contenus dans chaque base de données.

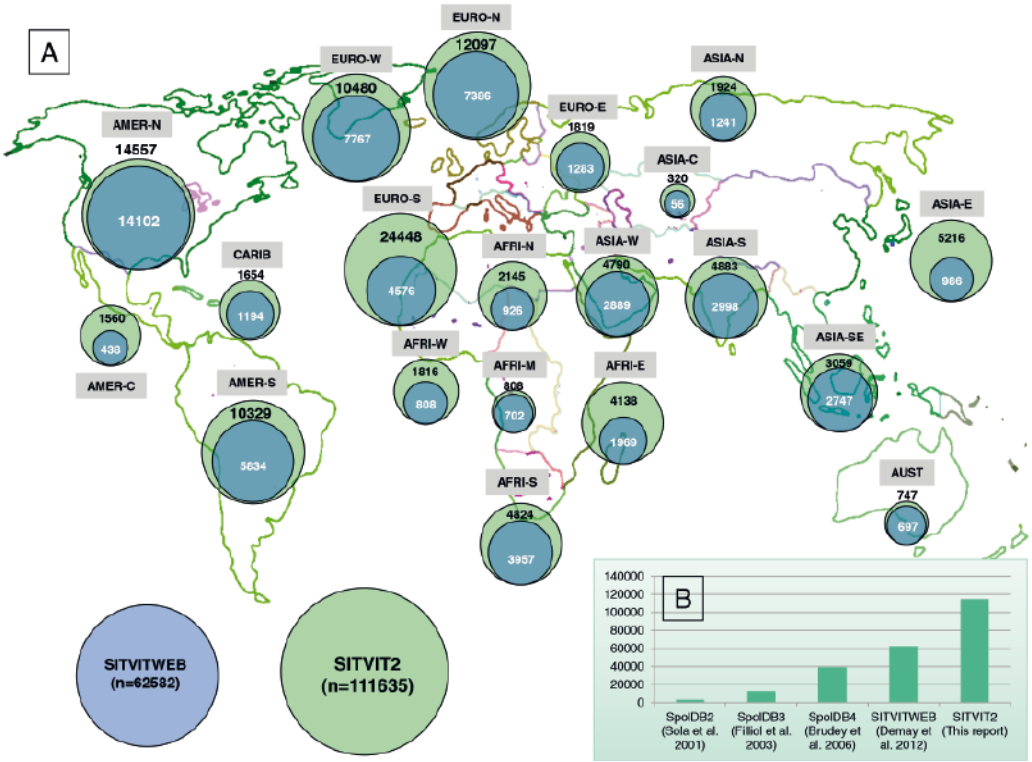
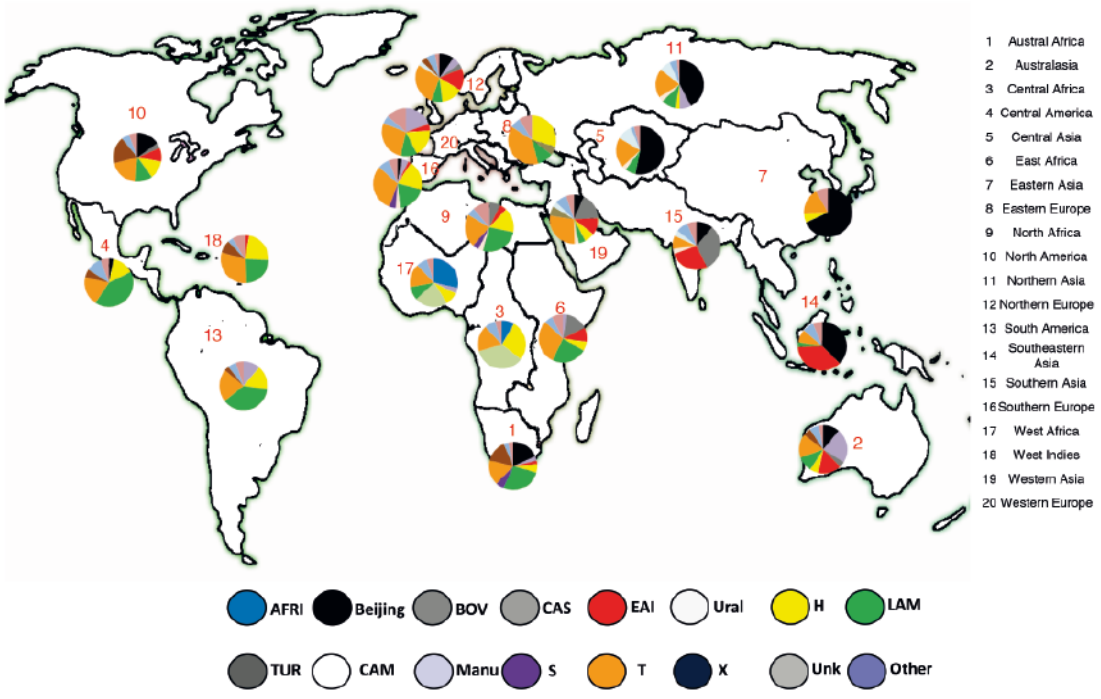


Figure 2 : Répartition mondiale des lignées contenues dans la base de données SITVIT2.



Réseaux

et logiciels décrits dans les études précédentes (Brudey *et al.,* 2006 ; Demay *et al.,* 2012), nous utilisons aussi les outils suivants pour l'analyse courante des données de SITVIT2 :
(a) Le logiciel STATA, version 12, pour des analyses descriptives et univariées.
(b) Le logiciel R, version 2.14.1, pour calculer l'Odds Ratios (OR) et les bornes de l'intervalle de confiance (IC) à 95 %.
(c) Le test du chi carré de Pearson et le test exact de Fisher pour comparer les fortes corrélations existant entre les données de génotypage (types communs/lignées) et les caractéristiques démographiques, épidémiologiques ou socioéconomiques (des valeurs p inférieures à 0,05 étant considérée statistiquement significatives).
(d) Des arbres de recouvrement minimal (ARM) sont construits à partir des données de génotypage (spoligotypes, profils MIRU) à l'aide du logiciel BioNumerics, version 6.6 (Applied Maths, Sint-Martens-Latem, Belgique). Il s'agit de graphes non orientés associés sur lesquels tous les profils sont reliés, avec le moins de liens possible entre les voisins les plus proches.
(e) Le logiciel SpolTools (http://www.emi.unsw.edu.au/spolTools) permet de dessiner des arbres Spoligoforest sur la base de l'algorithme de Fruchterman et Reingold ou d'une disposition hiérarchique (Reyes *et al.,* 2008 ; Tang *et al.,* 2008). On notera que, contrairement aux ARM, les arbres Spoligoforest sont des graphes orientés (et pas forcément associés) qui représentent les relations d'évolution existant entre les profils de spoligotypage des ascendants et des descendants.
(f) Le logiciel GraphViz, disponible sur http://www.graphviz.org (Ellson *et al.,* 2002), pour colorier les arbres Spoligoforest en fonction des lignées.
(g L'application WebLogo, version 2.8.2, disponible sur http://weblogo.berkeley.edu/ (Schneider and Stephens, 1990 ; Crooks *et al.,* 2004), pour évaluer et visualiser la diversité allélique des profils de spoligotypage en fonction de leurs lignées associées. Cette méthode de représentation adaptée au spoligotypage sur 43 espaceurs a été dénommée « Spoligologos » (Driscoll *et al.,* 2002). WebLogos comprend des piles de symboles qui permettent de représenter graphiquement la diversité comparative observée pour chaque espaceur de spoligotypage – une pile pour chacun des 43 espaceurs. La hauteur totale d'une pile indique le degré de conservation d'un espaceur donné (la lettre « n » indique la présence d'un espaceur et la lettre « o » son absence). Si un espaceur est systématiquement présent ou absent en une position donnée parmi les 43 disponibles (en d'autres termes, si 100 % des souches conservent la même présence ou la même absence en une position donnée), cela correspond à 4 bits ; la hauteur de chaque symbole dans la pile indique quant à elle la fréquence relative des espaceurs absents/présents en cette position.

Exemples d'études récentes utilisant la base de données SITVIT2

Il est bien entendu impossible de citer toutes les études réalisées avec SITVIT2 ; on peut cependant mentionner quelques exemples de travaux récents qui portent sur les corrélations entre lignées phylogénétiques et paramètres démographiques et épidémiologiques :
(a) Carte de répartition géographique des lignées de MTBC identifiées par spoligotypage dans plusieurs sous-régions d'Afrique et forte spécificité phylogéographique de *M. africanum* pour l'Afrique de l'Ouest, la Guinée-Bissau étant l'épicentre (Groenheit *et al.,* 2011).
(b) Preuve que les souches de MTBC potentiellement responsables de l'épidémie de TB qui a touché la Suède il y a un siècle appartenaient à un groupe de souches PGG2/3 d'évolution récente étroitement apparentées, limité à la Suède et aux pays immédiatement voisins (Groenheit *et al.,* 2012).
(c) Analyse des corrélations phylogénétiques entre le MTBC et la résistance aux antituberculeux au Pérou, suggérant une épidémie nosocomiale clonale prolongée des isolats MDR chez des patients infectés par le VIH (Sheen *et al.,* 2013).
(d) Cartographie phylogéographique de la TB en Finlande, révélant une forte ressemblance entre la structure de la population mondiale du MTBC et une structure décrite en Suède, en particulier la prédominance de la famille euro-américaine chez les patients âgés atteints de TB ; la principale différence résidant dans la lignée « Ural », qui a été observée en grandes proportions dans les cas d'origine finlandaise (et qui a aussi été observée en Russie, en Lettonie et en Estonie), mais pas en Suède (Smit *et al.,* 2013).
(e) Au niveau mondial, l'utilisation de la base de données SITVIT2 nous a permis de découvrir une corrélation plus significative entre les lignées X et LAM et une sérologie positive au VIH ; valeur p < 0,0001 (article sous presse).
(f) Plusieurs études ont mis en évidence la corrélation existant entre la lignée « Beijing » et une résistance excessive aux médicaments, y compris la TB MDR/XDR (van Soolingen *et al.,* 1995 ; Glynn *et al.,* 2002 ; Parwati *et al.,* 2010). Ainsi, nous avons récemment analysé les corrélations phylogénétiques (MTBC divisé en 2 groupes : « Beijing » et autres lignées) avec la résistance aux antibiotiques (quantifiée comme – pansusceptible, MDR, XDR ou autre) en utilisant la base de données SITVIT2 (Couvin et Rastogi, 2014). La **figure 3A** illustre la répartition de la résistance dans différentes sous-régions. Bien que la proportion de souches résistantes soit beaucoup plus élevée pour les souches « Beijing » que pour les autres souches à l'échelle mondiale, nous observons des variations importantes quant à la répartition mondiale de la résistance aux antituberculeux. La résistance est fortement corrélée aux souches « Beijing » par rapport aux autres souches en Russie, en Asie du Sud, en Asie du Sud-Est et dans les pays européens, mais pas en Amérique, en Asie de l'Ouest, en Chine et au Japon. Si l'on analyse l'évolution de la résistance dans le temps pour les souches « Beijing » (de 1998 à 2011, **figure 3B**), on observe une progression continue de la proportion de souches MDR et XDR (et une diminution relative des souches pansusceptibles) à l'échelle mondiale depuis 2003. Enfin et surtout, nous observons également qu'un profil de spoligotypage rare mais émergent, SIT190/Beijing, présente une corrélation plus significative avec la TB MDR que le profil SIT1/Beijing classique (valeur p < 0,0001).
(g) Concernant les informations obtenues grâce à un arbre Spoligoforest, nous souhaitons citer une étude récente menée à Bagdad en Iraq (Mustafa Ali *et al.,* 2014). Les résultats obtenus, sur un total de 270 isolats de MTBC, ont montré que 2 profils spécifiques, SIT1144/T1 et SIT309/CAS1-Delhi, prédominaient dans cette étude (6,3 % pour chaque profil). Les relations d'évolution existant entre les isolats iraquiens, visibles sur un arbre Spoligoforest avec disposition hiérarchique (**figure 4**), montrent clairement que la majeure partie de la TB observée dans l'Iraq d'après-guerre se limite à 2 groupes de souches de MTBC phylogénétiquement apparentées, appartenant aux lignées génotypiques T et CAS.
(h)Concernant les informations obtenues grâce au

Réseaux

Figure 3. Caractéristiques de résistance aux médicaments des lignées de *M. tuberculosis* « Beijing » et « non-Beijing » (A) et évolution de la résistance aux médicaments dans les isolats « Beijing » entre 1998 et 2011 (B)

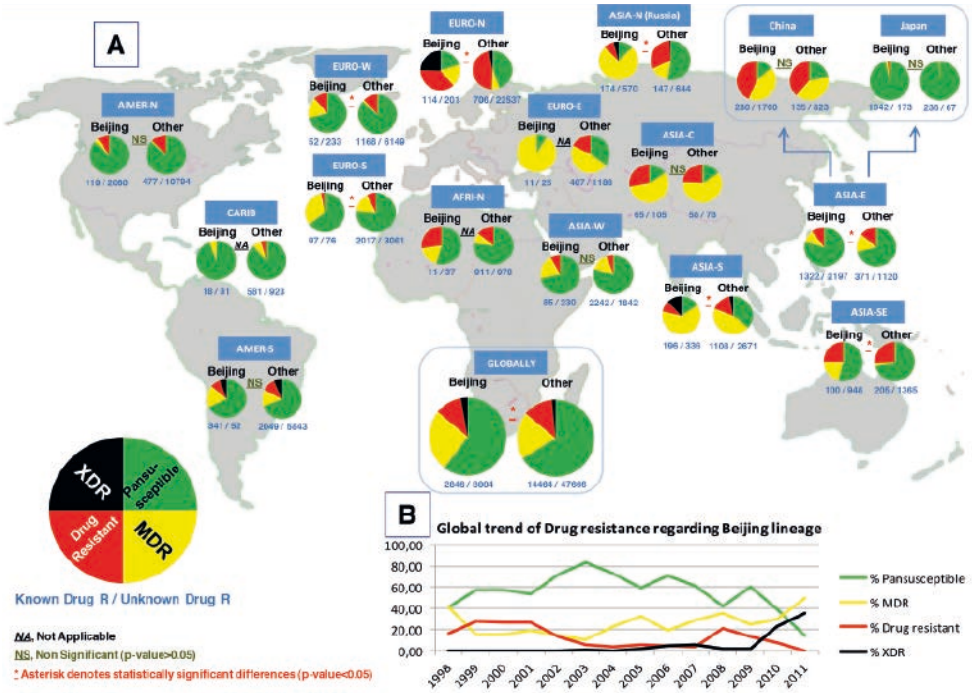
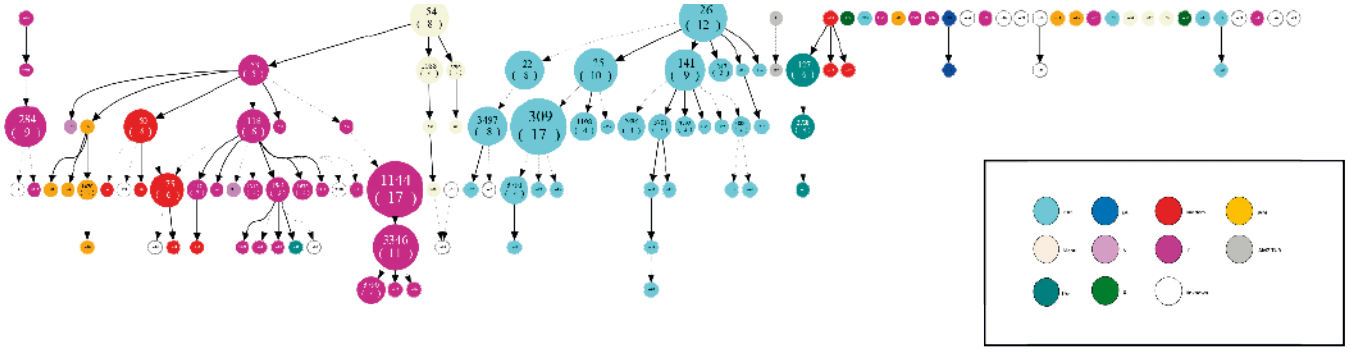
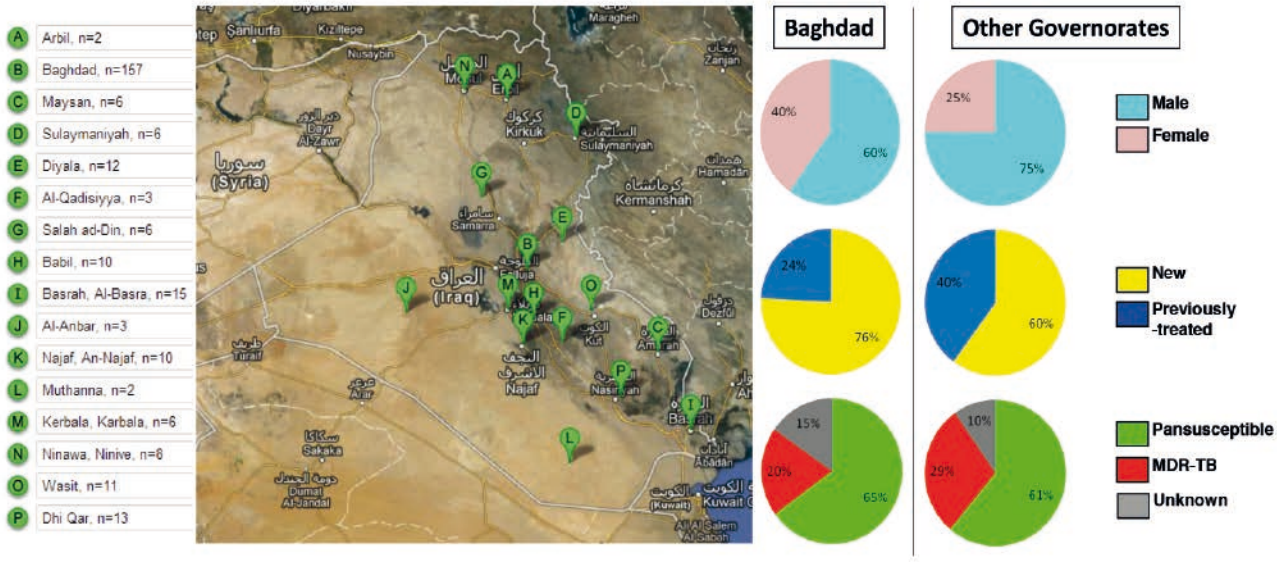


Figure 4. Graphique Spoligoforest, des parents aux descendants, lors d'une étude menée à Bagdad en Iraq (n = 270 isolats), avec une disposition hiérarchique. Sur cet arbre, chaque profil de spoligotypage est symbolisé par un nœud dont la taille est proportionnelle au nombre total d'isolats ayant ce profil. Les modifications (perte d'espaceurs) sont symbolisées par des flèches, entre les nœuds, qui pointent sur les spoligotypes des descendants. L'heuristique utilisée choisit une seule flèche entrante de poids maximal, au moyen d'un modèle de Zipf. Les traits pleins noirs relient les profils très similaires, à savoir ceux qui n'ont perdu qu'un seul espaceur (le poids maximal étant 1,0), tandis que les traits discontinus relient les poids compris entre 0,5 et 1, et les pointillés les poids inférieurs à 0,5. Il convient de noter que SIT309/CAS1-Delhi et SIT1144/T1 sont les deux plus gros nœuds (n = 17), suivis de SIT26/CAS1-Delhi (n = 12), SIT3346/T1 (n = 11) et SIT25/CAS1-Delhi (n = 10), qui sont les autres profils prédominants dans notre étude. Enfin, les isolats orphelins (doubles cercles) apparaissent principalement en positions terminales sur l'arbre ou sont des souches isolées sans liens avec les autres souches (figure élaborée à partir des données de Mustafa Ali et al., 2014).



Réseaux

Figure 5. Capture d'écran de Google Maps montrant la répartition des souches de MTBC en fonction de la ville d'isolement, à Bagdad et dans d'autres villes d'Iraq (n = 270 isolats), et comparaison des données en fonction de l'origine des patients et du rapport de masculinité, de la phase de traitement ou de la résistance aux médicaments à Bagdad et dans les autres gouvernorats (figure élaborée à partir des données de Mustafa Ali et al., 2014).



géoréférencement des données de génotypage avec une API Google, nous pouvons citer la même étude (Mustafa Ali *et al.,* 2014). Comme le montre la **figure 5**, il existe des différences significatives entre la ville de Bagdad et les autres villes iraquiennes, en matière d'informations démographiques et de données de résistance aux médicaments. En effet, avec une sex-ratio hommes/femmes de 1,49 à Bagdad, contre 3,04 dans les autres gouvernorats, la proportion de patients féminins est nettement supérieure dans la ville de Bagdad (valeur p = 0,009 ; Odds Ratio = 0,49 et IC 95 % [0,28 ; 0,86]). Le rapport entre les cas traités pour la première fois et les cas retraités varie considérablement d'un groupe à l'autre, la proportion de patients retraités étant plus élevée dans les autres gouvernorats iraquiens (valeur p = 0,007 ; OR = 2,1, IC 95 % [1,19 ; 3,62]). Enfin, le taux de TB MDR est supérieur dans les gouvernorats autres que Bagdad (p > 0,12 ; différence non statistiquement significative).

Conclusions

La collection de bases de données élaborées à l'IPG permet de disposer d'une vue d'ensemble de la situation mondiale en matière de TB. Avec la base de données SITVIT2 récemment mise à jour, nous avons par ailleurs pu décrire la circulation actuelle des souches de MTBC, sur la base de marqueurs de génotypage élargis, et nous avons également mis en évidence de fortes corrélations entre les lignées phylogénétiques de MTBC et les caractéristiques démographiques et épidémiologiques. Dans l'idéal, les prochaines améliorations devraient comprendre l'intégration d'autres marqueurs, comme les RD/LSP et les SNP, ainsi que des futures informations produites grâce au séquençage de nouvelle génération. Bien que de nombreuses caractéristiques évolutives et pathobiologiques de l'épidémie de TB actuelle restent à découvrir, la nouvelle collection de bases de données SITVIT constitue un outil de référence pour améliorer la surveillance épidémiologique de la TB et la lutte antituberculeuse.

Remerciements

Nous remercions les plus de cinq cents chercheurs qui ont fourni les données qui ont permis de constituer les bases de données SpolDB4, SITVITWEB et SITVIT2 (dont 190 ont fourni les données de génotypage de plus de 100 souches). Il convient de souligner que chaque souche et les données de génotypage respectives qui figurent dans chacune de ces bases de données font directement référence au chercheur concerné (liste exhaustive disponible sur demande). Nous remercions particulièrement Thierry Zozio, Julie Millet, Véronique Hill et Élisabeth Streit pour leurs réflexions très utiles. DC a reçu une bourse de doctorat du Fonds social européen, par l'intermédiaire du Conseil régional de Guadeloupe.

Bibliographie

1. Abadia E, Zhang J, dos Vultos T, Ritacco V, Kremer K, Aktas E, Matsumoto T, Refregier G, van Soolingen D, Gicquel B, Sola C. (2010). Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect Genet Evol.* 2010; 107: 1066–1074.

2. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. 2008.

Réseaux

Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol*. 46: 2692-2699.

3. Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, Campos-Herrero I, Mokrousov I, Niemann S, Kontsevaya I, Rastogi N, Samper S, Sng LH, Warren RM, Supply P. 2014. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. *J Clin Microbiol*. 52: 164-172.

4. Barry CE 3rd, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ, Young D. 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol*. 7: 845-855.

5. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. 1999. Analysis of rate of change of IS6110 RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. *J Infect Dis*. 180: 1238-1244.

6. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA*. 99: 3684-3689.

7. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C. 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6: 23.

8. CDC. 2010. Launch of TB Genotyping Information Management System (TB GIMS). *Morbidity and Mortality Weekly Report* (MMWR) March 19, 59(10); 300.

9. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet*. 45: 1176-1182.

10. Comas I, Homolka S, Niemann S, Gagneux S. 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS One*. 4: e7815.

11. Couvin D, Rastogi N. (2014). Tuberculosis – a global emergency: tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage. Tuberculosis (Edinb). Submitted for publication.

12. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14: 1188-1190.

13. Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, Heersma H, Lillebaek T, Quitugua T, Rastogi N, Skuce RA, Sola C, Van Soolingen D, Vincent V. 2001. Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standardised nomenclature. *Int J Tuberc Lung* Dis. 5: 216-219.

14. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N. 2012. SITVITWEB – a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect Genet Evol*. 12: 755-766.

15. Driscoll JR, Bifani PJ, Mathema B, McGarry MA, Zickas GM, Kreiswirth BN, Taber HW. 2002. Spoligologos: a bioinformatic approach to displaying and analyzing *Mycobacterium tuberculosis* data. *Emerg Infect Dis*. 8(11): 1306-9.

16. Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. 2002. Graphviz – Open Source Graph Drawing Tools. In: Mutzel P, Jünger M, Leipert S (Editors), Heidelberg: Springer-Verlag Berlin. pp. 483-484.

17. van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small PM. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol.* 31, 406-409.

18. Fang Z, Kenna DT, Doig C, Smittipat DN, Palittapongarnpim P, Watt B, Forbes KJ. 2001. Molecular evidence for independent occurrence of IS6110 insertions at the same sites of the genome of *Mycobacterium tuberculosis* in different clinical isolates. *J Bacteriol.* 183: 5279-5284.

19. Fenner L, Malla B, Ninet B, Dubuis O, Stucki D, Borrell S, Huna T, Bodmer T, Egger M, Gagneux S. (2011) "Pseudo-Beijing": evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PLoS One*. 6: e24737.

20. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, Alland D. 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for *a minima*l standard SNP set. *J Bacteriol*. 2006; 188: 759-772.

21. Filliol I, Driscoll JR, Van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Anh DD, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Källenius G, Kassa-Kelembho E, Koivula T, Ly HM, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, De Waard JH, Sola C, Rastogi N. 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis*. 2002 Nov; 8(11): 1347-9.

22. Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Kassa-Kelembho E, Ho ML, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, de Waard JH, Sola C, Rastogi N. 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. *J Clin Microbiol*. 41(5): 1963-1970.

23. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA*. 103(8): 2869-73.

24. García de Viedma D, Mokrousov I, Rastogi N. 2011. Innovations in the molecular epidemiology of tuberculosis. *Enferm Infecc Microbiol* Clin. 29 (Suppl 1): 8-13.

25. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. (2002). Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg Infect Dis*. 8(8): 843-849.

26. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. 1999. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol*. 32: 643-655.

27. Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, Riccardi F, Couvin D, Hill V, Rastogi N, Koivula T, Källenius G. 2011. The Guinea-Bissau family of *Mycobacterium tuberculosis* complex

Réseaux

revisited. *PLoS One*. 6(4): e18601.

28. Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S, Couvin D, Koivula T, Rastogi N, Källenius G. 2012. *Mycobacterium tuberculosis* strains potentially involved in the TB epidemic in Sweden a century ago. *PLoS One*. 7(10): e46848.

29. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis*. 2006; 193: 121-128.

30. Heersma HF, Kremer K, van Embden JD. 1998. Computer analysis of IS6110 RFLP patterns of *Mycobacterium tuberculosis. Methods Mol Biol.* 101: 395-422.

31. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun*. 59: 2695-2705.

32. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol*. 2008, 6(12): e311.

33. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci USA*. 2004; 101: 4871-4876.

34. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. (2008) Big data: The future of biocuration. *Nature*. 455(7209): 47-50.

35. Jagielski T, van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. 2014. Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria. *Biomed Res Int*. 2014 (Article ID 645802): 1-21.

36. Joshi KR, Dhiman H, Scaria V. (2013) tbvar: a comprehensive genome variation resource for *Mycobacterium tuberculosis*. Database Vol. 2013: article ID bat083.

37. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol.* 35: 907-914.

38. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res*. 11: 547-554.

39. Kisa O, Tarhan G, Gunal S, Albay A, Durmaz R, Saribas Z, Zozio T, Alp A, Ceyhan I, Tombak A, Rastogi N. (2012). Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey. *PLoS One*. 2012; 7(1): e30331.

40. Koro Koro F, Kamdem Simo Y, Piam FF, Noeske J, Gutierrez C, Kuaban C, Eyangoh SI. 2013. Population dynamics of *tuberculous Bacilli* in Cameroon as assessed by spoligotyping. *J Clin Microbiol.* 51: 299-302.

41. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent M. bovis. J Bacteriol. 178: 1274-1282.

42. Mokrousov I. 2012. The quiet and controversial: Ural family of *Mycobacterium tuberculosis. Infect Genet Evol.* 12: 619-629.

43. Mustafa Ali R, Trovato A, Couvin D, Al-Thwani AN, Borroni E, Dhaer FH, Rastogi N, Cirillo DM. 2014. Molecular Epidemiology and Genotyping of *Mycobacterium tuberculosis* Isolated in Baghdad. BioMed Research International, vol. 2014, Article ID 580981, 15 pages.

44. Parwati I, van Crevel R, van Soolingen D. 2010. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis*. 10: 103-111.

45. Radhakrishnan I, K MY, Kumar RA, Mundayoor S. 2001. Implications of low frequency of IS6110 in fingerprinting field isolates of *Mycobacterium tuberculosis* from Kerala, India. *J Clin Microbiol*. 39: 1683.

46. Rastogi N, Sola C. 2007. Chapter 2 – Molecular evolution of the *Mycobacterium tuberculosis* complex. In Tuberculosis 2007: from basic science to patient care, Edited by Palomino JC, Leao S, Ritacco V. 2007, 53-91, Amedeo Online Textbooks; http://pdf.flyingpublisher.com/tuberculosis2007.pdf

47. Reyes JF, Francis AR, Tanaka MM. 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. *BMC Bioinformatics*. 9: 496.

48. Rodriguez-Campos S, González S, de Juan L, Romero B, Bezos J, Casal C, Álvarez J, Fernández-de-Mera IG, Castellanos E, Mateos A, Sáez-Llorente JL, Domínguez L, Aranaz A; Spanish Network on Surveillance Monitoring of Animal Tuberculosis. 2012. A database for animal tuberculosis (mycoDB.es) within the context of the Spanish national program for eradication of bovine tuberculosis. *Infect Genet Evol.* 12(4): 877-82.

49. Rothschild BM, Martin LD, Lev G, Bercovier H, Bar-Gal GK, Greenblatt C, Donoghue H, Spigelman M, Brittain D. 2001. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis*. 33: 305-311.

50. Salamon H, Segal MR, Ponce de Leon A, Small PM. 1998. Accommodating error analysis in comparison and clustering of molecular fingerprints. *Emerg Infect Dis*. 4: 159-168.

51. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097-6100.

52. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP. 2012. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect Genet Evol.* 12: 789-797.

53. Sheen P, Couvin D, Grandjean L, Zimic M, Dominguez M, Luna G, Gilman RH, Rastogi N, Moore DA. 2013. Genetic diversity of *Mycobacterium tuberculosis* in Peru and exploration of phylogenetic associations with drug resistance. *PLoS One*. 8(6): e65873.

54. Smit PW, Haanperä M, Rantala P, Couvin D, Lyytikäinen O, Rastogi N, Ruutu P, Soini H. (2013). Molecular epidemiology of tuberculosis in Finland, 2008-2011. *PLoS One*. 8(12): e85027.

55. Smith NH, Upton P. 2012. Naming spoligotype patterns for the RD9-deleted lineage of the *Mycobacterium tuberculosis* complex; www.Mbovis.org. *Infect Genet Evol*. 12 (4), pp. 873–876.

56. Soares P, Alves RJ, Abecasis AB, Penha-Gonçalves C, Gomes MG, Pereira-Leal JB. (2013) inTB – a data integration platform for molecular and clinical epidemiological analysis of tuberculosis. *BMC Bioinformatics*. 14: 264.

57. Sola C, Devallois A, Horgen L, Maïsetti J, Filliol I, Legrand E, Rastogi N. 1999. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. *Emerg Infect Dis*. 5(3): 404-14.

58. Sola C, Filliol I, Gutierrez MC, Mokrousov I, Vincent V, Rastogi N. 2001. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. *Emerg Infect Dis*. 7(3): 390-396.

59. Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. 2001. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. *J Mol Evol*. 53: 680-689.

60. Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N. 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol*. 3: 125-133.

61. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD. (1995). Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. *J Clin Microbiol.* 33: 3234-3238.

62. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA*. 94: 9869-9874.

Réseaux

63. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J Clin Microbiol*. 39: 3563-3571.

64. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis. J Clin Microbiol*. 44: 4498-4510.

65. Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM. 2008. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. *Bioinformatics.* 24: 2414-415.

66. Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature*. 447: 279-283.

67. Zink AR, Sola C, Reischl U, Grabner W, Rastogi N, Wolf H, Nerlich AG. 2003. Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. *J Clin Microbiol*. 41: 359-367.

68.Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. *Nucleic Acids Res*. 38(Web Server issue): W326–W331.

69.WHO. 2006. The Stop TB Strategy: Building on and enhancing DOTS to meet the TB-related Millennium Development Goals. http://whqlibdoc.who.int/hq/2006/WHO_HTM_STB_2006.368_eng.pdf?ua=1.

70.WHO. 2013. Global tuberculosis report 2013.

www.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf

Article 03

Tuberculosis/ Original article

Tuberculosis – a global emergency: tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage

David Couvin, Nalin Rastogi*

WHO Supranational Tuberculosis Reference Laboratory, Tuberculosis and Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Abymes, Guadeloupe (France)

Running title: TB Global emergency: example of Beijing lineage

* Corresponding author. Mailing address: Institut Pasteur de Guadeloupe, Morne Jolivière, BP484, F97183 Abymes Cedex, Guadeloupe. Phone: +590-590-897661. Fax: +590-590-896941. E-mail: nrastogi@pasteur-guadeloupe.fr.

SUMMARY [298 words]

We argue in favor of a concerted and coordinated response to stop tuberculosis (TB) by monitoring global TB spread, drug-resistance surveillance and populations at risk using available molecular and web tools to identify circulating clones of *Mycobacterium tuberculosis* complex (MTBC). We took specific example of the Beijing lineage associated with worldwide emergence of both multiple, and extensively drug resistant (MDR/XDR)-TB. The study dataset (n=10850 isolates, 92 countries of patient origin) was extracted from our multimarker SITVIT2 database on MTBC genotyping (n=111635 isolates, 169 countries of patient origin). Epidemiological and demographic information in conjunction with spoligotyping (n=10850), MIRU-VNTR minisatellites (n=2896), and drug resistance (n=2846) data was mapped at macro-geographical (United Nations subregions) and country level, followed by statistical, bioinformatical, and phylogenetical analysis. The global male/female sex ratio was 1.96, the highest being 4.93 in Russia vs. range of 0.8-1.13 observed in Central America, Caribbean, Eastern Africa and Northern Europe ($p<0.0001$). The major patient age-group was 21-40 yrs worldwide except Japan (with majority of patients >60 yrs). Younger patients (age group 0-20 yrs) were more common in South America, South Asia, and Western Africa (25-33% of TB cases due to Beijing genotype). A continuous progression in the proportion of MDR and XDR strains is visible worldwide since 2003 and 2009 respectively. Pansusceptible TB mainly concerned older patients >60 yrs (44%) whereas Drug resistant, MDR and XDR-TB concerned patients preferentially aged 21-40 yrs (between 52 to 58%). Although the proportion of SIT1 pattern vs. other patterns was very high (93%); the proportion of MDR was highest for an emerging genotype SIT190 ($p<0.0001$). Lastly, proportion of pansusceptible strains was highest in Japan, while MDR/XDR strains were most common in Russia and Northern Europe. We underline remarkable macro/micro-geographical cleavages in phylogenetic and epidemiologic diversity of Beijing genotype, with phylogeographical specificity of certain genotypes.

Key words: *M. tuberculosis*; Beijing genotype; Epidemiology; Spoligotyping; MIRU-VNTR; Drug-resistance;

1. Introduction

Tuberculosis (TB) remains a major global health problem and one of the most important infectious diseases affecting mankind today. In 2012, an estimated 8.6 million people developed TB and 1.3 million died from the disease; poverty, malnutrition, overcrowding and immunosuppression are some risk factors which promote TB expansion [**1**]. A concerted and coordinated response to monitor and assess global TB spread, drug-resistance surveillance and populations at risk is urgently needed for global TB control. In the last decade, innovations in the molecular epidemiology of tuberculosis have helped define the global distribution of *Mycobacterium tuberculosis* lineages, to monitor the international spread of high-risk strains, to explore the evolutionary features of the bacterium, as well as to discriminate between events of recent transmission from those due to reactivation and to differentiate between recurrences due to reactivations vs. exogenous reinfections [**2**].

Much of these advances in micropopulation and macropopulation analysis of *M. tuberculosis* global epidemiology were made thanks to two major advances: PCR-based large-scale, high-throughput genotyping of *M. tuberculosis* and publicly available huge international databases in conjunction with web-tools [reviewed in **2-4**]. Although most of the earlier molecular epidemiological studies were performed using IS*6110*-RFLP [**5**], this labor-intensive methodology was replaced with alternative PCR-based strategies in the last decade. The most commonly used methodology for TB genotyping today essentially uses a combination of 2 methods – spoligotyping based on the polymorphism of the direct repeat (DR) locus [**6**] and MIRU-VNTR minisatellites used in 12–, 15–, or 24–loci formats [**7, 8**], and constitutes the gold-standard for optimal TB surveillance at national, regional, and global levels by "universal genotyping" of patient isolates. A query in PubMed on October 30th 2013 made for "IS*6110*-RFLP AND tuberculosis" returned 210 publications vs. 801 for "spoligotyping AND tuberculosis", 546 using the term "MIRU OR VNTR OR MLVA OR MIRU-VNTR AND tuberculosis", and 259 publications looking into the association of last 2 queries. In summary, the above mentioned PCR-based genotyping methods in conjunction with classical epidemiological investigations show adequate resolution for tracing TB transmission and predicting diverse strain lineages, and are used widely for TB surveillance worldwide today [**9–12**].

Furthermore, the numerical format of the spoligotype patterns and MIRU-VNTRs made it possible that the information generated is easily stored in huge international databases and made available to the international research community along with new web tools to study TB epidemiology and compare genotyping information worldwide; some examples include recently published SITVITWEB ([**13**]; http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE) and its earlier version SpolDB4 [**14**]; MIRU-VNTRplus ([**15**];http://www.miru-vntrplus.org/MIRU/index.faces); and TB-Lineage ([**16**]; http://tbinsight.cs.rpi.edu/run_tb_lineage.html).

Working on tools and methods to monitor, understand, and control the TB epidemic for last many years at Institut Pasteur de la Guadeloupe, we recently undertook the construction of a newer database named SITVIT2; the version under development already contains spoligotyping data on 111,635 *M. tuberculosis* complex clinical isolates from 169 countries of patient origin – almost twice as compared to the SITVITWEB developed and released publicly in 2012 (which incorporated spoligotyping data on 62,582 clinical isolates corresponding to 153 countries of patient origin [**13**]). Addition of new tools to the database and supplementary information allows global mapping of combined search results on *M. tuberculosis* genotypes, epidemiological and demographic information, and drug resistance.

In this paper, we evaluate the usefulness of such an approach by studying the specific example of the *M. tuberculosis* Beijing lineage which represents about 50% of all TB strains in East Asia and at least 13% of strains worldwide [**17**]. First reported in East Asian countries where they are highly endemic [**18, 19**] and later found to be disseminated throughout the world [**20, 21**], Beijing family strains are predominant in the former USSR countries, especially in Russia [**20–22**]. With an intrinsic advantage over other *M. tuberculosis* genotypes in terms of virulence (i.e., transmission, progression from latent to active tuberculosis, acquisition of drug resistance, or disease chronicity), Beijing lineage strains are notably associated with multiple drug resistant (MDR) and extensively drug resistant (XDR) TB [**17, 23, 24**]. It has been suggested that the study of Beijing genotype may reflect

hidden patterns of human migrations [**25**, **26**]; undoubtedly, the rapid global dissemination of this most successful clone of *M. tuberculosis* complex (MTBC) makes it a major cause of concern in public health. In this context, this paper will describe global mapping of *M. tuberculosis* Beijing lineage strains by analyzing its population structure and available demographic, epidemiologic, and drug-resistance data in the SITVIT2 database.

2. Materials and Methods

2.1 Brief description of the database

The detailed description of the SITVIT2 database, which is an updated version of the recently released SITVITWEB database [**13**], will be the subject of a separate publication (refer to **Table 1** for a summary of SITVITWEB vs. SITVIT2 databases). At the time of this study, SITVIT2 contained a total of 111,635 MTBC clinical isolates from 169 countries of patient origin. In this database, Spoligotype International Type (SIT) and MIRU International Type (MIT) designate identical patterns shared by 2 or more patient isolates, whereas “orphan” designates patterns reported for a single isolate that does not correspond to any of the patterns recorded in the repository of the SITVIT2 database. Furthermore, strains are classified in major phylogenetic clades assigned according to signatures provided earlier, which includes various MTBC members (AFRI, *M. africanum*; BOV, *M. bovis*; CANETTII, *M. canettii*; MICROTI, *M. microti*; PINI, *M. pinnipedii*), as well as for *M. tuberculosis* sensu stricto, i.e., the Beijing clade, the Central-Asian (CAS) clade, the East-African-Indian (EAI) clade, the Haarlem/Ural clades, the Latin-American-Mediterranean (LAM) clade, the Cameroon and Turkey lineages, the "Manu" family, the IS*6110*-low banding X clade, and the ill-defined T clade. Note that some spoligotypes previously classified as H3/H4 sublineages within Haarlem family were recently relabeled “Ural” [**27**]; these include patterns belonging to H4 sublineage that were relabeled "Ural-2", and some patterns previously classified as H3 sublineage but with an additional specific signature (presence of spacer 2, absence of spacers 29 to 31, and 33 to 36), that are now relabeled "Ural-1". With their definitive reclassification pending, we refer to these as H4/Ural-2 and H3/Ural-1 in SITVIT2 database. Furthermore, two LAM sublineages were recently raised to independent lineage level: LAM10-CAM as Cameroon lineage [**28**], and LAM7-TUR as Turkey lineage [**29, 30**]. In the context of the present paper, we will only focus on strains belonging to the Beijing clade (n=10,850 isolates from 92 countries of patient origin).

2.2 Genotyping

Genotyping methods used to define and analyze the Beijing strains in the international database were PCR-based spoligotyping [**6**], 12, 15 and 24-locus format of Mycobacterial Interspersed Repetitive Units – Variable Number of Tandem Repeats (MIRU-VNTRs) [**7**, **8**], and 5-locus exact tandem repeats (ETRs) [**31**]; the available results corresponded to 10,850 strains with spoligotyping, and 2,896 with MIRU-VNTRs (split in: 12-loci MIRU n=1,827; 15-loci MIRU n=1207; 24-loci MIRU n=420; and 5-loci ETR n=899). Throughout in this paper, the order of MIRU loci is as follows: 12-loci MIRU patterns: MIRU 2, 4, 10, 16, 20, 23, 24, 26, 27, 31, 39 and 40; 15-loci MIRU patterns : MIRU 4, 10, 16, 26, 31, and 40, ETR-A, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub30, and Mtub39; and 24-loci MIRU patterns: classical 12-loci pattern followed by ETR-A,

ETR-B, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub29, Mtub30, Mtub34, and Mtub39.

2.3 Statistical analyses

Hunter–Gaston Diversity Index (HGDI) was used through V-DICE tool (available at http://www.hpa-bioinformatics.org.uk/cgi-bin/DICI/DICI.pl) to compare typing methods (especially MIRU-VNTR data) and select the most discriminatory system [**32**]. The definition of HGDI derives from Simpson's diversity index which varies from 0 (no diversity) to 1 (highest diversity). In the present study, HGDI was used to determine the diversity of 5-locus ETRs and 12, 15, 24-locus MIRU-VNTR patterns. STATA software version 12 and MEGASTAT 12 (a statistical tool developed by The McGraw-Hill Companies which can be used as an add-in in Microsoft Excel) were used for descriptive and univariate analyses. Comparisons between different genotyping, demographic and epidemiological parameters were performed using Pearson's Chi-square test and Fisher's Exact Test. P values of <0.05 were considered as statistically significant.

2.4 Phylogenetic analyses and Bioinformatical tools

Due to a rather low discriminatory power of spoligotyping for Beijing lineage strains, Minimum Spanning Trees (MSTs) were preferably constructed based on MIRU-VNTR patterns using BioNumerics software version 6.6 (Applied Maths, Sint-Martens-Latem, Belgium). MSTs are connected undirected graphs in which all of the patterns are linked together with the fewest linkages between nearest neighbors.

SpolTools software (http://www.emi.unsw.edu.au/spolTools) was used to draw Spoligoforests tree based on the Fruchterman-Reingold algorithm [**33**]. Contrary to the MSTs, Spoligoforests are directed and not necessarily connected graphs. They permitted to highlight the evolutionary relationships between ascendant and descendant spoligotyping patterns.

The WebLogo application version 2.8.2 (available at http://weblogo.berkeley.edu/ [**34**, **35**]) was used to evaluate and visualize the allelic diversity of the 12-locus MIRU patterns in function of their region of isolation. WebLogos consist of stacks of symbols as graphical representations of the comparative diversity observed for individual MIRU loci – one stack for each MIRU loci. The overall height of the stack indicates the conservation of a given loci with a fixed number of copies at that position (i.e., if 100% of the strains conserve the same number of copies for a given MIRU loci, it corresponds to 4 bits), while the height of individual symbols within the stack indicates the relative frequency of number of copies of a given MIRU loci at that position. For our sample, we used WebLogo stack labels A – H for following copy numbers: A/1 copy, B/2 copies and so on (until H/8 copies).

2.5 Geographical distribution

The worldwide distribution of *M. tuberculosis* Beijing strains was studied both country wise (3 letter country codes according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3), and at macro-geographical level in subregions defined according to United Nations (available at http://unstats.un.org/unsd/methods/m49/m49regin.htm) as follows; Regions: AFRI (Africa), AMER

(Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). In this classification scheme, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that Russia was attributed a new sub-region by itself (Northern Asia) instead of including it among the rest of Eastern Europe.

3. Results and discussion

3.1 Database

The SITVIT2 database (n=111,635 MTBC clinical isolates from 169 countries of patient origin; **Table 1**) was used to extract a total of 48 spoligotyping patterns belonging to Beijing lineage (n=10,850 *M. tuberculosis* Beijing strains from 92 countries of origin of patients) in this investigation; a detailed description of the strains retained and their worldwide distribution is summarized in **Table 2**. As compared to other (non-Beijing) *M. tuberculosis* lineages (**Figure 1A**), the proportion of Beijing strains was higher in ASIA-E (n=3519/5216, or 67.47%), followed by ASIA-C (n=170/320, 53.13%), ASIA-N (n=744/1924, 38.67%), ASIA-SE (n=1048/3059, 34.26%), AFRI-S (n=823/4824, 17.06%), AMER-N (n=2179/14557, 14.97%), and ASIA-S (n=532/4883, 10.89%). The remaining subregions had a proportion of Beijing strains less than 10%. The male/female sex ratio distribution by subregions (**Figure 1B**) showed a global male/female sex ratio of 1.96 (1870 males/953 females) but with wide and statistically significant variations; it was highest at 4.93 in Russia vs. range of 0.8-1.13 observed in Central America, Caribbean, Eastern Africa and Northern Europe ($p<0.0001$). Statistically significant differences between sex ratios and age groups of the different subregions are highlighted in **Supplemental Figure S1**.

3.2 Drug-resistance

Drug resistance information was available for n=2846 strains (i.e. 26.23% of Beijing isolates) with the following distribution: pansusceptible, n=1713 (60.19%); strains resistant to at least one anti tuberculous drug, n=320 (11.24%); MDR n=736 (25.86%), and XDR n=77 (2.71%). The global distribution of drug resistance between Beijing strains vs. non-Beijing strains for different subregions (countries also shown for China and Japan) is illustrated in **Figure 2A** whereas the evolution of drug resistance over time for Beijing strains for a 14-year period (1998 to 2011) is summarized in **Figure 2B**. A continuous progression in the proportion of MDR and XDR strains (and a relative decrease of pansusceptible strains) is visible worldwide since 2003 (**Figure 2B**). Pansusceptible TB mainly concerned older patients >60 yrs (44%) whereas Drug resistant, MDR and XDR-TB concerned patients preferentially aged 21-40 yrs (between 52 to 58%). Although proportion of drug-resistant strains was significantly higher for Beijing vs. non Beijing strains globally, important variations in the distribution of drug-resistance were observed (**Figure 2A**): drug resistance was significantly linked to Beijing vs. non-Beijing strains in Russia, Southern Asia, Southeastern Asia, and European countries but not in Americas, Western Asia, China and Japan (despite the fact that the proportion of drug-resistant MTBC strains in China was significantly higher than in Japan). Hence,

although often perceived as being almost always associated with MDR-TB (essentially due to alarming drug-resistance reported in Russia and former USSR countries; [**23, 36**]); the data contained in the SITVIT2 database suggest that Beijing lineage is not always significantly more drug resistant as compared to other prevailing lineages within the same regions. Nevertheless, making a distinction between Beijing sublineages suggested a link between drug-resistance and certain genotypes. As already observed in recent studies from Colombia [**37, 38, 39**], an emerging SIT190 was well spread in Americas, particularly associated with MDR-TB and female patients. We observed that SIT190 was significantly more associated with MDR-TB than SIT1 (p-value<0.0001). Lastly, not enough genotyping and drug-resistance information was available for many of the African countries (note that available drug-resistance information from South Africa was intentionally omitted in the analysis shown in **Figure 2A** in order to avoid a selection bias since all the strains were recruited under specific investigations dealing almost exclusively with MDR-TB).

3.3 Distribution and demographic characteristics

The data on global distribution of Beijing strains by country of isolation is summarized in **Figure 3**, both in terms of percentage of total number of MTBC isolates recorded (**Figure 3A**) as well as absolute numbers (**Figure 3B**); followed by percent distribution among patients stratified in four age groups of 0–20, 21–40, 41–60, and >60 yrs (**Figure 3C**). The major patient age-group was 21–40 yrs worldwide except Japan with majority of patients >60 yrs. Younger patients (age group 0–20 yrs) were more common in South America, South Asia, and Eastern and Western Africa (25–33% of TB cases due to Beijing genotype). Regarding the global evolution of the proportion of Beijing vs. other MTBC strains in the database, it gradually increased over time, e.g., it was 6.67% in 1995 but more than doubled during 1998-2000 (15.8% of all strains), reaching a peak of 19.3% in 2008-2009. However, after massive inclusion of strains in 2010-2012 from many countries where Beijing was not highly predominant, the global proportion of Beijing vs. other lineage strains in the database settled at around 10% of all strains.

3.4 Beijing genotype and HIV coinfection

The number of patients with known HIV serology was relatively small (n=570/10,850 or 5.25%); among these HIV positive patients represented 16.67% of cases (n=95/570) vs. 83.33% of HIV negative cases (n=475/570). The mean age was 34.58 yrs for HIV positive vs. 31.67 yrs for HIV negative cases (statistically non-significant). Furthermore, no statistically significant differences were noted when comparing patients by age groups between HIV negative (n=402 with age information) and HIV positive patients (n=95 with age information); the figures being 59.95% vs. 77.89% in age group 20-39 yrs; 21.39% vs. 20% in age group 40-59 yrs, and 3.98% vs. 2.11% in age group ≥60 yrs (p-value>0.2). Note that in 402/570 patients with known age and HIV serology, none was < 20 yrs of age among HIV positive group vs. 59/402 among the HIV negative group. These data do not show a higher risk of developing TB specifically due to Beijing genotype among HIV coinfected patients. Nevertheless, the sex-ratio among the HIV-positive patients was highly associated with male gender (male to female sex-ratio of 3.32 vs. 1.98 for HIV negative patients; p-value=0.047). However, this observation must be taken with caution since the data available is not representative. Last but not

least, no statistically significant associations were found in our database when looking for an association between HIV sero-positivity and drug resistance information available, irrespective of pansusceptible, MDR, resistant to at least one drug, and XDR status of the strains (p-values >0.4).

3.5 Genotyping, discriminatory power and phylogenetical analysis

Regarding genotyping methods, spoligotyping allowed the distinction of 48 patterns for the Beijing lineage strains (**Table 2**): 15 orphans and 33 SITs, resulting in a low discriminatory power (HGDI, 0.128). Among the 48 spoligotype patterns, SIT1 (n=10130) represented 93.36% of all strains, underlining the necessity of a 2nd-line typing method such as MIRU-VNTRs to further discriminate the bulk of the Beijing strains. In our dataset, the global HGDI values significantly increased from 0.128 for spoligotyping alone to 0.781 for 5-loci ETRs; to 0.944 for 12-loci MIRUs; to 0.984 for 15-loci MIRUs, and to 0.990 for 24-loci MIRUs, corroborating the usefulness of MIRU-VNTRs to subtype Beijing strains.

The global diversity and geographic specificity of the Beijing strains could be specifically visualized thanks to the WebLogo application applied to the 12-loci MIRU-VNTRs data available (n=1827 strains; **Figure 4A**). As shown in this figure, the allelic diversity seen in conjunction with the discriminatory power (HGDI) clearly highlights certain geo-specificities as well as relatedness of 12-loci profiles in function of the prevailing Beijing strains. For example, one can note the similarity of WebLogos between Russia, Kyrgyzstan, South-Asia, and Eastern-Europe; which corresponded to a high proportion of MIT16 (pattern 223325153533) although with a low HGDI (in the range of 0.428 in Kyrgyzstan to 0.674 in ASIA-S; **Figure 4A**). On the other hand, HGDI was > 0.8 in all other regions, with a highest of 0.997 in China. With the exception of China, South America, Caribbean, and France, which were all characterized by specific WebLogos of their own, all other regions (particularly North-America, Northern-Europe, Japan, South Africa, and Southern-Europe), were characterized by the predominance of another 12-loci MIRU pattern 223325173533 (MIT17). The higher discriminatory powers of 12-loci MIRUs in conjunction with the high numbers of orphan strains in China (HGDI, 0.997) and Japan (HGDI, 0.945) corroborate the fact that Beijing strains were historically present in East-Asia and diverged over time, as opposed to their recent dissemination in Russia, likely within the course of the 20th century [**25**].

The above observation was further supported by the 12-loci MSTs drawn for selected countries (USA n=324, China n=280, Japan n=464, Russia n=331, and South Africa n=200) underlining evolutionary relationships of the Beijing strains in these settings (**Figure 4B**). We intentionally restricted the analysis to 5 most predominant MIRU patterns to easily visualize country-dependent specificities. In **Figure 4B** the first number inside the nodes indicates the MIT designation for clustered isolates followed by its percentage in brackets for a given country; while the peripheral ring around each MIT designation provides with its distribution in various countries in the database. The MSTs along with the distribution patterns of the MITs (**Figure 4B**) show that:

(i) Strains belonging to MIT17 were mostly isolated in Japan, followed by USA, Russia, South Africa, Sweden, Belgium, and at lower numbers in other countries such as Latvia, France, Spain, India, the Netherlands, Sudan, United Kingdom, Poland and Kyrgyzstan.

(ii) On the other hand, strains with MIT16 pattern were more common in Russia, Kyrgyzstan, USA, India, Belgium, Bangladesh, Latvia, Bulgaria, and isolated at lower numbers in Japan, France, Netherlands, Sweden, and Sudan.
(iii) Overall, MIT17 pattern was more spread (almost omnipresent) than MIT16.
(iv) Although MIT16 was highly predominant in Russia (representing 62.2% of MIRUs strains), MIT17 at almost 20% was as visible as it is in USA (25.3%), Japan (17.9%), and South Africa (15.5%).
(v) China was specific in not sharing any of these 2 major MITs with the rest of the world; it essentially contained MITs of its own (MIT1420, MIT1429, MIT1433, MIT1442, and MIT1468) that were specific to China.
(vi) The remaining MITs (19 in total) were distributed in varying proportions: MIT569 was the second most predominant MIT in USA (with 9.3% of strains) and was not yet found elsewhere; followed by MIT147 and MIT276 (single strains being reported from Japan and French Guyana respectively); MIT83 was present in USA and Russia as well as in Japan, South Africa, Kyrgyzstan, Belgium and France.
(vii) Interestingly, some probable paths of transmission could be highlighted, e.g., the majority of MITs from South Africa (except MIT17) were shared with USA, notably MIT254 which was found in South Africa and USA only. South Africa also shared at lower numbers some MIT profiles with United Kingdom, the Netherlands, China, Spain, Trinidad and Tobago, and Russia. Furthermore, MIT135 which is the second predominant MIT in Japan is also spread in USA, Russia, and Albania; MIT93 (the 3rd predominant MIT in Japan) is also found in Russia, Sweden and the Netherlands; MIT703 is common in Japan with 1 isolate in USA; MIT101 is common in Japan, South Africa and the USA.
(viii) We also looked at Single locus variants (SLV) vs. multi locus variants (MLV); SLVs were most commonly found in Russia and Japan showing highly linked variants of Beijing strains vs. exclusively MLVs in China showing extremely heterogeneous nature of the circulating strains. On the other hand, the USA was characterized essentially by MLVs with the exception of a single SLV on MIRU-26 allele, while South Africa was characterized by SLVs with the exception of a single MLV on MIRU-26 and MIRU-40 alleles.

These results show that evolutionary relationships vary more or less depending on the country of isolation of the Beijing strains, reflecting homogeneous (SLVs) vs. heterogeneous (MLVs) populations. Hence, despite the need of standard 24-loci MIRUs in conjunction with additional, hypervariable MIRU-VNTR loci for accurately discriminating closely related Beijing clones for TB molecular surveillance and follow-up [**40**], the 12-loci MIRU typing might still be useful for the scientific community to have a general idea of circulating strains worldwide – seeing the availability of huge amount of 12-loci data generated since the description of the methodology in 2001 [**7**]. Nevertheless, we also evaluated available 15- and 24-loci MIRU data in the database. Focusing on each MIRU-VNTR allele to calculate HGDI (**Supplemental Figures S2 to S4**), we found that some alleles (e.g. MIRU26, QUB4156, QUB11b, QUB26, Mtub21, Mtub04, MIRU31) were more discriminative than others. Indeed, as already underlined recently [**36**], some MIRU alleles seem to rather evolve by loss of number of copies (e.g., MIRU10, MIRU16, MIRU26, MIRU27, MIRU31, MIRU39, MIRU40, QUB11b, QUB26, QUB4156, ETR-A, Mtub04, Mtub29, and Mtub30; histograms marked by an

asterisk in **Supplemental Figure S2 to 4**) while others apparently evolve by gain of number of copies (e.g., MIRU-4, Mtub39).

Despite little diversity seen by spoligotyping for Beijing lineage strains, we also attempted to understand the evolutionary relationships between ascendant and descendant spoligotypes for the 48 patterns observed among 10,850 clinical isolates (33 SITs and 15 orphans). For this purpose, a Fruchterman-Reingold tree was drawn using the SpolTools software (**Supplemental Figure S5**). As expected, a huge central node corresponding to SIT1 (n=10,130 isolates) can be seen, followed by many smaller nodes. In this figure, Changes (loss of spacers) are represented by directed edges between nodes, and solid black lines link patterns that are very similar (maximum weigh being 1.0), i.e., loss of one spacer only. These concerned only 9/48 patterns shown clockwise as follows (number of strains for each pattern is shown in brackets): SIT796 (n=10), SIT255 (n=38); SIT1674 (n=7); SIT941 (n=24); SIT2610 (n=5); SIT632 (n=28); SIT190 (n=183); SIT265 (n=81); SIT621 (n=33). All other SITs were linked to the parental node (essentially SIT1) by dashed lines (weight comprised between 0.5 and 1), while none of the SITs were visibly linked by dotted lines (weight less than 0.5). Contrary to the MSTs, this tree is directed and does not necessarily connect all patterns, e.g., orphan 1 and orphan 6 in the figure. These observations globally show the close relatedness of the Beijing family strains worldwide.

However, the increasing diversity of Beijing strains is evident by drawing MSTs based on 12-, 15- and 24-loci MIRUs illustrated in **Supplemental Figure S6** in function of their country of isolation (represented by a different colors for each country). Each tree connects every pattern based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dashed and dotted lines) and circles representing each individual pattern. The length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed and gray dotted) denotes the number of allele changes between two patterns: solid lines, 1 to 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); gray dashed lines represent 4 changes; and gray dotted lines represent 5 or more changes. The size of the circle is proportional to the total number of isolates, highlighting unique isolates (smaller nodes) versus clustered isolates (bigger nodes). Due to lack of space, we have only marked the most predominant patterns by their MIT designations in the SITVIT2 database. These comprised 14 patterns for 12-loci MIRUs (12-loci MITs 17, 16, 135, 104, 569, 147, 93, 83, 254, 86, 99, 101, 703, and 1468: the reader is referred to **Figure 4B** for exact patterns); and 6 patterns for 15-loci MIRUs as follows: 15-MIT**415** (pattern 233753246834443), 15-MIT**11** (pattern 233753446834443), 15-MIT**4** (pattern 234753447253343), 15-MIT**142** (pattern 233553446824543), 15-MIT**6** (pattern 213753447854343), and 15-MIT**19** (pattern 233753446854343). Finally, 7 predominant 24-loci MITs were as follows: 24-MIT**25** (pattern 223325153533424682454433), 24-MIT**45** (pattern 223325171431424992254433), 24-MIT**27** (pattern 223325173533424672454433), 24-MIT**28** (pattern 223325153533424662454433), 24-MIT**4** (pattern 223325163531425693344433), 24-MIT**12** (pattern 223325173543424282452433), and 24-MIT**9** (pattern 223325173533424282452433).

The trees in **Supplemental Figure S6** highlight the geo-specificity of certain patterns and the similarity of others regarding the country of isolation. Irrespective of the MIRU typing formats used, one may notice the extremely diverse nature of the MIRU patterns from China. It was specific in not

sharing 12-MIT**16** and 12-MIT**17** with other countries (**Supplemental Figure S6-A**), and showed MITs exclusively found in China such as 12-MIT**1468** (in addition to other 12-MITs such as 1420, 1429, 1433, and 1442 specific to China; **Figure 4B**). It also corroborated some probable paths of transmission, e.g. 15-MIT**11**, 15-MIT**19**, and 15-MIT**6** that were shared between Peruvian and Japanese patients; and 15-MIT**142** shared between patients from India, Italy, Latvia, France, Bulgaria and Palestine (**Supplemental Figure S6-B**). Lastly, 24-MIT**25** was shared among patients from Russia, Latvia, Bulgaria and India, while 24-MIT**27** was shared among patients from Russia and Latvia (**Supplemental Figure S6-C**).

We also attempted to see how the strains in the database fit the definitions provided for tentatively discriminating the Beijing strains as modern, early ancient and late ancient based on the polymorphism of QUB4159 as reported recently [**41, 42**]. The results obtained (summarized in **Supplemental Table S1**) show that ancient Beijing sublineage has remained endemic to Japan (and to a lesser extent in China), while modern Beijing strains have spread worldwide.

4. Conclusions

Web tools that allow extracting knowledge contained within huge international databases are becoming increasingly important to visualize epidemiological and demographic characteristics of TB in order to control and monitor global TB epidemic [**4, 13-16, 33, 35, 43**]. Algorithms of Artificial Intelligence and Bayesian networks have already proved their efficiency in TB research [**44, 45**]. In this study, we attempted to give an insight into the recent developments of the multimarker TB genotyping database SITVIT2 developed and maintained at Institut Pasteur de la Guadeloupe, by taking the specific example of the *Mycobacterium tuberculosis* Beijing lineage strains which have today spread worldwide and are considered an example of one of the most successful emerging human pathogen [**17**]. Our results on worldwide mapping of the Beijing strains in conjunction with available demographic, epidemiologic, and genotypic data corroborate previous findings [**21, 25**] and further suggest that the Beijing epidemic may be tentatively split in 5 subgroups:

(i) Areas where it was present historically such as Eastern and South-Eastern Asia – the disease being split as endemic, prevalent among older patients with no evidence of drug resistance (Japan), and as being endemic, prevalent among younger patients, and often associated with drug resistance (China in particular);

(ii) Areas where Beijing genotype arrived between 17th – 19th centuries tentatively from coastal areas in Eastern-Asia [**25**], and is associated with younger age of the patients and drug resistance (South Africa). The same is probably true for Cuba [**21**] although both the proportion of Beijing vs. other strains is significantly lower than in South Africa, and the strains are less often linked to MDR-TB (results not shown).

(iii) Areas where the Beijing genotype arrived rather recently (19th-20th century) – the disease being split as predominantly caused by Beijing lineage often associated with MDR/XDR-TB in middle aged patients (Russia), vs. disease less frequently caused by Beijing strains (although associated with younger age) that are statistically not more frequently drug-resistant than other lineages (South America and Caribbean);

(iv) Areas where the Beijing strains arrived recently either due to political and trade relationships within the former communist block or during the political transition period in post-communist countries, mainly Eastern-Europe and former Soviet Union countries, prevalent among younger patients, and often associated with drug resistance;
(v) Areas where the modern Beijing strains have spread in conjunction with recent travel, trade and migration in 21st century, split in areas where the strains are sporadically linked to drug-resistance (North America), vs. areas where drug resistance (including MDR/XDR-TB) is frequent and statistically significant for Beijing strains as compared to other TB lineages (Western-, Southern- and Northern-Europe, and Southern-Asia), mostly involving younger age groups.

It may be concluded that the hotspots that represent a real threat about emergence/reemergence of Beijing strains are not necessarily countries with its highest prevalence (e.g., Japan), but rather the ones where this genotype is associated with younger age of patients, high demography and population density, and MDR/XDR-TB. Such examples include China, Indian sub-continent, and South-East Asian countries. Although, drug-resistance surveillance data in conjunction with genotyping results (needed to ascertain that the strain causing TB is Beijing or not) are not exhaustively available for all of African countries, the available data suggests that the proportion of Beijing strains is constantly increasing in West and South Africa, particularly among young patients (results not shown), which should be considered by utmost care by international health agencies.

Another major threat comes from the increased travel and migratory movements of the population worldwide in the last decades. For example, a recent report showed emergence of Beijing genotype in Saudi Arabia (4.5% of 1505 strains isolated during 2002-2005) associated with young age and drug resistance [**46**]. According to the World Tourism Organization, traveling for leisure (tourism) literally exploded since 1950 passing from 25 million to more than 1 billion in 2012 (http://www2.unwto.org/). Furthermore, the proportion of persons moving to work and escape poverty from 1995 to 2004 led to an important surplus of regional migratory balance in certain regions, noticeably in North America, Western Europe, Middle-East, while the highest deficit was observed in South-Asia, former soviet union countries, and Mexico, followed by Northern Africa and South-East Asia (metadata and map available through the European Environment Agency: http://www.eea.europa.eu/data-and-maps/figures/moving-to-work-and-escape-poverty).
Superimposing our data to these statistics leads us to target specifically Western Europe, North America and Middle East as regions particularly sensitive to a migratory threat of imported cases of disease.

Acknowledgements

The data dealt within this publication on 10,850 *M. tuberculosis* Beijing lineage strains was extracted from the SITVIT2 database, and corresponds to a total of 186 contributors (list available upon request). We are grateful to all the contributors who submitted their genotyping data to the successive versions of the databases maintained at Institut Pasteur de la Guadeloupe, namely SpolDB1 to SpolDB4 between 1999-2006, and more recently to SITVITWEB (published in 2012) and SITVIT2 (to be released in 2014). The database development was financed by the International Network of the Pasteur Institutes and the Regional Council of Guadeloupe (CR/08-1612: Biodiversité

et Risque Infectieux dans les modèles insulaires). David Couvin was awarded a Ph.D. fellowship by the European Social Funds through the Regional Council of Guadeloupe. We are thankful to Thierry Zozio and Julie Millet for helpful discussions.

Competing interests: None declared

References

1. World Health Organization (WHO) Global tuberculosis report 2013. www.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf
2. García de Viedma D, Mokrousov I, Rastogi N. Innovations in the molecular epidemiology of tuberculosis. Enferm Infecc Microbiol Clin. 2011; 29 Suppl 1:8-13. [Review].
3. Rastogi N, Sola C: Chapter 2 - Molecular evolution of the *Mycobacterium tuberculosis* complex. In Tuberculosis 2007: from basic science to patient care. Edited by Palomino JC, Leao S, Ritacco V. 2007:53–91. Amedeo Online Textbooks. http://eglobalmed.com/opt/AmedeoLiteratureGuide/www.tuberculosistextbook.com/tuberculosis2007.pdf
4. Shabbeer A, Ozcaglar C, Yener B, Bennett KP. Web tools for molecular epidemiology of tuberculosis. Infect Genet Evol. 2012; 12:767-781.
5. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small, PM. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 1993; 31:406-409.
6. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 1997; 35:907-914.
7. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J Clin Microbiol 2001; 39:3563-3571.
8. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 2006; 44:4498-4510.
9. Cardoso Oelemann M, Gomes HM, Willery E, Possuelo L, Batista Lima KV, Allix-Béguec C, Locht C, Goguet de la Salmonière YO, Gutierrez MC, Suffys P, Supply P. The forest behind the tree: phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country. PLoS One. 2011 Mar 25;6(3):e18256.
10. Vanhomwegen J, Kwara A, Martin M, Gillani FS, Fontanet A, Mutungi P, Crellin J, Obaro S, Gosciminski M, Carter EJ, Rastogi N. Impact of immigration on the molecular epidemiology of tuberculosis in Rhode Island. J Clin Microbiol 2011;49:834-844.

11. Marais BJ, Mlambo CK, Rastogi N, Zozio T, Duse AG, Victor TC, Marais E, Warren RM. Epidemic spread of multidrug-resistant tuberculosis in Johannesburg, South Africa. J Clin Microbiol. 2013 ;51:1818-1825.
12. Chaoui I, Zozio T, Lahlou O, Sabouni R, Abid M, El Aouad R, Akrim M, Amzazi S, Rastogi N, El Mzibri M. Contribution of spoligotyping and MIRU-VNTRs to characterize prevalent *Mycobacterium tuberculosis* genotypes infecting tuberculosis patients in Morocco. Infect Genet Evol. 2013 May 31. doi:pii: S1567-1348(13)00212-8. [In Press, Corrected Proof]
13. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N. SITVITWEB – a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol. 2012 ; 12:755-766.
14. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 2006;6:23.
15. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol. 2008; 46:2692-2699.
16. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. Infect Genet Evol 2012; 12:789-797.
17. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect Dis. 2010 ;10:103-111.
18. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. J Clin Microbiol 1995;33:3234-3238.
19. Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, Narvskaya O. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res 2005; 15:1357-1364.
20. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol 2002 Jan; 10:45-52.

21. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis. Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance. Emerg Infect Dis. 2006;12; 5:736-43.
22. Mokrousov I, Narvskaya O, Otten T, Vyazovaya A, Limeschenko E, Steklova L, Vyshnevskyi B. Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. Res Microbiol 2002; 153:629-37.
23. Toungoussova OS, Sandven P, Mariandyshev AO, Nizovtseva NI, Bjune G, Caugant DA. Spread of drug-resistant *Mycobacterium tuberculosis* strains of the Beijing genotype in the Archangel Oblast, Russia. J Clin Microbiol 2002; 40:1930-1937.
24. Drobniewski F, Balabanova Y, Nikolayevsky V, Ruddy M, Kuznetzov S, Zakharova S, Melentyev A, Fedorin I. Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. JAMA 2005; 293:2726-2731.
25. Mokrousov I. Genetic geography of *Mycobacterium tuberculosis* Beijing genotype: a multifacet mirror of human history? Infect Genet Evol 2008; 8:777-785.
26. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, van Soolingen D, Rüsch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog 2008 19;4(9):e1000160.
27. Mokrousov I. The quiet and controversial: Ural family of *Mycobacterium tuberculosis*. Infect Genet Evol. 2012; 12:619-629.
28. Koro Koro F, Kamdem Simo Y, Piam FF, Noeske J, Gutierrez C, Kuaban C, Eyangoh SI. Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping. J Clin Microbiol. 2013; 51:299-302.
29. Abadia E, Zhang J, dos Vultos T, Ritacco V, Kremer K, Aktas E, Matsumoto T, Refregier G, van Soolingen D, Gicquel B, Sola C. Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infect Genet Evol 2010; 107: 1066–1074.
30. Kisa O, Tarhan G, Gunal S, Albay A, Durmaz R, Saribas Z, Zozio T, Alp A, Ceyhan I, Tombak A, Rastogi N. Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey. PLoS One. 2012; 7(1):e30331.
31. Frothingham R, Meeker-O'Connell WA. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 1998; 144:1189-96.
32. Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 1988;26:2465-2466.
33. Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. Bioinformatics 2008; 24:2414-415.
34. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990; 18:6097-6100.
35. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004; 14:1188-1190.

36. Mokrousov I. Insights into the origin, emergence, and current spread of a successful Russian clone of *Mycobacterium tuberculosis*. Clin Microbiol Rev 2013; 26:342-360.
37. Murcia MI, Manotas M, Jiménez YJ, Hernández J, Cortès MI, López LE, Zozio T, Rastogi N. First case of multidrug-resistant tuberculosis caused by a rare "Beijing-like" genotype of *Mycobacterium tuberculosis* in Bogotá, Colombia. Infect Genet Evol. 2010; 10:678-81.
38. Nieto LM, Ferro BE, Villegas SL, Mehaffy C, Forero L, Moreira C, Rastogi N, van Soolingen D. Characterization of extensively drug-resistant tuberculosis cases from Valle del Cauca, Colombia. J Clin Microbiol. 2012; 50:4185-4187.
39. Villegas SL, Ferro BE, Perez-Velez CM, Moreira CA, Forero L, Martínez E, Rastogi N, Caminero JA. High initial multidrug-resistant tuberculosis rate in Buenaventura, Colombia: a public-private initiative. Eur Respir J. 2012; 40:1569-1572.
40. Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, Campos-Herrero I, Mokrousov I, Niemann S, Kontsevaya I, Rastogi N, Samper S, Sng LH, Warren RM, Supply P. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. J Clin Microbiol. 2013 Oct 30. [Epub ahead of print].
41. Wada T, Iwamoto T. Allelic diversity of variable number of tandem repeats provides phylogenetic clues regarding the *Mycobacterium tuberculosis* Beijing family. Infect Genet Evol 2009; 9:921-926.
42. Iwamoto T, Grandjean L, Arikawa K, Nakanishi N, Caviedes L, Coronel J, Sheen P, Wada T, Taype CA, Shaw MA, Moore DA, Gilman RH. Genetic diversity and transmission characteristics of Beijing family strains of *Mycobacterium tuberculosis* in Peru. PLoS One 2012;7(11):e49651.
43. Soares P, Alves RJ, Abecasis AB, Penha-Gonçalves C, Gomes MG, Pereira-Leal JB. inTB - a data integration platform for molecular and clinical epidemiological analysis of tuberculosis. BMC Bioinformatics. 2013 Aug 30;14:264.
44. Aandahl RZ, Reyes JF, Sisson SA, Tanaka MM. A model-based Bayesian estimation of the rate of evolution of VNTR loci in *Mycobacterium tuberculosis*. PLoS Comput Biol. 2012;8(6):e1002573.
45. Hill V, Zozio T, Sadikalay S, Viegas S, Streit E, Kallenius G, Rastogi N. MLVA based classification of *Mycobacterium tuberculosis* complex lineages for a robust phylogeographic snapshot of its worldwide molecular diversity. PLoS One. 2012;7(9):e41991.
46. Al Hajoj S, Rastogi N. The emergence of Beijing genotype of *Mycobacterium tuberculosis* in the Kingdom of Saudi Arabia. Ann Thorac Med. 2010; 53:149-152.

Legend to Figures

Figure 1: Phylogeographical distribution of Beijing vs. non-Beijing *M. tuberculosis* lineages, and demographical characteristics of the patients.

(A). Phylogeographical distribution of Beijing vs. non-Beijing strains in different subregions or countries. The size of the pie charts indicates the total number of strains; map also shows probable human migrations causing dispersal of Beijing lineage strains and its probable origin [**19**]; **(B).** Distribution of the male/female sex ratios and mean ages by UN subregions.

Figure 2. Drug resistance characteristics of Beijing vs. non-Beijing *M. tuberculosis* lineages (**A**) and evolution of drug resistance among Beijing isolates between 1998 to 2011 (**B**).

Figure 3. Global distribution of proportion of Beijing strains by country of isolation **(A)**, Global distribution of number of isolated strains by country **(B)**, and Distribution of age groups of patients (percentages) depending on subregions or countries **(C)**.

Figure 4. Global diversity of MIRU WebLogo patterns of the Beijing isolates (**A**) and evolutionary relationships of the predominant 12-loci patterns in countries having more than 100 isolates (**B**). Note that the inside color of the circle (in Figure B) indicates the number of copies for MIRU-26 locus (which was the most discriminatory locus in 12-loci typing scheme) as follows: blue, 5 copies; grey, 6 copies; pink, 7 copies.

Supplemental Figure S1.

Matrices of correlation indicating the statistical p-value of differences regarding the different subregions and between sex ratios **(A)** and age groups **(B)**

Supplemental Figures S2, S3 and S4.

Histograms representing the allelic diversities (number of copies) of each locus of 12-loci **(S2)**; 15-loci **(S3)**; and 24-loci MIRU-VNTRs **(S4)**.

Supplemental Figure S5.

Fruchterman-Reingold tree drawn using SpolTools (http://www.emi.unsw.edu.au/spolTools). each spoligotype pattern is represented by a node with area size being proportional to the number of isolates. Changes (loss of spacers) are represented by directed edges between nodes. The heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weigh being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5.

Supplemental Figure S6.

Minimum spanning trees (MSTs) illustrating evolutionary relationships between 12-loci **(A)**, 15-loci **(B)**, and 24-loci MIRU-VNTRs **(C)**, in function of their country of isolation (represented by a different color for each country). See text for more details.

Article 04 (en rédaction)

Database

SITVIT*Bovis* - a publicly available database and mapping tool to get a global overview of animal and human cases caused by *Mycobacterium bovis*

David Couvin[1], Iñaki Cervera-Marzal[1], Audrey David[1], Nalin Rastogi[1]*

[1]WHO Supranational TB Reference Laboratory, TB & Mycobacteria Unit,
Institut Pasteur de la Guadeloupe, Abymes Cedex, Guadeloupe (France)

ABSTRACT (302 words)

Mycobacterium bovis (*M. bovis*) is a mycobacterium that can cause tuberculosis (TB) disease in both animals and humans. This particular pathogen deserves to be intrinsically better understood in order to be better controlled.

SITVIT*Bovis* database incorporates genotyping and epidemiological data allowing to have a global overview of *M. bovis*/*M. caprae* worldwide genetic diversity based on 25,741 clinical isolates corresponding to 60 countries of patient origin and 75 countries of isolation. We reported a total of 1000 spoligotype patterns (corresponding to 25,741 clinical isolates) – grouped into 537 shared-types or spoligotype international types (SIT) containing 25,278 clinical isolates and 463 orphan patterns. The publicly accessible database also contained a total of 119 12-loci mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTRs) patterns from 344 clinical isolates – grouped into 41 shared-types or 12-MIRU international types (12-MIT); and a total of 60 15-loci MIRU patterns from 104 clinical isolates – grouped into 11 shared-types or 15-MIRU international types (15-MIT). Lastly, data on 5-locus exact tandem repeats (ETRs) were available on 415 isolates; a total of 109 different VNTR patterns were observed – split into 55 shared-types or VNTR International Types (VIT) containing 361 isolates and 54 orphans. Furthermore, significant differences were noted when comparing data on age and gender of human patients isolates.

Information on host of strains also provided notable mappings and phylogenies underlining several geographic specificities.

SITVIT*Bovis*, alongside with other databases dedicated to animal mycobacteria, further allow to have a global insight into *M. bovis*/*M. caprae* population structure and worldwide distribution. This publicly accessible database available at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis is a relevant tool which aims to improve our knowledge on specific bacteria threatening both humans and other mammals. The aim of this study was to improve our understanding of global BTB epidemiology, and facilitate genetic characterization of *M. bovis* population in several places of the world.

Key-words: Mycobacterium bovis, Database, Spoligotyping, MIRU-VNTRs, Evolution, Epidemiology, Host, bovine tuberculosis, human, animal.

---

* Corresponding author. Mailing address: Institut Pasteur de Guadeloupe, Morne Jolivière, BP484, F97183 Abymes Cedex, Guadeloupe. Phone: +590-590-897661. Fax: +590-590-896941. E-mail: nrastogi@pasteur-guadeloupe.fr.

Article 05

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 398484, 11 pages
http://dx.doi.org/10.1155/2014/398484

Research Article

Predicting *Mycobacterium tuberculosis* Complex Clades Using Knowledge-Based Bayesian Networks

Minoo Aminian,[1] David Couvin,[2] Amina Shabbeer,[1] Kane Hadley,[1] Scott Vandenberg,[3] Nalin Rastogi,[2] and Kristin P. Bennett[1,4]

[1] *Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[2] *WHO Supranational TB Reference Laboratory, Tuberculosis and Mycobacteria Unit, Institut Pasteur de la Guadeloupe, 97183 Abymes, Guadeloupe, France*
[3] *Department of Computer Science, Siena College, Loudonville, NY 12180, USA*
[4] *Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*

Correspondence should be addressed to Kristin P. Bennett; bennek@rpi.edu

Received 18 October 2013; Accepted 10 December 2013; Published 23 April 2014

Academic Editor: Tomasz Jagielski

Copyright © 2014 Minoo Aminian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We develop a novel approach for incorporating expert rules into Bayesian networks for classification of *Mycobacterium tuberculosis* complex (MTBC) clades. The proposed knowledge-based Bayesian network (KBBN) treats sets of expert rules as prior distributions on the classes. Unlike prior knowledge-based support vector machine approaches which require rules expressed as polyhedral sets, KBBN directly incorporates the rules without any modification. KBBN uses data to refine rule-based classifiers when the rule set is incomplete or ambiguous. We develop a predictive KBBN model for 69 MTBC clades found in the SITVIT international collection. We validate the approach using two testbeds that model knowledge of the MTBC obtained from two different experts and large DNA fingerprint databases to predict MTBC genetic clades and sublineages. These models represent strains of MTBC using high-throughput biomarkers called spacer oligonucleotide types (spoligotypes), since these are routinely gathered from MTBC isolates of tuberculosis (TB) patients. Results show that incorporating rules into problems can drastically increase classification accuracy if data alone are insufficient. The SITVIT KBBN is publicly available for use on the World Wide Web.

1. Introduction

Tuberculosis (TB) represents a reemerging serious health threat worldwide. TB is caused by the *Mycobacterium tuberculosis* complex (MTBC) bacterium. One-third of the world population is latently or actively infected with TB. Molecular epidemiology now plays a crucial role in the tracking and control of TB [1]. DNA fingerprinting methods have made it possible to distinguish between cases of recent transmission of TB and reactivation of latent infections. This has enabled the tracking of transmission routes and the timely identification of outbreaks. Thus, knowledge about the genotypes of prevailing strains has revolutionized traditional approaches to the epidemiology of TB. Moreover, the predominance of certain strains or groups of strains in certain host populations has been clearly observed [2]. Studies of the genetic and biogeographic diversity of the MTBC have revealed differences in the virulence, immunogenicity, and drug resistance of strains [2]. This has consequences for the development of control measures for TB.

The classification of MTBC strains into genetic groups or clades is important to help track transmission patterns and develop a better understanding of pathologic specificities in TB. Phylogeographic clades have been defined based on genetic similarities between strains and observed associations between groups of similar MTBC genotypes with host populations [3]. A variety of molecular techniques including the analysis of phylogenetically informative single nucleotide polymorphisms (SNPs) and long sequence polymorphisms (LSPs) are used to genotype MTBC strains [4]. Classification based on SNPs and LSPs is considered to be the gold standard. However, studies of such variations in DNA sequences

Figure 1: Example rules from SpolDB4. The rule column represents characteristic patterns specified by the visual rules as underlined subsequences in the spoligotype patterns. Each line corresponds to a rule. The underlined portions of the spoligotype must match exactly while the portions not underlined can take any value. All of these rules fire for the spoligotype 11011111111101111111000011111111100001111111, while three of the rules fire for 11111111111100111111000011111111100001111111.

of MTBC strains are not performed frequently for public health purposes. Spacer oligonucleotide typing (spoligotyping) and mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTRs) typing are two polymerase-chain-reaction- (PCR-) based DNA fingerprinting methods routinely used in the United States for genotyping all identified culture-positive TB cases. Large databases of spoligotypes have been collected. Each spoligotype for a strain is determined by the presence or absence of 43 specific spacers in the DR region, producing a 43-bit number. Each spacer separates two direct repeats. These strains have been assigned sublineage labels using mixed expert-based and bioinformatics approaches derived from visual rules applied to spoligotypes as shown in Figure 1.

These visual rules are based on the identification of characteristic deletions of one or more adjacent spacers in spoligotypes. Certain inferred mutations (deletions of blocks of adjacent spacers) in progenitor strains are considered to be lineage defining [5]. These deletions are conserved in all descendent strains since studies have shown that the mechanism of mutation observed in the direct repeat (DR) region involves loss of spacers, and spacers are rarely gained [6]. Additionally, the existence of these sublineages has been independently verified by clustering based on spoligotype and MIRU types of strains [7, 8]. Therefore, while it has been established that strains of TB belong to distinct sublineages, the definitions of these sublineages based on spoligotypes are not clear. The visual rules for a sublineage are generalizations of spoligotype patterns that belong to the sublineage. However, directly applying visual rules to spoligotype patterns can lead to multiple assignments of sublineage labels since spoligotype patterns may match patterns prescribed by more than one rule, and sometimes spoligotype patterns do not exactly match the patterns specified by any rule. This is an inherent limitation of a rule-based system, wherein rules need to be broad enough to capture general patterns but narrow enough to delineate classes. Additionally, spoligotyping is based on polymorphisms in a single locus, the DR region, and therefore has the potential for convergent evolution. Relying on specific subsequences within the spoligotypes for the study of genetic diversity is hence error prone.

This paper presents a hierarchical probabilistic graphical model, the knowledge-based Bayesian network (KBBN), that encodes rules of thumb and large training databases to classify data into given classes. Expert knowledge is modeled in the top level of variables in the BN representing the rules. The middle level variables represent the class and the lower level represents various features of interest. KBBN uses the strategy of not directly modeling the dependency of the rules on the features. This greatly reduces the model parameter space which helps reduce the amount of data required for training while capturing the knowledge in the rules.

The overall goal of this paper is to construct a model for predicting clades based on spoligotypes as determined by SITVIT using published rules from SPOLDB4 and other sources and to make this model available via the World Wide Web. For MTBC clade classification in this model, the visual rules of thumb are the top-level variables, the clades are the classes, and the 43 spacers that constitute the spoligotypes are the features. The KBBN for MTBC sublineages builds on the conformal Bayesian network previously designed for that domain. The structure of the KBBN encodes the knowledge base captured in the rules of thumb helping to improve overall accuracy while overcoming any potential problems such as ambiguous, inaccurate, or incomplete rules. A secondary goal is to assess the effectiveness of the KBBN in the MTBC lineage classification task. Thus we do extensive experiments on SITVIT as well as an additional testbed: CDC. The CDC test set consists of data and rules from the United States Centers for Disease Control and Prevention.

This paper is organized as follows. We first examine prior Bayesian networks for MTBC classification and then introduce the KBBN approach to incorporate rules. We then examine the rules of thumb and data associated with MTBC clades. We present results for the KBBN-SITVIT model and assess its accuracy. Finally, computational studies examine how KBBN can improve accuracy over Bayesian networks on the two KBBN testbeds (SITVIT and CDC). The results

show that KBBN is quite resilient to incomplete, inaccurate, or ambiguous rules and can obtain better performance than BN using less data.

Previously in the MTBC domain, other approaches to incorporate advice in the form of rules have been shown to improve discriminative learning models of MTBC major lineages and other problems [9]. However, those methods are limited to rules expressed in less-intuitive polyhedral form that requires preprocessing of data and rules.

The proposed KBBN model allows the existing rules of thumb to be incorporated with no modification resulting in improved classification over the predictions made with the rules or Bayesian networks alone. Also, unlike visual rules, the flexibility offered by the KBBN enables it to handle these common problems with the following rules of thumb.

(i) Incompleteness: rules only exist for some of the classes or only partially cover a class.

(ii) Ambiguity: multiple rules of thumb for different classes apply to the same exemplar. This frequently occurs if there is no precedence associated with the rules.

(iii) Inaccuracy: rules may incorrectly classify some exemplars.

Visual rules with precedence have been established for six major MTBC lineages [10]. A prior online knowledge-based support vector machine (SVM) approach combined these visual rules and precedence into a set of rules expressed in polyhedral form [9]. The method produced a high-accuracy SVM using much less data. However, as discussed in Section 5, this elegant work has several practical limitations that we sought to overcome in this study. First, expressing rules and precedence as polyhedral rules [9] can be challenging for a large number of rules. Second, the method works best with linear SVMs, but linear SVMs do not capture the underlying complexity of the biomarkers and their mechanism of evolution. This can be overcome by using nonlinear SVMs (SVMs using 3-degree polynomial kernels work very well), but then incorporating the polyhedral rules becomes even more challenging. Third, the complexity of training increases with the introduction of rules. Thus, the proposed design of the KBBN has the following salient features:

(i) incorporates rules easily without modification and without imposing precedence,

(ii) models known properties of the domain such as biomarkers and their mutation mechanisms,

(iii) provides an efficient training method for classes with and without rules,

(iv) achieves high prediction accuracy,

(v) overcomes ambiguity, incompleteness, and inaccuracy of the rules,

(vi) provides additional information about the effectiveness of each rule.

The overall approach produces a high quality model for predicting SITVIT clades which has been made available for use by other researchers.

2. Bayesian Network Background

A Bayesian network (BN) is a graphical representation of a probability distribution. Formally speaking, a BN is a directed acyclic graph $G(N, E)$ consisting of a set of nodes $X = \{x_i \mid x_i \in N\}$ to represent the variables and a set of directed links to connect pairs of nodes [11]. Each node has a conditional probability distribution that quantifies the probabilistic relation between the node and its parents such that for a network of k nodes

$$P(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} P(x_i \mid \text{parents}(x_i)). \quad (1)$$

Therefore, one can compute the full joint probability distribution from the information in the network. In other words, a well-represented Bayesian network can capture the complete nature of the relationship among a set of variables.

The SPOTCLUST Bayesian network was the first generative model used for analysis of MTBC sublineages [7]. SPOTCLUST uses mixture models based on spoligotypes to identify strain families of MTBC. SPOTCLUST models the asymmetric evolution of spacers using a Bayesian network with "hidden parents" [7]. The hidden parents of a lineage generate the members of the lineage. They capture the evolution of spoligotypes without generating the full phylogeny. A spacer in the hidden parent may be lost with small probability. A spacer that is absent in the parent is almost never gained. The design models the evolution mechanism of the DR region, allowing the Bayesian network to capture the deletions that are known to characterize spoligotype lineages. The hidden parent technique of SPOTCLUST is used for the spoligotype-associated parts of the KBBN model.

The conformal Bayesian network (CBN) is another generative model for analysis of both spoligotype and MIRU type data for MTBC strains [9, 12] (spoligotype CBN is shown in Figure 2(a)) originally designed for predicting major MTBC lineages. CBN captures the domain knowledge about the properties of spoligotypes and MIRU and uses this information to classify MTBC strain genotyping data into major lineages. CBN reflects the known mutation mechanisms of the spoligotypes and MIRU. With rare exceptions, ancestral strains have 2 or more repeats at MIRU24. Thus the top-level variable, M_{24}, indicates whether MIRU24 is less than two (indicating one of the modern lineages with high probability) or at least two (indicating one of the ancestral lineages with high probability).

One can think of the MIRU CBN model "generating" the data as follows. The value of locus MIRU24 generates the lineage, which in turn determines the number of repeats in the remaining MIRU loci. Thus, patterns in the occurrences of repeats at each locus for each lineage are captured. The lineage also generates the hidden parents of the lineage which in turn generate the spoligotype spacers. The MIRU24 determines the lineage priors.

We tried using the CBN model as designed for major lineages to classify MTBC genotyping data into sublineages. But using the single rule, *if MIRU*24 $\geq$ 2*, then lineage is ancestral*, as in the original CBN was not enough to generate a good

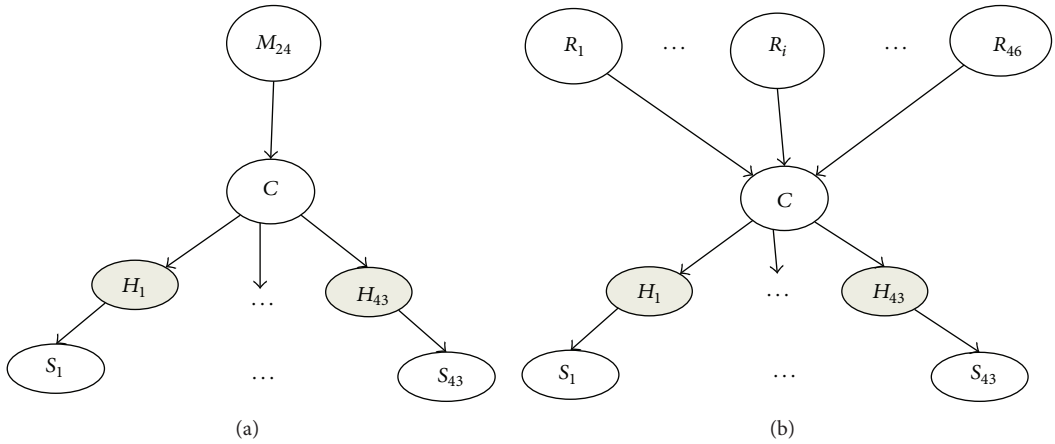


Figure 2: (a) The spoligotype conformal Bayesian network uses a single rule based on the number of repeats at the MIRU24 locus as the first level of a hierarchical Bayesian network. It uses the 43 spacers as features. CBN predicts the major lineage with high accuracy. (b) The KBBN uses multiple rules based on the presence of characteristic deletions as the first level of a hierarchical Bayesian network. As with the CBN, it uses the 43 spoligotype spacers.

model. KBBN grew out of the effort to incorporate all of the visual rules available from SpolDB4 [3] into CBN.

2.1. Knowledge-Based Bayesian Network. The knowledge-based Bayesian network (KBBN) is a hierarchical probabilistic graphical model which encodes the knowledge obtained from expert-defined rules derived from true observations with large databases to solve classification problems. KBBN incorporates the rules of thumb as high level variables or class priors in the Bayesian network and therefore combines the information obtained from rules of thumb with the information provided by a BN model specifically designed for the domain. The method is designed and tested on widely used simple BNs such as naïve Bayes and polytrees with polynomial time learning and inference algorithms. The KBBN for the tuberculosis domain uses CBN as its base and is used for both CDC and SITVIT testbeds.

KBBN, represented in Figure 2(b), is a novel hierarchical Bayesian network probability model for sublineage classification of MTBC. KBBN captures domain knowledge about the properties of spoligotypes and incorporates additional information provided by SpolDB4 or CDC rules to predict the class with high accuracy. The corresponding probability density function for the naïve KBBN model, shown in Figure 2(b), is

$$P\left(C, S_\Omega, R_\Psi\right) = \prod_{j\in\Omega}\left(\sum_{H_{j\in\{0,1\}}} P\left(S_j \mid H_j\right) P\left(H_j \mid C\right)\right) \times P\left(C \mid R_\Psi\right) P\left(R_\Psi\right), \quad (2)$$

where the random variable C represents the sublineage class, the random variable $S_\Omega = \{S_j \mid j \in \Omega\}$ with $\Omega = \{1, \ldots, 43\}$ represents the spoligotype spacers, and $R_\Psi = \{R_k \mid k \in \Psi\}$ represents the set of binary rules indicating whether each specific rule is fired. The spacer variables S and class variable C are assumed to follow binomial and multinomial distributions, respectively. The conditional probabilities of R given C are represented as a table which maps the set of possible combinations of rules fired in the data to the probability of each class. Laplacian smoothing is used.

For spoligotypes, we followed the SPOTCLUST model [7]. It captures the fact that spacers are lost but almost never gained, by introducing a variable for the unobserved hidden parent (H_j) and for each spacer S_j, both of which follow a binomial distribution. Given a 43-dimensional spoligotype S and its spacer position j, $S_j = 1$ if the spacer is present, and $S_j = 0$ if the spacer is absent. The probabilities of the spacer given the parent $P(S_j \mid H_j)$ are assumed to be known. As in Vitol et al., 2006 [7], we considered the probability of losing a spacer as 10^{-1} and the probability of gaining a spacer as 10^{-7}.

The KBBN assumes that the spoligotype hidden parents are conditionally independent given the sublineage. The conditional independence assumption of spacers is a model simplification previously made in the SPOTCLUST BN model. This conditional independence of the biomarkers in the BN model enables KBBN to conform to the set of available biomarkers without any expensive missing value computations. None of the genotyping variables in the BN are treated as unobserved except for the hidden parent spacers, which are always unobserved.

Using Bayes' rule, one can predict the sublineage for new data by determining the sublineage with maximum probability:

$$P\left(C \mid S_\Omega, R_\Psi\right) \propto \prod_{j\in\Omega}\left(\sum_{H_{j\in\{0,1\}}} P\left(S_j \mid H_j\right) P\left(H_j \mid C\right)\right) \times P\left(C \mid R_\Psi\right) P\left(R_\Psi\right). \quad (3)$$

3. Data Domains and Biology Rules

This study focused on creating a predictive model for clades that emulated SITVITWEB, a publicly available international

Table 1: SITVIT and CDC MTBC testbeds.

Testbed	Dataset	Size	Number of classes	Max class size	Min class size	Number of rules
SITVIT	Train	2714	69	390	1	69
	Test	7949	69	1107	1	69
	CV	2593	45	390	11	45
CDC	CV	1286	8	356	39	8

multimarker database for tuberculosis molecular epidemiology [13]. Different datasets were created for training, testing, and cross validation. To validate the approach we also used a dataset of isolates collected by the CDC for cross validation studies. The following two sections describe the datasets in detail. Table 1 summarizes them.

3.1. SITVIT Testbed. SITVIT-Train and SITVIT-Test are based on the SITVIT, a MTBC genotyping markers database provided by the Institut Pasteur de la Guadeloupe, and on the SpolDB4 rules that are published in Brudey et al., 2006 [3], plus one rule recently developed for the URAL1 clade. KBBN was trained on the SITVIT-Train dataset of 2714 records, each corresponding to a spoligotype and clade. There were 69 classes, the minimum sublineage size was 1, and the maximum sublineage size was 390. To test this model while keeping all classes we used SITVIT-Test, a large dataset based on SITVIT with 7949 records, each record corresponding to a spoligotype and clade. This dataset contained the same 69 classes as SITVIT-Train with different class distributions and again with the minimum class size of 1. SITVIT-Train and SITVIT-Test do not overlap so the total SITVIT dataset consists of 10633 distinct spoligotypes. To enable 10-fold cross validation (CV) with at least one spoligotype per class, the SITVIT-CV dataset was created which consists of the SITVIT-Train data restricted to the 45 classes with at least 11 spoligotypes each.

Note that some lineages have been reclassified while the KBBN model was under development. Two LAM sublineages were recently raised to lineage level: LAM10-CAM as the Cameroon lineage [14] and LAM7-TUR as the Turkey lineage [15, 16]. Some spoligotype patterns previously classified as H3 and H4 sublineages were relabeled "Ural" [17]. The latter include patterns belonging to H4 sublineage that were relabeled "Ural-2" and some patterns previously classified as H3 sublineage but with an additional specific signature (presence of spacer 2, absence of spacers 29 to 31 and 33 to 36), which are now relabeled "Ural-1." With their definitive reclassification pending, we hereby refer to these as H4-Ural-2, H3, and H3-Ural-1. Spoligotype patterns labeled as EAI and EAI5 were merged into a single group called EAI since one rule covers both patterns.

A sample of SpolDB4 rules is presented in Figure 1. Each line corresponds to a rule. The underlined portions of the spoligotype must match exactly while the portions not underlined can take any value. Note that in Brudey et al. [3] the rules are expressed using the octal coding of spoligotypes; here we express them in binary for simplicity. While these rules establish characteristic patterns for sublineages of MTBC, they are not exclusive and in some cases overlap. Up to 4 rules fired per example. The mode of the number of rules fired per record was 2. In practice, a precedence or order is introduced over the rules using expert knowledge so that unambiguous sublineage predictions are generated. However, this precedence has not been published for sublineages and is up to the individual user of the rules. The SpolDB4 rules have continued to evolve as new lineages such as H3-URAL-1 which are added and refined, and thus the exact rules that we used are provided in Supplement 1 in the Supplementary Materials available online at http://dx.doi.org/10.1155/2014/398484.

3.2. CDC-Sublineage and CDC Rules. The second dataset, CDC-Sublineage, examines 1286 MTBC isolates genotyped by spoligotyping and labeled with 8 sublineages. Dr. Lauren Cowan of the CDC was interviewed to obtain 8 rules of thumb. The data is a subset of 31,482 MTBC isolates genotyped by spoligotyping and 12-locus mycobacterial interspersed repetitive units (MIRU) typing with known lineages from a set collected by the CDC as part of routine TB surveillance in the United States from 2004 to 2009. Since only spoligotypes are used in the rules, the data for training were restricted to spoligotypes with labeled sublineages.

There are 8 rules expressed as numeric formulas based on the 43 spacers in the spoligotypes. For example, the rule for the Indo-Oceanic sublineage Manila (or EAI2-Manila in SpolDB4 rules) is

> If absence of spacers 3, 20, 21, 29–32, 34, sum(spacers33–36) > 0 and presence of spacers 2, 4, 19, 22, then Indo-Oceanic Manila.

There is one rule per sublineage. This dataset was preprocessed by adding an array of 8 bits, one bit per rule. The value of a bit was set to 1 if the rule was fired and zero otherwise. Note that the sublineage sizes are unequal. Overall, the minimum sublineage size was 39, the maximum sublineage size was 356, and the median was 138 records. The rules were ambiguous and no precedence was imposed. In some cases no rules fire for a record. A maximum of 2 rules is fired for each record. The mode of the number of rules fired per record was also 2. If multiple rules fire for a record and the sublineages determined conflict or if no rules fire, the record is considered to be misclassified. Details of the CDC rules can be found in Supplement 2.

4. SITVIT Experimental Results

In this section, we examine the effectiveness of the KBBN model for prediction of SITVIT classification results. Our experiments consist of two parts: (1) in-sample accuracy

Table 2: Training F-measure for KBBN trained on all 10633 SITVIT isolates.

Clade	F-measure	Clade	F-measure	Clade	F-measure
AFRI	0.800	H	0.736	PINI	0.750
AFRI_1	0.944	H1	0.924	PINI1	1.000
AFRI_2	0.908	H2	0.875	PINI2	0.667
AFRI_3	0.966	H3	0.915	S	0.976
Beijing	1.000	H3-Ural-1	0.873	T	0.926
BOV	0.948	H37Rv	0.958	T1-RUS2	0.778
BOV_1	0.993	H4-Ural-2	0.933	T2	0.953
BOV_2	1.000	LAM	0.947	T2-Uganda	0.991
BOV_3	0.644	LAM1	0.977	T3	0.964
BOV_4-Caprae	0.891	LAM11-ZWE	0.954	T3-ETH	0.65
Cameroon	0.929	LAM12-Madrid1	0.947	T3-OSA	0.626
CANETTI	1.000	LAM2	0.991	T4	0.988
CAS	0.937	LAM3	0.988	T4-CEU1	1.000
CAS1-Delhi	0.961	LAM4	0.970	T5	0.984
CAS1-Kili	0.973	LAM5	0.978	T5-Madrid2	1.000
CAS2	0.921	LAM6	0.856	T5-RUS1	0.949
EAI	0.982	LAM8	1.000	T-Tuscany	1.000
EAI1-SOM	0.986	Manu_ancestor	1.000	Turkey	0.928
EAI2-Manila	0.984	Manu1	0.991	X1	0.989
EAI2-Nonthaburi	1.000	Manu2	1.000	X2	0.963
EAI3-IND	0.963	Manu3	1.000	X3	0.995
EAI4-VNM	1.000	Microti	0.750	ZERO	0.800
EAI6-BGD1	0.989				
EAI7-BGD2	1.000			AVERAGE	**0.930**
EAI8-MDG	1.000				

of the SITVIT KBBN model trained using all available data and (2) out-of-sample accuracy of the SITVIT KBBN model trained on the SITVIT-Train and tested on the much larger SITVIT-Test set.The accuracy of the results was measured using the F-measure on the testing data (harmonic mean of precision and recall) averaged over the classes. The F-measure was selected since it effectively captures performance on the unbalanced multiclass data sets studied here. Reporting class accuracies/errors can be misleading for unbalanced classes such as those in the TB data. The minimum and maximum class sizes are reported in Table 1. The F-value was computed as

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (4)$$

where recall is the percentage of the isolates in a given clade correctly identified as being in that clade and precision is the percentage of isolates predicted to be in a clade that are actually in the clade.

4.1. SITVIT KBBN Model Accuracy. The SITVIT KBBN model was trained to predict 69 sublineages using the combined SITVIT-Train and SITVIT-Test data extracted from SITVITWEB along with the SPOLDB4 rules. Overall the model is very accurate; it correctly classifies 94.3% of all of the spoligotypes, achieving an average F-value of 0.93 across the 69 clades. Table 2 presents the in-sample results of SITVIT-KBBN for each clade. The errors that do occur primarily come from lack of specificity not sensitivity. The model achieves a sensitivity of greater than 82% on all of the clades, but the specificity is below 82% on 13 clades. The T clade, which is known to be ill defined, contributes errors leading to reduced specificity in a wide variety of clades including LAM6, T1-RUS, T3-ETH, T3-OSA, and AFRI. Within the *M. africanum* clades, AFRI is primarily confused with other *M. africanum* clades (AFRI_1, AFRI_2, and AFRI_3) which is an acceptable error. A few Cameroons, H3, and T isolates are mistakenly identified as AFRI. Many BOV isolates are assigned by the model as belonging to BOV_3 indicating that a more expansive definition of BOV_3 may be warranted. There are some minor confusions within the Haarlem sublineages H1, H2, and H3 combined with the new H4-URAL-2 and H3-URAL-1 sublineages. About 16% of H3 is assigned to other classes. This suggests that further refinement of the definition of these sublineages will be ongoing. Microti, PINI, and PINI2 have lower F-values, but this is partially due to the fact that these sublineages have only a few exemplars. More data is needed for these rare lineages to improve the model. The F-value of ZERO is reduced by 6 CAS misclassified as ZERO. The overall specificity averaged over the clades is 0.909 and the sensitivity is 0.965.

Table 3: Results of the F-measures of KBBN based on out of-sample test. The KBBN model was trained on SITVIT-Train (with 2714 records) and tested on SITVIT-Test with 7949 records. Overall average F-measure is 0.939.

Clade	F-measure	Clade	F-measure	Clade	F-measure
AFRI	0.889	H	0.942	PINI	0.667
AFRI_1	0.975	H1	0.977	PINI1	0.923
AFRI_2	0.926	H2	0.625	PINI2	0.522
AFRI_3	1.000	H3	0.944	S	0.956
Beijing	0.980	H3-Ural-1	0.887	T	0.969
BOV	0.981	H37Rv	1.000	T1-RUS2	0.956
BOV_1	0.996	H4-Ural-2	0.960	T2	0.991
BOV_2	1.000	LAM	0.949	T2-Uganda	1.000
BOV_3	1.000	LAM1	0.986	T3	0.969
BOV_4-Caprae	0.914	LAM1l-ZWE	0.976	T3-ETH	0.977
Cameroon	0.967	LAM12-Madrid1	1.000	T3-OSA	0.978
Canetti	0.500	LAM2	0.993	T4	0.984
CAS	0.990	LAM3	0.973	T4-CEU1	1.000
CAS1-Delhi	0.990	LAM4	0.967	T5	1.000
CAS1-Kili	0.846	LAM5	0.985	T5-Madrid2	1.000
CAS2	1.000	LAM6	0.889	T5-RUS1	0.883
EAI	0.989	LAM8	0.970	T-Tuscany	0.889
EAI1-SOM	1.000	Manu_ancestor	1.000	Turkey	0.941
EAI2-Manila	1.000	Manu1	0.995	X1	0.963
EAI2-Nonthaburi	0.933	Manu2	0.997	X2	0.944
EAI3-IND	1.000	Manu3	1.000	X3	0.971
EAI4-VNM	1.000	Microti	0.667	ZERO	0.800
EAI6-BGD1	1.000				
EAI7-BGD2	0.993			Average	**0.939**
EAI8-MDG	1.000				

4.2. Predictive Accuracy Results. To assess the out-of-sample predictive accuracy of the KBBN SITVIT model we trained the model on SITVIT-Train and tested it on SITVIT-Test. The model was very accurate overall achieving an average out-of-sample test F-value of 0.939, almost identical to the in-sample estimate of above 0.930. The average recall (percentage of the isolates in a given clade correctly identified as being in that clade) between all lineages is 97.5%, and the average precision (the percentage of isolates predicted to be in a clade that are actually in the clade) among all lineages is 91.9%. As shown in Table 3, the results for each clade are very similar to those reported in Table 2. The T clade and small rarer clades such as PINI variants and Microti account for much of the decrease in precision.

4.3. Model Validation. The next set of experiments evaluates the effectiveness of the KBBN approach with respect to other techniques and the effectiveness of incorporating rules. All experiments were done on both the SITVIT and CDC datasets to ensure that the results are not an artifact of a single dataset. For each dataset, first we used 10-fold stratified cross validation. Each training set was divided into 10 parts with 9 parts available as the training data for creation of models and 1 part held out as an independent test set. For all experiments the same test sets were employed, but the training dataset or the set of rules used may be varied. The accuracy of the results was measured using the F-measure on the testing data (harmonic mean of precision and recall) averaged over the classes. To facilitate a fair comparison, the data were constructed so that there are at least 10 records per class. In the SITVIT domain, this required removing clades that do not commonly infect human beings (e.g., PINI1 and PINI2). We refer to this subset of the SITVIT-Train dataset as SITVIT-CV. SITVIT-CV had 45 classes and 2593 records. The minimum sublineage size was 11, and the maximum sublineage size was 390 with a mode of 21 records.

4.3.1. Comparison with Other Techniques. We designed several sets of experiments on the two datasets SITVIT-CV and CDC-Sublineage to determine the following: if incorporating rules improved the performance of the Bayesian network over the performance of the BN or rules alone. The results were gathered for KBBN, BN, and the rules used alone. In addition, linear and nonlinear SVM results were provided as a baseline for comparison. The SVM implementation in WEKA (http://www.cs.waikato.ac.nz/ml/weka/) was used. The SVM kernels and parameters were selected using a grid search of 9-fold cross validated accuracy of the training set. The degree-three polynomial kernel and radial basis function kernels were found to work best. All SVM data was normalized

Table 4: Average F-measure of KBBN, BN, Rules-only, and SVM (nonlinear and linear) on two testbeds. While using Rules-only provides poor results, KBBN is able to provide results that are significantly better or at least not worse than BN and SVM on both domains. Results significantly different from KBBN at 5% significance level are shown in bold.

Dataset	Model				
	KBBN	BN	Rules-only	SVM nonlinear	SVM linear
SITVIT-CV	0.945	**0.771**	**0.345**	**0.903**	**0.914**
CDC-Sublineage	0.981	**0.934**	**0.312**	0.994	0.993

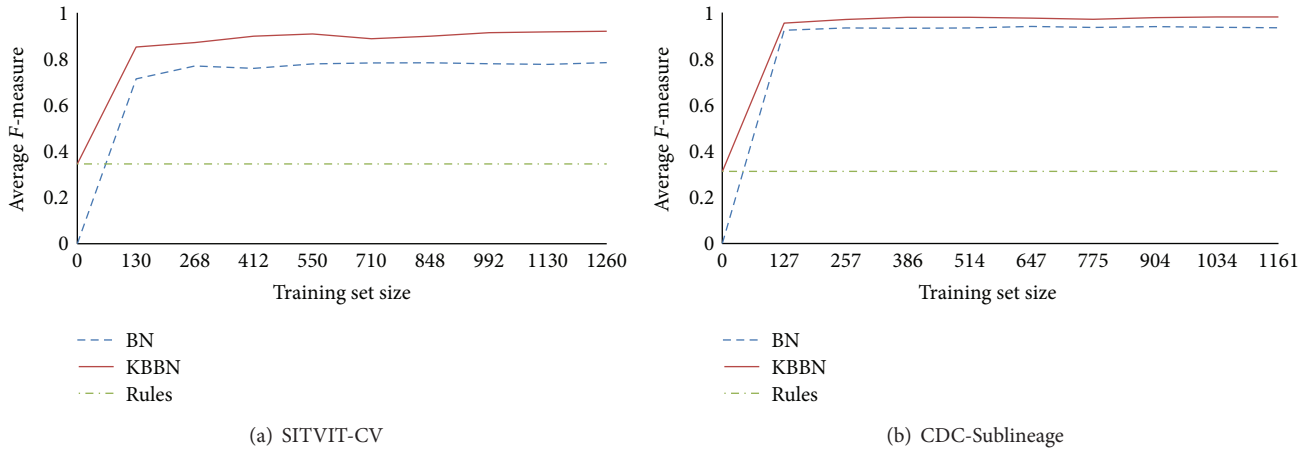


Figure 3: The result of adding rules to different training set sizes for the (a) SITVIT-CV and (b) CDC-Sublineage testbeds.

before training. Also, we are interested in the nature of the misclassification because it tells us about the potential inaccuracies in the definition of the lineages.

Table 4 compares the results of KBBN, BN, Rules-only, and SVM (nonlinear and linear) on the two testbeds. The rules themselves have very poor overall accuracy, but they led to improvements over the baseline BN accuracy on both datasets with statistically significant improvements on CDC-Sublineage and SITVIT-CV. The SVM results indicate that KBBN's accuracy is competitive with state-of-the-art nonlinear and linear classification methods. But note that KBBN, being a generative method, has many advantages over SVM such as availability of posterior probabilities of each class given the observation that can be interpreted as the confidence of the prediction, easier interpretation, and ease of incorporation of domain knowledge.

4.3.2. Effectiveness of Rules in Bayesian Network. Next, we designed several sets of experiments to determine the following: how the quality and quantity of rules and data affected the performance of KBBN. The basic underlying experimental design was the same for experiments across the two testbeds.

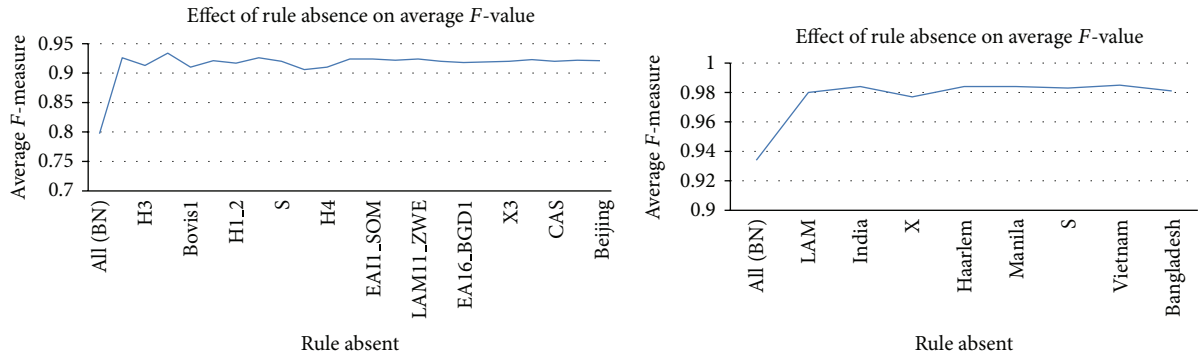
Our hypothesis was that KBBN can learn the concept faster with less data by adding rules. We wanted to show that rules can improve learning especially where you have less data. For each dataset, first we used 10-fold stratified cross validation. Next each training set was divided into 9 parts providing models using 1/9, 2/9,... or 9/9 of the training set and tested on the corresponding test set. The test sets were kept the same for different training set sizes. We measured the amount of F-value for different training set sizes with or without the rules and compared the result with the case of using no data at all (i.e., BN case) or only rules. The results are presented in Figure 3. Similar smaller testing set studies on CDC-Sublineage and SITVIT-CV found that KBBN always performs better than or as good as BN for all training set sizes.

To further examine the effect of incorporating rule sets and using incomplete rules, we performed two sets of experiments described in the following section: (1) using increasing percentages of the available rules and (2) using subsets of rules, removing rules for a given class at a time.

4.3.3. Removal of Rule Sets for a Class . In these experiments, we examined the effect of removing all the rules associated with a given class. We examined the KBBN accuracy and recorded the amount of average F-value between all classes after all the rules corresponding to a single class are removed. Again, 10-fold stratified cross validation was performed. The results are presented in Figure 4. "All (BN)" is when no rules are used in KBBN, which is equivalent to BN performance. Clearly, KBBN can lead to significant improvements compared to when no rules exist for entire classes of MTBC. We leave a more comprehensive study of when rules are most helpful for problems in other domains to future work.

5. Quality of Rules

KBBN can provide us with information about the quality of each rule. We studied posterior probabilities of rules given the class to provide insight into the utility and accuracy of each



(a) SITVIT-CV (b) CDC-Sublineage

Figure 4: Effect of removing rules for each class on the average F-value for (a) SITVIT-CV and (b) CDC-Sublineage.

Table 5: Posterior probability of each rule given class for CDC-Sublineage dataset. Blanks indicate 0.

	Class						
	Haarlem	LAM	S	X	India	Manila	Vietnam
Rule							
Haarlem	0.707				0.015		
LAM		1.000					
S			1.000		0.015	0.005	0.033
X				1.000	0.015		0.017
India					0.970	0.022	0.017
Manila						0.735	
Vietnam							0.283
No rule	0.297				0.015	0.243	0.700

rule. The $P(r \mid c)$ is of great interest because it tells us how good rule r is for a given class c. The posterior probability of the rules given the classes for the CDC-Sublineage data is presented in Table 5. The table includes a row for "No rule" indicating the probability of no rule getting fired. When no rule is fired a regular BN is used instead of KBBN. Note that the probabilities within columns may sum to more than 1 since rules are not mutually exclusive.

For CDC, the rule for LAM exactly corresponds to the class LAM on this data, since $P(\text{Rule} = \text{LAM} \mid \text{Class} = \text{LAM}) = 1$ and all other probabilities in the LAM row or column are 0. The rules for S and X correctly fire for their respective classes, but they also fire incorrectly for other lineages as indicated by the other entries in the S and X rows. The rules for Haarlem and Manila correctly predict their corresponding sublineages, but the fact that "No Rule" occurs 29.6% and 24.3% of the time, respectively, indicates that these rules fail to cover all members of their class. For the India class, the India rule is quite accurate, but the rules can be ambiguous as indicated by the multiple entries in the India column. Most Vietnams are not covered by any rules and for those that are covered the rules may be ambiguous.

We provide the posterior probability distribution of each rule given the sublineage for the SITVIT-CV dataset as a heat map in Figure 5. Good rules only have red on the diagonal. A rule fires for multiple classes if it has multiple red entries in a row. The rule set is ambiguous for a class if there are multiple red entries within a given class column. Notice that the rules that are fired for many classes with high probability

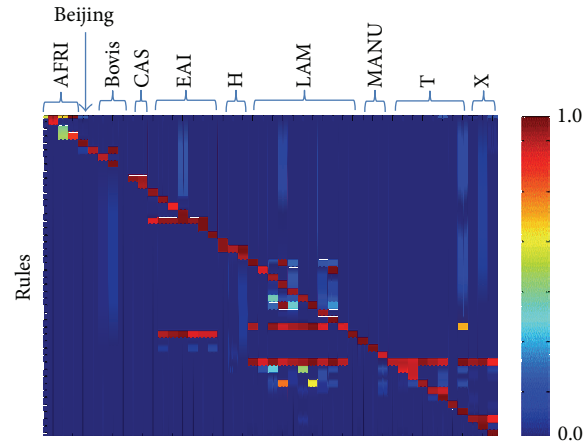


Figure 5: The heat map represents the posterior probability of each rule given the sublineage for the SITVIT dataset. A strong association of a rule in predicting a sublineage is shown with a red square while a blue square represents no relation. Here H includes URAL-1 and URAL-2 and LAM includes Turkey and Cameroon sublineages.

(e.g. T) are not very effective in indicating the associated class as opposed to Beijing which is an effective rule.

6. Discussion of Alternative Knowledge-Based Approaches

The KBBN has a great appeal over alternative knowledge-based approaches such as knowledge-based SVM (KBSVM) and knowledge-based neural networks (KBANN) [8, 9, 18]. The first advantage is that no special encodings of the rules are required. In KBANN, the rules are mapped into a neural network by converting the data to numeric form and designing appropriate nodes, links, and weights in the neural network. KBSVM requires each rule to be encoded as a polyhedral rule, such as $if\ x \in R^n\ satisfies\ Bx \leq d\ then\ class = 1$. In KBSVM, the process of converting rules to probabilities can greatly increase the number of rules. For example, for the task of predicting major lineages of MTBC, the 13 original logical rules published in a prior study [8] were mapped into 29 polyhedral rules. The added rules help capture the precedence of the original rules which made them mutually exclusive. There is no easy way to capture rule precedence in KBSVM or KBANN. KBSVM must add rules of the form "*if x satisfies condition A then x is not in Class y*." In KBBN, the data can be numeric or symbolic and each rule may be any arbitrary function of the observations to the classes. As reported in the preliminary study [9], KBBN works effectively on rules with and without precedence.

A second advantage of KBBN is the computational complexity of training. For the polytree type KBBN studied here, both training and inference can be done very efficiently in polynomial time. The MAP estimation of KBBN parameters has a closed form solution enabling globally optimal solutions to be found by simple counting algorithms. The only additional computation required over BN is representation of the probabilities of the classes given the rules. Unlike KBSVM and KBANN, no special purpose software is required for mapping and training KBBN beyond knowing which rules are fired for which example. KBSVM requires the solution of challenging nonconvex programs with many constraints and variables introduced for each rule over the original SVM. Similarly, KBANN also requires the solution of a nonconvex program of considerably greater complexity than the original ANN network due to the addition of weights and nodes to capture the rules. Special purpose software is needed to create the structure of the KBANN network, but any neural network training algorithm can be used to train it.

Additional benefits of KBBN over KBSVM and KBANN include that KBBN can be easily used for multiclass problems, it provides estimates of the posterior probabilities of the classes, and the resulting classification function is more transparent and explainable. KBSVM classification results published to date are limited to two-class problems and how to do efficient multiclass KBSVM remains an open research question. In both KBSVM and KBANN, the rules are used only to bias the construction of the prediction function and the prediction is typically a black box. In contrast, the KBBN probability density functions are readily interpretable as soft relaxations of the visual rules already used in TB. The posterior probabilities of the rules can be used to explain the effectiveness of these rules.

7. Conclusions

We have developed an effective classifier to predict SITVIT MTBC clades with high accuracy. The result is a publicly available web-based tool for SITVIT clade classification to support TB control and research efforts available for use at TB-INSIGHT (http://tbinsight.cs.rpi.edu/run_tb_lineage.html) and later on SITVIT2. We established that the clade estimates are robust by performing two out-of-sample testing experiments. Furthermore, the results on the two testbeds show that KBBN is a highly accurate classifier that can outperform methods based on rules or Bayesian networks trained on data alone and that meets or beats the performance of nonlinear and linear SVM models. KBBN proved to be robust to ambiguity, incompleteness, and inaccuracy of the rule set. The results here are limited to simple commonly used BN that are polytrees using MAP estimation, but future work is needed to examine the KBBN approach on more general BN models and algorithms.

As a general approach, KBBN has many attractive properties. It allows any type of rules to be incorporated into a Bayesian network with little increase in the model and training complexity. Prior knowledge-based SVM required manipulation of the rules, models, data, and/or kernel [8, 9]. There is no need to introduce precedence or resolve inconsistency of the rules for KBBN. The KBBN model can provide posterior class probabilities as well as information on how the rules were used and how classification decisions were made. We studied the posterior probability of rules given the class to provide insight into the utility of each rule. This underlines another advantage of KBBN as a generative model over its discriminative competitor models, like KBSVM. Thus KBBN offers a promising research direction for combining rule and data-driven predictive methods that may be applicable in many domains.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was made possible by and with the assistance of Dr. Jeffrey R. Driscoll and Dr. Lauren Cowan of the CDC. This work is supported by NIH R01LM009731. Thanks are due to Cagri Ozcaglar and Lei Yao for providing technical assistance.

References

[1] D. G. de Viedma, I. Mokrousov, and N. Rastogi, "Innovations in the molecular epidemiology of tuberculosis," *Enfermedades*

Infecciosas y Microbiología Clínica, vol. 29, supplement 1, pp. 8–13, 2011.

[2] S. Gagneux, K. DeRiemer, T. Van et al., "Variable host-pathogen compatibility in *Mycobacterium tuberculosis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 8, pp. 2869–2873, 2006.

[3] K. Brudey, J. R. Driscoll, L. Rigouts et al., "*Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpoIDB4) for classification, population genetics and epidemiology," *BMC Microbiology*, vol. 6, article 23, 2006.

[4] I. Filliol, A. S. Motiwala, M. Cavatore et al., "Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set," *Journal of Bacteriology*, vol. 188, no. 2, pp. 759–772, 2006.

[5] R. M. Warren, E. M. Streicher, S. L. Sampson et al., "Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data," *Journal of Clinical Microbiology*, vol. 40, no. 12, pp. 4457–4465, 2002.

[6] C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, and K. P. Bennett, "Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors," *BMC Genomics*, vol. 12, supplement 2, article S1, 2011.

[7] I. Vitol, J. Driscoll, B. Kreiswirth, N. Kurepina, and K. P. Bennett, "Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes," *Infection, Genetics and Evolution*, vol. 6, no. 6, pp. 491–504, 2006.

[8] G. Kunapuli, K. P. Bennett, A. Shabbeer, R. Maclin, and J. Shavlik, "Online knowledge-based support vector machines," in *Machine Learning and Knowledge Discovery in Databases*, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds., vol. 6322 of *Lecture Notes in Computer Science*, pp. 145–161, 2010.

[9] M. Aminian, C. Ozcaglar, A. Shabbeer, S. Vandenberg, K. Hadley, and K. P. Bennett, "Knowledge-based Bayesian network for the classification of *Mycobacterium tuberculosis* complex sublineages," in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB '11)*, pp. 201–208, August 2011.

[10] A. Shabbeer, L. S. Cowan, C. Ozcaglar et al., "TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 789–797, 2012.

[11] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press, 2009.

[12] M. Aminian, A. Shabbeer, and K. P. Bennett, "A conformal Bayesian network for classification of *Mycobacterium tuberculosis* complex lineages," *BMC Bioinformatics*, vol. 11, supplement 3, article S4, 2010.

[13] C. Demay, B. Liens, T. Burguière et al., "SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 755–766, 2012.

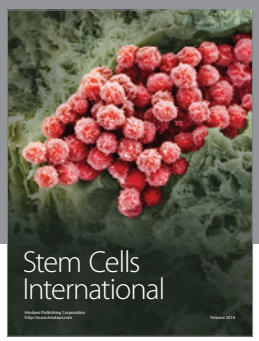
[14] F. K. Koro, Y. K. Simo, F. F. Piam et al., "Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping," *Journal of Clinical Microbiology*, vol. 51, no. 1, pp. 299–302, 2013.

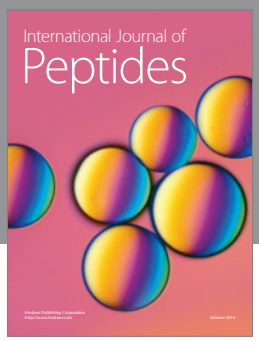
[15] E. Abadia, J. Zhang, T. dos Vultos et al., "Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method," *Infection Genetics and Evolution*, vol. 10, no. 7, pp. 1066–1074, 2010.

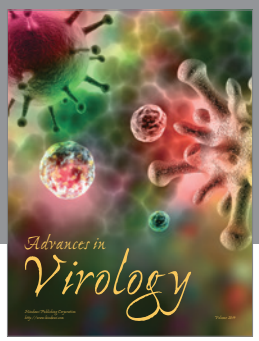
[16] O. Kisa, G. Tarhan, S. Gunal et al., "Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey," *PLoS ONE*, vol. 7, no. 1, Article ID e30331, 2012.

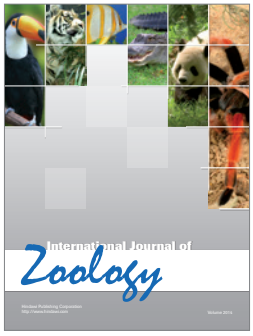
[17] I. Mokrousov, "The quiet and controversial: ural family of *Mycobacterium tuberculosis*," *Infection Genetics and Evolution*, vol. 12, no. 4, pp. 619–629, 2012.

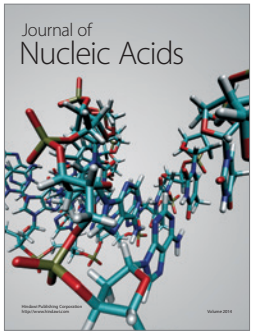
[18] G. Towell and J. Shavlik, "Knowledge-based artificial neural networks," *Artificial Intelligence*, vol. 70, no. 1, pp. 119–165, 1994.

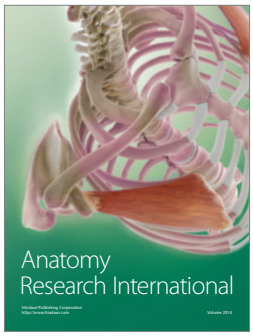


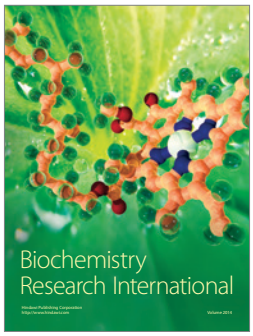


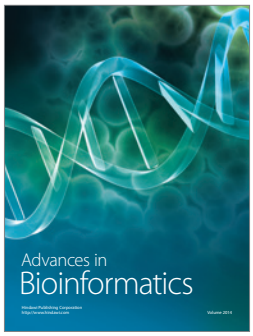


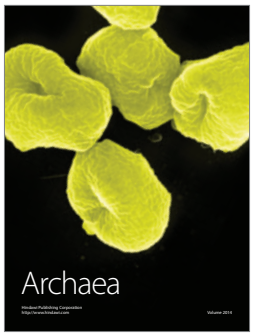


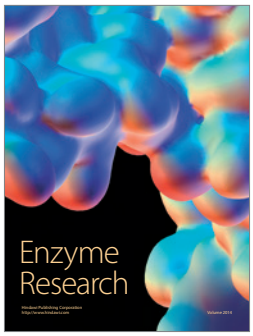


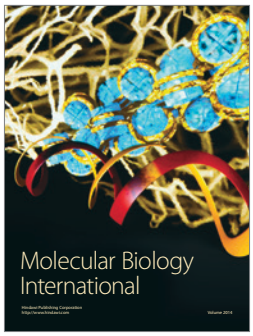












Hindawi

Submit your manuscripts at
http://www.hindawi.com

Article 06

Infection, Genetics and Evolution 14 (2013) 434–443

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Genetic diversity among multidrug-resistant *Mycobacterium tuberculosis* strains in Mexico

Armando Martinez-Guarneros [a,b], Nalin Rastogi [c], David Couvin [c], Alejandro Escobar-Gutierrez [a], Livia Maria Gonçalves Rossi [a], Carlos A. Vazquez-Chacon [a], Sandra Rivera-Gutierrez [b], Daniela Lozano [d], Arely Vergara-Castañeda [e], Jorge A. Gonzalez-y-Merchand [b], Gilberto Vaughan [a,*]

[a] *Instituto de Diagnóstico y Referencia Epidemiológicos, Secretaría de Salud, Mexico City, Mexico*
[b] *Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, Mexico City, Mexico*
[c] *WHO, Supranational Tuberculosis Reference Laboratory, Institut Pasteur de Guadeloupe, Abymes, France*
[d] *Centro de Investigación y Estudios Avanzados, Instituto Politécnico Nacional, Mexico City, Mexico*
[e] *Facultad de Ciencias Químicas, Universidad La Salle, Mexico City, Mexico*

ARTICLE INFO

Article history:
Received 2 October 2012
Received in revised form 11 December 2012
Accepted 14 December 2012
Available online 16 January 2013

Keywords:
Mycobacterium tuberculosis
Spoligotyping
MIRU-VNTR
MDR-TB

ABSTRACT

Tuberculosis is an important public health problem in Mexico. However, limited information about the genetic diversity of *Mycobacterium tuberculosis* strains circulating in the country is available. In this work, 109 multidrug-resistant (MDR) *M. tuberculosis* isolates collected in 23 different states of Mexico in 2003 were retrospectively characterized by spoligotyping and MIRU-VNTRs. All isolates, except for a single cluster containing two strains (subcluster E1), were split when information from the 12-loci MIRUs and spoligo-pattern was simultaneously analyzed. The discriminative power of 12-loci MIRU-VNTR and spoligotyping, by the Hunter-Gaston index, were 0.9998 and 0.9011, respectively. These findings suggest that almost all cases were epidemiologically unrelated. Instead, the genetic variations observed among these strains are suggestive of emergence of acquired drug-resistance during the course of treatment. The results suggest a high degree of genetic variability and a high frequency of SIT53 (T1 family) spoligotype among the MDR *M. tuberculosis* isolates included in the study.

Published by Elsevier B.V.

1. Introduction

Tuberculosis (TB) is still one of the most important causes of morbidity and mortality in the world (Supply et al., 2001). Although TB is a curable disease, it is considered a global public health problem affecting developing countries (Dye et al., 1999). Worldwide, approximately nine million new TB cases and 1.6 million deaths occur annually. In Mexico, the reported annual incidence of pulmonary TB is 14.27 cases per 100,000 inhabitants (Secretaria de Salud, 2005).

The emergence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB brings several concerns and jeopardizes the success of strategies aimed to control disease spread (Cox et al., 2008). It has been estimated that the global proportion of resistance among all cases is ~4.8%, with approximately 489,000 cases of MDR-TB occurring worldwide (WHO, 2010). Different studies have suggested that TB drug resistance is a growing problem in Mexico (Banerjee et al., 2008; Garcia-Garcia Mde et al., 2002; Martinez-Guarneros et al., 2012; Molina-Torres et al., 2010; Nava-Aguilera et al., 2011; Quitugua et al., 2002; Zenteno-Cuevas et al., 2012). Additionally, incompleteness of therapy among Latino individuals, which has been associated with idiosyncrasy and cultural barriers, seems to play role in treatment adherence and ultimately in treatment success (Kandula et al., 2004).

Limited information is available regarding the transmission routes of TB in Mexico. TB molecular epidemiology has rapidly become an important tool for the study of TB transmission and control (Daley, 2005). Several methods have been reported so far to assess the molecular epidemiology of TB (Kamerbeek et al., 1997; Reed et al., 2009). These methodologies allow the recognition of outbreaks, the correct typing of isolates, and the identification of transmission patterns in human populations. Among these methods, spoligotyping is a well-recognized method for the classification of TB isolates into distinct phylogenetic lineages. This is a PCR-based typing method targeting the 36-base pair direct repeat (DR) locus which generates distinct patterns based on the hybridization of 43 different oligonucleotides spacer sequences (Kamerbeek et al., 1997). Spoligotyping used in conjunction with mycobacterial interspersed repetitive units – variable number of

* Corresponding author. Address: Departamento de Investigaciones Inmunológicas, Instituto de Diagnósticos y Referencia Epidemiológicos, Carpio 470, Col. Santo Tomas, Mexico City 11340, Mexico. Tel.: +52 55 5342 7563; fax: +52 55 5342 1230. Current address: Centers for Diseases Control and Prevention, Atlanta, GA, USA.
E-mail address: gilvaughan@yahoo.com (G. Vaughan).

1567-1348/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.meegid.2012.12.024

DNA tandem repeats (MIRU-VNTRs), has emerged as a fast, reliable, and cost-effective alternative to the traditional IS*6110*-RFLP genotypic screening (http://www.miru-vntrplus.org/).

In this study, we assess the genetic diversity by spoligotyping and 12-MIRU-VNTRs of MDR *M. tuberculosis* isolates collected in Mexico in 2003

2. Materials and methods

2.1. Clinical samples and patients' characteristics

In this study we analyzed 109 MDR-TB isolates collected in Mexico from 23 different states in 2003 through the national tuberculosis surveillance as part of the national drug susceptibility testing (Fig. 1). Ethical reviews and informed consent approval were granted by the Ethical Committee of the Mexican Institute for Epidemiological Diagnosis and Reference. Informed consent was obtained from all patients enrolled in the study. All strains were isolated by the corresponding local state laboratory and subsequently submitted to the National Reference Laboratory. Subjects were asked to provide three serial sputa samples which were subjected to mycobacteria isolation by culture in Löwenstein-Jensen (LJ) slants.

2.2. Mycobacterial isolation

Sputum decontamination was performed by the Petroff's sodium hydroxide method, acid fast bacillus (AFB) Ziehl-Neelsen staining technique followed by culture on Löwenstein-Jensen and classical identification using the niacin, nitrate reduction and 68 °C catalase (David et al., 1989; Kent and Kubica, 1985). The samples (300 μl) were inoculated onto two different LJ slants. The slants were inspected during the first week to check for contamination and fast growers, and subsequently, every week for up to 8 weeks. Cultures that remained negative after eight weeks were considered negative. Colony morphology and acid-fast bacilli identification on smears by Ziehl-Neelsen technique were performed. Biochemical analyses for differentiation included nitrate reduction and niacin accumulation tests.

2.3. Drug susceptibility testing

Standard testing was performed for first-line anti-tuberculosis drugs (streptomycin [S], isoniazid [H], rifampicin [R], ethambutol [E], and pyrazinamide [Z]) as previously described (Canetti et al., 1963), followed by reconfirmation of MDR isolates using MGIT-960 method (Becton Dickinson, La Jolla, CA). *M. tuberculosis* strain H37Rv (ATCC, Manassas, VA), a susceptible strain to all first line drugs, was included as control in all experiments.

2.4. Genotyping methods

All *M. tuberculosis* strains were typed by spoligotyping and MIRU-VNTR analysis. Spoligotyping was carried out as described previously (Kamerbeek et al., 1997). *M. tuberculosis* H37Rv and *M. bovis* BCG (Pasteur strain) were used as controls.

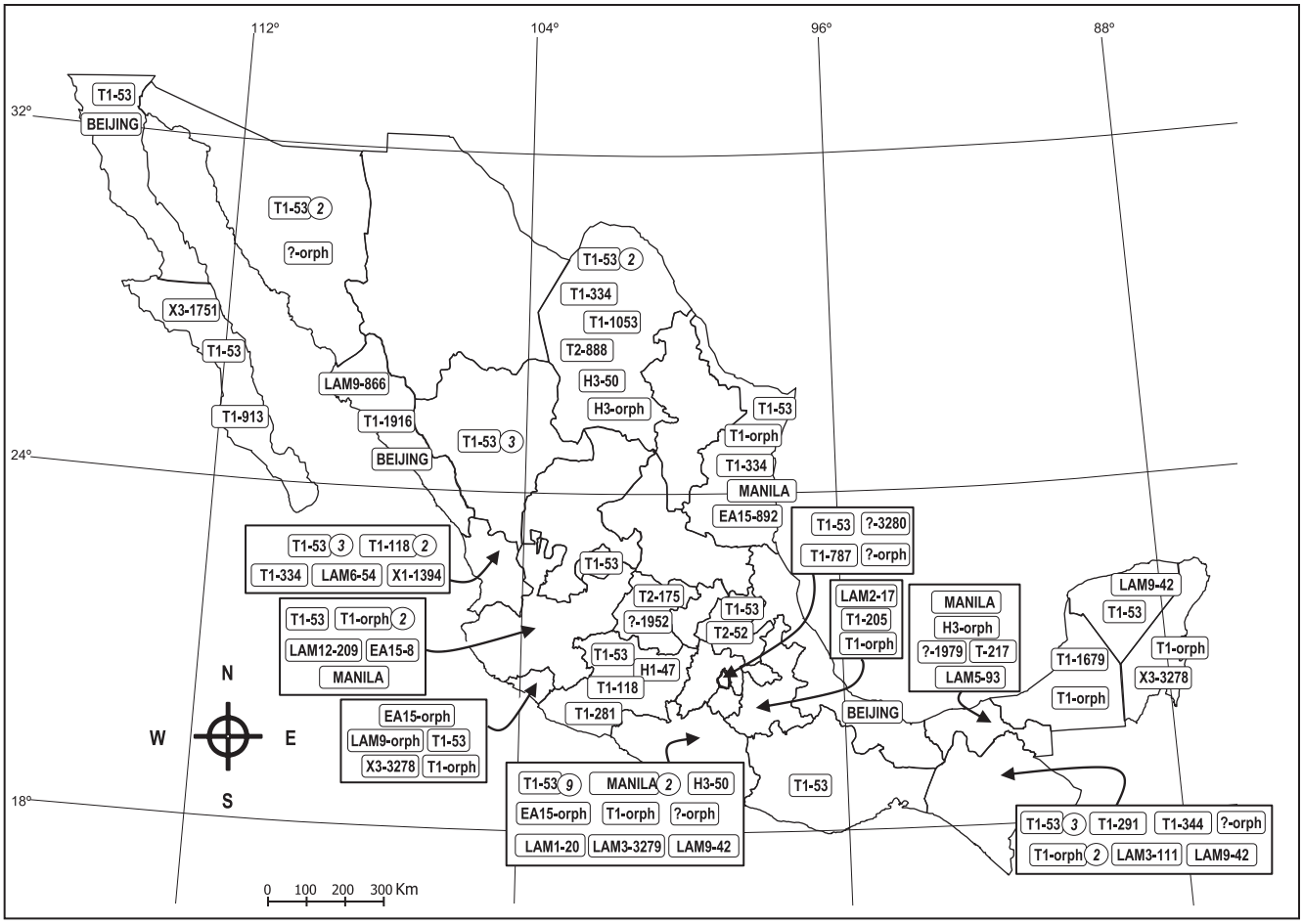


Fig. 1. TB spoligotype distribution in Mexico. The distribution of all spoligotypes identified in this study is depicted on the map (clade and SIT are shown for each type). The numbers inside the circles represent the number of isolates belonging to the corresponding type.

MIRU-VNTR typing was performed as described elsewhere (Supply et al., 2006; Supply et al., 2001). Genotypes were expressed as numerical codes representing the number of MIRU-VNTR for each locus. A dendrogram was constructed by the unweighted-pair group method using average linkages (UPGMA) using the MIRU-VNTRplus web application (http://www.miru-vntrplus.org/MIRU/index.faces) after pairwise comparison of strains by calculation of the Jaccard index. Both the MIRU-VNTR copy numbers and Spoligo-patterns were used to generate the corresponding tree.

A minimum spanning tree (MST) illustrating evolutionary relationships between MIRU12 patterns of the Beijing lineage strains in our study versus all the Beijing lineage strains with MIRU12 data from USA in the proprietary SITVIT2 database of Institut Pasteur de la Guadeloupe, was drawn using BioNumerics version 6.6 (Applied Maths NV, Sint-Maartens-Latem, Belgium).

2.5. Database comparison and geographical distribution of spoligotypes

Spoligotypes in binary format were converted to an octal code for comparison with the proprietary SITVIT2 database of Institut Pasteur de la Guadeloupe, which is an updated "in-house" version of the SITVITWEB database (Demay et al., 2012). In this database, SIT (Spoligotype International Type) and MIT (MIRU International Type) designate spoligotyping and MIRU patterns shared by two or more patient isolates, respectively, whereas "orphan" designates patterns reported for a single isolate. Major phylogenetic clades were assigned according to signatures provided in the database.

The discriminative power for both typing methods was calculated as reported elsewhere (Hunter and Gaston, 1988).

2.6. Statistical analysis

The $\chi 2$ and Kruskal–Wallis tests were used to determine the significant differences. *t*-test was used for age comparison among clustered and non-clustered isolates. Additionally, logistic regression and general linear model analysis was used to establish relationships between the corresponding spoligotype using gender and age as variables. Greenhouse-Geisser corrections were used when the model departed from the assumption of sphericity. All *p*-values were calculated using a 2-sided test.

3. Results

3.1. Characteristics of population studied

In 2003, 168 MDR TB isolates were reported in Mexico. Fifty-nine isolates were not included in the study for diverse reasons (poor growth, insufficient sample, etc.,). All remaining 109 isolates were identified as *M. tuberculosis*. The age of the patients ranged between 3 and 83 years, with an average of 45 years. The vast majority of the subjects (73%) were aged between 25 and 54 years. In this cohort, 61 (56%) male and 46 (44%) female individuals were included.

3.2. First-line drug susceptibility testing

The isolates were subjected to drug resistance testing. All 109 isolates were confirmed to be resistant to at least H and R. In addition, drug resistance testing to S, E and Z was also performed. Overall, resistance to other first-line drugs was similar (42, 47 and 44% for S, E and Z, respectively). The subdivision of 109 MDR strains in function of combined drug resistance was as follows: only H and R, 32/109 (29.4%); combined resistance to three drugs, 31/109 (28.4%); combined resistance to four drugs, 24/109 (22%); combined resistance to five drugs, 22/109 (20.2%).

3.3. Spoligotyping

All MDR TB isolates were subjected to 43-spacers spoligotyping. Sixty-two different spoligotypes were found. Among these, 22 orphan strains, obtained from 13 different states, were identified (Table 1). Ten strains corresponded to the T1 family while two isolates belonged to the EAI2-Manila family, and two more to the EAI5 family. One more strain belonged to the LAM9 family. Five isolates displayed unknown patterns with no matches to any of the major clades present in the database. Eighty-seven isolates, collected in 23 different states, were categorized into 40 shared-types (Table 2). Thirty-one strains exhibited unique SIT patterns. The remaining isolates formed 9 different clusters. SIT53, T1 Family, represented the predominant pattern, 34 different isolates accounting for 31.2% of all isolates in the study. The vast majority of the isolates belonged to the T and LAM families (Supplemental Table S1). Other minor families included the Beijing, EAI and H families. Table 3 depicts the predominant spoligotypes found among these isolates. Nine different SITs representing 56 strains were identified. As mentioned previously, SIT53-T1 was the predominant type (31.2% of isolates in our study), followed by SIT118-T1 (2.75%). SIT1-Beijing, SIT8-EAI5, SIT42-LAM9 and SIT334-T1 accounted for 2.75% each. Finally, SIT50-H3, SIT3277-EAI2-Manila and SIT3278-X3 represented 1.83% each. The discriminative power of the spoligotyping method, measured by the Hunter-Gaston index, was 0.9011.

3.4. MIRU-VNTR typing

The 12 loci MIRU-VNTR typing scheme allowed the identification of one subcluster (E1) which grouped two isolates from the State of Guerrero belonging to the T family (SIT53-T1; cluster E; Supplemental Table S1). Thus, all isolates were split by the 12-loci MIRU platform. Ninety-six isolates were not found in the database; and therefore, the strains were classified as "orphan" types, and 11 other strains were "unique". Clustering using both the MIRU-VNTR and spoligotyping identified two identical strains, 369 and 574 (Fig. 2). Both strains were recovered from the State of Guerrero and belonged to the T1 spoligotype. The Hunter-Gaston index for the 12-loci MIRU-VNTR was 0.9998.

The three Mexican Beijing strains identified in this study were compared to those circulating in the US. The comparative MST analysis using the 12-loci MIRU-VNTR patterns showed that the isolates collected in Mexico were significantly different from the US strains (Fig. 3).

3.5. Distribution of spoligotype and MIRU-VNTR type by age and gender

The statistical analysis showed no correlation between the corresponding type, either spoligotype or MIRU-VINT, and age or gender.

4. Discussion

In this study we have shown that SIT53 was the most predominant spoligotype found among the MDR TB strains included in this study. SIT53 has also been shown to be the most prevalent type in northern Mexico (Molina-Torres et al., 2010) and the second more prevalent type among HIV Mexican co-infected individuals (Lopez-Alvarez et al., 2010). Additionally, the southern parts of US have also shown the circulation of SIT53 which can be explained by the large population of Mexican origin residing in the region (Soini

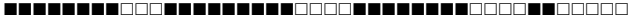
Table 1
Orphan strains (*n*=22) and corresponding spoligotyping defined lineages/sublineages and MIRU-VNTRs patterns found among a total of 109 MDRTB isolates of *M. tuberculosis* isolated from patients residing in Mexico.

Strain	Year	Sex/ Age	Isolation city	Drug resistance[a]					Spoligotype Description	Octal code	Lineage[b]	MIRU12	MIT[c]
				STR	INH	RIF	ETB	PZA					
17	2003	M/67	TAMAULIPAS	R	R	R	R	R		177777637760771	T1	224328163227	Orphan
187	2003	F/30	CHIAPAS	S	R	R	S	S		400067777760671	T1	224336263323	Orphan
346	2003	M/45	CAMPECHE	R	R	R	S	S		777777774160771	T1	223225163324	1196
347	2003	F /40	CHIAPAS	S	R	R	S	S		777177616040771	Unknown	213216132213	Orphan
579	2003	M/46	GUERRERO	S	R	R	S	S		400036777413771	EAI5	254328513322	Orphan
591	2003	M/31	COAHUILA	R	R	R	S	S		777777560020611	H3	228335141213	Orphan
771	2003	F/60	COLIMA	R	R	R	S	R		700077777413771	EAI5	210021234544	Orphan
878	2003	M/38	QUINTANA ROO	S	R	R	S	R		557777777760771	T1	223325123322	1054
942	2003	F/68	GUERRERO	S	R	R	S	S		777767477403671	EAI2-Manila	265437334543	Orphan
948	2003	M/43	JALISCO	S	R	R	R	R		777747777760671	T1	154238344532	Orphan
964	2003	F/25	TABASCO	R	R	R	S	S		777777774120771	H3	225324265433	Orphan
984	2003	F/39	PUEBLA	R	R	R	R	R		677777477060771	T1	335436154433	Orphan
1013	2003	M/18	COLIMA	S	R	R	S	R		777767607760671	LAM9	221202051020	Orphan
1014	2003	M/60	COLIMA	S	R	R	S	S		776247777760571	T1	224324153324	815
1052	2003	M/79	TAMAULIPAS	R	R	R	S	S		777767477413671	EAI2-Manila	254315223322	Orphan
1164	2003	M/19	SONORA	R	R	R	S	R		777767637741771	Unknown	223326273533	Orphan
1244	2003	M/28	GUERRERO	R	R	R	R	R		777367777740771	T1	223316273324	Orphan
1356	2003	M/32	JALISCO	R	R	R	S	R		777777670160771	T1	223325153324	125
1403	2003	F/47	CHIAPAS	S	R	R	R	R		777177677760771	T1	226648394108	Orphan
1423	2003	M/46	DURANGO	S	R	R	S	S		711067777706771	Unknown	225325453323	Orphan
1484	2003	M/64	MEXICO CITY	R	R	R	R	R		777776777767711	Unknown	224228274432	Orphan
1543	2003	M/47	GUERRERO	S	R	R	S	S		770077757400771	Unknown	273228224422	Orphan

[a] Drug resistance; S, Streptomycin; H, Isoniazid; R, Rifampin; E, Ethambutol; Z, Pyrazinamide; r, resistant; s, sensitive.
[b] Lineage designations for orphan patterns were done manually as Expert-based interpretations using revised SpolDB4 rules.
[c] MIT, 12-loci MIRU International Type in the SITVIT2 database.

Table 2
Description of 40 shared-types (SITs; *n*=87 isolates) and corresponding spoligotyping defined lineages/sublineages found among a total of 109 MDR *M. tuberculosis* strains isolated from patients residing in Mexico.

SIT[a]	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. database	Lineage[b]	Clustered vs. unique patterns[c]
1	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■■	000000000003771	3 (2.75)	0.03	Beijing	Clustered
8	■□□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■■	400037777413771	3 (2.75)	2.5	EAI5	Clustered
17	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	677737607760771	1 (0.92)	0.16	LAM2	Unique
19	■■□■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■■■	677777477413771	1 (0.92)	0.13	EAI2-Manila	Unique
20	■■□■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	677777607760771	1 (0.92)	0.13	LAM1	Unique
42	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	777777607760771	3 (2.75)	0.1	LAM9	Clustered
47	■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■■	777777774020771	1 (0.92)	0.07	H1	Unique
50	■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■■	777777777720771	2 (1.83)	0.07	H3	Clustered
52	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	777777777760731	1 (0.92)	0.13	T2	Unique
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777777777760771	34 (31.2)	0.64	T1	Clustered
64	■■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■■□□□□■■■■■■■■	777777607560771	1 (0.92)	0.3	LAM6	Unique
93	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	777737607760771	1 (0.92)	0.3	LAM5	Unique
111	■■■■■■■■□□□■■■□■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	776167607760771	1 (0.92)	3.23	LAM3	Unique
118	■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777767777760771	4 (3.67)	2.88	T1	Clustered
175	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■□□□□■■■□■■■	777777677760731	1 (0.92)	4.76	T2	Unique
205	■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	737777777760771	1 (0.92)	2.27	T1	Unique
209	■■■■■■■■□□□□□□■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	776017607760771	1 (0.92)	2.5	LAM12-Madrid1	Unique
217	■■■■■■■■■■■■□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■■	777736777760771	1 (0.92)	3.13	T	Unique
232	■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□■■■■■	777777760000171	1 (0.92)	11.11	Unknown	Unique
281	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■■	777775777760771	1 (0.92)	4	T1	Unique
291	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■□□□□■■■■■■■■	777777677760771	1 (0.92)	1.22	T1	Unique
334	■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	577777777760771	3 (2.75)	3.45	T1	Clustered
344	■■■□□□□□□□□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	700077777760771	1 (0.92)	5.26	T1	Unique
787	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□■■■■	777777777760071	1 (0.92)	16.67	T1	Unique
866	■□■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	577777607760771	1 (0.92)	2.33	LAM9	Unique
888	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■□□■■■	777777777760631	1 (0.92)	7.14	T2	Unique
892	■■■■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■■■	777777477413771	1 (0.92)	20	EAI5	Unique
913	■■■■■■■■■■■■■□□□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777743777760771	1 (0.92)	5.26	T1	Unique

1053		777767637760771	1 (0.92)	25	T1	Unique
1060		777777776160771	1 (0.92)	33.33	T1	Unique
1394		777776777760671	1 (0.92)	12.5	X1	Unique
1679		777767737760771	1 (0.92)	14.29	T1	Unique
1751		700066777760771	1 (0.92)	25	X3	Unique
1916		777701777760771	1 (0.92)	14.29	T1	Unique
1952		777777774000771	1 (0.92)	5.88	Unknown	Unique
1979		777777777400771	1 (0.92)	33.33	Unknown	Unique
3277*		777777477513771	2 (1.83)	100	EAI2-Manila	Clustered
3278*		700076717760771	2 (1.83)	66.67	X3	Clustered
3279*		776177607760600	1 (0.92)	50	LAM3	Unique
3280*		777777770000031	1 (0.92)	50	Unknown	Unique

[a] A total of 36/40 SITs matched a preexisting shared-type in the database, whereas 4/40 SITs (*n*=6 isolates) were newly created either within the present study or after a match with an orphan in the database. A total of 9 SITs containing 56 isolates were clustered within this study (2–34 isolates per cluster) while 31 strains were unique. Note that SITs followed by an asterisk indicates "newly created shared-type" (*n*=4 containing 6 isolates) due to 2 or more strains belonging to an identical new pattern within this study or after a match with an orphan in the database as follows. SIT designations followed by number of strains: 3277*; this study *n*=2; 3278* this study *n*=2, ESP *n*=1; 3279* this study *n*=1, NLD *n*=1; 3280* this study *n*=1, NLD *n*=1.

[b] Clade designations according to SITVIT2 using revised SpolDB4 rules; "Unknown" designates patterns with signatures that do not belong to any of the major clades described in the database.

[c] Clustered strains correspond to a similar spoligotype pattern shared by 2 or more strains "within this study"; as opposed to unique strains harboring a spoligotype pattern that does not match with another strain from this study. Unique strains matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as "orphan" (see Table 1).

Table 3
Spoligotype description of clustered MDR *M. tuberculosis* from clinical isolates (*n*=56) in this study, their worldwide distribution in the SITVIT2 database, followed by a summary of their 12-loci MIRU based subclustering of spoligotyping-defined clusters.

SPOLIGOTYPE CLUSTERSSIT (Clade) Octal NumberSpoligotype Description[a]	Number (%) in study	% in study vs. database	Distribution in Regions with ⩾ 3% of a given SIT[b]	Distribution in countries with ⩾ 3% of a given SITs[c]	MIRU CLUSTERSNb of subclusters and MIT designations[d]
1 (Beijing) 000000000003771 □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	3 (2.75)	0.03	ASIA-E 34.21, AMER-N 21.82, ASIA-SE 9.85, AFRI-S 8.97, ASIA-N 6.71, ASIA-S 3.7	USA 21.46, CHN 19.87, JPN 12.47 ZAF 8.97, RUS 6.72, VNM 4.22	No Subcluster (orphan n=3)
8 (EAI5) 400037777413771 ■□□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■	3 (2.75)	2.5	AMER-N 26.67, AFRI-E 25.83, EURO-N 24.17, ASIA-W 6.67, AFRI-S 5.0	USA 24.17, DNK 13.33, MOZ 10.0, TZA 9.17, GBR 8.33, ZMB 6.67, ZAF 5.0, SAU 5.0	No Subcluster (orphan n=2, MIT157 n=1)
42 (LAM9) 777777607760771 ■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	3 (2.75)	0.1	AMER-S 29.44, AMER-N 15.35, EURO-S 11.56, AFRI-N 9.97, EURO-W 6.35, AFRI-E 3.92, AFRI-S 3.68, EURO-N 3.44	USA 13.79, BRA 11.81, MAR 8.19, COL 7.12, ITA 6.08, ESP 3.89, VEN 3.85, ZAF 3.68	No Subcluster (orphan n=2, MIT742 n=1)
50 (H3) 777777777720771 ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	2 (1.83)	0.07	AMER-N 20.34, AMER-S 17.91, EURO-W 14.07, EURO-S 12.78, EURO-E 5.92, AFRI-N 4.73, EURO-N 4.7, AFRI-S 4.5, CARIB 3.78, ASIA-W 3.1	USA 19.52, BRA 7.76, AUT 6.77, ITA 6.05, ESP 6.05, ZAF 4.5, PER 4.11, CZE 4.08, SWE 3.16	No Subcluster (orphan n=1, MIT1451 n=1)
53 (T1) 777777777760771 ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	34 (31.2)	0.64	AMER-N 18.76, AMER-S 14.5, EURO-W 11.5, EURO-S 10.65, ASIA-W 7.79, AFRI-S 5.83, EURO-N 5.79, AFRI-N 4.13, ASIA-E 3.72, AFRI-E 3.51, EURO-E 3.03	USA 15.49, ITA 6.26, BRA 5.89, ZAF 5.7, TUR 4.07, AUT 4.02, PER 3.12	1 Subcluster (MIT1450 n=2), Unclustered (orphan n=30, MIT45 n=1, MIT1449 n=1)
118 (T1) 777767777760771 ■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■	4 (3.67)	2.88	EURO-W 17.27, ASIA-W 15.83, AMER-S 15.11, ASIA-E 10.07, EURO-S 9.35, AMER-N 8.63, AFRI-E 5.76, EURO-E 5.04	FXX 15.11, TUR 13.67, ITA 7.19, VEN 6.48, USA 5.76, CHN 5.76, MOZ 5.04, JPN 4.32, CZE 4.32, BRA 4.32, ARG 3.6	No Subcluster (orphan n=3, MIT333 n=1)
334 (T1) 577777777760771 ■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	3 (2.75)	3.45	EURO-S 21.84, ASIA-E 19.54, AMER-N 13.79, EURO-W 6.9, EURO-N 6.9, AMER-S 6.9, AFRI-W 4.6, AFRI-S 4.6, CARIB 3.45, ASIA-W 3.45	CHN 18.39, ITA 14.94, USA 10.35, ESP 6.9, ZAF 4.6, MEX 3.45, GUF 3.45, GNB 3.45, GEO 3.45, BEL 3.45	No Subcluster (orphan n=3)
3277* (EAI2-Manila) 777777477513771 ■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■□■■□□■□■■■■■■■■■	2 (1.83)	100	AMER-N 100.0	MEX 100.0	No Subcluster (orphan n=2)
3278* (X3) 700076717760771 ■■■□□□□□□□□□■■■■■□■■■■□□■■■■■■■■■■□□□□■■■■■■■	2 (1.83)	66.67	AMER-N 66.67, EURO-S 33.33	MEX 66.67, ESP 33.33	No Subcluster (orphan n=2)

[a] Note that SITs followed by an asterisk indicates a "newly created shared-type" among clustered isolates within this study.

[b] Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database. The definition of macro-geographical regions and sub-regions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia.

[c] The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ⩾5% of a given SITs as compared to their total number in the SITVIT2 database.

[d] Only results for subclusters are shown, individual MIT designations for unclustered isolates are not shown (please refer to Supplemental Table 1 for detailed MIRU patterns of individual isolates including orphan spoligotype patterns).

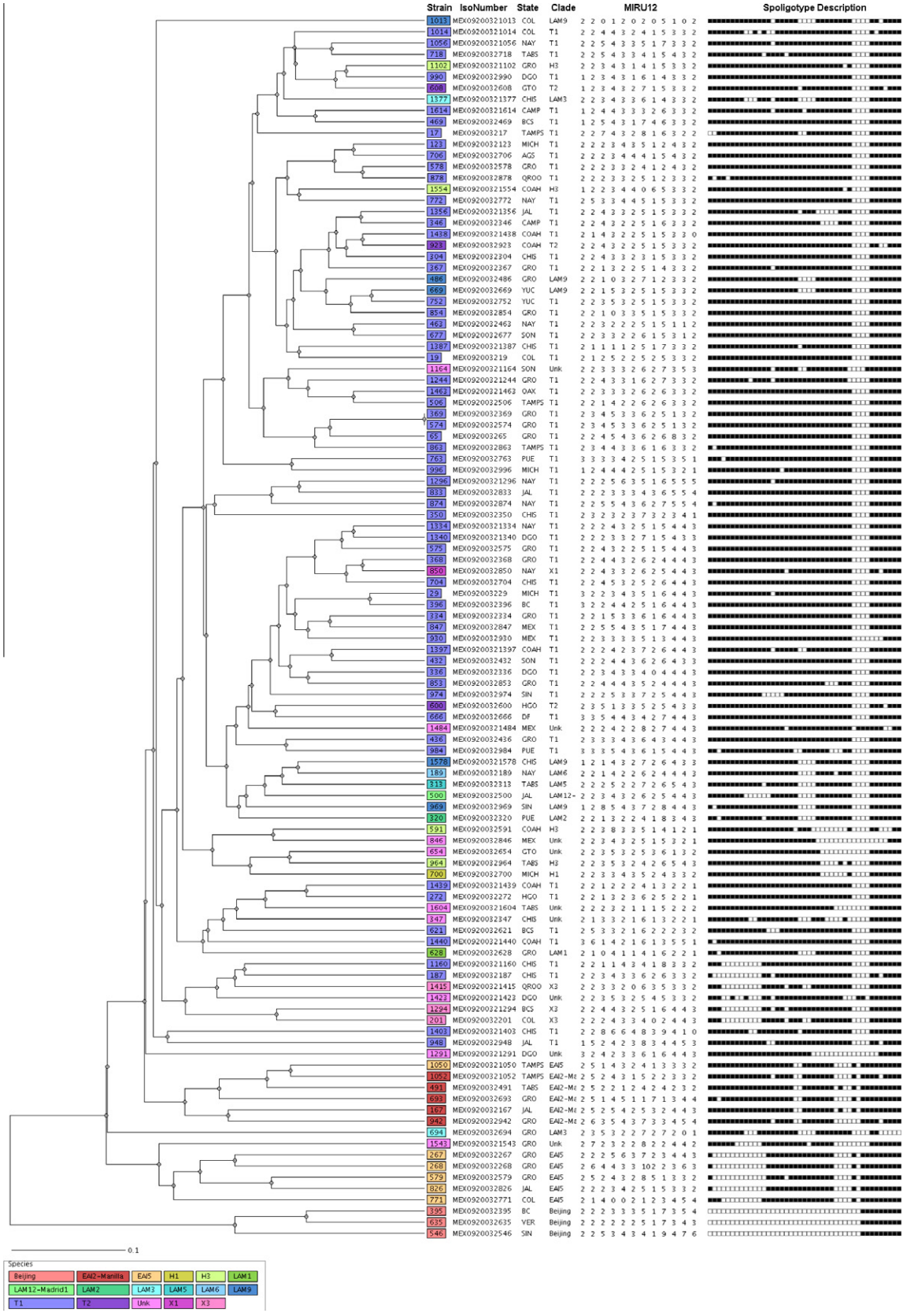


Fig. 2. MIRU and spoligotype data on clustering from 109 MDR-TB strains.

et al., 2000). Nevertheless, other studies have shown that SIT53 was more scarcely distributed in states along the Mexico-US border despite of including a significant number of Hispanic individuals (Chiang et al., 2009). SIT53 strains have also been identified in the US, Italy, Brazil, South Africa, Turkey, Australia and Peru (Table 3). The prevalence (17.85%) of this spoligotype

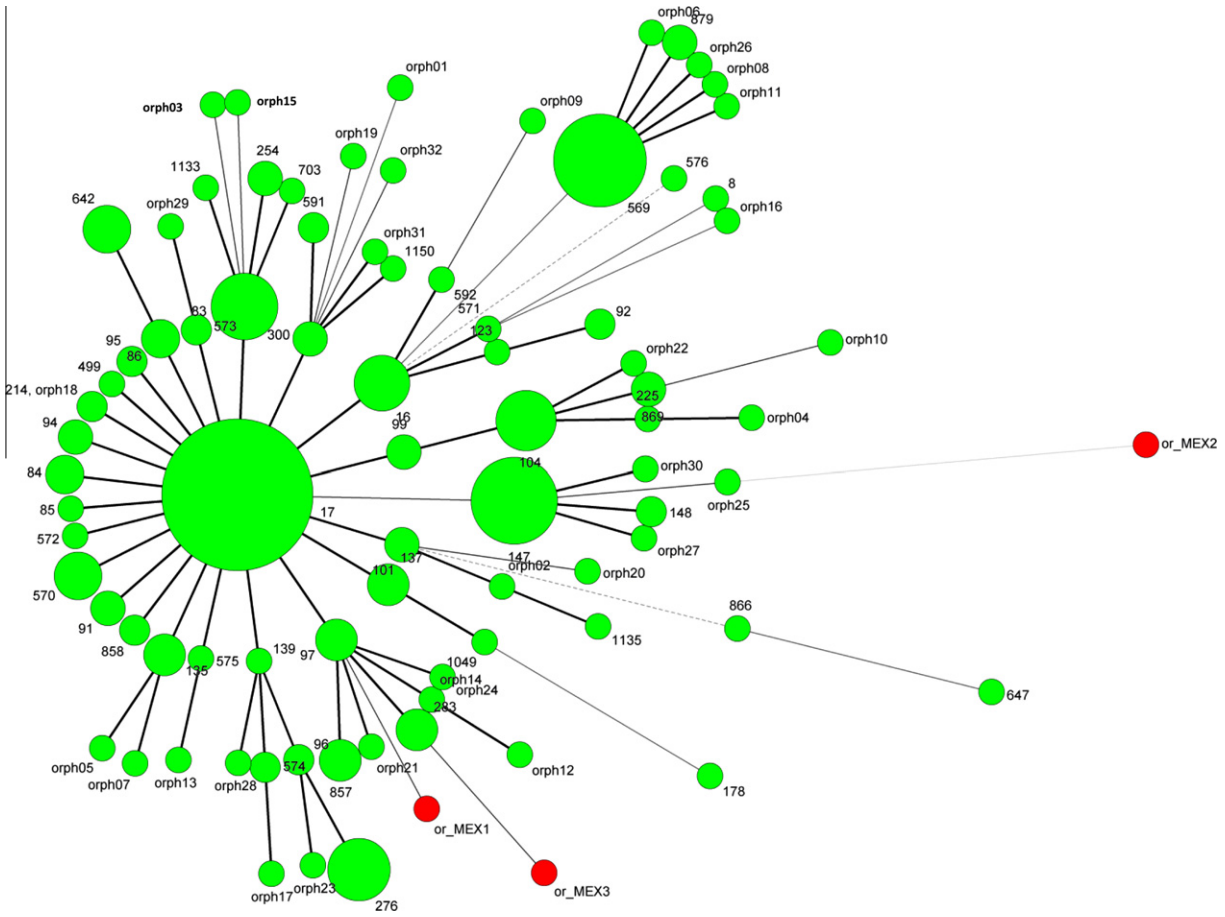


Fig. 3. A minimum spanning tree (MST) illustrating evolutionary relationships between MIRU12 patterns of the Beijing lineage strains in our study ($n = 3$; shown in red) versus all the Beijing lineage strains with MIRU12 data from USA ($n = 327$; shown in green) in the proprietary SITVIT2 database of Institut Pasteur de la Guadeloupe, consulted on December 10, 2012. The phylogenetic tree (drawn using BioNumerics version 6.6; Applied Maths NV, Sint-Maartens-Latem, Belgium) connects each MIRU12 pattern based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dashed and dotted lines) and circles representing each individual pattern. The length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed and gray dotted) denotes the number of allele changes between two patterns: solid lines, 1 to 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); gray dashed lines represent 4 changes; and gray dotted lines represent 5 or more changes. The size of the circle is proportional to the total number of isolates in our study, illustrating unique isolates (smaller nodes) versus clustered isolates (bigger nodes). The predominant MIRU12 patterns representing 2% or more in the US sample in the MST figure are as follows: MIT17 ($n = 82$, 25.1%), MIT569 ($n = 30$, 9.2%), MIT147 ($n = 26$, 8.0%), MIT83 ($n = 15$, 4.6%), MIT276 ($n = 13$, 4.0%), MIT104 ($n = 12$, 3.7%), MIT16 ($n = 10$, 3.1%), MIT642 ($n = 7$, 2.1%), MIT570 ($n = 7$, 2.1%). The 3 orphan patterns from Mexican study are labeled respectively or_MEX1 (IsoNumber MEX0920032395, MIRU pattern 223335173542); or_MEX2 (IsoNumber MEX0920032546, MIRU pattern 223434194765); and or_MEX3 (IsoNumber MEX0920032635, MIRU pattern 222225173432). As can be easily seen, the 3 Mexican orphan Beijing patterns are not directly related to the USA patterns.

worldwide shows the dominance of this type in different geographical region, including Europe and South Africa (Brudey et al., 2006; Garcia de Viedma et al., 2005).

Beijing strains are common in different geographical regions as well. However, in Mexico Beijing strains have been seldom identified (Martinez-Gamboa et al., 2008; Molina-Torres et al., 2010). Here, 3 Beijing strains (2.75%), isolated from three different states (Baja California, Sinaloa and Veracruz), were identified. Importantly, US bordering states have been shown to present a relatively high incidence of Beijing strains, accounting for 25% of the reported isolates (Soini et al., 2000). Interestingly, the results showed that the Beijing strains circulating in both countries are not genetically related and therefore do not share the same origin. Thus, further research is warranted to identify the origin of the Beijing strains circulating in Mexico.

Here, three TB strains belonging to the EAI2-Manila family, which are frequently found in Asia, were also identified (Douglas et al., 2003). Manila isolates have been previously reported circulating in Mexico (Martinez-Gamboa et al., 2008; Molina-Torres et al., 2010). It has been hypothesized that Manila strains were probably imported from the Philippines during the time when both countries were Spanish colonies (Molina-Torres et al., 2010), however, more research is required to correctly identify the origin of these strains.

As shown in Fig. 2, combinatory analysis using 12-loci MIRU-VNTR and spoligotyping provided sufficient resolution for the correct identification of TB strains in this study. Our approach was focused on the identification of straightforward "active" transmission among MDR TB cases. Thus, and if evidence of active transmission among primary MDR TB cases was observed, the use of 24-loci MIRU would have been required to further characterize the isolates to accurately establish the proportion of "real active transmission" among primary cases versus secondary MDR-TB cases between phylogenetically related strains.

The results showed that the TB lineages analyzed in this study exhibited a high degree of genetic diversity. This is not surprising considering the relatively small sample size in our study, which represent a small proportion of the circulating strains in Mexico. Our analysis assessed primarily the inter-state transmission since the number of specimens analyzed from each individual state was limited. This contributed at least to some degree to the high degree of genetic variability observed among geographically

distant strains. Importantly, the depth of our study was not enough to analyze the local transmission at the intra-state level, where most closely genetically related strains are likely to coexist.

Only two strains were phylogenetically related suggesting that the development of drug resistance among these two cases was most likely due to direct transmission between these individuals. More studies at the local level should be conducted to asses in more detail the transmission of MDR TB in the country. Additionally, new studies with more recent isolates should also be carried out to better understand the current situation of TB in Mexico.

In summary, this is the most comprehensive molecular study of MDR-TB carried out in Mexico. The molecular characterization of *M. tuberculosis* isolates is likely to improve our understanding on the molecular basis controlling TB transmission in the country. Our results highlight the importance of the implementation of a TB national molecular surveillance in Mexico.

Acknowledgements

David Couvin was awarded a Ph.D. fellowship by the European Social Funds through the Regional Council of Guadeloupe. This work was partially supported by CONACYT- SALUD-2010-C01-139760 and as part of the 7th Framework Program, European Commission, contract No. HEALTH-F3-2008-200999, and by grants CB-2010-156347, CONACYT and SIP20121271, IPN, México. JAG-y-M and SR-G are fellows of COFAA and EDI, IPN, México.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2012.12.024.

References

Banerjee, R., Allen, J., Westenhouse, J., Oh, P., Elms, W., Desmond, E., Nitta, A., Royce, S., Flood, J., 2008. Extensively drug-resistant tuberculosis in california, 1993–2006. Clin. Infect. Dis. 47, 450–457.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Canetti, G., Froman, S., Grosset, J., Hauduroy, P., Langerova, M., Mahler, H.T., Meissner, G., Mitchison, D.A., Sula, L., 1963. Mycobacteria: laboratory methods for testing drug sensitivity and resistance. Bull. World Health Organ. 29, 565–578.

Chiang, C.Y., Lee, J.J., Yu, M.C., Bai, K.J., Lin, T.P., Luh, K.T., 2009. Tuberculosis-related deaths without treatment. Int. J. Tuberc. Lung Dis. 13, 1563–1565.

Cox, H.S., Sibilia, K., Feuerriegel, S., Kalon, S., Polonsky, J., Khamraev, A.K., Rusch-Gerdes, S., Mills, C., Niemann, S., 2008. Emergence of extensive drug resistance during treatment for multidrug-resistant tuberculosis. N. Engl. J. Med. 359, 2398–2400.

Daley, C.L., 2005. Molecular epidemiology: a tool for understanding control of tuberculosis transmission. Clin. Chest Med. 26, 217–231, vi.

David, H., Levy-Frebault, V., Thorel, M.F., 1989. Méthodes de laboratoire pour Mycobactériologie clinique. Institut Pasteur, Paris, France.

Demay, C., Liens, B., Burguiere, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., Rastogi, N., 2012. SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect. Genet. Evol. 12, 755–766.

Douglas, J.T., Qian, L., Montoya, J.C., Musser, J.M., Van Embden, J.D., Van Soolingen, D., Kremer, K., 2003. Characterization of the Manila family of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 41, 2723–2726.

Dye, C., Scheele, S., Dolin, P., Pathania, V., Raviglione, M.C., 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA 282, 677–686.

Garcia-Garcia Mde, L., Ponce-De-Leon, A., Garcia-Sancho, M.C., Ferreyra-Reyes, L., Palacios-Martinez, M., Fuentes, J., Kato-Maeda, M., Bobadilla, M., Small, P., Sifuentes-Osornio, J., 2002. Tuberculosis-related deaths within a well-functioning DOTS control program. Emerg. Infect. Dis. 8, 1327–1333.

Garcia de Viedma, D., Bouza, E., Rastogi, N., Sola, C., 2005. Analysis of *Mycobacterium tuberculosis* genotypes in Madrid and identification of two new families specific to Spain-related settings. J. Clin. Microbiol. 43, 1797–1806.

Hunter, P.R., Gaston, M.A., 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 26, 2465–2466.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Kandula, N.R., Dworkin, M.S., Carroll, M.R., Lauderdale, D.S., 2004. Tuberculosis prevention in Mexican immigrants: limitations of short-course therapy. Am. J. Prev. Med. 26, 163–166.

Kent, P.T., Kubica, G.P., 1985. Public health Mycobacteriology a guide for the level III laboratory. US Department of Health and Human Services Centers for Disease Control, Atlanta, Georgia.

Lopez-Alvarez, R., Badillo-Lopez, C., Cerna-Cortes, J.F., Castillo-Ramirez, I., Rivera-Gutierrez, S., Helguera-Repetto, A.C., Aguilar, D., Hernandez-Pando, R., Samper, S., Gonzalez-y-Merchand, J.A., 2010. First insights into the genetic diversity of Mycobacterium tuberculosis isolates from HIV-infected Mexican patients and mutations causing multidrug resistance. BMC Microbiol. 10.

Martinez-Gamboa, A., Ponce-de-Leon, A., Galindo-Fraga, A., Bobadilla-del-Valle, M., Kato-Maeda, M., Robertson, B.D., Young, D.B., Small, P.M., Sifuentes-Osornio, J., 2008. Molecular analysis of *Mycobacterium tuberculosis* strains with an intact pks15/1 gene in a rural community of Mexico. Arch. Med. Res. 39, 809–814.

Martinez-Guarneros, A., Vaughan, G., Lozano, D., Vazquez-Chacon, C.A., Bäcker, C., Cazares, F., Ruiz-Tovar, K., Fonseca-Coronado, S., Carpio-Pedroza, J.C., Vergara-Castañeda, A., Escobar-Gutierrez, A., 2012. Multidrug-resistance (MDR) and extensively drug-resistance (XDR) tuberculosis in Mexico. Scand. J. Infect. Dis., Submitted.

Molina-Torres, C.A., Moreno-Torres, E., Ocampo-Candiani, J., Rendon, A., Blackwood, K., Kremer, K., Rastogi, N., Welsh, O., Vera-Cabrera, L., 2010. *Mycobacterium tuberculosis* spoligotypes in Monterrey, Mexico. J. Clin. Microbiol. 48, 448–455.

Nava-Aguilera, E., Lopez-Vidal, Y., Harris, E., Morales-Perez, A., Mitchell, S., Flores-Moreno, M., Villegas-Arrizon, A., Legorreta-Soberanis, J., Ledogar, R., Andersson, N., 2011. Clustering of *Mycobacterium tuberculosis* cases in acapulco: spoligotyping and risk factors. Clin. Dev. Immunol. 2011, 408375.

Quitugua, T.N., Seaworth, B.J., Weis, S.E., Taylor, J.P., Gillette, J.S., Rosas, I.I., Jost Jr, K.C., Magee Jr, D.M., Cox, R.A., 2002. Transmission of drug-resistant tuberculosis in Texas and Mexico. J. Clin. Microbiol. 40, 2716–2724.

Reed, M.B., Pichler, V.K., McIntosh, F., Mattia, A., Fallow, A., Masala, S., Domenech, P., Zwerling, A., Thibert, L., Menzies, D., Schwartzman, K., Behr, M.A., 2009. Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. J. Clin. Microbiol. 47, 1119–1128.

Secretaria de Salud, 2005. Boletines Epidemiologicos Anuales. Direccion General de Epidemiologia, Mexico City, Mexico.

Soini, H., Pan, X., Amin, A., Graviss, E.A., Siddiqui, A., Musser, J.M., 2000. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J. Clin. Microbiol. 38, 669–676.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., Locht, C., 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J. Clin. Microbiol. 39, 3563–3571.

WHO, 2010. Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response, in: WHO/HTM/TB/2010.3 (Ed.), World Helath Organization, 2010, Geneva.

Zenteno-Cuevas, R., Cuevas-Cordoba, B., Enciso, A., Enciso, L., Cuellar, A., 2012. Assessing the utility of three TaqMan probes for the diagnosis of tuberculosis and resistance to rifampin and isoniazid in Veracruz, Mexico. Can. J. Microbiol. 58, 318–325.

OPEN ACCESS Freely available online

PLOS ONE

Population Structure among *Mycobacterium tuberculosis* Isolates from Pulmonary Tuberculosis Patients in Colombia

Teresa Realpe[1,2,7], Nidia Correa[1,2,7], Juan Carlos Rozo[3,4,7], Beatriz Elena Ferro[3,7], Verónica Gomez[1], Elsa Zapata[1], Wellman Ribon[4,7,8], Gloria Puerto[4,7], Claudia Castro[4,7], Luisa María Nieto[3,7], Maria Lilia Diaz[5,7], Oriana Rivera[5,7], David Couvin[9], Nalin Rastogi[9], Maria Patricia Arbelaez[6,7], Jaime Robledo[1,2,7]*

1 Corporación para Investigaciones Biológicas, CIB, Medellín, Colombia, **2** Universidad Pontificia Bolivariana, Medellín, Colombia, **3** Centro Internacional de Entrenamiento e Investigaciones Médicas, CIDEIM, Cali, Colombia, **4** Instituto Nacional de Salud, Bogotá, Colombia, **5** Universidad del Cauca, Popayán, Colombia, **6** Universidad de Antioquia, Medellín, Colombia, **7** Centro Colombiano de Investigación en Tuberculosis, CCITB, Medellín, Colombia, **8** Universidad Industrial de Santander, Bucaramanga, Colombia, **9** WHO Supranational TB Reference Laboratory, TB & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Abymes Guadeloupe, France

Abstract

Background: Phylogeographic composition of *M. tuberculosis* populations reveals associations between lineages and human populations that might have implications for the development of strategies to control the disease. In Latin America, lineage 4 or the Euro-American, is predominant with considerable variations among and within countries. In Colombia, although few studies from specific localities have revealed differences in *M. tuberculosis* populations, there are still areas of the country where this information is lacking, as is a comparison of Colombian isolates with those from the rest of the world.

Principal Findings: A total of 414 *M. tuberculosis* isolates from adult pulmonary tuberculosis cases from three Colombian states were studied. Isolates were genotyped using *IS6110*-restriction fragment length polymorphism (RFLP), spoligotyping, and 24-locus Mycobacterial interspersed repetitive units variable number tandem repeats (MIRU-VNTRs). SIT42 (LAM9) and SIT62 (H1) represented 53.3% of isolates, followed by 8.21% SIT50 (H3), 5.07% SIT53 (T1), and 3.14% SIT727 (H1). Composite spoligotyping and 24-locus MIRU- VNTR minimum spanning tree analysis suggest a recent expansion of SIT42 and SIT62 evolved originally from SIT53 (T1). The proportion of Haarlem sublineage (44.3%) was significantly higher than that in neighboring countries. Associations were found between *M. tuberculosis* MDR and SIT45 (H1), as well as HIV-positive serology with SIT727 (H1) and SIT53 (T1).

Conclusions: This study showed the population structure of *M. tuberculosis* in several regions from Colombia with a dominance of the LAM and Haarlem sublineages, particularly in two major urban settings (Medellín and Cali). Dominant spoligotypes were LAM9 (SIT 42) and Haarlem (SIT62). The proportion of the Haarlem sublineage was higher in Colombia compared to that in neighboring countries, suggesting particular conditions of co-evolution with the corresponding human population that favor the success of this sublineage.

Citation: Realpe T, Correa N, Rozo JC, Ferro BE, Gomez V, et al. (2014) Population Structure among *Mycobacterium tuberculosis* Isolates from Pulmonary Tuberculosis Patients in Colombia. PLoS ONE 9(4): e93848. doi:10.1371/journal.pone.0093848

Editor: Axel Cloeckaert, Institut National de la Recherche Agronomique, France

Received November 19, 2013; **Accepted** March 8, 2014; **Published** April 18, 2014

Copyright: © 2014 Realpe et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by Colciencias grant No. 431–2004 for the Colombian Center for Tuberculosis Research, CCITB. Additional funding was provided by participating institutions: Corporación para Investigaciones Biológicas, Universidad de Antioquia, Centro Internacional de Investigaciones Médicas, Instituto Nacional de Salud and Universidad del Cauca. This work was also supported by the following Colombian Public Health Institutions: Dirección Seccional de Salud de Antioquia, Secretaría Departamental de Salud del Valle and Secretaría Departamental de Salud del Cauca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jrobledo@cib.org.co

Introduction

Tuberculosis (TB) continues to be a challenge to control. Although widespread and common efforts have had an impact in achieving declining numbers in global incidence for the first time in history, TB still causes 8.7 million new cases and 1.4 million deaths per year [1].

The worldwide population structure of *Mycobacterium tuberculosis* has been defined, linking specific lineages to human populations. Using comparative genomics and large sequence polymorphisms (LSPs), six phylogeographic lineages have been described and associated with human populations around the world [2]. For example, the East-Asian lineage is dominant in many countries of the Far East, while the Indo-Oceanic lineage occurs all around the Indian Ocean. The Euro-American lineage is clearly the most frequent lineage in Europe and the Americas. The relationships between these lineages and human populations are supported not only by studies with isolates from around the world, but also by the

tendency of each lineage to cause the disease in populations in specific urban cosmopolitan settings [3–6].

Genotyping techniques based on repetitive elements such as restriction fragment length polymorphism (RFLP) using *IS6110* [7], spoligotyping of clustered repetitive interspersed short palindromic repeats (CRISPR) [8], and mycobacterial interspersed repetitive units variable number tandem repeats (MIRU-VNTRs) [9], have been used to support epidemiological studies as well as to define the population structure in *M. tuberculosis* [10–12]. Spoligotyping and MIRU-VNTR have replaced the standard *IS6110*-RFLP due to the ease of implementation and standardization, and the availability of international databases for making comparisons [10,13,14]. These latter techniques have demonstrated a concordance with the assignation of major lineages as defined by LSPs [15–17], despite being subject to convergent evolution [18].

Colombia is the third most populated country in Latin America. The country has nearly 47 million inhabitants and has changed from being mainly a rural population at the beginning of the 20th century to a mostly urban population in the 21th century. The increase in urban population together with crowding and poor living conditions in the outskirts of major cities maintain favorable conditions for TB transmission. Despite the country's efforts to control the disease, the estimated incidence is 34 per 100,000 individuals in the population, corresponding to approximately 16,000 new cases per year [19], which poses a major task for the public health control of the disease. The overall incidence of TB for Colombia hides the dissimilarities between regions, reflecting differences in control measures as well as differences in transmission dynamics. These situations in turn, should influence the relationship established between human and *M. tuberculosis* populations.

Several studies have demonstrated the distribution of *M. tuberculosis* lineages and sublineages in Latin American countries, confirming the overwhelming predominance of the Euro-American lineage but with considerable variation in the distribution of sublineages or clades between and within countries [20–25]. In Colombia, studies performed on few specific locations also show a predominance of Euro-American lineages with differences among localities [26,27].

The aim of this study was to further assess the distribution of *M. tuberculosis* lineages and sublineages in Colombia and to gain a better understanding of the dynamics of the disease. *M. tuberculosis* isolates were obtained from patients with pulmonary tuberculosis from three different regions of Colombia. All isolates were genotyped by using *IS6110*-RFLP, spoligotyping, and 24-locus MIRU-VNTR. Then, associations between the main *M. tuberculosis* sublineages and the demographic and epidemiologic characteristics of patients were evaluated. The discriminatory power of the different genotyping methods was also calculated.

Methods

Ethics statement

All study procedures were approved by the Ethics Review Boards of the participating institutions who were in charge of recruiting the patients: Universidad de Antioquia, Centro Internacional de Entrenamiento e Investigaciones Médicas, CIDEIM, and Universidad del Cauca. All patients had a signed written consent previously approved by the ethics committee. When patients were less than 18 years old an informed written and signed consent was obtained with the additional approval and sign of one of the parents. All sign consents were kept in physical files locked under the custody of principal investigators to maintain the anonymity of patients. The study was also approved by regional and local health authorities in: Antioquia state and Medellin city, Valle del Cauca state and Cali city and Cauca state and Popayan city.

Study population

M. tuberculosis isolates were obtained from index tuberculosis patients belonging to three cohorts followed in three different cities in Colombia (Medellín, Cali, and Popayan) from March 2005 to 2008. These patients were part of a previous study performed in the same cities, were we studied factors associated with TB transmission among household contacts of patients with pulmonary tuberculosis [28].

Index cases were included consecutively from urban populations in cohorts from Medellin and Cali, whereas the smallest cohort included cases from Popayan as well as from smaller towns in Cauca state. An index case was included if the patient was older than 15 years and had at least one household contact as described previously [28]. Index cases were initially diagnosed based on clinical symptoms, signs, and chest-X rays, and confirmed by acid-fast bacilli (AFB) sputum examination using the Zielh-Nelsen stain, at the local health facility. A second sputum specimen was processed and cultured at the research laboratory designated in each city. Sputum samples were decontaminated with NaOH and N-acetyl-L-cysteine [29], cultured on an MGIT system (MGIT 960®) and/or solid Lowenstein-Jensen (LJ) culture media. Identification of AFB-positive cultures was performed by phenotypic methods such as niacin, nitrate and 68°C catalase tests [29]. Drug susceptibility testing for first line anti-TB drugs was performed using the proportion method in LJ [29]. *M. tuberculosis* isolates were frozen in 50% glycerol at −70°C until use. One isolate obtained from one AFB-positive smear sputum per patient was used for genotyping.

M. tuberculosis genotyping

Isolates were genotyped using spoligotyping, *IS6110*-RFLP, and 24-locus MIRU-VNTRs. For *IS6110*-RFLP genotyping, a standard protocol was used following international recommendations, which included a DNA extraction protocol [7,30]. Spoligotyping was performed following standard procedures [8], using a commercial source for membranes and reagents (Isogen Life Science, De Meern, the Netherlands).

MIRU-VNTR genotyping was performed using polymerase chain reaction (PCR) amplification of a standard set of 24 MIRU-VNTR loci with primers specific for the flanking regions of each VNTR region, and the detection of amplified PCR products was carried out by electrophoresis. From the gel images, the corresponding MIRU-VNTR bands were interpreted as copy numbers based on a reference table [9]. Two of the participating laboratories took part of the first and second multicenter proficiency studies of the Global Network for the Molecular Surveillance of Tuberculosis using MIRU-VNTR genotyping [31].

The role of participating laboratories was as follows: Mycobacteriology laboratory at University of Cauca in charge of culturing and identifying *M. tuberculosis* from patients in Popayan and surrounding towns (Cauca state). Mycobacteriology laboratory at Cideim was in charge of culturing and identifying *M. tuberculosis* from patients in Cali (Valle state) and performed 24-locus MIRU-VNTR in those isolates. Mycobacteriology laboratory at National Institute of Health in Bogota was in charge of performing drug susceptibility tests and genotyping by IS6110-RFLP, spoligotyping and 24-locus MIRU-VNTR to isolates from Cauca and Valle states. Mycobacteriology laboratory at CIB

performed culture and identification of isolates from patients from Medellin (Antioquia state) as well as genotyping using IS6110-RFLP, spoligotyping and 24-locus MIRU-VNTR.

Clustering analysis, allelic diversity, and discriminatory power

IS6110-RFLP, spoligotyping, and 24-locus MIRU-VNTRs results were analyzed using the BioNumerics software version 6.6 (Applied Maths. Sint-Martens-Latem, Belgium) to establish the relationships between different isolates of *M. tuberculosis*. Patterns of *IS6110*-RFLP were digitized and similarities were calculated using the Dice coefficient. MIRU-VNTR and spoligotyping data were entered as character type and analyzed using the categorical coefficient. Similarity trees and dendrograms were calculated using the unweighted pair group method with arithmetic averages (UPGMA). A cluster was defined as two or more *M. tuberculosis* isolates with identical patterns. The MIRU-VNTR allelic diversity (h) at a given locus was calculated as $h = 1 - \sum xi2\ [(n/n-1)]$, where xi is the frequency of the ith allele at the locus and n is the number of isolates [9,32]. To determine the discriminatory power (DP) of each genotyping method or a combination thereof, the Hunter-Gaston Discriminatory Index (HGDI) was calculated [33].

Minimum spanning trees (MST) were created using the Bionumerics software (Version 6.60) to explore the evolutionary relationship among spoligotyping and 24-locus MIRU-VNTR isolates. Spoligoforest trees were drawn to determine the parent-to-descendant spoligotypes in the group of isolates studied using the Fruchterman-Reingold algorithm and a hierarchical layout using the SpolTools software [http://www.emi.unsw.edu.au/spolTools, [34]], and reshaped and colored using the GraphViz software [http://www.graphviz.org].

Lineage assignation and comparison with an international database

Spoligotypes in binary format were converted to an octal code for comparison with the SITVIT2 proprietary database of Institut Pasteur de la Guadeloupe, which is an updated version of the previously released SpolDB4 and SITVITWEB databases [10,13]. At the time of the analysis, SITVIT2 contained genotyping information on about 110.000 *M. tuberculosis* clinical isolates from 160 countries of origin. In this database, a Spoligotype International Type (SIT) represents a spoligotyping pattern shared by 2 or more patient isolates, as opposed to "orphan," which does not match with another pattern in the SITVIT2 database. Major phylogenetic clades were assigned according to spoligotype signatures and using revised SpolDB4/SITVITWEB rules [10,13]. The sublineage distribution in cities from this study was also compared with those from two other cities of Colombia (Buenaventura and Bogotá), for which data were available in the SITVIT2 database. We also compared the distribution of the predominant SITs in the present study with the available data for 3 neighboring countries (Venezuela, Brazil, and Peru) in the SITVIT2 database.

Descriptive statistics were used to show the distribution of lineages and SITs per cohort of patients. STATA version 12 (STATA Corp. USA) was used for statistical analysis. Association of clades and SITs with demographic and epidemiological characteristics (human immunodeficiency virus [HIV] serology, sex), susceptibility to first-line drugs, and number of *IS6110*-RFLP copies, as well as differences in distribution according to cohorts, were calculated using Pearson's Chi-square test when more than 80% of the data had values greater than 5 and Fisher's Exact Test for the remaining data with smaller values (where at least 20% of data had values less than 5).

All study procedures and written consent forms were approved by the Ethics Review Boards of the participating institutions.

Results

Four hundred and fourteen *M. tuberculosis* isolates were studied from index cases included in three cohorts followed in three different Colombian cities for a period of three years (2005 to 2008). The median age of the patients was 39.1 years (range 15 to 96 years), and 42.8% of them were female. Bacillus Calmette-Guerin (BCG) vaccination was confirmed in 75.6% of patients and 1.8% were sero-positive for HIV. Most of the isolates (75.1%) were pan-susceptible to anti-TB drugs, 12.1% exhibited some drug resistance, and 4.6% were resistant to both isoniazid and rifampin (multi-drug resistant, MDR).

A total of 84 spoligotypes were identified; these included 20 orphan patterns that have not yet been reported to the SITVIT2 database (Table 1). The other 64 patterns matched a preexisting shared type in the database (50/64 SITs containing 374 isolates) or created a new shared type (14/64 SITs containing 20 isolates) within this study or after a match with a previously reported orphan in the SITVIT2 database (Table 2). Furthermore, 25 out of 64 pre-existing SITs containing 355 isolates were clustered (2 to 124 isolates per cluster), corresponding to 85.75% of all isolates. The number of unclustered isolates was 59 (39 isolates with unique SITs plus 20 orphan isolates) out of 414, or 14.25%.

SIT 42 (LAM9) with 124 isolates and SIT 62 (H1) with 97 isolates represented 29.9% and 23.4% of the total isolates, respectively (Table 2). These two SITs accounted for 3.78% and 17.7% of the isolates when compared with the total number of isolates in the SITVIT2 database, and together represented more than 10% of the isolates in South America, North America, and Southern Europe. In contrast, SIT207 (H3) with 8 isolates and SIT727 (H1) with 13 isolates represented 25.8% and 34.2% of the isolates in the SITVIT2 database; these SITs have been reported mostly in South America and North America (Table 3, see also table S3 for comparison of sublineages distribution with neighboring countries).

MSTs were constructed based on spoligotype patterns and 24-locus MIRU-VNTRs. Figure 1A shows MSTs based on spoligotypes in which two major groups were evident and belonged to the Haarlem and LAM sublineages, and included most of the isolates (SIT62 and SIT42, respectively, were the most frequent in these lineages). Other isolates were grouped as the ill-defined T sublineage (with SIT 53 as the most frequent) and X sublineage (with SIT91 as the most frequent). More distance was evident among isolates that integrate with the Haarlem sublineage than in those integrating with the LAM sublineage. In contrast, when MSTs were constructed using 24-locus MIRU-VNTR, isolates belonging to LAM appeared more distant than those integrating with the Haarlem sublineage. However, 24-locus MIRU-VNTRs better grouped isolates into major lineages such as LAM, Haarlem, S, T, and X; unique isolates belonging to the African sublineage and East African-Indian sublineage were clearly separated (Figure 1B). MSTs combining spoligotyping and MIRU-VNTR are shown in Figure 1C. There was agreement in the manner in which every genotyping method grouped isolates in the major sublineages. The 24-locus MIRU-VNTR analysis of common SITs (SIT42, SIT62, and SIT50) clearly showed that they are composed of very closely related isolates, which were distinguished by only one or two allele changes. Spoligoforest trees generated by means of the Fruchterman-Reingold algorithm and

Table 1. Orphan strains (n = 20) and corresponding spoligotyping defined lineages/sublineages recorded among a total of 414 *M. tuberculosis* strains from Colombian patients.

IsoNumber	Year	Strain	Spoligotype Description	Octal code	Lineage*	Label	MIRU24	24-MIT	Drug R**	Isolation city	Sex/Age	IS6110 RFLP	HIV
COL012005100005	2005	00005	■■■■■■■■□□■■■■□■■■■■■■■■□□□□□□□□□□■■■□■■■	776367770000731	Unknown	o01	2223251433224243523142233	o	1	Medellin	M/24	8	(-)
COL012005100028	2005	00028	■■□■■■■■■□■■■■□■■■■■□□□□■■■■■■■■□□□□■■□■■■■	677367607760671	LAM1	o02	2232261933212242624341 32	o	1	Medellin	F/40	9	(-)
COL012005300032	2005	00032	■■■■■■■□■□■■■□■■■■■■■■■■■■■■■■□□□□■■■■■■■■	775357777760771	T1	o03	2223251533251245222242 33	191	3 (INH, SM)	Medellin	F/83	7	(-)
COL012005100066	2005	00066	■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■□□□□■■■■■■■	777757777020771	H3	o04	2253251533233235632244 33	o	1	Medellin	F/26	8	(-)
COL012005100068	2005	00068	■■■■■□■■■■■■■■■■■■■■□□□□■□■■■■■■□□□□■■■■■■■	767777605760771	LAM9	o05	1243261533202242523341 32	o	1	Medellin	F/17	10	(-)
COL012005100071	2005	00071	■□■■□■■■■□■■■□□■■■■■■■■■■■■■■■■□□□□■■□■□■	557347777760661	T1	o06	2223251533251245222242 33	191	1	Medellin	M/41	6	(-)
COL012005100081	2005	00081	■■■■■■■■■□■■■□■■■■■■■■■■■■■■■■□□□□■□□□□■■■■■■■	777357777020771	H3	o07	2253251533233233731244 33	o	1	Medellin	M/35	8	(-)
COL012005000161	2005	00161	■■■■■■■■■□■■■□■■■■■■□□□□■■■■■■■■□□□□■■□■■■■	777357607760671	LAM9	o08	1233261532162242744341 32	o	0	Medellin	M/34	10	(-)
COL012005100178	2005	00178	■■■■■■■■■□■■■■□□□■□■■■■■■■■■■■■■□□□□■■□■■■■	777361377760671	T4	o09	2232251533245244533242 33	o	1	Medellin	M/46	9	(-)
COL012006100203a	2006	00203a	■■■□□□□□□□□□□■■■■□■■■■■■■□■■■■■■□□□□■■■■■■■	700036775760771	X3	o10	2243251533223660434433	o	1	Cali	F/23	2	(-)
COL012005000246b	2005	00246b	■■■■■■□□□■■■■■■■■■■■■■□■■■■■■■■■■■■□□■■■■	770777757777471	AFRI	o11	2262242435214545132434 34	o	0	Medellin	M/60	6	(-)
COL012006100302	2006	00302	■■■■■■■□□■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	774777607760771	LAM9	o12	1243261533202241622341 32	o	1	Cali	M/42	11	(-)
COL012006100313	2006	00313	■■■■■□■■■■■■■■■■■■■■■□■■■■■■□■□□□□■■■■■■■	767777767720771	H3	o13	1253251533233335632344 34	o	1	Medellin	F/22	8	(-)
COL012006100340	2006	00340	■■■■■■■■■■■■■■■■■□■■■■■■□□□□□□■□□□□■■■□■■■	777777374020731	H1	o14	2253251533232334742 44 24	197	1	Medellin	M/27	8	(-)
COL012006000348	2006	00348	■■■■■■■■■■■■□□□□■□□■■■■■■■■■■■□■□□□□■■■■■■■	777741177720771	H3	o15	2253131533233235732243 3	209	0	Medellin	M/41	11	(-)
COL012006100408	2006	00408	■■■□□■■■■■■□■■■■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	717737774020731	H1	o16	2253251533243334733442 3	204	1	Medellin	M/26	12	(-)
COL012006100422	2006	00422	■□■■■■■■□□■■■■■■■■■■■■■□□□□■■■■■■□□□□■■■□■■■	576377703740731	S	o17	2523251533244243522142 34	o	1	Medellin	F/45	13	(-)
COL012006000473b	2006	00473b	■■■■■■■■□□■■■■■■■■■■■■□□□□□■■■■■□□□□□■■■□■■■	776377701740731	S	o18	2223251633243245222242 33	226	0	Popayan	M/60	9	(-)
COL012006100486b	2006	00486b	■■□■■■■■■■■□■■■■■■■■□□□□□■■■■■■■■□□□□■■■■■■■	677737603760771	T3	o19	2222261532212244823342 33	o	1	Popayan	M/91	8	(-)
COL012007100524	2007	00524	□□□□□□□□□□□□□□□■■■■■■■■□□□□□□□□□□□□■■■□■■■	000003770000731	Unknown	o20	2323252533244243523142 34	o	1	El tambo	F/36	10	(-)

* Lineage designations for orphan patterns were done manually as Expert-based interpretations using revised SpolDB4 rules.

** Drug resistance (Drug R) information is shown as: 0, unknown; 1, pansusceptible; 2, MDRTB, i.e., combined resistance to INH-RIF (with or without resistance to other drugs); 3, any other resistance(s); 4, proven XDRTB.

doi:10.1371/journal.pone.0093848.t001

Table 2. Description of 64 shared-types (SITs; n = 394 isolates) and corresponding spoligotyping defined lineages/sublineages starting from a total of 414 *M. tuberculosis* strains isolated from Colombian patients.

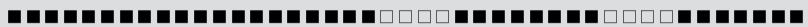
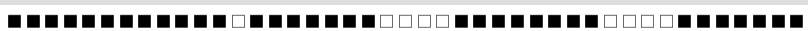
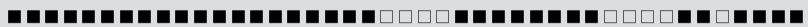
SIT*	Spoligotype Description	Octal Code	Nb in study	% in this study	% in study vs. Database	Lineage**	Clustered vs. Unique patterns***
2		000000004020771	2	0.48	0.51	H2	Clustered
3		000000007720771	1	0.24	0.95	H3	Unique
4		000000007760771	1	0.24	0.28	Unknown	Unique
8		400037777413771	1	0.24	0.55	EAI5	Unique
17		677737607760771	2	0.48	0.3	LAM2	Clustered
20		677777607760771	7	1.69	0.86	LAM1	Clustered
33		776177607760771	3	0.72	0.27	LAM3	Clustered
37		777737777760771	2	0.48	0.43	T3	Clustered
42		777777607760771	124	29.95	3.78	LAM9	Clustered
45		777777764020771	6	1.45	5.66	H1	Clustered
47		777777774020771	2	0.48	0.13	H1	Clustered
50		777777777720771	34	8.21	1.02	H3	Clustered
53		777777777760771	21	5.07	0.34	T1	Clustered
62		777777774020731	97	23.43	17.7	H1	Clustered
64		777777607560771	1	0.24	0.27	LAM6	Unique
86		777777737760771	1	0.24	1.09	T1	Unique
91		700036777760771	5	1.21	1.98	X3	Clustered
93		777737607760771	1	0.24	0.28	LAM5	Unique
102		777703777760771	1	0.24	1.37	T	Unique
106		776177400000171	1	0.24	0.76	Unknown	Unique
124		777777777700771	2	0.48	3.03	Unknown	Clustered
130		776177607760731	1	0.24	0.93	LAM3	Unique
162		777777607760671	1	0.24	3.23	LAM9	Unique
167		777777777660771	1	0.24	1.37	T1	Unique
207		767777777720771	8	1.93	25.81	H3	Clustered
230		777777603760771	1	0.24	5.56	LAM9	Unique
237		777777777700000	1	0.24	0.93	Unknown	Unique
254		777760007760771	1	0.24	0.57	T5-RUS1	Unique
316		777777770020731	1	0.24	2.13	H3	Unique
379		777741017760771	2	0.48	9.52	Unknown	Clustered
384		777777674020771	1	0.24	11.11	H1	Unique
447		777760377760771	1	0.24	12.5	T1	Unique
727		777737774020731	13	3.14	34.21	H1	Clustered

Table 2. Cont.

SIT*	Spoligotype Description	Octal Code	Nb in study	% in this study	% in study vs. Database	Lineage**	Clustered vs. Unique patterns***
753	■□□■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■	477777607760771	1	0.24	5	LAM1	Unique
881	■■■■■■■□□■■■■■■■■■■■■■□□□□□□□□□□□□■■□■■	776377770000731	6	1.45	24	Unknown	Clustered
1243	■■■■■■■■□■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■	777377777720771	1	0.24	6.25	H3	Unique
1277	■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■□□□□■■■■■■	777777207760771	1	0.24	3.85	LAM9	Unique
1283	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□□□□■■■■■■	777777777020771	1	0.24	11.11	H3	Unique
1486	□■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■□□□□■■■■■■	300036777760771	1	0.24	33.33	X3	Unique
1533	■■■■■■■■■■■■□□■■■■■■■■■■■■■□■□□□□■■■■■■	777763777720771	2	0.48	28.57	H3	Clustered
1536	■■■■■■■□□■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■	776377607760771	1	0.24	4.76	LAM9	Unique
1561	■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□□■■■■■■	777777774020371	1	0.24	20	H1	Unique
1804	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■□■□□□□■■■■■■	777777577720771	2	0.48	28.57	H3	Clustered
1832	■■■■■■■■■■■■■■□■■■■□□□□■■■■■■■□□□□■■■■■■	777773607760771	1	0.24	5	LAM9	Unique
2025	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■■■■■	777777777700371	1	0.24	9.09	T	Unique
2114	■■■■■■■■■■■■■□□□■■■■■■■■■■■■■□□□□■■■■■■	777761777760771	1	0.24	14.29	T1	Unique
2497	■■□■■■■■■■□■■■■■■■■□□□□■■■■■■■□□□□■■■■■■	677677607760771	1	0.24	16.67	LAM1	Unique
2865	■■■■■■■■■■■■■■■■■■■□□□□■■■□□■□□□□■■■■■■	777777607460771	1	0.24	25	LAM9	Unique
2940	■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■□□■■■■	777777607760471	2	0.48	28.57	LAM9	Clustered
3010	■■■■■■■□□■■■■■■■■■■■■■■■■■■□□□□□■■□■■■	776377777740731	3	0.72	33.33	S	Clustered
3018*	■■■■■■□□□□□□□□■■□■■■■■■■■■■■■□□□□■■■■■■	774006777760771	1	0.24	50	X1	Unique
3019*	■□■■■■■■■■■■□■■■■■■■□□□□■■■■■■■□□□□■■■■■■	577737607760771	1	0.24	25	LAM5	Unique
3020*	■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■□□□□■■■■■■	777777607360771	1	0.24	25	LAM9	Unique
3022*	■■■□□□□□□□□□■■■■■■■■■■■■□□□□□□■□□□□■■□■■	700077774020731	3	0.72	100	H1	Clustered
3023*	■■■□□■■■□□□■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■	716177607760771	1	0.24	50	LAM3	Unique
3039*	■■■■■■■■■■■■■□□□■□■■■■■■■■■■■□□□□■■■■■■	777761377760771	3	0.72	75	T4	Clustered
3040*	■■■■■■■■□■■■■□■■■■■□□□□■■■■■■■□□□□■■□■■■	777367607760671	1	0.24	50	LAM9	Unique
3041*	■■□■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■□□□□□	677777607760700	2	0.48	66.67	LAM	Clustered
3042*	■■■■■■■■■■■■■■■■■■□■■■□□□□□□■□□□□■■□■■■	777777674020731	1	0.24	50	H1	Unique
3043*	■■■□□□■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■	743777774020771	1	0.24	50	H1	Unique
3044*	■■■■■■■■■■■■■■■■■■■□□□□■■■■□□□□□□□□■■■■■	777777607700071	2	0.48	100	Unknown	Clustered
3045*	■■□■■■■■■■■□■■■■■■■□□□□□□□□□□□□□□□□□□□□	677737600000000	1	0.24	50	Unknown	Unique
3311*	□□□□□□■■□□□■■■■■■■■□□□□■■■■■■■□□□□■■■■■■	006177607760771	1	0.24	33.33	LAM3	Unique
3462*	■■■■■■■■■■■■□□□□□□■■■■■■■■■■■□□□□■■■■■■	777740377760771	1	0.24	50	T1	Unique

* newly created SITs due to 2 or more strains belonging to an identical new pattern within this study or after a match with an orphan in the database; SIT designations followed by number of strains: 3018* this study n = 1, USA n = 1; 3019* this study n = 1, BRA n = 2, SEN n = 1; 3020* this study n = 1, FXX n = 1, GNB n = 2; 3022* this study n = 3; 3023* this study n = 1, CUB n = 1; 3039* this study n = 3, COL n = 1; 3040* this study n = 1, ESP n = 1; 3041* this study n = 2, BRA n = 1; 3042* this study n = 1, VEN n = 1; 3043* this study n = 1, DEU n = 1; 3044* this study n = 2; 3045* this study n = 1, VEN n = 1; 3311* this study n = 1,COL n = 2; 3462* this study n = 1, FXX n = 1;

** Lineage designations according to SITVIT2 using revised SpolDB4 rules; "Unknown" designates patterns with signatures that do not belong to any of the major lineages described in the database.

*** Clustered strains correspond to a similar spoligotype pattern shared by 2 or more strains "within this study"; as opposed to unique strains harboring a spoligotype pattern that does not match with another strain from this study. Unique strains matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as "orphan" (see Table 1).

doi:10.1371/journal.pone.0093848.t002

Table 3. Description of clusters containing 5 or more isolates in this study, and their worldwide distribution in the SITVIT2 database.

SIT (Lineage) Octal Number Spoligotype Description	Octal Number	Number (%) in study	% in study vs. Database	Distribution in Regions with ≥3% of a given SIT *	Distribution in countries with ≥3% of a given SIT **
20 (LAM1) 677777607760771 ■■□■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	677777607760771	7 (1.69)	0.86	AMER-S 28.48, AMER-N 19.61, EURO-W 11.34, AFRI-S 10.97, EURO-S 9.74, CARI 5.3, AFRI-E 4.32	USA 19.61, BRA 18.13, NAM 7.65, PRT 6.04, FXX 6.04, VEN 5.18, ZAF 3.33, HTI 3.21, ESP 3.08
42 (LAM9) 777777607760771 ■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	777777607760771	124 (29.95)	3.78	AMER-S 29.77, AMER-N 12.1, EURO-S 11.55, EURO-W 9.66, AFRI-N 8.74, EURO-N 3.81, AMER-C 3.63, AFRI-E 3.63, AFRI-S 3.23	BRA 12.22, USA 12.1, COL 7.77, MAR 7.19, ITA 6.67, FXX 5.15, ESP 3.41, VEN 3.38, ZAF 3.23
45 (H1) 777777764020771 ■■■■■■■■■■■■■■■■■■■■■□■□□□□□□□■□□□□■■■■■■■	777777764020771	6 (1.45)	5.66	EURO-W 24.53, CARI 24.53, AMER-S 16.04, EURO-S 11.32, AMER-N 9.43, EURO-N 5.66	AUT 16.98, MTQ 13.21, USA 9.43, ITA 8.49, BRA 8.49, GLP 7.55, COL 7.55, DEU 3.77
50 (H3) 777777777720771 ■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	777777777720771	34 (8.21)	1.02	AMER-S 18.02, EURO-W 17.87, AMER-N 17.87, EURO-S 11.74, EURO-E 5.62, EURO-N 4.35, AFRI-N 4.32, AFRI-S 4.11, CARI 3.54, ASIA-W 3.27	USA 17.84, BRA 7.18, FXX 7.0, AUT 6.19, ITA 5.53, ESP 5.53, PER 4.81, ZAF 4.11, CZE 3.72
53 (T1) 777777777760771 ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777777777760771	21 (5.07)	0.34	EURO-W 16.1, AMER-N 13.88, AMER-S 12.39, EURO-S 9.68, ASIA-W 7.52, EURO-N 5.49, AFRI-S 5.11, AFRI-E 4.64, ASIA-E 4.36, AFRI-N 3.62, EURO-E 3.36, AMER-C 3.33	USA 13.58, FXX 8.11, ITA 5.49, BRA 5.28, ZAF 5.0, TUR 3.57, AUT 3.52, CHN 3.18
62 (H1) 777777774020731 ■■■■■■■■■■■■■■■■■■■■■■□□□□□□□■□□□□■■■□■■■	777777774020731	97 (23.43)	17.7	AMER-S 36.5, EURO-W 16.06, EURO-S 12.41, AMER-N 12.41, AFRI-E 3.65, AFRI-W 3.29	COL 34.12, USA 12.41, FXX 12.04, ITA 7.48, ESP 3.83, GMB 3.1
91 (X3) 700036777760771 ■■■□□□□□□□□□■■■■□■■■■■■■■■■■■■□□□□■■■■■■■	700036777760771	5 (1.21)	1.98	AMER-N 49.8, AMER-S 22.13, CARI 13.83, EURO-S 6.32, EURO-N 4.35	USA 46.64, HTI 11.46, PER 9.88, ESP 6.32, COL 4.74, GUF 3.95, CAN 3.16
207 (H3) 767777777720771 ■■■■■□■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	767777777720771	8 (1.93)	25.81	AMER-S 38.71, AMER-N 29.03, AMER-C 16.13, EURO-W 9.68, CARI 3.23, ASIA-W 3.23,	USA 29.03, COL 29.03, MEX 16.13, BRA 6.45, AUT 6.45, SAU 3.23, PER 3.23, MTQ 3.23, FXX 3.23,
727 (H1) 777737774020731 ■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	777737774020731	13 (3.14)	34.21	AMER-S 68.42, AMER-N 18.42, AFRI-W 5.26	COL 63.16, USA 18.42, NGA 5.26
881 (Unknown) 776377770000731 ■■■■■■■■□□■■■■■■■■■■■■■■□□□□□□□□□□□□■■■□■■■	776377770000731	6 (1.45)	24	AMER-S 76.0, AMER-N 16.0, EURO-S 8.0,	COL 72.0, USA 16.0, VEN 4.0, ITA 4.0, ESP 4.0,

* Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database. The definition of macro-geographical regions and sub-regions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia.

** The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ≥3% of a given SITs as compared to their total number in the SITVIT2 database.

doi:10.1371/journal.pone.0093848.t003

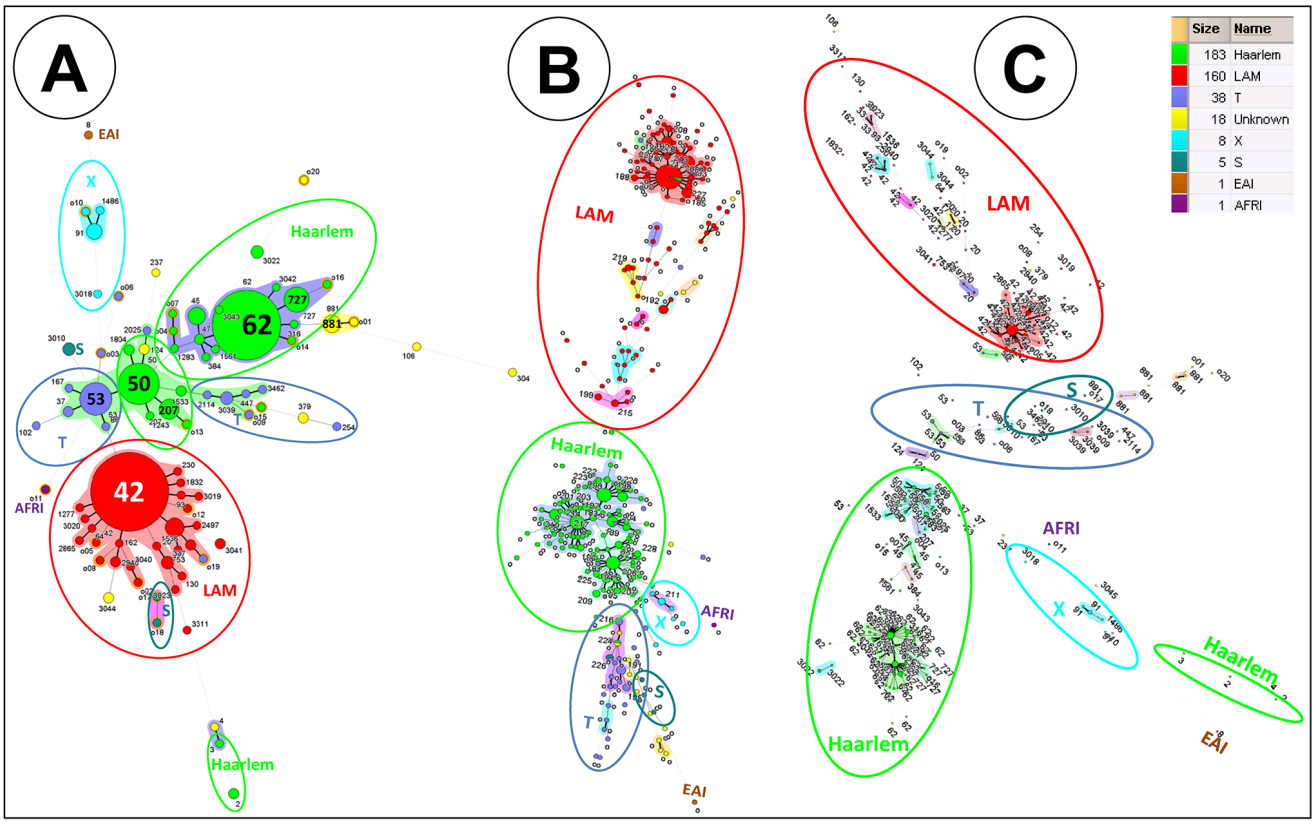


Figure 1. Minimum spanning tree (MST) illustrating evolutionary relationships between *M. tuberculosis* spoligotypes identified in our study (**A**). MST constructed with spoligotyping. (B) MST constructed with 24-locus MIRU-VNTR (**C**) Composite MST with spoligotyping and MIRU-VNTRs markers. MST were constructed on all isolates (n = 414, including 20 orphan patterns). The phylogenetic tree connects each genotype based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dashed and dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines, 1 or 2 or 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); gray dashed lines represent 4 changes; and gray dotted lines represent 5 or more changes. The size of the circle is proportional to the total number of isolates in our study, illustrating unique isolates (smaller nodes) versus clustered isolates (bigger nodes). The color of the circles indicates the phylogenetic lineage to which the specific pattern belongs. Note that orphan patterns are circled in orange. Patterns colored in yellow indicate a strain with an unknown signature (unclassified).
doi:10.1371/journal.pone.0093848.g001

hierarchical layout (Figure S2 A and B) confirmed the dominance of SIT42 (LAM) and SIT62 (Haarlem). The SIT42 (LAM) cluster was the largest node evolved from SIT53 (T1), from which multiple spoligotypes arose. The second largest spoligotype SIT62 (H1) appears to derive originally from SIT53 (T1) and more recently from SIT50 (H3), finally giving rise to a lower amount of SITs.

An evolutionary MST based on spoligotypes as a function of several associated characteristics showed significant differences between predominant SITs (>2%) and drug resistance (unknown not included) ($p<0.001$) (Figure S1). It is worth noting that all strains belonging to SIT45 (H1) were MDR (6 out of 6). The difference between predominant SITs (>2%) and the three higher *IS6110*-RFLP copy number (8, 9, and 11 bands) was also significant ($p = 0.011$). Significant differences were found between predominant SITs and HIV-positive serology ($p = 0.044$). The proportion of HIV-positive patients was greater among isolates belonging to SIT727 (H1) (2 HIV-positive out of 13) and SIT53 (T1) (2 HIV-positive out of 19)(Figure S1). Unknown HIV-status was not included in the analysis. No significant differences were noted when comparing the sex ratios of all predominant SITs ($p>0.5$).

The phylogeographical distribution of *M. tuberculosis* lineages around Colombia is shown in Figure 2A. Our data showed that LAM represented 39.6% of the isolates from Medellín (Antioquia State), 39.1% of the isolates from Cali (Valle state) and 24.0% of the isolates from several towns in Cauca state. The Haarlem sublineage was found to comprise 48.7% of the group of isolates from Medellín and 39.0% of the group of isolates from Cali, but was not represented in isolates from Cauca. In contrast, the ill-defined T sublineage makes up 40.0% of the isolates from Cauca, including the town of Popayan (the main city), as compared to the proportion of isolates from this sublineage in Medellín (6.4%) and Cali (11.0%). No isolates belonging to Beijing sublineage were identified in our study. Genotyping data available in the SITVIT2 database from other two cities in Colombia (Buenaventura and Bogotá), showed a predominance of isolates belonging to the LAM and Haarlem sublineages. Other lineages such as T, X, and S are less represented in these two cities with the exception of the Beijing sublineage which was frequent in the city of Buenaventura.

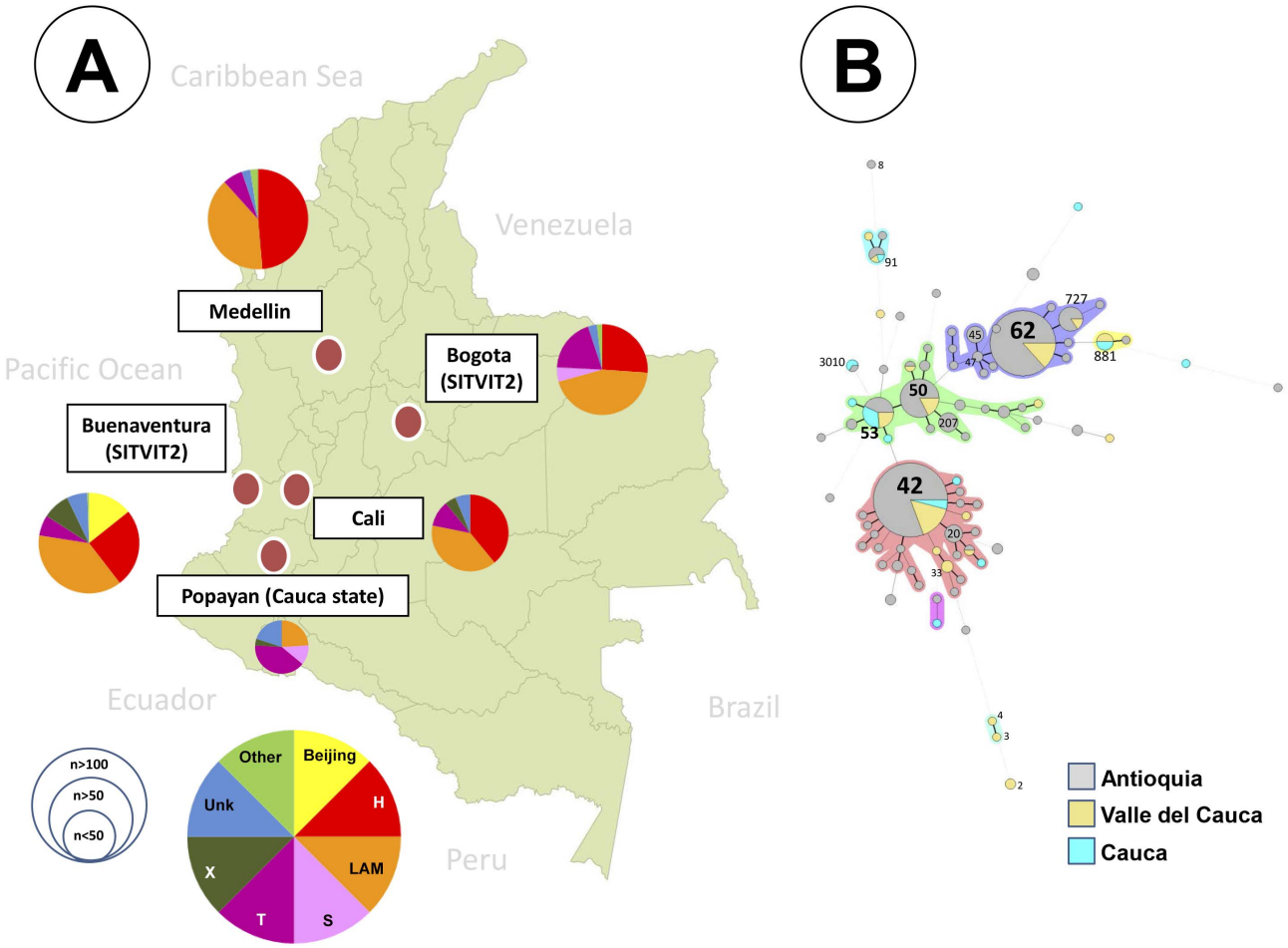


Figure 2. Phylogeographical distribution of *M. tuberculosis* sublineages identified in our study and MST according to patients' origin grouped by cities or states. (**A**) The Map shows the cities of our study and others with their corresponding pie representing the proportion of *M. tuberculosis* sublineages; Distribution of sublineages among strains belonging to the 3 sites of study (**Antioquia**, **Valle del Cauca** and **Cauca**) vs. strains contained in international database SITVIT2 for cities Bogota and Buenaventura. (**B**) MST illustrating evolutionary relationships between *M. tuberculosis* spoligotypes in our study in function of states. The cities have been located into their corresponding states: Medellin is located in **Antioquia**; Cali is located in **Valle del Cauca**; and Popayan, Caldono, Morales, El Tambo, and Piendamo are located in **Cauca state**. The map was obtained from http://www.uxabilidad.com/recursos/mapa-politico-de-colombia-envectores.html which is available as a public domain.
doi:10.1371/journal.pone.0093848.g002

An MST based on spoligotypes and the state in which the isolates were obtained (Figure 2B) revealed a close evolutionary relationship with the main spoligotypes found in this study, which were LAM (SIT42) and Haarlem sublineages (SIT62) mostly in the states of Antioquia and Valle. In contrast, isolates from Cauca state were more distantly related even among the most frequently identified, the ill-defined T sublineage. The only exception was SIT53 (T1), which was found in greater proportion in Cauca state than in other areas of the country. Furthermore, the difference in the sublineage distribution of isolates from the three states of the country reported in this study was significant ($p<0.001$). The analysis based on 24-locus MIRU-VNTR showed that 52.3% of isolates from Medellin were grouped into 40 clusters (2 to 29 isolates per cluster), 9.4% of isolates from Cali were grouped into 3 clusters (2 isolates each) and 8% of isolates from Popayán and surrounding towns were grouped in one cluster (2 isolates).

Comparative DP was calculated for the three genotyping methods used in this study. The method with the highest DP (0.9916) was 24-locus MIRU-VNTR, followed by *IS6110*-RFLP (0.9868) and then spoligotyping (0.8414). The DP obtained using the combination of the three genotyping methods was slightly higher than that observed for 24-locus MIRU-VNTRs (0.9918 vs. 0.9916). We also evaluated different combinations of MIRU-VNTRs and calculated their corresponding DP. Eight-locus MIRU-VNTR with a allelic diversity greater than 0.6 showed a discriminatory power of 0.9771, while 15-locus MIRU-VNTR with the highest allelic diversity, showed a DP of 0.9855 slightly above of the recommended set of 15-locus MIRU-VNTRs [35], that showed a discriminatory power of 0.9847 (see Table S1). The allelic diversity for the different MIRU-VNTR loci was evaluated using Hunter-Gaston diversity analysis. Locus QUB11b showed the greatest allelic diversity with a diversity index of 0.780 (CI 0.767–0.793), whereas locus 20 showed the lowest diversity index 0.033 (CI 0.009–0.056) (see Table S2).

Discussion

This study presents a phylogeographic panorama of the *M. tuberculosis* population structure in Colombia. The isolates analyzed

from three states showed the LAM and Haarlem clades as being dominant, grouping 82.8% of them mostly in urban settings (Medellín and Cali cities). Other studies carried out in Colombia, based on spoligotyping and deposited in the SITVIT2 database, showed the same predominance of the Haarlem and LAM sublineages. One of these studies was in Bogotá, the major urban setting of Colombia, in which the proportion of the LAM, Haarlem, and T clades were 49.3%, 25.0%, and 13.8% respectively [26]. This study also reported SIT42 (LAM9), SIT62 (H1), and SIT53 (T1) as the major clusters comprising 45.8% of isolates. By contrast, our present study showed the same SITs (42, 62, 53) comprising 58.4% of isolates, of which SIT42 (LAM9) and SIT62 (H1) accounted for 53% of isolates.

The LAM, Haarlem, ill-defined T, X, and S sublineages belong to the Euro-American lineage or lineage 4, one of the six major lineages for *M. tuberculosis* that have been described around the world [2,36,37]. This lineage, although present in several regions of the world, is predominant in Europe and America. In Latin America, this lineage is dominant and have been reported by several studies with considerable variation among countries. In our study 37.4% of isolates belonged to LAM. This sublineage appeared to be most common in Brazil (46%), Venezuela (53%), and Peru (28.3%). In the three countries that share borders with Colombia, the proportion of Haarlem is quite variable; our data shows a proportion of 41% among the studied isolates, in contrast to those seen in Venezuela (5%), Brazil (12%), and Peru (28%) [20–23,25].

When comparing the main SITs found in our study with their frequencies in neighboring countries, there is a significant difference in the proportion of SIT42 (LAM9): 29.9% for our study versus 11.8% for Venezuela, 8.8% for Brazil, and 5.6% for Peru (based on the SITVIT2 database). The differences are more striking in the case of SIT62 (H1), which is one of the two most endemic in Colombia, when compared to the same neighboring countries, with 23.4% of isolates belonging to this SIT versus 0.53%, 0.02%, and 0.08% for Venezuela, Brazil, and Peru, respectively (SITVIT2 database) (see Table S3).

Contrary to the sublineage distribution observed in isolates from the main urban settings, those obtained from Cauca state were grouped predominantly (40%) in the ill-defined T sublineage, with no isolates belonging to Haarlem. The clear difference among the distribution of sublineages in Cauca state compared to that in the urban settings of Valle and Antioquia (Cali and Medellín) might be explained by the smaller group of isolates studied, and by the human origin of these isolates. Most of the cases from Cauca state were from patients living in smaller urban and rural areas located in the south of the country, which is characterized by a higher proportion of indigenous population. These facts suggest differences in transmission conditions as well as host factors that ultimately may affect the successful establishment of a particular *M. tuberculosis* lineage in a given human population.

Analysis of *M. tuberculosis* isolates from Buenaventura city (Valle state), a seaport in the South Pacific coast of Colombia, have identified isolates belonging to the LAM and Haarlem sublineages, but also isolates belonging to the Beijing sublineage (SITVIT2 database). This was an unusual finding compared to our study, in which no Beijing isolates were identified. This sublineage, although described for the first time in 1998 in this Colombian seaport city, has only been reported since then from patients whose origin is from this same city, or patients with the same origin but diagnosed in the country's inland major urban settings [38–40]. In agreement with these data, despite the human migration from Asia, where the Beijing isolates are very frequent, they represent a proportion about 5% or less of isolates in Latin America, according to several reports [20–25].

There was significant association between MDR-TB and HIV status with particular spoligotypes. For example, six isolates belonging to SIT45 (H1) were MDR. Analysis of these isolates using 24-locus MIRU-VNTRs revealed that they were very closely related, but only grouped four of them into two clusters. This finding may represent a particular transmission focus, because all were isolates from patients in Medellín, rather than showing a particular predisposing trend of this spoligotype to develop as MDR. No clear association has been found in terms of predominant lineages or sublineages and MDR among different studies in Latin America [25,41,42]. Despite this, the Beijing sublineage has been associated with a high proportion of drug resistant isolates in several parts of the world [43,44], including Colombia [27]. Moreover, the M strain (Haarlem 2) has been linked to large MDR-TB outbreaks in Argentina [45].

MST analysis provided a detailed picture of genetic distances among *M. tuberculosis* isolates studied based on spoligotyping and MIRUs. Using both genotyping methods facilitated a better assignment of the major and dominant groups belonging to the Euro-American sublineages LAM and Haarlem (lineage 4). This was in agreement with previous studies that demonstrated the utility of this approach in assigning clades and sublineages, particularly within the Euro-American lineage [15]. The hierarchical layout and Fruchterman-Reingold analysis based on spoligotyping interestingly showed that SIT42 (LAM9) and SIT62 (H1), the more conspicuous SITs found in this study, were derived initially from SIT 53 (T1) and lately evolved to the more dominant type. The reason behind the expansion of these particular SITs in the studied isolates and populations, over co-existing non-dominant SITs, might suggest a conjunction of social changes such as accelerated population growth in impoverished sub-urban settings facilitating the transmission of the disease, with mostly still unknown microbe characteristics that allowed the adaptation of particular *M. tuberculosis* isolates to specific human populations. An example of successful *M. tuberculosis* isolates in a particular population was published recently, linking the success of some of them to phenotypic characteristics such as slower growth in monocytes and the ability to elicit a less inflammatory response [46].

A more detailed look at the spoligotyping and 24-locus MIRU-VNTRs composite MST for SIT42 (LAM9) and SIT62 (H1) showed a MIRU-VNTR multiplicity of clusters: 12 clusters in SIT62 and 15 clusters in SIT42, along with unique isolates. Most isolates belonging to these two spoligotypes were very closely related, since they were differentiated by one or two MIRU-VNTR allele changes, supporting the notion that recent expansion and evolution of these groups of isolates have occurred in accordance with the rate of mutation calculated for MIRU-VNTR [17]. The analysis based on 24-locus MIRU-VNTR showed that there was a greater percentage of clustering in isolates from Medellin (52.3%) than in Cali (9.4%) and Popayan and surrounding towns (8.0%). This suggests a more active and ongoing transmission in Medellín (the largest group of isolates studied) than in the other two areas.

A practical utility of the major discriminatory power of the 24-locus MIRU-VNTR set over spoligotyping and *IS6110*-RFLP is its use as an epidemiological marker to distinguish between a diversity of isolates, including those associated with specific transmission chains. This is particularly useful in settings with high endemicity and disease transmission, as observed in one of urban settings studied. Supporting the epidemiological use of MIRU-VNTR in the population studied is the finding that most of the clusters

identified by this method were circumscribed to one of the two urban centers studied. In addition, we found that MIRU23, ETRB, and Mtub34, which are excluded from the recommended 15-locus MIRU-VNTR for epidemiological studies [35], had allelic diversity index values higher than 0.5. This finding might lead us to consider the use of these loci in different genotyping studies, especially in areas with *M. tuberculosis* lineage distribution similar to that observed in this study.

In summary, this study shows the distribution of *M. tuberculosis* lineages and sublineages in several regions in Colombia, with an important dominance of LAM and Haarlem belonging to lineage 4, particularly in two major urban settings (Medellín and Cali). Two dominant spoligotypes were LAM9 (SIT42) and Haarlem1 (SIT62). The use of 24-locus MIRU-VNTR showed the best discriminatory power and proved useful in epidemiological studies in which the Euro-American lineage is prevalent. The proportion of the Haarlem sublineage was higher in Colombia compared to that in neighboring countries, suggesting the presence of particular conditions of co-evolution with the corresponding human population that favor the success of this sublineage.

Supporting Information

Figure S1 A minimum spanning tree (MST) illustrating evolutionary relationships between the *M. tuberculosis* spoligotypes in our study in function of studied parameters. (**A**) Drug resistance; (**B**) *IS6110*-RFLP; (**C**) HIV Serology; (**D**) Sex ratio.Difference between predominant SITs (>2%) including SIT45 vs. Drug resistance (Code 0 Unknown not included) is very significant ($p<0.001$); note that all strains belonging to SIT45/H1 are MDR. The difference between predominant SITs>2% and the 3 Major *IS6110* RFLP No of Bands (8, 9 and 11) is significant (with a p-value = 0.011). The difference between predominant SITs and HIV serology is significant (p = 0.044), note that the proportion of HIV positive patients is more visible among strains belonging to SIT727/H1 (number of HIV positive = 2/13) and SIT53/T1 (n = 2/19). Missing HIV status values have not been taken into account. No significance difference was observed when comparing sex ratios of all predominant SITs (p value>0.5).
(TIF)

Figure S2 A representation of parent to descendant spoligotypes within our study sample (n = 414 isolates) as seen through Spoligoforest trees drawn using the SpolTools software (available through http://www.emi.unsw.edu.au/spolTools; Reyes et al. 2008), and reshaped and colored using the GraphViz software (http://www.graphviz.org; J. Ellson et al. 2002). (**A**) Tree drawn using Fruchterman Reingold algorithm (**B**) tree drawn using a Hierarchical Layout. In both trees, each spoligotype pattern from the study is represented by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. The heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weigh being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5. Note that in both trees, SIT42/LAM9 is the biggest node (n = 124, 29.95%), followed by SIT62/H1 (n = 97, 23.43%), SIT50/H3 (n = 34, 8.21%), SIT53/T1 (n = 21, 5.07%) and SIT727/H1 (n = 13, 3.14%), which are other predominant patterns in our study. On the other hand, orphan isolates (double circled), appear mostly at terminal positions on the tree, or are isolated strains without interconnections with the other strains.
(TIF)

Table S1 Comparative discriminatory power of three genotyping methods used in 414 *M. tuberculosis* isolates from Colombia.
(DOCX)

Table S2 Allelic diversity using Hunter-Gaston Diversity index for 24 MIRU-VNTR loci genotyping in 414 *M. tuberculosis* isolates from Colombia.
(DOCX)

Table S3 Distribution of the proportion of predominant SITs in study as compared to their distribution in neighboring countries Venezuela (n = 935), Brazil (n = 4556), and Peru (n = 1296), recorded in the SITVIT2 database.
(DOC)

Author Contributions

Conceived and designed the experiments: JR NC BF WR TR. Performed the experiments: EZ VG LMN OR CC GP JCR NC TR. Analyzed the data: JR MPA DC NR. Contributed reagents/materials/analysis tools: MLD NR JR MPA. Wrote the paper: JR NR DC BF JCR LMN NC TR.

References

1. WHO (2012) Global Tuberculosis Report 2012. Geneva, Switzerland: World Health Organization (WHO).
2. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, et al. (2006) Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 103: 2869–2873.
3. Baker L, Brown T, Maiden MC, Drobniewski F (2004) Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. Emerg Infect Dis 10: 1568–1577.
4. Gagneux S, Small PM (2007) Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. The Lancet Infectious Diseases 7: 328–337.
5. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM (2004) Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci U S A 101: 4871–4876.
6. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, et al. (2009) Major Mycobacterium tuberculosis lineages associate with patient country of origin. J Clin Microbiol 47: 1119–1128.
7. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, et al. (1993) Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 31: 406–409.
8. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, et al. (1997) Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 35: 907–914.
9. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, et al. (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol 44: 4498–4510.
10. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, et al. (2006) Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6: 23.
11. Ghosh S, Moonan PK, Cowan L, Grant J, Kammerer S, et al. (2012) Tuberculosis genotyping information management system: enhancing tuberculosis surveillance in the United States. Infect Genet Evol 12: 782–788.
12. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN (2006) Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev 19: 658–685.
13. Demay C, Liens B, Burguiere T, Hill V, Couvin D, et al. (2012) SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol 12: 755–766.
14. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D (2010) MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. Nucleic Acids Res 38: W326–331.

15. Cardoso Oelemann M, Gomes HM, Willery E, Possuelo L, Batista Lima KV, et al. (2011) The forest behind the tree: phylogenetic exploration of a dominant Mycobacterium tuberculosis strain lineage from a high tuberculosis burden country. PLoS One 6: e18256.
16. Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, et al. (2011) Strain classification of Mycobacterium tuberculosis: congruence between large sequence polymorphisms and spoligotypes. Int J Tuberc Lung Dis 15: 131–133.
17. Wirth T, Hildebrand F, Allix-Beguec C, Wolbeling F, Kubica T, et al. (2008) Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog 4: e1000160.
18. Comas I, Homolka S, Niemann S, Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One 4: e7815.
19. WHO website available: https://extranet.who.int/sree/Reports?op = Replet&name = %2FWHO_HQ_Reports%2FG2%2FPROD%2FEXT%2FTBCountryProfile&ISO2 = CO&LAN = EN&outtype = html. (Accesed 2013 september 10).
20. Abadia E, Sequera M, Ortega D, Mendez MV, Escalona A, et al. (2009) Mycobacterium tuberculosis ecology in Venezuela: epidemiologic correlates of common spoligotypes and a large clonal cluster defined by MIRU-VNTR-24. BMC Infect Dis 9: 122.
21. Candia N, Lopez B, Zozio T, Carrivale M, Diaz C, et al. (2007) First insight into Mycobacterium tuberculosis genetic diversity in Paraguay. BMC Microbiol 7: 75.
22. Gomes HM, Elias AR, Oelemann MA, Pereira MA, Montes FF, et al. (2012) Spoligotypes of Mycobacterium tuberculosis complex isolates from patients residents of 11 states of Brazil. Infect Genet Evol 12: 649–656.
23. Miranda SS, Carvalho Wda S, Suffys PN, Kritski AL, Oliveira M, et al. (2011) Spoligotyping of clinical Mycobacterium tuberculosis isolates from the state of Minas Gerais, Brazil. Mem Inst Oswaldo Cruz 106: 267–273.
24. Ritacco V, Lopez B, Cafrune PI, Ferrazoli L, Suffys PN, et al. (2008) Mycobacterium tuberculosis strains of the Beijing genotype are rarely observed in tuberculosis patients in South America. Mem Inst Oswaldo Cruz 103: 489–492.
25. Sheen P, Couvin D, Grandjean L, Zimic M, Dominguez M, et al. (2013) Genetic diversity of Mycobacterium tuberculosis in Peru and exploration of phylogenetic associations with drug resistance. PLoS One 8: e65873.
26. Cerezo I, Jimenez Y, Hernandez J, Zozio T, Murcia MI, et al. (2012) A first insight on the population structure of Mycobacterium tuberculosis complex as studied by spoligotyping and MIRU-VNTRs in Bogota, Colombia. Infect Genet Evol 12: 657–663.
27. Ferro BE, Nieto LM, Rozo JC, Forero L, van Soolingen D (2011) Multidrug-resistant Mycobacterium tuberculosis, Southwestern Colombia. Emerg Infect Dis 17: 1259–1262.
28. del Corral H, Paris SC, Marin ND, Marin DM, Lopez L, et al. (2009) IFNgamma response to Mycobacterium tuberculosis, risk of infection and disease in household contacts of tuberculosis patients in Colombia. PLoS One 4: e8257.
29. Kent B, Kubica G (1985) Public Health Mycobacteriology. A guide for the level III laboratory. Atlanta, GA: Center for Disease Control (CDC).
30. van Soolingen D, de Haas PE, Kremer K (2001) Restriction fragment length polymorphism typing of mycobacteria. Methods Mol Med 54: 165–203.
31. de Beer JL, Kremer K, Kodmon C, Supply P, van Soolingen D, et al. (2012) First worldwide proficiency study on variable-number tandem-repeat typing of Mycobacterium tuberculosis complex strains. J Clin Microbiol 50: 662–669.
32. Ferdinand S, Valetudie G, Sola C, Rastogi N (2004) Data mining of Mycobacterium tuberculosis complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. Res Microbiol 155: 647–654.
33. Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 26: 2465–2466.
34. Reyes JF, Francis AR, Tanaka MM (2008) Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 9: 496.
35. Oelemann MC, Diel R, Vatin V, Haas W, Rusch-Gerdes S, et al. (2007) Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J Clin Microbiol 45: 691–697.
36. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, et al. (2010) Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet 42: 498–503.
37. Coscolla M, Gagneux S (2010) Does M. tuberculosis genomic diversity explain disease diversity? Drug Discov Today Dis Mech 7: e43–e59.
38. Laserson KF, Osorio L, Sheppard JD, Hernandez H, Benitez AM, et al. (2000) Clinical and programmatic mismanagement rather than community outbreak as the cause of chronic, drug-resistant tuberculosis in Buenaventura, Colombia, 1998. Int J Tuberc Lung Dis 4: 673–683.
39. Murcia MI, Manotas M, Jimenez YJ, Hernandez J, Cortes MI, et al. (2010) First case of multidrug-resistant tuberculosis caused by a rare "Beijing-like" genotype of Mycobacterium tuberculosis in Bogota, Colombia. Infect Genet Evol 10: 678–681.
40. Nieto LM, Ferro BE, Villegas SL, Mehaffy C, Forero L, et al. (2012) Characterization of extensively drug-resistant tuberculosis cases from Valle del Cauca, Colombia. J Clin Microbiol 50: 4185–4187.
41. Gonzalo X, Ambroggi M, Cordova E, Brown T, Poggi S, et al. (2011) Molecular epidemiology of Mycobacterium tuberculosis, Buenos Aires, Argentina. Emerg Infect Dis 17: 528–531.
42. Imperiale BR, Zumarraga MJ, Di Giulio AB, Cataldi AA, Morcillo NS (2013) Molecular and phenotypic characterisation of Mycobacterium tuberculosis resistant to anti-tuberculosis drugs. Int J Tuberc Lung Dis 17: 1088–1093.
43. de Steenwinkel JE, ten Kate MT, de Knegt GJ, Kremer K, Aarnoutse RE, et al. (2012) Drug susceptibility of Mycobacterium tuberculosis Beijing genotype and association with MDR TB. Emerg Infect Dis 18: 660–663.
44. European Concerted Action on New Generation Genetic M, Techniques for the E, Control of T (2006) Beijing/W genotype Mycobacterium tuberculosis and drug resistance. Emerg Infect Dis 12: 736–743.
45. Ritacco V, Lopez B, Ambroggi M, Palmero D, Salvadores B, et al. (2012) HIV infection and geographically bound transmission of drug-resistant tuberculosis, Argentina. Emerg Infect Dis 18: 1802–1810.
46. Mathema B, Kurepina N, Yang G, Shashkina E, Manca C, et al. (2012) Epidemiologic consequences of microvariation in Mycobacterium tuberculosis. J Infect Dis 205: 964–974.

OPEN ACCESS Freely available online

PLOS ONE

Genetic Diversity of *Mycobacterium tuberculosis* in Peru and Exploration of Phylogenetic Associations with Drug Resistance

Patricia Sheen[1], David Couvin[2], Louis Grandjean[1,3,4], Mirko Zimic[1], Maria Dominguez[1], Giannina Luna[1], Robert H. Gilman[1,4,5], Nalin Rastogi[2], David A. J. Moore[1,3,4,5]*

1 Laboratorio de Enfermedades Infecciosas, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru, **2** WHO Supranational TB Reference Laboratory, TB and Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Guadeloupe, France, **3** LSHTM TB Centre and Department of Clinical Research, London School of Hygiene and Tropical Medicine, London, United Kingdom, **4** Department of Infectious Diseases and Immunity, Imperial College, London, United Kingdom, **5** Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America

Abstract

Background: There is limited available data on the strain diversity of *M tuberculosis* in Peru, though there may be interesting lessons to learn from a setting where multidrug resistant TB has emerged as a major problem despite an apparently well-functioning DOTS control programme.

Methods: Spoligotyping was undertaken on 794 strains of M tuberculosis collected between 1999 and 2005 from 553 community-based patients and 241 hospital-based HIV co-infected patients with pulmonary tuberculosis in Lima, Peru. Phylogenetic and epidemiologic analyses permitted identification of clusters and exploration of spoligotype associations with drug resistance.

Results: Mean patient age was 31.9 years, 63% were male and 30.4% were known to be HIV+. Rifampicin mono-resistance, isoniazid mono-resistance and multidrug resistance (MDR) were identified in 4.7%, 8.7% and 17.3% of strains respectively. Of 794 strains from 794 patients there were 149 different spoligotypes. Of these there were 27 strains (3.4%) with novel, unique orphan spoligotypes. 498 strains (62.7%) were clustered in the nine most common spoligotypes: 16.4% SIT 50 (clade H3), 12.3% SIT 53 (clade T1), 8.3% SIT 33 (LAM3), 7.4% SIT 42 (LAM9), 5.5% SIT 1 (Beijing), 3.9% SIT 47 (H1), 3.0% SIT 222 (clade unknown), 3.0% SIT1355 (LAM), and 2.8% SIT 92 (X3). Amongst HIV-negative community-based TB patients no associations were seen between drug resistance and specific spoligotypes; in contrast HIV-associated MDRTB, but not isoniazid or rifampicin mono-resistance, was associated with SIT42 and SIT53 strains.

Conclusion: Two spoligotypes were associated with MDR particularly amongst patients with HIV. The MDR-HIV association was significantly reduced after controlling for SIT42 and SIT53 status; residual confounding may explain the remaining apparent association. These data are suggestive of a prolonged, clonal, hospital-based outbreak of MDR disease amongst HIV patients but do not support a hypothesis of strain-specific propensity for the acquisition of resistance-conferring mutations.

Citation: Sheen P, Couvin D, Grandjean L, Zimic M, Dominguez M, et al. (2013) Genetic Diversity of *Mycobacterium tuberculosis* in Peru and Exploration of Phylogenetic Associations with Drug Resistance. PLoS ONE 8(6): e65873. doi:10.1371/journal.pone.0065873

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received December 20, 2012; **Accepted** April 29, 2013; **Published** June 24, 2013

Copyright: © 2013 Sheen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: DAJM was supported by Wellcome Trust Fellowships (064672/Z/01/Z and 078067/Z/05/Z; www.wellcome.ac.uk). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.moore@lshtm.ac.uk

Introduction

Molecular fingerprinting of *M. tuberculosis* (MTB) permits investigation of the epidemiology of tuberculosis to a previously unattainable level of detail, revealing insights into the differential transmission success of strains whilst observation and analysis of this epidemiology can generate testable hypotheses about strain biology [1]. There is limited data on the epidemiology and strain diversity of *M tuberculosis* in Peru [2], though there may be interesting lessons to learn from a setting where multidrug resistant TB has emerged as a major problem despite an apparently well-functioning DOTS control programme [3]. Here we report the results of an exercise to spoligotype all the strains of a large bank of well-characterized MTB strains derived from research projects conducted in Lima, Peru between 1999 and 2005, conduct spoligotype-based phylogenetic analyses and explore phylogenetic associations with HIV infection and drug resistance.

Methods

Strain bank

The sampling framework for this study was opportunistic, making use of a strain bank of anonymised but phenotypically well characterized strains of *M tuberculosis* collected in the course of four

clinical research studies conducted amongst adults with pulmonary tuberculosis in hospital and community-based studies in Lima, Peru. All studies had been reviewed and approved by the Institutional Review Board of Universidad Peruana Cayetano Heredia (Lima, Peru). Spoligotyping was undertaken on 794 strains of *M tuberculosis* (1 per patient) which came from (1) an unselected community-based cohort from south Lima (Feb 1999–May 2002) (n = 329), (2) a hospital-based cohort from an HIV unit (May 1999–Feb 2005) (n = 241), (3) an unselected community-based cohort from north Lima (June 2003–July 2004) (n = 155), (4) a selected community-based cohort from east Lima of TB patients reporting conventional risk factors for drug-resistant TB (Nov 2003–Oct 2005) (n = 69). Recruitment for each of these studies has been reported previously [4,5,6]. Strains were stored at −70°C in the Laboratorio de Investigación de Enfermedades Infecciosas of Universidad Peruana Cayetano Heredia (UPCH) in Lima. Available patient data was limited to gender, age and HIV status; strain data included first line drug susceptibility profile and name of source study with date of collection of original clinical sample.

Spoligotyping was undertaken at UPCH in batches over several months and films were interpreted by two independent readers; for the rare occasions where there was lack of independent agreement and subsequent failure to resolve discrepancies between both readers spoligotyping was repeated and the new film was used. Phylogenetic analyses and the construction of phylogenetic trees and spoligoforests (drawn using the SpolTools software available through http://www.emi.unsw.edu.au/spolTools; [7,8]) permitted identification of clusters and orphan strains by comparison with the SITVIT2 database (Institut Pasteur de Guadeloupe). The minimum spanning tree (MST), based on spoligotyping patterns, was drawn using BioNumerics software version 3.5. The MST is an undirected connected graph which links all the strains together with the fewest possible linkages between nearest neighbours. Contrarily to the MST, the spoligoforest trees are directed graphs which only evolve by loss of spacers. In these trees, nodes are not necessarily all connected (indeed, in case of too many changes between two strains, there are no edges linking them. In combination with SpolTools software, GraphViz software (http://www.graphviz.org) [9] was used to colour the orphan strains on the spoligoforest trees. Strains were categorized into spoligotype international types (SIT) and clades for the purpose of reporting strain diversity within the strain bank. In a univariate analysis odds ratios were computed for associations between strain groupings (by clade and by SIT separately) and patient gender, HIV status, strain year of origin, isoniazid mono-resistance, rifampicin mono-resistance and multidrug resistance (with each compared to drug susceptible reference group); in subsequent multivariate logistic regression only those clade or SIT associations with a p value$<$0.1 on univariate analysis were included in the model.

Results

Study population

Mean patient age was 31.9 years (range 15–78, no gender difference), 63% were male and 30.4% were known to be HIV+. HIV infection was significantly more frequent amongst males (OR 3.00, 95% CI 2.1–4.3). Rifampicin mono-resistance, isoniazid mono-resistance and multidrug resistance (MDR) were identified in 4.7%, 8.7% and 17.3% of tested strains respectively (12 rifampicin and 20 isoniazid results were unavailable).

Strain diversity

Of 794 strains from 794 patients there were 149 different spoligotypes identified. Of these there were 27 strains with novel, unique orphan spoligotypes, 718 which mapped to spoligotypes already described at least twice, and 49 which were newly created shared types either within the present study or after a match with an orphan in the database. Descriptions of the orphan strain spoligotypes and the demographics of the source patients are given in table 1. In table 2 the 122 already known spoligotype international types (SITs) and corresponding lineages detected in this strain set are shown along with the frequency that each occurred in this strain set as compared with the comparison global database. (The complete original dataset of 794 spoligotypes with accompanying clinical data is made available in table S1; the comparative frequency of the predominant SITs in this study with those reported elsewhere in Latin America is shown in table S2).

498 strains (62.7% of all 794) were clustered in the nine most common spoligotypes; 16.4% SIT 50 (clade H3), 12.3% SIT 53 (clade T1), 8.3% SIT 33 (LAM3), 7.4% SIT 42 (LAM9), 5.5% SIT 1 (Beijing), 3.9% SIT 47 (H1), 3.0% SIT 222 (clade unknown), 3.0% SIT1355 (LAM), and 2.8% SIT 92 (X3) (table 3).

The phylogenetic relationships between strains are illustrated in minimum spanning trees (Figure 1) which demonstrate that PGG2/3 (H, LAM, T, X, S) strains are highly predominant (representing 83.8% of all strains), most notably H (n = 228, 28.7%), LAM (n = 225, 28.3%) and T (n = 161, 20.3%). (figure S1 includes SIT numbers which can be seen by zooming in on pdf file). The spoligoforests shown in Figure 2 (and figure S2) highlight (regardless of layout technique) that SIT50 (H3) is the largest node (n = 130), followed by SIT53 (T1, n = 98), SIT33 (LAM3, n = 66), SIT42 (LAM9, n = 59) and SIT1 (Beijing, n = 44).

Strain clade associations

There was no predominant spoligotype associated with MDR amongst TB patients without HIV co-infection (table 4) - the odds of MDR were highest in those with disease caused by the LAM9 spoligotype SIT 42 though this was not statistically significant. Amongst patients with HIV co-infection this spoligotype was associated with by far the highest odds of MDR (87.5% of HIV patients with SIT42 disease had MDR); the T1 spoligotype SIT 53 was also associated with a increased odds of MDR, though only amongst patients with HIV (60.0% of HIV patients with SIT53 disease had MDR compared to 14.0% of HIV uninfected patients).

HIV infection was strongly associated with MDRTB in this analysis. This association was significantly reduced (though incompletely mitigated) after adjustment for confounders, an effect largely mediated by inclusion of SIT42 and SIT53 (Table 5). Indeed even after adjustment for HIV and other covariates SIT42 and SIT53 were independently associated with MDRTB, though not with either isoniazid or rifampicin resistance. On univariate analysis male gender was associated with MDRTB but this effect was entirely driven by the increased HIV prevalence in males and disappeared after adjustment in the multivariate model. Neither study site nor year of sample collection were associated with drug resistance in the multivariate model.

Discussion

In this report of strain diversity from Peru covering the period 1999–2005 3.4% of spoligotypes observed were from novel, orphan strains. The nine most frequently observed spoligotypes (out of 149 observed) accounted for over 60% of all disease and the eight of these also featured amongst the nine most frequently

Table 1. Description of spoligotypes with corresponding spoligotyping defined lineages/sublineages and demographic information for 27 orphan strains identified from amongst 794 strains of *M. tuberculosis* isolated from adults with pulmonary tuberculosis in Lima, Peru.

Year	Strain	Spoligotype Description	Octal code	Lineage *	Drug-R **	Sex/Age	HIV Serology
2003	0075	■□□□□□■■■■■■■■■■■■□□□□■■□□□□□□□□□□□□■■■	407777606000031	Unknown	1	M/48	
2003	0095	■■■■■■■□□□■■■■■■■■□□□□□□□□■□■■■□□□□■■■■■■■	776177600560771	T1	1	F/27	
2003	0112	■□□■■■■■■■■■■■■■■■■□□■■■■■■■■■□□□■■■□■■■	477777717760731	T2	1	M/50	
2003	0127	■■■■■■■■■■■□■■□■■■■■■■■■■■■■■■□□□□■■■□■■■	777733777760731	T	1	F/53	
2003	0129	■■□□■■■■□□□■■■■■■■■□□□□□□□□□□□□□□□□□□■■■■■	636177400000171	Unknown	1	F/78	
2004	0165	■■■□■■■■■■■■■■■■■■□□□□■■■■□■■■□□□■■■□■■■	737777607560731	LAM6	1	F/43	
2004	0168	■■■□□□□□□□□□■■■■□■■■■■■■■■■■■□□□□□■■■■■■■	700036777740771	X3	2	M/30	
2004	0173	■■■■■■■■□■■□□■■■■■■■■■■■■■■■■■□□□□■■■■■■■	776637777760771	T1	1	M/50	
2004	0178	■■■■■■■□□□■■■■■■■■■■■■■■■■□□■■■■□□□□■■■□■■	774377776360751	T1	1	M/18	
2004	0186	■■■■■■■■■■■■■■■■■■■□■■□■■■■■■■□□□□□□■■■■■■■	777777557700771	Unknown	1	F/39	
1999	0228	■■□□□□□■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	603777607760771	LAM1	1	M/26	HIV+
1999	0239	■■■■■■■□□□■■□□□□■■■□□□□■■■■■■■□□□□■■■□■■■	776141607760731	LAM3	1	M/26	HIV+
2000	0279	■■■■■■■■■■■■■■■■■■■□□□□■■□■□■■■□□□□■■■□■■■	777777606560731	LAM6	1	F/22	HIV+
2000	0290	□■■■■■■■■■■■■■■■■■■□□□□■■□□□□□□□□□□□□□■■■	377777606000031	Unknown	1	F/29	HIV+
2001	0325	■■■■■■■■■■■■■■■■■■■■■□■□□□□□■□□□□□■■■■■■■	777777775020371	H3	1	M/40	HIV+
1999	0350	■■□■■■□□□□□□□□□□□□□□□□□□□□□□■■□□□□□■■■■■■■	670000000160771	T1	1	M/24	
1999	0367	■■■■■■■■■■■■■■■■■□■■■■■■■□■■■■■■□□□□■■■□□□□□	777776775760700	X1	1	M/0	
2000	0478	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■■□■■■	777777777700731	Unknown	1	F/0	
2000	0498	■■■■■■■□□■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	774777777760731	T2	1	M/0	
2000	0519	■■■■■■■■■■■■■■■■■□■■■■■■■□■■■■■□□□□■■□□□□□□	777776775760600	X2	1	F/0	
2000	0570	■■■□□□■□□□□□□□□□■■□■■■□□■■■■■■■■□□□□■□□□□□□	704003347760400	T4	1	M/0	
2001	0593	■■■■■■■■■■■■■□□□■■■■□■■■■■■■■■■■□□□□■■■■■■■	777743677760771	T1	1	M/23	
2002	0668	■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□■■■■■■■	617777777700771	Unknown	1	M/25	
2002	0671	■■■■□□□□■■■■■■■■■□□□□□■■■■■■■□■■■□□□□■■■■■■■	741774077560771	T1	1	F/36	
2002	0705	■■■■■■■■□□■■■■■■■■■■■■■□□□■■■■■□□□□■■■■■■■	776377761760771	S	1	M/28	HIV+
2003	0719	■■□□□■■■■■■□■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	617737777720771	H3	1	M/32	HIV+
2004	0796	■■■■■■■■■■■■■■■■■□□□□■■■■■■■□■■■□□□□■■■□□■■	777774077560711	T1	1	M/34	HIV+

*Lineage designations for orphan patterns were done manually as Expert-based interpretations using revised SpoIDB4 rules.

**Drug-R code: 1, pansusceptible; 2, MDR (combined resistance to INH-RIF); 3, any other resistance; 4, XDR-TB (combined resistance to INH, RIF, fluoroquinolones, and 1 of 3 injectable drugs, i.e., capreomycin, kanamycin, or amikacin).

doi:10.1371/journal.pone.0065873.t001

Table 2. Description of 122 shared-types (SITs; n = 767 isolates) and corresponding spoligotyping defined lineages/sublineages starting from a total of 794 *M. tuberculosis* strains isolated from adults with pulmonary tuberculosis in Lima, Peru.

SIT *	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. database	Lineage**	Clustered vs. unique patterns***
1	□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■	000000000003771	44 (5.54)	0.46	Beijing	Clustered
4	□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■□□□□■■■■■■	000000007760771	2 (0.25)	0.6	Unknown	Clustered
11	■□□■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□□□■■■■	477777777413071	1 (0.13)	0.18	EAI3-IND	Unique
19	■■□■■■■■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■	677777477413771	1 (0.13)	0.12	EAI2-Manilla	Unique
20	■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	677777607760771	4 (0.5)	0.51	LAM1	Clustered
33	■■■■■■■□□□■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	776177607760771	66 (8.31)	6.05	LAM3	Clustered
36	■■■■■■■■■■□■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777737777720771	4 (0.5)	3.51	H3	Clustered
39	■■■■■■■■■■■■■■■■■□■■■□□■■■■■■■■□□□□■□□■■■■	777777347760471	2 (0.25)	1.49	T4-CEU1	Clustered
42	■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	777777607760771	59 (7.43)	1.92	LAM9	Clustered
46	■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□	777777770000000	1 (0.13)	0.54	Unknown	Unique
47	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	777777774020771	31 (3.9)	2.24	H1	Clustered
49	■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■□■■■■	777777777720731	7 (0.88)	4.43	H3	Clustered
50	■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	777777777720771	130 (16.37)	4.17	H3	Clustered
51	■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□□□□□	777777777760700	1 (0.13)	0.38	T1	Unique
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777777777760771	98 (12.34)	1.74	T1	Clustered
54	■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■	777777777763771	1 (0.13)	0.46	MANU2	Unique
58	■■■■■■■■■■■■■■■■■■■■■□■■□■■■■■■■■□□□□■■■■■■■	777777557760771	8 (1.01)	4.94	T5-Madrid2	Clustered
60	■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■□■■■	777777607760731	2 (0.25)	0.51	LAM4	Clustered
62	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	777777774020731	1 (0.13)	0.2	H1	Unique
64	■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■□□□□■■■■■■■	777777607560771	6 (0.76)	1.75	LAM6	Clustered
73	■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	777737777760731	2 (0.25)	0.84	T	Clustered
78	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□□■■	777777777760711	1 (0.13)	1.61	T	Unique
86	■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■□□□□■■■■■■■	777777737760771	1 (0.13)	1.19	T1	Unique
91	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■□□□□■■■■■■■	700036777760771	22 (2.77)	8.76	X3	Clustered
92	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■□□□□■■■■■■■	700076777760771	5 (0.63)	1.18	X3	Clustered
93	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	777737607760771	12 (1.51)	3.55	LAM5	Clustered
95	■■■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■□□□□■■■□■■■	777777607560731	2 (0.25)	4.55	LAM6	Clustered
99	■■■■□■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	757777777720771	2 (0.25)	2.94	H3	Clustered
106	■■■■■■■□□□■■■■■■■■■□□□□□□□□□□□□□□□□□□■■■■	776177400000171	2 (0.25)	1.55	Unknown	Clustered
119	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■■■■■	777776777760771	4 (0.5)	0.37	X1	Clustered
130	■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■□□□□■■■□■■■	776177607760731	7 (0.88)	6.67	LAM3	Clustered
132	■■■■■■■■■■■■■■■■■■■□□□□■■□□□□□□□□□□□□□■■	777777606000031	1 (0.13)	6.25	Unknown	Unique
177	□■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	377777607760771	1 (0.13)	1.08	LAM9	Unique

Table 2. Cont.

SIT *	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. database	Lineage**	Clustered vs. unique patterns***
183	■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■□■□□□□■■■■■■	777777377720771	3 (0.38)	5.77	H3	Clustered
190	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■□■■■	000000000003731	1 (0.13)	0.56	Beijing	Unique
211	■■■■■■■□□□■□■■■■■■■□□□■■■■■■■■□□□□■■■■■■■	776137607760771	1 (0.13)	1.23	LAM3	Unique
215	■■■■■■■■■■■■□□□□■■■■■■■■■■■■□■■■■□□□□■■■■■■■	777703777360771	1 (0.13)	14.29	T	Unique
216	■■■■■■■■■■■□□■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	777717607760771	3 (0.38)	15	LAM5	Clustered
219	■■■■■■■■■■■■□□□□□■■■■■■■■■■■■□□□□■■■■■■■	777740777760771	5 (0.63)	13.16	T1	Clustered
222	■■■■■■■■■■■■■■□□□□□■■■■■■□■■■□□□□■■■■■■■	777774077560771	24 (3.02)	46.15	Unknown	Clustered
237	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□	777777777700000	4 (0.5)	3.88	Unknown	Clustered
239	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□■■■	777777777760031	2 (0.25)	3.77	T2	Clustered
283	■■■■■■■■■■■■■■■■■■■■■■■■■□□□■□□□□□□□■□□□□■■■■■■	777777704020771	1 (0.13)	1.75	H1	Unique
373	■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■□□□□■■■■■■■	777777767760771	2 (0.25)	3.33	T1	Clustered
384	■■■■■■■■■■■■■■■■■■■■■■■■■□■■■□□□□□□■□□□□■■■■■■■	777777674020771	1 (0.13)	11.11	H1	Unique
390	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■□□□□■■■■■■■	777777777620771	3 (0.38)	10.34	H3	Clustered
396	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■□■■■□□□□□■■■■■■■	777777607560371	2 (0.25)	11.76	LAM6	Clustered
418	■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	617777777720771	9 (1.13)	64.29	H3	Clustered
430	■■■□□□■□□□□□□□□□■■□■■■□□■■■■■■■■□□□□■□□■■■■	704003347760471	4 (0.5)	21.05	T4-CEU1	Clustered
450	■■■■■■■■■■■■■■■■■■■■■□■■■■■■■□□□□□□□□□□□□□□□□	777776770000000	14 (1.76)	15.05	Unknown	Clustered
469	□□□■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	077777607760771	1 (0.13)	3.57	LAM1	Unique
489	■■■■■□■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■□□□□■	767776777760601	1 (0.13)	9.09	X2	Unique
512	■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■□■□□□□■■■■■■■	777777707720771	4 (0.5)	13.79	H3	Clustered
534	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■□□□□□□□□□□□□□	777777607400000	1 (0.13)	8.33	LAM	Unique
546	■■■□□□□□□□□□□■■■■□■■■■■■■■■■□■■■□□□□■■■■■■■	700036777560771	1 (0.13)	7.14	X3	Unique
559	■■■■■■■□□■■■■■■■■■■■■■■■■□□□■■■■■□□□□■■■■■■■	776377770760771	1 (0.13)	16.67	S	Unique
620	■■■■■■■■■■■■■■■■■■■■■■■■■□□■□□□□□□■□□□□■■■■■■■	777777744020771	2 (0.25)	14.29	H1	Clustered
644	□■□□□□□□□□□□□□□□■■□■■■■■■■□■□□□□■■■□□□□□□	200003377207600	1 (0.13)	3.45	BOV_4-CAPRAE	Unique
740	■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■□■□□□□■■■■■■■	777777747720771	1 (0.13)	7.69	H3	Unique
748	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■□□	777777777720760	1 (0.13)	20	H3	Unique
777	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■□□□□■■■■■■■	777777777420771	1 (0.13)	2.08	H3	Unique
784	■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■■	776377777760731	1 (0.13)	1.82	S	Unique
786	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□	777777760000000	3 (0.38)	23.08	Unknown	Clustered
826	■■□■■■■■■■■□■■■■■□■□□□□■■■■■■■■□□□□■■■■■■■	677737207760771	4 (0.5)	44.44	LAM2	Clustered
849	■■□□■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	637777777720771	2 (0.25)	20	H3	Clustered
867	■■■■■■■■■■■□■■■■■■■□□□□■■■■□■■■□□□□■■■□■■■	777737607560731	1 (0.13)	7.69	LAM	Unique
893	■■■■■■■■■■■■■■■■■■■□□□□□■■■■■■■■■□□□□■■■■■■■	777777017760771	3 (0.38)	33.33	T	Clustered

Table 2. Cont.

SIT *	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. database	Lineage**	Clustered vs. unique patterns***
914	■■■■■■■□□■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	776377777720771	2 (0.25)	5.26	Unknown	Clustered
1080	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■■□□□■	777776777760701	1 (0.13)	14.29	X1	Unique
1105	■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777773777760771	8 (1.01)	29.63	T1	Clustered
1122	■■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	767777777760771	1 (0.13)	2.38	T1	Unique
1139	■■■■■■■■■■■■■■■■■■■□■■■■■□□□□□□■□□□□■■■■■■■	777777374020771	2 (0.25)	25	H1	Clustered
1150	■■■□□□□□□□□□■■■■□■■■■■■■■□□■■■■□□□□■■■■■■■	700036776360771	2 (0.25)	50	X3	Clustered
1177	■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■□□□□□□□□□□□□	777777767700000	1 (0.13)	25	Unknown	Unique
1214	■■■■■■■■■■□□□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777617777760771	1 (0.13)	5	T3	Unique
1220	■■■■□■■□□□■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■□□□□	754377607760700	1 (0.13)	33.33	LAM	Unique
1230	■■■■■■■■■■■■■□■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	777757774020771	1 (0.13)	20	H1	Unique
1235	■■■■■■■■■■■■■■■■■■■■■■■■■□□■□□■■■□■□□□□■■■■■■■	777777744720771	1 (0.13)	33.33	H3	Unique
1354	■□■■■■■■□□□■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	576177607760771	2 (0.25)	33.33	LAM3	Clustered
1355	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□■■■■□■■■□□□□■■■□■■■	777777407560731	24 (3.02)	50	LAM	Clustered
1356	■■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■	776377777760751	2 (0.25)	11.76	S	Clustered
1476	■■■□■■■■□□□■■■■■■■■■■□□□□■■■■■■□□□□□□□□■■■■■	736177607700171	5 (0.63)	41.67	AFRI_2	Clustered
1525	□■■■■■■■□□□■□■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	376137607760771	1 (0.13)	20	LAM3	Unique
1552	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■□□■■■	777777774020631	1 (0.13)	16.67	H1	Unique
1563	■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■□■■■■■□□□□■■■■■■■	777777636760771	1 (0.13)	33.33	T1	Unique
1617	■■■□□□□□□□□□□■■■■□■■■■■■■■□■■■■■□□□□■■■■■■■	700036776760771	1 (0.13)	33.33	X3	Unique
1681	□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	377777774020771	1 (0.13)	20	H1	Unique
1708	■■■■■■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	776777607760771	1 (0.13)	33.33	LAM9	Unique
1999	■■■■■■■■■■■□□□□□■■■□□□□■■■■■■■■■□□□□■■■■■■■	777701607760771	1 (0.13)	16.67	LAM5	Unique
2028	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□□□□□□■□□□□■■■■■■■	777777604020771	1 (0.13)	20	Unknown	Unique
2054	■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■■□□□□□■■■■■■	677777607760371	1 (0.13)	20	LAM	Unique
2179	■■■■■■■□□■■■■■■■■■■■■■■■■□□□□□■■■□□□□■■■■■■■	776377760360771	1 (0.13)	20	S	Unique
2230	■■■■■■■■■■■□□□□□□□□□□■■■■■■■■■■■□□□□■■■■■■■	777700077760771	1 (0.13)	25	Unknown	Unique
2274	□□□□■■■■■■■■■■■■■■■■□□□□□■■■■□■■■□□□□■■■□■■■	037777407560731	7 (0.88)	63.64	LAM	Clustered
2381	■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777677777720771	1 (0.13)	12.5	H3	Unique
2383	■■■■□■■■■■■□■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	757737607760771	1 (0.13)	20	LAM5	Unique
2626	■■■■■■■□□□■■■■■■■■■■■□□□□■■■■■□□□□□□□■■■■■■■	776177607700171	1 (0.13)	25	LAM3	Unique
2744	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□■■■■■■■	777777777700071	1 (0.13)	33.33	Unknown	Unique
2885	■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■□■□□□□■■■■■■■	777777477720771	1 (0.13)	9.09	H3	Unique
2916	■■■■■■■□□■■■■■■■■■■■■■■■■□■■■■□□□□■■■■■■■	776377777360771	1 (0.13)	25	S	Unique
2961*	■□□□□□□□□□□□□□□□□□□■■□□■■■■■■□□□□■□■■■■■	400000071760571	2 (0.25)	100.0	Unknown	Clustered

Table 2. Cont.

SIT *	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. database	Lineage**	Clustered vs. unique patterns***
3000*	■■■■□■■■■■■■■■■■■■□□□□□□■■■■□■■■□□□□■■■□■■■	757777407560731	3 (0.38)	100.0	LAM	Clustered
3001*	■■■■■■■■■■■■■■■■□□□□□■■■■■□□□□□■□□□□■■■■■■■	777774077020771	8 (1.01)	88.9	H3	Clustered
3004*	■■■■■■■□□■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□	776377777000000	2 (0.25)	100.0	Unknown	Clustered
3005*	■■□□□□□■■■■■■■■■■■■■■■□■■■■■■■□■■■□□□□■■■■■■■	603777677560771	2 (0.25)	100.0	T	Clustered
3006*	■□■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■□■■■	577777607760731	2 (0.25)	66.7	LAM4	Clustered
3007*	■■■■■■■■□□■■■■■■■■■■■■■■■□□□□■■■□□□□■■■■■■■	776377770360771	1 (0.13)	50.0	S	Unique
3008*	■■■■■■■□■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	776777774020771	1 (0.13)	50.0	H1	Unique
3009*	■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□□□■■■■■■	777773777760171	2 (0.25)	100.0	T1	Clustered
3010*	■■■■■■■■□□■■■■■■■■■■■■■■■■■□□□□■■■□■■■	776377777740731	1 (0.13)	50.0	S	Unique
3011*	■■■■□□□■■■■■■■■■□□□□□■■■■■■□■■■□□□□■■■□■■■	743774077560731	4 (0.5)	100.0	Unknown	Clustered
3012*	■■■■■■■■■■■■■■■■□□□□□■■□□□□□□□□□□□□□■■■■■	777774070000171	2 (0.25)	100.0	Unknown	Clustered
3013*	□■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■■■■■	377777377760771	3 (0.38)	66.7	T4	Clustered
3014*	■■□■■■■■■■■■■■■■■■■□□□□■■■■■■□□□□□□■■■■■■■	677777607700771	1 (0.13)	50.0	Unknown	Unique
3015*	□□□■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■□□□■	077776777760701	2 (0.25)	100.0	X1	Clustered
3016*	■■□□□□□□□□□□□□□□□□□□□□□□□□□■□■■■□□□□■■■■■■	600000000560771	1 (0.13)	50.0	Unknown	Unique
3017*	■■■■■■■■■■■■■■■■■■■■■■■■■□□■□■□□□□■■■■■■■	777777777120771	3 (0.38)	100.0	H3	Clustered
3089*	□■■□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	337777777760771	2 (0.25)	66.67	T1	Clustered
3168*	■■■□■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■□■■■	737777777720731	1 (0.13)	50	H3	Unique
3431*	■■■■■■■■■■■■■■■■■□■■■■■■■□■■■■□□□□■■■■■■■	777776775720771	2 (0.25)	100	H3	Clustered
3432*	■■■□□□□□□□□□□□□■■■□□□□■■■■■■■□□□□■■■■■■■	700003607760771	2 (0.25)	100	LAM3	Clustered
3433*	■■■■■■■□□□■■■■■■■■■□□□□■■■□□■■□□□□■■■■■■■	776177607460771	2 (0.25)	66.67	LAM3	Clustered

*A total of 100/122 SITs (n = 718) matched a preexisting shared-type in the database, whereas 22/122 SITs (n = 49 isolates) were newly created either within the present study or after a match with an orphan in the database. A total of 66 SITs containing 711 isolates were clustered within this study (2 to 130 isolates per cluster), while 56 SITs contained a unique strain within this study.

Note that SITs followed by an asterisk indicates "newly created shared-type" (n = 22 containing 49 isolates) due to 2 or more strains belonging to an identical new pattern within this study or after a match with an orphan in the database. SIT designations followed by number of strains: 2961* this study (n = 2); 3000* this study (n = 3); 3001* this study (n = 8) and USA (n = 1); 3004* this study (n = 2); 3005* this study (n = 2); 3006* this study (n = 2) and South Africa (n = 1); 3007* this study (n = 1) and USA (n = 1); 3008* this study (n = 1) and USA (n = 1); 3009* this study (n = 2); 3010* this study (n = 1) and USA (n = 1); 3011* this study (n = 4); 3012* this study (n = 2); 3013* this study (n = 3) and USA (n = 1); 3014* this study (n = 1) and Argentina (n = 1); 3015* this study (n = 2); 3016* this study (n = 1) and Panama (n = 1); 3017* this study (n = 3); 3089* this study (n = 2) and Mexico (n = 1); 3168* this study n = 1, Sweden (n = 1); 3431* this study (n = 2); 3432* this study (n = 2); 3433* this study (n = 2), BRA (n = 1).

**Lineage designations according to SITVIT2 using revised SpolDB4 rules; "Unknown" designates patterns with signatures that do not belong to any of the major clades described in the database.

***Clustered strains correspond to a similar spoligotype pattern shared by 2 or more strains "within this study"; as opposed to unique strains harboring a spoligotype pattern that does not match with another strain from this study. Unique strains matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as "orphan" (see Table 1).

doi:10.1371/journal.pone.0065873.t002

Table 3. Description of clusters composed of predominant shared types (defined as SITs representing >2% strains, n = 16) in our study and their worldwide distribution in the SITVIT2 database.

SIT (Clade) Octal Number Spoligotype Description	Number (%) in study	% in study vs. SITVIT2	Distribution in Regions with ≥3% of a given SITs *	Distribution in Countries with ≥3% of a given SITs **
50 (H3) 777777777720771 ■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	130 (16.37)	4.17	AMER-N 19.94, AMER-S 18.5, EURO-W 13.78, EURO-S 12.56, EURO-E 5.78, EURO-N 4.66, AFRI-N 4.62, AFRI-S 4.4, CARI 3.76, ASIA-W 3.02	USA 19.08, BRA 7.68, AUT 6.62, ITA 5.91, ESP 5.91, PER 4.46, ZAF 4.4, CZE 3.98, SWE 3.08
53 (T1) 777777777760771 ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	98 (12.34)	1.74	AMER-N 18.23, AMER-S 13.2, EURO-W 11.32, EURO-S 10.5, ASIA-W 7.61, EURO-N 5.95, AFRI-S 5.54, AFRI-E 5.03, ASIA-E 4.73, AFRI-N 3.93	USA 14.73, ITA 5.95, BRA 5.72, ZAF 5.42, TUR 3.87, AUT 3.82, CHN 3.45, MEX 3.18
33 (LAM3) 776177607760771 ■■■■■■■□□□■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	66 (8.31)	6.05	AFRI-S 29.79, AMER-S 25.76, AMER-N 15.03, EURO-S 12.65, EURO-W 5.04, AMER-C 4.49	ZAF 29.79, USA 14.57, BRA 12.1, ESP 8.16, PER 6.23, ARG 5.23, HND 3.94, ITA 3.85
42 (LAM9) 777777607760771 ■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	59 (7.43)	1.92	AMER-S 31.52, AMER-N 14.87, EURO-S 10.9, AFRI-N 9.34, EURO-W 6.25, EURO-N 4.07, AFRI-E 3.87, AFRI-S 3.45	BRA 13.05, USA 12.92, COL 8.3, MAR 7.68, ITA 5.69, ESP 3.64, VEN 3.61, ZAF 3.45
1 (Beijing) 000000000003771 □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■	44 (5.54)	0.46	ASIA-E 34.21, AMER-N 20.99, ASIA-SE 9.48, AFRI-S 8.63, ASIA-N 7.22, ASIA-S 4.67, EURO-N 3.23	USA 20.65, CHN 19.77, JPN 12.0, ZAF 8.63, RUS 7.22, VNM 4.06
47 (H1) 777777774020771 ■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	31 (3.9)	2.24	AMER-N 19.32, EURO-W 17.81, EURO-S 15.21, AMER-S 12.76, EURO-E 7.43, EURO-N 7.14, AFRI-N 4.11, ASIA-W 3.89	USA 17.16, ITA 9.3, AUT 9.08, BRA 7.93, CZE 4.25, ESP 4.04, SWE 3.82, MAR 3.17
222 (Unknown) 777774077560771 ■■■■■■■■■■■■■■■■■■■■■□□□□□■■■■■■■□■■■□□□□■■■■■■■	24 (3.02)	46.15	AMER-S 51.92, AMER-N 34.62, EURO-S 9.62, EURO-N 3.85	PER 50.0, USA 32.69, ESP 5.77, SWE 3.85, ITA 3.85
1355 (LAM) 777777407560731 ■■■■■■■■■■■■■■■■■■■■■□□□□□■■■■□■■■□□□□■■■□■■■	24 (3.02)	50	AMER-S 56.25, EURO-S 27.08, AMER-N 14.58	PER 54.17, ITA 20.83, USA 14.58, ESP 6.25
91 (X3_var) 700036777760771 ■■■□□□□□□□□□□□■■■■□■■■■■■■■■■■■□□□□■■■■■■■	22 (2.77)	8.76	AMER-N 51.0, AMER-S 21.91, CARI 13.94, EURO-S 6.38, EURO-N 4.38	USA 47.01, HTI 11.55, PER 9.56, ESP 6.38, COL 4.78, GUF 3.98, CAN 3.19

*Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database. The definition of macro-geographical regions and sub-regions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia.

**The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ≥3% of a given SITs as compared to their total number in the SITVIT2 database.

doi:10.1371/journal.pone.0065873.t003

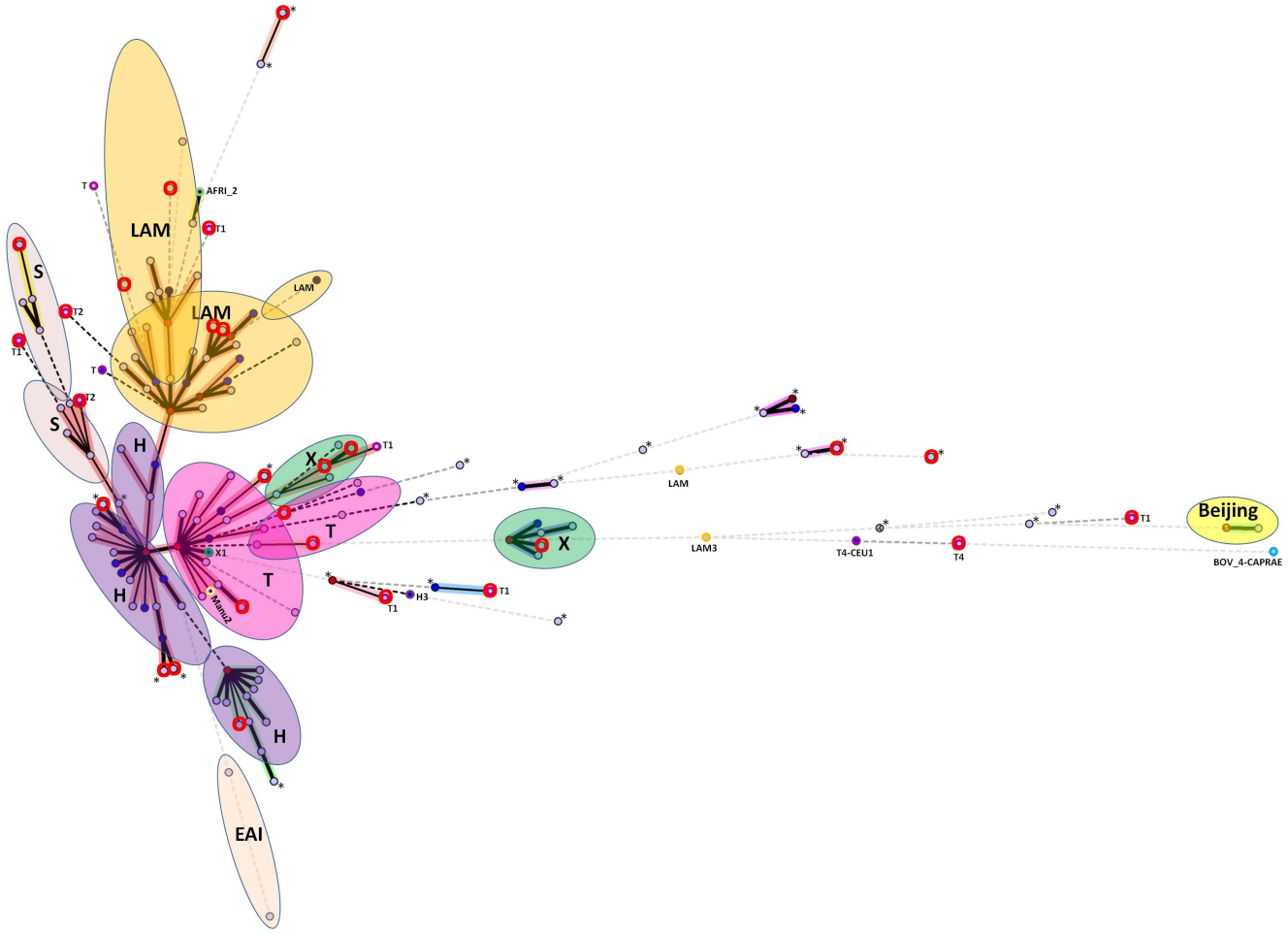


Figure 1. A minimum spanning tree (MST) illustrating evolutionary relationships between the Peruvian spoligotypes (n = 794). The phylogenetic tree connects each genotype based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, black dotted and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines, 1 or 2 changes (thicker ones indicate a single change, while the thinner ones indicate 2 changes); dotted lines, three or more changes (black dotted for 3, and grey dotted for 4 or more changes). The color of the circles is proportional to the number of clinical isolates in our study, illustrating unique isolates (sky blue) versus clustered isolates (Blue, 2–5 strains; dark blue, 6–9 strains; Bordeaux, 10–19 strains; Red, 20 and more). Note that orphan patterns are circled with the letter "o" in red. Patterns marked by an asterisk (*) indicate a strain with an unknown signature (unclassified).
doi:10.1371/journal.pone.0065873.g001

observed in a previous study in north Lima [2]; 5.5% were SIT1/Beijing family. Strains belonging to Haarlem, T, LAM and Beijing families predominated, and drug-resistance was not shown to be associated with any specific family, including Beijing, findings consistent with the single previous report from Peru [2].

With the exception of the Beijing family strains, recently examined in greater detail [10] very few PGG1 strains (AFRI, BOV, EAI, Manu) were found in this study (n = 9, 1.1%). One may notice that these PGG1 strains are not located at central positions on the trees (spoligoforests and MST). Instead, they mostly occupied terminal leaves of the trees (in the MST), or were isolated with few or no connections with other strains in the spoligoforests.

However, PGG2/3 group (n = 665, 83.8%) strains which are predominant in the study occupied a more visible and central position on the trees. Spoligoforest trees have been used to highlight the predominance of some specific well known shared types (SIT). These trees can also give us an overview on the parental links that probably exist between strains belonging to different lineages. For example, one may notice on the top left of the hierarchical layout figure (2A), that SIT19/EAI2-Manila may lead to SIT1/Beijing, through loss of many spacers.

The MST very well shows the similarity (or the distance) between each strain, and clearly defines the major evolution of the MTB lineages present in the study. For example, one can notice that in Figure 1, the Beijing family group is very far from the strains present in the central nodes of strains belonging to the PGG2/3 group (Euro American). At the very bottom of the MST, we can note the presence of the only two strains belonging to EAI lineage (SIT19/EAI2-Manila and SIT11/EAI3-IND).

The spoligoforest tree demonstrates that most of the orphans belong to the modern PGG2/3 group (H, T, LAM, T, X, S). The orphan strains are mostly located at terminal positions on the trees or are located in the top right layer of the hierarchical layout as isolated strains without interconnections with the other strains.

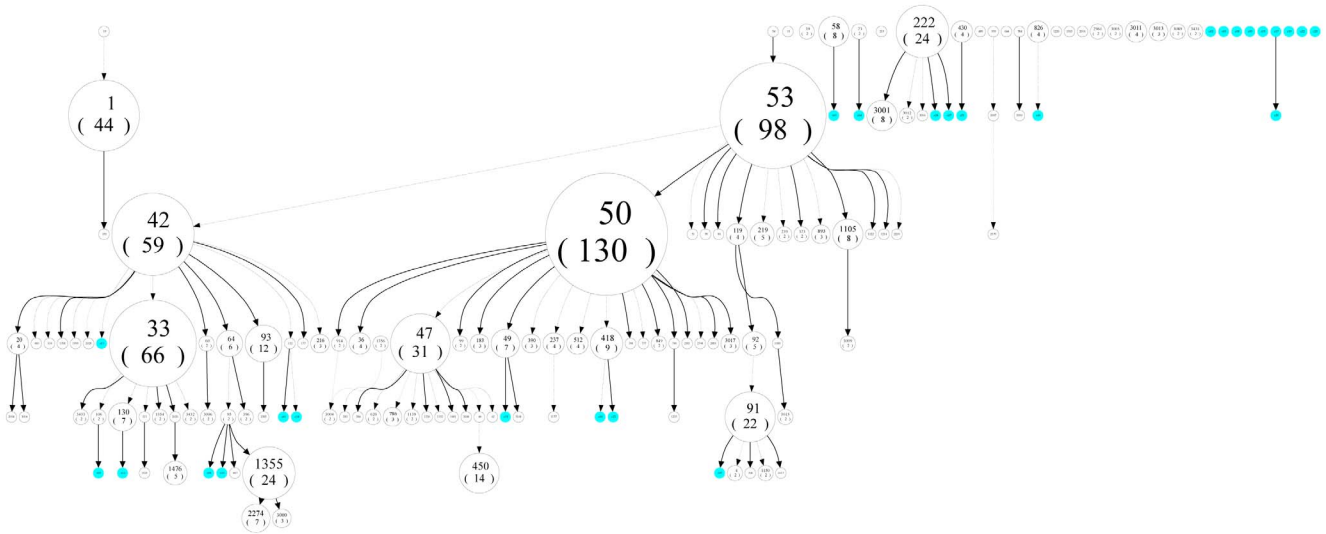


Figure 2. Discrete spoligotypes relationships for all isolates (n = 794) presented through spoligoforest tree drawn as a ''hierarchical layout'' using the SpolTools software (available through http://www.emi.unsw.edu.au/spolTools; Reyes et al. 2008 [11]). Each spoligotype pattern from the study is represented by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. In this representation, the heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weight being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5. Orphan isolates, indicated in cyan, are isolated strains without interconnections with the other strains. This presentation illustrates for example the parental links for PGG2/3 strains such as SIT53 and SIT42, showing how SIT53 may be considered as the precursor of all other modern PGG2/3 patterns. SIT53 leads to SIT50/H3 by the loss of spacer 31, and it leads to SIT42 by the loss of four spacers (spacers 21–24), which in turn leads to SIT1355/LAM via SIT64/LAM6 then SIT95/LAM6. Through other spacer deletions, SIT53 leads to SIT91/X3 via SIT119/X1 and SIT92/X3. Lastly, SIT222/Unknown has no parental SITs in our study.
doi:10.1371/journal.pone.0065873.g002

Indeed, none of the orphans explicitly belonged to the PGG1 group (considering that some orphans have an unknown lineage).

Given the celebrated performance of the Peruvian National Tuberculosis Control Programme in demonstrating the effectiveness of DOTS in bringing about a reduction in TB incidence [3], the emergence of MDRTB in Peru as a major threat might be viewed as surprising [11]; conventional wisdom suggests MDR is driven by weak healthcare systems. We were therefore interested in exploring whether other factors, such as strain-specific biological propensity for resistance, might be relevant. In unselected community-based TB patients (largely HIV-negative) there was no association observed between drug resistance and specific spoligotypes. However amongst patients with HIV recruited from a hospital setting MDR was particularly frequently seen amongst the SIT42 and SIT53 strains. After multivariate analysis to control for the effects of HIV infection, gender and year

Table 4. Strain spoligotype (SIT) frequency by patient HIV status and strain MDR status.

	HIV negative		HIV positive		SIT total number of isolates[1] (% MDR[2])
SIT (of clade)	**MDR**	**Non-MDR**	**MDR**	**Non-MDR**	
50 (H3)	7	96	3	22	128 (7.8%)
53 (T1)	8	49	24	16	97 (33.0%)
33 (LAM3)	1	53	0	11	65 (1.54%)
42 (LAM9)	4	15	35	5	59 (66.1%)
1 (Beijing)	3	32	1	8	44 (9.1%)
47 (H1)	2	24	0	5	31 (6.5%)
22 (unknown)	2	16	2	4	24 (16.7%)
1355 (LAM)	1	12	4	6	23 (21.7%)
91 (X3)	1	14	1	6	22 (9.1%)
others	24	183	12	70	289 (12.5%)
total	53	494	82	153	782 (17.3%)

[1]12 of 794 isolates lacked either MDR or HIV status data, leaving 782 here as the denominator.
[2]Percentage of SIT-specific isolates with MDR.
doi:10.1371/journal.pone.0065873.t004

Table 5. Associations with drug resistance.

	Isoniazid monoresistance		Rifampicin monoresistance		MDR	
	Unadjusted OR (95% CI)	Adjusted[1] OR (95% CI)	Unadjusted OR (95% CI)	Adjusted[1] OR (95% CI)	Unadjusted OR (95% CI)	Adjusted[1] OR (95% CI)
HIV	1.05 (0.61–1.81)	1.05 (0.59–1.87)	1.63 (0.83–3.20)	1.70 (0.83–3.49)	**5.00 (3.38–7.38)**	**3.42 (2.21–5.29)**
Male gender	1.18 (0.69–2.01)	1.20 (0.69–2.08)	1.21 (0.60–2.44)	1.07 (0.51–2.24)	**1.51 (1.01–2.27)**	1.13 (0.70–1.80)
Year	1.04 (0.91–1.18)	1.04 (0.91–1.19)	**1.20 (1.01–1.43)**	1.18 (0.98–1.41)	0.99 (0.90–1.09)	1.03 (0.92–1.16)
Clustered SIT	0.92 (0.55–1.54)	0.94 (0.54–1.64)	**0.38 (0.19–0.75)**	**0.43 (0.20–0.91)**	**1.77 (1.17–2.67)**	0.69 (0.41–1.18)
SIT42 (LAM9)	0.86 (0.64–1.16)	0.87 (0.63–1.19)	0.76 (0.46–1.25)	0.81 (0.48–1.37)	**1.89 (1.63–2.18)**	**1.95 (1.64–2.32)**
SIT53 (T1)	1.12 (0.78–1.59)	1.12 (0.76–1.65)	0.78 (0.43–1.42)	0.96 (0.50–1.85)	**1.67 (1.32–2.11)**	**2.16 (1.60–2.91)**

OR = odds ratio; 95% CI = 95% confidence interval; clustered SIT = in 9 most common SITs (accounting for 63% of all strains); bold type indicates statistically significant associations.

[1]all adjusted ORs incorporate SIT42 and SIT53 and clustered SIT variable and into model.

doi:10.1371/journal.pone.0065873.t005

the effect size increased; given the lack of such an association in the community one hypothesis to explain this would be that this is highly suggestive of a prolonged nosocomial clonal outbreak with strains of these two spoligotypes. The alternative hypothesis of a biological predisposition of these specific strains to acquire drug resistance-conferring mutations is much less likely given the absence of an association with isoniazid or rifampicin mono-resistance. It is noteworthy that the association of HIV with MDR, though diminished after adjustment for SIT42 and SIT53, remained significant indicating that an outbreak with strains from these two spoligotypes is insufficient to explain the whole HIV-MDR association. We cannot exclude the possibility of residual confounding as the explanation for this apparent association.

There are acknowledged limitations of the data presented here. Most importantly our sampling strategy was opportunistic, making use of a strain bank derived from several studies with different designs so the study populations differed and though all relevant subgroups (community and hospital based, HIV infected and uninfected) were included the sample could not be considered representative. There are advantages in having a strain bank which is delinked from patient identifiers but the drawback is that only limited clinical data is available and returning to clinical notes for further detail is not possible – it would have been interesting to differentiate between new and retreatment cases and to investigate patient outcomes by strain, for example. Finally, because the samples were all from studies in adults we were unable to describe the strains causing paediatric disease thus missing an opportunity to clearly identify currently/recently circulating strains, and because all strains were from pulmonary TB patients we were unable to investigate whether extrapulmonary disease phenotype was associated with any particular strain in Peru as has been suggested elsewhere [12].

There are important strengths in our strain bank: (1) each sample evaluated here is a single strain from a unique patient – though serial strains are available in the bank, for this analysis we were careful to only examine one strain per patient, (2) availability of drug susceptibility data and HIV status for every strain enabled the analysis we report here with very few missing values, in contrast to an earlier report for which only 70% of strains had drug susceptibility data and HIV status was not reported [2], (3) the spread of strains includes diverse but well characterized patient demographic groups which are also geographically spread across metropolitan Lima (home to one third of the population of Peru and more than 75% of the incident TB), (4) the collection reported here span a time period of 6 years (indeed the bank continues to accumulate strains to the present day, extending the collection to more than 13 years) enabling investigation of temporal trends (none were found here).

In summary, we report the strain distribution of *M tuberculosis* isolates in Lima, Peru, highlight a significant proportion of novel spoligotypes, and hypothesize a prolonged, clonal, hospital-based outbreak of MDR disease amongst HIV patients but no evidence to support a hypothesis of strain-specific propensity for the acquisition of resistance-conferring mutations.

Supporting Information

Figure S1 MST from manuscript Figure 1 presented in PDF format which (through zooming in) enables reading of SIT labels.
(PDF)

Figure S2 Discrete spoligotypes relationships for all isolates (n = 794) presented through a Fruchterman Reingold spoligoforest tree drawn using the SpolTools software (available through http://www.emi.unsw.edu.au/spolTools; Reyes et al. 2008 [11]). Each spoligotype pattern from the study is represented by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. In this representation, the heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weight being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5. Orphan isolates, indicated in cyan, appear at terminal positions on the tree, as isolated strains without interconnections with the other strains.
(PDF)

Table S1 Detailed genotyping and drug-resistance data and demographic information on *M. tuberculosis* strains (n = 794) isolated from adults with pulmonary tuberculosis in Lima, Peru.
(PDF)

Table S2 A comparison of the proportion of the most predominant SITs found in Peru as compared to neighbouring countries (Brazil, Colombia) and regions (Central America and Caribbean), recorded in the SITVIT2 database as consulted on 9 April 2013.
(PDF)

Acknowledgments

We thank Jorge Coronel and the late and sorely missed Luz Caviedes for their contribution to the original laboratory work.

Author Contributions

Conceived and designed the experiments: PS LG MZ NR DM. Performed the experiments: PS DC MD GL NR. Analyzed the data: PS DC LG MZ MD NR DM. Contributed reagents/materials/analysis tools: PS DC LG RG NR DM. Wrote the paper: PS LG MZ GL RG NR DM.

References

1. Demay C, Liens B, Burguiere T, Hill V, Couvin D, et al. (2012) SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol 12: 755–766.
2. Taype CA, Agapito JC, Accinelli RA, Espinoza JR, Godreuil S, et al. (2012) Genetic diversity, population structure and drug resistance of Mycobacterium tuberculosis in Peru. Infect Genet Evol 12: 577–585.
3. Suarez PG, Watt CJ, Alarcon E, Portocarrero J, Zavala D, et al. (2001) The dynamics of tuberculosis in response to 10 years of intensive control effort in Peru. J Infect Dis 184: 473–478.
4. Kawai V, Soto G, Gilman RH, Bautista CT, Caviedes L, et al. (2006) Tuberculosis mortality, drug resistance, and infectiousness in patients with and without HIV infection in Peru. Am J Trop Med Hyg 75: 1027–1033.
5. Moore DA, Evans CA, Gilman RH, Caviedes L, Coronel J, et al. (2006) Microscopic-observation drug-susceptibility assay for the diagnosis of TB. N Engl J Med 355: 1539–1550.
6. Moore DA, Mendoza D, Gilman RH, Evans CA, Hollm Delgado MG, et al. (2004) Microscopic observation drug susceptibility assay, a rapid, reliable diagnostic test for multidrug-resistant tuberculosis suitable for use in resource-poor settings. J Clin Microbiol 42: 4432–4437.
7. Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM (2008) spolTools: online utilities for analyzing spoligotypes of the Mycobacterium tuberculosis complex. Bioinformatics 24: 2414–2415.
8. Reyes JF, Francis AR, Tanaka MM (2008) Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 9: 496.
9. Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G (2002) Graphviz - Open Source Graph Drawing Tools. In: Mutzel P, Jünger M, Leipert S, editors. Heidelberg: Springer-Verlag Berlin. pp. 483–484.
10. Iwamoto T, Grandjean L, Arikawa K, Nakanishi N, Caviedes L, et al. (2012) Genetic diversity and transmission characteristics of Beijing family strains of Mycobacterium tuberculosis in Peru. PLoS One 7: e49651.
11. Zignol M, van Gemert W, Falzon D, Sismanidis C, Glaziou P, et al. (2012) Surveillance of anti-tuberculosis drug resistance in the world: an updated analysis, 2007–2010. Bull World Health Organ 90: 111–119D.
12. Svensson E, Millet J, Lindqvist A, Olsson M, Ridell M, et al. (2011) Impact of immigration on tuberculosis epidemiology in a low-incidence country. Clin Microbiol Infect 17: 881–887.

Article 09

Infection, Genetics and Evolution 14 (2013) 39–45

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Population structure and circulating genotypes of drug-sensitive and drug-resistant *Mycobacterium tuberculosis* clinical isolates in São Paulo state, Brazil

Maria Conceição Martins [a,*], Carmen M. Saraiva Giampaglia [a], Rosângela S. Oliveira [a], Vera Simonsen [a], Fábio Oliveira Latrilha [a], Letícia Lisboa Moniz [a], David Couvin [b], Nalin Rastogi [b], Lucilaine Ferrazoli [a]

[a] *Núcleo de Tuberculose e Micobacterioses, Centro de Bacteriologia, Instituto Adolfo Lutz, São Paulo, Brazil*
[b] *WHO Supranational TB Reference Laboratory, Institut Pasteur de la Guadeloupe, Abymes, France*

ARTICLE INFO

Article history:
Received 20 June 2012
Received in revised form 4 October 2012
Accepted 26 October 2012
Available online 28 November 2012

Keywords:
Mycobacterium tuberculosis
Population-structure
Spoligotyping
Brazil
Database

ABSTRACT

São Paulo is the most populous Brazilian state and reports the largest number of tuberculosis cases in the country annually (over 18,500). This study included 193 isolates obtained during the 2nd Nationwide Survey on *Mycobacterium tuberculosis* Drug Resistance that was conducted in São Paulo state and 547 isolates from a laboratory based study of drug resistance that were analyzed by the Mycobacteria Reference Laboratory at the Institute Adolfo Lutz. Both studies were conducted from 2006 to 2008 and sought to determine the genetic diversity and pattern of drug resistance of *M. tuberculosis* isolates (MTC) circulating in São Paulo. The patterns obtained from the spoligotyping analysis demonstrated that 51/740 (6.9%) of the isolates corresponded to orphan patterns and that 689 (93.1%) of the isolates distributed into 144 shared types, including 119 that matched a preexisting shared type in the SITVIT2 database and 25 that were new isolates. A total of 77/144 patterns corresponded to unique isolates, while the remaining 67 corresponded to clustered patterns (n = 612 isolates clustered into groups of 2–84 isolates each). The evolutionarily ancient PGG1 lineages (Beijing, CAS1-DEL, EAI3-IND, and PINI2) were rarely detected in São Paulo and comprised only 13/740, or 1.76%, of the total isolates; all of the remaining 727/740, or 98.24%, of the MTC isolates from São Paulo state were from the recent PGG2/3 evolutionary isolates belonging to the LAM, T, S, X, and Haarlem lineages, i.e., the Euro-American group. This study provides the first overview of circulating genotypes of *M. tuberculosis* in São Paulo state and demonstrates that the clustered shared types containing seven or more *M. tuberculosis* isolates that are spread in São Paulo state included both resistant and susceptible isolates.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Tuberculosis (TB) continues to be a major cause of disease and death and is a significant economic burden in the Americas. According to the World Health Organization's (WHO) report on global TB control from 2009, Brazil had an estimated 92,102 new cases of TB in 2007 and an estimated incidence of 48/100,000 in the population. These cases accounted for 31% of all TB cases in the WHO's Latin American Region (http://www.usaid.gov/our_work/global_health/id/tuberculosis/countries/lac/brazil_profile.html). São Paulo state had over 18,500 cases and a reported incidence of 40.5/100,000 in 2010 (CVE-São Paulo, 2011). Although the incidence of TB in São Paulo state is close to the average for Brazil (Bombarda and Galesi, 2006; WHO, 2011), it nonetheless requires special attention because of the high number of patients; the increased number of multidrug-resistant (MDR) tubercle bacilli cases, which are defined by simultaneous drug resistance to isoniazid (I) and rifampin (R) (i.e., n = 1056 cases in 2007); and the relatively high prevalence of HIV (12%) in patients with TB in Brazil (http://www.stoptb.org/assets/images/about/tbl_burden.gif). As of October 2008, extensively drug-resistant TB (XDR-TB, which is defined as MDR-TB, i.e., resistance to I and R and simultaneous resistance to any of the fluoroquinolones and at least one of the three injectable second-line drugs) had also been found in Brazil (Araújo-Filho et al., 2008).

Global and local surveillance of *Mycobacterium tuberculosis* drug resistance is beneficial to TB control programs because information regarding resistance is important for effective control of new TB transmissions. The first World Survey covering the period from 1995 to 1996 showed that MDR-TB represented 1.3% of total TB cases detected in Brazil (WHO, 1997). The 2nd Nationwide Survey on *M. tuberculosis* Drug Resistance, which covered the period from

* Corresponding author. Tel.: +55 11 32412472; fax: +55 11 30682986.
E-mail addresses: martinsesouza@uol.com.br (M.C. Martins), hrgiampa@uol.com.br (Carmen M. Saraiva Giampaglia), rosangela.bio.pqc@gmail.com (R.S. Oliveira), vsimonsen@ial.sp.gov.br (V. Simonsen), folatrilha@hotmail.com (F.O. Latrilha), leticialisboa17@gmail.com (L.L. Moniz), dcouvin@pasteur-guadeloupe.fr (D. Couvin), nrastogi@pasteur-guadeloupe.fr (N. Rastogi), lferrazoli@ial.sp.gov.br (L. Ferrazoli).

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2012.10.015

2006 to 2008, found that MDR-TB accounted an estimated 0.90% of all newly notified TB cases (ranging from 0.60% to 1.4%; WHO, 2011). Similar observations in the last decade have highlighted the global increase in drug-resistant tubercle bacilli, thereby facilitating and promoting the use of molecular markers to study TB epidemiology. The global TB control objectives were recently redefined to add an active research component to routine TB diagnostics and a treatment approach because a better understanding of the disease and the bacterium that causes it may facilitate the design of new approaches to control TB (http://www.stoptb.org/global/plan).

Molecular typing of *M. tuberculosis* greatly assists our understanding of the epidemiology and diversity of this pathogen. Restriction fragment length polymorphism analysis using IS*6110* (IS*6110*-RFLP), PCR-based methods, such as mycobacterial interspersed repetitive units–variable-number tandem repeats (MIRU–VNTR) and Spacer Oligonucleotide Typing (spoligotyping) are considered standard methodologies for elucidating the dissemination of *M. tuberculosis*. The spoligotyping which is an easy method has replaced the cumbersome IS*6110*-RFLP method because it allows simple determination of the circulating clades (genotypic families, lineages) of tubercle bacilli in a specific region versus their worldwide circulation (van Embden et al., 1993; Brudey et al., 2006).

Among these newer approaches, a fine study of the macropopulation structure of *M. tuberculosis* clinical isolates is now possible thanks to new innovations that were developed to study TB molecular epidemiology (reviewed in García de Viedma et al., 2011) and the existence of huge international *M. tuberculosis* genotyping databases, such as the SITVIT2 proprietary database at the Pasteur Institute of Guadeloupe, which was recently used in the first nationwide study to describe the spoligotyping-based genetic diversity of *M. tuberculosis* complex clinical isolates in Brazil (Gomes et al., 2011). Although the emerging drug-resistant strains of TB may further exacerbate the already complex TB situation in Brazil (Bombarda and Galesi, 2006; Dalcolmo et al., 2007; WHO, 1997), this preliminary study from Gomes et al. (2011) did not explore the distribution of drug resistance patterns in relation to genotypic lineages. Because a correlation between genotype, the resistance data, and the respective prevalence of clones region by region may provide more insight into the global circulation of complex *M. tuberculosis* genotypic clones (Borsuk et al., 2005; Ritacco et al., 2011), we decided to analyze by spoligotyping the population structure of *M. tuberculosis* clinical isolates and the existence of resistance-associated genotypes. In our opinion, this is an important step in understanding TB and in the implementation of the appropriate control measures in São Paulo, Brazil.

2. Material and methods

2.1. Patients and bacterial isolates

This study included a total of 740 *M. tuberculosis* complex isolates; 193 isolates were obtained during the 2nd Nationwide Survey on *M. tuberculosis* Drug Resistance that was conducted in São Paulo state between 2006 and 2008 and 547 isolates were obtained from a laboratory based study of drug resistance analyzed by the Mycobacteria Reference Laboratory at Institute Adolfo Lutz during the same time period.

2.2. Drug susceptibility testing

The BD BACTEC-Mycobacteria Growth Indicator Tube MGIT960 system (MGIT 960)™ was used to establish the susceptibility patterns to I (0.1 μg/ml), R (1.0 μg/ml), ethambutol – E (5.0 μg/ml) and streptomycin – S (1.0 μg/ml). For each isolate, 0.5 ml of bacterial inoculum was added into four MGIT 960 tubes containing the drugs and one growth control MGIT 960 tube without drug. All of the inoculated tubes were placed into the fluorometric BACTEC MGIT 960 system and the instrument reported the susceptibility results automatically. (Giampaglia et al., 2007) As previously reported, the susceptibility to pyrazinamide (P) was determined using a pyrazinamidase test with a medium containing pyrazinamide 0.100 g/L. After de bacterial growth in the medium, 1 ml of 1% freshly prepared ferrous ammonium sulfate was added in the tube. A reading was made after 15 min, a pink color indicate pyrazinamidase positive test (OPAS–OMS, 1986).

2.3. Spoligotyping and database analysis

Spoligotyping was performed using a previously published method that is based on DR locus polymorphisms (Kamerbeek et al., 1997). Briefly, the entire DR locus in the *M. tuberculosis* genome was amplified with the Dra and Drb primers, and the Dra primer was biotinylated at the 5′ end. The amplified products were hybridized to a set of 43 oligonucleotides that were covalently bound to a nylon membrane. The hybridized DNA was detected using chemiluminescence, the presence of a spacer was visualized on X-ray film, and spoligotypes were recorded in a binary format as black (positive hybridization) or empty (negative hybridization) squares. For database analysis, spoligotypes were converted to octal codes and entered into the SITVIT2 proprietary database at the Pasteur Institute of Guadeloupe, which is an updated version of the previously released SpolDB4 database (Brudey et al., 2006). During the present study, SITVIT2 database contained genotype information on approximately 87,000 *M. tuberculosis* clinical isolates from 160 countries. In this database, SIT (Spoligotype International Type) designates spoligotypes that are shared by two or more patient isolates as opposed to "orphan" isolates, which are single isolates. The major phylogenetic clades were assigned according to the signatures provided in the SpolDB4 database, which defined 62 genetic lineages/sub-lineages; these include specific signatures for various *M. tuberculosis* complex (MTC) species such as *Mycobacterium bovis*, *Mycobacterium microti*, *Mycobacterium caprae*, *Mycobacterium pinipedii* and *Mycobacterium africanum* and rules that define the major lineages/sublineages for *M. tuberculosis* sensu stricto. The latter include the Beijing clade, the Central Asian (CAS) clade and two sublineages, the East African–Indian (EAI) clade and nine sublineages, the Haarlem (H) clade and three sublineages, the Latin American–Mediterranean (LAM) clade and 12 sublineages, the "Manu" family and three sublineages, the S clade, the IS*6110*-low-banding X clade and four sublineages and an ill-defined T clade (defined by default) and five sublineages.

2.4. Distinction between ancestral, ancient, and modern lineages

The lineages were tentatively linked to evolutionary "ancient" and "recent" lineages of tubercle bacilli as defined by the Principal Genetic groups (PGG), which are based on *KatG463–gyrA95* polymorphisms (Sreevatsan et al., 1997) that were inferred from the reported linking of specific spoligotype patterns to PGG1, -2 or -3 (Soini et al., 2000; Rastogi and Sola, 2007). Furthermore, based on the previously reported presence or absence of an *M. tuberculosis* specific deletion 1 (TbD1, a 2-kb sequence), the lineages were classified as "ancestral" (TbD1+) or "modern" TbD1-negative (TbD1$^-$) MTC isolates. This classification superimposes well with the previous PGG classification (Brosch et al., 2002). As reviewed recently, all MTC species other than *M. tuberculosis* sensu stricto and the EAI lineage are TbD1+; hence, only two groups of MTC are considered to be the classical human pathogens. The EAI and the *M. africanum* lineages are considered ancestral (Rastogi and

Sola, 2007). Thus, the MTC isolates that were used in this study were tentatively identified as TbD1+/PGG1 (ancestral: *M. africanum* and the EAI lineage of *M. tuberculosis*), TbD1$^-$/PGG1 (ancient: Beijing, CAS), and TbD1$^-$/PGG 2–3 (modern/evolutionary recent: LAM, T, S, X, Haarlem). Last but not least, as described recently (Gagneux et al., 2006; Gagneux and Small, 2007), the latter subgroup (TbD1$^-$/PGG 2–3) was also found to superimpose perfectly with the "Euro-American" phylogenetic group based on large sequence polymorphisms (LSP). These terms are used throughout this paper to denote the lineages observed in the present study.

2.5. Quality control

M. tuberculosis H37Rv (ATCC 27294) and *M. bovis*–BCG (obtained from the culture collection of the Mycobacteria Reference Laboratory at the Institute Adolfo Lutz, São Paulo state, Brazil) were included as quality controls for the phenotypic and genotypic tests.

3. Results

This study included 740 isolates, 483 (65.3%) of which were obtained from the Metropolitan São Paulo area, 125 (16.9%) of which were obtained from 64 interior towns, 58 (17.8%) were obtained from metropolitan Baixada Santista area and 74 (10.0%) of which were obtained from the prisons in São Paulo state. In addition to a high proportion of resistant isolates from the laboratory-based studies (547–73.9%), we also analyzed 193 (26.1%) isolates from the 2nd Nationwide Survey of Drug-Resistant TB in São Paulo.

Among the isolates from the laboratory based studies of resistance, 308 (56.3%) isolates were detected in the metropolitan São Paulo, 173 (31.6%) in the interior towns and 66 (12.1%) were detected in the prisons of São Paulo state. The majority of the isolates obtained during the 2nd Nationwide Survey of Drug Resistance were from the metropolitan São Paulo area [175 (90.7%)], and the remaining isolates were from interior towns [10 (5.2%)] and the prisons in São Paulo state [8 (4.1%)].

When the patterns obtained from the spoligotyping analysis of these isolates were compared with those described in the SITVIT2 database, a total of 51/740 (6.9%) of the isolates corresponded to orphan patterns (Supplemental Table S1). Lineage interpretations for these orphan patterns were performed manually by expert-based interpretations using revised SpolDB4 rules; this method designated the lineages of these orphan isolates as follows: LAM $n = 21$, T $n = 14$ (including 1 isolate from the T5-Madrid2 sublineage), Haarlem (H) $n = 4$, S family $n = 1$, EAI (EAI5 sublineage) $n = 1$, and unknown signatures, $n = 10$ (Supplemental Table S1). The remaining 689/740 (93.1%) of the isolates were distributed into the 144 shared types including 119 patterns that matched a preexisting shared type in the SITVIT2 database and 25 that were newly created (Supplemental Table S2); 77/144 of the patterns corresponded to unique isolates, while the remaining 67 patterns corresponded to clustered patterns ($n = 612$ isolates clustered into groups of 2–84 isolates each). Thus, the total level of clustering corresponded to 612/740 strains, or 82.7% of all isolates. The analysis of these isolates by shared types/sublineages showed that only four sublineages contained >5% of the isolates [SIT50/H3 ($n = 84$, 11.35%), SIT 53/T1 ($n = 62$, 8.38%), SIT 42/LAM9 ($n = 48$, 6.49%), and SIT17/LAM2 ($n = 46$, 6.22%)].

When the isolates were analyzed by lineage (including the orphans), the following distribution pattern was observed for the 740 isolates: Beijing ($n = 7$, 0.95%), CAS1-DEL ($n = 1$, 0.13%), EAI3-IND ($n = 4$, 0.54%), Haarlem ($n = 153$, 20.68%), LAM ($n = 331$, 44.73%), PINI2 ($n = 1$, 0.13%), S ($n = 11$, 1.49%), T ($n = 148$, 20%), X ($n = 24$, 3.24%), and unknown ($n = 60$, 8.11%). The evolutionarily ancient PGG1 lineages (Beijing, CAS1-DEL, EAI3-IND, and PINI2) were rarely detected and made up only 13/740, or 1.76%, of the isolates; all of the remaining 727/740, or 98.24%, of the MTC isolates in São Paulo state belonged to the recent evolutionary PGG2/3 isolate lineages LAM, T, S, X, and Haarlem, i.e., the Euro-American group (see Supplemental Table S3 for details).

The most predominant lineages and shared types in our study were the following: H1/SIT47 $n = 24$, H3/SIT50 $n = 84$, LAM1/SIT20 $n = 20$, LAM2/SIT17 $n = 46$, LAM3/SIT33 $n = 36$, LAM4/SIT60 $n = 23$, LAM6/SIT64 $n = 31$, LAM9/SIT42 $n = 48$, T1/SIT53 $n = 62$, T1/SIT51 $n = 21$, and X2/SIT317 $n = 13$. Thus, very high spoligotype diversity was documented for the LAM lineage. Out of the 12 sublineages that have been reported worldwide for the LAM family (Brudey et al., 2006), a total of eight sublineages were observed in our recruitment sample. Similarly, all of the T sublineages that have been reported were present in our study sample.

Lastly, the 25 newly created SITs that matched another orphan patterns in the database or the SITs that were composed of two or more strains belonging to a new pattern within our study were as follows (Supplemental Table S2): SITs 3113, 3114, 3116, 3117, and 3119–3123 contain two or more isolates in the present study, as opposed to SITs 3112, 3115, 3144, 3147, and 3150, which contain 1–3 isolates in the present study and matched an orphan pattern from some region in Brazil. The remaining shared types (SITs 3124–3128, 3142, 3143, 3145, 3146, 3148, and 3149) contained 1–2 isolates in the present study and matched an orphan pattern from another country (see Supplemental Table S2 for details).

Table 1 presents the description of clusters containing seven or more isolates, 15 SITS (453 isolates – 61.2%), and their worldwide distribution in the SITVIT2 database. The distribution ($\geqslant 3\%$) of these SITs by countries showed that 14 of them had already been described in Brazil, with the exception of SIT 1-Beijing. The majority (53.0%) of clustered isolates belonged to four SITs: 50-H3, 53-T1, 42-LAM9 and 17-LAM2.

Table 2 shows that the drug susceptibility pattern of clustered spoligotypes correlated with gender and age. The mean age of the patients ranged from 31.0 years (SIT 4-unknown) to 46.0 years (SIT 137-X2). The male/female sex ratio observed ranged from 1.2 (SIT 1905-T1) to 12.0 (SIT 137-X2).

The drug susceptibility pattern of clustered isolates showed 349 (77.1%) drug-resistant isolates that contained a significant percentage of MDR (31.1%) and mono-resistant isolates (40.4%). In the following five SITs, MDR isolates made up more than 40.0% of the SIT: SIT 1-Beijing, SIT 4-unknown, SIT 60-LAM 4, SIT 177-LAM 9, and SIT 137-X2.

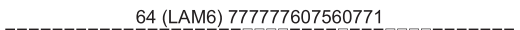
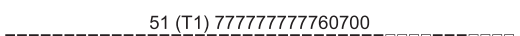
SIT 51-T1 had the lowest percentage of MDR isolates (4.8%) but the highest percentage of mono-resistant isolates (71.4%), primarily with resistance to S. Over 45.0% of the isolates in the following three SITs were mono-resistant: SIT 1905-T1 63.6% (only I), SIT 53-T1 51.6% (I = 16.1%, S = 30.6%), and SIT 47-H1 45.8% (I = 25.0%, S = 8.3%, R = 8.3%).

The 25 newly created SITs included 46 (6.2%) isolates. Twelve (26.1%) of the isolates were susceptible to all drugs tested, 9 (19.6%) of the isolates were MDR, 20 (43.5%) of the isolates were mono-resistant (10 to S, 7 to I, 2 to R, 1 to E) and 5 (10.9%) of the isolates were resistant to other drugs. Among the 13 (52.0%) new SITs (presenting 2–5 isolates), the designated lineages were LAM-12 isolates, T and S – 4 isolates each, H3 – 2 isolates and unknown – 12 isolates. The resistance profile contained the following six MDR isolates: SIT 3113-H3 (1), SIT 3114-LAM 4 (1), SIT 3119-LAM2 (2), SIT 3122-S (1), and SIT 3124-LAM3 (1).

The 51 orphan isolates contained 10 (19.6%) isolates that were susceptible to all drugs tested, 8 (15.7%) isolates that were MDR, 28 (54.9%) isolates that were mono-resistant (19 to I, 6 to S, 2 to R, and 1 to P) and 5 (9.8%) isolates were resistant to other drugs.

The analysis of 547 isolates from laboratory based study of resistance showed that 11 isolates showed no resistance to the

Table 1
Description of clusters containing seven or more isolates in this study, and their worldwide distribution in the SITVIT2 database (predominant shared-types are shown in decreasing order of their occurrence).

SIT (Lineage) octal number spoligotype description[a]	Number (%) in study	% In study versus database	Distribution in regions with ⩾3% of a given SIT[b]	Distribution in countries with ⩾3% of a given SITs[b]
50 (H3) 777777777720771	84 (11.35)	2.9	AMER-N 21.16, AMER-S 18.85, EURO-W 14.8, EURO-S 12.17, EURO-E 6.22, AFRI-N 4.98, AFRI-S 4.74, CARIB 3.98, ASIA-W 3.25	USA 20.54, BRA 8.16, AUT 7.12, ESP 6.36, ITA 5.08, ZAF 4.74, PER 4.32, CZE 4.3
53 (T1) 777777777760771	62 (8.38)	1.25	AMER-N 19.19, AMER-S 15.62, EURO-W 12.38, EURO-S 9.68, ASIA-W 8.4, AFRI-S 6.28, EURO-N 4.81, AFRI-N 4.45, AFRI-E 3.79, EURO-E 3.26	USA 16.69, BRA 6.34, ZAF 6.14, ITA 4.95, TUR 4.4, AUT 4.33, PER 3.36, SAU 3.16, FXX 3.0
42 (LAM9) 777777607760771	48 (6.49)	1.69	AMER-S 29.92, AMER-N 15.39, EURO-S 11.75, AFRI-N 10.13, EURO-W 6.46, AFRI-E 3.4, AFRI-S 3.74	USA 14.01, BRA 12.0, MAR 8.33, COL 7.23, ITA 6.18, ESP 3.95, VEN 3.92, ZAF 3.74
17 (LAM2) 677737607760771	46 (6.22)	7.3	AMER-S 60.64, AMER-N 19.52, CARIB 8.41, EURO-S 5.71	BRA 29.21, VEN 28.41, USA 18.9, ESP 4.6, HTI 3.49
33 (LAM3) 776177607760771	36 (4.86)	3.36	AFRI-S 30.35, AMER-S 24.46, AMER-N 15.22, EURO-S 12.9, EURO-W 5.14, AMER-C 4.58	ZAF 30.35, USA 14.85, BRA 12.05, ESP 8.31, ARG 5.32, PER 5.23, HND 4.02, ITA 3.92
64 (LAM6) 777777607560771	31 (4.19)	9.31	AMER-S 51.95, AMER-N 29.43, EURO-S 5.41, EURO-W 3.9	BRA 41.14, USA 28.23, GUF 5.71, PRT 4.2
20 (LAM1) 677777607760771 	26 (3.51)	3.49	AMER-S 28.63, AMER-N 23.12, AFRI-S 11.96, EURO-S 10.62, EURO-W 7.66, CARIB 5.78, AFRI-E 4.7, AFRI-W 3.23	USA 21.37, BRA 18.95, NAM 8.33, PRT 6.59, VEN 5.65, ZAF 3.63, HTI 3.5, ESP 3.36
47 (H1) 777777774020771	24 (3.24)	1.9	AMER-N 20.63, EURO-W 19.53, EURO-S 13.36, AMER-S 12.81, EURO-E 8.14, AFRI-N 4.51, ASIA-W 4.27, EURO-N 4.11, AFRI-M 3.16	USA 18.81, AUT 9.96, BRA 8.38, ITA 7.04, CZE 4.66, ESP 4.43, MAR 3.48, NLD 3.08
60 (LAM4) 777777607760731	23 (3.11)	5.88	AFRI-S 41.69, AMER-S 22.0, AFRI-N 7.67, AFRI-W 5.88, AMER-N 5.63, EURO-S 4.6, EURO-W 4.35	ZAF 41.69, BRA 14.58, USA 4.86, MAR 4.86, VEN 3.84, GMB 3.33
51 (T1) 777777777760700	21 (2.84)	8.64	AMER-S 26.34, AMER-N 18.11, EURO-S 14.4, EURO-W 13.17, CARI 12.76, ASIA-SE 8.23	USA 16.87, BRA 16.46, ITA 11.52, AUT 9.88, MYS 7.82, HTI 7.0, GLP 4.53, VEN 4.12, GUF 3.7
137 (X2) 777776777760601 	13 (1.76)	1.31	AMER-N 74.87, EURO-N 7.34, AFRI-S 6.53, AMER-S 4.62	USA 73.57, ZAF 6.53, GBR 5.53, BRA 4.32
4 (Unknown) 000000007760771	12 (1.62)	4.17	EURO-S 16.32, AMER-N 15.97, AMER-S 15.63, ASIA-W 11.11, EURO-W 10.76, AFRI-S 10.42, AFRI-E 5.56, EURO-E 4.17, EURO-N 3.47	USA 12.5,ZAF 10.42, TUR 10.07, ITA 9.03, BRA 9.03, BGR 3.82, CAN 3.47, BEL 3.13
1905 (T1) 777777777460771	11 (1.49)	55	AMER-S 95.0, EURO-S 5.0	BRA 85.0, PER 5.0, ITA 5.0, ARG 5.0
177 (LAM9) 377777607760771	9 (1.22)	11.39	AMER-S 60.76, EURO-S 17.72, AFRI-N 6.33, ASIA-W 3.8, AMER-N 3.8	BRA 49.37, ESP 11.39, ITA 6.33, USA 3.8, SAU 3.8, PRY 3.8, MAR 3.8
1 (Beijing) 000000000003771	7 (0.95)	0.11	AMER-N 30.23, ASIA-SE 13.68, AFRI-S 12.451, ASIA-E 11.01, ASIA-N 8.22, ASIA-S 5.13, ASIA-W 3.89,EURO-N 3.19	USA 29.79, ZAF 12.45, RUS 8.22, JPN 8.05, VNM 5.86, MYS 3.60

[a] SIT and lineage designations are shown following SITVIT2 proprietary database of Institut Pasteur de la Guadeloupe. Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database. The definition of macro-geographical regions and sub-regions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in four sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia.

[b] The three letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ⩾5% of a given SITs as compared to their total number in the SITVIT2 database.

Table 2
Drug susceptibility patterns of clusters containing seven or more *M. tuberculosis isolates*, correlated with gender and age of the patients from São Paulo state, Brazil.

SIT (lineage) – No. of isolates	Drug resistant patterns, No. (%)						Gender, No. (%)		Age (years)	
	MDR	Monoresistance			Other-resistance	Susceptible	Male	Female	Mean	Median
		I	R	S						
1 (Beijing) – 7	3 (42.8)	1 (14.3)	0	1 (14.3)	0	2 (28.6)	5 (71.4)	2 (28.6)	41.0 (25–58)	37.0
177 (LAM9) – 9	4 (44.4)	1 (11.1)	0	1 (11.1)	0	3 (33.3)	7 (77.8)	2 (22.2)	38.0 (26–50)	38.0
1905 (T1) – 11	3 (27.3)	7 (63.6)	0	0	0	1 (9.1)	6 (54.5)	5 (45.5)	38.0 (17–54)	51.0
4 (U) – 12	7 (58.3)	4 (33.3)	0	0	0	1 (8.3)	8 (66.7)	4 (33.3)	31.0 (20–51)	29.0
137 (X2) – 13	6 (46.1)	4 (30.8)	0	2 (15.4)	0	1 (7.7)	12 (92.3)	1 (7.7)	46.0 (20–71)	48.0
51 (T1) – 21	1 (4.8)	3 (14.3)	0	12 (57.1)	1 (4.8)	4 (19.0)	16 (76.2)	5 (23.8)	40.0 (25–66)	41.0
60 (LAM4) – 23	15 (65.2)	5 (21.7)	0	0	0	3 (13.0)	13 (56.5)	10 (43.5)	33.0 (10–62)	31.0
47 (H1) – 24[a]	4 (16.7)	6 (25.0)	2 (8.,3)	2 (8.3)	3 (12.5)	6 (25.0)	18 (75.0)	6 (25.0)	34.3 (20–54)	33.0
20 (LAM1) – 26	6 (23.1)	4 (15.4)	2 (7.7)	3 (11.5)	3 (11.5)	8 (30.8)	18 (69.2)	8 (30.8)	39.8 (21–69)	40.0
64 (LAM6) – 31	9 (29.0)	11 (35.5)	0	1 (3.2)	1 (3.2)	9 (29.0)	20 (64.5)	11 (34.5)	38.3 (16–74)	37.0
33 (LAM3) – 36	11(30.5)	4 (11.1)	3 (8.3)	7 (19.4)	2 (5.5)	9 (25.0)	20 (55.6)	16 (44.4)	39.7 (17–70)	40.0
17 (LAM2) – 46	18 (39.1)	14 (30.4)	0	3 (6.5)	1 (2.2)	10 (21.7)	36 (78.3)	10 (21.7)	36.5 (15–76)	33.0
42 (LAM9) – 48	13 (27.1)	7 (14.6)	2 (4.2)	5 (10.4)	6 (12.5)	15 (31.2)	32 (66.7)	16 (33.3)	36.5 (12–63)	35.5
53 (T1) – 62[a]	10 (16.1)	10 (16.1)	2 (3.2)	19 (30.6)	3 (4.8)	17 (27.4)	44 (71.0)	18 (29.0)	35.0 (15–66)	34.0
50 (H3) – 84[b]	31 (36.9)	16 (19.0)	2 (2.4)	14 (16.7)	5 (5.9)	15 (17.8)	65 (77.4)	19 (29.6)	36.3 (16–64)	34.0

Abbreviations: I, isoniazid; R, rifampin; SM, streptomycin; MDR, multidrug-resistant.
[a] One isolate with monoresistance to ethambutol.
[b] One isolate with monoresistance to pyrazinamide.

drugs tested. The 536 remaining isolates showed 210 (39.2%) MDR isolates with the following profiles: IR – 94 (17.5%), IRS – 34 (6.3%), IRP – 26 (4.8%), IRE – 19 (3.5%), IRPS – 15 (2.8%), IRES – 8 (1.5%), IRPE – 8 (1.5%) and IRPES – 6 (1.1%). The 281 mono-resistant isolates presented I – 134 (47.7%), S – 109 (38.8%), R – 28 (9.9%), E and P – 5 (1.8% each). The other-resistances were composed by 45 (8.4%) isolates with the following profiles: double resistance IS – 25 (4.7%), IP – 6 (2.1%), IE – 5 (1.8%), RS – 2 (0.7%), RP – 1 (0.3%) and triple resistance IES – 4 (1.4%), IPS – 2 (0.7%).

Among resistant isolates, 321 (59.9%) were found in the clustered SITS and 134 (25.0%) of them were MDR isolates and showed the following profiles IR – 57 (42.5%), IRS – 26 (19.4), IRP – 21 (15.7%), IRPS – 11 (8.2%), IRPE – 6 (4.5%), IRPES – 5 (3.7%), IRE and IRES – 4 (3.0%) each.

Isolates from the 2nd Nationwide Survey of Drug-Resistant TB in São Paulo included 156 (80.8%) susceptible isolates and 37 (19.2%) isolates with various resistant profiles. Twenty-seven isolates (73.0%) were mono-resistant [I-14 (51.8%), S-8 (29.6), E-2 (7.4%), P-2 (7.4%), R-1 (3.7%)] eight (21.6%) were double resistant [IR-2, ES-1, and IS-5] and two (5.4%) isolates were triple resistant [IRP-1, and IES-1)]. The SITs containing the three MDR isolates were 42, 91 (IR) and orphan (IRP).

4. Discussion

This study provides the first overview of circulating genotypes of drug-sensitive and drug-resistant *M. tuberculosis* in São Paulo state and shows the predominance of the LAM lineage, which accounts for 44.5% of the isolates. This percentage is similar to that described in South America (45.0%) by Ritacco et al. (2011) and to the average obtained by Gomes et al. (2011) from 11 states in Brazil (46.2%). However, it differs from that described by other studies in the same country. Mendes et al. (2011) reported a frequency of 36.1% of LAM among 93 isolates from the metropolitan São Paulo area, and Gomes et al. (2011), who provided the first insight into the population structure of TB isolates in Brazil, reported that the lowest percentage of the LAM lineage (35.0%) had been observed among 180 isolates from São Paulo state. Noguti et al. (2010) showed the lowest percentage (34.7%) of this lineage in Maringá (93 isolates), and Miranda et al. (2011) showed the highest percentage (55.3%) of this lineage in Minas Gerais (114 isolates).

In this study, the H and T lineages accounted for 21.5% and 18.7% of the 740 isolates, respectively, which is different from the rates described by Mendes et al. (2011), who studied isolates from the metropolitan São Paulo region and found that T accounted for 33.0% and H accounted for 15.3% of the isolates. However, our results are similar to those described by Noguti et al. (2010), who studied isolates from Maringá and found that H accounted for 22.5% of the isolates and T accounted for 15.3% of the isolates. Other studies also showed a higher percentage of the T lineage when compared to the H lineage. In particular, Miranda et al. (2011) described the lowest percentage of the H lineage (7.0%) in Minas Gerais.

Although SIT 1905-T1 accounted for a relatively low percentage (1.49%) of clustered isolates in our study, it showed a high phylogenetical specificity for São Paulo since 55.0% of all SIT1905 isolates in the SITVIT2 databases were from the present study; further countrywide distribution showed that 85.0% of such isolates were described from Brazil (Table 1).

As cited by Ritacco et al. (2011) the frequency of the Beijing genotype was found to be below 1.0% in most of the surveyed South American countries. This study showed a low percentage of this lineage in São Paulo state, which is similar to the results observed in previous studies from other states in Brazil. Nevertheless, Mendes et al., (2011) reported that the Beijing genotype comprised 5.2% of the lineages in the metropolitan São Paulo area.

A group of SITs found in Brazil at a very low frequency could represent imported genotypes due to the international migratory flux (Gomes et al., 2011). Some genotypes were observed at a very low frequency in our study, six had already been described in Brazil, including SIT 11 (EA13-IND), SIT 18 (X2), SIT 58 (T5-Madrid2), SIT 254 (T5-RUS1), SIT 863 (PIN12), and SIT 1983 (EA13-IND). However, three of the SITs were described for the first time in this study, including SIT 22 (CAS1-Delhi), SIT 39 (T4-CEU1), and SIT 209 (LAM 12-Madrid1).

The results from laboratory based study showed a higher percentage of some clusters in the metropolitan São Paulo area when compared to the interior towns: SIT 53 (56.8%), SIT 1 and 17 (60%), SIT 20 (64.7%), SITs 137, 177, 1905 (66.7%), and SIT 64 (71.4%) and SIT 60 (85.7%).

The incidence of TB in the metropolitan Baixada Santista area is twice the average of the incidence in the interior towns. Isolates from this area comprised 40.0% of the isolates in SIT 33 and 18.2% of the isolates in SIT 50. However, despite this, the laboratory-based study showed no isolates from SITS 1, 42, 177, and 1905.

We observe that SIT 137 included isolates detected primarily in males (92.3%) with mean age of 40 years, but detailed patient records were not available to conclude the association of clustering with gender.

On the other hand, the analysis of 193 isolates from the 2nd Nationwide Survey showed that 74 (47.4%) of the susceptible isolates were distributed among clusters composed of seven or more isolates. No resistant isolates were found in the clustered SIT 1, 60, 137 and 177. Eighteen (48.6%) resistant isolates were distributed in 10 SITs that were primarily mono-resistant to I (8 isolates) and S (5 isolates). Two isolates that were mono-resistant to I were detected in SIT 50, and two isolates resistant to S were detected in SIT 51. Only one MDR isolate was included in the clustered SIT 42-IR. The MDR-TB rate of 1.5%, as reported by the WHO (2011) for Brazil, suggests low transmission of MDR strains among newly diagnosed patients. Among the three MDR isolates, two were from the interior towns of Santos (SIT 42) and São José dos Campos (orphan isolate) and one was from prison (SIT 91).

In a nutshell, a comparison of the spoligotyping patterns obtained in our study with those described in the SITVIT2 database as well as with similar data presented in other Brazilian studies highlights: (i) the predominance of LAM lineage, (ii) the high phylogenetical specificity of a minor clone for São Paulo area (SIT1905-T1; 1.49% of all isolates in our study which represented 55% of all SIT1905 isolates in the SITVIT2 database), (iv) overall similar percentages of resistant and susceptible isolates among clustered SITs, (v) a higher percentage of some clusters in the metropolitan São Paulo area when compared to the interior towns, e.g., the incidence of TB in Santos is twice the average of the incidence in the interior towns (with a predominance of SIT 137 in males, 92.3%; with a mean age of 40). Nonetheless, some limitations of our study need to be acknowledged when interpreting the results. The study was performed with *M. tuberculosis* isolates obtained from the 2nd Nationwide Survey of Drug Resistant TB in São Paulo and high proportion of resistant isolates from the laboratory-based studies which can cause a selection bias. In São Paulo the culture for mycobacteria and susceptibility testing is performed according to criteria defined by Program to Control Tuberculosis such as history of MDR-TB contact, previous treatment, homeless, prisoner, occupational healthy and of HIV patients. Despite these limitations, our results are consistent with the literature, and they allowed showing the population structure of *M. tuberculosis* clinical isolates of São Paulo.

An additional comment may be made regarding the choice of spoligotyping used alone for genotyping that may not be sufficient to make definite conclusions about the probable subdivision of the clusters containing seven or more *M. tuberculosis* isolates spread in São Paulo state (Table 2), including both resistant and susceptible isolates, as well as acquisition of drug-resistance during treatment. Indeed, the results obtained with the isolates from laboratory-based studies of resistance showed the diversity of *M. tuberculosis* genotypes causing resistant tuberculosis with high percentage of isolates with resistance to I, S and MDR distributed in the clustered SITs. Among the MDR isolates, 42.5% showed resistance only to IR and 57.5% showed resistance to one or more drugs in addition to IR, which suggests the acquisition of resistance during treatment. In this regard, it is worth mentioning that the isolates from the 2nd Nationwide Survey showed a similar percentage of resistant and susceptible isolates distributed in the clustered SITs. As 90% of these isolates were obtained from new cases (according to principles of patient recruitment in this global project); this result may indicate that both resistant and susceptible isolates are now circulating among newly infected patients. However, only future studies using MIRU-typing would allow to split some of the big spoligotyping defined clusters found in this study, and shed new light not only on genotypes with potential epidemiological links, but also to pinpoint the proportion of primary drug resistance in newly-infected TB patients versus secondary drug-resistance that is emerging due to a lack of observance during treatment.

Acknowledgements

This investigation was supported by the Brazilian Research Council/CNPq (Proc. No. 410480/2006-1). We also thank the ICO-HRTA AIDS/TB, FIC/NIH#5U2R TW006883-02 projects for support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2012.10.015.

References

Araújo-Filho, J.A., Vasconcelos Jr, A.C., Sousa, E.M., Silveira, C., Ribeiro, E., Kipnis, A., Junqueira-Kipnis, A.P., 2008. Extensively drug-resistant tuberculosis: a case report and literature review. Braz. J. Infect. Dis. 12, 447–452.

Bombarda, S., Galesi, V.M.N., 2006. Tuberculose no Estado de São Paulo. Bol. Pneumol. Paulista. 31, 7–9.

Borsuk, S., Dellagostin, M.M., Madeira, S.G., Lima, C., Boffo, M., Mattos, I., Almeida da Silva, P.E., Dellagostin, O.A., 2005. Molecular characterization of *Mycobacterium tuberculosis* isolates in a region of Brazil with a high incidence of tuberculosis. Microbes Infect. 7, 1338–1344.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D., Cole, S.T., 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc. Natl. Acad. Sci. U.S.A. 99, 3684–3689.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaïa, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rüsch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

CVE – Centro de Vigilância Epidemiológica São Paulo, 2011. Divisão de Controle da Tuberculose. <http://www.cve.saude.sp.gov.br/htm/tb/tb_num/tb_cn10.htm> (accessed 29.09.11).

Dalcolmo, M.P., Andrade, M.K.N., Picon, P.D., 2007. Multiresistant tuberculosis in Brazil: history and control. Rev. Saúde Pública. 41 (Suppl. 1), 34–42.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006. Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U.S.A. 103, 2869–2873.

Gagneux, S., Small, P.M., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect. Dis. 7, 328–337.

García de Viedma, D., Mokrousov, I., Rastogi, N., 2011. Innovations in the molecular epidemiology of tuberculosis. Enferm. Infecc. Microbiol. Clin. 29 (Suppl. 1), 8–13.

Giampaglia, C.M., Martins, M.C., Vieira, C.B., Vinhas, S.A., Telles, M.A., Palaci, M., Marsico, A.G., Hadad, D.J., Mello, F.C., Fonseca, L.S., Kritski, A., 2007. Multicentre evaluation of an automated BACTEC 960 system for susceptibility testing of *Mycobacterium tuberculosis*. Int. J. Tuberc. Lung Dis. 11, 986–991.

Gomes, H.M., Elias, A.R., Oelemann, M.A., Pereira, M.A., Montes, F.F., Marsico, A.G., Kritski, A., Filho, L.D., Caldas, P.C., Possuelo, L.G., Cafrune, P., Rossetti, M.L., Lucena, N., Saad, M.A., Cavalcanti, H.R., Leite, C.Q., Brito, R.C., Lopes, M.L., Lima, K., Souza, M., Trindade, R.D., Zozio, T., Sola, C., Rastogi, N., Suffys, P.N., 2011. Spoligotypes of *Mycobacterium tuberculosis* complex isolates from patients residents of 11 states of Brazil. Infect. Genet. Evol. (Epub 2011 Sep 1).

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., Van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Mendes, N.H., Melo, F.A., Santos, A.C., Pandolfi, J.R., Almeida, E.A., Cardoso, R.F., Berghs, H., David, S., Johansen, F.K., Espanha, L.G., Leite, S.R., Leite, C.Q., 2011. Characterization of the genetic diversity of *Mycobacterium tuberculosis* in São Paulo city, Brazil. BMC Res. Notes 29, 269.

Miranda, S.S., Carvalho, W.S., Suffys, P.N., Kritski, A.L., Oliveira, M., Zarate, N., Zozio, T., Rastogi, N., Gicquel, B., 2011. Spoligotyping of clinical *Mycobacterium tuberculosis* isolates from the state of Minas Gerais, Brazil. Mem. Inst. Oswaldo Cruz 106, 267–273.

Noguti, E.N., Leite, C.Q., Malaspina, A.C., Santos, A.C., Hirata, R.D., Hirata, M.H., Mamizuka, E.M., Cardoso, R.F., 2010. Genotyping of *Mycobacterium tuberculosis* isolates from a low-endemic setting in northwestern state of Paraná in Southern Brazil. Mem. Inst. Oswaldo Cruz 105, 79–85.

OPAS–OMS, 1986. Manual de Normas y procedimientos tecnicos para la Bacteriologia de la tuberculosis. Parte III. Sensibilidad del *Mycobacterium tuberculosis* a las drogas. La identificatión de micobactérias. Nota técnica 28. Centro Panamericano de Zoonosis, Buenos Aires, Argentina.

Rastogi, N., Sola, C. 2007. Molecular evolution of the *Mycobacterium tuberculosis* complex. In: Palomino, J.C., Leao, S., Ritacco, V. (Eds.), Tuberculosis 2007: From Basic Science to Patient Care. Amedeo Online Textbooks, pp. 53–91, <http://www.tuberculosistextbook.com/index.htm>.

Ritacco, V., Iglesias, M.J., Ferrazoli, L., Monteserin, J., Dalla Costa, E.R., Cebollada, A., Morcillo, N., Robledo, J., de Waard, J.H., Araya, P., Aristimuño, L., Díaz, R., Gavin, P., Imperiale, B., Simonsen, V., Zapata, E.M., Jiménez, M.S., Rossetti, M.L., Martin, C., Barrera, L., Samper, S., 2011. Conspicuous multidrug-resistant *Mycobacterium tuberculosis* cluster strains do not trespass country borders in Latin America and Spain. Infect. Genet. Evol. (Epub 2011 Jun 21).

Soini, H., Pan, X., Amin, A., Graviss, E.A., Siddiqui, A., Musser, J.M., 2000. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J. Clin. Microbiol. 38, 669–676.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser, J.M., 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. U.S.A. 94, 9869–9874.

van Embden, J.D., Cave, D.T., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., Small, P.M., 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31, 406–409.

World Health Organization, 1997. Anti-tuberculosis drug resistance in the world. The WHO/IUATLD Global Project on anti-tuberculosis drug resistance surveillance, 1994–1997. World Health Organization, Geneva, Switzerland WHO/TB/97.229.

World Health Organization, 2011. WHO Report 2011. Global tuberculosis control. Annex 2 country profiles (22 high-burden countries). <http://who.int/tb/publications/global_report/en/index.html> (accessed 04.11.11).

OPEN ACCESS Freely available online

PLoS one

The Guinea-Bissau Family of *Mycobacterium tuberculosis* Complex Revisited

Ramona Groenheit[1,2], Solomon Ghebremichael[1], Jenny Svensson[1], Paulo Rabna[3,4], Raffaella Colombatti[5,6], Fabio Riccardi[7], David Couvin[8], Véronique Hill[8], Nalin Rastogi[8], Tuija Koivula[1,9], Gunilla Källenius[9]*

1 Department of Preparedness, Swedish Institute for Infectious Disease Control, Solna, Sweden, **2** Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden, **3** Laboratório Nacional de Saúde Pública, Bissau, Guinea-Bissau, **4** Bandim Health Project, Indepth Network, Bissau, Guinea-Bissau, **5** Hospital "Raoul Follereau", Bissau, Guinea-Bissau, **6** Department of Pediatrics, Azienda Ospedaliera-Università di Padova, Padova, Italy, **7** Department of Public Health, University of "Tor Vergata", Rome, Italy, **8** WHO Supranational TB Reference Laboratory, Tuberculosis and Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Guadeloupe, France, **9** Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden

Abstract

The Guinea-Bissau family of strains is a unique group of the *Mycobacterium tuberculosis* complex that, although genotypically closely related, phenotypically demonstrates considerable heterogeneity. We have investigated 414 *M. tuberculosis* complex strains collected in Guinea-Bissau between 1989 and 2008 in order to further characterize the Guinea-Bissau family of strains. To determine the strain lineages present in the study sample, binary outcomes of spoligotyping were compared with spoligotypes existing in the international database SITVIT2. The major circulating *M. tuberculosis* clades ranked in the following order: AFRI (n = 195, 47.10%), Latin-American-Mediterranean (LAM) (n = 75, 18.12%), ill-defined T clade (n = 53, 12.8%), Haarlem (n = 37, 8.85%), East-African-Indian (EAI) (n = 25, 6.04%), Unknown (n = 12, 2.87%), Beijing (n = 7, 1.68%), X clade (n = 4, 0.96%), Manu (n = 4, 0.97%), CAS (n = 2, 0.48%). Two strains of the LAM clade isolated in 2007 belonged to the Cameroon family (SIT61). All AFRI isolates except one belonged to the Guinea-Bissau family, i.e. they have an AFRI_1 spoligotype pattern, they have a distinct RFLP pattern with low numbers of IS*6110* insertions, and they lack the regions of difference RD7, RD8, RD9 and RD10, RD701 and RD702. This profile classifies the Guinea-Bissau family, irrespective of phenotypic biovar, as part of the *M. africanum* West African 2 lineage, or the AFRI_1 sublineage according to the spoligtyping nomenclature. Guinea-Bissau family strains display a variation of biochemical traits classically used to differentiate *M. tuberculosis* from *M. bovis*. Yet, the differential expression of these biochemical traits was not related to any genes so far investigated (*narGHJI* and *pncA*). Guinea-Bissau has the highest prevalence of *M. africanum* recorded in the African continent, and the Guinea-Bissau family shows a high phylogeographical specificity for Western Africa, with Guinea-Bissau being the epicenter. Trends over time however indicate that this family of strains is waning in most parts of Western Africa, including Guinea-Bissau (p = 0.048).

Citation: Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, et al. (2011) The Guinea-Bissau Family of *Mycobacterium tuberculosis* Complex Revisited. PLoS ONE 6(4): e18601. doi:10.1371/journal.pone.0018601

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received November 28, 2010; **Accepted** March 7, 2011; **Published** April 20, 2011

Copyright: © 2011 Groenheit et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was funded by grants from the Swedish International Development Cooperation Agency, the Swedish Heart-Lung Foundation, and the Swedish Research Council. The SITVIT2 database project benefited from research grants from the Regional Council of Guadeloupe (grant CR/08-1612), and by United States National Institutes of Health (grant R01LM009731). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gunilla.kallenius@ki.se

Introduction

Guinea-Bissau is a West African country with a population of approximately 1.6 million people. It is a country with one of the highest tuberculosis (TB) burdens in the world. In previous studies of *Mycobacterium tuberculosis* complex isolates from Guinea-Bissau, we identified a unique group of strains that we designated the Guinea-Bissau family of strains [1,2]. Based on a genotypic analysis of a subset of Guinea-Bissau family isolates, and a comparison of known corresponding genes in mycobacteria outside the *M. tuberculosis* complex, we proposed that strains of the Guinea-Bissau family belong to a unique branch of the *M. tuberculosis* complex tree positioned between classical *M. tuberculosis* and classical *M. bovis* [2], clustering together within one of the *M. africanum* subtype I branches [3]. The most intriguing fact about the Guinea-Bissau family of strains is that, although genotypically closely related, they demonstrate a considerable heterogeneity of phenotypic traits [1,2], with a variation in niacin production, pyrazinamide (PZA) resistance, and nitrate reductase activity, that are basic phenotypic criteria for differentiating *M. tuberculosis* and *M. bovis*. Yet, the differential expression of these biochemical traits by the Guinea-Bissau family of strains have not been related to any genes so far investigated [2].

Since the first *M. tuberculosis* strains from the late 1980s and early 1990s were isolated in Guinea-Bissau [1,4] there has been a war in the country. This caused the TB program to dissolve, the Reference TB hospital was out of function and the TB laboratory at the Laboratório Nacional de Saúde Pública (LNSP), Bissau, was destroyed and no drug resistance surveillance could be performed. Culture of sputum samples has only recently been made possible

again. We have now collected a new sample of *M. tuberculosis* complex isolates from patients with pulmonary TB and compared them to previously collected (published and unpublished) isolates in terms of genotype. Thus, we hereby reinvestigated *M. tuberculosis* complex clinical isolates from Guinea-Bissau (n = 414) collected between 1989 and 2008 with emphasis on the phylogeography, of the Guinea-Bissau family of strains, and their differential expression of biochemical traits.

Materials and Methods

Ethics Statement

The Research Committee at Karolinska Institutet exempted this retrospective study from review because it describes bacterial samples collected during the course of routine surveillance, and waived the need for consent due to the fact that the samples received could only be combined with the sex, age, and country of birth of the patients from which the strains were isolated.

Patients

The isolates were collected consecutively as part of epidemiological TB studies in Guinea-Bissau. The patients had been referred to the Raoul Follereau Reference TB hospital in Bissau, Guinea-Bissau from various part of the country. The Hospital is the TB National Reference Center, therefore receives patients from the entire country, namely patients referred from regional hospitals or health centers and also patients coming by themselves. Patients come from the entire country (Bissau, Gabu, Bafata, Cacheu regions).

Study A. During 1989–1994, 229 isolates from the same number of patients with positive cultures for the *M. tuberculosis* complex were collected as part of an epidemiological TB study in Guinea-Bissau [1,2,4,5,6].

Study B. As part of a survey of drug resistance, a systematic collection of *M. tuberculosis* complex isolates was performed between 1994–1998. Approximately 700 sputum samples from the same number of patients with suspected pulmonary TB were collected at the Raoul Follereau Reference TB hospital. Positive culture for the *M. tuberculosis* complex was confirmed for 192 samples.

Study C. During 2006, 2007 and 2008 approximately 400 sputum samples from the same number of patients were collected. Out of these, 85 were culture positive for the *M. tuberculosis* complex.

Mycobacterium tuberculosis complex isolates

Isolates included in study A, B and late part of study C were cultured at the LNSP as previously described [1]. Culture in the early part of study C was performed at the TB laboratory, Karolinska University Hospital, Stockholm, Sweden. All *M. tuberculosis* complex isolates were transported to the Reference TB laboratory at the Swedish Institute for Infectious Disease Control, Solna, Sweden, for further molecular characterization.

Spoligotyping

Spoligotyping relied on the amplification of the polymorphic direct repeat region [7] to obtain hybridization patterns of the amplified DNA using multiple synthetic spacer oligonucleotides, which are covalently bound to a membrane (Isogen Bioscience BV, The Netherlands).

Database comparison of spoligotypes

Spoligotypes in binary format were entered in the SITVIT2 database (Pasteur Institute of Guadeloupe), which is an updated version of the previously released SpolDB4 database [8]. Of 414 isolates 229 had been previously entered by us into the database, and 185 were added in conjunction with this study. At the time of the present study, SITVIT2 contained genotyping information on about 75,000 *M. tuberculosis* clinical isolates from 160 countries of origin. In this database, SIT (Spoligotype International Type) designates spoligotypes shared by two or more patient isolates, as opposed to "orphan" which designates patterns reported for a single isolate. Major phylogenetic clades were assigned according to signatures provided in SpolDB4, which defined 62 genetic lineages/sub-lineages. These include specific signatures for various *M. tuberculosis* complex species such as *M. bovis, M. microti, M. caprae, M. pinipedii,* and *M. africanum*, as well as rules defining major lineages/sublineages for *M. tuberculosis* sensu stricto. These include the Beijing clade, the Central Asian (CAS) clade and two sublineages, the East African-Indian (EAI) clade and nine sublineages, the Haarlem (H) clade and three sublineages, the Latin American-Mediterranean (LAM) clade and 12 sublineages, the "Manu" family and three sublineages, the S clade, the IS*6110*–low-banding X clade and four sublineages, and an ill-defined T clade with five sublineages.

IS*6110* RFLP analyses

A standardized method for IS*6110* RFLP analysis [9] was used. DNA was extracted and RFLP typing was performed using the insertion sequence IS*6110* as a probe and *Pvu*II as the restriction enzyme. Visual bands were analyzed using the BioNumerics software version 5.10 (Applied Maths, Kortrijk, Belgium). Fingerprint patterns were compared by the un-weighted pair-group method of arithmetic averaging using the Jaccard coefficient. Dendrograms were constructed to show the degree of relatedness among strains according to a previously described algorithm [10], and similarity matrixes were generated to visualize the relatedness between the banding patterns of all isolates.

PCR for RD702

Genomic DNA (10–20 ng) for RD702, which delineates the branch of *M. africanum* West African 2 lineage [11,12] was amplified in a volume of 25 µL, which contained premixed *Taq*, Q solution (both from Qiagen), and primers RD702F (TTCCGAGGACCCGTTGTTGAGTGC) and RD702R (GGGCGGGTTGGGTTGCTGGTC) at a final concentration of 0.2 µmol/L [3,11]. Conditions used for the amplification were as follows: 94°C for 3 min, followed by 35 cycles each of 94°C for 1 min, 64°C for 1 min, and 72°C for 1 min. PCR products of ~ 2 kb for the control strain H37Rv and ~ 700 bp for the isolates belonging to *M. africanum* West African 2 lineage were visualized on a 0.8% agarose gel [12].

Single-nucleotide polymorphism in the *narGHJI* operon

A mutation in the *narGHJI* operon has been reported to be responsible for differential nitrate reductase activity of *M. tuberculosis* (-215T) versus *M. bovis* (-215C) [13]. The single-nucleotide polymorphism (SNP) at position -215 within the nitrate reductase (*narGHJI*) operon promoter in *M. tuberculosis* complex isolates with different NO_3 activity was investigated. The isolates were characterized by PCR-RFLP using the previously published primers LC66 (AACCGACGGTGTGGTTGAC) and LC67 (ATCTCGATGGATGGGCGTC) [13]. The PCR products obtained were digested using the restriction endonuclease Sau3AI (Roche-Biomedicals, Meylan, France). This enzyme specifically cuts at the GATC sequence overlapping the *M. bovis*-like -215C *narGHJI* promoter site, producing two bands in PCR-RFLP of the LC66-LC67 PCR fragment. In the case of the -215T sequence,

Table 1. Sublineage distribution and associated SIT designations for each of the lineages in the study.

Lineage	Sublineage	SIT
AFRI n = 195 (47.10%)	AFRI_1 n = 194	***SIT181*** n = 96
		SIT187 n = 46
		orphan n = 26
		SIT536 n = 6
		SIT188 n = 3
		SIT318 n = 2
		SIT326 n = 2
		SIT525 n = 2
		SIT530 n = 2
		SIT532 n = 2
		SIT3104 n = 2
		SIT3118* n = 2
		SIT324 n = 1
		SIT537 n = 1
		SIT3130 n = 1
	AFRI_3 n = 1	orphan n = 1
LAM n = 75 (18.12%)	LAM9 n = 43	***SIT42*** n = 36
		SIT2201 n = 2
		SIT3020* n = 2
		SIT766 n = 1
		SIT866 n = 1
		orphan n = 1
	LAM1 n = 14	SIT20 n = 11
		SIT534 n = 3
	LAM4 n = 9	***SIT60*** n = 7
		SIT828 n = 1
		orphan n = 1
	LAM5 n = 3	***SIT93*** n = 1
		orphan n = 2
	LAM2 n = 2	SIT604 n = 2
	LAM3 n = 2	SIT4 n = 1
		SIT3101* n = 1
	LAM10-CAM n = 2	***SIT61*** n = 2
T clade n = 53 (12.80%)	T1 n = 41	***SIT53*** n = 10
		SIT244 n = 7
		SIT334 n = 3
		SIT804 n = 3
		orphan n = 4
		SIT3100* n = 3
		SIT196 n = 2
		SIT522 n = 2
		SIT535 n = 2
		SIT230 n = 1
		SIT521 n = 1
		SIT611 n = 1
		SIT801 n = 1
		SIT888 n = 1
	T n = 8	SIT73 n = 5
		SIT3129* n = 2
		SIT2440 n = 1
	T5 n = 3	orphan n = 2
		SIT44 n = 1
	T2 n = 1	orphan n = 1
Haarlem n = 37 (8.94%)	H1 n = 22	***SIT47*** n = 18
		SIT62 n = 1
		SIT531 n = 1
		SIT2030* n = 1
		orphan n = 1
	H3 n = 15	***SIT50*** n = 12
		SIT533 n = 2
		SIT75 n = 1
EAI n = 25 (6.04%)	EAI5 n = 20	SIT527 n = 9
		SIT528 n = 4
		orphan n = 4
		SIT458 n = 2
		SIT236 n = 1
	EAI-SOM n = 4	SIT529 n = 3
		SIT538 n = 1
	EAI6-BGD n = 1	SIT129 n = 1
Unknown signatures n = 13 (3.14%)	Unk n = 13	orphan n = 6
		SIT2445 n = 2
		SIT3131* n = 2
		SIT523 n = 1
		SIT1200 n = 1
		SIT1204 n = 1
Beijing n = 7 (1.69%)	Beijing n = 7	***SIT1*** n = 7
× clade n = 4 (0.97%)	×3 n = 2	***SIT92*** n = 2
	×1 n = 2	***SIT119*** n = 2
Manu n = 4 (0.97%)	Manu2 n = 2	***SIT54*** n = 1
		orphan n = 1
	Manu1 n = 1	SIT3132* n = 1
	Manu-ancestor n = 1	***SIT523*** n = 1
CAS n = 2 (0.48%)	CAS1-DEL n = 2	SIT3111 n = 2

Newly-defined SITs are shown by an asterisk while the prototypes defining a sublineage in SITVIT2 database are shown in ***bold italics***.
doi:10.1371/journal.pone.0018601.t001

the restriction site is absent and as a result the PCR-RFLP of the *narGHJI* amplicon yields a single band.

Geographical distribution of spoligotypes and lineages

The distribution of spoligotypes in the SITVIT2 database, in relation to their geographical origin was recorded for regions representing ≥5% of a given SIT as compared to their total number in the SITVIT2 database. The various macro-geographical regions and sub-regions were defined according to the

specifications of the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm). Correlation of spoligotype families and principal genetic groups (PGG) based on *KatG463-gyrA95* polymorphisms [14] was inferred from the reported linking of specific spoligotype patterns to PGG1 (ancient lineages), and PGG2/3 (modern lineages) as previously described [8,15,16].

For statistical analysis of trends over time the Chi square test for trend was performed using PASW v 18.

Results

In study A, spoligotyping and RFLP was performed on 229 isolates. In study B, adequate amounts of DNA, in order to perform RFLP, were retrieved from 132 isolates, while it was possible to perform spoligotyping on 138 isolates. In study C, spoligotyping could be performed on 47 isolates, while sufficient amounts of DNA were available for RFLP typing of 41 isolates. In regard to the remaining positive samples they were either too contaminated or the isolates had died during transportation. All isolates that were subjected to RFLP were also spoligotyped. For single-nucleotide polymorphism in the *narGHJI* operon, 28 isolates were selected from RFLP cluster B:5 [1] in study A.

Spoligotyping

Altogether, 414 isolates were available for spoligotyping, of which 336 isolates were grouped into 42 clusters (2–96 isolates per cluster), accounting for a very high clustering rate of 81.2%, while the remaining 78 (18. 8%) strains did not cluster. Seven clusters included more than ten isolates each.

Study A. Out of 229 isolates, 82 different spoligotypes were obtained, of which 168 (73.4%) were clustered into 21 spoligo-clusters, comprising 2–61 isolates per cluster. The remaining 61 (26.6%) spoligopatterns were unique i.e. the isolates did not cluster with other patient isolates. The 229 isolates are identical to those reported in [1], however a mistake in the recording of one spoligotype (isolate IH-218) has been corrected here. Also, nine isolates with unique spoligotypes (summarized in Table 2 of reference [1]), were not mentioned in the Results section of the 1999 paper.

Study B. Of 138 isolates, 56 different spoligotypes were obtained, of which 98 (71.0%) were clustered into 15 spoligo-clusters, comprising 2–26 isolates per cluster. The remaining 40 (29.0%) spoligopatterns were unique i.e. the isolates did not cluster with other patient isolates.

Study C. Of 47 isolates, 26 different spoligotypes were obtained, of which 29 (61.7%) were clustered into 8 spoligo-clusters, comprising 2–9 isolates per cluster. The remaining 18 (38.3%) spoligopatterns were unique i.e. the isolates did not cluster with other patient isolates.

The spoligotyping results from the SITVIT2 database comparison are summarized in **Table 1**. Of 120 different patterns for the 414 strains studied 50 patterns corresponded to orphan strains that were unique among the 75,000 strains recorded in the SITVIT2 database (Supplementary **Table S1**), as opposed to 70 patterns corresponding to 364 clinical isolates that corresponded to shared-types (Supplementary **Table S2**), i.e. an identical pattern shared by two or more patients worldwide (within this study, or matching another strain in the SITVIT2 database). A SIT number was attributed to each pattern according to the SITVIT2 database.

When the overall repartition of strains according to major *M. tuberculosis* genotypic families was defined (**Table 1**), the major circulating *M. tuberculosis* clades ranked in the following order: AFRI (n = 195, 47.10%) > LAM (n = 75, 18.12%) > ill-defined T clade (n = 53, 12.80%) > Haarlem (n = 37, 8.85%) > EAI (n = 25, 6.04%) > Unknown (n = 12, 2.87%) > Beijing (n = 7, 1.68%) > X clade (n = 4, 0.96%) > Manu (n = 4, 0.97%), CAS (n = 2, 0.48%). Two strains of the LAM clade, both isolated in 2007, belonged to the Cameroon family (SIT61) [17].

The most predominant lineage in Guinea-Bissau was AFRI, which accounted for as high as 47.10% of all TB cases. Surprisingly all the cases with AFRI with the exception of a single orphan isolate classified as AFRI_3, were caused by the AFRI_1 sublineage split into two spoligopatterns – SIT181 (the prototype of the AFRI_1 sublineage) and SIT187. As compared to the SITVIT2, these two SITs were highly predominant in Guinea-Bissau since they alone accounted for 36.92% and 52.87% of all such strains in the database (414 of 423 isolates from Guinea-Bissau in the database originating from this study) with extremely high phylogeographical specificity for Western Africa (74.62 and 74.71% respectively, (Supplementary **Table S3**).

Another interesting observation was a pattern belonging to the EAI5 sublineage (SIT527, n = 9 strains in this study), with 100% of the cases being identified so far exclusively in Guinea-Bissau. These lineages are characteristic of the evolutionary ancient PGG1 strains. The remaining predominant SITs (SIT42 and SIT20 belonging to the LAM9 and LAM1 sublineages of LAM clade, and SIT47 and SIT50 belonging to the H1 and H3 sublineages of Haarlem clade) were characteristic of the evolutionary-recent PGG2/3 subgroups. The last predominant spoligotype (SIT53) also belonged to the evolutionary-recent T1 sublineage, nonetheless, the "T" family does not represent a genotypic lineage in a strict evolutionary sense since it was defined by default to include strains that could not be classified in one of the established lineages with well-established phylogeographical specificity such as the Haarlem, LAM, CAS, and EAI lineages [8].

Thus if one adds all the PGG1 isolates in Guinea-Bissau, i.e., AFRI n = 195, EAI n = 25, Beijing n = 7, Manu n = 4, CAS n = 2, and one unknown lineage pattern (SIT1200) that could be nonetheless classified as PGG1, the total number of ancestral strains in Guinea-Bissau amounts to 234/414 or 56.52% of all strains.

To get an indication of the distribution of the Guinea-Bissau family in Africa, and in spite of the fact that the SITVIT2 is not absolutely representative in terms of proportional representation, we decided to compare all the data obtained in Guinea-Bissau, and compared it to isolates in the SITVIT2 from various subregions of Africa following United Nations geoscheme, i.e., Eastern Africa, Middle Africa, Northern Africa, Southern Africa, and Western Africa (which included Guinea-Bissau). The results obtained for a global collection of 11956 clinical isolates in the database are summarized in **Figure 1**, and highlight major differences in the population structure of circulating tubercle bacilli in Africa. It should be noted that to facilitate comparison, results were only tabulated for all major lineages with a cut-off at 3% (i.e., strains belonging to all minor lineages below the cut-off 3% are pooled together and shown as "Other"). Furthermore, all spoligotyping signatures that are not yet associated to a well-defined genotypic lineage in SITVIT2 were designated as "Unknown".

As for other subregions, this distribution in Guinea-Bissau does not take into account the minor lineages below the cut-off of 3% that included Beijing, X, Manu and CAS lineages.

RFLP

Out of the 402 isolates investigated, 193 different RFLP patterns were obtained, of which 268 (66.7%) isolates were clustered into 61 RFLP-clusters, comprising 2–45 isolates per cluster. The remaining 132 (32.8%) RFLP patterns were unique i.e. the isolates did not cluster with other patient isolates. Of the

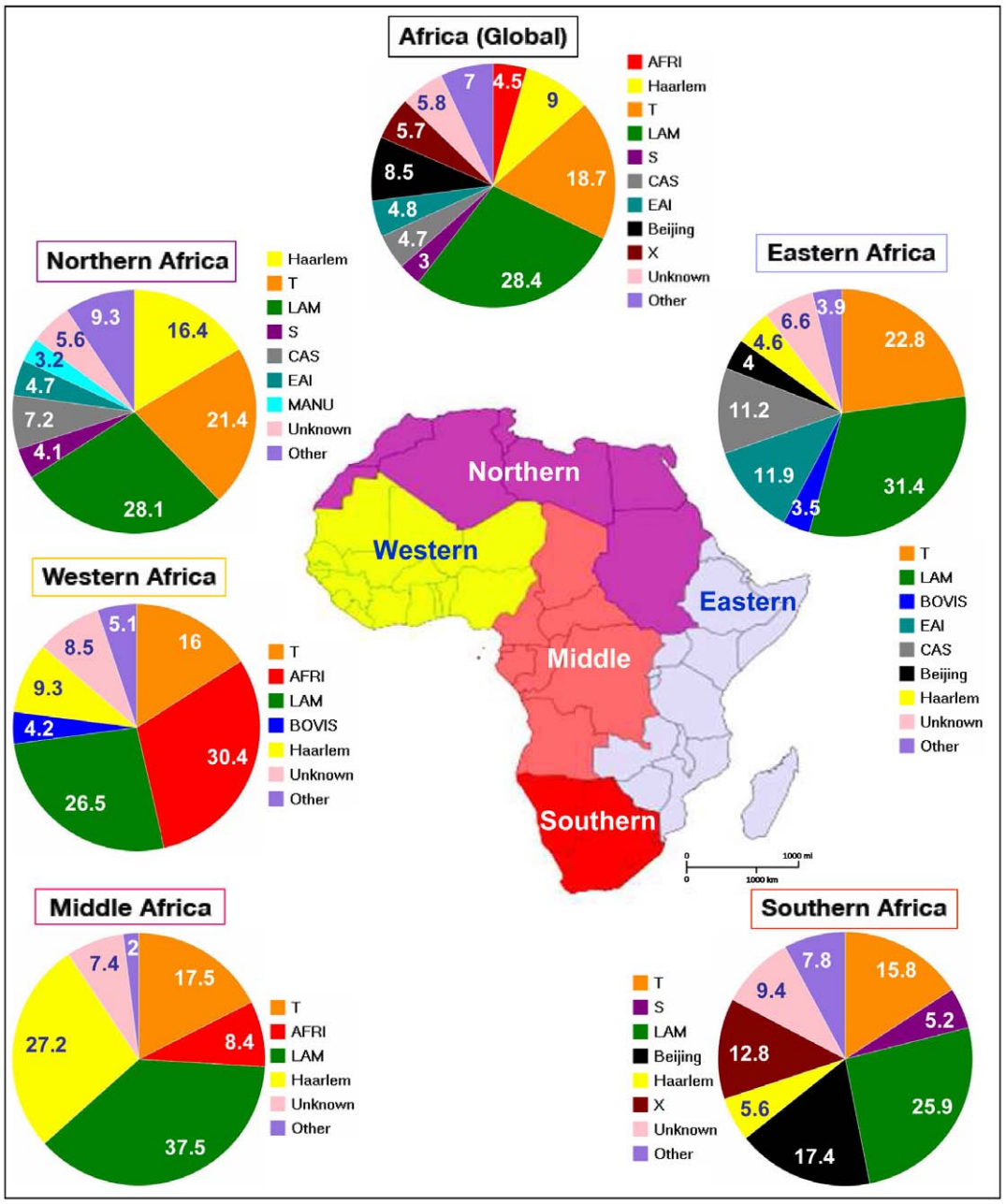


Figure 1. Geographical distribution of spoligotyping-based genotypic lineages of *Mycobacterium tuberculosis* in various subregions of Africa (n = 11956 clinical isolates) in the SITVIT2 database.
doi:10.1371/journal.pone.0018601.g001

RFLP-typed isolates with an *M. africanum* spoligotype 146/193 (75.6%) isolates, and 121/209 (57.9%) of the non-*M. africanum* isolates were clustered into 24 and 37 RFLP-clusters respectively. Of those with an *M. africanum* spoligotype 125/193 had 5 bands or less (median 5 bands) and 38/209 of the non-*M. africanum* isolates had 5 bands or less (median 9 bands).

RD702

Of the 195 isolates with a *M. africanum* signature 192 isolates were analyzed for RD702; they all showed a PCR product of ~ 700 bp indicating that RD702 is absent in the Guinea-Bissau family isolates.

Single-nucleotide polymorphism in the *narGHJI* operon

Twenty-eight Guinea-Bissau isolates, representing all five biovars [4], were investigated. Six of the isolates were nitrate reductase positive, while 22 were negative. The *M. tuberculosis* H37Rv reference strain served as a control. All 28 isolates showed a single PCR product with a size of approximately 155 bp. For all but one of the Guinea-Bissau family strains, irrespective of nitrate reductase activity, the subsequent digestion with Sau3AI, specifically cutting at the GATC sequence, produced two bands (~ 90 and ~ 70 bp) showing that these strains harbored the -215C promoter SNP. For the *M. tuberculosis* H37Rv reference strain and one Guinea-Bissau strain (nitrate positive) cleavage was not

successful, resulting in a single, uncut band of approximately 155 bp.

Trends over time

The proportion of AFRI_1 significantly decreased over time when the three studies A, B and C were compared (P-value = 0.048). Thus in study A 118 (52%) of 229 isolates were AFRI, in study B 60 (44%) of 138 isolates were AFRI, and in study C 18 (38%) of 47 isolates were AFRI. The AFRI_3 isolate was found in 1996 (study B).

Discussion

The Guinea-Bissau family is a member of the *M. africanum* West African 2 lineage Traditionally, *M. africanum* strains were defined according to their phenotypic traits and geographic origin, and were accordingly named West African (subtype I) and East African (subtype II) *M. africanum* [4,18]. Based on genetic definitions so called *M. africanum* subtype II strains are now regarded as phenotypic variants of *M. tuberculosis* [19,20,21]. *M. africanum* subtype I (West African *M. africanum*) strains are defined on the basis of particular genotypic characteristics that place them in the *M. tuberculosis* complex phylogenetic tree [2,19,22,23]. Thus, genomic deletions, or regions of difference (RDs), can effectively distinguish them from *M. tuberculosis* and *M. bovis*. Isolates of West African *M. africanum* lack RD9 and are positive for RD12 [19,20]. We subsequently refer to *M. africanum* subtype I as *M. africanum*. *M. africanum* strains can be further subdivided into two main lineages, West African 1 and West African 2 [12,20]. West African 1 isolates lack RD711, while West African 2 isolates lack RD 7, 8, 10 and 702 [12,24]. By spoligotyping *M. africanum* is defined by loss of spacers 8, 9 and 39. Three sublineages are defined based on finer differences: AFRI_1 by loss of spacers 7–9, and 39, AFRI_2 by loss of spacers 8–12, 21–24, and 37–39, and AFRI_3 by loss of spacers 8–12, and 37–39. These subdivisions correspond to LSP-defined classification as follows: AFRI_1 corresponds to West African 2, while AFRI_2 and AFRI_3 collectively constitute the West African 1 lineages.

This study of *M. tuberculosis* complex isolates from Guinea-Bissau, collected during two decades, shows that TB in Guinea-Bissau is essentially caused by predominant genotypes characteristic of the Western and Middle African region, with the highest prevalence of *M. africanum* (AFRI_1/West African 2) in the whole African continent. Forty-seven percent of all isolates were of what we call the Guinea-Bissau family, a member of the *M. africanum* AFRI_1/West African 2 lineage.

Thus, they have an AFRI_1 spoligotype pattern [1] (this study), they have a distinct RFLP pattern with low numbers of IS*6110* insertions [1] (this study) and they lack the four regions of difference RD7, RD8, RD9 and RD10 [2], as well as RD701 [3] and RD702 (this study). This profile classifies the Guinea-Bissau family strains, irrespective of phenotypic biovar, as part of the *M. africanum* West African 2 lineage [24] or the AFRI_1 sublineage according to the spoligotyping nomenclature. This family shows an extremely high phylogeographical specificity for Western Africa, with Guinea-Bissau being the epicenter.

Other lineages. In addition to the AFRI isolates 25 isolates were of the "ancient" EAI lineage, and the remaining strains were "modern" *M. tuberculosis* strains. These latter strains most likely represent strain importation from countries where such strains predominate. Only a few isolates were of the Beijing lineage, which is relatively rare in Africa, except for parts of South Africa [25].

Historically, Guinea-Bissau is a former colony of Portugal, a country with a high incidence of TB caused by the classical human *M. tuberculosis* type. Thus, contacts between the populations of Guinea-Bissau and Portugal have probably been more numerous than contacts between the populations of Guinea-Bissau and other African countries. We therefore speculate that these *M. tuberculosis* strains may have their origin from within Portugal [1]. Indeed, some of the isolates had RFLP patterns that were identical to isolates from Portuguese TB patients [1].

An interesting observation was the exclusive presence of an EAI5 sublineage (SIT527) in Guinea-Bissau. LAM10-CAM strains were found in a very low proportion, with only 2/414 (0.5%) of such isolates, interestingly both strains belonged to LAM10-CAM prototype SIT61. Yet they are predominant in Middle Africa and Western Africa, with phylogeographical specificity of LAM10-CAM for Cameroon and neighboring countries in West Africa [8,26], and a recent study showed that LAM10-CAM prototype SIT61 was also the most predominant spoligotype in Cotonou, Benin [27].

Another interesting feature was the total absence of *M. bovis* in our dataset. Similarly, the finding of four Manu strains in Guinea-Bissau is noteworthy. The "Manu" lineage was initially described as a new family from India in 2004 [28], and later similar strains in small numbers were reported in a study from Madagascar [29]. Soon afterwards, it was tentatively subdivided into Manu-1 (deletion of spacer 34), Manu-2 (deletion of spacers 33–34) and Manu-3 (deletion of spacers 34–36) sub-lineages, with the suggestion that it could represent an ancestral clone of PGG1 strains [8]. More recently, Manu lineage strains were reported from Saudi Arabia, [30] Tunisia [31], and Egypt [32]. In the latter study, Manu lineage strains represented as high as 27% of all isolates, with the new addition of SIT523 (all 43 spacers present) designated as "Manu-ancestor" [32]. However, it was recently suggested that some Manu strains may eventually correspond to a "mixed" pattern due to concomitant Beijing and Euro-American lineage strains (the latter comprising H, LAM, X, and T lineages in spoligotyping defined clades), a possibility that may be investigated by RD105 analysis of Manu pattern isolates [33].

Phenotype and genotype

Although the *M. africanum*-specific deletions each have putative biological consequence, they do not directly account for these mentioned variable biochemical characteristics. However, genomic variations other than the deletion of large sequence polymorphisms (LSPs) may be relevant for the observed phenotypic variability. Small polymorphisms, single nucleotide polymporhisms (SNPs), gene rearrangements, or genomic regions absent from H37Rv but present in *M. africanum* (duplications, insertions) are overlooked when applying RD analysis. As a result, phenotypic heterogeneity among *M. africanum* may be due to unidentified genetic polymorphisms.

Classical *M. bovis* strains, in contrast to *M. tuberculosis* strains, are naturally resistant to PZA and lack PZase activity. A point mutation in the *pncA* gene has been thought to be the cause of the defective PZase activity and PZA resistance in classical *M. bo*vis strains [34]. We have investigated isolates of the Guinea-Bissau family (Study A), and found that irrespective of phenotypic PZase activity, they all had the *M. tuberculosis* genotype in the *pncA* gene [1] (data not shown). This was true also for isolates lacking PZase activity, indicating that the variation in phenotypic PZase activity is not due to a variation in the *pncA* gene.

The differential nitrate reductase (NO_3) activity of *M. tuberculosis* versus *M. bovis* has been reported to be due to a SNP in the *narGHJI* operon promoter [13]. The Guinea-Bissau strains tested,

except for one, all carried the -215C *narGHJI* promoter genotype, as defined by the two-band pattern observed in PCR-RFLP. This genotype is present in *M. bovis* and some ancestral *M. tuberculosis* [35], and has been connected with a reduced nitrate reductase production. Expression of *narGHJI* is typically induced under anaerobic conditions, anaerobic nitrate reductase activity being expressed in both -215T and -215C genotypes. However, anaerobic nitrite reductase activity is considerably lower in *M. bovis*, which carries the -215C genotype, compared to *M. tuberculosis*. Diagnostic nitrate reductase activity is generally measured as the rapid accumulation of nitrite by the bacilli within two hours. *M. bovis* typically needs a longer time to accumulate detectable amounts of nitrite from nitrate [13]. Thus, the variable expression of nitrate reductase by *M. africanum* may be due to the narrow time-detection limit of nitrite defined by the test used and not a genotypic difference between individual *M. africanum* strains.

The differential expression of these biochemical traits is therefore not related to any genes that have been reported to differentiate between *M. bovis* and *M. tuberculosis*. Instead, we hypothesize that Guinea-Bissau strains may more readily switch genes on and off during its adaptation to hosts of different species (and hence broaden their host tropism) [4].

Will the Guinea-Bissau family of strains disappear from West Africa?

This analysis shows that there exists a clear gradient in the African continent; *M. africanum* being predominant in Western and Middle Africa, EAI and CAS in Northern and Eastern Africa, Beijing and X essentially limited to Southern Africa, while LAM, Haarlem, and T are distributed throughout.

During the 15-year period since isolates were first collected, the proportion of isolates of the Guinea-Bissau family have diminished over time. This is in agreement with trends in some other West African countries, Cameroon [17] and Burkina Faso [36], where the proportion of *M. africanum* isolates have drastically decreased during the last decades. *M. africanum* appears at various frequencies in West Africa. In Sierra Leone a recent study identified 23/97 (24%) isolates as *M. africanum* [37], which appears to be similar to that reported from a study in 1992/1993 [18]. In The Gambia, 38% of isolates were *M. africanum* [38] and in Ghana 29% [39]. In the 1970's, most cases of TB in Cameroon were caused by *M. africanum* [40], while three decades later only 9% of cases were reported to be caused by *M. africanum* [17]. This change appears to be due to the recent expansion of the so-called Cameroon family of *M. tuberculosis* [17,26]. In Burkina Faso, *M. africanum* was not detected in a recent survey [36] although *M. africanum* strains (defined by phenotypic characteristics) were reported in about 20% of cases two decades earlier [36,41,42]. Interestingly, during 2007, two strains of the Cameroon family had found their way into our setting.

The reason for the relative decline of *M. africanum* in West Africa could be a lower transmission capacity or virulence. There is a considerable uncertainty about the virulence of *M. africanum*. In studies from The Gambia [11,38,43,44] and Ghana [39] this issue has been addressed. No difference in virulence, as assessed by the severity of radiological presentation, was found between the West African 1 and West African 2, and *M. africanum* was seen to closely resemble *M. tuberculosis* in pathology [39]. The pathology in patients with smear positive TB caused by *M. africanum* (West African 2) were found to be as severe as that of *M. tuberculosis* patients [44]. However, when comparing *M. tuberculosis* and *M. africanum* (West African 2) the rate of progression to disease was higher in household contacts of index patients with *M. tuberculosis* compared to those with *M. africanum* infection [43]. Interestingly, *M. africanum* elicited an attenuated T cell response to early secreted antigenic targets in patients with TB and their household contacts [11].

Another reason for the decline of *M. africanum* in West Africa could be that certain "modern" lineages possess advantages in their ability to disseminate within a community, in relation to more "ancient" lineages such as *M. africanum*. Of particular interest in this context is the appearance of Beijing isolates, as well as the recent isolation of two strains of the Cameroon family in our study. Beijing strains have recently been shown to have emerged and rapidly expanded in South Africa [25].

In conclusion the Guinea-Bissau family of strains, irrespective of phenotypic biovar, is part of the *M. africanum* West African 2 lineage, or the AFRI_1sublineage according to the spoligotyping nomenclature. The variation of biochemical traits classically used to differentiate *M. tuberculosis* from *M. bovis* is not related to any genes so far investigated (*narGHJI* and *pncA*). The Guinea-Bissau family shows an extremely high phylogeographical specificity for Western Africa, with Guinea-Bissau being the epicenter. Trends over time however indicate that this family of strains is waning in most parts of Africa, including Guinea-Bissau.

Supporting Information

Table S1 Description of the orphan strains (n = 50) and corresponding spoligotyping defined lineages found among a total of 414 *M. tuberculosis* clinical isolates from studies performed in Guinea Bissau.
(DOC)

Table S2 Description of 70 shared types containing 364 *M. tuberculosis* complex clinical isolates from Guinea Bissau (GNB).
(DOC)

Table S3 Description of predominant SITs (representing 10 or more strains) in Guinea Bissau (GNB), and their worldwide distribution.
(DOC)

Acknowledgments

We thank the staff at the Laboratório Nacional de Saúde Pública, Bissau, the TB National Reference Hospital "Raoul Follereau", Bissau and the TB laboratory at the Karolinska University Hospital, Stockholm for culturing assistance. We also thank Biniyam Ghebru for running the *narGHJI* PCR, Lina Benson for statistical help and Thierry Zozio and Thomas Burguière (Institut Pasteur de Guadeloupe) for helping with mapping of the spoligotype patterns using the SITVIT2 database.

Author Contributions

Conceived and designed the experiments: RG SG NR TK GK. Performed the experiments: RG SG JS PR RC FR DC VH TK. Analyzed the data: RG SG PR RC FR DC VH NR TK GK. Contributed reagents/materials/analysis tools: PR RC FR NR. Wrote the paper: RG SG PR RC FR DC VH NR TK GK.

References

1. Kallenius G, Koivula T, Ghebremichael S, Hoffner SE, Norberg R, et al. (1999) Evolution and clonal traits of Mycobacterium tuberculosis complex in Guinea-Bissau. J Clin Microbiol 37: 3872–3878.

2. Koivula T, Ekman M, Leitner T, Lofdahl S, Ghebremicahel S, et al. (2004) Genetic characterization of the Guinea-Bissau family of Mycobacterium tuberculosis complex strains. Microbes Infect 6: 272–278.

3. Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, et al. (2004) Genomic analysis distinguishes Mycobacterium africanum. J Clin Microbiol 42: 3594–3599.
4. Hoffner SE, Svenson SB, Norberg R, Dias F, Ghebremichael S, et al. (1993) Biochemical heterogeneity of Mycobacterium tuberculosis complex isolates in Guinea-Bissau. J Clin Microbiol 31: 2215–2217.
5. Dias F, Michael SG, Hoffner SE, Martins L, Norberg R, et al. (1993) Drug susceptibility in Mycobacterium tuberculosis of a sample of patients in Guinea Bissau. Tuber Lung Dis 74: 129–130.
6. Naucler A, Winqvist N, Dias F, Koivula T, Lacerda L, et al. (1996) Pulmonary tuberculosis in Guinea-Bissau: clinical and bacteriological findings, human immunodeficiency virus status and short term survival of hospitalized patients. Tuber Lung Dis 77: 226–232.
7. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, et al. (1997) Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 35: 907–914.
8. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, et al. (2006) Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6: 23.
9. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, et al. (1993) Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 31: 406–409.
10. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD (1991) Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol 29: 2578–2586.
11. de Jong BC, Hill PC, Brookes RH, Gagneux S, Jeffries DJ, et al. (2006) Mycobacterium africanum elicits an attenuated T cell response to early secreted antigenic target, 6 kDa, in patients with tuberculosis and their household contacts. J Infect Dis 193: 1279–1286.
12. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, et al. (2006) Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 103: 2869–2873.
13. Stermann M, Bohrssen A, Diephaus C, Maass S, Bange FC (2003) Polymorphic nucleotide within the promoter of nitrate reductase (NarGHJI) is specific for Mycobacterium tuberculosis. J Clin Microbiol 41: 3252–3259.
14. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, et al. (1997) Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 94: 9869–9874.
15. Gutierrez MC, Ahmed N, Willery E, Narayanan S, Hasnain SE, et al. (2006) Predominance of ancestral lineages of Mycobacterium tuberculosis in India. Emerg Infect Dis 12: 1367–1374.
16. Molina-Torres CA, Moreno-Torres E, Ocampo-Candiani J, Rendon A, Blackwood K, et al. Mycobacterium tuberculosis spoligotypes in Monterrey, Mexico. J Clin Microbiol 48: 448–455.
17. Niobe-Eyangoh SN, Kuaban C, Sorlin P, Cunin P, Thonnon J, et al. (2003) Genetic biodiversity of Mycobacterium tuberculosis complex strains from patients with pulmonary tuberculosis in Cameroon. J Clin Microbiol 41: 2547–2553.
18. Haas WH, Bretzel G, Amthor B, Schilke K, Krommes G, et al. (1997) Comparison of DNA fingerprint patterns of isolates of Mycobacterium africanum from east and west Africa. J Clin Microbiol 35: 663–666.
19. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, et al. (2002) A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99: 3684–3689.
20. Niemann S, Kubica T, Bange FC, Adjei O, Browne EN, et al. (2004) The species Mycobacterium africanum in the light of new molecular markers. J Clin Microbiol 42: 3958–3962.
21. Asiimwe BB, Koivula T, Kallenius G, Huard RC, Ghebremichael S, et al. (2008) Mycobacterium tuberculosis Uganda genotype is the predominant cause of TB in Kampala, Uganda. Int J Tuberc Lung Dis 12: 386–391.
22. Mostowy S, Cousins D, Brinkman J, Aranaz A, Behr MA (2002) Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. J Infect Dis 186: 74–80.
23. Huard RC, Fabre M, de Haas P, Lazzarini LC, van Soolingen D, et al. (2006) Novel genetic polymorphisms that further delineate the phylogeny of the Mycobacterium tuberculosis complex. J Bacteriol 188: 4271–4287.
24. Gagneux S, Small PM (2007) Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis 7: 328–337.
25. Cowley D, Govender D, February B, Wolfe M, Steyn L, et al. (2008) Recent and rapid emergence of W-Beijing strains of Mycobacterium tuberculosis in Cape Town, South Africa. Clin Infect Dis 47: 1252–1259.
26. Niobe-Eyangoh SN, Kuaban C, Sorlin P, Thonnon J, Vincent V, et al. (2004) Molecular characteristics of strains of the cameroon family, the major group of Mycobacterium tuberculosis in a country with a high prevalence of tuberculosis. J Clin Microbiol 42: 5029–5035.
27. Affolabi D, Anyo G, Faihun F, Sanoussi N, Shamputa IC, et al. (2009) First molecular epidemiological study of tuberculosis in Benin. Int J Tuberc Lung Dis 13: 317–322.
28. Singh UB, Suresh N, Bhanu NV, Arora J, Pant H, et al. (2004) Predominant tuberculosis spoligotypes, Delhi, India. Emerg Infect Dis 10: 1138–1142.
29. Ferdinand S, Sola C, Chanteau S, Ramarokoto H, Rasolonavalona T, et al. (2005) A study of spoligotyping-defined Mycobacterium tuberculosis clades in relation to the origin of peopling and the demographic history in Madagascar. Infect Genet Evol 5: 340–348.
30. Al-Hajoj SA, Zozio T, Al-Rabiah F, Mohammad V, Al-Nasser M, et al. (2007) First insight into the population structure of Mycobacterium tuberculosis in Saudi Arabia. J Clin Microbiol 45: 2467–2473.
31. Namouchi A, Karboul A, Mhenni B, Khabouchi N, Haltiti R, et al. (2008) Genetic profiling of Mycobacterium tuberculosis in Tunisia: predominance and evidence for the establishment of a few genotypes. J Med Microbiol 57: 864–872.
32. Helal ZH, Ashour MS, Eissa SA, Abd-Elatef G, Zozio T, et al. (2009) Unexpectedly high proportion of ancestral Manu genotype Mycobacterium tuberculosis strains cultured from tuberculosis patients in Egypt. J Clin Microbiol 47: 2794–2801.
33. Viegas SO, Machado A, Groenheit R, Ghebremichael S, Pennhag A, et al. Molecular diversity of Mycobacterium tuberculosis isolates from patients with pulmonary tuberculosis in Mozambique. BMC Microbiol 10: 195.
34. Scorpio A, Zhang Y (1996) Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. Nat Med 2: 662–667.
35. Goh KS, Rastogi N, Berchel M, Huard RC, Sola C (2005) Molecular evolutionary history of tubercle bacilli assessed by study of the polymorphic nucleotide within the nitrate reductase (narGHJI) operon promoter. J Clin Microbiol 43: 4010–4014.
36. Godreuil S, Torrea G, Terru D, Chevenet F, Diagbouga S, et al. (2007) First molecular epidemiology study of Mycobacterium tuberculosis in Burkina Faso. J Clin Microbiol 45: 921–927.
37. Homolka S, Post E, Oberhauser B, George AG, Westman L, et al. (2008) High genetic diversity among Mycobacterium tuberculosis complex strains from Sierra Leone. BMC Microbiol 8: 103.
38. de Jong BC, Hill PC, Brookes RH, Otu JK, Peterson KL, et al. (2005) Mycobacterium africanum: a new opportunistic pathogen in HIV infection? Aids 19: 1714–1715.
39. Meyer CG, Scarisbrick G, Niemann S, Browne EN, Chinbuah MA, et al. (2008) Pulmonary tuberculosis: virulence of Mycobacterium africanum and relevance in HIV co-infection. Tuberculosis (Edinb) 88: 482–489.
40. Huet M, Rist N, Boube G, Potier D (1971) [Bacteriological study of tuberculosis in Cameroon]. Rev Tuberc Pneumol (Paris) 35: 413–426.
41. Ledru S, Cauchoix B, Yameogo M, Zoubga A, Lamande-Chiron J, et al. (1996) Impact of short-course therapy on tuberculosis drug resistance in South-West Burkina Faso. Tuber Lung Dis 77: 429–436.
42. Rey JL, Villon A, Menard M, Le Mao G, Albert JP (1979) [Present state of resistance to antibiotics by tuberculous bacilli in Africa (author's transl)]. Med Trop (Mars) 39: 149–154.
43. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, et al. (2008) Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. J Infect Dis 198: 1037–1043.
44. de Jong BC, Hill PC, Aiken A, Jeffries DJ, Onipede A, et al. (2007) Clinical presentation and outcome of tuberculosis patients infected by M. africanum versus M. tuberculosis. Int J Tuberc Lung Dis 11: 450–456.

Article 11

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 238532, 12 pages
http://dx.doi.org/10.1155/2014/238532

Research Article

Strain Diversity of *Mycobacterium tuberculosis* Isolates from Pulmonary Tuberculosis Patients in Afar Pastoral Region of Ethiopia

Mulugeta Belay,[1,2] Gobena Ameni,[1] Gunnar Bjune,[2] David Couvin,[3] Nalin Rastogi,[3] and Fekadu Abebe[2]

[1] *Aklilu Lemma Institute of Pathobiology, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia*
[2] *Section for International Health, Department of Community Medicine, Institute for Health and Society, University of Oslo, P.O. Box 1130, Blindern, 0318 Oslo, Norway*
[3] *WHO Supranational TB Reference Laboratory, TB & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, 97183 Abymes, France*

Correspondence should be addressed to Mulugeta Belay; mulg2002@yahoo.com

Received 24 November 2013; Accepted 22 January 2014; Published 6 March 2014

Academic Editor: Tomasz Jagielski

Copyright © 2014 Mulugeta Belay et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Data on genotypic diversity of *Mycobacterium tuberculosis* complex (MTBC) is important to understand its epidemiology, human adaptation, clinical phenotypes, and drug resistance. This study aimed to characterize MTBC clinical isolates circulating in a predominantly pastoralist area in Ethiopia, a country where tuberculosis is the second leading cause of mortality. Culture of sputum samples collected from a total of 325 pulmonary TB suspects was done to isolate MTBC. Spoligotyping was used to characterize 105 isolates from culture positive slopes and the result was compared with an international database. Forty-four spoligotype patterns were observed to correspond to 35 shared-types (SITs) containing 96 isolates and 9 orphan patterns; 27 SITs containing 83 isolates matched a preexisting shared-type in the database, whereas 8 SITs ($n = 13$ isolates) were newly created. A total of 19 SITs containing 80 isolates were clustered within this study (overall clustering of 76.19%). Three dominant lineages (T, CAS, and Manu) accounted for 76.19% of the isolates. SIT149/T3-ETH was one of the two most dominant sublineages. Unlike previous reports, we show that Manu lineage strains not only constitute a dominant lineage, but are also associated with HIV infection in Afar region of Ethiopia. The high level of clustering suggests the presence of recent transmission that should be further studied using additional genotyping markers.

1. Introduction

Tuberculosis (TB) is among the 10 major causes of death [1] and it is only outranked by HIV/AIDS among infectious causes of death worldwide [2]. Despite the global effort towards controlling TB for the last 20 years, the burden remains high with over 8.7 million incident cases and 1.4 million deaths in 2011 [3]. The HIV epidemics and the spread of MDR-TB are among the major obstacles for the successful control of TB and, in this regard, Africa is lagging behind in achieving to halve the 1990s mortality by 2015 [3].

Ethiopia is among the 22 high TB as well as the 27 high MDR-TB burden countries [3]. According to the Ministry of Health, TB is the second leading cause of death [4]. Apart from other factors, the HIV epidemics and the emergence of MDR-TB have contributed to the high TB burden in the country. According to the recent national TB prevalence survey, the prevalence of smear positive pulmonary TB is highest among pastoral communities [5].

Genotypic variations among *M. tuberculosis* strains as well as the existence of human genetic polymorphism linked to TB have resulted in the changing relationship between *M. tuberculosis* and the human host. Such changes have complicated TB control efforts [6]. Data on strain diversity of mycobacterial isolates is important to understand the transmission dynamics and phylogeographical distribution of dominant circulating strains of *M. tuberculosis*.

This study was part of a major project on molecular epidemiology, clinical epidemiology, and immunology of TB in a pastoral community and their livestock in Ethiopia. Some

studies have reported the dominant strains among human isolates in central highlands of Ethiopia [7–11]; besides, strains isolated from goats and camels have been reported from Afar region recently [12, 13]. This study reports strains of *M. tuberculosis* isolates circulating in the Afar pastoral communities.

2. Methods

2.1. Study Area. The study area has been described elsewhere [14]. Briefly, the study was conducted in two public (Awash Health Centre and Dubti Hospital) health facilities in Afar region and three private (Selam hospital, Bati Hospital, and Amir Higher Clinic) health facilities in Dessie town. The reason for including these private health facilities is because they provide diagnostic services for a substantial number of TB patients coming from Afar region.

2.2. Study Design and Study Participants. A health facility-based cross-sectional study was conducted between September 2009 and March 2010. A total of 325 pulmonary TB suspects (≥18 years of age) who were residents of Afar region during the study period and came to the selected health facilities with cough lasting more than 2 weeks were included consecutively. TB patients who were already taking anti-TB drugs were excluded.

2.3. Data Collection. A semi-structured, pretested questionnaire was used to collect data on basic sociodemographic characteristics as well as clinically relevant symptoms. Three sputum samples (spot-morning-spot) were collected and smear microscopic examination for AFB was done at the respective health facilities according to the national TB and Leprosy control programme guideline [15]. The rest of the sputa were stored at +4°C and transported to Aklilu Lemma Institute of Pathobiology (ALIPB), Addis Ababa University, within 1 week. Participants were tested for HIV according to the national guideline [16].

2.4. Mycobacterial Culture. The three sputa from each participant were pooled at ALIPB laboratory; culture was done according to the WHO guideline [17]. Briefly, sputum samples were homogenized and decontaminated with equal volume of 4% NaOH and shaken for 15 minutes at room temperature. Subsequently, it was centrifuged at 3000 rpm for 15 minutes. After the supernatant was poured off, the sediment was neutralized with 2 N HCL and 2–4 loopfuls of the centrifuged sediment were inoculated into four slopes of LJ medium. Inspection of media was done every week for growth until 8 weeks. Based on colony morphology and smear microscopy, those with growth were identified and two colonies were transferred to eppendorf tube with 300 μL distilled water. Mycobacterial genomic DNA was extracted by heating isolates at 80°C for 60 minutes. Heat-killed isolates were stored at −20°C until spoligotyping was done.

2.5. Spoligotyping and Database Comparison. Spoligotyping of 105 isolates was performed using a commercially available kit from Ocimum Biosolutions Company, Iisselstein, The Netherlands, according to the company's instructions and as described previously [9, 18]. Briefly, the direct repeat (DR) region was amplified by a Thermal Cycler PCR machine (VWR International) using oligonucleotides and primers derived from this region. The amplified product was hybridized to a set of 43 immobilized oligonucleotides, each corresponding to one of the unique spacer. Hybridized DNA was detected by chemiluminescence method (Amersham Biosciences, Little Chalfont, UK) and by exposure to X-ray film (Hyperfilm ECL, Amersham Biosciences), as specified by the manufacturer. The hybridization patterns were converted into binary and octal formats and compared with previously reported strains in the in-house SITVIT2 proprietary database of Institut Pasteur de la Guadeloupe, which is an updated version of the recently released SITVITWEB database [19]. In this database, a Spoligotype International Type (SIT) designates a pattern shared by 2 or more patient isolates, whereas "orphan" represents a pattern reported for a single isolate. At the time of this comparison, SITVIT2 contained genotyping data on more than 110,000 MTBC clinical isolates from 160 countries of patient origin, and >3500 SITs. Major phylogenetic clades were essentially assigned according to signatures provided earlier [19]; these included specific signatures for various MTBC members as well as rules defining major lineages/sublineages for *M. tuberculosis* sensu stricto. The latter included the Beijing clade, the Central-Asian (CAS) clade and its 2 sublineages, the East-African-Indian (EAI) clade with its 9 sublineages, the Haarlem clade and its 3 sublineages, the Latin-American-Mediterranean (LAM) clade and its 10 sublineages, the "Manu" family and its 3 sublineages, the IS*6110*-low banding X clade and its 3 sublineages, and the ill-defined T clade and its 5 sublineages. The high phylogeographical specificity of LAM10-CAM prototype SIT61 for Cameroon has led to its designation as the Cameroon lineage [20], whereas LAM7-TUR lineage was tentatively reclassified as the Turkey lineage [21].

We also studied the worldwide distribution of all major spoligotyping clusters in our study (clusters containing 3 or more isolates) by interrogating the database for their distribution in macrogeographical regions and subregions according to the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm; for further details please refer to the footnote of Table 3). Lastly, the distribution of predominant spoligotype patterns in this study (n = 105 strains) was further compared to all other strains reported from Ethiopia in the international database (n = 1507 strains excluding the present study) as well as the five African subregions (n = 13656 strains), and the neighboring Western Asian countries (n = 4790 strains; for further details please refer to the footnote of Table 4).

Minimum spanning trees (MSTs) illustrating evolutionary relationships between the *M. tuberculosis* spoligotypes and various sociodemographic variables were constructed using BioNumerics software version 6.6 (Applied Maths, Sint-Martens-Latem, Belgium). MST is an undirected network in which all of the samples are linked together with the fewest possible linkages between nearest neighbors. Lastly, Spoligoforest trees which represent another way to illustrate

Table 1: Orphan strains ($n = 9$) and corresponding spoligotyping defined lineages/sublineages recorded among a total of 105 *M. tuberculosis* strains isolated in Afar Region, Northeast Ethiopia.

Iso number	Year	Strain	Spoligotype description	Octal code	Lineage	SIT
ETH0720100D42	2010	D42	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■□□□□■■□□□■■	777777377760611	T4	O1
ETH0720100D56	2010	D56	■■■■■■■■■□■■■□■■■□■■■■■■■■■■■■□□□□■■■■■■■	777356777760771	X1	O2
ETH0720100D168	2010	D168	■■■■■■■■□■■■■□□■■■■■■■■■■■■■■■■■□□■■■■■■■■■	776717777763771	Manu2	O3
ETH0720100A22596	2010	A22596	■■■■■■■■□□□□□□□■□□■■■■■■■■■■■■■□□□□■■■■■■■	776004777760771	T1	O4
ETH0720100B36	2010	B36	■■■□□□□■■■■■■■■■■■■■□□■■■□□■■■■■□■■■■■■■■	703777747175771	Unknown	O5
ETH0720100B55	2010	B55	■■■■■■□□□□■■■■■■■■■□□■■■■■■■■■■■■■■■□■■■■	770377477777671	AFRI	O6
ETH0720100B74	2010	B74	■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	777777427760771	T1	O7
ETH0720100AM6	2010	AM6	■■■■■■□□□■■■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■	761737777760771	T3	O8
ETH0720100S22	2010	S22	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■□■■■■	777777777763671	Manu2	O9

*Lineage designations for orphan patterns were done manually as expert-based interpretations using revised SpolDB4 rules.

probable strain evolutionary relationships between spoligotypes were drawn [22, 23] using spolTools online utilities available through http://www.emi.unsw.edu.au/spolTools. As opposed to the MSTs, the spoligoforest trees are directed and only evolve by loss of spacers. GraphViz software (http://www.graphviz.org/ [24]) was used to color the strains based on their lineages on the spoligoforest trees.

2.6. Ethics Statement. The study has been ethically cleared by the Norwegian Ethics Committee and the Ethiopian National Ethics Committee. Written informed consent was obtained from each study participant.

2.7. Statistical Analysis. Chi-square test using STATA software version 12 was used to evaluate whether or not significant association may exist between lineages or SITs with sociodemographic and epidemiological characteristics. Pearson's Chi-square test was used when more than 80% of data had an expected value greater than 5 and Fisher's Exact Test for remaining data with smaller expected values (at least 20% of data having values less than 5). P values less than 0.05 were considered as statistically significant.

3. Results

3.1. Demographic Characteristics of Study Participants. Among 325 pulmonary TB suspects, a total of 105 patients were confirmed to have culture positive pulmonary TB and included for further analysis. The median age of patients was 29 (IQR: 22.5–40.0) years and the majority were males with a male to female ratio of 1.92. Thirty-nine (37%) of culture confirmed TB patients were positive for acid fast bacilli on sputum smear examination. Besides, HIV result was available for 95 of the patients and 39% of them were found to be HIV co-infected. The majority (77.1%) of the study participants were new cases of pulmonary TB. Pastoralists represented 42.9% of the study participants. The proportion of HIV positives among ethnic Afar (8.7%) was significantly lower compared to other ethnic groups (28.0–47.5%) ($P < 0.001$).

3.2. Genetic Diversity and Distribution of Isolates. Overall, 103 (98.1%) isolates had spoligotypes characteristic of *M. tuberculosis* complex and 2 (1.9%) isolates had spoligotypes characteristic of *M. bovis*. A total of 9 isolates showed orphan patterns (Table 1), whereas 16 strains were unique making the total number of unclustered isolates to be 25 (23.81%). On the other hand, 19 SITs containing 80 isolates (76.19%) were clustered within this study (2 to 10 isolates per cluster). Besides, a total of 44 spoligotype patterns were observed and, therefore, the overall diversity of the isolates was 41.9%. A total of 27 SITs containing 83 isolates matched a preexisting shared-type in the SITVIT2 database, whereas 8 SITs ($n = 13$ isolates) did not match any of the previous isolates in the database and hence were newly created (Table 2).

The majority (76.19%) of the isolates belonged to the following three dominant lineages: T ($n = 43$, 40.95%), Manu ($n = 20$, 19.05%), and CAS ($n = 17$, 16.19%). The other minor lineages in this study include X ($n = 6$, 5.71%), Haarlem ($n = 4$, 3.81%), EAI ($n = 4$, 3.81%), LAM ($n = 3$, 2.86%), BOV ($n = 2$, 1.90%), AFRI ($n = 1$, 0.95%), Turkey ($n = 1$, 0.95%), and S ($n = 1$, 0.95%) (Table 2). Three isolates (2.86%) had patterns with signatures that do not belong to any of the major lineages described in the database and hence labelled as "unknown" (Tables 1 and 2). No Beijing strains were identified in this study.

The dominant SITs in this study included SIT 149 of the T3-ETH lineage, SIT 37 of the T3 lineage, SIT 53 of the T1 lineage, and SIT 54 of the Manu2 lineage all together accounting for 36.2% of the isolates (Table 3). On comparing the dominant isolates from this study with the isolates in the SITVIT2 database from Ethiopia, the five African regions, and neighboring Western Asia, statistically significant differences were observed (Table 4). Significantly higher proportions of T2 ($P < 0.01$), X1 ($P < 0.01$) and H ($P = 0.024$) lineages/sublineages were found among isolates in this study compared to their representations in the SITVIT2 database from Ethiopia. Notably, the Manu lineage was significantly higher compared to previous isolates from Africa and neighboring Western Asia ($P < 0.01$). Besides, this lineage has not been reported to SITVIT2 database from Ethiopia so far.

Table 2: Description of 35 shared-types (SITs; $n = 96$ isolates) and corresponding spoligotyping defined lineages/sublineages starting from a total of 105 *M. tuberculosis* strains isolated in Addis Ababa (Ethiopia).

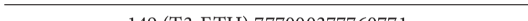
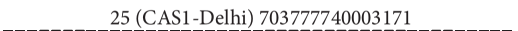
SIT*	Spoligotype description	Octal number	Nb (%) in study	% in study versus database	Lineage**	Clustered versus unique patterns***
10	■□□■■■■■■■■■■■■■□■□■■■■■■■□□□□■□■■■■■■■■■	477777277413771	1 (0.95)	1.14	EAI8-MDG	Unique
20	■■□■■■■■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	677777607760771	2 (1.9)	0.24	LAM1	Clustered
25	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□■■□□■■■■■	703777740003171	4 (3.81)	0.64	CAS1-Delhi	Clustered
26	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■■	703777740003771	4 (3.81)	0.26	CAS1-Delhi	Clustered
34	■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	776377777760771	1 (0.95)	0.12	S	Unique
37	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777737777760771	10 (9.52)	2.04	T3	Clustered
41	■■■■■■■■■■■■■■■■■■□□□□□■□□■■■■■■■□□□□■■■■■■■	777777404760771	1 (0.95)	0.21	LAM7-TUR	Unique
50	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	777777777720771	1 (0.95)	0.03	H3	Unique
52	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	777777777760731	6 (5.71)	0.65	T2	Clustered
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777777777760771	9 (8.57)	0.14	T1	Clustered
54	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■	777777777763771	9 (8.57)	3.56	Manu2	Clustered
100	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■	777777777773771	1 (0.95)	1.22	Manu1	Unique
119	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■■■■■■	777776777760771	3 (2.86)	0.28	X1	Clustered
121	■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■□■□□□□■■■■■■■	777777775720771	3 (2.86)	5.77	H3	Clustered
149	■■■■■■■■■□□□□□□□□□■■■■■■■■■■■■■■■■□□□□■■■■■■■	777000377760771	10 (9.52)	2.07	T3-ETH	Clustered
241	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□□□□■■	777777777760411	1 (0.95)	3.7	T1	Unique
247	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□■■■□□■■■■	703777740003471	2 (1.9)	5.13	CAS1-Delhi	Clustered
289	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□■■■□■■■■■	703777740003571	2 (1.9)	15.38	CAS1-Delhi	Clustered
336	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■■□■■■	777776777760731	2 (1.9)	1.71	X1	Clustered
523	■■	777777777777771	3 (2.86)	6.25	Manu_ancestor	Clustered
804	■□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	477777777760771	1 (0.95)	3.85	T1	Unique
1264	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□□	703777740000000	1 (0.95)	4.35	CAS1-Delhi	Unique
1338	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■	777777777777671	1 (0.95)	16.67	Unknown	Unique
1967	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□■■■	703777740000031	2 (1.9)	28.57	CAS	Clustered
2693	■■■□□□□■■■■■■■□□■■■■■■□□□□□□□□□□□□■■■■■■■■■	703763740003771	1 (0.95)	11.11	CAS1-Delhi	Unique
2781	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■	777777777767771	1 (0.95)	11.11	Unknown	Unique
2820	■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■□□□□■■■□■■■	777777767760731	1 (0.95)	14.29	T2	Unique
3896*	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□	777737777760000	1 (0.95)	50	T3	Unique
3897*	■■■■■■■■■□■■■□□■■■■■■■■■■■■■■■■■■□□■■■■□■■■■	777347777763671	3 (2.86)	75	Manu2	Clustered
3898*	■■□■■■■■■■■■■■■■■□□□■■■■■■■□□□□■□■■■■■■■■■	677777077413771	1 (0.95)	50	EAI2	Unique
3899*	■■■□□□□■■■■■■■■■■■■■■□□□□□□□□□□□□□■■■■■■■■	703777700001771	1 (0.95)	50	CAS	Unique
3900*	■■■■■■■■■□■■■□□■■■■■■■■■■■■■■■■■■□□■■■■■■■■■	777347777763771	2 (1.9)	66.67	Manu2	Clustered
3901*	□□□□□□□□□■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□□	000773777777600	2 (1.9)	66.67	BOV	Clustered
3902*	■■■■■■■■■■■■■■■■■■■□□■□■■■■■□□□□■□■■■■■■■■■	777777457413771	2 (1.9)	66.67	EAI6-BGD1	Clustered
3903*	■■□■■■■■■■■■■■■■■■■□□□□□■■■■■■■■■□□□□■■■■■■■	677777407760771	1 (0.95)	50	LAM1	Unique

A total of 27/35 SITs containing 83 isolates matched a preexisting shared-type in the database, whereas 8/35 SITs ($n = 13$ isolates) were newly created. A total of 19/35 SITs containing 80 isolates were clustered within this study (2 to 10 isolates per cluster) while 16/35 SITs containing 16 strains were unique (for total unique strains, one should add to this number the 9 orphan strains which brings the number of unclustered isolates in this study to 25/105 (23.81%) and clustered isolates to 80/105 (76.19%)). Note that SITs followed by an asterisk indicates "newly created" SITs due to 2 or more strains belonging to an identical new pattern within this study or after a match with an orphan in the database; SIT designations followed by number of strains: 3896 this study $n = 1$. DEU $n = 1$; 3897* this study $n = 3$. USA $n = 1$; 3898* this study $n = 1$. AUS $n = 1$; 3899* this study $n = 1$. NPL $n = 1$; 3900* this study $n = 2$. CHN $n = 1$; 3901* this study $n = 2$. FXX $n = 1$; 3902* this study $n = 2$. MYS $n = 1$; 3903* this study $n = 1$. BRA $n = 1$.

**Lineage designations according to SITVIT2 using revised SpolDB4 rules; "unknown" designates patterns with signatures that do not belong to any of the major lineages described in the database.

***Clustered strains correspond to a similar spoligotype pattern shared by 2 or more strains "within this study"; as opposed to unique strains harboring a spoligotype pattern that does not match with another strain from this study. Unique strains matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as "orphan" (see Table 1).

Table 3: Description of clusters containing 3 or more isolates in this study and their worldwide distribution in the SITVIT2 database.

SIT (Lineage) Octal number spoligotype description	Nb (%) in study	% in study versus database	Distribution in regions with ≥3% of a given SITs*	Distribution in countries with ≥3% of a given SITs**
37 (T3) 777737777760771	10 (9.52)	2.04	AFRI-E 20.37, EURO-N 12.63, EURO-W 12.22, ASIA-W 11.2, AMER-N 9.16, ASIA-E 6.52, EURO-S 6.11, EURO-E 4.68, ASIA-S 3.87, AMER-S 3.67, AFRI-S 3.46	ETH 17.72, USA 8.15, SWE 5.5, SAU 5.3, CHN 5.3, FXX 4.48, ITA 4.28, ZAF 3.46, DNK 3.46, FIN 3.06
149 (T3-ETH) 777000377760771 	10 (9.52)	2.07	AFRI-E 58.06, EURO-N 21.69, ASIA-W 7.03, EURO-W 6.41, AMER-N 4.75	ETH 58.06, DNK 11.78, SAU 6.61, USA 4.75, SWE 4.55, NLD 3.51
53 (T1) 777777777760771	9 (8.57)	0.14	EURO-W 15.71, AMER-N 13.54, AMER-S 12.08, EURO-S 9.45, EURO-N 7.51, ASIA-W 7.34, AFRI-S 4.99, AFRI-E 4.67, ASIA-E 4.25, AFRI-N 3.53, EURO-E 3.28, AMER-C 3.25	USA 13.25, FXX 7.91, ITA 5.36, BRA 5.15, ZAF 4.88, TUR 3.49, AUT 3.44, CHN 3.1
54 (Manu2) 777777777763771	9 (8.57)	3.56	ASIA-W 16.6, ASIA-S 14.63, ASIA-E 14.63, AFRI-N 14.23, AMER-N 7.91, AFRI-S 6.32, AFRI-E 5.14, ASIA-N 3.95, AMER-S 3.56	CHN 14.63, IND 13.83, EGY 13.44, USA 7.91, SAU 7.91, ZAF 6.32, IRQ 5.53, RUS 3.95, ETH 3.56
52 (T2) 777777777760731	6 (5.71)	0.65	EURO-W 19.63, ASIA-E 15.29, EURO-N 13.88, AMER-N 12.04, EURO-S 5.32, ASIA-W 5.32, AFRI-E 4.99, EURO-E 4.34, AFRI-M 3.58, AMER-C 3.47	CHN 11.82, USA 11.61, SWE 9.76, FXX 9.0, BEL 4.88, ITA 3.47, JPN 3.25, ETH 3.25
25 (CAS1-Delhi) 703777740003171	4 (3.81)	0.64	ASIA-W 27.05, AFRI-E 24.96, AFRI-N 13.53, ASIA-S 12.4, AMER-N 8.37, EURO-N 5.8, EURO-W 4.35	ETH 23.83, SAU 15.14, SDN 11.11, USA 8.37, IND 7.89, OMN 6.76, IRQ 4.67
26 (CAS1-Delhi) 703777740003771	4 (3.81)	0.26	ASIA-S 50.0, AMER-N 14.47, ASIA-W 14.07, EURO-W 5.61, AFRI-E 4.95, EURO-N 4.36, EURO-S 3.3	IND 30.19, USA 14.47, PAK 9.51, OMN 7.2, BGD 5.48, SAU 5.15, IRN 4.1, ITA 3.17, ETH 3.1
119 (X1) 777776777760771 	3 (2.86)	0.28	AMER-N 58.87, AFRI-S 13.22, AMER-C 9.89, AMER-S 3.97, EURO-N 3.51	USA 58.23, ZAF 13.22, MEX 9.61
121 (H3) 777777775720771	3 (2.86)	5.77	ASIA-E 38.46, AFRI-E 19.23, EURO-W 11.54, EURO-E 7.69, EURO-N 5.77, ASIA-W 5.77, AMER-N 5.77, EURO-S 3.85	TWN 36.54, ETH 19.23, FXX 11.54, CZE 7.69, USA 5.77, SAU 5.77
523 (Manu_ancestor) 777777777777771	3 (2.86)	6.25	ASIA-E 27.08, ASIA-SE 16.67, AMER-N 10.42, EURO-W 8.33, ASIA-W 8.33, AFRI-W 8.33, AFRI-E 6.25, AFRI-S 4.17	USA 10.42, SAU 8.33, JPN 8.33, CHN 8.33, NGA 6.25, MYS 6.25, KOR 6.25, ETH 6.25, ZAF 4.17, VNM 4.17, TWN 4.17, THA 4.17, BEL 4.17
3897 (Manu2) 777347777763671	3 (2.86)	75	AFRI-E 75.0, AMER-N 25.0	ETH 75.0, USA 25.0

*Worldwide distribution (analysis made on August 18, 2013) is reported for regions with more than 3% of given SITs as compared to their total number in the SITVIT2 database. The definition of macrogeographical regions and subregions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 subregions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new subregion by itself (Northern Asia) instead of including it among the rest of Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern, and Southeastern Asia.

**The 3-letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ≥3% of given SITs as compared to their total number in the SITVIT2 database.

Table 4: Distribution in Ethiopia and neighboring subregions of 11 predominant spoligotype patterns in this study (clusters containing 3 or more isolates), as seen through the updated SITVIT2 database[a].

Octal code (SIT) lineage	Ethiopia			Distribution in five African subregions and neighboring Western Asia in SITVIT2[b]											
	This study N/t (%)	SITVIT2 N/t (%)	P value	AFRI-E N/t (%)	P-value	AFRI-N N/t (%)	P value	AFRI-W N/t (%)	P value	AFRI-M N/t (%)	P value	AFRI-S N/t (%)	P value	ASIA-W N/t (%)	P value
37 (T3) 777737777760771	10/105 (9.52)	77/1507 (5.11)	0.072	90/4138 (2.17)	0.0003*	7/2145 (0.33)	<0.0001*	6/1985 (0.30)	<0.0001*	0/808 (0.00)	<0.0001*	17/4580 (0.37)	<0.0001*	55/4790 (1.15)	<0.0001*
149 (T3-ETH) 777000377760771	10/105 (9.52)	271/1507 (17.98)	0.056	271/4138 (6.55)	0.26	1/2145 (0.05)	<0.0001*	0/1985 (0.00)	<0.0001*	0/808 (0.00)	<0.0001*	1/4580 (0.02)	<0.0001*	34/4790 (0.71)	<0.0001*
53 (T1) 777777777760771	9/105 (8.57)	154/1507 (10.22)	0.62	283/4138 (6.84)	0.52	221/2145 (10.30)	0.60	109/1985 (5.49)	0.21	24/808 (2.97)	0.006*	312/4580 (6.81)	0.51	459/4790 (9.58)	0.75
54 (Manu2) 777777777763771	9/105 (8.57)	0/1507 (0.00)	<0.0001*	4/4138 (0.10)	<0.0001*	36/2145 (1.68)	<0.0001*	3/1985 (0.15)	<0.0001*	0/808 (0.00)	<0.0001*	16/4580 (0.35)	<0.0001*	42/4790 (0.88)	<0.0001*
52 (T2) 777777777760731	6/105 (5.71)	24/1507 (1.59)	0.0036*	40/4138 (0.97)	0.001*	6/2145 (0.28)	<0.0001*	6/1985 (0.30)	<0.0001*	33/808 (4.08)	0.44	21/4580 (0.46)	<0.0001*	49/4790 (1.02)	0.0013*
25 (CAS1-Delhi) 703777740003171	4/105 (3.81)	144/1507 (9.56)	0.066	151/4138 (3.65)	0.79	84/2145 (3.92)	1.0	2/1985 (0.10)	<0.0001*	1/808 (0.12)	0.0009*	3/4580 (0.07)	<0.0001*	168/4790 (3.51)	0.79
26 (CAS1-Delhi) 703777740003771	4/105 (3.81)	43/1507 (2.85)	0.55	71/4138 (1.72)	0.12	16/2145 (0.75)	0.013*	0/1985 (0.00)	<0.0001*	0/808 (0.00)	0.0002*	6/4580 (0.13)	<0.0001*	213/4790 (4.45)	1.0
119 (X1) 777776777760771	3/105 (2.86)	1/1507 (0.07)	0.001*	4/4138 (0.10)	0.0005*	5/2145 (0.23)	0.005*	2/1985 (0.10)	0.0012*	0/808 (0.00)	0.0016*	143/4580 (3.12)	1.0	9/4790 (0.19)	0.002*
121 (H3) 777777775720771	3/105 (2.86)	7/1507 (0.46)	0.024*	7/4138 (0.17)	0.002*	0/2145 (0.00)	0.0001*	0/1985 (0.00)	0.0001*	0/808 (0.00)	0.0016*	0/4580 (0.00)	<0.0001*	3/4790 (0.06)	0.0002*
523 (Manu_ancestor) 777777777777771	3/105 (2.86)	0/1507 (0.00)	0.0003*	0/4138 (0.00)	<0.0001*	1/2145 (0.05)	0.0004*	4/1985 (0.20)	0.004*	0/808 (0.00)	0.0016*	2/4580 (0.04)	0.0001*	4/4790 (0.08)	0.0003*
3897 (Manu2) 777347777763671	3/105 (2.86)	0/1507 (0.00)	0.0003*	0/4138 (0.00)	<0.0001*	0/2145 (0.00)	0.0001*	0/1985 (0.00)	0.0001*	0/808 (0.00)	0.0016*	0/4580 (0.00)	<0.0001*	0/4790 (0.00)	<0.0001*

[a] Asterisk denotes statistically significant differences ($P < 0.05$) as compared to the total number of strains for a given pattern in this study ($n = 105$ strains). For statistical comparison of differences observed, note that Pearson's Chi-square test was used when more than 80% of data had an expected value greater than 5, and Fisher's Exact Test for remaining data with smaller values (at least 20% of data having values less than 5).

[b] The definition of subregions is according to the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm). The table summarizes distribution in this study ($n = 105$) versus all strains reported from Ethiopia in the international database ($n = 1507$, excluding the present study), the five African subregions ($n = 13656$), and neighboring Western Asian countries (ASIA-W, $n = 4790$). AFRI-E: Burundi, Comoros, Ethiopia, Kenya, Madagascar, Malawi, Mauritius, Mozambique, Reunion, Tanzania, Uganda, Zambia, and Zimbabwe; AFRI-N: Algeria, Egypt, Libyan Arab Jamahiriya, Morocco, Sudan, and Tunisia; AFRI-W: Benin, Burkina Faso, Gambia, Guinea, Guinea-Bissau, Ivory Coast, Mali, Mauritania, Nigeria, Senegal, and Sierra Leone; AFRI-M: Cameroon, Central African Republic, Chad, and Democratic Republic of Congo; AFRI-S: Botswana, Namibia, South Africa, and Swaziland; ASIA-W: Armenia, Azerbaijan, Georgia, Iraq, Israel and Palestinian territories, Oman, Saudi Arabia, Turkey, and Yemen (SITVIT2 database comparison made on August 21st 2013 by David Couvin and Nalin Rastogi).

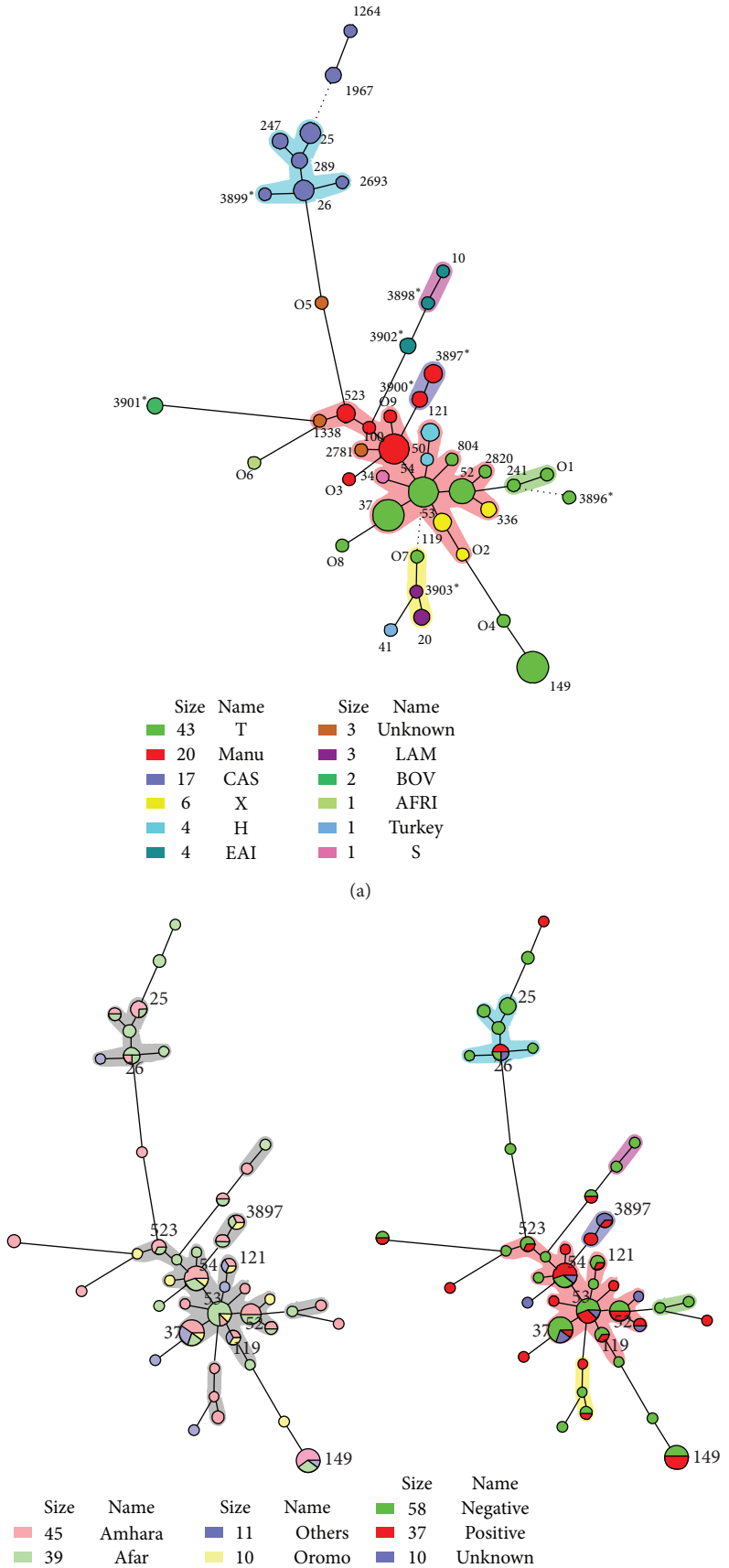


Figure 1: Continued.

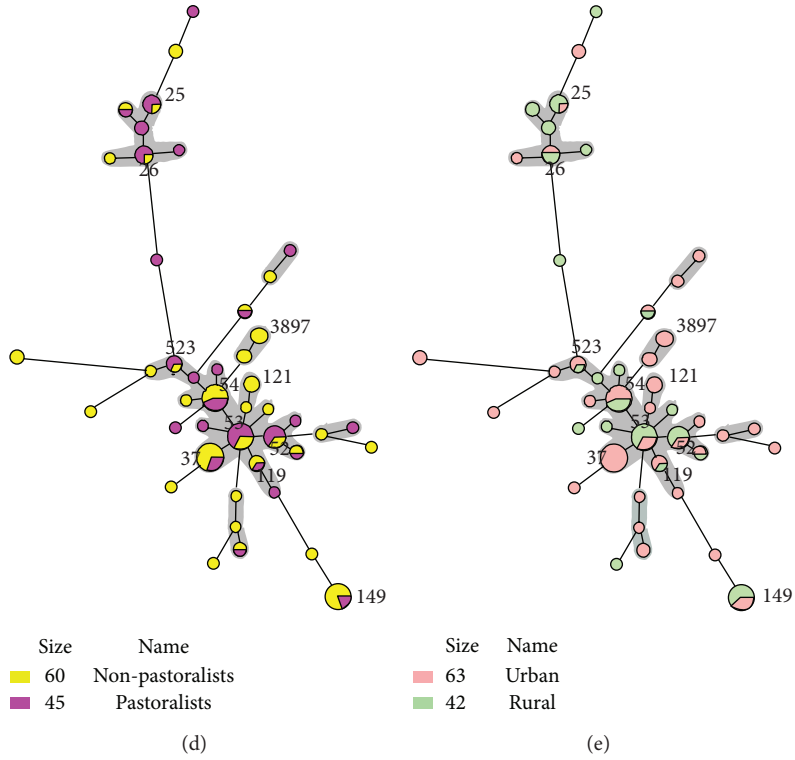


Figure 1: A minimum spanning tree (MST) illustrating evolutionary relationships between the *M. tuberculosis* spoligotypes (a) and various parameters studied such as ethnicity of patients (b), HIV serology (c), occupation (d), and residence (e). Note that the MSTs were constructed using BioNumerics 6.6 on all spoligotypes ($n = 105$). The phylogenetic tree connects each pattern based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous versus dashed and dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed, and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines, 1 or 2 or 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); grey dashed lines represent 4 changes; and dotted lines represent 5 or more changes. The size of the circle is proportional to the total number of isolates in our study, illustrating unique isolates (smaller nodes) versus clustered isolates (bigger nodes). The colour of the circles indicates the phylogenetic lineage to which the specific pattern belongs.

On the other hand, there was no significant difference in the distribution of T1, T3, T3-ETH, and CAS1-Delhi lineages compared to previous isolates from Ethiopia ($P > 0.05$).

A minimum spanning tree (MST) illustrating evolutionary relationships among the strains has been constructed (Figure 1). In this evolutionary tree, MANU strains were closest to the central node of the unrooted tree which is represented by SIT53, whereas CAS strains are further away from the central node of the tree. On spoligoforest trees (Figure 2), SIT149/T3-ETH and SIT37/T3 are the biggest nodes (each with 10 isolates), followed by SIT53/T1 and SIT54/Manu2 (each with 9 isolates); SIT52/T2 ($n = 6$), SIT25, and SIT26/CAS1-Delhi (each with 4 isolates) are the other dominant isolates in our study.

On investigating the association between strains and HIV co-infection, strains belonging to Manu2 lineage were significantly associated with HIV infection compared to strains belonging to CAS lineage (Fisher's Exact Test, $P = 0.019$). However, strain clustering was not significantly associated with HIV infection and other sociodemographic characteristics of the study participants.

4. Discussion

Analysis of strain diversity and comparison with international database is key to have an insight into the global distribution of *M. tuberculosis* strains. In this regard, we describe the strain diversity of *M. tuberculosis* complex among clinical isolates of pulmonary TB patients from a predominantly pastoralist area in Northeast Ethiopia and compared the strains with SITVIT2 database strains.

Among 105 isolates, the majority (98.1%) of the isolates were identified as *M. tuberculosis* complex with only 2 (1.9%) isolates being *M. bovis*. This finding is in agreement with previous reports from Ethiopia [25, 26]. Unlike a previous study from Ethiopia [27] which reported a significant contribution (17%) of bovine TB among TB lymphadenitis patients, the contribution of *M. bovis* to human TB seems to be minimal and mainly restricted to the pastoral communities as supported by current evidence [25, 26].

In our study, the overall clustering was 76.19% indicating a high rate of recent transmission in the study area. Although this high level of spoligotyping based clustering should be

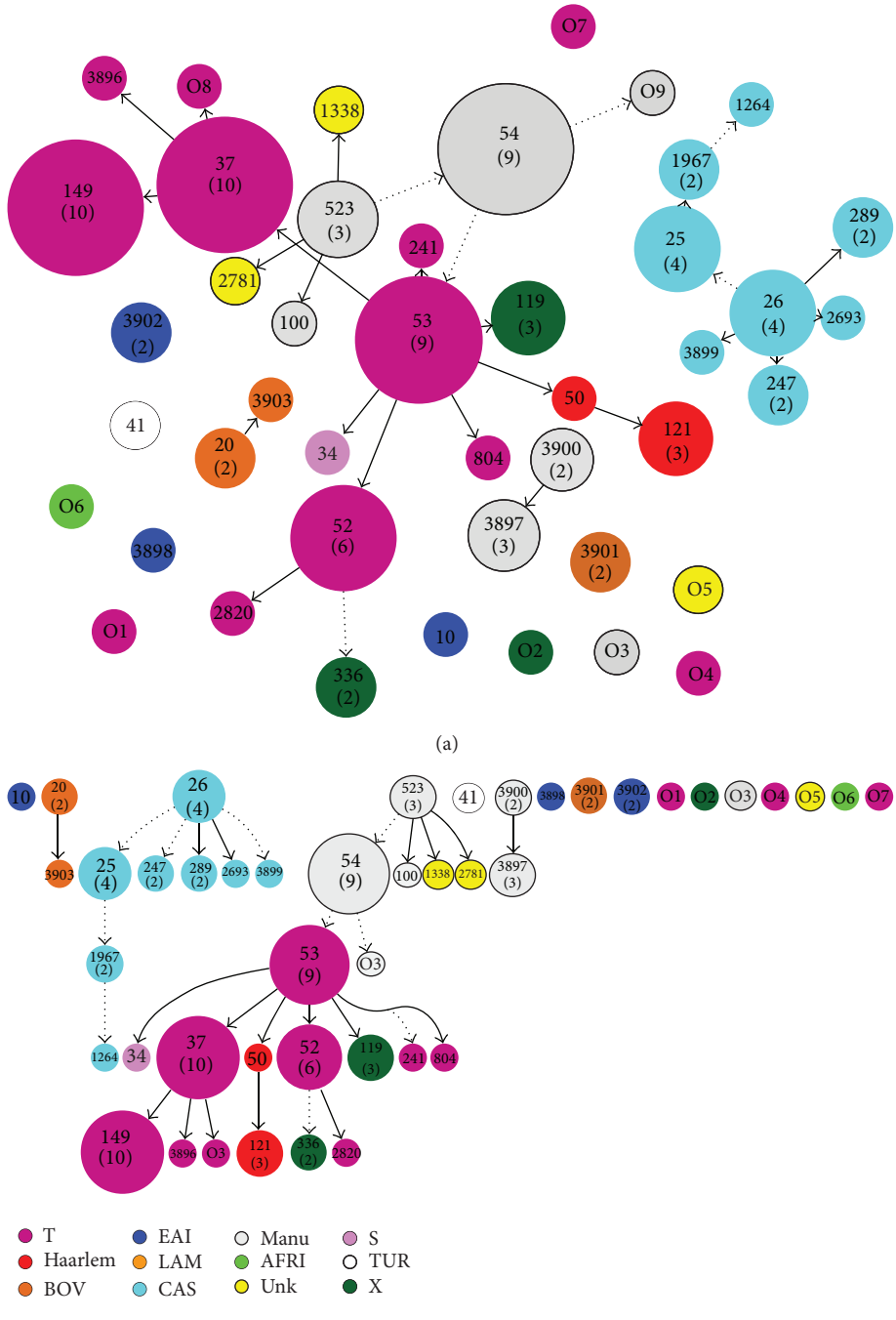


FIGURE 2: Spoligoforest trees were drawn using the spoligotyping data (n = 105 isolates) and the SpolTools software ([22]; available through on http://www.emi.unsw.edu.au/spolTools) and reshaped and colored manually by the GraphViz software (http://www.graphviz.org/; [24]). Two different trees were drawn using the Fruchterman Reingold algorithm (a) and the Hierarchical Layout (b). Note that the trees illustrate each spoligotype pattern from our study by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. The heuristic used selects a single inbound edge with a maximum weight using a Zipf model. The significance of the edges is the same for Hierarchical Layout and Fruchterman Reingold trees. Solid black lines link patterns that are very similar, that is, loss of one spacer only (maximum weigh being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1 and dotted lines a weight less than 0.5. In both trees, one can denote that SIT149/T3-ETH and SIT37/T3 are the biggest nodes (n = 10), followed by SIT53/T1 and SIT54/Manu2 (n = 9), SIT52/T2 (n = 6) and SIT25, and SIT26/CAS1-Delhi (n = 4), which are the other predominant patterns in our study.

further studied using additional genotyping markers, it is comparable with a previous study from Addis Ababa [8]. The recent national TB prevalence survey also argues for an ongoing TB epidemic mainly affecting the young probably indicating a defect in the TB control programme [28]. Besides, we have previously reported [29] that TB patients in Afar region suffered long diagnostic delays contributing to continued transmission with strain circulation.

Three dominant lineages were identified in this study and in agreement with previous reports [8, 9], the ill-defined T-lineage is the most dominant lineage accounting for 40.95% of the isolates. Among the T-lineages, T3-ETH is the most common lineage accounting for 25.6% of the T-lineage in this study. Similarly, previous studies from Ethiopia [8, 9, 11] reported a high proportion of T3-ETH among isolates indicating that this lineage is one of the dominant lineages in the country. CAS-Delhi lineage is the other dominant lineage in this study accounting for 16.19% of the isolates. In agreement with our finding, this lineage has been reported in previous studies as one of the dominant lineages in Ethiopia [7–9, 11] as well as in neighboring Djibouti [30]. EAI lineage was also isolated in smaller proportion in the study area. Both CAS and EAI are mainly prevalent in Middle East and Central Asia (Table 3). Two hypotheses could explain the presence of these lineages in Ethiopia as well: (i) it could have emerged in Ethiopia and migrated to Middle East and Central Asia, a hypothesis in agreement with the suggestion that East Africa is the cradle of *M. tuberculosis* complex species [31]; this is supported by a recent evidence [32] that MTBC co-evolved with the modern human host and migrated from Africa (particularly East Africa) to Asia and other parts of the world; (ii) alternatively, it could be due to migration of this lineage from Middle East and Central Asia to Ethiopia due to the recent human migrations from these areas to Ethiopia as suggested by a previous study [7].

Interestingly, Manu is one of the dominant lineages in our study contributing to 19.05% of the isolates. In agreement with this, investigators from Egypt reported a high proportion (27%) of Manu among their isolates [33]. In Ethiopia, only one study [9] from the Central highlands reported SIT 54 contributing to 8.33% of the isolates which is similar to our finding for SIT 54. However, Manu has not been reported from Ethiopia in other previous studies [7, 8, 11] as well as in the SITVIT2 database in which SIT 54 (Manu2) has been mainly reported from South and East Asia, Middle East including Egypt, and America. Furthermore, a recent study on a relatively large number of isolates mainly from Djibuti [34] did not find Manu lineage. The diverse distribution of strains in the study areas included in Ethiopia so far implies the need to map the distribution of *Mycobacterium tuberculosis* in the country. Interestingly, Manu2 lineage was found to be significantly associated with HIV infection although the reason for such association is not apparent from the current study.

Regarding the evolutionary relationship among strains, as opposed to CAS strains, the Manu strains in the MST tree are the ones closest to the dominant T lineage as reported before [33]. Besides, the T1 strain represented by its prototype SIT53 constitutes the central node of this unrooted tree and SIT 52 of T2 and SIT 37 of T3 are the other big nodes closest to the central node (Figure 1(a)). SIT149 (T3-ETH), which is the commonest lineage in Ethiopia, is further from the central node.

In conclusion, this study identified the presence of high level of clustering suggesting the presence of recent transmission in the study area. Nonetheless, one limitation of the present study is the fact that Mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTR) [35, 36] that allows splitting certain spoligotyping defined clusters in smaller subclusters [37] could not be used in the present study. Considering a recent study showing that some of the Manu lineage strains could result from mixed or polyclonal infections [38], and seeing the higher than usual proportion of Manu lineage strains in our study, it might be important to systematically perform MIRU typing of all spoligotyping clusters in future studies in the Afar pastoral region. We suggest that a nationwide study aiming to map the population structure of *M. tuberculosis* complex should be planned using spoligotyping, MIRU-VNTRs, and SNP in Ethiopia.

Conflict of Interests

The authors declared that they have no competing interests.

Authors' Contribution

Mulugeta Belay was involved in the design, data collection and laboratory work, statistical analysis, and interpretation of data and drafted the manuscript; Fekadu Abebe was involved in the design, interpretation, and critical revision of the paper; Gunnar Bjune was involved in the design, interpretation, and critical revision of the paper; Gobena Ameni was involved in data collection, spoligotyping of the isolates, and critical revision of the paper. Nalin Rastogi and David Couvin analyzed the strains and compared them with SITVIT2 database; besides, both were involved in the critical revision of the paper. All authors have read and approved the paper for submission.

Acknowledgments

This study was supported by the Norwegian Programme for Development, Research and Education (NUFU) (Project no. NUFUPRO.2007/10198), and the Norwegian Research Council (Project no. 196397/S50). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the paper. The authors' sincere gratitude goes to Mr. Adane Worku at Aklilu Lemma Institute of Pathobiology for carrying out spoligotyping. Laboratory technicians working at the respective health facilities deserve our sincere gratitude for collecting samples. The authors would like to acknowledge participants for consenting to participate in the study.

References

[1] D. Bleed, C. Dye, and M. C. Raviglione, "Dynamics and control of the global tuberculosis epidemic," *Current Opinion in Pulmonary Medicine*, vol. 6, no. 3, pp. 174–179, 2000.

[2] T. R. Frieden, T. R. Sterling, S. S. Munsiff, C. J. Watt, and C. Dye, "Tuberculosis," *The Lancet*, vol. 362, no. 9387, pp. 887–899, 2003.

[3] WHO, *Global Tuberculosis Report*, WHO, Geneva, Switzerland, 2012.

[4] Federal Ministry of Health, *Guidelines For Clinical and Programmatic Management of TB, Leprosy and TB/HIV in Ethiopia*, Federal Ministry of Health, Addis Ababa, Ethiopia, 12th edition, 2012.

[5] Z. Alebachew, "EthiopianNational TB Prevalence survey 2010–2011: Preliminary Result," http://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/meetings/lille_oct11_ethiopia.pdf.

[6] C. Dye and B. G. Williams, "The population dynamics and control of tuberculosis," *Science*, vol. 328, no. 5980, pp. 856–861, 2010.

[7] B. Tessema, J. Beer, M. Merker et al., "Molecular epidemiology and transmission dynamics of *Mycobacterium tuberculosis* in Northwest Ethiopia: new phylogenetic lineages found in Northwest Ethiopia," *BMC Infectious Diseases*, vol. 13, article 131, 2013.

[8] A. Mihret, Y. Bekele, A. G. Loxton et al., "Diversity of *Mycobacterium tuberculosis* isolates from new pulmonary tuberculosis cases in Addis Ababa, Ethiopia," *Tuberculosis Research and Treatment*, vol. 2012, Article ID 892079, 7 pages, 2012.

[9] L. Garedew, A. Mihret, G. Mamo et al., "Strain diversity of mycobacteria isolated from pulmonary tuberculosis patients at Debre Birhan Hospital, Ethiopia," *International Journal of Tuberculosis and Lung Disease*, vol. 17, no. 8, pp. 1076–1081, 2013.

[10] A. Deribew, G. Abebe, L. Apers et al., "Prevalence of pulmonary TB and spoligotype pattern of *Mycobacterium tuberculosis* among TB suspects in a rural community in Southwest Ethiopia," *BMC Infectious Diseases*, vol. 12, article 54, 2012.

[11] M. Agonafir, E. Lemma, D. Wolde-Meskel et al., "Phenotypic and genotypic analysis of multidrug-resistant tuberculosis in Ethiopia," *International Journal of Tuberculosis and Lung Disease*, vol. 14, no. 10, pp. 1259–1265, 2010.

[12] G. Mamo, F. Abebe, Y. Worku et al., "Tuberculosis in goats and sheep in afar pastoral region of ethiopia and isolation of *Mycobacterium tuberculosis* from goat," *Veterinary Medicine International*, vol. 2012, Article ID 869146, 8 pages, 2012.

[13] G. Mamo, G. Bayleyegn, T. S. Tessema et al., "Pathology of camel tuberculosis and molecular characterization of its causative agents in pastoral regions of Ethiopia," *PLoS ONE*, vol. 6, no. 1, Article ID e15862, 2011.

[14] M. Belay, G. Bjune, G. Ameni, M. Abebe, and F. Abebe, "Serodiagnostic performance of rESAT-6-CFP-10 in the diagnosis of pulmonary tuberculosis in Ethiopia," *Mycobacterial Diseases*, vol. 1, no. 2, article 103, 2011.

[15] Federal Ministry of Health, *Tuberculosis, Leprosy and TB/HIV Prevention and Control Programme*, Federal Ministry of Health, Addis Ababa, Ethiopia, 4th edition, 2008.

[16] HAPCO, *Guidelines for HIV Counselling and Testing in Ethiopia*, HAPCO, Addis Ababa, Ethiopia, 2007.

[17] WHO, *Laboratory Services in Tuberculosis Control: Culture. Part III*, WHO, Geneva, Switzerland, 1998.

[18] J. Kamerbeek, L. Schouls, A. Kolk et al., "Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology," *Journal of Clinical Microbiology*, vol. 35, no. 4, pp. 907–914, 1997.

[19] C. Demay, B. Liens, T. Burguière et al., "SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 755–766, 2012.

[20] F. Koro Koro, Y. Kamdem Simo, F. F. Piam et al., "Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping," *Journal of Clinical Microbiology*, vol. 51, no. 1, pp. 299–302, 2013.

[21] O. Kisa, G. Tarhan, S. Gunal et al., "Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey," *PLoS ONE*, vol. 7, no. 1, Article ID e30331, 2012.

[22] J. F. Reyes, A. R. Francis, and M. M. Tanaka, "Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes," *BMC Bioinformatics*, vol. 9, article 496, 2008.

[23] C. Tang, J. F. Reyes, F. Luciani, A. R. Francis, and M. M. Tanaka, "spolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex," *Bioinformatics*, vol. 24, no. 20, pp. 2414–2415, 2008.

[24] J. G. E. Ellson, E. Koutsofios, S. C. North, and G. Woodhull, "Graphviz and Dynagraph—Static and Dynamic Graph Drawing Tools," 2002, http://www.graphviz.org/Documentation/EGKNW03.pdf.

[25] R. Firdessa, S. Berg, E. Hailu et al., "Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis," *Emerging Infectious Diseases*, vol. 19, no. 3, pp. 460–463, 2013.

[26] B. Gumi, E. Schelling, S. Berg et al., "Zoonotic transmission of tuberculosis between pastoralists and their livestock in southeast Ethiopia," *EcoHealth*, vol. 9, no. 2, pp. 139–149, 2012.

[27] D. Kidane, J. O. Olobo, A. Habte et al., "Identification of the causative organism of tuberculous lymphadenitis in Ethiopia by PCR," *Journal of Clinical Microbiology*, vol. 40, no. 11, pp. 4230–4234, 2002.

[28] A. Kebede, Z. Alebachew, F. Tsegaye et al., "Background document 6: The First Ethiopian National Population-based TB Prevalence Survey," http://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/meetings/tf5doc06_keyresults-prevalencesurveyethiopia2010.pdf.

[29] M. Belay, G. Bjune, G. Ameni, and F. Abebe, "Diagnostic and treatment delay among Tuberculosis patients in Afar Region, Ethiopia: a cross-sectional study," *BMC Public Health*, vol. 12, article 369, 2012.

[30] S. Godreuil, F. Renaud, M. Choisy et al., "Highly structured genetic diversity of the *Mycobacterium tuberculosis* population in Djibouti," *Clinical Microbiology and Infection*, vol. 16, no. 7, pp. 1023–1026, 2010.

[31] G. M Cristina, S. Brisse, R. Brosch et al., "Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*," *PLoS Pathogens*, vol. 1, no. 1, article e5, 2005.

[32] I. Comas, M. Coscolla, T. Luo et al., "Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans," *Nat Genet*, vol. 45, no. 10, pp. 1176–1182, 2013.

[33] Z. H. Helal, M. S. E.-D. Ashour, S. A. Eissa et al., "Unexpectedly high proportion of ancestral manu genotype *Mycobacterium tuberculosis* strains cultured from tuberculosis patients in Egypt," *Journal of Clinical Microbiology*, vol. 47, no. 9, pp. 2794–2801, 2009.

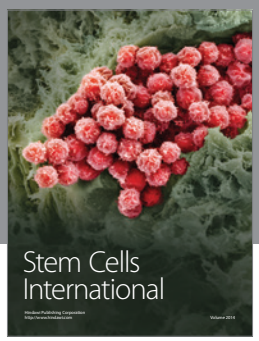
[34] Y. Blouin, Y. Hauck, C. Soler et al., "Significance of the identification in the horn of Africa of an exceptionally deep branching *Mycobacterium tuberculosis* clade," *PLoS ONE*, vol. 7, no. 12, Article ID e52841, 2012.

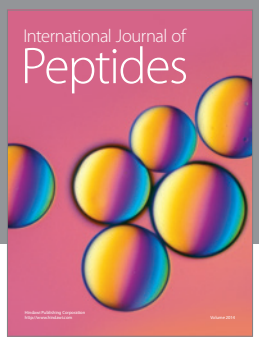
[35] P. Supply, S. Lesjean, E. Savine, K. Kremer, D. van Soolingen, and C. Locht, "Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units," *Journal of Clinical Microbiology*, vol. 39, no. 10, pp. 3563–3571, 2001.

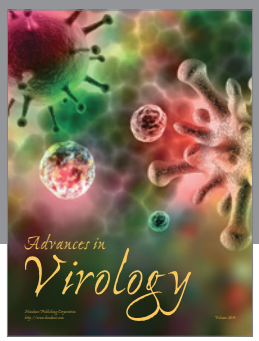
[36] P. Supply, C. Allix, S. Lesjean et al., "Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*," *Journal of Clinical Microbiology*, vol. 44, no. 12, pp. 4498–4510, 2006.

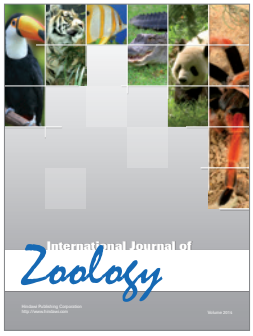
[37] C. Sola, I. Filliol, E. Legrand et al., "Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics," *Infection, Genetics and Evolution*, vol. 3, no. 2, pp. 125–133, 2003.

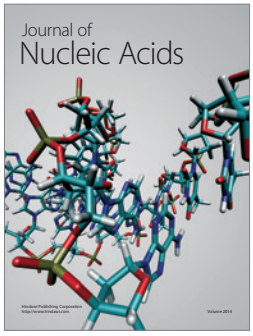
[38] L. C. O. Lazzarini, J. Rosenfeld, R. C. Huard et al., "*Mycobacterium tuberculosis* spoligotypes that may derive from mixed strain infections are revealed by a novel computational approach," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 798–806, 2012.

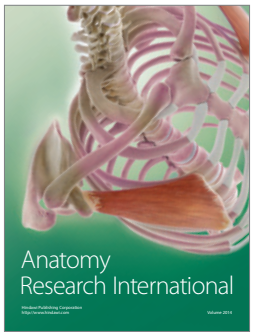


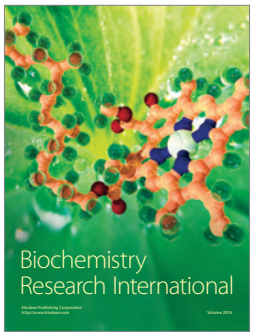


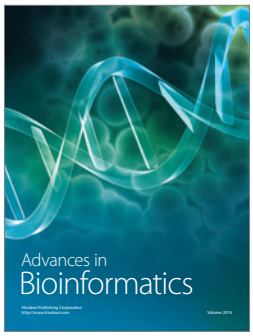


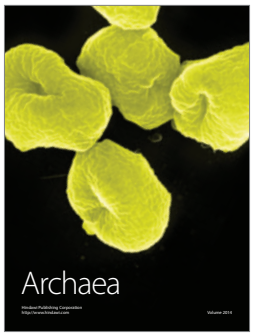


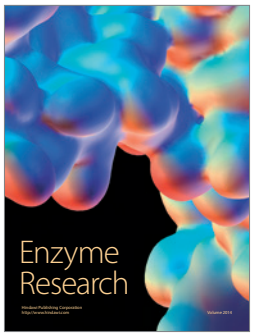


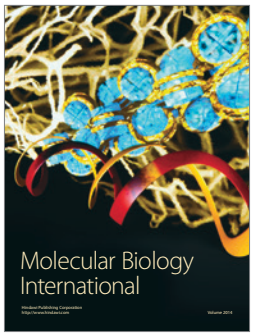












Hindawi

Submit your manuscripts at
http://www.hindawi.com

Article 12

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 580981, 15 pages
http://dx.doi.org/10.1155/2014/580981

Research Article

Molecular Epidemiology and Genotyping of *Mycobacterium tuberculosis* Isolated in Baghdad

Ruqaya Mustafa Ali,[1] Alberto Trovato,[2] David Couvin,[3] Amina N. Al-Thwani,[1] Emanuele Borroni,[2] Fahim H. Dhaer,[4] Nalin Rastogi,[3] and Daniela M. Cirillo[2]

[1] *Genetic Engineering and Biotechnology Institute for Post Graduate Studies, University of Baghdad, 10070 Jadriyah, Baghdad, Iraq*
[2] *Emerging Bacterial Pathogens Unit, San Raffaele Scientific Institute (HSR), via Olgettina, 60 20132 Milano, Italy*
[3] *Institut Pasteur de la Guadeloupe, Abymes, 97183 Guadeloupe, France*
[4] *Ministry of Agriculture, Al Nidhal Street, Baghdad 5157, Iraq*

Correspondence should be addressed to Daniela M. Cirillo; cirillo.daniela@hsr.it

Received 7 November 2013; Revised 17 January 2014; Accepted 18 January 2014; Published 26 February 2014

Academic Editor: Tomasz Jagielski

Copyright © 2014 Ruqaya Mustafa Ali et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tuberculosis (TB) remains a major health problem in Iraq but the strains responsible for the epidemic have been poorly characterized. Our aim was to characterize the TB strains circulating in Bagdad (Iraq). A total of 270 *Mycobacterium tuberculosis* complex (MTBC) strains isolated between 2010 and 2011 from TB patients attending the Center of Chest and Respiratory diseases in Baghdad were analyzed by Spoligotyping. The analysis indicated that 94.1% of the isolates belong to known genotype clades: CAS 39.6%, ill-defined T clade 29.6%, Manu 7.4%, Haarlem 7%, Ural 4.1%, LAM 3.3%, X 0.7%, LAM7-TUR 0.7%, EAI 0.7%, S 0.7%, and unknown 5.9%. Comparison with the international multimarker database SITVIT2 showed that SIT 309 (CAS1-Delhi) and SIT1144 (T1) were the most common types. In addition, 44 strains were included in SITVIT2 database under 16 new Spoligotype International Types (SITs); of these, 6 SITs (SIT3346, SIT3497, SIT3708, SIT3790, SIT3791, and SIT3800) ($n = 32$ strains) were created within the present study and 10 were created after a match with an orphan in the database. By using 24-loci MIRU-VNTR-typing on a subset of 110 samples we found a high recent transmission index (RTI) of 33.6%. In conclusion, we present the first unifying framework for both epidemiology and evolutionary analysis of *M. tuberculosis* in Iraq.

1. Introduction

Tuberculosis (TB) is an ancient disease that currently represents an immense global health challenge. In 2011, WHO estimated that globally there were 8.7 million new cases of active TB leading to 1.4 million reported deaths [1].

According to the 2012 report of the Ministry of Health (MOH), the incidence rate of TB in Iraq was 45/100,000, with 13,860 new TB cases and 1140 of previously treated cases. The Iraqi laboratory network includes 124 district smear microscopy laboratories and one national reference laboratory located in Baghdad performing cultures and drug susceptibility testing of *M. tuberculosis* [2].

In the last decades, a large number of different molecular methods based on DNA fingerprints have been developed. The usefulness of these methods has been demonstrated primarily as epidemiological markers to discriminate the pathogen at the genus, species, and subspecies level. The level of strain differentiation is of crucial importance for the study of transmission dynamics, determining whether the infection is caused by single strain or by multiple strain and if recurrence of the disease is due to treatment failure or infection with new strain of *M. tuberculosis* [3, 4]. Spoligotyping, targeting the Direct Repeat locus and Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeats (MIRU-VNTRs) typing, has been shown to be a valuable alternative to IS*6110* [5, 6]. An optimized 24-loci MIRU-VNTR typing scheme has been proposed as international standard [7, 8]. In addition to their use for tracing TB transmission at the strain level, MIRU-VNTR markers are

also phylogenetically more informative, especially in the 24-locus format and can therefore be used to predict grouping into strain lineage [9, 10].

Although TB is still a public health problem in Iraq, there is little information about the genetic characteristics of the isolates driving the epidemic. A better knowledge of the molecular characteristics of *M. tuberculosis* complex isolates could contribute to better understanding of the transmission dynamics of the disease within the country and can guide interventions to control the epidemic. The aim of this study is to determine molecular epidemiology features of *M. tuberculosis* isolates in Baghdad, as well as in other Iraqi governorates, to assess their transmission dynamics.

2. Materials and Methods

2.1. Study Population. This study was conducted in Baghdad at the Center of Chest and Respiratory diseases. A total of 270 isolates were collected between 2010 and 2011 representing approximately 40% of new and previously treated TB patients. This study was approved by the local ethical committee.

2.2. Culture and Drug Susceptibility Test. Diagnostic specimens were cultured and isolated on Lowenstein-Jensen (LJ) media after decontamination. Drug Susceptibility Tests (DST) against first-line anti-TB drugs rifampicin (RFP), isoniazid (INH), streptomycin (SM), and ethambutol (EMB) were combined with LJ medium at the following concentrations: RFP 40 μg/mL, INH 0.2 μg/mL, SM 4 μg/mL, and EMB 2 μg/mL. DST was used to detect the drug resistance of *M. tuberculosis* by the proportion method [11].

2.3. Identification of Mycobacterium tuberculosis. Mycobacterium tuberculosis isolates were identified on the basis of colony morphology, growth rate, pigmentation properties, niacin accumulation, nitrate reduction, thiophene-2-carboxylic acid hydrazide (TCH), and para-nitrobenzoic acid (PNB) test [11].

2.4. DNA Extraction. DNA was extracted from cultures by the standard Cetyltrimethyl ammonium-bromide (CTAB) method. The DNA was stored in TE buffer (10 mM Tris, 1 mM EDTA) at −20°C until use [12].

2.5. Spoligotyping. All genotyping methods were performed at the Emerging Bacterial Pathogens Unit, WHO/IUATLD Supra-National Reference TB Laboratory, San Raffaele Scientific Institute (FCSR). Spoligotyping analysis was performed by using commercial kit (Ocimun Biosolutions) as described by Kamerbeek et al. [13]. The 43 spacers between the direct repeats in the target region were amplified by using DRa biotinylated at 5′ end and DRb primers. The PCR products were hybridized to a membrane containing 43 oligonucleotides by reverse line blotting. *M. tuberculosis* H37Rv and *M. bovis* BCG were used as positive controls in each run. Spoligotyping results were converted into octal code and analyzed by using the SITVIT2 proprietary database of the Pasteur Institute of Guadeloupe, which is an updated version of the previously released SpolDB4 and SITVITWEB databases [14, 15].

2.6. MIRU-VNTR. Standardized 24-loci MIRU-VNTR typing [7] was performed using the MIRU-VNTR typing kit (Genoscreen, Lille, France). PCR products were run with Genescan 1200LIZ size standard (Applied Biosystems, California, and USA) on ABI3730 sequencer. Sizing of the PCR fragments and assignment of MIRU-VNTR alleles were done by Gene Mapper software version 3.7 (Applied Biosystems, California, USA). In order to deine clusters and to build an UPMGA tree, we used the MIRU-VNTRplus web application available at http://www.miru-vntrplus.org/MIRU/index.faces. The allelic diversity of the strains was determined by using the Hunter Gaston Discriminatory Index (HGDI).

2.7. Interpretation of Typing Results. Spoligotypes in binary format and MIRU patterns in 24-digit codes were entered in the SITVIT2 database. At the time of the comparison, it contained genotyping data on more than 100,000 MTBC strains isolated from 160 countries of patient origin. In this database, "SIT" (Spoligotype International Type) designates spoligotyping shared by two or more patient isolates, "MIT" (MIRU International Type) designates 24-loci MIRU patterns shared by two or more patient isolates, and "orphan" designates patterns reported for a single isolate. Major spoligotyping-based phylogenetic clades were assigned according to revised signatures provided in SITVITWEB [15]. These clades include specific signatures for *M. tuberculosis* complex members and rules defining major lineages/sublineages for *M. tuberculosis* sensu stricto, that is, the Central Asian (CAS) clade and 2 sublineages, the East-African-Indian (EAI) clade and 9 sublineages, the Haarlem (H) clade and 3 sublineages, the Latin-American-Mediterranean (LAM) clade and 12 sublineages note that two LAM sublineages were recently raised to lineage level: LAM10-CAM as the Cameroon lineage [16] and LAM7-TUR as the Turkey lineage [17], the "Manu" family and 3 sublineages, the IS*6110*-low-banding X clade and 4 sublineages, and an ill-defined T clade with 5 sublineages. The recently described "Ural" family, subdivided into 2 sublineages (Ural-1 and Ural-2), replaced some spoligotype patterns previously classified as H3 and H4 [18].

2.8. Phylogenetic Analysis. The evolutionary relationships among all the observed spoligotypes were studied by drawing minimum spanning trees (MSTs) with BioNumerics Software version 6.6 (Applied Maths, Sint-Martens-Latem, Belgium).

3. Results

3.1. Studied Population. This study was performed over a period of 13 months on 270 *M. tuberculosis* complex strains from Iraqi patients. The related demographic information and drug resistance patterns obtained are summarized in Table 1. One hundred and fifty-seven (58.14%) of the cases included in the study were from Baghdad, 113 (41.85%) were from other governorates in Iraq but diagnosed in Baghdad.

Table 1: Demographic characteristics of the population studied and drug resistance profiles.

Parameters	Origin of patients		Total
	Baghdad (%)	Other governorates (%)	
Total no. of strain	157 (58.1)	113 (41.9)	270
Sex			
Male	94 (59.9)	85 (75.2)	179
Female	63 (40.1)	28 (24.8)	91
Sex ratio M : F	1.49	3.04	1.97
Age group (yr) (no.)			
0–20	17 (10.8)	4 (3.5)	21
21–40	82 (52.2)	71 (62.8)	153
41–60	48 (30.6)	25 (22.1)	73
>60	10 (6.4)	13 (11.5)	23
Smear microscopy (no.)			
Positive	98 (62.4)	69 (61.1)	167 (61.8)
Negative	59 (37.6)	44 (38.9)	103 (38.2)
Treatment			
New cases	119 (75.8)	68 (60.2)	187 (69.3)
P.T.C	38 (24.2)	45 (39.8)	83 (30.7)
Drug susceptibility			
Susceptible to all drugs	102 (65)	69 (61.1)	171 (63.3)
MDR (INH – RIF)	31 (19.8)	33 (29.2)	64 (23.7)
Unknown	24 (15.3)	11 (9.7)	35 (13.0)

As shown in the table, male patients are predominants (male to female sex ratios varied from 1.49 to 3.04 depending on the origin of the patients, from Baghdad or other governorates, resp.) and 56.7% of the TB patients included in the study are between 21 and 40 years old with the mean age 36.3. As for smear microscopy, 167 tested positive and 103 were negative. 187 patients had newly TB diagnosed, while 83 patients were previously treated cases (PTC). Lastly, 171 cases were susceptible to all antituberculosis drugs, whereas 64/270 (23.7%) patients had MDR. 35 patients had an unknown resistance profile.

3.2. Distribution of Phylogenetic Clades. Spoligotyping results are summarized in Tables 2 and 3. A total of 117 different patterns were observed among 270 strains studied; a total of 53 patterns corresponding to 53 strains belonged to orphan patterns not yet reported to the SITVIT2 database (Table 2), whereas the remaining 64 patterns ($n = 217$) corresponded to SITs, of which 48 patterns ($n = 173$ strains) matched a preexisting shared type in the SITVIT2 database, while 16 SITs ($n = 44$ strains) were newly created. Of these, 6 SITs (SIT3346, SIT3497, SIT3708, SIT3790, SIT3791, and SIT3800) ($n = 32$ strains) were created within the present study and 10 were created after a match with an orphan in the database (SIT3275 from France metropolitan area, SIT3789 from Iraq, SIT3792, SIT3793, and SIT3794 from India, SIT3795 and SIT3799 from Brazil, SIT3796 from Pakistan, SIT3797 from Mexico, and SIT3798 from Iraq).

Nearly 94.1% of the isolates belonged to known genotype clades. These include, in decreasing order: CAS 39.6%, ill-defined T clade 29.6%, Manu 7.4%, Haarlem 7%, Ural 4.1%, LAM 3.3%, X 0.7%, LAM7-TUR 0.7%, EAI 0.7%, S 0.7%, and unknown 5.9%. This observation on the complex diversity of *M. tuberculosis* in Iraq was further supported by the minimum spanning tree (MST) illustrated in Figure 1, which was constructed on all isolates ($n = 270$, including 53 orphan patterns).

As shown in Figure 1, the CAS lineage strains constitute the biggest group of strains infecting Iraqi patients; the major shared types being SIT309, 26, 141, 25, 3497, and 22. At a considerable distance is seen a separate group of strains made up of smaller nodes: T lineage (SIT1144, 3346, 284, 118, 53), Haarlem (SIT75, 50), LAM (SIT42, 1470), Ural (SIT127), and Manu (SIT54). Even smaller nodes concerned other lineages such as Turkey, S, X, and the ancestral EAI lineages. Indeed, as many as 129/270 (47.7%) strains from this study belonged to Principal Genetic Group 1 (PGG1), according to the *KatG-gyrA* polymorphism-based classification of Sreevatsan et al. [19], and these strains represent all of known lineages associated with PGG1 strains (CAS, Manu, and EAI with the exception of Beijing). However, evolutionary modern PGG2/3 strains constitute a high proportion in this study such as ill-defined T, Haarlem, Ural, LAM, LAM7-TUR, S,

Table 2: Orphan strains ($n = 53$) and corresponding spoligotyping defined lineages/sublineages recorded among a total of 270 *M. tuberculosis* strains (one isolate per patient) from Iraqi patients.

Iso number	Year	Strain	Spoligotype description	Octal code	Lineage*	Label	Drug R**	Isolation city	Sex/age
IRQ02201004351	2010	4351	■■□■■■■■■■■■■■■■□■■■■■■■■□□□□■□■■■■■■■■	677777377413771	EAI5	O01	0	Wasit	F/60
IRQ02201023876	2010	3876	■■■□□□□■■■■■■■□■■■■■■■□□□□□□□□□□□□■■□□□■■■■	703767740003071	CASI-Delhi	O02	2	Baghdad	M/29
IRQ02201015527	2010	5527	□□□□■■■■■■■■□■□■■■■■■■■■■■■■■■■□□□□■■■■■■■	037727777760771	T3	O03	1	Babil	F/29
IRQ02201003954	2010	3954	□□□□■■■■■■■■□■□■■■■■■■■■■■■■■■■□□□□■■■■■■■□	037657777741770	Unknown	O04	0	Baghdad	M/22
IRQ02201113313	2011	3313	■□□□■■■■■■■■□■□■■■■■■■■■■■■■■■■□□□□■■■■■■■	437727777760771	T3	O05	1	Sulaimaniya	M/35
IRQ02201114243	2011	4243	■□■□■■■■■■■■□□■■■■■■■■■■■■■■■■□□□□■■■■■■■	537637777760771	T1	O06	1	Baghdad	F/35
IRQ02201014888	2010	4888	□□■■■■■■■■■■■■□■■■□■■■■■■■■□□□□□□□□■■□■■■■■	177767377400671	Unknown	O07	1	Diyala	M/33
IRQ02201016952	2010	6952	■□■■■■■■■■■■■■■■■■■■■■■■■■□■□■□□□□■■■■■■■	577777777520771	H3	O08	1	Baghdad	M/35
IRQ02201017197	2010	7197	■□■■■□■■■■■■■□■□■■■■■■■■■■■■□□□□□□□□■■■■■■■	557727777400771	Unknown	O09	1	Baghdad	M/39
IRQ02201115282	2011	5282	■■■■□■■■■■■■■□■■■■■■■□□□□□□□□□□□■■□□■■■■■	757767740003171	Unknown	O10	1	Baghdad	M/35
IRQ02201115448	2011	5448	■■■■□■■■■■■■■■■■■■■■■□□□□□■■■■■■□□■■■■■■■■	757777740763771	Manu2	O11	1	Diyala	M/63
IRQ02201013167	2010	3167	■■■□□□□■■■■□□■□■■■■□□□□□□□□□□□□□□□■■■■■■■■	703627400003771	CAS	O12	1	Baghdad	F/16
IRQ02201003924	2010	3924	■■■□□□□■■■■□□■■■■■■□□□□□□□□□□□□□□□■■■■■■■■	703637400003771	CAS	O13	0	Baghdad	M/48
IRQ02201005145	2010	5145	■□□■■■■■■■■■■□■■■■■□□□□■■■■■■■■□□□□■■■■■■■	477767607760771	LAM1	O14	0	Najaf	F/35
IRQ02201004981	2010	4981	■□■■■□■■■■■■■□□■■■■■□□□□■■■■■■■□□□□□■■□■□■	557747607740661	LAM9	O15	0	Dhi-qar	M/37
IRQ02201025861	2010	5861	■■□□■■■■■■■■□■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	637737777760771	T3	O16	2	Al-Qadisiya	M/29
IRQ02201026647	2010	6647	■■■■■■■■■■■■■■■□■■■■■□□■■■■□■■■■■□□□□■■■■■■■	777767636760771	T1	O17	2	Basra	F/29
IRQ02201015057	2010	5057	■■■■■■■■■■■■■■□■□□□□□■■■■■■■■■■■□□□□■■■■■■■	777764077760771	T1	O18	1	Baghdad	M/47
IRQ02201024897b	2010	4897b	■■□■■■■■■■■■■■■□■■■□■■■■■■■■□□□□■□■■■■■■■■	677767377413771	EAI5	O19	2	Basra	M/21
IRQ02201024258	2010	4258	■□■□□□□■■■■■■■□■■■■■■■□□□□□□□□□□□□■■■■■■■□■	503767740003761	CASI-Delhi	O20	2	Baghdad	M/54
IRQ02201014287	2010	4287	■■■■■■■■■■■■□■□■■■■■■■■■■■■■□□□■□□□□■■■■■■■	777727777420771	H3 (Ural-1)	O21	1	Dhi-qar	F/51
IRQ02201016432	2010	6432	■■■■■■■□■■■■■■□■■■■■□□□□□□□■■■■■□□□□■■■■■■■	775767600760771	T1	O22	1	Baghdad	M/60
IRQ02201014433	2010	4433	■□■□□□□■■■■□■■□■■■■■■□□□□□□□□□□□□□■■■■■■■■	503667700003771	CAS	O23	1	Babil	M/23
IRQ02201016038	2010	6038	■■■■□■■■■■■■□■■■■■■□□■■■■■■■■■□□■■■■■■■■■	757767637763771	Manu2	O24	1	Baghdad	M/28
IRQ02201013391	2010	3391	■■■■■■■■■■■■□■□■■■■■■■■■■■□□□□□□□□□■■■■■■■	777727777000771	Unknown	O25	1	Baghdad	M/31
IRQ02201113592	2011	3592	■□■□□□□■■■■■■■□■■■■□□□□□□□□□□□□□□□□■■■■■■■■	503767400001771	CAS	O26	1	Baghdad	M/21
IRQ02201113234	2011	3234	■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□■□□■□■■■■■■■	777767777722771	Manu2	O27	1	Dhi-qar	M/40
IRQ02201004179	2010	4179	■■□■■■■■■■■■■□■■■■■□□■■■■■■□■■■□□■■■■■■■■■	677767637563771	Manu2	O28	0	Baghdad	M/25
IRQ02201016082	2010	6082	■■■■■■■■■■■■■■□■■□■■□■□□□□■■■■■■■■□□□□■■■■■■■	777766607760771	LAM9	O29	1	Baghdad	F/24
IRQ02201103164	2011	3164	■■■■■■■■■■■■■■□■■□■■□□■■■■■■■■■■□□□□■■■■■■■	777766637760771	X1	O30	0	Baghdad	M/24
IRQ02201012729	2010	2729	■■■□□□□□■■■■■■□■■■■■■□□□□□□□□□□□□□□■■□□□■■■■	701767700003071	CAS	O31	1	Dhi-qar	M/65
IRQ02201015095	2010	5095	■□□□□□□■■■■■■■□■■■■□□□□□□□□□□□□□□□□■■■■■■■■	403767400001771	Unknown	O32	1	Najaf	M/70
IRQ02201014584	2010	4584	■□■■■■■■■■■■■□■■■■■■■■■■■□■□■□□□□■■■■■■■■	577767777520771	H3	O33	1	Baghdad	F/18
IRQ02201004326	2010	4326	■■■■■■■■■■■■■□■■■■■■■□□□□□□□□□□□□■■■■■■■■	777767770000771	Unknown	O34	0	Baghdad	M/25
IRQ02201113543	2011	3543	■□■□□□□■■■■■■□■■■■■■■□□□□□□□□□□□□■■□□□■■■■	501767740003071	CAS	O35	1	Baghdad	M/22
IRQ02201026360b	2010	6360b	■□■■■■■■■■■■■□■■■■■□□□□□□■■■■■■□□□■■■■■■■■	577767600760771	T1	O36	2	Basra	M/35
IRQ02201014885	2010	4885	■□■■■■■■■■■□■■■■■■■■■■■■□■□■□□□□■■■■■■■■	577737777520771	H3	O37	1	Baghdad	M/29
IRQ02201026915	2010	6915	■■■■□□□■■■■■■■□■■■■■■■□□□□□□□□□□□■■□□■■■■■	743767740003171	Unknown	O38	2	Salah	M/37
IRQ02201015996	2010	5996	■■■■□■■■■■■■■□■■■■□□□□□□□■■■■□□□□■■■■■■■■	757767400741771	Unknown	O39	1	Baghdad	M/20

Table 2: Continued.

Iso number	Year	Strain	Spoligotype description	Octal code	Lineage*	Label	Drug R**	Isolation city	Sex/age
IRQ02201014761	2010	4761		500000000000000	CAS	O40	1	Baghdad	F/56
IRQ02201113275	2011	3275		703767400003071	CAS	O41	1	Salah	M/30
IRQ02201004309	2010	4309		503767740003000	CAS1-Delhi	O42	0	Salah	M/37
IRQ02201016005	2010	6005		757767777720771	H3	O43	1	Baghdad	F/54
IRQ02201123429	2011	3429		501727740003071	CAS	O44	2	Baghdad	M/53
IRQ02201113672	2011	3672		703767740000171	CAS	O45	1	Dhi-qar	M/58
IRQ02201017297	2010	7297		703727740003111	CAS1-Delhi	O46	1	Kerbala	M/24
IRQ02201013852	2010	3852		703767740003661	CAS1-Delhi	O47	1	Baghdad	M/20
IRQ02201015137	2010	5137		575767737760671	T1	O48	1	Baghdad	M/25
IRQ02201012104	2010	2104		757777744743771	Unknown	O49	1	Dhi-qar	M/65
IRQ02201013811	2010	3811		757767777743771	Unknown	O50	1	Baghdad	F/35
IRQ02201113484	2011	3484		577727777420771	H4 (Ural-2)	O51	1	Baghdad	F/53
IRQ02201112326	2011	2326		557767777743771	Unknown	O52	1	Dhi-qar	M/56
IRQ02201016511	2010	6511		555727604643071	Unknown	O53	1	Wasit	M/30

*Lineage designations for orphan patterns were done manually as expert-based interpretations using revised SpolDB4 rules.

**Drug resistance (drug R) information is shown as 0, unknown; 1, non-MDR; 2, MDRTB, that is, combined resistance to INH − RIF (with or without resistance to other drugs); 3, any other resistance(s); 4, proven XDRTB, that is, resistance to INH + RIF + a fluoroquinolone + any 1 of 3 injectable 2nd-line drugs (capreomycin, kanamycin, amikacin).

Table 3: Description of 64 shared types (SITs; $n = 217$ isolates) and corresponding spoligotyping defined lineages/sublineages starting from a total of 270 *M. tuberculosis* strains isolated from Iraqi patients.

SIT*	Spoligotype description	Octal number	Nb in study	% in study	% in study versus database	Lineage**	Clustered versus unique patterns***
4		000000007760771	1	0.37	0.28	Unknown	Unique
20		677777607760771	1	0.37	0.12	LAM1	Unique
22		703777400001771	8	2.96	10.26	CAS1-Delhi	Clustered
25		703777740003171	10	3.7	1.77	CAS1-Delhi	Clustered
26		703777740003771	12	4.44	0.87	CAS1-Delhi	Clustered
34		776377777760771	1	0.37	0.12	S	Unique
41		777777404760771	1	0.37	0.21	LAM7-TUR	Unique
42		777777607760771	1	0.37	0.03	LAM9	Unique
47		777777774020771	1	0.37	0.07	H1	Unique
50		777777777720771	6	2.22	0.18	H3	Clustered
53		777777777760771	5	1.85	0.08	T1	Clustered
54		777777777763771	8	2.96	3.31	Manu2	Clustered
64		777777607560771	1	0.37	0.28	LAM6	Unique
75		777767777720771	6	2.22	8.7	H3	Clustered
117		777767777760731	2	0.74	8	T2	Clustered
118		777767777760771	6	2.22	3.9	T1	Clustered
124		777777777700771	1	0.37	1.54	Unknown	Unique
127		577777777420771	6	2.22	3.08	H4 (Ural-2)	Clustered
141		703767740003771	9	3.33	28.13	CAS1-Delhi	Clustered
186		777767404760771	1	0.37	8.33	LAM7-TUR	Unique
247		703777740003471	3	1.11	8.11	CAS1-Delhi	Clustered
284		037637777760771	9	3.33	5.11	T1	Clustered
309		703767740003171	17	6.3	80.95	CAS1-Delhi	Clustered
485		703777400003771	1	0.37	4	CAS1-Delhi	Unique
516		777777777360771	1	0.37	5.56	T1	Unique
568		777767477760771	1	0.37	20	T-H37Rv	Unique
610		777767774020731	1	0.37	14.29	H1	Unique
831		776367777760771	1	0.37	7.69	S	Unique
878		777776777760571	1	0.37	4	X1	Unique
954		703677740003771	1	0.37	6.67	CAS1-Delhi	Unique
1088		777767777763771	4	1.48	33.33	Manu2	Clustered
1144		777777600760771	17	6.3	41.46	T1	Clustered
1198		703737740003171	4	1.48	10.53	CAS1-Delhi	Clustered
1318		577767777760771	2	0.74	25	T1	Clustered
1470		777776607760771	2	0.74	10	LAM9	Clustered
1547		777727777760771	3	1.11	27.27	T3	Clustered
1630		777777601760771	1	0.37	16.67	T1	Unique
1638		577767777763771	1	0.37	33.33	Manu2	Unique
1655		777723777760771	1	0.37	2.7	T3	Unique

Table 3: Continued.

SIT*	Spoligotype description	Octal number	Nb in study	% in study	% in study versus database	Lineage**	Clustered versus unique patterns***
1679		777767737760771	2	0.74	20	T1	Clustered
1913		777767757760771	1	0.37	25	T5	Unique
2032		037627777760771	1	0.37	20	T1	Unique
2230		777700077760771	1	0.37	16.67	Unknown	Unique
2359		703677740003171	1	0.37	4.55	CAS1-Delhi	Unique
2686		703767740003471	4	1.48	66.67	CAS1-Delhi	Clustered
2691		037737777760771	1	0.37	20	T3	Unique
2707		701767740003671	1	0.37	33.33	CAS	Unique
2728		577767777420771	3	1.11	50	H4 (Ural-2)	Clustered
3275*		700037777760731	1	0.37	50	T2	Unique
3346*		777767600760771	11	4.07	100	T1	Clustered
3497*		703767400001771	8	2.96	100	CAS	Clustered
3708*		703767740003761	3	1.11	100	CAS1-Delhi	Clustered
3789*		777777740763771	2	0.74	66.67	Manu2	Clustered
3790*		777767600360771	4	1.48	100	T1	Clustered
3791*		703727740003171	4	1.48	100	CAS1-Delhi	Clustered
3792*		703767640003771	2	0.74	66.67	CAS1-Delhi	Clustered
3793*		557767777763771	1	0.37	50	Manu2	Unique
3794*		703767740000771	1	0.37	50	CAS	Unique
3795*		277767607760771	1	0.37	50	LAM1	Unique
3796*		703767740000371	1	0.37	50	CAS	Unique
3797*		777777774120771	1	0.37	50	H3	Unique
3798*		777717777760371	1	0.37	50	T1	Unique
3799*		777767614760771	1	0.37	50	T1	Unique
3800*		703767400003771	2	0.74	100	CAS	Clustered

A total of 48/64 SITs containing 173 isolates matched a preexisting shared type in the database, whereas 16/64 SITs ($n = 44$ isolates) were newly created. A total of 31/64 SITs containing 184 isolates were clustered within this study (2 to 17 isolates per cluster) while 33/64 SITs containing 33 strains were unique (for total unique strains, one should add to this number the 53 orphan strains, which brings the number of unclustered isolates in this study to 86/270 or 31.85% and clustered isolates to 184/270 or 68.15%). Note that SITs followed by an asterisk indicates "newly created" SITs due to 2 or more strains belonging to an identical new pattern within this study or after a match with an orphan in the database; SIT designations followed by number of strains: 3275 this study $n = 1$, FXX $n = 1$; 3346* this study $n = 11$; 3497* this study $n = 8$; 3708* this study $n = 3$; 3789* this study $n = 2$, IRQ $n = 1$; 3790* this study $n = 4$; 3791* this study $n = 4$; 3792* this study $n = 2$, IND $n = 1$; 3793* this study $n = 1$, IND $n = 1$; 3794* this study $n = 1$, IND $n = 1$; 3795* this study $n = 1$, BRA $n = 1$; 3796* this study $n = 1$, PAK $n = 1$; 3797* this study $n = 1$, MEX $n = 1$; 3798* this study $n = 1$, IRQ $n = 1$; 3799* this study $n = 1$, BRA $n = 1$; 3800* this study $n = 2$.

**Lineage designations according to SITVIT2 using revised SpolDB4 rules; "unknown" designates patterns with signatures that do not belong to any of the major lineages described in the database.

***Clustered strains correspond to a similar spoligotype pattern shared by 2 or more strains "within this study," as opposed to unique strains harboring a spoligotype pattern that does not match with another strain from this study. Unique strains matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as "orphan" (see Table 2).

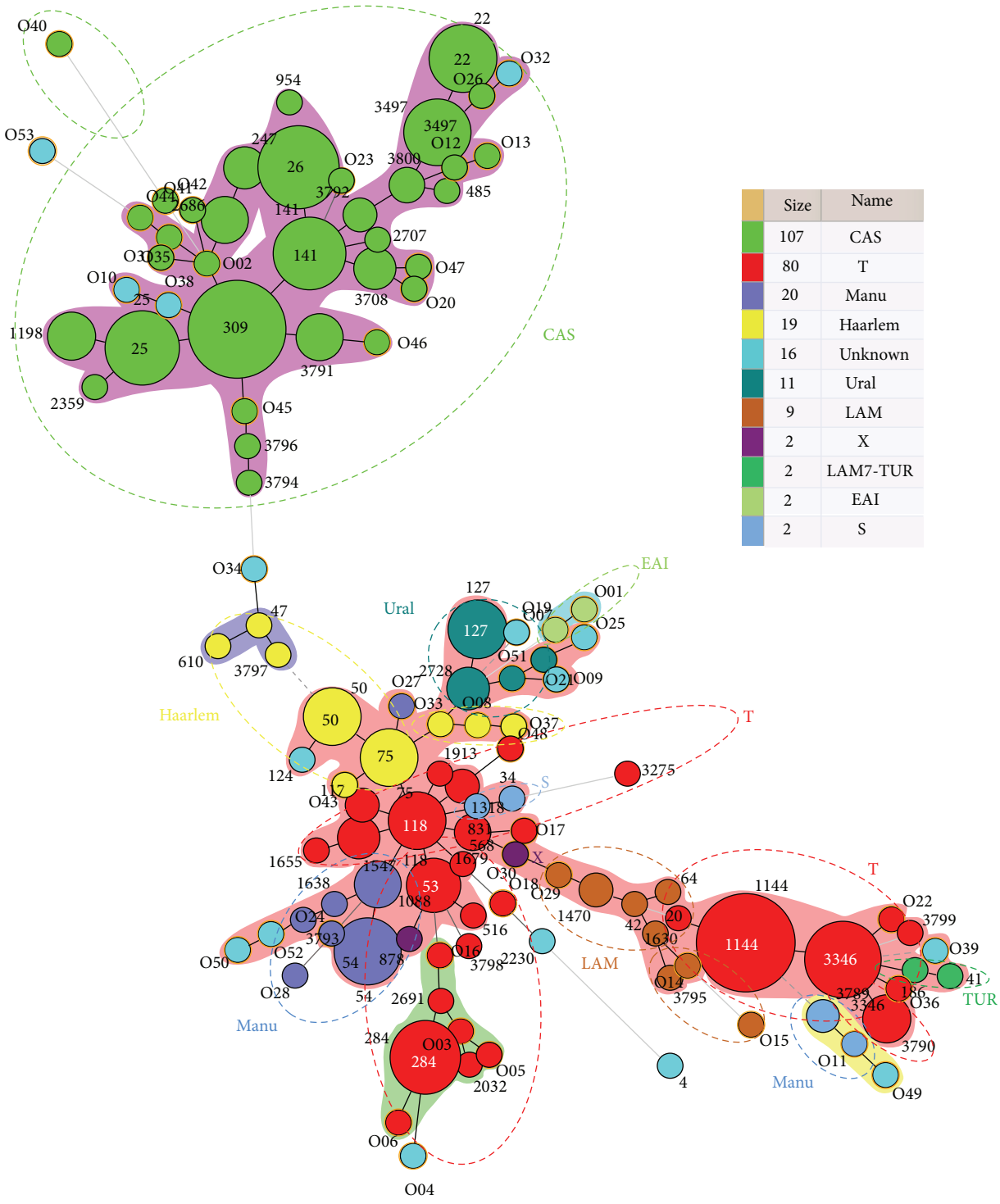


Figure 1: A minimum spanning tree (MST) illustrating evolutionary relationships between the *Mycobacterium tuberculosis* spoligotypes. This kind of tree connects each genotype based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous versus dashed and dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed, and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines represent 1, 2, or 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); gray dashed lines represent 4 changes; and gray dotted lines represent 5 or more changes. The size of the circle is proportional to the total number of isolates in our study, illustrating unique isolates (smaller nodes) versus clustered isolates (bigger nodes). The color of the circles indicates the phylogenetic lineage to which the specific pattern belongs. Note that orphan patterns are circled in orange. Patterns colored in cyan-blue indicate a strain with an unknown signature (unclassified).

and X. These strains accounted for 125/270 (46.3%) of the studied strains.

A total of 31/64 SITs containing 184 strains were clustered within this study (2 to 17 strains per cluster) while 33/64 SITs containing 33 strains were unique. For total unique strains, one should add to this number the 53 orphan strains, which brings the number of unclustered strains in this study to 86/270 or 31.85% and clustered strains to 184/270 or 68.15%. 23 clusters are shown in Table 3, along with 8 additional clusters (2 to 11 strains per cluster) represented by the newly created shared types.

The two largest clusters of 17 strains were composed of SIT309 (CAS1-Delhi lineage) and SIT1144 (T1 lineage).

Description of clusters containing 5 or more isolates in this study, and their worldwide distribution in the SITVIT2 database, is detailed in Table 4 and will be commented further under the discussion section.

The MIRU-VNTR analysis detected a total of 73 MIRU patterns from 110 strains using the full 24 MIRU-VNTR locus set, including 17 clusters and 56 unique (Figure 2). Allelic diversity for each locus was calculated in order to determine the discriminatory power of these loci in a combined group for the *Mycobacterium tuberculosis* population studied. Based on their discriminatory index (HGDI), 7 loci (MIRU02, MIRU04, MIRU20, Mtub29, ETR-B, MIRU24, and MIRU27) showed poor discriminatory power (HGDI < 0.3). Seven loci (MIRU 23, Qub11b, Mtub30, Mtub34, Mtub39, MIRU 39, and QUB 4156) discriminated the isolated moderately ($0.3 \leq$ HGDI ≤ 0.6). Lastly, 10 loci (Mtub 04, ETR-C, MIRU 40, MIRU 10, MIRU 16, Mtub21, ETR-A, MIRU 26, MIRU 31, and QUB-26) were highly discriminative (HGDI > 0.6). In this study the locus QUB-26 was found to be the most discriminatory allele in order to distinguish between strains (HGDI of 0.83). Conversely, locus MIRU-24 was found to be the least discriminatory with an HGDI of 0.03. The Recent Transmission Index (RTI_{n-1}) was found at 33.6% showing evidence of ongoing transmission.

3.3. Drug Resistance Patterns. Both the drug resistance patterns and the treatment status of the patients (new versus retreated cases) were studied in detail on all the 270 strains included in this study in function of their spoligotyping-based genotypic lineages, and the results were concomitantly exploited to draw a minimum spanning tree (MST) shown in Figure 3. Of the 270 strains studied, 171 (63.3%) were sensitive to all five of the first-line drugs tested, 64 (23.70%) were MDR, while the drug-susceptibility information was not available for 35 (12.96%) of strains. Regarding the treatment status of the patients, 187 (69.26%) were new cases while 83 (30.74%) were previously treated. It is noteworthy that all the 64 MDR cases were exclusively found among the retreated patients, bringing the proportion of MDR isolates in this group to 77.1% the difference of drug resistance between new versus retreated cases being highly significant (P value < 0.0001). As shown in Figure 3(a), a high frequency of MDR-TB was associated with SIT53/T1 which contained 60% of MDR strains. Interestingly, this same shared type was also exclusively associated with retreated cases (Figure 3(b)). Although the clustering rate between drug susceptible and drug resistant isolates did not vary significantly (P value > 0.4), differences were noted when comparing predominant SITs. Indeed, one may notice a relatively high frequency of MDR-TB among patterns related to SIT53/T1 (60% of MDR) in Figure 3(a), which was significantly higher than other SITs such as SIT26/CAS1-Delhi (16.67% of MDR), P value $= 0.0372$, and SIT141/CAS1-Delhi (11.11% of MDR), P value $= 0.0030$. Lastly, although the rate of MDR-TB was slightly higher among patients from cities other than Baghdad (Table 1), the difference was not statistically significant (P value $= 0.1224$).

4. Discussion

In this study we characterized, by spoligotyping, 270 *M. tuberculosis* isolates collected from patients diagnosed in Baghdad/Reference TB laboratory in a 13-month period. One hundred and ten samples were also characterized by 24-loci MIRU-VNTR.

All TB cases reported in this study were caused by *M. tuberculosis*. The molecular investigation of strain by spoligotyping did not show the specific indicator for other members of *M. tuberculosis* complex such as *M. bovis*. This situation has been described in other settings and countries such as by Godreuil et al. in west African countries [20], Nakajima et al. in Bangladesh [21], and Viegas in Mozambique [22]. Spoligotyping of the 270 *M. tuberculosis* strains revealed 31 clusters consisting of 184 strains, with clustering rate 57%, whereas 86 strains were unique.

One hundred and seven (39.6%) of the 270 studied strains were CAS belonging to different SITs. CAS has also been identified as predominant family in Saudi Arabia (22.5%) and also the predominance of Delhi genogroup in Iran [23, 24] was reported. In Pakistan this lineage represents 61% of the total [25]. And CAS has been also identified by recent study as a predominant strain in north India [26]. The data from Turkey suggested that there is no dominant *M. tuberculosis* clade such as what has been observed in Asia and former USSR republics [17]. CAS clade is followed, in decreasing order, by ill-defined T clade 29.6%, Manu 7.4%, Haarlem 7%, Ural 4.1%, LAM 3.3%, X 0.7%, LAM7-TUR 0.7%, EAI 0.7%, S 0.7%, and unknown 5.9% (Table 4). These observations emphasize the complex diversity of circulating *M. tuberculosis* strains in Iraq that could reflect the different transmission pathways occurring within the country. Besides, it has been suggested that particular lineages of *M. tuberculosis* might be adapted to specific human population [27, 28]. Indeed, many strains of this study belonged to Principal Genetic Group (PGG1), according to the *KatG-grA* polymorphism-based classification of Sreevatsan et al. [19], and these strains represent all of the known lineages associated with PGG1 with the exception of the Beijing clade. However, evolutionary modern PGG 2/3 strains were also found with different distribution.

In this study 110 strains were classified into 73 MIRU patterns of which 17 were clusters and 56 unique strains. The high level of Recent Transmission Index (RTI), at 33.6%,

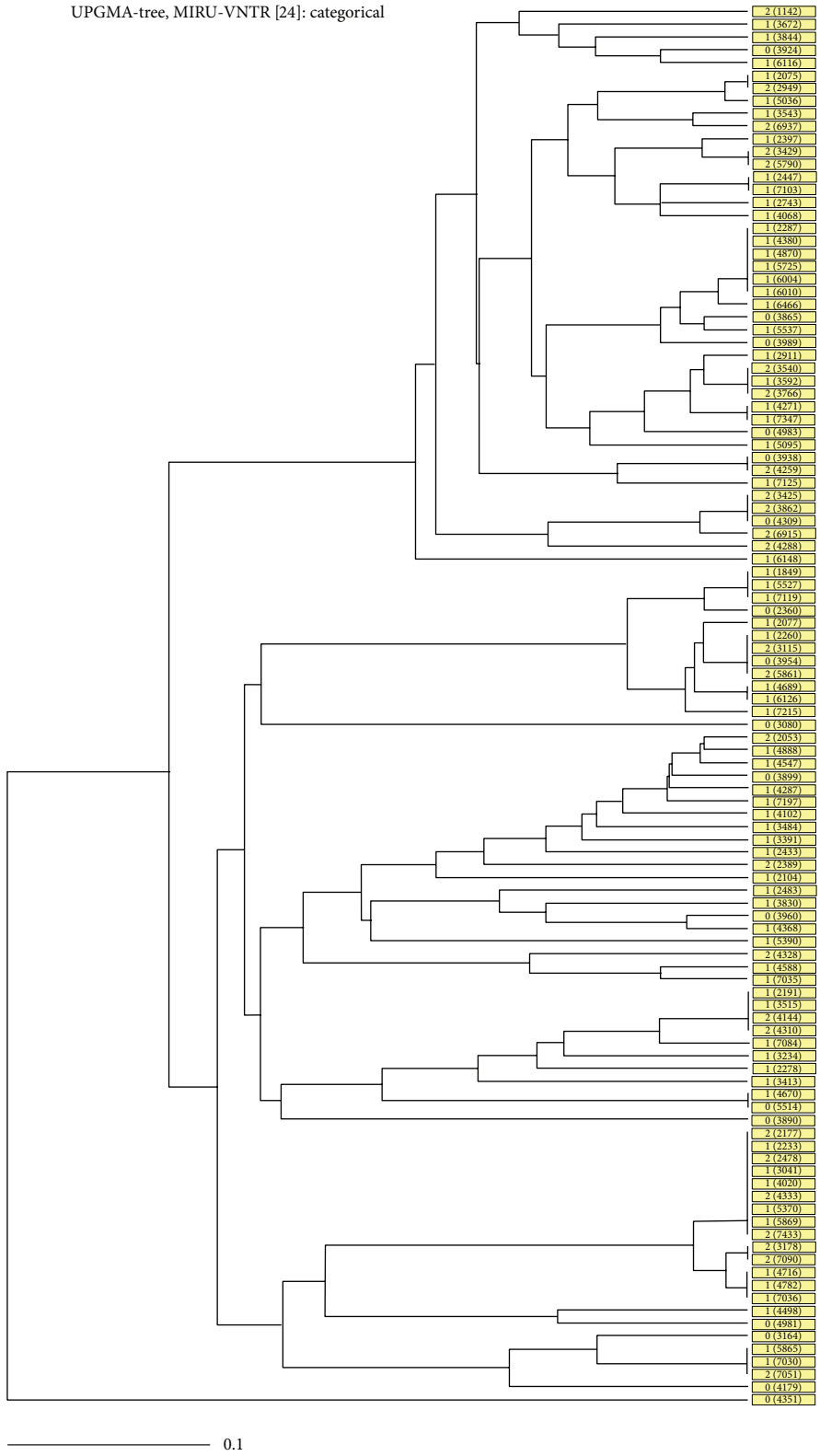


Figure 2: UPGMA tree based on MIRU 24 pattern of the subset of 110 samples. Drug resistance (Drug R) information is shown as 0, unknown; 1, non-MDR; 2, MDR-TB, that is, combined resistance to INH-RIF (with or without resistance to other drugs).

Table 4: Description of clusters containing 5 or more isolates in this study and their worldwide distribution in the SITVIT2 database.

SIT (lineage) octal number Spoligotype description	Number (%) in study	% in study versus database	Distribution in regions with ≥3% of a given SIT*	Distribution in countries with ≥3% of a given SIT**
22 (CAS1-Delhi) 703777400001771	8 (2.96)	10.26	ASIA-W 65.39, ASIA-S 12.82, AFRI-E 7.69, EURO-N 5.13, EURO-W 3.85	SAU 38.46, IRQ 25.64, IRN 7.69, TZA 3.85
25 (CAS1-Delhi) 703777740003171	10 (3.7)	1.77	AFRI-E 26.68, ASIA-W 22.26, AFRI-N 14.84, ASIA-S 13.6, AMER-N 9.19, EURO-N 4.95, EURO-W 4.77	ETH 25.44, SAU 16.61, SDN 12.19, USA 9.19, IND 8.66, IRQ 5.12, IRN 3.0
26 (CAS1-Delhi) 703777740003771	12 (4.44)	0.87	ASIA-S 54.08, AMER-N 15.81, ASIA-W 7.51, EURO-W 6.14, AFRI-E 5.13, EURO-N 4.19, EURO-S 3.61	IND 33.0, USA 15.81, PAK 10.4, SAU 5.63, BGD 5.42, IRN 4.48, ITA 3.47, NLD 3.11, ETH 3.11
50 (H3) 777777777720771	6 (2.22)	0.18	AMER-S 18.03, EURO-W 17.88, AMER-N 17.88, EURO-S 11.75, EURO-E 5.62, EURO-N 4.36, AFRI-N 4.33, AFRI-S 4.12, CARI 3.55, ASIA-W 3.19	USA 17.85, BRA 7.18, FXX 7.0, AUT 6.19, ITA 5.53, ESP 5.53, PER 4.81, ZAF 4.12, CZE 3.73
53 (T1) 777777777760771	5 (1.85)	0.08	EURO-W 16.17, AMER-N 13.93, AMER-S 12.43, EURO-S 9.72, ASIA-W 7.19, EURO-N 5.51, AFRI-S 5.13, AFRI-E 4.66, ASIA-E 4.38, AFRI-N 3.64, EURO-E 3.37, AMER-C 3.34	USA 13.64, FXX 8.14, ITA 5.51, BRA 5.3, ZAF 5.02, TUR 3.59, AUT 3.54, CHN 3.19
54 (Manu2) 777777777763771	8 (2.96)	3.31	ASIA-W 16.53, ASIA-S 15.29, ASIA-E 15.29, AFRI-N 14.88, AMER-N 8.26, AFRI-S 6.61, ASIA-N 4.13, AMER-S 3.72	CHN 15.29, IND 14.46, EGY 14.05, USA 8.26, SAU 8.26, ZAF 6.61, IRQ 5.79, RUS 4.13
75 (H3) 777767777720771	6 (2.22)	8.7	AMER-S 18.84, ASIA-W 15.94, EURO-W 11.59, ASIA-S 10.15, EURO-S 7.25, AMER-N 7.25, AFRI-M 7.25, AFRI-E 5.8, EURO-E 4.35, CARI 4.35, AFRI-W 4.35	IRN 10.15, FXX 10.15, BRA 10.15, IRQ 8.7, USA 7.25, CMR 7.25, TUR 4.35, MDG 4.35, GUF 4.35, ARG 4.35
118 (T1) 777767777760771	6 (2.22)	3.9	ASIA-W 18.83, EURO-W 17.53, AMER-S 13.64, EURO-S 9.74, ASIA-E 9.74, AMER-N 5.2, AFRI-E 5.2, EURO-E 4.55, AMER-C 4.55	FXX 15.58, TUR 12.34, ITA 6.49, VEN 5.84, CHN 5.84, USA 5.2, MOZ 4.55, JPN 3.9, IRQ 3.9, CZE 3.9, BRA 3.9, MEX 3.25, ARG 3.25
127 (Ural-2) 577777777420771	6 (2.22)	3.08	ASIA-S 54.36, EURO-W 16.41, ASIA-W 11.8, EURO-N 5.64, ASIA-E 5.13	IRN 49.23, NLD 7.69, SAU 6.15, IRQ 5.13, CHN 5.13, SWE 4.62, AUT 4.62, DEU 3.59
141 (CAS1-Delhi) 703767740003771	9 (3.33)	28.13	ASIA-S 37.5, ASIA-W 31.25, EURO-W 12.5, AMER-N 12.5, ASIA-SE 3.13, AFRI-E 3.13	IND 34.38, IRQ 28.13, USA 12.5, FXX 9.38, SAU 3.13, PAK 3.13, NLD 3.13, MOZ 3.13, MMR 3.13
284 (T1) 037637777760771	9 (3.33)	5.11	ASIA-W 42.61, EURO-E 22.73, EURO-W 15.91, EURO-N 7.39, EURO-S 3.41	BGR 22.73, TUR 19.89, SAU 13.07, IRQ 9.09, SWE 6.25, AUT 5.68, FXX 3.41
309 (CAS1-Delhi) 703767740003171	17 (6.3)	80.95	ASIA-W 85.71, AFRI-M 9.52, AFRI-E 4.76	IRQ 80.95, CAF 9.52, SAU 4.76, ETH 4.76
1144 (T1) 777777600760771	17 (6.3)	41.46	ASIA-W 90.24, AMER-N 4.88	IRQ 90.24, USA 4.88
3346 (T1) 777767600760771	11 (4.07)	100	ASIA-W 100.0	IRQ 100.0

Table 4: Continued.

SIT (lineage) octal number Spoligotype description	Number (%) in study	% in study versus database	Distribution in regions with ≥3% of a given SIT*	Distribution in countries with ≥3% of a given SIT**
3497 (CAS) 703767400001771 	8 (2.96)	100	ASIA-W 100.0	IRQ 100.0

*Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database. The definition of macrogeographical regions and subregions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania) are subdivided in E (eastern), M (middle), C (central), N (northern), S (southern), SE (south eastern), and W (western). Furthermore, CARIB (caribbean) belongs to Americas, while Oceania is subdivided in 4 subregions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new subregion by itself (northern Asia) instead of including it among the rest of eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in central, eastern, and south eastern Asia.

**The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ≥3% of a given SITs as compared to their total number in the SITVIT2 database.

FIGURE 3: A minimum spanning tree (MST) illustrating evolutionary relationships between the *M. tuberculosis* spoligotypes in our study in function to studied parameters. (a) Drug resistance and (b) treatment status, that is, new versus retreated cases.

indicates that the most cases are due to recent transmissions rather than reactivation of *M. tuberculosis* infections. In our study, ten loci (Mtub 04, ETR-C, MURU 40, MIRU 10, MIRU 16, Mtub21, ETR-A, MIRU 26, MIRU 31, and QUB-26) were highly discriminatory (HGDI > 0.6), seven loci (MIRU23, Qub11b, Mtub30, Mtub34, Mtub39, MIRU 39, and QUB 4156) moderately discriminate (0.3 ≤ HGDI ≤ 0.6), and seven loci (MIRU02, MIRU04, MIRU20, Mtub 29, ETR-B, MIRU 24, and MIRU27) showed poor discriminatory power (HGDI < 0.3).

MIRU-VNTR allelic results have been correlated with definition of ancestral and modern MTB lineages, with the presence of single allele in locus 24 being related to a modern strain type. We found that 98% of our strains contained only single repeat at locus 24, further confirming their modern lineage. This is comparable with previous reports for CAS strain from Bangladesh and Singapore [29, 30] and also with the studies from Pakistan and Bulgaria as supported by the finding that 62% of their CAS family strains contained only one allele at the locus 24 [31, 32]. Moreover, the relative discriminatory powers of particular VNTR loci vary depending on the strain in question [33–37].

This study found that in Iraqi population, the characteristic of MDR in *M. tuberculosis* is mostly acquired as a result of treatment failure, due to irregularity in taking of drug (anti-TB), neglect, and incorrect prescriptions. Although the extremely high level of MDR among previously treated patients might indicate transmission chains within the population, molecular epidemiology revealed that except for SIT53/T1 genotype, no significant differences were found and the MIRU analysis in the subset of strains did not show clusters of exclusively MDR strains. The fact that identical MIRU patterns were shared both by MDR and non-MDR strains, and that the isoniazid and rifampicin resistance patterns were independent of their genotypes, suggests that MDR strains most probably emerged due to the selective pressure because of problems in adherence to treatment (in addition to other environmental factors), (high population, poor housing, overcrowding, and malnutrition).

The data of this study provide important baseline information on the genetic diversity of *M. tuberculosis* in Iraq. Therefore, it could be used to monitor change in the transmission pattern of tuberculosis. Spoligotyping has been proved useful for categorizing strains into different families and can be used as an initial technique to be subsequently followed by MIRU-VNTR. This study showed that 24-loci MIRU-VNTR typing offers a higher discriminatory power. Iraq needs to conduct epidemiological survey by using conventional and genotyping methods in order to provide adequate data that can be used for the formulation of control strategies of tuberculosis transmission.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Nalin Rastogi and Daniela M. Cirillo contributed equally to this paper.

Acknowledgment

Ruqaya Mustafa Ali would like to thank Baghdad University, Iraq, for funding and support.

References

[1] WHO, "Global tuberculosis control," WHO Report, WHO, Geneva, Switzerland, 2012.

[2] Ministry of Health of Iraq, "New diagnosis of tuberculosis in Iraq," Tech. Rep., Ministry of Health of Iraq, Baghdad, Iraq, 2012.

[3] B. Mathema, N. E. Kurepina, P. J. Bifani, and B. N. Kreiswirth, "Molecular epidemiology of tuberculosis: current insights," *Clinical Microbiology Reviews*, vol. 19, no. 4, pp. 658–685, 2006.

[4] I. S. Kontsevaya, V. V. Nikolayevsky, and Y. M. Balabanova, "Molecular epidemiology of tuberculosis: objectives, methods, and prospects," *Molecular Genetics, Microbiology and Virology*, vol. 26, no. 1, pp. 1–9, 2011.

[5] P. Supply, S. Lesjean, E. Savine, K. Kremer, D. van Soolingen, and C. Locht, "Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units," *Journal of Clinical Microbiology*, vol. 39, no. 10, pp. 3563–3571, 2001.

[6] L. S. Cowan, L. Diem, T. Monson et al., "Evaluation of a two-step approach for large-scale, prospective genotyping of *Mycobacterium tuberculosis* isolates in the United States," *Journal of Clinical Microbiology*, vol. 43, no. 2, pp. 688–695, 2005.

[7] P. Supply, C. Allix, S. Lesjean et al., "Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*," *Journal of Clinical Microbiology*, vol. 44, no. 12, pp. 4498–4510, 2006.

[8] A. Thong-On, N. Smittipat, T. Juthayothin et al., "Variable-number tandem repeats typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS6110 in Thailand," *Tuberculosis*, vol. 90, no. 1, pp. 9–15, 2010.

[9] T. Wirth, F. Hildebrand, C. Allix-Béguec et al., "Origin, spread and demography of the *Mycobacterium tuberculosis* complex," *PLoS Pathogens*, vol. 4, no. 9, Article ID e1000160, 2008.

[10] M. Cardoso Oelemann, H. M. Gomes, E. Willery et al., "The forest behind the tree: phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country," *PloS ONE*, vol. 6, no. 3, Article ID e18256, 2011.

[11] H. David, V. Levy-Frébault, and M. F. Thorel, *Méthodes de laboratoire de mycobactériologie clinique*, Commission des Laboratoires de Référence et d'Expertise de l'Institut Pasteur, Paris, France, 1989.

[12] D. van Soolingen, P. W. M. Hermans, P. E. W. de Haas, D. R. Soll, and J. D. A. van Embden, "Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis," *Journal of Clinical Microbiology*, vol. 29, no. 11, pp. 2578–2586, 1991.

[13] J. Kamerbeek, L. Schouls, A. Kolk et al., "Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology," *Journal of Clinical Microbiology*, vol. 35, no. 4, pp. 907–914, 1997.

[14] K. Brudey, J. R. Driscoll, L. Rigouts et al., "*Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology," *BMC Microbiology*, vol. 6, article 23, 2006.

[15] C. Demay, B. Liens, T. Burguière et al., "SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 755–766, 2012.

[16] F. Koro Koro, Y. K. Simo, F. F. Piam et al., "Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping," *Journal of Clinical Microbiology*, vol. 51, no. 1, pp. 299–302, 2013.

[17] O. Kisa, G. Tarhan, S. Gunal et al., "Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey," *PLoS ONE*, vol. 7, no. 1, Article ID e30331, 2012.

[18] I. Mokrousov, "The quiet and controversial: ural family of *Mycobacterium tuberculosis*," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 619–629, 2012.

[19] S. Sreevatsan, X. Pan, K. E. Stockbauer et al., "Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9869–9874, 1997.

[20] S. Godreuil, G. Torrea, D. Terru et al., "First molecular epidemiology study of *Mycobacterium tuberculosis* in Burkina Faso," *Journal of Clinical Microbiology*, vol. 45, no. 3, pp. 921–927, 2007.

[21] C. Nakajima, Z. Rahim, Y. Fukushima et al., "Identification of *Mycobacterium tuberculosis* clinical isolates in Bangladesh by a species distinguishable multiplex PCR," *BMC Infectious Diseases*, vol. 10, article 118, 2010.

[22] S. O. Viegas, A. Machado, R. Groenheit et al., "Molecular diversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mozambique," *BMC Microbiology*, vol. 10, article 195, 2010.

[23] S. A. Al-Hajoj, "Tuberculosis in Saudi Arabia: can we change the way we deal with the disease?" *Journal of Infection and Public Health*, vol. 3, no. 1, pp. 17–24, 2010.

[24] M. Jafarian, M. Aghali-Merza, P. Farnia, M. Ahmadi, M. R. Masjedi, and A. A. Velayati, "Synchronous comparison of *Mycobacterium tuberculosis* epidemiology strains by MIRU-VNTR and MIRU-VNTR and spoligotyping technique," *Avicenna Journal of Medical Biotechnology*, vol. 2, no. 3, pp. 145–152, 2010.

[25] M. Tanveer, Z. Hasan, A. R. Siddiqui et al., "Genotyping and drug resistance patterns of *M. tuberculosis* strains in Pakistan," *BMC Infectious Diseases*, vol. 8, article 171, pp. 2–10, 2008.

[26] M. Varma-Basil, S. Kumar, J. Arora et al., "Comparison of spoligotyping, mycobacterial interspersed repetitive units typing and IS6110-RFLP in a study of genotypic diversity of *Mycobacterium tuberculosis* in Delhi, North India," *Memórias do Instituto Oswaldo Cruz Rio de Janeiro*, vol. 106, no. 5, pp. 524–535, 2011.

[27] Z. H. Helal, M. S. E.-D. Ashour, S. A. Eissa et al., "Unexpectedly high proportion of ancestral manu genotype *Mycobacterium*

tuberculosis strains cultured from tuberculosis patients in Egypt," *Journal of Clinical Microbiology*, vol. 47, no. 9, pp. 2794–2801, 2009.

[28] S. Gagneux, K. DeRiemer, T. Van et al., "Variable host-pathogen compatibility in *Mycobacterium tuberculosis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 8, pp. 2869–2873, 2006.

[29] S. Banu, S. V. Gordon, S. Palmer et al., "Genotypic analysis of *Mycobacterium tuberculosis* in bangladesh and prevalence of the Beijing strain," *Journal of Clinical Microbiology*, vol. 42, no. 2, pp. 674–682, 2004.

[30] Y.-J. Sun, R. Bellamy, A. S. G. Lee et al., "Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to examine genetic diversity of *Mycobacterium tuberculosis* in Singapore," *Journal of Clinical Microbiology*, vol. 42, no. 5, pp. 1986–1993, 2004.

[31] A. Ali, Z. Hasan, M. Tanveer et al., "Characterization of *Mycobacterium tuberculosis* central Asian Strain 1 using mycobacterial interspersed repetitive unit genotyping," *BMC Microbiology*, vol. 7, article 76, 2007.

[32] V. Valcheva, I. Mokrousov, O. Narvskaya, N. Rastogi, and N. Markova, "Utility of new 24-locus variable-number tandem-repeat typing for discriminating *Mycobacterium tuberculosis* clinical isolates collected in Bulgaria," *Journal of Clinical Microbiology*, vol. 46, no. 9, pp. 3005–3011, 2008.

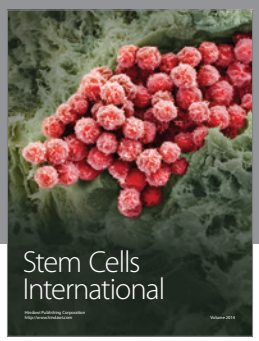
[33] Y. Murase, S. Mitarai, I. Sugawara, S. Kato, and S. Maeda, "Promising loci of variable numbers of tandem repeats for typing Beijing family *Mycobacterium tuberculosis*," *Journal of Medical Microbiology*, vol. 57, no. 7, pp. 873–880, 2008.

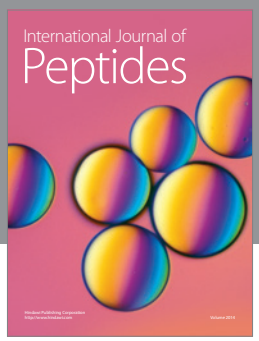
[34] I. Comas, S. Homolka, S. Niemann, and S. Gagneux, "Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies," *PLoS ONE*, vol. 4, no. 11, Article ID e7815, 2009.

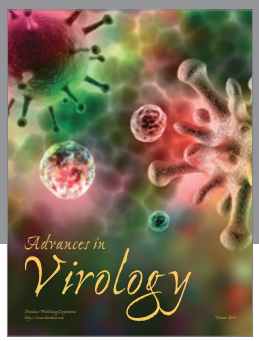
[35] S. Ferdinand, G. Valétudie, C. Sola, and N. Rastogi, "Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families," *Research in Microbiology*, vol. 155, no. 8, pp. 647–654, 2004.

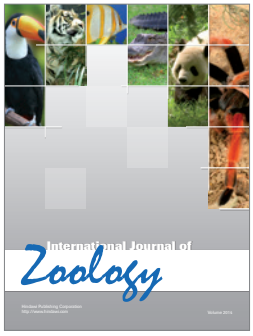
[36] M. Aminian, A. Shabbeer, and K. P. Bennett, "A conformal Bayesian network for classification of *Mycobacterium tuberculosis* complex lineages," *BMC Bioinformatics*, vol. 11, supplement 3, article S4, 2010.

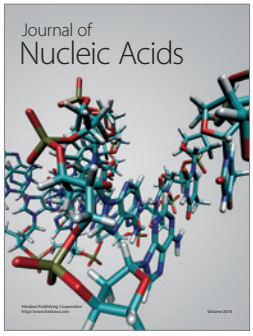
[37] N. H. Mendes, F. A. F. Melo, A. C. B. Santos et al., "Characterization of the genetic diversity of *Mycobacterium tuberculosis* in São Paulo city, Brazil," *BMC Research Notes*, vol. 4, article 269, 2011.

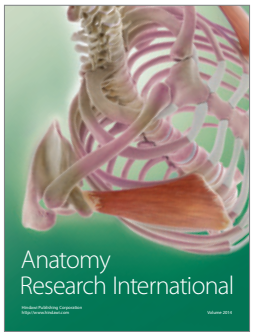


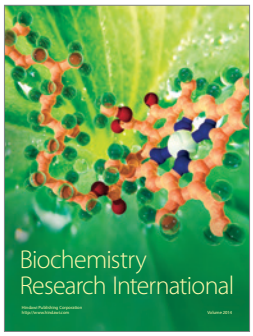


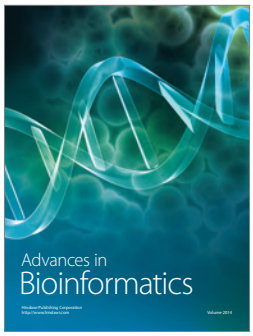


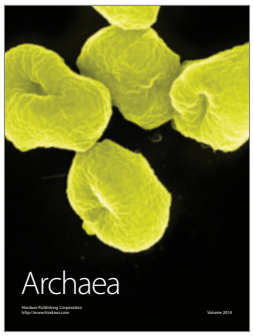


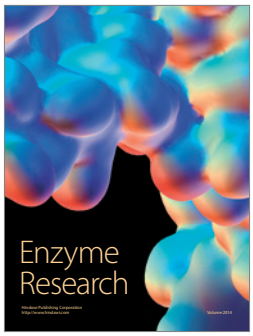


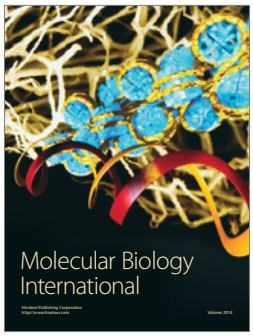












Hindawi

Submit your manuscripts at
http://www.hindawi.com

Article 13

The International Journal of Tuberculosis and Lung Disease

Phylogenetic associations with drug resistance of Mycobacterium tuberculosis isolates in a pediatric population.

Journal:	*The International Journal of Tuberculosis and Lung Disease*
Manuscript ID:	IJTLD-02-14-0173.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Arora, Jyoti; National Institute of Tuberculosis and Respiratory Diseases, Microbiology Sidiq, Zeeshan; National Institute of Tuberculosis and Respiratory Diseases, Microbiology SHARMA, SANGEETA; National Institute of Tuberculosis and Respiratory Diseases, Pediatric Pulmonology Singhal, Ritu; National Institute of Tuberculosis and Respiratory Diseases, Microbiology Bhalla, Manpreet; National Institute of Tuberculosis and Respiratory Diseases, Microbiology Couvin, David; Institut Pasteur, Supra National Laboratory Sarin, Rohit; National Institute of Tuberculosis and Respiratory Diseases, TB Chest Rastogi, Nalin; Institut Pasteur, Supra National Laboratory Myneedu, Vithal; National Institute of Tuberculosis and Respiratory Diseases, Microbiology
Key Words:	Pediatric, Spoligotyping, Phylogenetic

SCHOLARONE™
Manuscripts

Phylogenetic associations with drug resistance of *Mycobacterium tuberculosis* isolates in a pediatric population.

Referral
Jyoti Arora [a], arojyoti@gmail.com
Zeeshan Sidiq [a], sidiqueezeeshan@yahoo.co.in
Sangeeta Sharma[b], sangeetasharma2000@gmail.com
Ritu Singhal [a], ritugo@hotmail.com
Manpreet Bhalla [a], mnprtbhalla@yahoo.com
David Couvin[b], dcouvin@pasteur-guadeloupe.fr
Rohit Sarin [a], drsarin@yahoo.com
Nalin Rastogi [c*], nrastogi@pasteur-guadeloupe.fr
Vithal Prasad Myneeedu [a*], tbmicro@gmail.com

[a] Department of Microbiology, National Institute of Tuberculosis and Respiratory Diseases, Sri Aurobindo Marg, New Delhi-110030, India
[b] Department of Paediatrics, National Institute of Tuberculosis and Respiratory Diseases, Sri Aurobindo Marg, New Delhi-110030, India
[c] WHO Supranational TB Reference Laboratory, TB & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Abymes, France

Running title: TB genotypes in pediatric patients

*** Corresponding Authors: V.P. Myneedu / N. Rastogi**
– V.P. Myneedu, Department of Microbiology, Lala Ram Sarup Institute of Tuberculosis and Respiratory Diseases, Aurobindo Marg, New Delhi-110030, India. tbmicro@gmail.com
– N. Rastogi, WHO Supranational TB Reference Laboratory, Institut Pasteur de la Guadeloupe, Morne Jolivière, BP 484, 97183-Abymes, France. nrastogi@pasteur-guadeloupe.fr

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

For Review Only

Introduction

Despite being curable, Tuberculosis (TB) remains a global public health concern in developing countries. Over the years the prevention and management of TB in adults has gained importance but, little attention has been given to children with multiple drug resistant (MDR) TB or children in contact with it. Furthermore, reliable epidemiological data on the prevalence of pediatric TB from endemic areas is scarce due to diagnostic difficulties, limited resources, poor sensitivity of currently available diagnostic modalities, and lack of data from laboratories. According to the World Health Organization (WHO) report at least half a million children become ill with TB each year and nearly 74 000 HIV-uninfected children die from it every year.[1] The Revised National Tuberculosis Control Program (RNTCP) status report states that nearly 40% of the Indian population is infected with the TB bacillus and 10% of new cases occur in pediatric population.[2] Considering that pediatric TB might serve as a surrogate marker to measure recent transmission and as an indicator of the effectiveness of TB control programs; we thought it desirable to use DNA fingerprinting as a tool to understanding of ongoing TB transmission and evaluate the phylogeographical specificities of the prevalent *Mycobacterium tuberculosis (M. tuberculosis)* lineages as compared to neighboring countries.

Material and methods

Study population and bacteriological investigations: The study was conducted in the Department of Microbiology, National Institute of Tuberculosis and Respiratory Diseases, New Delhi, India. The institute is a tertiary care hospital as well as the National Reference Laboratory for the RNTCP, duly accredited through regular rounds of proficiency testing for first and second line drug susceptibility by the WHO Supranational TB Reference Laboratory, Institute of Tropical Medicine, Antwerp; Belgium. Majority of the pulmonary patients for whom drug susceptibility testing (DST) is performed are retreatment cases (either relapse or failure cases). The MDR TB rate among culture positive patients is high and falls between 55% – 60%.

All pediatric patients (n=111) who attended the OPD during the study period of 2009-2011 and whose sputum samples were sent to Microbiology department for liquid culture/DST were included in this study. Samples were processed by NaLC-NaOH method and cultured using MGIT960 liquid culture system (Becton Dickinson, Sparks, MD) in a BSL-3 laboratory, followed by routine mycobacterial identification. DST of positive cultures containing strains identified as MTBC was performed as per manufacturer's instructions at the following critical concentrations: rifampicin (RIF) 1.0 µg/ml, isoniazid (INH) 0.1 µg/ml, streptomycin (SM) 1.0 µg/ml, ethambutol (EMB) 5.0 µg/ml.[3] Since the study was performed retrospectively on de-identified reference laboratory samples ethics approval was not required.

Spoligotyping and database comparison: Spoligotyping was performed with a commercially available kit as described earlier.[4] Spoligotypes in octal format were entered in the SITVIT2 proprietary database of the Institut Pasteur de la Guadeloupe, which is an updated version of

For Review Only

the SpolDB4 and SITVITWEB databases.[5, 6] The patterns from this study were compared to those within the SITVIT2 database (n= 87,000 MTBC clinical isolates from 160 countries of patient origin). In this database, SIT (Spoligotype International Type) designates spoligotypes shared by two or more patient isolates, as opposed to "orphan" which designates patterns reported for a single isolate.

Phylogenetical analysis: The MTBC lineages observed were discriminated as ''ancient'' vs. ''modern'' lineages as defined by principal genetic groups (PGG) based on *KatG463-gyrA95* polymorphism,[7] and the reported linking of specific spoligotype patterns to PGG1, PGG2/3 (the latter being easily characterized by the lack of spacers 33–36).[8,9]

A minimum spanning tree (MST) was constructed using the BioNumerics software version 3.5 based on spoligotyping data. It is a unidirectional network in which all the isolates are linked together with the fewest possible linkages between nearest neighbors. A Spoligoforest tree was also drawn as a hierarchical layout using the SpolTools software to visualize parent to descendant spoligotypes (available through http://www.emi.unsw.edu.au/spolTools; [10, 11]. Note that as opposed to MST, a Spoligoforest tree is a directed graph in which all patterns are not necessarily connected

Results

Demographic characteristics and drug-resistance:

In the study population the distribution patterns regarding age groups in years (male to female sex-ratios) were: 0-5 n=1 (0/1); 6-10 n=13 (7/6 or 1.17); 11-15 n=97 (22/75 or 0.29). Majority of children, 97/111 (87.38%) were in the age group of 11-15 years with high prevalence of disease in females (73.87%; p-value ≤0.0001).The mean age was 13.4 years, with male to female sex-ratio 29/82 or 0.35. Ziehl Neelsen smear microscopy was positive for acid-fast bacilli (AFB) in 78/111 or 70.27 % samples with following culture results: smear +ve/culture +ve, 77/111(69.36%); smear –ve/culture +ve, 21/111 (18.9%); smear +ve/ culture –ve, 1/111 (0.9%); smear –ve/culture –ve, 9/111 (8.10%); and smear –ve/ culture contaminated 3/111 (2.71%). Thus cultures were positive for MTBC in 97/111 (87.4%) of patients.

DST results showed that 31/97 (32%) strains were pansusceptible while 66/97 (68%) were resistant to at least one drug; out of which 55/97 (56.7%) were resistant to at least INH and rifampicin RIF (MDR). The individual drug resistance patterns observed were: SM only 2, INH only 3, RIF only 0, EMB only 0, INH+RIF 6, INH+SM 4, INH+EMB 0, SM+INH+RIF 18, SM+INH+RIF+EMB 25, INH+RIF+EMB 6.

Spoligotyping: Interpretable results were obtained on a total of 90/97 MTBC isolates leading to a total of 29 different patterns; 22 patterns correspond to shared-types (SITs) in the database, and 7 patterns belong to orphan strains. **Table 1** summarizes the spoligotypes obtained and exploration of phylogenetic associations as genotypic lineages vs. drug resistance. Regarding clustering, a total of 66/90 (73.3%) isolates were clustered (among 5 major SITs containing 2–28 strains per SIT, **Table 2**) whereas 24/90 (26.6%) strains were unique. The largest cluster was SIT1/Beijing (28/90 or 31.1%), followed by SIT26/CAS1-Delhi (27/90 or 30%), SIT53/T1 (6/90 or 6.7%), SIT25/CAS1-Delhi (3/90 or 3.3%), and SIT250/Beijing (2/90 or 2.2%). Thus taken as a lineage irrespective of shared-type designations, CAS was the most predominant clade (n=40), followed by Beijing (n=30), EAI (n=5), ill-defined T (n=7), and Haarlem (n=2); whereas 5 isolates showed unknown lineage signatures.

For Review Only

Worldwide distribution of spoligotypes: **Table 2**, shows that with 64.2% of SIT26/CAS1-Delhi lineage strains being reported from South and West Asia (ASIA-S 50.1%, ASIA-W 14.1%), this lineage is characterized by a high phylogeographic specificity for the Indian sub-continent (ISC) since 45.3% of such strains are isolated from three ISC countries, namely India (30.3%), Pakistan (9.5%) and Bangladesh (5.5%).

The distribution of SIT1/Beijing among macro-geographical regions and sub-regions in the SITVIT2 database is predominantly linked to East, South-East, and North Asia with 48.6% of strains (ASIA-E 32%, ASIA-SE 9.8%, ASIA-N 6.8%) vs. only 5% in South Asia. Furthermore, none of the ISC countries including India features in the list of countries with ≥ 3% of SIT1/Beijing genotype (**Table 2**); nonetheless with 28/90 or 31.1% of all strains among pediatric patients, this MTBC pattern has apparently reached India, and is at present being transmitted actively.

Another interesting case concerns the remarkable presence of rarely seen spoligotype in India previously – the SIT53/T1 pattern that was found in 6/90 or 6.7% of pediatric TB cases. Although considered a ubiquitous pattern, this spoligotype was rarely reported from ISC countries including India. Hence, this MTBC pattern has apparently reached India recently via travelers from one of these countries or through Indian migrants returning home from host countries.

The last case worth-mentioning is SIT250/Beijing pattern with only 2/90 or 2.2% of pediatric patients, which is a rarely observed Beijing strain since it represents only 36/10850 or 0.33% isolates in the SITVIT2 database as compared to the prototype SIT1/Beijing pattern (n=10130/10850 or 93.4% isolates).

Distribution of spoligotyping-defined lineages and evolutionary relationships: Majority of isolates, 81(90%) belonged to PGG1. The exceptions included modern PGG2/3 strains represented by 9 (10%) strains which belonged to Haarlem (n=2) and ill defined T clade (n=7). The major genogroups in the MST and Spoligoforests corresponded to CAS, Beijing, EAI and T. The major clustered patterns (SIT1/Beijing, SIT26/CAS1-Delhi, and SIT53/T1) constituting the central nodes of their respective lineage groups were all represented by their prototypes in the international database, suggesting that the bulk of ongoing pandemic was caused by most predominant and well-characterized representatives of the respective MTBC lineages globally (Figure 1). Nonetheless, diversity observed for CAS lineages strains (thirteen patterns for CAS, **Table 1**), corroborates the historical presence of the CAS genotype in Delhi and its ongoing evolution.[12, 13] On the other hand, the diversity of the few EAI strains (n=5, Table 1) and the fact that all these strains were unlinked, may suggest occasional contacts with South Indian travelers and/or migrants since the EAI lineage strains are more frequent in the Southern India.[13, 14]

Distribution of clustered spoligotype patterns in neighboring countries: As summarized in **Table 3**, we also studied the distribution of clustered *M. tuberculosis* spoligotype patterns (2 to 28 strains/cluster) with those recorded in the updated SITVIT2 database for neighboring countries. Saudi Arabia although not a neighbor, was selected due to frequent travel and work related links with Indian migrant workers.

Discussion

TB has been found to be the most common infectious disease amongst children admitted to hospitals in India.[15] The high prevalence and incidence of TB in children in developing countries is

For Review Only

probably due to poor socio-economic conditions, malnutrition, overcrowding, younger population, human immunodeficiency virus (HIV) co-infection and a high prevalence of TB in adults acting as contacts.[16,17,18] Since children are the reservoir of infection, epidemiological studies may help to understand recent transmission and evaluate the effectiveness of control programs. Majority of participants, (87.38%) were in the age group of 11-15 years with high prevalence of disease in females. The reason for this could be less attentive care of girls in India who are often malnourished and ignored as compared to boys who are benefited by a better social status and care in society. We speculate that the girls were brought to the hospital much later than boys when the disease became complicated, a finding also reflected in previous studies. [19, 20]

The predominant MTBC lineages in our study belonged to the PGG1 phylogenetical groups: CAS (n=41 or 45.56%), Beijing (n=30, 33.33%), and EAI (n=5 or 5.56%). Among Beijing family, its prototype SIT1 was represented by 28 strains while among the CAS family, its CAS1-Delhi prototype SIT26 represented 27 strains. On the other hand, the "modern" PGG2/3 strains were represented by 9 (10%) strains: Haarlem (n=2), and ill-defined T (n=7). High proportion of SIT1/Beijing strains in our study (**Table 3**) which at n=28/90 or 31.11% is significantly higher than the rest of India (below 10%). It is also higher than the figures reported from Bangladesh, Pakistan, Nepal, Sri Lanka, and Saudi Arabia (comprised between 3 – 16% of all strains); it is similar to the proportion observed in Myanmar (31%), but remains below China (61%) where the Beijing genotype has been present historically.[21, 22] Furthermore, a rare and newly emerging clone of Beijing genotype SIT250/Beijing was also found in this study (present at 2.22%). It was inexistent among all the neighboring countries, and has essentially been found so far in China, Taiwan, Vietnam, Indonesia, USA and India. Only 3 strains have yet been reported from India out of 2647 strains reported to the SITVIT2 database; it is noteworthy that 2/3 such strains have been found in our pediatric population – highlighting the recent introduction of this Beijing strain in India. It is noteworthy that the Beijing genotypes were not so commonly isolated in other studies from Delhi or Mumbai.[12, 13, 23] Considering the fact that this genotype is the major cause of outbreaks involving drug resistant variants worldwide,[24] and is highly associated with drug resistance in our study (25/30 or 83.3% of the spoligotyped Beijing strains were drug resistant in our study, and 22/30 or 73.3% showed MDR-TB with or without association to other drugs), our findings a major cause of concern for health authorities.

Regarding the other predominant lineage observed (**Table 3**), the proportion of SIT26/CAS1-Delhi in this study at n= 27/90 or 30% was higher than for rest of Delhi and India (below 20%), almost triple of Bangladesh (7.35%), and equaled the proportion found in Nepal (22.45%) and Pakistan (37.4%); and the fact that within India, SIT26/CAS1-Delhi is the most common spoligotype in North region in general (Singh et al., 2007).[12,13] Another potentially emerging CAS1-Delhi clone in our study might be SIT25/CAS1-Delhi (present at 3.33%), which is almost inexistent among neighboring countries but present at around 5% in Saudi Arabia (**Table 3**), with a high phylogeographical specificity for Middle-East, namely Ethiopia, Saudi Arabia, Sudan, India, Oman and Iraq (see distribution pattern in **Table 2**). Lastly, the proportion of SIT53/T1 strains at 6/90 or 6.67% was more than the rest of Delhi, India, Bangladesh, Pakistan, and Myanmar (1 to 3%). It equals their proportion in Nepal, China and Saudi Arabia, and might be an emerging MTBC clone which is of concern seeing its association with MDR-TB in certain areas.[25, 26] However, the proportion of EAI lineage strains was

For Review Only

very low in our study (5.56%) vs. for the rest of New Delhi (25.27%) as seen through the SITVIT2 database. Future studies will be needed to verify if the ancestral EAI lineage strains of south-Indian origin are not currently being replaced by the arrival of more virulent MTBC Beijing strains in Northern India.[5, 6, 14]

In our previous study on Extensively Drug Resistant (XDR) isolate, [27] SIT1 was the largest cluster with 21 isolates (42%), followed by SIT26 (10 isolates; 20%) and SIT53 (8 isolates; 16%). The same SITs are predominant though with different percentages in pediatric TB patients, which reflects the ongoing transmission of the drug resistant strains in the community from adults to children – a finding similar to a recent report from Ethiopia.[28] Unfortunately, the main limitation of our study remains the fact that purely epidemiological conclusions could not be drawn due to lack of details on contact history of the patients. Furthermore, since the hospital is a referral centre sampling bias cannot be ruled out.

Conclusion: Our preliminary findings on the phylogenetic associations with drug resistance of pulmonary MTBC isolates in a pediatric population show that both CAS and Beijing genotype are locally prevalent and actively transmitted. These considerations justify a focus on the proper management of childhood TB in Delhi. Screening for children who are in close contact with index cases/drug resistant cases has a great potential to reduce the burden of TB disease by helping to identify cases early, and thereby preventing further spread of the disease.

Acknowledgments

Thanks are due to Banwari Lal and Rachna for providing technical assistance. Zeeshan Sidiq gratefully acknowledges the fellowship received from Indian Council of Medical Research (ICMR). David Couvin was awarded a Ph.D. fellowship by the European Social Funds through the Regional Council of Guadeloupe.

References

1. World Health Organization (2013) WHO Global TB Report 2013.

2. Central TB Division (2010) RNTCP status report. Nirman Bhavan, New Delhi, India Directorate of General Health Services, Ministry of Health and Family Welfare, 2006. Revised National TB Programme, TB India. http://www.tbcindia.org

3. Kent PT, Kubica GP. Public health mycobacteriology, A guide for the level III laboratory, Centers for Disease Control, Atlanta, Ga, 1985.

4. Kamerbeek J, Schouls L, Kolk A,et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 1997; 35: 907–914.

5. Brudey K, Driscoll JR, Rigouts L, et al. *Mycobacterium tuberculosis complex* genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 2006; 6:23.

6. Demay C, Liens B, Burguière T, et al. SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol 2012; 12(4): 755-766.

For Review Only

7. Sreevatsan S, Pan X, Stockbauer KE, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 1997; 94(18): 9869-9874.

8. Brosch R, Gordon SV, Marmiesse M, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A 2002; 99(6): 3684-3689.

9. Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, Musser JM. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J Clin Microbiol 2000; 38(2): 669-676.

10. Reyes JF, Francis AR, Tanaka MM. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 2008; 9: 496.

11. Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM. spolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. Bioinformatics 2008; 24(20): 2414–2415.

12. Singh UB, Suresh N, Bhanu NV, et al. Predominant tuberculosis spoligotypes, Delhi, India. Emerg Infect Dis 2004; 10: 1138–1142.

13. Singh UB, Arora J, Suresh N, et al. Genetic biodiversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in India. Infect Genet Evol 2007; 7(4): 441-448.

14. Joseph BV, Soman S, Radhakrishnan I, et al. Molecular epidemiology of *Mycobacterium tuberculosis* isolates from Kerala, India using IS*6110*-RFLP, spoligotyping and MIRU-VNTRs. Infect Genet Evol 2013; 16: 157-164.

15. Kabra S K, Lodha R, Seth V. Category based treatment of tuberculosis in children. Indian Pediatr 2004; 41: 927–937.

16. Marais BJ, Hesseling AC, Gie RP, Schaaf HS, Beyers N. The burden of childhood tuberculosis and the accuracy of community-based surveillance data in an endemic area. Int J Tuberc Lung Dis 2006; 10: 259–263.

17. Steinbrook R. HIV in India – a complex epidemic. N Engl J Med 2007; 356: 1089–1093.

18. Salazar G E, Schmitz T L, Cama R, et al. Pulmonary tuberculosis in children in a developing country. Pediatrics 2001; 108(2): 448-453.

19. Sharma S, Sarin R, Khalid UK, Singla N, Sharma PP, Behera D. The DOTS strategy for treatment of paediatric pulmonary tuberculosis in South Delhi, India. Int J Tuberc Lung Dis. 2008; 12(1): 74-80.

20. Arora VK, Gupta R. Directly observed treatment for tuberculosis. Indian J Pediatr 2003; 70(11): 885-889.

21. Mokrousov I, Ly HM, Otten T, et al. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res 2005; 15(10): 1357-1364.

22. Wan K, Liu J, Hauck Y, et al. Investigation on *Mycobacterium tuberculosis* diversity in China and the origin of the Beijing clade. PLoS One 2011; 6(12): e29190.

23. Chatterjee A, D'Souza D, Vira T, et al. Strains of *Mycobacterium tuberculosis* from western Maharashtra, India, exhibit a high degree of diversity and strain-specific associations with drug resistance, cavitary disease, and treatment failure. J Clin Microbiol 2010; 48(10): 3593-3599.

24. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of *Mycobacterium tuberculosis* W Beijing family strains.Trends Microbiol 2002; 10: 45-52.

For Review Only

25. Molina-Torres CA, Moreno-Torres E, Ocampo-Candiani J, et al. *Mycobacterium tuberculosis* spoligotypes in Monterrey, Mexico. J Clin Microbiol 2010; 48(2): 448-455.

26. Martinez-Guarneros A, Rastogi N, et al. Genetic diversity among multidrug-resistant *Mycobacterium tuberculosis* strains in Mexico. Infect Genet Evol 2013; 14: 434-443.

27. Arora J, Bhalla M, Sidiq Z, et al. Predominance of Beijing genotype in Extensively Drug Resistant *Mycobacterium tuberculosis* isolates from a tertiary care hospital in New Delhi, India. Int J Mycobacteriology 2013; 2: 109-113.

28. Workalemahu B, Berg S, Tsegaye W, et al. Genotype diversity of *Mycobacterium isolates* from children in Jimma, Ethiopia. BMC Res Notes 2013; 4;6(1): 352.

Legend to Figure

Figure 1: An illustration of potential evolutionary relationships between the *M. tuberculosis* spoligotypes from pediatric TB patients in Delhi as seen through a Minimum spanning tree (A), and a representation of parent to descendant spoligotypes determined by Spoligoforests illustrated by a Hierarchical Layout of SITs and orphan patterns (B), and by Fruchterman Reingold representation of SITs alone (C).

For Review Only

Table 1: Exploration of phylogenetic associations with drug resistance of pediatric *M. tuberculosis* strains (n=90). Note that patterns were grouped in corresponding lineages.

Lineage groupings	Octal Code (SIT) Lineage Spoligotype pattern [a]	Total (%) in study	% in study vs. database	S	R	MDR
Beijing (n=30)	000000000003771 (SIT1) Beijing	**28 (31.11)**	**0.27**	5 (16.7)	25 (83.3)	23 (76.7)
	000000000000371 (SIT250) Beijing	**2 (2.22)**	**5.56**			
CAS (n=41)	703777740003771 (SIT26) CAS1-Delhi	**27 (30.0)**	**1.79**	15 (36.6)	26 (63.4)	20 (48.8)
	703777740003171 (SIT25) CAS1-Delhi	**3 (3.33)**	**0.49**			
	703767740003771 (SIT141) CAS1-Delhi	1 (1.11)	4.55			
	703777740003371 (SIT428) CAS1-Delhi	1 (1.11)	4.76			
	703777400003771 (SIT485) CAS1-Delhi	1 (1.11)	4.17			
	702777740003771 (SIT1092) CAS1-Delhi	1 (1.11)	4.17			
	703737740003771 (SIT1343) CAS1-Delhi	1 (1.11)	7.69			
	703777340003771 (SIT1590) CAS1-Delhi	1 (1.11)	12.5			
	703307740003771 (Orphan-o2) CAS1-Delhi	1 (1.11)	100			
	700377740003771 (SIT288) CAS2	1 (1.11)	0.9			
	703777700003171 (SIT2392) CAS	1 (1.11)	11.11			
	703737000003371 (Orphan-o3) CAS	1 (1.11)	100			
	703775000000171 (Orphan-o6) CAS	1 (1.11)	100			
T (n=7)	777777777760771 (SIT53) T1	**6 (6.67)**	**0.09**	3 (42.9)	3 (42.9)	3 (42.9)
	777737777760771 (SIT37) T3	1 (1.11)	0.22			
EAI (n=5)	477777777413071 (SIT11) EAI3-IND	1 (1.11)	0.18	2 (40.0)	3 (60.0)	2 (40.0)
	777777777413731 (SIT48) EAI1-SOM	1 (1.11)	0.23			
	777700777413700 (SIT763) EAI5	1 (1.11)	8.33			
	474000077413771 (Orphan-o1) EAI5	1 (1.11)	100			
	777347777413600 (Orphan-o7) EAI5	1 (1.11)	100			
Haarlem (n=2)	777777777720771 (SIT50) H3	1 (1.11)	0.03	0 (0.0)	1 (50.0)	1 (50.0)
	777757777720771 (SIT764) H3	1 (1.11)	10			
Unknown (n=5)	703777747770371 (SIT27) Unknown	1 (1.11)	4.35	2 (40.0)	3 (60.0)	2 (40.0)
	703777707770371 (SIT1887) Unknown	1 (1.11)	5.88			
	700000000000771 (SIT3188) Unknown	1 (1.11)	33.33			
	777775000000000 (Orphan-o4) Unknown	1 (1.11)	100			
	700167760001771 (Orphan-o5) Unknown	1 (1.11)	100			

[a] The seven orphan strains are also marked as o1 to o7 as in the Spoligoforest tree shown in Figure 1B. Note that the tentative lineage determination for the orphan strains was done manually based on the SpolDB4 rules revised according to SITVITWEB. The database IsoNumber of the respective orphan strains is as follows: o1, IND1220103938; o2, IND12201021691; o3, IND12201121839; o4, IND12201132165; o5, IND12201112534; o6, IND12201123048; o7, IND12201113169. Data for clustered strains are marked in bold.

[b] MDR, multidrug resistant TB. Note that no DST information was available in case of a single strain each for H and T lineage groups.

For Review Only

Table 2: Description of clustered spoligotype patterns (2 to 28 strains per cluster) observed among pediatric *M. tuberculosis* clinical isolates (n=90) in Delhi, corresponding Spoligotype International Type (SIT) numbers, genotypic lineages, and their worldwide distribution in the SITVIT2 database.

SIT (Lineage) Octal Number Spoligotype Description *	Total (%) in study / vs. Database	Distribution in Regions with ≥3% of a given SIT [a]	Distribution in countries with ≥ 3% of a given SITs [b]
000000000003771 (SIT1) Beijing	28 (31.11); 0.27%	ASIA-E 32.0, AMER-N 19.56, ASIA-SE 9.83, AFRI-S 8.07, ASIA-N 6.78, ASIA-S 5.01, EURO-N3.56, AMER-S 3.47	USA 19.31, CHN 18.49, JPN 11.22, ZAF 8.07, RUS 6.78, VNM 3.80, PER 3.07
703777740003771 (SIT26) CAS1-Delhi	27 (30.0); 1.79%	ASIA-S 50.13, AMER-N 14.5, ASIA-W 14.11, EURO-W 5.63, AFRI-E 4.70, EURO-N 4.37, EURO-S 3.31	IND 30.27, USA 14.5, PAK 9.54, OMN 7.22, BGD 5.5, SAU 5.17, IRN 4.11, ITA 3.18
777777777760771 (SIT53) T1	6 (6.67); 0.09%	EURO-W 15.75, AMER-N 13.57, AMER-S 12.11, EURO-S 9.47, EURO-N 7.53, ASIA-W 7.35, AFRI-S 5.0, AFRI-E 4.53, ASIA-E 4.26, AFRI-N 3.54, EURO-E 3.28, AMER-C 3.25	USA 13.28, FXX 7.93, ITA 5.37, BRA 5.16, ZAF 4.88, TUR 3.49, AUT 3.44, CHN 3.11
703777740003171 (SIT25) CAS1-Delhi	3 (3.33); 0.49	ASIA-W 27.63, AFRI-E 24.84, AFRI-N 13.82, ASIA-S 12.66, AMER-N 8.55, EURO-N 4.61, EURO-W 4.44	ETH 23.68, SAU 15.46, SDN 11.35, USA 8.55, IND 8.06, OMN 6.91, IRQ 4.77
000000000000371 (SIT250) Beijing	2 (2.22); 5.56	ASIA-E 38.89, AMER-N 30.56, ASIA-SE 13.89, ASIA-S 11.11	USA 30.56, CHN 22.22, TWN 13.89, IND 8.33, VNM 5.56, IDN 5.56

[a] Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database (***data updated on September 3rd 2013***). The definition of macro-geographical regions and sub-regions is according to the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm); Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia.

[b] The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ≥5% of a given SITs as compared to their total number in the SITVIT2 database.

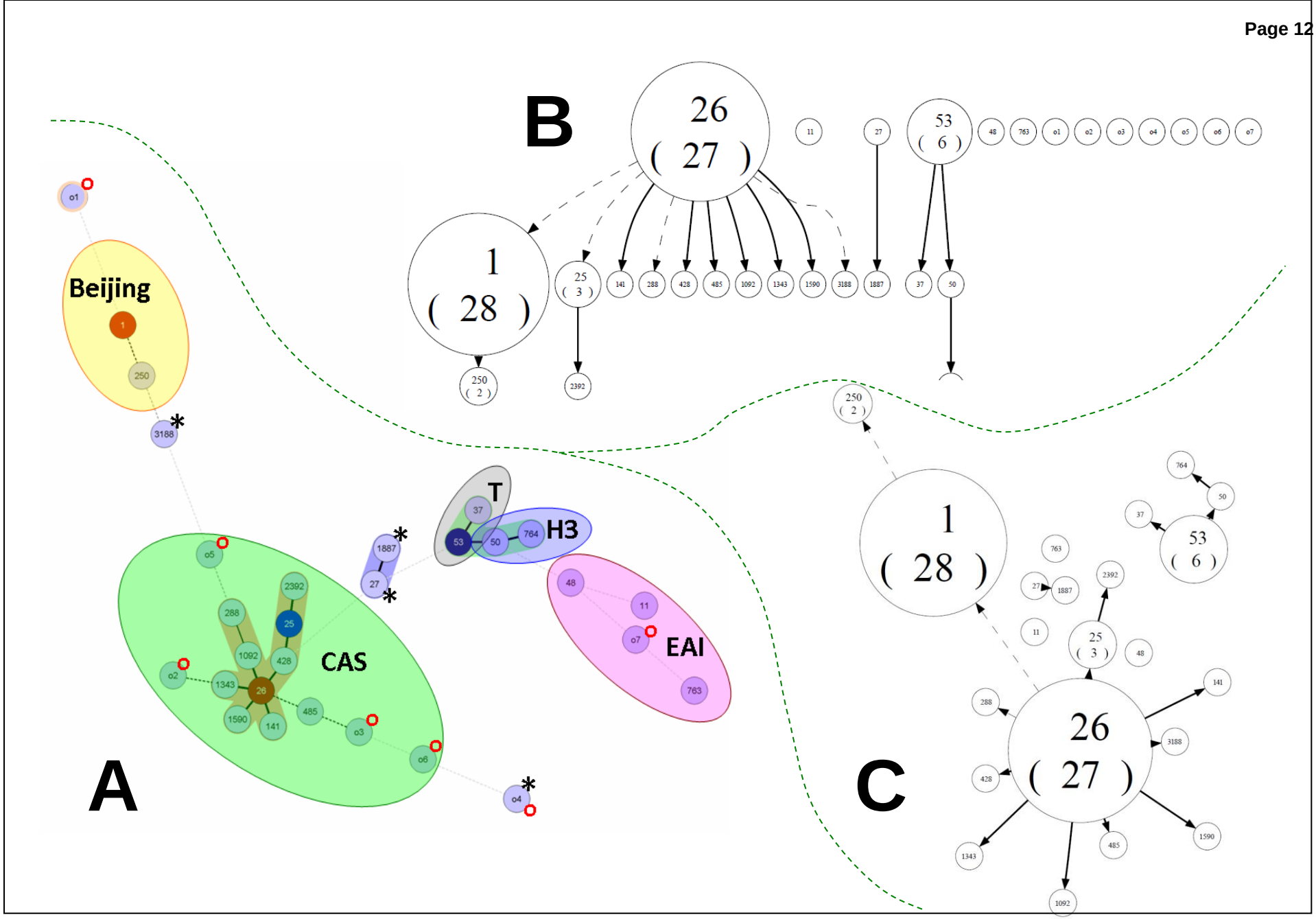
For Review Only

Table 3: Distribution in neighboring countries of spoligotype patterns corresponding to clustered *M. tuberculosis* spoligotype patterns (2 to 28 strains per cluster) observed among pediatric *M. tuberculosis* clinical isolates (n=90) in Delhi, as seen through the updated SITVIT2 database. [a]

	India			Distribution in neighboring countries (SITVIT2)						
	A	B	C							
Octal Code (SIT) Lineage	This study N/t (%)	New Delhi SITVIT2, N/t (%)	All India SITVIT2, N/t (%)	Bangladesh N/t (%)	Pakistan N/t (%)	Sri Lanka N/t (%)	Nepal N/t (%)	China N/t (%)	Myanmar N/t (%)	Saudi-Arabia N/t (%)
000000000003771 (SIT1) Beijing	28/90 (31.11)	209/2148 (9.73)	262/2647 (9.90)	181/1130 (16.02)	23/385 (2.97)	14/100 (14.0)	8/49 (16.33)	1885/3072 (61.36)	98/310 (31.61)	76/1818 (4.18)
703777740003771 (SIT26) CAS1-Delhi	27/90 (30.0)	420/2148 (19.55)	457/2647 (17.26)	83/1130 (7.35)	144/385 (37.40)	0/100 (0)	11/49 (22.45)	4/3072 (0.13)	5/310 (1.61)	78/1818 (4.29)
777777777760771 (SIT53) T1	6/90 (6.67)	78/2148 (3.63)	92/2647 (3.48)	38/1130 (3.36)	11/385 (2.86)	2/100 (2.0)	2/49 (4.08)	194/3072 (6.31)	3/310 (0.97)	157/1818 (8.64)
703777740003171 (SIT25) CAS1-Delhi	3/90 (3.33)	44/2148 (2.05)	49/2647 (1.85)	4/1130 (0.35)	6/385 (0.78)	0/100 (0)	1/89 (1.12)	1/3072 (0.03)	0/310 (0)	94/1818 (5.17)
000000000000371 (SIT250) Beijing	2/90 (2.22)	0/2148 (0)	3/2647 (0.11)	1/1130 (0.09)	0/385 (0)	0/100 (0)	0/49 (0)	8/3072 (0.26)	0/310 (0)	0/1818 (0)

[a] Distribution in neighboring countries was performed by comparing the patterns reported in the SITVIT2 database on September 3rd 2013. Note that the data shown for New Delhi (Column B) does not take in account the data on 90 isolates from this study (Column A), however the data shown for all over India (column C) takes in account all data available on September 3rd 2013.

For Review Only


B
26
(27)
1
(28)
25
(3)
53
(6)
250
(2)
11
27
48
763
o1
o2
o3
o4
o5
o6
o7
141
288
428
485
1092
1343
1590
3188
1887
37
50
2392
Beijing
1
250
3188
o1
T
37
53
50
764
H3
1887
27
48
11
o7
EAI
763
o5
2392
288
25
1092
428
CAS
o2
1343
26
1590
141
485
o3
o6
o4
A
C
1
(28)
764
50
37
53
(6)
763
27
1887
2392
11
25
(3)
48
141
288
26
(27)
3188
428
1343
485
1590
1092

Note: The MST connects each *M. tuberculosis* spoligotype based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dashed and dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines, 1 or 2 or 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); gray dashed lines represent 4 changes; and gray dotted lines represent 5 or more changes. The color of the circle indicates number of isolates in our study: unique isolates (sky blue) versus clustered isolates (deep blue, 2-5 strains; red, 6 strains and more). The 7 orphan strains (highlighted by small red circles) are marked as o1 to o7, while strains with an unknown lineage signature are marked by an asterisk. The representation of parent to descendant spoligotypes determined by Spoligoforests is illustrated by a Hierarchical Layout of SITs and orphan patterns in **Figure 1B**, and by Fruchterman Reingold representation (SITs only) in **Figure 1C**.). In Spoligoforests, each spoligotype pattern from the study is represented by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. In this representation, the heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weigh being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5. The tree in **Figure 1B** represents the changes hierarchically, i.e., the more the strains evolve (lose spacers), the more they are present in the down layouts; however, in case of too many changes between two patterns, there are no links/edges linking them (like some of the orphan patterns shown as the smallest circles on the top). The Fruchterman Reingold tree (**Figure 1C**) essentially represents the same relationships as in the Hierarchical layout, but the nodes containing larger proportion of strains are centered. Considering that the patterns evolve by loss of spacers, the proximity of the nodes represents similarity between them. Note that this figure was drawn without orphan patterns to better show the SIT numbers (in circles) followed by the number of strains for each pattern (shown in brackets). The interpretation of the links (solid black, dashed, or dotted) is the same as for Hierarchical Layout.

Article 14

Infection, Genetics and Evolution 12 (2012) 718–724

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

The population structure of drug-resistant *Mycobacterium tuberculosis* clinical isolates from Sichuan in China

Yuding Zhao [a], Qin Feng [a], Ke Tang [a], Congcong Zhang [a], Honghu Sun [a], Tao Luo [b], Zhirong Yang [a], David Couvin [c], Nalin Rastogi [c], Qun Sun [a,*]

[a] Key Laboratory of Bio-resources and Eco-environment, Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610064, PR China
[b] Chengdu Anti-Tuberculosis Hospital, Chengdu, Sichuan 610016, PR China
[c] WHO Supranational TB Reference Laboratory, Institut Pasteur de la Guadeloupe, Abymes 97183, Guadeloupe, France

ARTICLE INFO

Article history:
Available online 2 October 2011

Keywords:
Mycobacterium tuberculosis
Spoligotyping
MIRU-VNTRs
Drug-resistance
Genetic diversity
Sichuan

ABSTRACT

China ranks second next to India among 22 high-burden countries despite decades' effort on tuberculosis (TB) control. The Sichuan province today contains the second-largest number of TB cases among Chinese provinces, where the prevalence of drug-resistant TB, especially MDR-TB, is much higher than the average level in eastern China. In this study, the population structure and the transmission characteristics of drug-resistant TB in Sichuan province were studied by spoligotyping and 24-locus Mycobacterial interspersed repetitive units-variable number tandem DNA repeats (MIRU-VNTR), applied to a total of 306 clinical isolates. Spoligotyping-based analysis showed that Beijing family represented 69.28% of all isolates and constituted the largest group (66.24%) of MDR-TB in Sichuan. The remaining isolates, accounting for 33.76% of MDR isolates, belonged to the ill-defined T family, Manu2, H3, LAM9, and other minor unassigned clades. The discriminatory power evaluated for spoligotyping was poor (HGI = 0.595), but high for 24-locus MIRU-VNTRs (HGI = 0.999). The number of the most discriminatory loci ($h > 0.6$) was 12, including locus 424, 802, 960, 1644, 1955, 2163b, 2996, 3007, 3192, 3690, 4348 and 4052. It was concluded that 24-locus MIRU-VNTRs could be a more discriminatory tool for differentiating clinical isolates from Sichuan region. The small clustering size obtained from the current population structure analysis suggested that the high prevalence of drug-resistant TB in this region might be attributed partially to the acquired resistance due to inappropriate drug use rather than active transmission of drug-resistant TB (primary resistance).

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Tuberculosis (TB) is still one of the most prevalent infectious diseases in the world with an estimated 9.4 million incident cases of TB in 2009 (WHO, 2010). China is one of the 22 high-burden countries in the world with total TB cases of 1.3 million, and a mortality rate of 11.54%. These figures rank China at the second place for total number of TB cases, only behind India over the world (Duanmu, 2002; WHO, 2010). Although a variety of efforts have been directed towards the control of TB transmission for decades in China, imbalance of economy development and prevalence of drug-resistant TB still make China the focus of worldwide TB control programme. Sichuan province, located in the southwest of China with more than 87,000,000 populations, has become the second largest TB case province in recent years (Yang et al., 2008). Previous study released a significant increase in the prevalence of multidrug resistant (MDR)-TB in Sichuan during 1985–2000 (Liu et al., 2002). Documents of TB Control Project Office of Sichuan Province showed the rate of MDR-TB in sputum smear-positive patients at 28.3% and the proportion of MDR-TB among patients harboring strains showing any drug-resistance at 74% (Yang et al., 2008). The prevalence of drug-resistant TB, especially MDR-TB, in this area is much higher than the average level of other provinces in eastern region of China (CDC, 2006; Liu et al., 2003, 2002; NTSG, 2002; Yang et al., 2008). In addition, the relatively high level of migration and co-infection with HIV, as well as a relatively poorly organized medical service in charge of patient follow-up during anti-TB treatment, makes state projected TB control almost a goal out of reach in near future, unless efficient surveillance of drug-resistant TB in Sichuan steps in to support TB-control strategies.

In this context, the molecular epidemiological studies greatly facilitate human efforts to detect epidemic links among TB cases, differentiate reactivation from re-infection, identify the transmission of drug-resistant isolates, as well as clarify the laboratory cross-contamination (Jiao et al., 2008). Previous gold-standard for molecular

* Corresponding author. Address: 29# Wangjiang Road, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610064, PR China. Tel.: +86 28 8541 8810; fax: +86 28 8546 0487.
E-mail address: qunsun@scu.edu.cn (Q. Sun).

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.09.022

epidemiological studies relied on IS*6110* restriction fragment length polymorphism (IS*6110*-RFLP). Unfortunately, time-consuming and technique-demanding performance of this method has limited its application in population-based epidemiological studies. Moreover, it takes several weeks to culture bacteria in order to acquire large enough quantities of DNAs needed for southern hybridization, and analysis of IS*6110* fingerprints needs special software. Mycobacterial interspersed repetitive units-variable-number tandem repeats (MIRU-VNTR) analysis, a PCR-based discriminatory typing method, meets the need of easy performance and high-throughput detection (Mazars et al., 2001). In addition, numerical results obtained from MIRU-VNTR analysis can be exchanged among laboratories globally (Maes et al., 2008). A novel set of 24-locus MIRU-VNTR analysis, an updated version of original 12-locus MIRU-VNTR, proposed by Supply et al. (2006), achieved great discriminatory power by the implementation of more loci scattered throughout genome of *Mycobacterium tuberculosis* (Supply et al., 2000; Valcheva et al., 2008). Especially when combined with spoligotyping, discriminatory power of 24-locus MIRU-VNTR was reported to be equal or higher to that of IS*6110*-RFLP (Maes et al., 2008; Oelemann et al., 2007; Phyu et al., 2009; Shamputa et al., 2010).

Though molecular epidemiological studies on TB have been initiated in some regions in China, a detailed analysis of genetic diversity of *M. tuberculosis* isolates, especially population structure of drug-resistant TB in Sichuan, the second-largest TB case region, is still not available. On the other hand, 24 MIRU scheme is not sufficient for discrimination of the Beijing genotype strains when compared to IS6110-RFLP as was shown in Beijing, China (Jiao et al., 2008), Russia (Mokrousov et al., 2008) and Japan (Iwamoto et al., 2007). In this study, we investigated the population structure of drug-resistant *M. tuberculosis* isolates, and genotyped 306 drug-resistant isolates from Sichuan province by applying spoligotyping and 24-locus MIRU-VNTRs. The discriminatory power of each approach used alone or in combination was also evaluated.

2. Materials and methods

2.1. Bacterial isolates and drug susceptibility testing (DST)

All isolates in this study represented 43.47% of total 708 drug-resistant *M. tuberculosis* isolates from 2,723 pulmonary tuberculosis patients from different regions of Sichuan province who were diagnosed in Chengdu anti-TB hospital, the only professional anti-tuberculosis hospital in Sichuan, from January 2008 to August 2009. Samples were obtained from 204 male and 102 female of age ranging from 11 to 78 years. Clinical data were obtained from the subjects' medical records without any invasion of patients' privacy. Furthermore, no results from this study influenced the treatment of patients. Specimens were collected and disposed in accordance with WHO guidelines (WHO, 1997). Culture positive isolates were cultured on Löwenstein Jensen (LJ) slants at 37 °C and subjected to drug susceptibility testing described elsewhere (Canetti et al., 1969) by using proportion method with following drug concentrations – streptomycin, 10 μg/ml; isoniazid, 0.2 μg/ml; rifampin, 40 μg/ml and ethambutol, 2 μg/ml. Culture with paranitrobenzoic acid and 2-thiopnene carboryic acid hydrazide were also used to identify the isolates bacteriologically. A total of 306 *M. tuberculosis* clinical isolates with drug-resistance to any first line anti-tuberculosis antibiotics (streptomycin, rifampin, isoniazid and ethambutol) were randomly selected. Genome DNA from clinical isolates was extracted using the CTAB method (van Soolingen et al., 1991).

2.2. Genotyping

Spoligotyping was performed as described by Kamerbeek et al. (1997). Briefly, PCR-amplified biotin-labeled DR locus is hybridized against an array of 43 different immobilized DR spacers in a nylon membrane. The resulting hybridization signals were revealed by chemiluminescence. The profiles were compared with SITVIT2 database (Pasteur Institute of Guadeloupe), which is an updated version of the previously released SpolDB4 database (Brudey et al., 2006) (available online at http://www.pasteurguadeloupe.fr:8081/SITVITDemo). At the time of the present study, the SITVIT2 database contained genotyping data on almost 86,000 *M. tuberculosis* isolates from 160 countries of origin (spoligotyping and mycobacterial interspersed repetitive-unit-variable-number tandem repeats [MIRU-VNTRs] for about 75,000 and 11,000 isolates, respectively). Major phylogenetic clades were assigned according to the signatures provided in the SpolDB4 database, which defined 62 genetic lineages/sublineages (Brudey et al., 2006). For MIRU-VNTR genotyping, all the 24-loci were amplified with corresponding primers described by Supply et al. (2006). Briefly, 1 μl template DNA was added up to a total 25 μl PCR reaction volume including 12.5 μl 2*Taq* PCR master mix (Tiangen, Beijing, China) and 10 pmol of each primer. The H37Rv isolate was run as an additional control of the performance of the method. The PCR products were detected by electrophoresis in 2.0% agarose gel using a 100-bp DNA ladder (TAKARA) as molecular marker. The size analysis of PCR fragments and assignment of the various VNTR alleles were achieved by Quantity One (version 4.6.2) software (BIO-RAD Laboratories) and by comparison with the correspondence tables kindly provided by Philip Supply (Institut Pasteur de Lille, Lille, France).

2.3. Statistical analysis

The allelic diversity of the VNTR loci was calculated using $h = 1-\sum x_i^2\ [n/(n-1)]$, where x_i is an allele's frequency at the locus and n is the number of isolates. The Hunter–Gaston index (HGI) was calculated as described previously (Hunter and Gaston, 1998) and was used to evaluate the level of discriminatory power of the typing methods. Dendrogram was generated using program available online at www.miru-vntrplus.org (Allix-Béguec et al., 2008; Weniger et al., 2010) with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). Fully identical patterns of *M. tuberculosis* isolates from different patients were assigned to same cluster. The clustering rate was defined as $(n_c-c)/n$, where n_c is the total number of clustered cases, c is the number of clusters, and n is the total number of cases in the sample.

3. Results

3.1. Drug-resistance profile of clinical isolates

Classification of all isolates investigated in this study according to patients' age and gender were described in Table 1. Results of drug susceptible testing (Supplementary Table 1) showed that the number of isolates with mono-resistance to streptomycin, isoniazid, rifampin and ethambutol were 30 (9.8%), 10 (3.27%), 19 (6.21%) and 4 (1.31%), respectively, while a total of 157 (51.3%) isolates were multi-drug resistant (resistance to both rifampin and isoniazid) ones. The remaining 86 (28.1%) isolates showed one of other resistant patterns (non-MDR).

3.2. Spoligotyping-based diversity of isolates

As summarized in Table 2, spoligotyping resulted in 9 orphan patterns and 52 shared types ($n = 297$ isolates) distributing in five lineages; note that 10 isolates showed patterns that were not assigned to any known major lineages in the SITVIT2 database, and were labeled as "unknown". Seven genotypes SIT1, SIT53, SIT334,

Table 1
Classification of 306 drug-resistant *M. tuberculosis* isolates from Sichuan based on age and gender.

Parameters	Number (%) of total drug-resistant isolates investigated (this study)	Number (%) of MDR isolates investigated (this study)	Number (%) of total drug-resistant isolates (n = 708)	Number (%) of total isolates (n = 2723)
Age				
⩽20	27 (8.82)	14 (4.58)	62 (8.76)	278 (10.21)
21–40	172 (56.21)	91 (29.74)	370 (52.26)	1402 (51.49)
41–60	83 (27.12)	44 (14.38)	197 (27.82)	711 (26.11)
⩾61	24 (7.84)	9 (2.94)	79 (11.16)	332 (12.19)
Gender				
Male	204 (66.67)	110 (35.95)	483 (68.22)	1884 (69.19)
Female	102 (33.33)	47 (15.36)	225 (31.78)	839 (30.81)

SIT52, SIT632, SIT54 and SIT1096 found in current study, with at least 4 isolates included in each type, were distributed worldwide (Table 3). A total of 212 isolates were Beijing genotype from our observation. As expected, Beijing family was predominant group that represented 69.28% of all isolates. Among these isolates, 194 (63.4%) isolates corresponded to the typical spoligotype pattern (SIT1) of the Beijing family (absence of the first 34 spacers and presence of the last 9 spacers). The remaining 18 isolates were characterized by difference of one or more spacers assigned to Beijing family. Fifty-seven isolates belonging to ill-defined T family were observed, which constituted 18.63% of all isolates. Within this family, SIT53/T1, SIT334/T1and SIT52/T2 were most frequent. Eighteen isolates were assigned to three other families, Manu2, H3, and LAM9, respectively.

Distribution of drug-resistant patterns in clades (Table 4) indicated that the largest group of MDR isolates (104 isolates, 33.97%) belonged to Beijing family in this study, and 92 (30.01%) isolates with identical spoligotype (SIT1) were found to be the predominant one in this group. The second-largest group of MDR isolates (32 isolates, 10.46%) was found among T family. Besides MDR pattern, varied resistance including mono-resistance and Non-MDR was found among both Beijing family and T family. No mono-resistance was observed in strains labeled as "Unknown" lineages and Manu2 family. Six MDR isolates, one Non-MDR and two STE-mono-resistant isolates corresponded to orphan spoligotypes. Two families, H3 and LAM9 represented only a small proportion of drug-resistant isolates in this study.

3.3. Discriminatory power of genotyping methods

The discriminatory power of each genotyping method is described in Table 5. Spoligotyping showed a poor discriminatory power (HGI = 0.595); among 61 different types, spoligotyping identified 42 unique isolates (13.73%) and 264 (86.27%) isolates grouped in 19 clusters, with a high clustering rate of 80.1%. The largest cluster including 194 isolates was assigned to the typical Beijing family (ST1). Compared with spoligotyping, original set of 12-MIRUs alone identified 218 unique isolates and 88 isolates in 31 clusters with a significantly higher increased discriminatory power (HGI = 0.997). When used in combination with spoligotyping, 12-MIRU achieved an HGI value of 0.998, with a reduced clustering rate of 14.38%. The same resolution (HGI = 0.999) was observed from both sets of 15-locus VNTR and 24-locus VNTR analysis applied alone or in combination with spoligotyping.

The allelic diversity among 24 MIRU-VNTR loci (Table 6) differed from 0 to 0.837 in this study. Highly diverse (h > 0.6) VNTR loci among both Beijing and non-Beijing isolates were found to be MIRU16, MIRU26, MIRU39, Mtub04, Mtub21, Mtub39, QUB-11b, and QUB26. However, based on the value of allelic diversity (h > 0.6), five additional loci, including MIRU04 (h = 0.629), MIRU27 (h = 0.648), MIRU31 (h = 0.669), MIRU40 (h = 0.712) and ETR-A (h = 0.662), were more diverse among non-Beijing isolates than in Beijing isolates, although one locus (QUB-4156) in Beijing isolates (h = 0.611) was highly diverse compared with in non-Beijing isolates (h = 0.489) and all isolates (h = 0.581) in this study.

4. Discussion

Previous studies (Dong et al., 2010; Li et al., 2011) described the distribution of genotypes of *M. Tuberculosis* isolates from some areas in China. However, a detailed analysis of the population structure of *M. tuberculosis* isolates from Sichuan, the second-largest TB case province where the prevalence of drug-resistant TB is much higher than the average level observed in eastern China, was not available. Consequently, the present study not only attempted to map the population structure of drug-resistant *M. Tuberculosis* isolates from Sichuan, but also assessed the utility of 24-locus MIRU-VNTR in differentiating these drug-resistant isolates.

Population structure of *M. tuberculosis* complex isolates in this study showed that Beijing family represented by its prototype SIT1 was the most predominant genotype in Sichuan province. Beijing family, which is distributed globally, represents a high proportion of *M. tuberculosis* isolates causing ongoing TB transmission in Russia, Central Asia and East Asia (Mokrousov, 2008; Niemann et al., 2009; Wada et al., 2009). Up to date, this family has been reported to represent about 13% isolates worldwide and more than 50% of isolates in Asia (Kim et al., 2001; Park et al., 2000; Parwati et al., 2010), with even higher proportions (up to 92.59%) in Beijing and its surrounding areas in China (Dong et al., 2010). Our study showed a concordance with this the range reported above. Nonetheless, our results showed a relatively higher diversity of *M. tuberculosis* complex population structure as compared with that in northern China, including Beijing, Jilin and Gansu, where the prevalence of Beijing isolates exceeds 80% (Dong et al., 2010). The results in present study showed a higher proportion of Beijing isolates from Sichuan than that (54.8%) reported by Li et al. (2011), which probably was due to the difference of sampling time and size, locations, as well as the interest that we mainly focus on drug-resistant isolates. However, the similar distribution of lineages of non-Beijing isolates and the dominance of ill-defined T family among non-Beijing families in Sichuan were observed between our study and that of Li et al. (2011). Taken together, these results may help us to understand the phylogeography of *M. tuberculosis* circulating in this area. Interestingly, the prevalence of Beijing family in three provinces neighboring Sichuan at north, including Tibet (90.38%), Gansu (87.5%) and Shanxi (80%) (Dong et al., 2010), is much higher than that in Sichuan, which probably was due to Sichuan's geographic distance from the north as well as its diversity structure of ethnic groups and climatic specificity. In our study, isolates among T family, Manu2, H3, LAM9 and other unassigned clades represented 30.72% of all isolates, and these families accounted for 33.76% of all MDR isolates (or 17.32% of all isolates). Available publications (Liu et al., 2008; Shi et al., 2007; Tong et al., 2010) reported similar proportion of population of Beijing family in three above provinces as Dong et al. (2010), while non-Beijing isolates account less than 20% of the population of MDR isolates in these areas in northern China. Hence these "non-Beijing" lineages are spreading in southern China, and

Table 2
Description of 52 shared-types (SITs; n = 297 isolates) and 9 orphan isolates with corresponding spoligotyping defined lineages/sublineages starting from a total of 306 isolates of *M. tuberculosis* in Sichuan.

SIT[a]	Spoligotype description	Octal number	Number (%) in study	% In study vs. database	Lineage[b]	Clustered vs. unique patterns[c]
1	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■	000000000003771	194 (63.4)	2.12	Beijing	Clustered
37	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777737777760771	1 (0.33)	0.29	T3	Unique
42	■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	777777607760771	1 (0.33)	0.03	LAM9	Unique
44	■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■□□□□■■■■■■■■	777777757760771	1 (0.33)	0.52	T5	Unique
50	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■■	777777777720771	3 (0.98)	0.1	H3	Clustered
52	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■■	777777777760731	8 (2.61)	1.06	T2	Clustered
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777777777760771	11 (3.59)	0.21	T1	Clustered
54	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■■	777777777763771	4 (1.31)	1.95	Manu2	Clustered
118	■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777767777760771	1 (0.33)	0.75	T1	Unique
159	■■■■■■■■■■■■■■□□□□□□□□□□■■■■■■■■■■□□□□■■■■■■■■	777740017760771	1 (0.33)	1.47	Unknown	Unique
189	■■■■■■■■■■■■■■□□□□■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777741777760771	1 (0.33)	10	T1	Unique
190	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■□■■■	000000000003731	3 (0.98)	2.14	Beijing	Clustered
196	■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	677777777760771	1 (0.33)	1.41	T1	Unique
250	□□■■■■■■	000000000000371	1 (0.33)	3.13	Beijing	Unique
265	□□■■□■■■■■	000000000003371	1 (0.33)	1.47	Beijing	Unique
278	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■□■	777777777760761	1 (0.33)	9.09	T1	Unique
291	■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■□□□□■■■■■■■■	777777677760771	1 (0.33)	1.23	T1	Unique
294	■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■■	577777777720771	2 (0.65)	7.69	H3	Clustered
334	■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	577777777760771	9 (2.94)	10.71	T1	Clustered
393	■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777757777760771	3 (0.98)	9.68	T1	Clustered
462	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■□□□□■■■■■■■■	777777777560771	1 (0.33)	2.08	T1	Unique
511	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□■□□□□■■■■■■■■	777777700020771	1 (0.33)	2.44	H3	Unique
583	■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■■	777737777763771	1 (0.33)	33.33	Manu2	Unique
621	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■□■■■■■■■	000000000002771	1 (0.33)	3.23	Beijing	Unique
632	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■□■■■■■	000000000003571	5 (1.63)	20.83	Beijing	Clustered
780	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■■■■■■	777777777600771	1 (0.33)	10	Unknown	Unique
803	■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■■□□□□■■■■■■■■	777740007760771	1 (0.33)	3.23	LAM9	Unique
804	■□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	477777777760771	1 (0.33)	4.55	T1	Unique
834	■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777767777740771	1 (0.33)	12.5	T1	Unique
848	■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	737777777760731	1 (0.33)	3.85	T2	Unique
1096	■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■	577777777763771	4 (1.31)	44.44	Manu2	Clustered
1122	■■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	767777777760771	1 (0.33)	2.56	T1	Unique
1168	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■□□■■■	000000000003631	1 (0.33)	20	Beijing	Unique
1302	■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	577777777760731	3 (0.98)	16.67	T2	Clustered
1451	■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□	777760000000000	1 (0.33)	7.14	Unknown	Unique
1475	■■■■■■■■■□■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777357777760771	1 (0.33)	16.67	T1	Unique
1578	■■■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	767777777760731	1 (0.33)	20	T2	Unique
1674	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■□■	000000000003761	1 (0.33)	14.29	Beijing	Unique
1690	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■□■■■■■■■	777777777762771	1 (0.33)	25	Manu2	Unique
1890	■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■■■■□□□□■■■□■■■	777777707760731	1 (0.33)	20	T2	Unique
2276	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■■■	777777777743771	2 (0.65)	50	Unknown	Clustered
2979	□□□■■■■■	000000000000171	2 (0.65)	28.57	Beijing	Clustered
3199	■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	771777777760731	1 (0.33)	33.33	T2	Unique
3233*	■■■□□□□□□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	700377777760731	3 (0.98)	100	T	Clustered
3234*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□■■■■■■■	777777777600371	2 (0.65)	66.67	Unknown	Clustered
3235*	■■□□□□■■■■■■■■■■■■■■■■■□□□□□□□■■□□□□■■■□■■■	607777760060731	2 (0.65)	100	T	Clustered
3236*	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□■■■	000000000002031	1 (0.33)	50	Beijing	Unique
3237*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□■■	777777777400011	1 (0.33)	50	Unknown	Unique
3238*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■□□□□	777777777740300	2 (0.65)	100	T	Clustered
3239*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□■■■■■■	777760000000731	1 (0.33)	50	Unknown	Unique
3240*	■■■■■■■■■□□□■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	776167777760731	1 (0.33)	50	Unknown	Unique
3241*	□□□■■■□□□□□	000000000003600	2 (0.65)	66.67	Beijing	Clustered
Orphan	■■■■■■■■■■■■■■■■■■■■■■■□□□■□□□■■□□□□■■■■■■■■■■	777777704303771	1 (0.33)	100	Unknown	Unique
Orphan	■■□□■■	600000000000011	1 (0.33)	100	Unknown	Unique
Orphan	■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■□□□□■□■■■■■■	777740007760571	1 (0.33)	100	LAM	Unique
Orphan	■■□■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■■■■■■■	667777776360771	1 (0.33)	100	T1	Unique
Orphan	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■	777737777760571	1 (0.33)	100	T3	Unique
Orphan	■■■■■□■■■■■■■□□□□□□□□□□□□■■■■■■■■■□□□□■■■■■■■	767740003760771	1 (0.33)	100	Ambiguous (T/LAM)	Unique
Orphan	■■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■■□□□□■□■□■■■	777757777760531	1 (0.33)	100	T2	Unique
Orphan	■□□■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	477477777760731	1 (0.33)	100	T2	Unique
Orphan	■□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■■	477777777763771	1 (0.33)	100	Manu2	Unique

[a] A total of 43/52 SITs matched a preexisting shared-type in the database, whereas 9/52 SITs (n = 15 isolates) were newly created either within the present study or after a match with an orphan in the database. A total of 19 SITs containing 264 isolates were clustered within this study (2–194 isolates per cluster) while 33 isolates were unique. Note that SITs followed by an asterisk indicates ''newly created shared-type'' (n = 9 containing 15 isolates) due to 2 or more isolates belonging to an identical new pattern within this study or after a match with an orphan in the database. SIT designations followed by number of isolates: 3233*; this study n = 3; 3234* this study n = 2, LKA n = 1; 3235* this study n = 2; 3236* this study n = 1, CHN n = 1; 3237* this study n = 1, USA n = 1; 3238* this study n = 2; 3239* this study n = 1, FXX n = 1; 3240* this study n = 1, CHN n = 1; 3241* this study n = 2, USA n = 1.

[b] Clade designations according to SITVIT2 using revised SpolDB4 rules; ''Unknown'' designates patterns with signatures that do not belong to any of the major clades described in the database. Note that lineage designations for orphan patterns were done manually as expert-based interpretations using revised SpolDB4 rules.

[c] Clustered isolates correspond to a similar spoligotype pattern shared by 2 or more isolates ''within this study''; as opposed to unique isolates harboring a spoligotype pattern that does not match with another isolate from this study. Unique isolates matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as ''orphan''.

Table 3
Description of clusters containing four or more isolates of *M. tuberculosis* from Sichuan and their worldwide distribution in the SITVIT2 database.

SIT (Clade) Octal number spoligotype description	Number (%) in study	% In study vs. database	Distribution in regions with ⩾3% of a given SITs[a]	Distribution in countries with ⩾3% of a given SITs[b]
1 (Beijing) 000000000003771	194 (63.4)	2.12	ASIA-E 34.32, AMER-N 21.86, ASIA-SE 9.89, AFRI-S 9.0, ASIA-N 6.74, ASIA-S 3.71	USA 21.53, CHN 19.93, JPN 12.51, ZAF 9.0, RUS 6.74, VNM 4.23
52 (T2) 777777777760731	8 (2.69)	1.06	EURO-W 18.17, AMER-N 16.45, EURO-N 14.72, ASIA-E 10.48, ASIA-W 6.23, EURO-S 6.1, EURO-E 5.31, AFRI-M 4.38	USA 14.19, SWE 11.94, CHN 6.23, BEL 5.84, FXX 5.31, JPN 3.98, ITA 3.98, NLD 3.58, CZE 3.32, CMR 3.18, AUT 3.05
53 (T1) 777777777760771	11 (3.7)	0.21	AMER-N 18.55, AMER-S 14.84, EURO-W 11.76, EURO-S 9.2, ASIA-W 7.97, AFRI-S 5.97, EURO-N 5.93, AFRI-N 4.23, ASIA-E 3.81, AFRI-E 3.60, EURO-E 3.10	USA 15.85, BRA 6.02, ZAF 5.83, ITA 4.70, TUR 4.17, AUT 4.11, PER 3.20, SAU 3.0
54 (Manu2) 777777777763771	4 (1.31)	1.95	AFRI-N 17.56, ASIA-E 16.59, ASIA-W 12.20, ASIA-S 11.71, AMER-N 9.76, AFRI-S 7.81, ASIA-N 4.88, AMER-S 4.39	EGY 16.59, CHN 16.59, IND 10.73, USA 9.76, SAU 9.76, ZAF 7.81, RUS 4.88
334 (T1) 577777777760771	9 (3.03)	10.71	EURO-S 22.62, ASIA-E 20.24, AMER-N 10.71, EURO-N 7.14, AMER-S 7.14, EURO-W 7.14, AFRI-S 4.76, CARI 3.57, ASIA-W 3.57	CHN 19.05, ITA 15.48, USA 10.71, ESP 7.14, ZAF 4.76, GUF 3.57, GNB 3.57, GEO 3.57, BEL 3.57
632 (Beijing) 000000000003571	5 (1.63)	20.83	ASIA-E 79.17, ASIA-S 8.33, EURO-W 4.17, AUST 4.17, ASIA-W 4.17	CHN 58.33, JPN 16.67, IND 8.33, TWN 4.17, GEO 4.17, BEL 4.17
1096 (Manu2) 577777777763771	4 (1.31)	44.44	ASIA-E 77.78, ASIA-S 22.22	CHN 77.78, IND 22.22

[a] Worldwide distribution is reported for regions with more than 3% of a given SITs as compared to their total number in the SITVIT2 database. The definition of macro-geographical regions and sub-regions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) is according to the United Nations; Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). Furthermore, CARIB (Caribbean) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that in our classification scheme, Russia has been attributed a new sub-region by itself (Northern Asia) instead of including it among rest of the Eastern Europe. It reflects its geographical localization as well as due to the similarity of specific TB genotypes circulating in Russia (a majority of Beijing genotypes) with those prevalent in Central, Eastern and South-Eastern Asia.

[b] The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; countrywide distribution is only shown for SITs with ⩾5% of a given SITs as compared to their total number in the SITVIT2 database.

Table 4
Distribution of drug-resistant patterns among various genotypic lineages observed for *M. tuberculosis* isolates from Sichuan province, China.

Family	No. (%) of isolates with drug-resistant patterns					
	MDR	Non-MDR	STR	INH	RIF	EMB
Beijing	104 (33.97)	61 (19.94)	21 (6.86)	8 (2.61)	16 (5.23)	2 (0.65)
T	32 (10.46)	15 (4.9)	4 (1.31)	2 (0.65)	3 (0.98)	1 (0.33)
Unknown	9 (2.94)	1 (0.33)	0	0	0	0
Orphan	6 (1.96)	1 (0.33)	2 (0.65)	0	0	0
Manu2	4 (1.31)	6 (1.96)	0	0	0	0
H3	1 (0.33)	2 (0.65)	2 (0.65)	0	0	1 (0.33)
LAM9	1 (0.33)	0	1 (0.33)	0	0	0
Total	157 (51.3)	86 (28.1)	30 (9.8)	10 (3.27)	19 (6.21)	4 (1.31)

Table 5
Discriminatory power of different typing methods used in Sichuan province, China.

Method	No. of types	No. of unique isolates	No. of clustered isolates	No. of clusters	Cluster size (no. of isolates)	Clustering rate (%)	HGI
Spoligotyping	61	42	264	19	2–194	80.1	0.595
12-Loci MIRU	249	218	88	31	2–7	18.63	0.997
15-Loci VNTR	296	286	20	10	2	3.27	0.999
24-Loci VNTR	300	294	12	6	2	1.96	0.999
12-Loci MIRU and spoligotyping	262	235	71	27	2–7	14.38	0.998
15-Loci VNTR and spoligotyping	299	292	14	7	2	2.29	0.999
24-Loci VNTR and spoligotyping	303	300	6	3	2	0.65	0.999

Table 6
Allelic diversity of the 24 MIRU-VNTRs loci in 306 drug-resistant *M. tuberculosis* isolates from Sichuan province, China.

VNTR locus	VNTR alias	Number of alleles	Range of repeats	Allelic diversity (h) for		
				Beijing isolates (n = 212)	Non-Beijing isolates (n = 94)	All isolates (n = 306)
154	MIRU02	3	1–3	0	0.052	0.016
580	MIRU04	7	1–6,8	0.179	0.629	0.356
960	MIRU10	6	0–5	0.588	0.582	0.628
1644	MIRU16	6	0–5	0.662	0.67	0.669
2059	MIRU20	3	1–3	0.023	0.092	0.048
2531	MIRU23	7	1,3–8	0.178	0.169	0.178
2687	MIRU24	1	1	0	0	0
2996	MIRU26	10	1–10	0.72	0.728	0.79
3007	MIRU27	5	0–4	0.596	0.648	0.622
3192	MIRU31	6	1–6	0.561	0.669	0.671
4348	MIRU39	7	0–6	0.728	0.682	0.749
802	MIRU40	5	1–5	0.529	0.712	0.63
2165	ETR-A	5	1–5	0.293	0.662	0.503
2461	ETR-B	3	1–3	0.023	0.472	0.226
577	ETR-C	5	2–6	0.104	0.189	0.128
424	Mtub04	8	0–6,8	0.725	0.694	0.75
1955	Mtub21	9	1–9	0.709	0.719	0.805
2347	Mtub29	5	2–6	0.069	0.093	0.079
2401	Mtub30	5	2–6	0.138	0.245	0.437
3171	Mtub34	4	1–4	0.111	0.151	0.127
3690	Mtub39	6	1–6	0.608	0.776	0.691
2163b	QUB-11b	14	1–10,19–22	0.775	0.837	0.821
4052	QUB-26	11	1–11	0.745	0.814	0.783
4156	QUB-4156	8	0–7	0.611	0.489	0.581

account for the high level of prevalence of MDR-TB in addition to the Beijing family in Sichuan.

MIRU-VNTR genotyping has been recently reported to be a rapid and powerful tool in discriminating *M. tuberculosis* isolates in different settings, including Bulgaria, Russia, South Korea and Beijing (Jiao et al., 2008; Mokrousov et al., 2008; Shamputa et al., 2010; Valcheva et al., 2008). Because of the genetic homogeneity of the Beijing family, spoligotyping may poorly differentiate Beijing isolates (van Soolingen et al., 1995) or isolates from a region where Beijing family is predominant (Lv et al., 2009). In present study, we used 24-locus MIRU-VNTR set combined with spoligotyping to genotype all isolates investigated. In this study, an unexpectedly high resolution (HGI = 0.997) was obtained even for the original set of 12-locus MIRU analysis, although 12-locus MIRU was found to be not discriminatory enough in previous studies (Kam et al., 2005; Maes et al., 2008). When 24-locus MIRU-VNTR was fully applied, it presented an excellent discriminatory power (HGI = 0.999) in differentiating all isolates by reducing clustering rate from 18.63% to 1.96%. When different sets of MIRU-VNTR were used in combination with spoligotyping, clustering rate was reduced too. However, combination with spoligotyping only reduced clustering rate slightly for 12-MIRU, 15-VNTR and 24-VNTR, respectively, indicating that spoligotyping did not contribute much to differentiate all isolates in this study.

Taking into account that the collection of our samples was from Sichuan, one of the largest provinces in China of a land area near half million square kilometers and population more than 80 million, while our collection showed a high clonal stability of MIRU-VNTR used to identify the epidemiological links among Sichuan patients, the same MIRU-VNTR patterns of isolates (clustered isolates) from different patients would be epidemiologically regarded as the consequences of transmission among patients. In our study, among 12 clustered isolates identified by 24-locus MIRU-VNTR, 10 isolates in five clusters belonged to Beijing family, while 2 belonged to Beijing and Manu2 family, respectively (Supplementary Fig. 1). Among clustered cases, two isolates (SC241 and SC245) with identical patterns of 24-locus MIRU-VNTR, had same profiles of drug resistance, however, only 2 isolates (SC244 and SC248) with fully identical 24-locus MIRU-VNTR patterns and spoligotypes had same profiles of drug resistance. Theoretically, if the isolates within one cluster (with the same genotype) have the same profile of drug resistance, high level of prevalence of drug-resistant TB should be contributed by the transmission of drug-resistance. On the contrary, prevalence should be attributed by obtained resistance or inappropriate used of anti-TB drugs. Therefore, the small clustering size observed from different sets of MIRU-VNTR applied alone or in combination with spoligotyping in this study suggested that the acquired resistance or inappropriate use of anti-TB drugs, rather than the transmission of drug-resistance, might account for the high level of prevalence of drug-resistant TB in Sichuan.

In conclusion, this investigation on the genetic diversity of drug-resistant *M. tuberculosis* isolates in Sichuan showed that the Beijing family isolates are predominant in our setting. In addition, other genotypic lineages such as T family, Manu2, H3, LAM9 and other minor unassigned clades are responsible for high prevalence of MDR-TB, in addition to Beijing family. The present study also showed that 24-locus MIRU-VNTR could be used as a powerful tool for differentiating these drug-resistant isolates. Last but not least, our data suggest that the bulk of drug-resistant TB in Sichuan might be primarily due to the acquired resistance rather than the transmission of drug-resistant TB.

Acknowledgements

The authors are grateful to the financial support by National Nature Science Foundation of China (30972524), and the Department of Science and Technology of Sichuan Province of China (2009SZ0015). We also thank Philip Supply for kindly providing the VNTR manual and the correspondence tables. D.C. and N.R. acknowledge help of their colleagues Véronique Hill and Thierry Zozio (Institut Pasteur de la Guadeloupe) for helping with the SITVIT2 database.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2011.09.022.

References

Allix-Béguec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S., 2008. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J. Clin. Microbiol. 46, 2692–2699.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Canetti, G., Fox, W., Khomenko, A., Mahler, H.T., Menon, N.K., Mitchison, D.A., Rist, N., Smelev, N.A., 1969. Advances in techniques of testing mycobacterial drug sensitivity, and the use of sensitivity tests in tuberculosis control programmes. Bull. World Health Organ. 41, 21–43.

Centers for Disease Control, Prevention (CDC), 2006. Emergence of *Mycobacterium tuberculosis* with extensive resistance to second-line drugs-worldwide, 2000–2004. MMWR Morb. Mortal. Wkly. Rep. 55, 301–305.

Dong, H.Y., Liu, Z.G., Lv, B., Zhang, Y.Y., Liu, J., Zhao, X.Q., Liu, J.H., Wan, K.L., 2010. Spoligotypes of *Mycobacterium tuberculosis* from different provinces of China. J. Clin. Microbiol. 48, 4102–4106.

Duanmu, H.J., 2002. Report on fourth national epidemiological sampling survey of tuberculosis. Chin. J. Tubercu. Respire. Dis. 25, 3–7.

Hunter, P.R., Gaston, M.A., 1998. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 26, 2465–2466.

Iwamoto, T., Yoshida, S., Suzuki, K., Tomita, M., Fujiyama, R., Tanaka, N., Kawakami, Y., Ito, M., 2007. Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on Mycobacterium tuberculosis strains predominated by the Beijing family. FEMS Microbiol. Lett. 272, 282–283.

Jiao, W.W., Mokrousov, I., Sun, G.Z., Guo, Y.J., Vyazovaya, A., Narvskaya, O., Shen, A.D., 2008. Evaluation of new variable-number tandem-repeat systems for typing *Mycobacterium tuberculosis* with Beijing genotype isolates from Beijing. China J. Clin. Microbiol. 46, 1045–1049.

Kam, K.M., Yip, C.W., Tse, L.W., Wong, K.L., Lam, T.K., Kremer, K., Au, B.K., van Soolingen, D., 2005. Utility of mycobacterial interspersed repetitive unit typing for differentiating multidrug-resistant *Mycobacterium tuberculosis* isolates of the Beijing family. J. Clin. Microbiol. 43, 306–313.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., van Embden, J., 1997. Rapid detection and simultaneous strain differentiation of *Mycobacterium tuberculosis* for diagnosis and tuberculosis control. J. Clin. Microbiol. 35, 907–914.

Kim, S.J., Bai, G.H., Lee, H., Kim, H.J., Lew, W.J., Park, Y.K., Kim, K., 2001. Transmission of *Mycobacterium tuberculosis* among high school students in Korea. Int. J. Tuberc. Lung Dis. 5, 824–830.

Li, X., Xu, P., Shen, X., Qi, L.H., DeRiemer, K., Mei, J., Gao, Q., 2011. Non-Beijing strains of *Mycobacterium tuberculosis* in China. J. Clin. Microbiol. 49, 392–395.

Liu, J., Jiang, S., Cheng, S., 2003. Analysis of current status of tuberculosis control and its strategy in China. J. Chinese Antituberc. Assoc. 25, 129–131.

Liu, Y., Duan, W., Xiong, G., Wu, J., Cao, H., Xia, L., 2002. The dynamic change of drug resistance of tuberculosis in Sichuan Province. J. Prev. Med. Inf. 18, 9–10.

Liu, Z.G., Zhang, X.M., Zhang, Y.Y., Zeng, L.C., Lu, B., Wang, X.L., Zhao, X.Q., Du, X.L., Han, Y.L., Wan, K.L., 2008. The status of drug-resistance on *Mycobacterium tuberculosis* isolated from Xi'an City and its correlation with the genotyping of the Beijing family. Chin. J. Zoonoses 24, 435–438.

Lv, B., Liu, M., Li, Z.N., Liu, Z.G., Zhao, X.Q., Wan, K.L., 2009. Analysis on genotyping of 159 strains of *Mycobacterium tuberculosis* isolated clinically in someareas of China with MLVA-19 Loci. Dis. Surveill. 24, 359–362 (in Chinese).

Maes, M., Kremer, K., Van Soolingen, D., Takiff, H., de Waard, J.H., 2008. 24-Locus MIRU-VNTR genotyping is a useful tool to study the molecular epidemiology of tuberculosis among Warao Amerindians in Venezuela. Tuberculosis 88, 490–494.

Mazars, E., Lesjean, S., Banuls, A.L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C., Supply, P., 2001. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. Proc. Natl. Acad. Sci. USA 98, 1901–1906.

Mokrousov, I., 2008. Genetic geography of Mycobacterium tuberculosis Beijing genotype: a multifacet mirror of human history? Infect Genet Evol. 8, 777–785.

Mokrousov, I., Narvskaya, O., Vyazovaya, A., Millet, J., Otten, T., Vishnevsky, B., Rastogi, N., 2008. *Mycobacterium tuberculosis* Beijing genotype in Russia: in search of informative variable-number tandem-repeat loci. J. Clin. Microbiol. 46, 3576–3584.

National Technical Steering Group of the Epidemiological Sampling Survey for Tuberculosis (NTSG), 2002. Report on nationwide random survey for the epidemiology of tuberculosis in 2000. J. Chin. Antituberc. Assoc. 24, 65–108.

Niemann, S., Köser, C.U., Gagneux, S., Plinke, C., Homolka, S., Bignell, H., Carter, R.J., Cheetham, R.K., Cox, A., Gormley, N.A., Kokko-Gonzales, P., Murray, L.J., Rigatti, R., Smith, V.P., Arends, F.P., Cox, H.S., Smith, G., Archer, J.A., 2009. Genomic Diversity among Drug Sensitive and Multidrug Resistant Isolates of *Mycobacterium tuberculosis* with Identical DNA Fingerprints. PLoS ONE 4, e7407.

Oelemann, M.C., Diel, R., Vatin, V., Haas, W., Rüsch-Gerdes, S., Locht, C., Niemann, S., Supply, P., 2007. Assessment of an optimized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J Clin. Microbiol. 45, 691–697.

Park, Y.K., Bai, G.H., Kim, S.J., 2000. Restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolated from countries in the Western Pacific region. J. Clin. Microbiol. 38, 191–197.

Parwati, I., van Crevel, R., van Soolingen, D., 2010. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect. Dis. 10, 103–111.

Phyu, S., Stavrum, R., Lwin, T., Svendsen, Ø.S., Ti, Ti., Grewal, H.M.S., 2009. Predominance of *Mycobacterium tuberculosis* EAI and Beijing lineages in yangon. Myanmar. J. Clin. Microbiol. 47, 335–344.

Shamputa, I.C., Lee, J., Allix-Béguec, C., Cho, E.J., Lee, J.I., Rajan, V., Lee, E.G., Min, J.H., Carroll, M.W., Goldfeder, L.C., Kim, J.H., Kang, H.S., Hwang, S., Eum, S.Y., Lee, H., Park, S.K., Supply, P., Cho, S.N., Via, L.E., Barry, C.E., 2010. Genetic diversity of *Mycobacterium tuberculosis* isolates from a tertiary care tuberculosis hospital in South Korea tertiary care tuberculosis hospital in South Korea. J. Clin. Microbiol. 48, 387–394.

Shi, L., Yang, M., Pourcel, C., Dong, H.Y., Nimapengduo., Zhang, Y.Y., Lang, Z., Lv, B., Jiang, Y., Wei, S.Z., Fan, X.J., Wan, K.L., 2007. Genotyping study of 216 *Mycobacterium tuberculosis* strains isolated from the patients in tibet with MLVA and spoligotyping. Chin. J. Microbiol. Immunol. 27, 711–718.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unitvariable- number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., Locht, C., 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. 36, 762–771.

Tong, C.X., Zhao, X.Q., Ma, J.J., Liu, Z.G., Cao, W.J., Yang, Y.H., Lu, B., Jiang, Y., Wan, K.L., 2010. Analysis on the 228 *Mycobacterium tuberculosis* with MLVA. China Med. Herald. 7, 97–98.

Valcheva, V., Mokrousov, I., Narvskaya, O., Rastogi, N., Markova, N., 2008. Utility of new 24-locus variable-number tandem-repeat typing for discriminating *Mycobacterium tuberculosis* clinical isolates collected in Bulgaria. J. Clin. Microbiol. 46, 3005–3011.

van Soolingen, D., Qian, L., de Haas, P.E., Douglas, J.T., Traore, H., Portaels, F., Qing, H.Z., Enkhsaikan, D., Nymadawa, P., van Embden, J.D., 1995. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. J. Clin. Microbiol. 33, 3234–3238.

van Soolingen, D., Hermans, P.W., de Haas, P.E., SOLL, D.R., van Embden, J.D.A., 1991. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J. Clin. Microbiol. 29, 2578–2586.

Wada, T., Iwamoto, T., Maeda, S., 2009. Genetic diversity of the *Mycobacterium tuberculosis* Beijing family in East Asia revealed through refined population structure analysis. FEMS. Microbiol. Lett. 291, 35–43.

Weniger, T., Krawczyk, J., Supply, P., Niemann, S., Harmsen, D., 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res. 38 (Suppl.), W326–W331.

World Health Organisation, 2010. Global tuberculosis control: WHO report 2010. World Health Organization, Geneva, Switzerland.

World Health Organization and International Union against Tuberculosis and Lung Disease, 1997. Guidelines for surveillance of drug resistance in tuberculosis. WHO/TB96.216. World Health Organization, Geneva, Switzerland.

Yang, X.Y., Zhang, N.M., Diao, X., Mao, X., Li, Y.P., 2008. Epidemiological analysis of pulmonary tuberculosis in Sichuan Province, China, 2000–2006. Int. J. Infect. Dis. 12, 534–554.

OPEN ACCESS Freely available online

PLOS ONE

Mycobacterium tuberculosis Strains Potentially Involved in the TB Epidemic in Sweden a Century Ago

Ramona Groenheit[1]*, Solomon Ghebremichael[1], Alexandra Pennhag[1], Jerker Jonsson[1], Sven Hoffner[1], David Couvin[2], Tuija Koivula[1,3], Nalin Rastogi[2], Gunilla Källenius[3]

1 Department of Preparedness, Swedish Institute for Communicable Disease Control, Solna, Sweden, **2** Tuberculosis and Mycobacteria Unit, WHO Supranational TB Reference Laboratory, Institut Pasteur de la Guadeloupe, Guadeloupe, France, **3** Department of Clinical Science and Education, Karolinska Institutet, Södersjukhuset, Stockholm, Sweden

Abstract

A hundred years ago the prevalence of tuberculosis (TB) in Sweden was one of the highest in the world. In this study we conducted a population-based search for distinct strains of *Mycobacterium tuberculosis* complex isolated from patients born in Sweden before 1945. Many of these isolates represent the *M. tuberculosis* complex population that fueled the TB epidemic in Sweden during the first half of the 20th century.

Methods: Genetic relationships between strains that caused the epidemic and present day strains were studied by spoligotyping and restriction fragment length polymorphism.

Results: The majority of the isolates from the elderly population were evolutionary recent Principal Genetic Group (PGG)2/3 strains (363/409 or 88.8%), and only a low proportion were ancient PGG1 strains (24/409 or 5.9%). Twenty-two were undefined. The isolates demonstrated a population where the Euro-American superlineage dominated; in particular with Haarlem (41.1%) and T (37.7%) spoligotypes and only 21.2% belonged to other spoligotype families. Isolates from the elderly population clustered much less frequently than did isolates from a young control group population.

Conclusions: A closely knit pool of PGG2/3 strains restricted to Sweden and its immediate neighbours appears to have played a role in the epidemic, while PGG1 strains are usually linked to migrants in todays Sweden. Further studies of these outbreak strains may give indications of why the epidemic waned.

Citation: Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S, et al. (2012) *Mycobacterium tuberculosis* Strains Potentially Involved in the TB Epidemic in Sweden a Century Ago. PLoS ONE 7(10): e46848. doi:10.1371/journal.pone.0046848

Editor: Philip Supply, Institut Pasteur de Lille, France

Received November 21, 2011; **Accepted** September 10, 2012; **Published** October 8, 2012

Copyright: © 2012 Groenheit et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was financed by grants to Gunilla Källenius from Konung Gustaf V:s och Drottning Victorias Frimurarestiftelse, the Swedish Heart-Lung Foundation (http://www.hjart-lungfonden.se) and the Swedish Vetenskapsrådet (http://www.vr.se). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ramona.groenheit@smi.se

Introduction

Tuberculosis (TB) is globally a major cause of morbidity and mortality, with a majority of cases occurring in low income countries. As estimated by the WHO currently one third of the worlds population is infected with bacteria of the *Mycobacterium tuberculosis* complex, and ten million cases of active TB disease occur each year, resulting in almost two million deaths annually. The increasing spread of TB has also been paralleled by a rapid increase in multi-drug resistant TB in many parts of the world, making the disease in several instances practically incurable.

Distinct *M. tuberculosis* strains have been associated with large outbreaks of TB. Also in the Nordic countries there are outbreaks of specific strains of *M. tuberculosis*. During the last decades, a specific strain of *M. tuberculosis* has emerged rapidly in Denmark [1], another outbreak has been recorded in Norway [2] and one of the largest outbreaks ever recorded in a low endemic country is ongoing in Sweden [3,4,5]. Little is known about the *M. tuberculosis* population that dominated Sweden a hundred years ago. It does however appear that this bacterial population has been successfully reduced from representing the major public health problem to its current level of near elimination.

Today the Nordic countries are high-income nations with low prevalence of TB. In 2011, the TB incidence in Sweden was 6.3/100,000 population and only 11% of the TB patients were born in Sweden. The number of cases is almost entirely dependent on migration from countries with high TB incidence and the group of elderly Swedish born patients with reactivated TB infection is decreasing [6]. However, less than a century ago the prevalence of TB in the Nordic countries (Denmark, Finland, Norway and Sweden) was among the highest in the world. In 1905 the overall estimated TB incidence in Sweden was 890/100,000 population [7] higher than in most high incidence countries of sub-Saharan Africa today.

Ten percent of otherwise healthy persons infected with *M. tuberculosis* are estimated to progress to active disease, with the highest risk during the first two years after infection. A combination of bacterial and patient factors influence the risk for induction of active disease in patients with latent infection, and many who were infected may develop TB after decades of

infection [8]. In the elderly population in Sweden many are still latently infected with TB, and some develop active TB. In the cohorts born before 1945 most subjects presumably have latent TB infection (LTBI). Using a TB incidence among newborns of <1/100,000 population as an indicator of interrupted TB transmission must be interpreted with caution but by doing so one can estimate that since around 1967 [9,10] most active TB cases occurring in the elderly Swedish-born population can be seen as reactivation of LTBI.

In the past decades our understanding of the molecular genetics of *M. tuberculosis* has further expanded. One of the major achievements using DNA fingerprinting techniques has been the implementation of population based transmission surveillance. Geographically defined lineages of *M. tuberculosis* have been identified [11] and specific genetically highly conserved groups of strains of *M. tuberculosis* have attracted special attention.

Worldwide, few collections of isolates from the 20th century are available. At the Swedish Institute for Communicable Disease Control (SMI) clinical *M. tuberculosis* isolates have been preserved since the 1980s. This consequently provides us with an exceptional possibility to perform population-based studies of the transmission of TB. In this study we conduct the first systematic population-based search for distinct strains of *M. tuberculosis* isolated from elderly patients born in Sweden before 1945. The isolates represent strains most likely acquired in Sweden during the last 60–100 years, and many of these isolates may represent the *M. tuberculosis* population that fueled the TB epidemic in Sweden during the first half of the 20th century.

Materials and Methods

Ethics Statement

At the SMI, clinical *M. tuberculosis* complex strains are routinely collected for disease surveillance. The current study describes a bacterial collection and the bacterial genotypes could only be combined with the sex, age, and country of birth for the patients from which the strains were isolated. Ethical approval was therefore not required.

Bacterial Isolates

M. tuberculosis complex isolates obtained from all six Swedish TB laboratories in Gothenburg, Linköping, Malmö/Lund, Stockholm and Umeå during the years 1994–2009 were studied. They represent all strains from patients born in Sweden before 1945 that have been preserved at SMI during the sampling period and that did not cluster with any patients born after 1945 and/or were foreign born. During the same sampling period all isolates that had been preserved at SMI from patients born in Sweden in 1985 or later were also analysed as a control group. The patients were identified as born in Sweden through the national TB Register. The isolates had been stored at −70C°.

Drug Susceptibility Testing

In Sweden, all isolates are tested for susceptibility to the first-line drugs isoniazid (INH), rifampicin (RIF), ethambutol (EMB) and pyrazinamide (PZA) using the BACTEC 460TB or the MGIT 960 liquid culture and drug susceptibility testing systems according to the instructions of the manufacturer. During the major part of the study all isolates were also tested for susceptibility to streptomycin (SM), except for the years 2004–2009, when the Linköping and Stockholm laboratories stopped testing for SM-resistance, since SM no longer was used for treatment of TB patients in Sweden. All laboratories had taken part in the external quality assurance program for drug susceptibility testing of *M. tuberculosis* offered by the Swedish TB reference laboratory at SMI.

Spoligotyping

All isolates were characterized by spoligotyping, which characterizes the polymorphic direct repeat region of the *M. tuberculosis* chromosome [12]. The patterns obtained by spoligotyping were compared by visual examination and computer assisted analyses by use of the BioNumerics version 6.6 software (Applied Maths, Kortrijk, Belgium). The spoligotypes were also compared with those contained in the international database SITVIT2, an updated version of the previously published SpolDB4 database [11] (http://www.pasteur-guadeloupe.fr:8081/SITVITDemo) which defines 62 spoligotype families/subfamilies of *M. tuberculosis* complex isolates. The SITVIT2 database contains to date genotyping data on more than 86,000 clinical isolates from 160 countries of origin, with more than 3,000 spoligo-international-types (SITs; a pattern shared by two or more patient isolates). The BioNumerics software version 3.5 was used to build spoligotyping-based minimum spanning trees (MST). MST is an undirected network in which all of the samples are linked together with the fewest possible linkages between nearest neighbors. Using this approach, one considers that all intermediate stages are present within the sample analyzed by first including the individual that shows the greatest number of possible linkages to other individuals in the population studied. We used this method to highlight the links between the spoligotype families differing by changes observed in their direct variable repeats. To evaluate the distribution of orphan strains in order to get information on their specific evolution two trees were generated - one for SITs, and another for all patterns, i.e. SITs and orphans pooled together. We also created spoligoforest trees [13] to illustrate probable strain evolutionary relationships between spoligotypes. Contrarily to the MSTs the spoligoforest trees are directed and only evolve by loss of spacers. Lastly, the major *M. tuberculosis* genotypic families were also linked to "ancient" and "modern" lineages of tubercle bacilli as defined by Principal Genetic Groups (PGG) defined by *katG463-gyrA95* polymorphism [14], and inferred from the reported linking of specific spoligotype patterns to PGG1, 2 or 3 grouping [15,16,17,18,19]. For statistical analyses the isolates were categorised into three groups, modern [consisting of the Haarlem (H), T, Latin-American and Mediterranean (LAM), X and S families], ancient [consisting of Beijing, East-African-Indian (EAI) and Central-Asian (CAS) families] and [*M. bovis*/*M. bovis* like and Manu].

IS*6110* RFLP

The isolates were cultured on Löwenstein Jensen medium, DNA was extracted and RFLP typing was performed using the insertion sequence IS*6110* as a probe and *Pvu*II as the restriction enzyme [20]. Visual bands were analyzed using the BioNumerics version 6.6 software. On the basis of the molecular sizes of the hybridizing fragments and the number of IS*6110* copies of each isolate, fingerprint patterns were compared by the un-weighted pair-group method of arithmetic averaging using the Jaccard coefficient. Dendrograms were constructed to show the degree of relatedness among strains according to a previously described algorithm [21] and similarity matrixes were generated to visualize the relatedness between the banding patterns of all isolates. The RFLP patterns were entered into the RFLP database at SMI, which at the time of this study contained 3951 isolates that had been isolated in Sweden. Strains with identical RFLP-patterns (100% similarity) were judged to belong to a cluster.

Statistical Method

Mean and standard deviation (SD) were calculated for age of the patient at diagnosis and year of birth by isolate group. Age and year of birth were not normally distributed and therefore possible differences in these variables for isolate group and gender were investigated with the non-parametric Kruskal-Wallis test. Post-hoc tests for isolate group were performed for the three pairs using the Kruskal-Wallis test (Mann-Whitney U test). The chi^2-test was used to test for association between two categorical variables. The level of significance was set to 0.05 (two-sided) and all analyses were performed using R v 2.9.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

In total, 409 isolates from 242 (59.2%) men and 167 (40.8%) women born in Sweden before 1945 were analysed. These patients were born in Sweden between the years 1908–1945 (Table S1). A total of 9.8% were born before 1914, 19.6% were born in 1915–1919, 27.1% were born in 1920–1924, 17.1% were born in 1925–1929, 11.0% were born in 1930–1934, 7.3% were born in 1935–1939 and 8.1% were born in 1940–1945. At diagnosis the patients were 52–98 years old, with a mean age of 78.1 years. The 58 patients in the young Swedes control group were born between the years 1985–2008 (Tabel S2). A total of 24.1% were born in 1985–1989, 37.9% were born in 1990–1994, 17.2% were born in 1995–1999, 8.6% were born in 2000–2004 and 12.1% were born in 2005–2008. At diagnosis the patients were 0–23 years old, with a mean age of 10.9 years.

Drug Resistance

Among the elderly Swedes, information on drug resistance was obtained for 404/409 isolates. Of those, 38 isolates (9.3%) were resistant to one or more of the drugs SM (n = 3), INH (n = 14), RIF (n = 4), EMB (n = 1) and PZA (n = 22). The large number of isolates resistant to PZA is explained by the inclusion of *M. bovis* isolates in the study. Five isolates were resistant to more than one drug, and of those four isolates were multidrug resistant. Among the 58 young Swedes in the control group, 17 isolates (29.3%) were resistant to one or more of the drugs SM (n = 5), INH (n = 16), RIF (n = 2), EMB (n = 1) and PZA (n = 1). Five isolates were resistant to more than one drug, and of those two were multidrug resistant. As some laboratories during the study period stopped testing for SM resistance, only 207/409 and 17/58 of the isolates were analysed for SM resistance.

Spoligotyping

Out of 409 isolates, 173 different spoligo patterns were obtained, of which 277 (67.7%) were clustered in 41 spoligo clusters comprising 2–56 strains per cluster. The remaining 132 (32.3%) spoligo patterns were unique i.e. the isolates did not cluster with other patient isolates. When compared with SITVIT2, the majority, 364 clinical isolates, were shared-types (Table 1), i.e. had an identical pattern shared by two or more isolates worldwide (within this study, or matching another strain in the SITVIT2 database). A SIT number was attributed to each pattern according to the SITVIT2 database. Forty-five patterns corresponded to orphan strains that were unique among the 86,000 strains recorded in the SITVIT2 database (Table 2). The isolates demonstrated a highly homogenous population where the modern H and T clades dominated. The absolute majority (n = 363, 88.8%) were evolutionary recent PGG2/3 strains, including H (n = 168, 41.1%), T (n = 154, 37.7%), LAM (n = 32), S (n = 8) and X (n = 1) isolates (Table 1 and S3). Only 24 (5.9%) were evolutionary ancient PGG1 strains (3 Beijing, 3 CAS1-Delhi, 4 EAI, 3 Manu and 11 *M. bovis/M. bovis* like isolates) (Table 1 and S4). Twenty-two spoligotyping signatures that are not yet associated to a well-defined spoligotype familiy in SITVIT2 were designated as "Unknown" (Table 1 and 2). The most common spoligotypes were SIT50 (n = 56, 13.7%) of the H3 subfamily, SIT53 (n = 43, 10.5%) of the T1 subfamily, SIT47 (n = 33, 8.1%) of the H1 subfamily and SIT42 (n = 14, 3.4%) of the LAM9 subfamily (Table S3). In addition to the T1 subfamily prototype, SIT53, two more T clade SITs, SIT153 (n = 11, 2.7%) and SIT37 (n = 7, 1.7%) were among the seven predominant SITs. Four of the *M. bovis* isolates (SIT691) lacked spacer 11, in addition to spacers 3, 9, 16, and 39 to 43. Of the 45 orphan strains, 22 were of the T spoligotype family, 17 of the H family, 2 *M. bovis/M. bovis* like, 1 LAM family and 3 of unknown family. Significant for all except the two *M. bovis* strains was that they all lacked spacers 33–36 (signature of SIT53).

The 168 patients with H family isolates were born between 1908–1945 (median 1924), with a mean age of 77.8 years at diagnosis (range 52–98), the 154 patients with T family isolates were born between 1910–1945 (median 1923), with a mean age of 78.5 years at diagnosis (range 53–97), and the 32 patients with LAM family isolates were born between 1913–1943 (median 1923), with a mean age of 78.9 years at diagnosis (range 63–93). The three CAS1-Delhi strains were isolated from patients born in 1921, 1938, and 1943 with a mean age of 67.7 years (57, 71, and 75 years). The four patients with EAI isolates were born later than the patients with H, T and LAM isolates. They were all except one born in the 1940s: three EAI2-Manilla isolates from patients born 1911, 1943 and 1943 and one EAI5 isolate from a patient born 1943. The patients were also younger (mean age 68.0) than average at diagnosis (57, 62, 66 and 87 years). The three patients with Beijing isolates were all diagnosed between 2007 and 2008, i.e. in the later part of the study, at a mean age of 80.7 years. We tested the hypothesis that the 363 "modern" isolates of the H, T, LAM, X and S families differed in patient characteristics compared to the 24 "ancient" isolates split into two groups [group one (n = 10): Beijing, EAI and CAS and group two (n = 14): *M. bovis/M. bovis* like and Manu] with regard to the age of the patients at diagnosis and date of birth. Patients with "modern" isolates of the H, T, LAM, X and S families were significantly older at diagnosis and were born significantly earlier than patients with "ancient" isolates.

For phylogenetical analyses, we drew two separate MSTs (Figure 1) to summarize the possible evolutionary relationships between all the genotypes obtained. Figure 1A is based only on SITs, while Figure 1B shows combined data for both SITs and orphan patterns pooled together. The SIT-based MST (Figure 1A) shows a tree split into distinct families: the top section displays the ancient PGG1 strains (EAI, Bovis, Manu and CAS1-Delhi families) whereas the bottom shows the evolutionary modern PGG2/3 strains belonging to the H, T and LAM families. As summarized in Table 2, it should be noticed that only 2/47 orphan strains were PGG1 (1 BOV-1, and 1 BOV_LIKE), the rest being evolutionary modern PGG2/3. The spoligoforests generated (see Figures S1 and S2, and legends for detailed comments) highlighted the predominance of PGG2/3 group which are well represented (as large, visible nodes). The rare ancestral PGG1 strains (belonging to the EAI, Manu, and CAS families) were mostly located in the top layer of the hierarchical layout as isolated strains without interconnections with others (Figure S1).

Interestingly, the MST shown after combining orphans with SITs in Figure 1B is overlapping with the tree shown in Figure 1A in the sense that the two ancestral PGG1 orphans were grouped

Table 1. Description of 127 shared-types (SITs; n = 364 isolates, 2–56 isolates per cluster) and corresponding spoligotyping families/subfamilies starting from a total of 409 *M. tuberculosis* complex strains isolated from Swedish patients born before 1945.

SIT*	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. Database	Spoligotype family**	Clustered vs. unique patterns***
1		000000000003771	3 (0,73)	0,03	Beijing	Clustered
2		000000004020771	1 (0,24)	0,25	H2	Unique
4		000000007760771	1 (0,24)	0,29	Unknown	Unique
19		677777477413771	3 (0,73)	0,35	EAI2-Manilla	Clustered
20		677777607760771	4 (0,98)	0,49	LAM1	Clustered
26		703777740003771	1 (0,24)	0,08	CAS1-Delhi	Unique
29		760001400000171	1 (0,24)	2,44	LAM	Unique
32		776000000000171	2 (0,49)	1,92	Unknown	Clustered
34		776377777760771	6 (1,47)	0,73	S	Clustered
36		777737777720771	3 (0,73)	2,34	H3	Clustered
37		777737777760771	7 (1,71)	1,55	T3	Clustered
39		777777347760471	1 (0,24)	0,71	T4-CEU1	Unique
40		777777377760771	1 (0,24)	0,65	T4	Unique
42		777777607760771	14 (3,42)	0,44	LAM9	Clustered
44		777777757760771	1 (0,24)	0,48	T5	Unique
45		777777764020771	2 (0,49)	1,89	H1	Clustered
47		777777774020771	33 (8,07)	2,26	H1	Clustered
49		777777777720731	12 (2,93)	6,98	H3	Clustered
50		777777777720771	56 (13,69)	1,71	H3	Clustered
51		777777777760700	1 (0,24)	0,36	T1	Unique
52		777777777760731	3 (0,73)	0,33	T2	Clustered
53		777777777760771	43 (10,51)	0,72	T1	Clustered
54		777777777763771	1 (0,24)	0,45	MANU2	Unique
62		777777774020731	4 (0,98)	0,73	H1	Clustered
64		777777607560771	2 (0,49)	0,56	LAM6	Clustered
65		777777777760471	1 (0,24)	0,95	T1	Unique
78		777777777760711	1 (0,24)	1,56	T	Unique
97		777777760760771	1 (0,24)	14,29	T1	Unique
99		757777777720771	1 (0,24)	1,33	H3	Unique
100		777777777773771	1 (0,24)	1,32	MANU1	Unique
123		777777776360771	1 (0,24)	4,17	T1	Unique
134		777777777720631	1 (0,24)	1,92	H3	Unique
151		777767774020771	1 (0,24)	4	H1	Unique
153		757777777760731	11 (2,69)	14,86	T2	Clustered

Table 1. Cont.

SIT*	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. Database	Spoli-gotype family**	Clustered vs. unique patterns***
167	■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■□□□□■■■■■■■	777777777660771	3 (0,73)	4,17	T1	Clustered
172	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□■■■■■■■	777777777740771	1 (0,24)	1,82	T1	Unique
183	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■□■□□□□■■■■■■■	777777377720771	1 (0,24)	1,85	H3	Unique
191	□□■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	177777777760771	1 (0,24)	5,56	T1	Unique
205	■■■□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	737777777760771	1 (0,24)	1,92	T1	Unique
209	■■■■■■■□□□□□□■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	776017607760771	1 (0,24)	2,44	LAM12-Madrid1	Unique
218	■■■■■■■■■■□■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	777737774020771	3 (0,73)	8,33	H1	Clustered
241	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□□□□■■	777777777760411	5 (1,22)	21,74	T1	Clustered
254	■■■■■■■■■■■■□□□□□□□□□□■■■■■■■■■□□□□■■■■■■■	777760007760771	1 (0,24)	0,6	T5-RUS1	Unique
262	■■■■■■□□■■■■■■■■■■■■■■■□□□■□□□□■■■■■■■	774777777420771	1 (0,24)	0,57	H3	Unique
276	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■■□□□□■■■■■■■	777777770060771	1 (0,24)	2,78	T1	Unique
280	■■■■■□□□□□□□□□□□■■■■■■■■■■■■■■□□□□■■■■■■■	770000777760771	2 (0,49)	2,94	T1-RUS2	Clustered
283	■■■■■■■■■■■■■■■■■■■■■□□□■□□□□□□■□□□□■■■■■■■	777777704020771	1 (0,24)	1,64	H1	Unique
353	■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■□□□□■■■■■■■	777777774760771	2 (0,49)	7,41	T1	Clustered
356	■■■□□□□■■■■■■■■■■■■■□□□□□□□□□□□□□□■■■■■■■	703777600001771	1 (0,24)	7,14	CAS1-Delhi	Unique
373	■■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■	777777767760771	1 (0,24)	1,52	T1	Unique
383	■■■■■■■□□□■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	776377774020771	1 (0,24)	7,14	H1	Unique
393	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777757777760771	2 (0,49)	5,56	T1	Clustered
452	■■■■■■■■■■■■■■■■■■■■■□□□□■□□■■■□□□□■■■■■■■	777777606360771	1 (0,24)	6,25	LAM9	Unique
453	■■■■■■■■■■■■■■■■■■■■■□■■■■■■■□■■■□□□□■■■■■■■	777777677560771	1 (0,24)	6,25	T1	Unique
463	■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■□■■■■■■■	777777777720571	2 (0,49)	11,76	H3	Clustered
485	■■■□□□□■■■■■■■■■■■■■□□□□□□□□□□□□□□□■■■■■■■	703777400003771	1 (0,24)	4,35	CAS1-Delhi	Unique
511	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□■□□□□■■■■■■■	777777700020771	4 (0,98)	9,09	H3	Clustered
515	■■■■■■■■■■■■■■■■■■■■■■■□■■■■■□□□□■■■□■■■	777777776760731	1 (0,24)	7,14	T2	Unique
523	■■■	777777777777771	1 (0,24)	2,27	MANU_ancestor	Unique
524	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■□□■■	777777777720711	1 (0,24)	8,33	H3	Unique
533	■■■■■■■■■□□□□□□□□■■■■■■■■■■■■□■□□□□■■■■■■■	777400777720771	1 (0,24)	33,33	H3	Unique
602	■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□■■■■■■■	777777770000771	2 (0,49)	1,96	Unknown	Clustered
612	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■	777777777760751	2 (0,49)	7,41	T1	Clustered
620	■■■■■■■■■■■■■■■■■■■■■■□□■□□□□□□■□□□□■■■■■■■	777777744020771	1 (0,24)	7,14	H1	Unique
628	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■□□□	777777777760760	1 (0,24)	6,67	T1	Unique
649	□■■■■■■■■■■■■■■■■■■■■■□□□■□□■■□■■□□□□■■■■■■■	377777423560771	1 (0,24)	5,26	T1	Unique
691	■■□■■■■■□■□■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□□□	676573777777600	4 (0,98)	0,55	BOV_1	Clustered
732	■■■■■■■■■■■■■■■□□■■■■■■■■■■■■■■■■□□□□■■■■■■■	777763777760771	1 (0,24)	4,17	T1	Unique

Table 1. Cont.

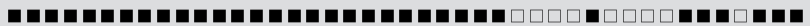
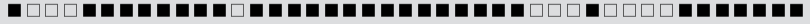
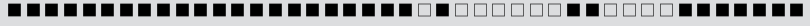
SIT*	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. Database	Spoli-gotype family**	Clustered vs. unique patterns***
748		777777777720760	1 (0,24)	20	H3	Unique
766		777761007760771	2 (0,49)	4,17	LAM9	Clustered
781		000000004020731	1 (0,24)	7,69	H2	Unique
801		774777777760771	1 (0,24)	11,11	T1	Unique
837		777777777020731	1 (0,24)	33,33	H	Unique
848		737777777760731	4 (0,98)	13,79	T2	Clustered
888		777777777760631	1 (0,24)	6,25	T1	Unique
913		777743777760771	1 (0,24)	5	T1	Unique
917		777577777760771	1 (0,24)	10	T1	Unique
918		770000607760771	4 (0,98)	22,22	Unknown	Clustered
921		437737777420771	1 (0,24)	14,29	H4	Unique
926		773777777760771	1 (0,24)	3,13	T1	Unique
928		777777750060771	7 (1,71)	35	T	Clustered
946		777777740020771	1 (0,24)	5	H	Unique
947		777777777413631	1 (0,24)	10	EAI5	Unique
963		777777737720771	1 (0,24)	8,33	H3	Unique
964		437777607760771	4 (0,98)	19,05	LAM1	Clustered
1107		777700007760771	1 (0,24)	20	LAM	Unique
1173		770000777760731	1 (0,24)	14,29	T1-RUS2	Unique
1185		676673777777600	1 (0,24)	33,33	BOV_1	Unique
1211		576377777760771	1 (0,24)	5,56	S	Unique
1560		703777777760771	1 (0,24)	14,29	T1	Unique
1561		777777774020371	1 (0,24)	25	H1	Unique
1611		776777777720771	1 (0,24)	12,5	H3	Unique
1621		737737777760731	1 (0,24)	20	T	Unique
1626		777777776760771	2 (0,49)	16,67	T1	Clustered
1807		777777774020571	1 (0,24)	11,11	H1	Unique
1915		776377737760771	1 (0,24)	25	S	Unique
2201		177777607760771	1 (0,24)	12,5	LAM9	Unique
2644		177777777720771	1 (0,24)	14,29	H3	Unique
2721		777777700000171	1 (0,24)	25	Unknown	Unique
2912		777777350060771	1 (0,24)	33,33	T	Unique
2954		677777600060771	1 (0,24)	33,33	T	Unique
3038		617777777760601	1 (0,24)	33,33	T	Unique

Table 1. Cont.

SIT*	Spoligotype Description	Octal Number	Number (%) in study	% in study vs. Database	Spoligotype family**	Clustered vs. unique patterns***
3091	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□■■■	777777777740031	1 (0,24)	14,29	T	Unique
3139	■□■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	577737777760771	1 (0,24)	20	T3	Unique
3168*	■■■□■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■□■■■	737777777720731	3 (0,73)	75	H3	Clustered
3169*	■□□■■□□□□□□□■■□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■	460060000000771	2 (0,49)	66,67	Unknown	Clustered
3170*	■■□■■■■■□■■■■■■□■■■■■■■■■■■■□■■■■■■■■■■□□□□□	676773776777600	3 (0,73)	75	BOV_1	Clustered
3171*	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□■□□□□■■■■■■■	777777417720771	1 (0,24)	50	H3	Unique
3173*	■■■■■■□□□□□□□□□□□■■■■■■■■■□■■■■□□□□■■■■■■■	770000777360771	2 (0,49)	66,67	T1-RUS2	Clustered
3175*	■■■■■■■□□□□□■□□□□■■■■■■■■■■■■■■■□□□□■■■■■■■	776020777760771	1 (0,24)	50	Unknown	Unique
3176*	■■■■■■■■■□□□■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	777437777720771	1 (0,24)	50	H3	Unique
3177*	■■■■■■□□□□□□□□□□□□■□□□□■■■■■■■□□□□■■■■■■■	770000207760771	1 (0,24)	50	Unknown	Unique
3179*	■■■■□■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■□■■■	757777777720731	2 (0,49)	50	H3	Clustered
3180*	■□□■■■□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■■■■■■■	470000777760771	2 (0,49)	66,67	T1-RUS2	Clustered
3181*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■□■■■	777777777520731	1 (0,24)	50	H3	Unique
3182*	■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□□□■■■■■	777776777760171	1 (0,24)	50	X1	Unique
3183*	□□□□□□□□□□□□□□□□□□□■■■■■■■■■■■□□□□■■■■■■■	000000077760771	1 (0,24)	33,33	Unknown	Unique
3184*	■■□■■■■■□■■■■■■□■■■■■■□□■■■■■■■■■■■■□□□□□□	676773747777600	1 (0,24)	50	BOV_1	Unique
3185*	■■■■■□□□■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	761777777720771	2 (0,49)	100	H3	Clustered
3186*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□■■■■■□□□□■■■■■■■	777777764760771	1 (0,24)	50	T	Unique
3187*	■□□■■□□□□□□□■■□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■	460060000000371	1 (0,24)	50	Unknown	Unique
3188*	■■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■	700000000000771	1 (0,24)	50	Unknown	Unique
3189*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□■■■□■□□□□■■■■■■■	777777765720771	1 (0,24)	50	H3	Unique
3190*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□□□■■■	777777747760431	1 (0,24)	50	T	Unique
3191*	■□□□□□□□□□□□■■□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■	400060000000771	1 (0,24)	50	Unknown	Unique
3199*	■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	771777777760731	1 (0,24)	33,33	T2	Unique
3276*	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□■■■■	777777770000171	1 (0,24)	33,33	Unknown	Unique

A total of 104 SITs containing 332 isolates matched a preexisting shared type in the SITVIT2 database, whereas 23 SITs (n = 32 isolates) were newly-created either within the present study or after a match with an orphan in the database.

SIT followed by an asterisk indicates ''newly created shared-type'' (n = 24 containing 34 isolates) due to two or more strains belonging to an identical new pattern within this study or after a match with an orphan in the database. SIT designations followed by number of strains: 3168 this study n = 3, PER n = 1; 3169* this study n = 2, USA n = 1; 3170* this study n = 3, SWE n = 1; 3171* this study n = 1, BRA n = 1; 3172* this study n = 1, VEN n = 1; 3173* this study n = 2, POL n = 1; 3174* this study n = 2; 3175 this study n = 1, USA n = 1; 3176* this study n = 1, ZAF n = 1; 3177* this study n = 1, GLP n = 1; 3178* this study n = 1, ALB n = 1; 3179* this study n = 2, FXX n = 2; 3180* this study n = 2, USA n = 1; 3181* this study n = 1, USA n = 1; 3182* this study n = 1, AUS n = 1; 3183* this study n = 1, FXX n = 1, ALB n = 1; 3184* this study n = 1, SWE n = 1; 3185* this study n = 2; 3186* this study n = 1, SWE n = 1; 3187* this study n = 1, SWE n = 1; 3188* this study n = 1, BEL n = 1; 3189* this study n = 1, DEU n = 1; 3190* this study n = 1, DEU n = 1; 3191* this study n = 1, SWE n = 1; 3199* this study n = 1, CHN n = 1, JPN n = 1; 3276* this study n = 1, COL n = 1, ITA n = 1.

**Spoligotype family designations according to SITVIT2 using revised SpolDB4 rules; ''Unknown'' designates patterns with signatures that do not belong to any of the major clades described in the database.

***Clustered strains correspond to a similar spoligotype pattern shared by two or more strains ''within this study''; as opposed to unique strains harboring a spoligotype pattern that does not match with another strain from this study. Unique strains matching a preexisting pattern in the SITVIT2 database are classified as SITs, whereas in case of no match, they are designated as ''orphan''.

doi:10.1371/journal.pone.0046848.t001

Table 2. Orphan strains (n = 45) and corresponding spoligotyping families/subfamilies found among a total of 409 *M. tuberculosis* complex strains isolated from Swedish patients born before 1945.

Isolate	Sex	Age	Spoligotype Description	Octal code	Spoligotype family
BTB09-066	F	92	■■□■■■■■□■■■■■□□□■□□□□■■■■■■■■■■■■■■■■□□□□□	676761037777600	BOV_LIKE
BTB99-387	M	83	■■□■■■■■□■■■■■■□□■■■■■■■■■■■■■■■■■■■■■□□□□□	676771777777600	BOV-1
BTB07-032	M	87	■■■■■■■■■□□□□□□□□□□□□□□■□□□□□□□■□□□□■■■■■■	777400004020771	H1
BTB01-233	F	78	■■■■■■■■■□□□□□□□□□□□□□□■■■■■□■□□□□■■■■■■	777400007720771	H3
BTB01-359	M	82	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■□□□□□	777777777720600	H3
BTB05-050	M	70	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■□■□■■■	777777777720531	H3
BTB05-143	M	63	■■■■■■■■■■■□□□□□■□■■■■■■■■■■■□■□□□□■■■■■■	777701377720771	H3
BTB05-149	F	63	■■■■■■■■■■■■■■■■■□■■■■■■■■□□■□■□□□□■■■■■■	777777377120771	H3
BTB05-507	M	78	□□□□■■■■■■■□■■■■□□■■■■■■■■■■■□□■□□□□■■■■■■	037674777620771	H3
BTB06-380	M	88	■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□■□□□□■■■■■■	777600000020771	H3
BTB06-563	M	79	■■■■■■■■□□□□□□□□□□□□□□□□□□□□□□■□□□□■□■■■■■	777000000020571	H3
BTB06-654	F	83	■■■■■■■■□□□□■■■■■■■■■■□□■■■■■□■□□□□■■□■■■■	776077747720671	H3
BTB07-067	F	64	■■□□□□□□□□□□□□□□□□■■■■■■■■■■■□■□□□□■■■■□■■	600000377720751	H3
BTB07-320	F	82	■■■■■■■■■■■■□■■■■■■■□■■■■■■■■■□□□■□□□□■■■■■■	777737577420771	H3
BTB09-374	F	79	■■□□□□□□□□■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■	600377777720771	H3
BTB09-435	F	98	■■■■■■■■■□□□□□□□□□□□□□□■□□□□■□■□□□□■■■■■■	777400004120771	H3
BTB99-379	M	55	■■■■■■■■■■■■□■■■■■■■■■■■■■□□■■■■■■□■□□□□■■■■■■	777737747720771	H3
S96-202	M	82	■□■□□□□□□■■■■■■□□■■■■■□□■□□□□□□□■□□□□■■■□■■■	500771744020731	H3
BTB00-127	M	89	□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■□□□■□□□□■■■□■■■	000000007420731	H4
BTB05-475	M	85	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■□	677777607760730	LAM
BTB04-450	M	83	■■■■■■■■■■■■■■■■□□■□■■■□■□□□□□□■■□□□□■■■■■■	777771350060771	T
BTB05-005	F	61	■■■■■■□□□□□□□□□□□□■■□■■■■■■■■■■■■□□□□■■■□■■■	770000677760731	T
BTB98-539	M	77	■■■■■■□□□□□□□□□□□□■■□□□□■■■■■■■■■□□□□■■■□■■■	770000607760731	T
BTB00-056	F	76	■■■■■■■■■■■■■■■■■■■■■□□■■■■■■■■□□□□■□□□□■■	777777747760411	T1
BTB00-059	M	83	■■■■■■■■■■■■■■■■■■■□■■■■■□□■■■■■■■■□□□□■□□■■■	777775747760471	T1
BTB01-392	M	72	■■■■■■■■■■□□□□□□□□□□□□□□□□□■■■■■□□□□■■■□□■■	777600001760711	T1
BTB05-487	M	87	□□□□□□□□□□□□□□□□□□□□□■■■■■□■■■■■□□□□■■■■■■	000000176760771	T1
BTB07-129	F	93	■■■□■■□■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	733777777760731	T1
BTB07-182	M	84	■■□■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■□□□□□■■■■■	677777777360371	T1
BTB07-232	M	83	■■■■■■■■■■■■■■■■■■■■■□■■■■■□□■■□□□□■■■■■■■	777777757460771	T1
BTB08-128	F	75	■■■■■■□■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	773777037760771	T1
BTB09-443	F	93	□□□□□□□□□□■■■■■■■■■□□■■■■■■■■■□□□□■■■■■■■	000177717760771	T1
BTB09-515	M	77	■■■■■■■■■■■■■■□■■■■■■■□■□□□□□□■■□□□□■■■■■■■	777767750060771	T1
BTB98-359	F	80	■■■■■■■■■■■■■■■■■■■■□□□□□■□□□□■■□□□□□□■■■■■	777777602060371	T1

Table 2. Cont.

Isolate	Sex	Age	Spoligotype Description	Octal code	Spoligotype family
BTB98-532	F	85	□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■	37767777760771	T1
BTB98-547	M	60	■■□□□□□■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■	60377777760771	T1
BTB99-138	F	80	■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□□	73777777760720	T1
BTB99-476	F	85	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■	77777777360371	T1
S96-295	F	64	■■■■■□□□□□□□■■■■■■■■■■■■■■■■■■■□□□□■■■■■	70000643760771	T1-RUS2
BTB08-174	F	89	■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■	75777767760731	T2
BTB07-190	M	77	■■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■	70777344560471	T4-CEU1
BTB99-125	F	61	■■■■□■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■	75777757760711	T5
BTB04-418	F	89	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■	77777747700771	Unknown
BTB09-069	F	80	■■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■□□	74000000000031	Unknown
BTB98-543	M	69	■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□	77637777740000	Unknown

*Spoligotype family designations for orphan patterns were done manually as Expert-based interpretations using revised SpolDB4 rules.
doi:10.1371/journal.pone.0046848.t002

together with their own family (*M. bovis*), while evolutionary modern PGG2/3 (H, T, LAM) SITs grouped with their own orphans. Almost all of the evolutionary-recent orphan strains appeared at distant (terminal) positions within their respective genotypic families on the tree and not as the central nodes that were essentially defined by spoligotype family prototypes such as SIT50 for H3 and SIT53 for T1 subfamilies. This observation suggests that most of the orphan strains that existed in Sweden a century ago, when TB was highly endemic, probably underwent extinction because they evolved as terminal members within these families that were unfit to proliferate. Indeed, in the Fruchterman-Reingold spoligoforest tree, many PGG2/3 orphans are not linked to other nodes (Figure S2).

IS*6110* RFLP

Out of 409 isolates, 375 different RFLP patterns were obtained, of which 53 (13.0%) were clustered in 19 RFLP clusters comprising 2–9 isolates per cluster (Figure 2). Most clusters were small and the majority comprised only two individuals (n = 15). The remaining 356 (87.0%) RFLP patterns were unique i.e. the strains did not cluster with other patient strains in the SMI database. A clustering comparison was also made with isolates from patients born in Sweden in 1985 or later. It was found that present day isolates cluster much more frequently. Among 58 isolates from these young Swedes, 16 (27.6%) belonged to a cluster compared to 53/409 (13.0%) among the elderly Swedes (chi^2-test p-value 0.0033). All clusters (n = 5) among the young Swedes also contained a foreign born patient with an isolation date prior to the young Swedish born patients.

Discussion

In this study of *M. tuberculosis* complex isolates from patients born in Sweden before 1945 the majority of the isolates should represent the epidemic strains circulating at the time of the down slope of the epidemic that took place in Sweden the last centuries. The shape of the epidemic curve for TB is the same as that for any other infectious disease, if one adjusts the time scale to allow for the roughly 300-year duration of a TB epidemic [22]. Rates of mortality, morbidity and latent infections increase rapidly, peak successively, and decline at a prolonged, exponentially decelerated pace [22]. In Europe the epidemic started over three centuries ago, and the TB morbidity and mortality began to fall well before introduction of BCG vaccine and effective treatment. It began in the late 16th century in England, where it peaked around the 1780s [10]. From England it spread rapidly eastwards, to developing cities on the continent, where the curves peaked in the first decades of the 19th century. In Scandinavia, the peaks in mortality were reached in the last decades of the 19th century [10], and the incidence declined to 147 per 100.000 population in 1950 and to 1.4 in 2009 in the Swedish-born population. A similar shape of the epidemic curve can be observed in the neighbouring Scandinavian countries. The present TB situation in Sweden is largely influenced by migration as only 11% of the patients diagnosed in 2011 were born in Sweden. Approximately 40% of all isolates typed in 2011 belonged to a cluster [6]. The low rate of isolates clustered by RFLP in the elderly population studied here supports the concept that most isolates represent reactivation cases, without active TB transmission, and thus represent the past epidemic. Indeed, a significant difference in clustering was found when this elderly Swedish population was compared to a young Swedish population.

We found a highly homogenous bacterial population with a domination of the H (41.1%), T (37.7%), and LAM (7.8%)

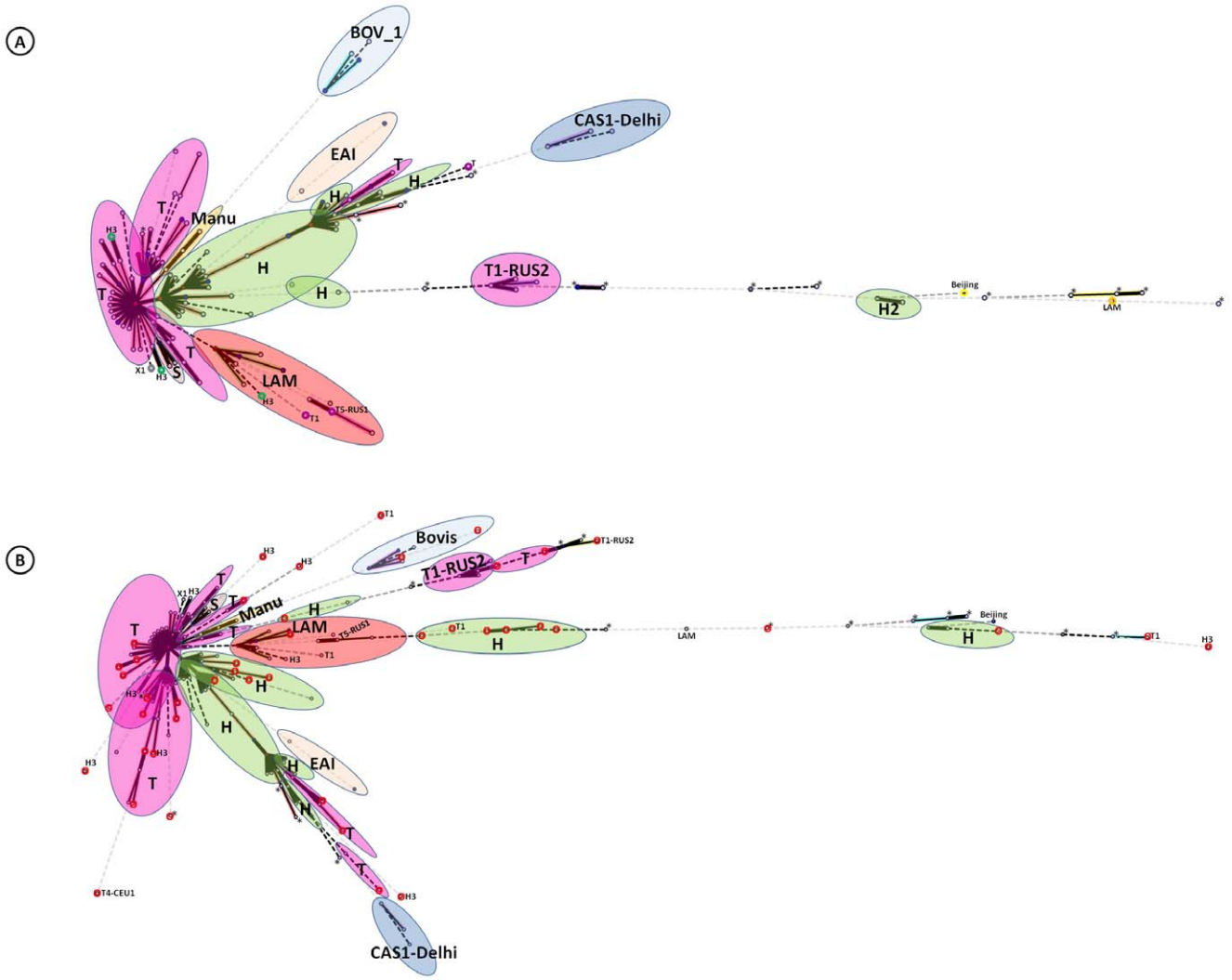


Figure 1. A minimum spanning tree (MST) illustrating possible evolutionary relationships between the Swedish spoligotypes obtained for *M. tuberculosis* complex strains causing the TB epidemic a century ago. (**A**) **SITs alone** (**B**) **all patterns (including the orphan patterns) pooled together.** The tree connects each genotype based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, black dotted and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines, 1 or 2 change (thicker ones indicate a single change, while the thinner ones indicate 2 changes); dotted lines, three or more changes (black dotted for 3, and grey dotted for 4 or more changes). The color of the circles is proportional to the number of clinical isolates in our study, illustrating unique isolates (sky blue) versus clustered isolates (blue, 2–5 strains; dark blue, 6–9 strains; bordeaux, 10–19 strains; red, 20 and more). Abbreviation: PGG, Principal Genetic Group. Note that orphan patterns in Fig. 1B are circled in red while patterns marked by an asterisk (*) indicate an strain with an unknown signature (unclassified).
doi:10.1371/journal.pone.0046848.g001

families, which belong to the modern Euro-American group of strains and includes all the spoligotype families predominating in the Western world, such as H, LAM, the ill-defined T group, X and S. Only 24 isolates did not belong to modern families. This high prevalence of modern H, T and LAM strains is similar to the prevalence among isolates from patients born before 1950 in Norway, where a total of 40% of 213 isolates belonged to the T family and 35% to the H family [23]. As the Norwegian study, our study only included isolates displaying RFLP patterns not previously present in our database indicating an unlikely recent transmission. Both studies demonstrated that the isolates were of a highly homogenous population (T and H family) with low rate of diversity. The two major spoligotypes in our study were SIT50 of the H3 subfamily, and SIT53 of the T1 subfamily. Although they have been designated to two different subfamilies, they only differ in one spacer. In addition to the lack of spacers 33–36 in both types SIT50 also lacks spacer 31. SIT53 is the prototype pattern of the T family and is widely spread in the world [11]. The T family defines a polyphyletic group of strains belonging to the Euro-American superlineage. It does not represent a lineage in a strict evolutionary sense since it was defined by default [11]. Although, some subfamilies belonging to the T group have a known phylogeographical specificity, i.e. T-Tuscany, T1-RUS2 (Russia), T2-Uganda, T3-ETH (Ethiopia), T3-OSA (Osaka), T4-CEU1 (Central Europe), T5-Madrid2, the remaining T1 to T5 spoligo-prototypes are not monophyletic [24]. In a molecular analysis of *M. tuberculosis* DNA from a family of 18th century Hungarians two spoligotypes were identified, corresponding to SIT50 and SIT53

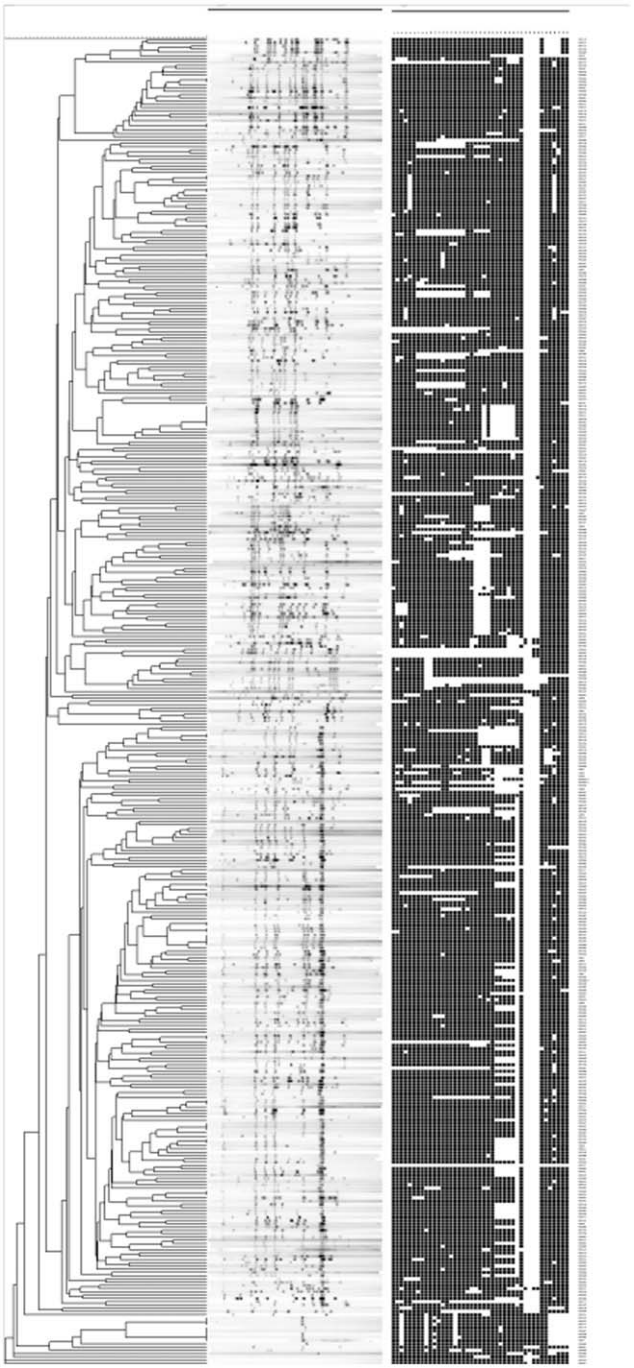


Figure 2. IS*6110* Restriction Fragment Length Polymorphism and spoligotyping dendrogram of 409 *M. tuberculosis* complex strains from 409 patients born in Sweden before 1945.
doi:10.1371/journal.pone.0046848.g002

[25]. As in our study SIT53 is found in elderly patients in other parts of the world: in a study from Venezuela, patients with SIT53 had a significantly higher mean age compared to all other patients [26], and in Mexico SIT53 was significantly more prevalent in elderly patients, especially in females [27]. Polymorphism in the direct repeat region in clinical isolates appears to be the result of successive deletions of single discrete direct variant repeats (DVRs) or of multiple contiguous DVRs from a primordial direct repeat region [28]. Thus the lack of spacer block 33–36 in SIT53 is a very large signature that defines almost all evolutionary modern PGG2/3 *M. tuberculosis* strains. It also corresponds to the large sequence polymorphism-based broader "Euro-American super-lineage". Thus it can be seen as the baseline structure among evolutionary-modern *M. tuberculosis* complex strains (defined by prototype SIT53). From it might have evolved most of other "modern" strains by loss of spacers, including the H family strains. The H baseline therefore is defined by prototype SIT50 of the H3 subfamily (which has a characteristic absence of spacer 31). In addition to the T1 subfamily prototype, SIT53, two more T clade SITs (SIT37 and SIT153) were among the seven predominant SITs. It can be speculated that they separately derived from the prototype SIT53 by loss of spacer 13 (SIT37), or the loss of spacers 5 and 40 (SIT153). However, to know whether such differences are caused by a single or multiple event(s), one would need complementary investigations.

We only found 24 ancestral PGG1 *M. tuberculosis* strains, including isolates of the EAI2-Manila and CAS families. EAI2-Manila isolates are more commonly found in Asian countries, such as Indonesia or the Philippines, where they account for high percentages of the *M. tuberculosis* isolates. Ancestral PGG1 strains are usually linked to migrants in todays Sweden [4,29], indicating they are more recently introduced in Sweden. The majority of the isolates collected in Sweden present in the database at SMI are from foreign born patients. Only three isolates were of the recently emerging Beijing genotype. The majority of patients with drug resistant Beijing strains in Sweden are foreign born [30]. The three patients with Beijing isolates were all diagnosed between 2007 and 2008, i.e. in the later part of the study, at a mean age of 80.7 years. Since the Beijing family is recently introduced in Sweden [30], it is possible that these cases are recent infections and not reactivation of earlier infections.

In the 1930s and 1940s bovine TB was considered to be a significant zoonosis in Europe, including Sweden, and *M. bovis* was thought to be responsible for more than 50% of cervical lymphadenitis cases in children [31] through milk and dairy products from infected herds. In 1958, Sweden was declared free of bovine TB after an extensive national eradication campaign, by the long term application of a test-and-slaughter policy that removed infected cattle [32]. Thus, the 11 *M. bovis* isolates are all from patients born before 1958, indeed the youngest persons with *M. bovis* infection were born in 1911. The four SIT691 isolates have previously been shown to contain the RDEU1 deletion [33] which is specific for the clonal complex European 1 (Eu1), the dominant *M. bovis* complex isolated from cattle in the Republic of Ireland and the UK.

The fact that patients with isolates of the H, T and LAM families were significantly older at diagnosis, and were born significantly earlier than patients with "ancient" isolates further supports the hypothesis that the closely knit H, T, and LAM isolates represent the old TB epidemic in Sweden, and probably the whole of Scandinavia. In the study from Venezuela, patients with SIT53 were not only older but were more commonly smear negative. The authors draw the conclusion that SIT53 strains may be less virulent and associated with reactivation of past infections in older patients. These associations provoke a number of questions. If SIT53 were really less virulent, why is it still the sixth most common spoligotype in SITVIT2, causing 4% of cases? Could it have been a very common strain in the past, that is now more apt at latency and reactivation than person-to-person transmission, and will its prevalence decrease over time? One could speculate that SIT53 represents a progenitor of the strains causing the epidemic in Sweden and Norway, where mutations

causing attenuation of the outbreak strains are illustrated by the further loss of spacers in e.g. SIT50, and SIT153.

Although spoligotyping is known to show marked homoplasy [34], we find the fact interesting that evolutionary recent orphan strains of the PGG2/3 grouping essentially appeared at terminal positions within their respective genotypic families on the MST. The fact that none constituted the central nodes within their own major genotypic families nor appeared to play an important role in the interconnection of prevailing families, suggests why they are probably not at the heart of the *M. tuberculosis* biodiversity observed in Sweden. It is possible that these orphan strains correspond to strains that were simply not represented well enough in the Swedish TB epidemic to have continued on infecting people. However, we believe that most of these orphan strains probably underwent extinction because they were the most vulnerable and unfit strains to be able to continue infecting new hosts during the epidemic peak of the disease a century ago. A detailed genetic analysis might discover a link to a lack of fitness of these strains. In conclusion the study of genetic variability within natural populations of pathogens may provide insight into bacterial evolution and pathogenesis. The characteristics of the bacterial population studied here are those of an old outbreak under extinction, with the superimposition of new characteristics due to cases of importation and recent transmission.

Supporting Information

Figure S1 Hierarchical layout based spoligoforest tree showing only the shared-type SIT patterns. The spoligoforest trees are drawn using http://www.emi.unsw.edu.au/spolTools In this figure each spoligotype pattern is represented by a circle with the area proportional to the number of isolates with that pattern (in the circles SIT numbers are shown followed by the number of isolates with that pattern). Note that contrary to the Minimum spanning trees, the Spoligoforest trees are directed, and only evolve by loss of spacers. Comments on Figure S1: The Hierarchical layout based tree shown in Figure S1 represents hierarchically the changes between *Mycobacterium tuberculosis* complex strains; the more the strains evolve (lose spacers), the more they are present in the down layouts (bottom of the figure, like tree leaves). The patterns which have not lost many alleles/spacers are located in the upper layouts. However, in case of too many changes between two spoligotype patterns, there are no links/edges linking them (like some of the orphan patterns shown as the smallest circles on the top). In this method, the authors used a heuristic method that selects a single inbound edge with a maximum weight using a Zipf model; solid black lines link patterns which have a maximum weight of distance (very similar: loss of one spacer). Dashed line represents a link of weight comprised between 0.5 and 1. And dotted line represents a link of weight less than 0.5. In our study sample, one can notice the predominance of evolutionary recent *M. tuberculosis* spoligotype families (Haarlem, ill-defined T, and LAM): SIT50/H3 (n = 56), SIT53/T1 (n = 43), SIT47/H1) (n = 33); followed by SIT42/LAM9 (n = 14). Regarding some interesting evolutionary conclusions, one may notice; (i) on the top left of the figure, that SIT26/CAS1-Delhi leads to SIT1/Beijing as a second generation descendant (via SIT485/CAS1-Delhi); (ii) on the top center of the figure, patterns belonging to Manu family may evolve to highly represented (and ubiquitous) SIT53/T1 spoligotype family.
(PDF)

Figure S2 Fruchterman Reingold algorithm based spoligoforest tree of all spoligotype patterns including orphans. The spoligoforest trees are drawn using http://www.emi.unsw.edu.au/spolTools In this figure each spoligotype pattern is represented by a circle with the area proportional to the number of isolates with that pattern (in the circles SIT numbers are shown followed by the number of isolates with that pattern). Note that contrary to the Minimum spanning trees, the Spoligoforest trees are directed, and only evolve by loss of spacers. Comments on Figure S2: The Fruchterman Reingold algorithm based tree shown in Figure S2 essentially represents the same relationships as in the Hierarchical layout, but the nodes containing a huge number of strains are centered and more visible in the figure. Considering that the patterns evolve by loss of spacers, the proximity of the nodes represents similarity between them. Note that this figure was drawn without orphan patterns to better show the SIT numbers (in circles) followed by the number of strains for each pattern (shown in brackets). The interpretation of the links (solid black, dashed, or dotted) is the same as for Hierarchical layout.
(PDF)

Table S1 Patients born in Sweden between the years 1908–1945.
(DOCX)

Table S2 Patients born in Sweden between the years 1985–2008.
(DOCX)

Table S3 Description of predominant SITs (patterns representing ≥2% strains in our study), and their worldwide distribution.
(DOCX)

Table S4 Description of rare PGG1 spoligotype patterns in this study (n = 15 patterns containing 24 isolates), and their worldwide distribution.
(DOCX)

Acknowledgments

We thank the Swedish TB laboratories for providing the strains and for fruitful discussions, Lina Benson for important help with the statistical analysis and Stefan Svenson for valuable suggestions for the initiation and design of the study.

Author Contributions

Conceived and designed the experiments: GK. Performed the experiments: RG SG AP DC NR. Analyzed the data: RG SG AP JJ SH DC TK NR GK. Contributed reagents/materials/analysis tools: JJ DC NR. Wrote the paper: RG SG AP SH TK NR GK.

References

1. Lillebaek T, Dirksen A, Kok-Jensen A, Andersen AB (2004) A dominant Mycobacterium tuberculosis strain emerging in Denmark. Int J Tuberc Lung Dis 8: 1001–1006.
2. Dahle UR, Sandven P, Heldal E, Caugant DA (2003) Continued low rates of transmission of Mycobacterium tuberculosis in Norway. J Clin Microbiol 41: 2968–2973.
3. Kan B, Berggren I, Ghebremichael S, Bennet R, Bruchfeld J, et al. (2008) Extensive transmission of an isoniazid-resistant strain of Mycobacterium tuberculosis in Sweden. Int J Tuberc Lung Dis 12: 199–204.
4. Ghebremichael S, Petersson R, Koivula T, Pennhag A, Romanus V, et al. (2008) Molecular epidemiology of drug-resistant tuberculosis in Sweden. Microbes Infect 10: 699–705.

5. Sandegren L, Groenheit R, Koivula T, Ghebremichael S, Advani A, et al. (2011) Genomic stability over 9 years of an isoniazid resistant Mycobacterium tuberculosis outbreak strain in Sweden. PLoS One 6: e16647.
6. Smittskyddsinstituet (2011) Statistik för tuberkulos.
7. Tamm G, Dovertie G., Johansson J., Nylander O., Östberg G. (1907) Betänkande och förslag angående tuberkulossjukvårdens ordnande i riket.
8. Lillebaek T, Dirksen A, Baess I, Strunge B, Thomsen VO, et al. (2002) Molecular evidence of endogenous reactivation of Mycobacterium tuberculosis after 33 years of latent infection. J Infect Dis 185: 401–404.
9. Winqvist N, Bjork J, Miorner H, Bjorkman P (2011) Long-term course of Mycobacterium tuberculosis infection in Swedish birth cohorts during the twentieth century. Int J Tuberc Lung Dis 15: 736–740.
10. Winqvist N (2011) Dynamics of tuberculosis infection in Sweden [Doktorsavhandling]. Malmö: Lunds universitet. 155 p.
11. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, et al. (2006) Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6: 23.
12. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, et al. (1997) Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 35: 907–914.
13. Reyes JF, Francis AR, Tanaka MM (2008) Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. Bmc Bioinformatics 9.
14. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, et al. (1997) Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 94: 9869–9874.
15. Gagneux S, Small PM (2007) Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infectious Diseases 7: 328–337.
16. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, et al. (2006) Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences of the United States of America 103: 2869–2873.
17. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, et al. (2002) A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99: 3684–3689.
18. Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, et al. (2000) Characterization of Mycobacterium tuberculosis isolates from patients in Houston, Texas, by spoligotyping. J Clin Microbiol 38: 669–676.
19. Rastogi N, Sola C Molecular evolution of the Mycobacterium tuberculosis complex. Amadeo Online Textbooks 2007. Available: http://www.tuberculosistextbook.com/index.htm. Accessed 19 September 2011.
20. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, et al. (1993) Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 31: 406–409.
21. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD (1991) Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol 29: 2578–2586.
22. Grigg ER (1958) The arcana of tuberculosis with a brief epidemiologic history of the disease in the U.S.A. IV. Am Rev Tuberc 78: 583–603.
23. Kinander W, Bruvik T, Dahle UR (2009) Dominant Mycobacterium tuberculosis lineages in elderly patients born in Norway. PLoS One 4: e8373.
24. Demay C, Liens B, Burguiere T, Hill V, Couvin D, et al. (2012) SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol 12: 755–766.
25. Fletcher HA, Donoghue HD, Taylor GM, van der Zanden AG, Spigelman M (2003) Molecular analysis of Mycobacterium tuberculosis DNA from a family of 18th century Hungarians. Microbiology 149: 143–151.
26. Abadia E, Sequera M, Ortega D, Mendez MV, Escalona A, et al. (2009) Mycobacterium tuberculosis ecology in Venezuela: epidemiologic correlates of common spoligotypes and a large clonal cluster defined by MIRU-VNTR-24. BMC Infect Dis 9: 122.
27. Molina-Torres CA, Moreno-Torres E, Ocampo-Candiani J, Rendon A, Blackwood K, et al. (2010) Mycobacterium tuberculosis spoligotypes in Monterrey, Mexico. J Clin Microbiol 48: 448–455.
28. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, et al. (1991) Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. Infect Immun 59: 2695–2705.
29. Svensson E, Millet J, Lindqvist A, Olsson M, Ridell M, et al. (2011) Impact of immigration on tuberculosis epidemiology in a low-incidence country. Clin Microbiol Infect 17: 881–887.
30. Ghebremichael S, Groenheit R, Pennhag A, Koivula T, Andersson E, et al. (2010) Drug resistant Mycobacterium tuberculosis of the Beijing genotype does not spread in Sweden. PLoS One 5: e10893.
31. Cosivi O, Meslin FX, Daborn CJ, Grange JM (1995) Epidemiology of Mycobacterium bovis infection in animals and humans, with particular reference to Africa. Rev Sci Tech 14: 733–746.
32. Szewzyk R, Svenson SB, Hoffner SE, Bolske G, Wahlstrom H, et al. (1995) Molecular epidemiological studies of Mycobacterium bovis infections in humans and animals in Sweden. J Clin Microbiol 33: 3183–3185.
33. Smith NH, Berg S, Dale J, Allen A, Rodriguez S, et al. (2011) European 1: A globally important clonal complex of Mycobacterium bovis. Infect Genet Evol 11: 1340–1351.
34. Comas I, Homolka S, Niemann S, Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS One 4: e7815.

OPEN ACCESS Freely available online

PLOS ONE

Molecular Epidemiology of Tuberculosis in Finland, 2008-2011

Pieter Willem Smit[1,2*], Marjo Haanperä[2], Pirre Rantala[2], David Couvin[3], Outi Lyytikäinen[2], Nalin Rastogi[3], Petri Ruutu[2], Hanna Soini[2]

1 European Public Health Microbiology Training Programme, (EUPHEM), European Centre for Disease Prevention and Control (ECDC), Stockholm, Sweden, **2** Department of Infectious Disease Surveillance and Control, National Institute for Health and Welfare, Helsinki, Finland, **3** WHO Supranational TB Reference Laboratory, Tuberculosis & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Abymes, France

Abstract

In industrialized countries the majority of tuberculosis (TB) cases are linked to immigration. In Finland, most cases are still Finnish born but the number of foreign born cases is steadily increasing. In this 4-year population based study, the TB situation in Finland was characterized by a genotypic analysis of *Mycobacterium tuberculosis* isolates. A total of 1048 *M. tuberculosis* isolates (representing 99.4% of all culture positive cases) were analyzed by spoligotyping and MIRU. Spoligotype lineages belonging to the Euro-American family were predominant among the Finnish isolates, particularly T (n=346, 33.0%) and Haarlem (n=237, 22.6%) strains. The lineage signature was unknown for 130 (12.4%) isolates. Out of the 17 multi-drug resistant TB strains, 10 (58.8%) belonged to the Beijing lineage. In total, 23 new SIT designations were given and 51 orphan strains were found, of which 58 patterns were unique to Finland. Phylogeographical TB mapping as compared to neighboring countries showed that the population structure in Finland most closely resembled that observed in Sweden. By combining spoligotyping and MIRU results, 98 clusters comprising 355 isolates (33.9%) were found. Only 10 clusters contained both Finnish and foreign born cases. In conclusion, a large proportion of the *M. tuberculosis* isolates were from Finnish born elderly patients. Moreover, many previously unidentified spoligotype profiles and isolates belonging to unknown lineages were encountered.

Citation: Smit PW, Haanperä M, Rantala P, Couvin D, Lyytikäinen O, et al. (2013) Molecular Epidemiology of Tuberculosis in Finland, 2008-2011. PLoS ONE 8(12): e85027. doi:10.1371/journal.pone.0085027

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received August 23, 2013; **Accepted** November 20, 2013; **Published** December 26, 2013

Copyright: © 2013 Smit et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: Pieter.smit@thl.fi

Introduction

Finland had a very severe tuberculosis (TB) epidemic in the beginning of the 20th century, with respiratory TB mortality rates of 250-300 per 100,000 [1]. After the Second World War, TB mortality rates started to decline reaching 51 per 100,000 in 1951. This rapid decrease was related to an improved standard of living, effective TB control program and availability of TB medication [1]. In 2001, Finland became a low incidence country when TB incidence fell below 10 per 100,000, and in 2012 the TB incidence was 5/100,000.

However, the epidemiological situation in Finland is different from neighboring low incidence countries. In Sweden, Norway and Denmark, a transition took place from the native population to foreigners, that currently account up to 80% of the new TB cases [2–4]. In Finland, most cases are still Finnish born but the number of foreign born cases is steadily increasing over time. In our neighboring countries in the East and South, the TB incidence is high and predominantly associated with *Mycobacterium tuberculosis* isolates belonging to the Beijing genotype family [5]. Multidrug resistant (MDR) TB is a significant problem in Russia and Estonia, while in Finland this remains a rare occurrence [6–8]. Puustinen et al. showed that *M. tuberculosis* isolates belonging to the Beijing lineage were infrequently encountered in Finland [9]. However, MDR-TB cases will most likely increase in Finland as well, because of the rapid spread of the Beijing lineage across Europe and increasing immigration to Finland.

To better understand the current epidemiology of TB in a country with a history of high TB incidence, the aim of the study was to evaluate the current TB situation in Finland by genotypic characterization of *M. tuberculosis* complex isolates. Our first objective was to use the typing data to provide a general overview of the genotypes retrieved from TB cases in Finland, and compare the situation in Finland to our neighboring countries.

Materials and Methods

Since 1995, the reporting of approximately 70 infectious diseases, including TB, to the National Infectious Disease Registry has been mandatory for physicians and clinical microbiology laboratories in Finland [8]. In addition, all *M. tuberculosis* complex isolates are sent to the Mycobacterial Reference Laboratory at the National Institute for Health and Welfare (THL) for species confirmation, genotyping, drug susceptibility testing, and strain collection.

Isolates

Between 2008 and 2011, 1397 patients were notified with TB, and cultures with *M. tuberculosis* complex were obtained from 1054 patients. The data in this study were analyzed within the epidemiological research purposes authorized by the Finnish Communicable Diseases Act (clause 4 in section 40a, Finlex, Valtion säädöstietopankki. Available at: http://www.finlex.fi/fi/laki/alkup/2003/20030935 (accessed 23.08.2013). Therefore, ethical approval was deemed unnecessary.

From the culture positive patients with a valid personal identification code (Finnish social security number), 1048 (99.4%) isolates were successfully genotyped at the reference laboratory, and included in this study. The isolates were cultured on Löwenstein – Jensen medium for further analysis. All isolates were characterized by spoligotyping, as described previously [9]. The spoligotype patterns were uploaded and assignment of the Spoligo International Type (SIT) designations was done using the international SITVIT2 database of the Institute Pasteur of Guadeloupe, which is an updated version of the SITVITWEB database [10]. When a spoligotype was not found in the SITVIT2 database, it was classified as a unique spoligotype. The term "orphan spoligotype" refers to an isolate having a spoligotype that was present only once in our dataset. When multiple isolates (i.e. 2 or more) were obtained with a unique spoligotype pattern, a new SIT number was assigned by Institute Pasteur. For the unique patterns, lineage designations were given manually by expert-based interpretations using revised SpolDB4 rules. A Finnish designation "F-type" was assigned to the Finnish unique orphan spoligotypes. Fifteen loci mycobacterial interspersed repetitive units -variable number of tandem repeats (MIRU VNTR (MIRU)) was performed at Genoscreen (Lille, France) on all isolates.

Genotypic lineages (i.e. clades or families) were obtained from the SITVIT2 database (accession date: May 16th 2013), designated in the database as: the Beijing clade, the Central Asian (CAS) clade and 2 sublineages, the East African-Indian (EAI) clade and 9 sublineages, the Haarlem (H) clade and 3 sublineages, the Latin American-Mediterranean (LAM) clade and 12 sublineages, the ancestral "Manu" family and 3 sublineages, the S clade, the IS*6110*-low-banding X clade and 3 sublineages, an ill-defined T clade with 5 sublineages, as well as the recently described Ural lineage (with 2 sublineages Ural-1 and Ural-2).

Statistical analysis

Data regarding patients, clinical presentation, and *M. tuberculosis* complex isolates were obtained from the laboratory database and from the National Infectious Diseases Registry, and analyzed in STATA 12 (Statacorp LP, USA). When multiple samples were collected from the same patient, the first isolate with a genotype result was used. Spoligotypes in binary format and MIRU results were entered into BioNumerics software version 6.6 (Applied Maths Inc., Sint-Martens-Latem, Belgium) for cluster analysis. Minimum spanning tree (MST) was constructed by permutation resampling (200x) with highest resampling proportion. Fischer exact and Pearson Chi_2 test were performed to assess the similarity between lineages.

Results

Of the 1048 patients, 733 (69.9%) were born in Finland, 264 (25.2%) were foreign born and for 51 (4.9%) patients the place of birth was unknown. 481 patients were ≥65 years old (45.9%) and the median age of the patients was 61 years (range: 0-103), and 601 (57.3%) patients were male.

Bacterial isolates

Of the 1048 *M. tuberculosis* complex isolates, 1044 (99.6%) were *M. tuberculosis*, three were *Mycobacterium africanum*, and one was *Mycobacterium bovis*. With MIRU typing, 722 distinct profiles were obtained of which 149 (20.6%, 475 isolates) were recognized and 573 (79.4%) were not recognized in the MIRU-VNTR plus website [11].From the 273 distinct spoligotypes that were identified, 215 (78.7%) were recognized in the SITVIT2 database while 58 (21.2%) were not recognized (unique). Out of the 58 unique spoligotypes, 23 new SITs (41 isolates) were created using revised SpolDB4 rules. Of the new SITs and unique orphanspoligotype patterns, 68 isolates were only detected in Finland (F-type), 67.6% (46/68) of these were obtained from Finnish born patients. 51 strains remained orphan. An overview of these strains can be found in Table S1 and Figure S1.

M. tuberculosis genotypic lineages

The most common TB lineages found from the Finnish isolates were T (n=346, 33.0%) and Haarlem (n=237, 22.6%) (Table 1 and Figure 1). For 130 isolates (12.4%) the lineage could not be defined (i.e. was unknown). Out of the 130 isolates, 103 (79.2%) were obtained from Finnish born patients. Based on phylogeographical mapping, the distribution of lineages closely resembles the distribution seen in Sweden (Figure 2). The main difference was for the Ural lineage, which according to the SITVIT2 database, was not regularly found in Sweden. Both Ural-1 and Ural-2 (42 and 18 isolates respectively) were predominantly found among Finnish born cases (50/60, 83.3%). Additionally, isolates not belonging to any known lineage (i.e. undefined clade) were found more commonly among Finnish born (n=103, 79.2%) than foreign born patients (n=25, 19.2%) (p: <0.05).

A minimum spanning tree using spoligotyping and MIRU as categorical data shows the phylogenetic similarity of the *M.*

Table 1. An overview of the lineages found in Finland, 2008-2011.

Lineage	Total	Finnish-born	Foreign-born	Unknown origin	Median age	Pulmonary	Extra-pulmonary	Pan-susceptible	Drug resistant	MDR
AFRI	3	0	2	1	33	1	2	3	0	0
Beijing	55	11	40	4	34	41	14	31	14	10
BOV	1	0	1	0	25	0	1	0	1	0
Cameroon	1	0	1	0	42	1	0	1	0	0
CAS	44	6	33	5	27	22	22	33	8	3
EAI	70	5	53	12	31	34	36	53	17	0
Haarlem	237	214	20	3	73	143	94	217	19	0
LAM	77	65	10	2	55	57	20	73	3	1
Manu	1	0	1	0	24	0	1	1	0	0
S	14	6	5	3	50	7	7	12	1	1
T	346	268	63	15	66	201	145	325	19	1
Turkey	3	3	0	0	64	0	3	3	0	0
Unknown	130	103	25	2	64	75	55	115	14	1
Ural	60	50	6	4	71	40	20	56	4	0
X	6	2	4	0	66	3	3	6	0	0
	1048	733	264	51		625	423	929	100	17

doi: 10.1371/journal.pone.0085027.t001

tuberculosis strains. The 68 isolates with a spoligotype only found in Finland (F-type) were mostly unlinked and seemed phylogenetically remote to SIT strains (Figure S1 ; Hierarchical layout shows that F-types appear at terminal position in this Spoligoforest).

Clustering

A total of 98 clusters comprising 873 isolates (83.2%) were detected by using spoligotyping alone. By combining spoligotyping and MIRU profiles, 791 distinct combinations were obtained. The 98 combined clusters contained 355 isolates (33.9% of all isolates). The largest cluster, consisting of 20 isolates, belonged to SIT42/LAM9 lineage. All were Finnish patients. Only 10 of the 98 (10.2%) clusters contained both Finnish and foreign born patients. MIRU has a higher discriminatory power than spoligotyping, usually fragmenting spoligotype clusters. However, our results indicate that 39 MIRU patterns consisting of 220 isolates (5.4% of all MIRU patterns, 21.0% of all isolates) were divided by spoligotyping. Overall, 65 (25.0%) of the foreign born patients and 273 (37.2%) of the Finns were clustered by using spoligotyping and MIRU combined.

Beijing lineage and drug resistance

Among the 1048 isolates analyzed, 117 (11.2%) were drug resistant and the most common resistance was to isoniazid only (28 out of 117 isolates/ 23.9%). Seventeen (1.6%) were multidrug resistant (MDR-TB). Out of the 17 MDR-TB cases, 7 cases (41.2%) were Finnish born of which one was ≥65 years old. Six out of the 17 isolates (35.3%) were clustered, and only one cluster contained more than one MDR-TB patient.

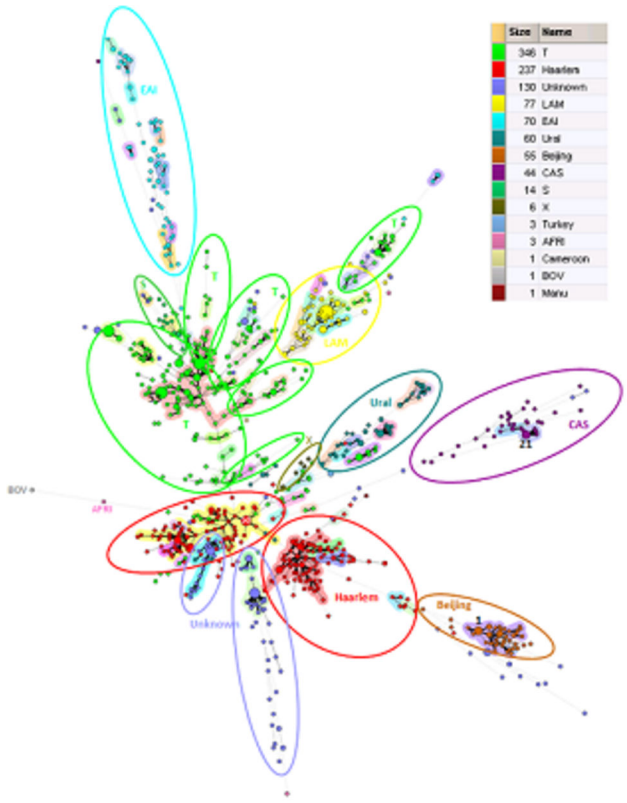


Figure 1. Cluster analysis of all isolates found in this study. Colors of nodes depict the lineage of the strain.

doi: 10.1371/journal.pone.0085027.g001

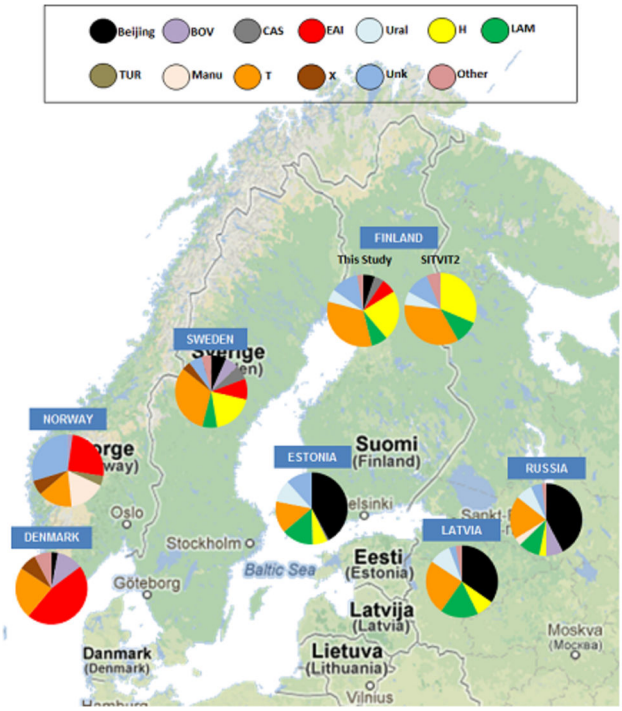


Figure 2. Phylogeographical distribution of *M. tuberculosis* lineages in neighboring countries. Colors of nodes depict the lineage of the strain. Note that the proportion of Ural lineage (notably Ural-1) is not negligible in the following countries: Russia, Latvia, Estonia, and Finland [11].

doi: 10.1371/journal.pone.0085027.g002

Discussion

This study provides an update of the molecular epidemiology of TB in Finland. The 4-year nationwide data showed that the majority of the TB cases were Finnish born and a large proportion was aged 65 or older. This is uncommon for a low TB incidence country, as can be seen, for example in Sweden or Denmark where the vast majority of TB cases are immigrants [2,4]. This is primarily explained by a different immigration situation of these countries, as in Sweden and Denmark 10% of the population is born abroad, while the corresponding figure in Finland is less than 3%. This could also be the reason why there seems to be less transmission in Finland between foreign and Finnish born cases, compared to Sweden and Denmark [3,4].

In Finland, a relatively high clustering rate (33.9%) was noted compared to Sweden (23.6% using MIRU-VNTR and spoligotyping) [12]. Clustering rate was higher among Finnish born than immigrant cases, which was also found in Denmark using a different methodology [4]. The interpretation of clustering in population based studies is complicated, particularly if it covers stable urban and rural populations. While clustering is associated with recent transmission in urban settings, it is more complicated to interpret on population levels due to possible predominance of certain well-preserved strains in the population [13,14]. Additionally, Finland – and other Nordic countries- had endured a very high tuberculosis incidence before the second world war [1]. As reported from Sweden and Norway, certain lineages (Haarlem and T) were predominately found among elderly cases, suggesting that these lineages were most likely associated with the epidemic during the first half of the 20th century [15,16]. As found in this study, Haarlem and T were also the most common lineages found among Finnish born TB cases. The relatively high rate of clustering could partly be explained by well conserved strains in the Finnish population.

The overall analysis shows that the isolates belonged to many different genotypic lineages, reflecting the wide variety of *M. tuberculosis* strains circulating in Finland. The most common lineages were T and Haarlem as in the case in the rest of Western Europe [10], although for many isolates the lineage was unknown. Based on phylogeographical analysis, the distribution of TB lineages in Finland seems to be similar to Sweden, but not to the other neighboring countries. The main difference between the lineages in Sweden and Finland was the presence of the Ural lineage in Finland (5.7% of isolates). The relatively recently named Ural lineage [17] was primarily found among Finnish born cases. This lineage is common in Estonia, Russia, and Latvia (Figure 2). Certain lineages such as Beijing, which is dominant in Russia, was barely seen in Finland 12 years ago and is now one of the more common lineages [9]. MDR-TB is still rare and most of the isolates belonged to the Beijing lineage.

Almost all lineages were seen in both foreign born and Finnish born patients, but only one large cluster was shared by the two groups, suggesting that transmission of TB between Finns and foreigners does not occur frequently. In-depth cluster analysis and epidemiological research is warranted to investigate this conclusively.

Although the majority of our isolates could be genotyped and classified successfully, many of the spoligotypes and MIRU profiles found in this study were not recognized in the international databases. We identified 68 isolates with spoligopatterns that were exclusively encountered in Finland (F-type), and of those 51 were orphan spoligotypes. This could be due to information lacking in the databases or could reflect the unique nature of Finnish *M. tuberculosis* isolates. These strains were genetically distinct from the more common SIT strains, but more research is needed to explore this in-depth.

In conclusion, a large proportion of the *M. tuberculosis* isolates seen in Finland originated from Finnish born elderly patients. Moreover, many previously unidentified spoligotype profiles and isolates belonging to unknown lineages were encountered. Based on these study results, the TB situation in Finland is different from other low incidence countries. More research and in-depth assessment is needed to obtain a better insight in the unusual TB situation in Finland.

Supporting Information

Figure S1. Hierarchical layout of spoligotypes, Finland, 2008-2011. A representation of parent to descendant spoligotypes within our study sample (n=1048 isolates) as seen through Spoligoforest trees drawn using the SpolTools

software (available through http://www.emi.unsw.edu.au/spolTools), and reshaped and colored using the GraphViz software (available through: http://www.graphviz.org). The tree shown was drawn using a Hierarchical Layout where the F-types and orphan strains are highlighted in green. In this tree, each spoligotype pattern from the study is represented by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. The heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weigh being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5. Note that SIT53/T1 constitutes the biggest node (n=135), followed by SIT47/H1 (n=65), SIT1/Beijing (n=55), SIT49/H3 (n=43) and SIT50/H3 (n=40), which are other predominant patterns in Finland. On the other hand, F-types and/or orphan isolates appear mostly at terminal positions on the tree, or as isolated strains without interconnections with the other strains.
(PDF)

Table S1. Orphan and F-type strains (n=68) and corresponding spoligotyping.
(XLSX)

Table S2. Overview of spoligotypes found in Finland, 2008-2011.
(XLSX)

Acknowledgements

The authors would like to thank Anna Musku, Tiina Nurmio, Elisa Rehnberg, and Arto Talvitie for their contribution in analyzing the isolates. Additionally, we would like to thank Aftab Jasir for critically reading the manuscript.

Author Contributions

Conceived and designed the experiments: PWS MH DC OL NR HS P. Ruutu. Performed the experiments: PWS HS OL MH P. Rantala. Analyzed the data: PWS NR DC HS. Contributed reagents/materials/analysis tools: PWS DC NR P. Rantala MH HS P. Ruutu. Wrote the manuscript: PWS NR HS OL MH P. Ruutu P. Rantala.

References

1. TUBERCULOSIS mortality in Finland (1955) Bull World Health Organ 12: 211–246. PubMed: 14351976.
2. Dahle UR, Eldholm V, Winje BA, Mannsåker T, Heldal E (2007) Impact of immigration on the molecular epidemiology of Mycobacterium tuberculosis in a low-incidence country. Am J Respir Crit Care Med 176: 930–935. Available: http://ajrccm.atsjournals.org/content/176/9/930.short. Accessed 29 January 2013 doi:10.1164/rccm.200702-187OC. PubMed: 17673698.
3. Svensson E, Millet J, Lindqvist A, Olsson M, Ridell M et al. (2011) Impact of immigration on tuberculosis epidemiology in a low-incidence country. Clin Microbiol Infect 17: 881–887. doi:10.1111/j.1469-0691.2010.03358.x. PubMed: 20825440.
4. Kamper-Jørgensen Z, Andersen AB, Kok-Jensen A, Bygbjerg IC, Andersen PH et al. (2012) Clustered tuberculosis in a low-burden country: nationwide genotyping through 15 years. J Clin Microbiol 50: 2660–2667. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3421533&tool=pmcentrez&rendertype=abstract. Accessed 26 March 2013
5. Mokrousov I, Vyazovaya A, Otten T, Zhuravlev V, Pavlova E et al. (2012) Mycobacterium tuberculosis population in northwestern Russia: an update from Russian-EU/Latvian border region. PLOS ONE 7: e41318. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3402494&tool=pmcentrez&rendertype=abstract. Accessed 14 January 2013 doi:10.1371/journal.pone.0041318. PubMed: 22844457.
6. Chernyaeva E, Dobrynin P, Pestova N, Matveeva N, Zhemkov V et al. (2012) Molecular genetic analysis of Mycobacterium tuberculosis strains spread in different patient groups in St. Petersburg, Russia. Eur J Clin Microbiol Infect Dis 31: 1753–1757. Available: http://www.ncbi.nlm.nih.gov/pubmed/22160887. Accessed 23 January 2013 doi:10.1007/s10096-011-1497-2. PubMed: 22160887.
7. Mäkinen J (2011) Extremely high prevalence of multidrug resistant tuberculosis in Murmansk, Russia: a population-based study. Eur J Clin Microbiol Infect Dis 30: 1119–1126. Available: http://www.springerlink.com/index/G277852522406822.pdf. Accessed 18 January 2013 PubMed: 21394425.
8. Vasankari T, Soini H, Liippo K, Ruutu P (2012) MDR-TB in Finland--still rare despite the situation in our neighbouring countries. Clin Respir J 6: 35–39. Available: http://www.ncbi.nlm.nih.gov/pubmed/21554562. Accessed 2 April 2013 doi:10.1111/j.1752-699X.2011.00242.x. PubMed: 21554562.
9. Puustinen K, Marjamäki M, Rastogi N, Sola C, Filliol I et al. (2003) Characterization of Finnish Mycobacterium tuberculosis Isolates by Spoligotyping. J Clin Microbiol 41: 1525–1528. doi:10.1128/JCM.41.4.1525. PubMed: 12682140.
10. Demay C, Liens B, Burguière T, Hill V, Couvin D et al. (2012) SITVITWEB--a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol 12: 755–766. Available: http://www.ncbi.nlm.nih.gov/pubmed/22365971. Accessed 26 March 2013 doi:10.1016/j.meegid.2012.02.004. PubMed: 22365971.
11. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D (2010) MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. Nucleic Acids Res 38: W326–W331. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2896200&tool=pmcentrez&rendertype=abstract. Accessed 1 July 2013 doi:10.1093/nar/gkq351. PubMed: 20457747.
12. Brudey K, Gordon M, Moström P, Jonsson B, Sola C et al. (2004) Molecular Epidemiology of Mycobacterium tuberculosis in Western Sweden Molecular Epidemiology of Mycobacterium tuberculosis in Western Sweden. J Clin Microbiol 42: 3046–3051. doi:10.1128/JCM.42.7.3046. PubMed: 15243058.
13. Braden C, Templeton G (1997) Interpretation of restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolates from a state with a large rural population. J Infect 175: 1446–1452. Available: http://jid.oxfordjournals.org/content/175/6/1446.short. Accessed 5 December 2013
14. Easterbrook PJ, Gibson A, Murad S, Lamprecht D, Ives N et al. (2004) High Rates of Clustering of Strains Causing Tuberculosis in Harare , Zimbabwe□: a Molecular Epidemiological Study. J Clin Microbiol 42: 4536–4544. doi:10.1128/JCM.42.10.4536. PubMed: 15472306.
15. Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S et al. (2012) Mycobacterium tuberculosis strains potentially involved in the TB epidemic in Sweden a century ago. PLOS ONE 7: e46848. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3466202&tool=pmcentrez&rendertype=abstract. Accessed 24 January 2013 doi:10.1371/journal.pone.0046848. PubMed: 23056484.
16. Kinander W, Bruvik T, Dahle UR (2009) Dominant Mycobacterium tuberculosis lineages in elderly patients born in Norway. PLOS ONE 4: e8373. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2791218&tool=pmcentrez&rendertype=abstract. Accessed 26 March 2013 doi:10.1371/journal.pone.0008373. PubMed: 20020043.
17. Mokrousov I (2012) The quiet and controversial: Ural family of Mycobacterium tuberculosis. Infect Genet Evol 12: 619–629. Available: http://www.ncbi.nlm.nih.gov/pubmed/22036706. Accessed 22 May 2013 doi:10.1016/j.meegid.2011.09.026. PubMed: 22036706.

Article 17

Infection, Genetics and Evolution 26 (2014) 58–64

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Four year longitudinal study of *Mycobacterium tuberculosis* complex isolates in a region of North-Eastern Italy

Loredana Fallico [a,b,1], David Couvin [c,1], Marta Peracchi [b], Michela Pascarella [d], Elisa Franchin [a,b], Enrico Lavezzo [a], Mario Rassu [d], Riccardo Manganelli [a,b,*], Nalin Rastogi [c], Giorgio Palù [a,b]

[a] Department of Molecular Medicine, University of Padova, Padova, Italy
[b] U.O. Microbiologia e Virologia, Azienda Ospedaliera di Padova, Padova, Italy
[c] Supranational TB Reference Laboratory, Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Abymes, Guadeloupe, France
[d] Department of Microbiology, S. Bortolo Hospital, Vicenza, Italy

ARTICLE INFO

Article history:
Received 21 January 2014
Received in revised form 30 April 2014
Accepted 4 May 2014
Available online 10 May 2014

Keywords:
Tuberculosis
Epidemiology
Spoligotyping
MIRU-VNTRs
Italy
Transmission

ABSTRACT

Recent reports have suggested a change of *Mycobacterium tuberculosis* complex genetic diversity in Western Europe due to an increasing proportion of imported cases of tuberculosis (TB). This study analyzed a total of 705 *M. tuberculosis* strains isolated from 2006 to 2009 in Veneto, a North-Eastern Italian region, to see the impact of foreign-born cases vs. Italian patients on prevailing TB epidemiology. Strains were genotyped using spoligotyping followed by comparison with international genotyping database SITVIT2. Six spoligotyping clusters with suspected phylogeographical specificity for imported cases, were typed by 15-loci MIRUs for a finer characterization. Overall, 410 (58.16%) strains were isolated from foreign-born patients, while 295 (41.84%) were isolated from Italian patients. Older patients (>70 years, *i.e.*, 46.4% of cases) predominated among Italians while younger age groups prevailed among foreign-born patients. Our results suggest that despite a high proportion of reactivation of latent TB infection in elderly Italian-born patients, active TB transmission between foreign-born and Italian patients may be ongoing, and argue in favor of an increased TB surveillance among immigrants to combat TB epidemic in Italy.
© 2014 Elsevier B.V. All rights reserved.

1. Introduction

A hundred and thirty-one years after the isolation of *Mycobacterium tuberculosis*, this important pathogen still represents one of the main scourges of humanity claiming more than one million deaths every year (Glaziou et al., 2013; Kaufmann, 2003; Zumla et al., 2013). It was earlier believed that tuberculosis (TB) resulted from the infection with a highly conserved strain of this microorganism; nonetheless recent studies clarified that *M. tuberculosis* has evolved in large, geographically constrained, large clonal populations of genotypically different lineages which might have different epidemiological and virulence impact (Gagneux et al., 2006; Gagneux and Small, 2007; Mathema et al., 2006; Rastogi and Sola, 2007).

This, together with the large human reservoirs of latent TB, makes TB epidemiology very complex and of fundamental importance to understand the development of the disease in the world. The epidemiology of TB is quite different in low- and high-income countries (Glaziou et al., 2013; Kaufmann, 2003; Zumla et al., 2013): its incidence is usually higher in the former due to poor living conditions with a majority of cases occurring among a relatively younger age group, while it is mostly restricted to elderly patients among high-income countries (Groenheit et al., 2012). Large movements of migrant populations from low- to high-income countries further complicated matters in the last decades by favoring the appearance of *M. tuberculosis* complex (MTBC) clones (and lineages) typical of low-income countries (Garcia de Viedma et al., 2011; Inigo et al., 2013). This reflects the Italian situation (Morandi et al., 2011) in which the notification rate of tuberculosis was 7.7 in 2007, 7.4 in 2008 and 7.1 in 2009 (with 4525, 4418 and 4244 cases, respectively). The amount of foreign-born patients however, increased in the period from 1999 to 2008 from 22% to 46%. In Veneto, a North-Eastern Italian region, the notification rate in the same 3 years was of 13.7, 12.3 and 11.2, respectively (Morandi et al., 2011).

Unfortunately, only a few studies are available regarding the *M. tuberculosis* population structure in Italy (Garzelli et al., 2010;

* Corresponding author at: Department of Molecular Medicine, University of Padova, Via Gabelli 63, 35121 Padova, Italy. Tel.: +39 049 827 2366.
E-mail address: riccardo.manganelli@unipd.it (R. Manganelli).
[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.meegid.2014.05.004
1567-1348/© 2014 Elsevier B.V. All rights reserved.

Lari et al., 2007; Sola et al., 2001, 2005). The aim of this study was therefore to analyze the genotypic structure of the MTBC strains isolated in Veneto, in order to evaluate the impact of foreign-born patients on prevailing TB epidemic, and to pinpoint probable cases of infection from the allochthonous to the autochthonous population or *vice versa*.

2. Materials and methods

2.1. Patients and data collection

The population studied included all patients with pulmonary or extrapulmonary TB who were reported to the Padova Hospital, Veneto Regional Reference Center for *Mycobacteria*, from January 1st 2006 to December 31st 2009. The MTBC isolates ($n = 705$ strains; 128 isolated in 2006, 177 in 2007, 192 in 2008, and 208 in 2009), were isolated from patients residing in the Veneto Region, including the provinces of Padova, Vicenza, Rovigo, Venice and Belluno. When sequential isolates were available for the same patient, only the first one was included in the study. The increase in numbers of the isolates from 2006 to 2009 may be partly explained by the improvement of the efficiency of the Regional Reference Center at Padova Hospital in collecting the isolates, and approached 100% in 2009. The routine demographic data collected included age, sex, race or ethnicity, country of birth and address at the time of diagnosis. Specific information concerning TB disease included the date of diagnosis, site or sites of disease, and results of microbiologic studies. Sampling and testing were strictly made as part of diagnosing the disease and drug resistance solely for patient benefit, and no sampling or investigations were oriented by research needs. For the present population-based genotyping study, the MTBC isolates were retrospectively collected; hence it was considered as a laboratory study and ethical approval from institutional IRB was not required. The correlation between genotypes and epidemiological and demographic data was performed in a de-identified manner. Confidentiality was further ensured by having dedicated health department personnel to remove all personal identification information prior to data analysis.

2.2. Routine bacterial identification and drug-susceptibility testing

Mycobacteria Growth Indicator Tube (MGIT), containing 7 ml of modified Middlebrook 7H9 Broth base, was used for the detection and recovery of MTBC isolates using the BACTEC MGIT960 (Becton Dickinson). Molecular identification was performed using Accuprobe MTBC culture identification test (Gen-Probe) and/or the GenoType Mycobacterium CM test (Hain Lifescience). First-line drug-susceptibility testing to streptomycin (STR, 1.0 μg/ml), isoniazid (INH, 0.1 μg/ml), rifampin (RIF, 1.0 μg/ml), ethambutol (EMB, 5.0 μg/ml) and pyrazinamide (PZA, 100 μg/ml) was performed using BACTEC MGIT960 according to manufacturer's instructions. *M. tuberculosis* ATCC 27294 was used as a control of susceptibility to all first-line anti-tubercular drugs. Note that resistance to any drug was reconfirmed by repetition; *e.g.*, in case of resistance to STR, INH or EMB, the susceptibility confirmation test was run using both the standard drug concentration and higher concentrations as follows: STR (4.0 μg/ml), INH0.4 μg/ml, and EMB (7.5 μg/ml). For all strains showing multidrug-resistance (MDR; *i.e.*, combined drug resistance to INH and RIF with or without other drugs), 2nd-line drugs were tested using the proportion method on Middlebrook 7H10 agar, following NCLS recommendations as follows: capreomycin 10 μg/ml; ethionamide 5 μg/ml; kanamycin 5 μg/ml; amikacin 4 μg/ml; ofloxacin 2 μg/ml; para-aminosalicylic acid 2 μg/ml; D-cycloserine 25 μg/ml; ciprofloxacin: 0.1 μg/ml.

2.3. Genotyping, database comparison, and phylogenetical analysis

For spoligotyping, 5–10 ng of genomic mycobacterial DNA were subjected to a previously described protocol to study polymorphism of the Direct Repeat (DR) locus (Kamerbeek et al., 1997). *M. tuberculosis* H37Rv and *Mycobacterium bovis* BCG Pasteur were used as reference strains. For database comparison and assignation of major phylogenetic clades, spoligotypes in octal format were entered in an Excel spreadsheet and compared to the international SITVIT2 proprietary database of the Pasteur Institute of Guadeloupe, which is an updated version of previously released SpolDB4 (Brudey et al., 2006) and SITVITWEB (Demay et al., 2012) databases. Note that with their definitive reclassification pending, we refer to LAM10-CAM (prototype SIT61) as the Cameroon lineage (Koro Koro et al., 2013), LAM7-TUR as the Turkey lineage (Kisa et al., 2012), and some spoligotypes previously classified among H3/H4 sublineages as H3/Ural-1 and H4/Ural-2 (Mokrousov, 2012).

For selected spoligotyping clusters, 15-loci MIRU-VNTR analysis was performed as reported earlier (Oelemann et al., 2007; Supply et al., 2006) using the publicly available protocol from the MIRU-VNTRplus website (http://www.miru-vntrplus.org/MIRU/files/MIRU-VNTRtypingmanualv6.pdf). In-house genotyping results were regularly validated using the commercially available MIRU-VNTR Typing Kit (GenoScreen), accordingly manufacturer's instruction. Amplicons size was determined by an automated ABI 3130XL DNA analyzer (Applied Biosystems) using GeneScan LIZ 1200 (Applied Biosystems) as size standard. The 15-loci MIRUs are provided in the order according to SITVIT2 database: MIRU-4, MIRU-10, MIRU-16, MIRU-26, MIRU-31, MIRU-40, ETR-A, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub30, Mtub39.

Worldwide distribution of all major spoligotyping clusters (defined as clusters containing at least isolates) was analyzed by interrogating the database for their distribution in macro-geographical regions and sub-regions according to the United Nations (http://unstats.un.org/unsd/methods/m49/m49regin.htm). Phylogenetical analysis was performed using the BioNumerics software version 6.6 (Applied Maths). Minimum spanning trees (MST) based on both spoligotyping and 15-loci MIRU typing patterns were drawn to highlight the links between the spoligotype families with different epidemiologic and demographic parameters. Lastly, a Spoligoforest tree was drawn as a hierarchical layout to illustrate probable parent to descendant relationships between spoligotypes (Reyes et al., 2008; Tang et al., 2008). GraphViz software (http://www.graphviz.org) (Ellson et al., 2002) was used to color the strains based on their lineages on the Spoligoforest trees.

2.4. Statistical analysis

STATA statistical software version 12 and MegaStat12 were used for both descriptive and univariate analyses. Comparison of lineages and SITs with demographic and epidemiological parameters were calculated using Pearson's Chi-square test when more than 80% of data had a value greater than 5 and Fisher's Exact Test for remaining data with smaller values (at least 20% of data having values less than 5). *p*-values between 0.04 and 0.06 were considered as marginally significant while *p*-values <0.04 were considered as statistically significant. Odds Ratios (OR) and 95% Confidence Interval were added to better certify the significance of observed differences. Note that only groups counting number of strains >30 were taken into account to calculate the *p*-values and the odds ratios.

3. Results

3.1. Patient population

During the period spanning from 2006 to 2009, we characterized a total of 705 *M. tuberculosis* complex strains isolated in different areas of Veneto, a region of North-Eastern Italy. Available demographic and epidemiological information regarding patient population is summarized in Table S1, and graphically represented in Figs. 1 and S1. One hundred twenty-eight strains were collected in 2006, 177 in 2007, 192 in 2008 and 208 in 2009. The ratio between Italian-born and foreign-born patients was 0.72, while the male/female ratio was 1.34. Regarding the age distribution in the two subpopulations, the majority of Italian-born patients were >60 years (63.75% of patients), while the majority of foreign-born patients were significantly younger (range 21–40 years for 71% of patients; Fig. 1A). The ratio of isolates resistant to at least one first line drug was 14.2% in 2006, 10.9% in 2007, 12.7% in 2008 and 18.4% in 2009; while that of MDR strains was 3.9% in 2006, 1.2% in 2007, 3.2% in 2008 and 3.9% in 2009 (Table S1). The distribution of resistant and sensitive strains between Italian and foreign born patients did not vary significantly (Fig. 1B). Lastly, most of the foreign-born patients originated from 4 macro-regions: 32.1% were from EURO-E, mostly from Romania, 17.1% from AFRI-N, mostly from Morocco, 14.4% from ASIA-S, mostly from India, and 14.2% from AFRI-W (Table S1, Fig. S1).

3.2. Spoligotyping analysis and worldwide distribution

Spoligotyping of the 705 strains lead to 244 distinct patterns (Tables S2 and S3); Among the 81 clusters observed, clusters containing >1% of strains, *i.e.*, containing 7 or more isolates ($n = 14$ clusters containing 343 isolates; 7–89 isolates per cluster, corresponding to 343/542 or 63.3% of clustered isolates) were considered as major spoligotype patterns, and their worldwide distribution in the SITVIT2 database in the updated database is summarized in Table S4. Among them the most represented SITs where (i) SIT53/T1, $n = 89$ strains; (ii) SIT42/LAM9, $n = 44$; (iii) SIT47/H1, $n = 40$ isolates; (iv) SIT50/H3, $n = 37$; (v) SIT1/Beijing, $n = 29$; and (vi) SIT26/CAS1-Delhi, $n = 22$.

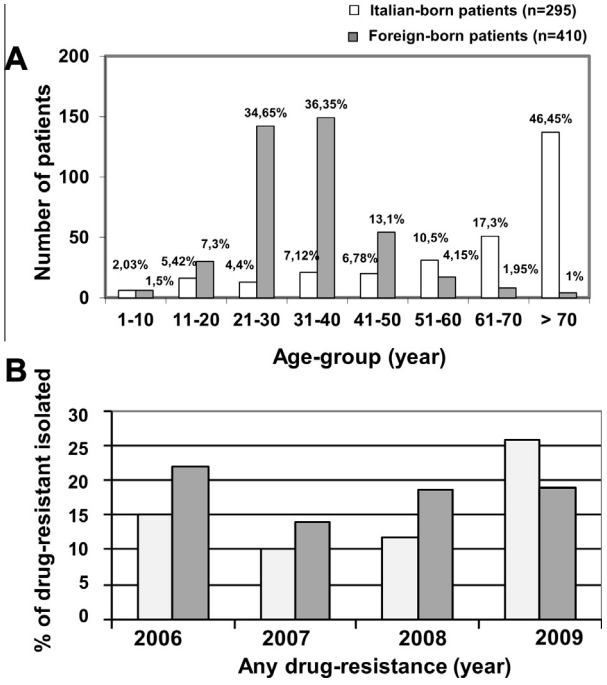


Fig. 1. Graphic representation of age groups and drug-resistance distribution in the patient population and origin of the foreign-born patients. (A) Distribution of age groups between Italian and foreign born patients, percent in specific population is reported. (B) Proportion of any drug resistant by year of isolation and origin of patients.

3.3. Phylogenetical analysis

A finer analysis of evolutionary relationship of the *M. tuberculosis* spoligotypes identified in our study sample is illustrated in the minimum spanning tree (MST) shown in Fig. 2A. The phylogenetic tree connects each genotype based on degree of changes required to go from one allele to another. This figure clearly shows that most of the ongoing tuberculosis epidemic in Veneto is organized around selected major SITs (Table S4), each being characterized by its prototype in the central node (with the exception of SIT62/H1 and SIT2754/T1), and followed by evolutionary related minor SITs in its vicinity. This observation about the importance of prevalent strains with well-characterized prototype signatures in TB epidemic in Veneto is further supported by the Spoligoforest tree shown in Fig. S2.

3.4. Correlation of phylogenetic, demographic, and drug-resistance data

A superposition of phylogenetic trees with demographic and drug-resistance data is illustrated in Fig. 2B–D; with the exception of SIT34/S family which was more frequent in female patients (marginally significant; p-value 0.0545 as compared to the most predominant SIT53/T1), gender-dependent distribution did not vary significantly for other spoligotypes (Fig. 2B). Regarding origin of patients, SIT34/S and SIT482/BOV_1 were more frequent among patients with Italian origin, as opposed to strains belonging to SIT1/Beijing, SIT26/CAS1-Delhi, SIT37/T3, SIT61/Cameroon, SIT47/H1 and SIT42/LAM9, that were more frequently isolated from foreign-born patients (p-value <0.00001; Fig. 2C). Lastly, no particular linkage between predominant SITs and drug resistance was found, however, one may note that an important proportion of MDR-TB was found in strains belonging to SIT1/Beijing and SIT26/CAS1-Delhi lineages as compared with most predominant SIT53/T1 (p-value <0.001; Fig. 2D). All *M. bovis* strains were found to be monoresistant to pyrazinamide as expected.

The origin of patients and frequency of isolation in various collection centers is summarized in Table 1. Interestingly, while some SITs were similarly distributed over the territory, some had a characteristic distribution pattern such as the SIT1/Beijing which was isolated with higher frequency in Rovigo and SIT26/CAS1-Delhi in Vicenza. Indeed, one could almost conclude on the phylogeographical specificity of the major SITs simply looking at Table 1. Interestingly, certain lineages such as Haarlem, LAM and CAS were similarly shared between foreign-born and Italian whereas S lineage prevailed significantly among Italians patients. This may indirectly suggest their active transmission between these communities, even if further typing using 24-loci MIRU or whole genome sequencing will be required to confirm this hypothesis. On the other hand some lineages have remained confined to the foreign-born ethnical groups; this is particularly true for SIT61/Cameroon and SIT262/Ural-1.

When the isolation site was taken into consideration, we found that BOVIS, CAS, and EAI lineages were significantly associated with extrapulmonary TB as compared to pulmonary disease (Table S5). Furthermore, as compared to the major SIT53/T1, we found that SIT4/Unknown SIT26/CAS1-Delhi and SIT482/BOV_1 were prevalently associated to extrapulmonary disease (Table S6).

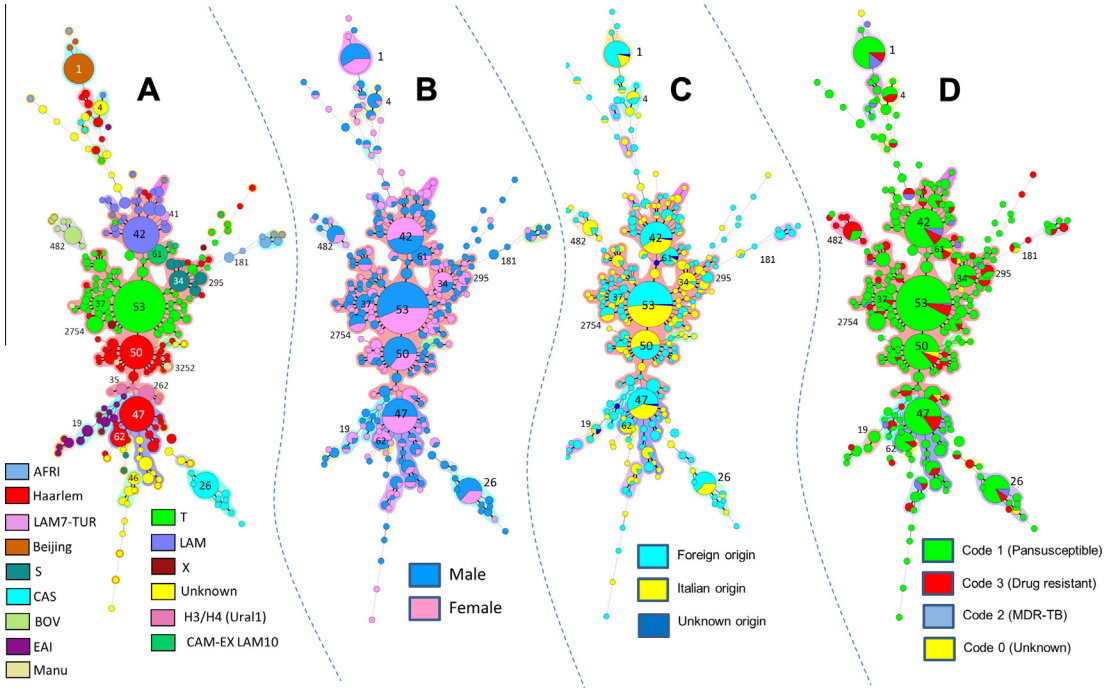


Fig. 2. A minimum spanning tree (MST) illustrating evolutionary relationships between the *M. tuberculosis* genotypes isolated in Italy. (A) MST constructed on all isolates including the orphan patterns ($n = 705$); (B) MST showing the distribution of sex ratio among all patterns; (C) MST showing the distribution regarding the origin of patient; (D) MST showing the distribution regarding the information of drug resistance for all spoligotypes. The phylogenetic tree connects each genotype based on degree of changes required to go from one allele to another. The structure of the tree is represented by branches (continuous vs. dashed and dotted lines) and circles representing each individual pattern. Note that the length of the branches represents the distance between patterns while the complexity of the lines (continuous, gray dashed and gray dotted) denotes the number of allele/spacer changes between two patterns: solid lines, 1 or 2 or 3 changes (thicker ones indicate a single change, while the thinner ones indicate 2 or 3 changes); gray dashed lines represent 4 changes; and gray dotted lines represent 5 or more changes. The size of the circle is proportional to the total number of isolates in our study, illustrating unique isolates (smaller nodes) versus clustered isolates (bigger nodes). The color of the circles indicates the phylogenetic lineage to which the specific pattern belongs. Note that orphan patterns are circled in orange in Fig. 2A. Patterns colored in yellow indicate a strain with an unknown signature (unclassified). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To look at the bigger picture we also analyzed the data considering the major *M. tuberculosis* lineages irrespective of SIT labeling: although only 3 predominant families (T, Harlem, and LAM) included over 68% of the isolates; out of 15 groups analyzed, statistically significant differences in distribution among Italian-born vs. foreign-born patients was found for 8 lineages (Table 2). Finally, we also studied the phylogeographical distribution of the major *M. tuberculosis* lineages detected in our study according to patient origin by mapping all the isolates of this study vs. those present in the international database SITVIT2. The justification for this comparison came from our previous findings that imported cases of active TB or those assumed to result from reactivation of latent TB infection in foreign-born patients are easily reflected by the global distribution of major *M. tuberculosis* lineages in the global database (Vanhomwegen et al., 2011). As shown in Fig. S4A, the data underlines that almost all immigrants were infected by *M. tuberculosis* genotypic lineages prevailing in their country of origin, the only exception being the presence of a significant proportion of S-family strains among immigrants from South America and Caribbean since such strains were not previously reported in these regions. Interestingly, the proportion of strains belonging to CAS lineage has doubled in this study as compared to previous data of strains isolated in Italy in SITVIT2 database (Fig. S4B).

3.5. 15-Loci MIRU typing of spoligotyping clusters with high phylogeographical specificity for imported TB cases

15-Loci MIRUs analyses are considered enough discriminatory for epidemiological purposes, especially when used in combination with spoligotyping. (Supply et al., 2006; Alonso-Rodriguez et al., 2008), even if the possibility to slightly overestimated recent transmission rate is present. So, we further typed 6 spoligotyping clusters with proven phylogeographical specificity for imported cases of disease ($n = 83$; SIT1/Beijing, SIT19/EAI2-Manila, SIT26/CAS1-Delhi, SIT34/S, SIT61/LAM10-CAM, and SIT181/AFRI_1) by 15-loci MIRUs for a finer characterization of transmission links (Alonso-Rodriguez et al., 2008). We also selected an essentially Italian *M. bovis* cluster (SIT482/BOV_1, $n = 10$) since spoligotyping alone has been reported to be of poorly discriminative for *M. bovis* strains (Boniotti et al., 2009). The detailed results are summarized in Table S7, and show that starting from 7 clusters (clusters A to G, $n = 93$ strains; 3–29 isolates per cluster), only 6 clusters ($n = 15$ strains; 2–4 strains per cluster) contained identical MIRU and Spoligotype profiles:

(i) Cluster A1 (MIRU pattern: 233753446822543/MIT259) corresponding to 3 SIT1/Beijing strains, all isolated from Chinese patients in 2006 from Vicenza; (ii) Cluster A2 (MIRU pattern: 233563446724543/MIT266) also corresponding to 2 SIT1/Beijing strains that were isolated from an Italian and a Moldavian patient in 2007 from Vicenza; (iii) Cluster C1 (MIRU pattern: 264753222853422/MIT268) corresponding to 2 SIT26/CAS1-Delhi strains isolated from an Italian and an Indian patient in 2008 from Vicenza; (iv) Cluster D1 (MIRU pattern 333534343143123/MIT260) corresponding to 4 SIT34/S-family strains isolated exclusively from Italian patients between 2006 and 2009 from Vicenza; (v) Cluster E1 (MIRU pattern 233531446522322/MIT265) corresponding to 2 SIT61/Cameroon strains isolated from a Liberian and a Nigerian patients from Padova and Rovigo, respectively, and (vi) Cluster F1 (MIRU pattern 272451754434444/MIT264) corresponding to 2 SIT181/AFRI_1 strains isolated from an Italian and a Ghanaian patient from Padova and Vicenza, respectively.

Table 1
Origin of patients for major SITs characterized, and frequency of isolation (%) in various collection centers.

SIT/Lineage	Country of origin of patients[a]	Italians (%)	% Italians/ Europeans	Proportion (%) in different collection centers					
				Camposampiero (*n* = 66)	Conegliano (*n* = 45)	Mestre (*n* = 60)	Padova (*n* = 205)	Rovigo (*n* = 59)	Vicenza (*n* = 146)
1/Beijing (*n* = 29)	**CHN 10**, IND 1, ITA 5, **MDA 9**, PAK 1, PER 1, UKR 1, NA 1	17.24	33.3	1.5	4.4	5.0	3.9	10.2	4.8
4/Unknown (*n* = 7)	ALB 1, **ITA 4**, MAR 1, SEN 1	57.10	80	0.0	0.0	0.0	1.9	0.0	0.7
26/CAS1-Delhi (*n* = 22)	BGD 2, **IND 9**, **ITA 8**, PAK 3	36.40	100	0.0	0.0	1.7	2.9	0.0	9.6
34/S Family (*n* = 14)	GRC 1, **ITA 13**	92.80	92.3	3.0	2.2	1.7	1.5	0.0	4.8
37/T3 (*n* = 11)	BGD 1, **ITA 3**, MAR 1, PAK 1, **ROM 4**, SOM 1	27.30	42.8	0.0	0.0	0.0	3.9	0.0	2.0
42/LAM9 (*n* = 44)	AFG 1, BRA 1, CUB 1, HND 1, **ITA 17**, **MAR 17**, NGA 1, ROM 1, SEN 2, UKR 1, NA 1	36.40	89.4	7.6	11.1	6.7	4.4	3.4	4.8
47/H1 (*n* = 40)	ALB 1, DZA 1, **ITA 16**, MAR 4, MDA 2, MKD 1, **ROM 12**, SCG 1,TUN 1, NA 1	40.00	48.5	10.6	2.2	8.3	4.4	5.8	4.4
50/H3 (*n* = 37)	BIH 2, **ITA 20**, **MAR 4**, MDG 1, MKD 1, POL 1, **ROM 5**, SCG 1, SEN 1, TGO 1	54.00	66.7	7.7	13.3	5.0	3.9	5.1	4.1
53/T1 (*n* = 89)	BGD 1, CIV 2, ECU 1, ETH 1, GHA 1, IND 2, **ITA 42**, **MAR 10**, **ROM 24**, SEN 2, TUN 1, TZA 1, NA 1	47.20	63.6	15.1	11.1	5.0	14.1	20.3	9.6
61/Cameroon (*n* = 13)	BFA 1, GHA 1, ITA 1, LBR 1, **NGA 7**, SEN 1, NA 1	7.70	100	3.0	0.0	0.0	2.4	5.1	2.1
62/H1 (*n* = 11)	COL 1, **ITA 5**, **ROM 4**, SEN 1	45.40	55.5	3.0	0.0	3.3	0.5	0.0	1.4
262/Ural-1 (*n* = 8)	MDA 2, **ROM 6**	0.00	NA	0.0	2.2	3.3	0.5	3.4	1.4
482/BOV_1 (*n* = 10)	**ITA 8**, MAR 1, TUN 1	80.00	100	6.1	2.2	0.0	1.5	0.0	0.0
2754/T1 (*n* = 9)	ALB 1, **ITA 4**, NGA 1, PAK 1, ROM 2	44.40	57.1	0.0	0.0	0.0	3.4	0.0	1.4

NA, not available.
[a] The 3 letter country codes are according to http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3; the names of countries of origin of a majority of patients are highlighted in bold.

These data were then used to calculate the rate of recent TB transmission with the "n-1" method (Small et al., 1994), which resulted in a relatively low transmission rate of 9.7%. However, since the molecular clock of MIRUs is considerably faster than that of spoligotyping, single locus variants (SLVs) may arise during recent TB transmission both due to rare and stochastic mutational events and inherent genetic drift in clonal populations (Supply et al., 2006), hence it is accepted today to consider a SLV as being clustered provided that 2 strains share the same spoligotype pattern. Using this parameter, we calculated the rate of recent transmission at 22.6%.

4. Discussion

This work represents the first survey of the *M. tuberculosis* strains circulating in Veneto, a North-Eastern Italian region during a 4 year period (from 2006 to 2009); a total of 705 strains were collected in our Center from the hospitals of the cities of Padova, Vicenza, Rovigo, Mestre (Venice) and of minor centers of their hinterlands. As expected for a low tuberculosis incidence country with active immigration from medium and high incidence countries, most of the Italian-born patients were >60 years old, clearly suggesting reactivation of latent disease, while most of the foreign born patients were significantly younger suggesting actively transmitted disease acquired in their countries of origin.

The results obtained during this investigation further established that most of the ongoing TB epidemic in Veneto is organized around selected major SITs, each being characterized by its prototype in the central node (with the exception of SIT62/H1 and SIT2754/T1), and followed by evolutionary related minor SITs in its vicinity (Table 2A and Fig. S2). A temporal analysis of the SITs isolated during the 4 year period evidenced that while some SITs were stably present in the region (as SIT42, SIT47, and SIT50), other SITs had a positive trend, as SIT62 which represented 0.8% and 0.6% of the isolates in the first two years (1 isolate/year), rising to 1.6% (3 isolates) and 2.3% (6 isolates) in the last two years (Table S1). Similarly, SIT262 was not collected in the first two years and represented 0.5% of the isolates in 2008 (1 isolate) and 3.4% of the isolates in 2009 (7 isolates). It will be interesting to see if this trend will continue or not in the next years. If the distribution of the spoligotypes was considered according to the city of isolation, some SITs were almost equally distributed in the territory (as SIT42 or SIT50), while others were unequally distributed such as SIT1/Beijing which was slightly more represented in Rovigo, or SIT26/CAS1-Delhi that was exclusively present in Vicenza reflecting the presence of large Chinese and Indian communities in the two cities, respectively.

Finally, we also analyzed the correlation between SITs and site of isolation and found that SIT4/unknown, SIT26/CAS1-Delhi and SIT482/BOVIS_1 were significantly less frequently associated with respiratory samples. While association of MTBC isolates of BOVIS and CAS genotypic lineages with extrapulmonary disease was previously reported (Lari et al., 2009), this is the first time that we document it for SIT4. Further work involving higher number of patients and isolates is needed to confirm these findings.

The presence of several SIT482/BOV_1 (BCG) infections in Italians, can be explained due to BCG-treatment of bladder-cancer patients: BCG Immunotherapy by intra-vesicular delivery of *M. bovis* BCG has been established as the most effective adjuvant treatment for preventing local recurrences and tumor progression in bladder cancer patients for the last more than three decades (Alexandroff et al., 1999; Brandau and Suttmann, 2007), since the foreign born population living in Italy has a low median age, it is possible to hypothesize that the proportion of individuals suffering from bladder cancer in this part of the population is lower than that found in Italians. However, this hypothesis remains to be validated since the history of these patients is not known.

A comparison of our data with worldwide phylogeographic distribution of *M. tuberculosis* lineages using the SITVIT2 database established that most of the immigrants were infected by

Table 2
Distribution and comparison of the proportion of *M. tuberculosis* lineages among Italian-born and Foreign-born patients.

Lineages	Italian-born	Foreign-born	*p*-values	Odds ratios [95% CI]	Total (%)
T	97[a] (33)[b]	100 (24)	0.013*	1.52 [1.07; 2.14]	197 (28)
Harlem	70 (24)	91 (22)	0.47	1.09 [0.75; 1.58]	161 (23)
Ural	0 (0.00)	16 (4)	NA	NA	16 (2)
LAM	37 (13)	50 (12)	0.17	1.03 [0.64; 1.66]	87 (12)
Cameroon	1 (0.4)	15 (4)	NA	NA	16 (2)
Turkey	0 (0)	3 (0.1)	NA	NA	3 (0.4)
Unknown	30 (10)	33 (8)	0.33	1.29 [0.74; 2.25]	63 (9)
CAS	11 (4)	24 (6)	0.20	0.62 [0.27; 1.35]	35 (5)
S	25 (8)	8 (2)	<0.0001*	4.64 [1.99; 12.09]	33 (5)
EAI	3 (1.)	22 (5)	NA	NA	25 (4)
Beijing	5 (2)	26 (6)	0.003*	0.26 [0.08; 0.69]	31 (4)
BOV	10 (3)	6 (1)	NA	NA	16 (2)
AFRI	2 (0.7)	13 (3)	NA	NA	15 (2)
MANU	4 (1)	0 (0)	NA	NA	4 (0.6)
X	0 (0)	3 (0.1)	NA	NA	3 (0.4)

[a] Total number of isolates.
[b] % of the isolates.
* Asterisk denotes statistically significant differences ($p < 0.05$); Pearson's Chi-square test was used when more than 80% of data had an expected value greater than 5, and Fisher's Exact Test for remaining data with smaller values (at least 20% of data having expected values less than 5). Note that only lineages having number of strains >30 were taken into account to calculate the *p*-values and the odds ratios (lineages having number of strains lesser than 30 were notified NA, not applicable).

M. tuberculosis lineages prevailing in their country of origin. This observation suggests that most of the foreign-born TB patients in Veneto were infected prior to their arrival in Italy.

Understandably, the distinction between Euro-American phylogeographical lineage strains among Italian patients vs. "foreign-origin" strains essentially imported from Asia and Africa in our study, was very straightforward and allowed to easily trace phylogeographical distinction of strains causing the disease in Italian vs. foreign-born patients (Tables 1 and 2). However, if the isolates of T, Harlem, Ural, LAM, Cameroon, Turkey, S and X families were grouped together as a larger Euro-American group of strains (Gagneux et al., 2006; Gagneux and Small, 2007; Rastogi and Sola, 2007), these were almost similarly prevalent in Italian and foreign born patients (78.0% and 69.8%, respectively). Interestingly, in a previous study performed in Tuscany a region of central Italy, the prevalence of Euro-American isolates was 77.7% in Italian born patients vs. 57.9% in foreign-born patients (Garzelli et al., 2010). Moreover, the proportion of TB cases due to MTBC genotypic families associated with distant geographical areas, *i.e.*, EAI, Beijing, CAS, and *Mycobacterium africanum* that accounted for 3.9% of Italian-born TB cases in Tuscany (Garzelli et al., 2010) has almost doubled among Italians in our study to 21/295 or 7.12% (Table 2). Whether this observation might be explained on the basis of the fact that waves of immigration to Veneto started before Tuscany and/or higher numbers of immigrants and TB transmission rates remains speculative until further investigations evaluate effect of immigration in changing MTBC genotypic diversity in Italy.

Fifteen-loci MIRU typing of 6 spoligotyping clusters with proven phylogeographical specificity showed that certain *M. tuberculosis* lineage clones are shared among foreign-born and Italian patients in Veneto region suggesting the possibility of ongoing active transmission between these 2 subpopulations (Table S7). On the other hand, other lineage clones have remained confined to the foreign-born ethnical groups without evidence of transmission to Italian patients (*e.g.*, SIT1/Beijing cluster A1 among 3 Chinese patients), and *vice versa* (SIT34/S-family cluster D1 among 4 Italian patients, out of which 3 young Italian patients were attending the same school).

In summary, although the 15-loci MIRU typing resulted in a relatively low transmission rate of 9.7%; considering that SLVs may frequently arise during recent TB transmission due to a considerably faster molecular clock of MIRUs than that of spoligotyping (Garcia de Viedma et al., 2011; Rastogi and Sola, 2007; Supply et al., 2006), we can hypothesize a rate of recent transmission of 22.6%. However, since this typing scheme might slightly overestimate the rate of transmission, these data need to be confirmed in future studies by comparing clusters retained after 15-loci MIRUs using 24-loci MIRUs (Supply et al., 2006) in conjunction with whole genome sequencing (Rastogi, 2013; Roetzer et al., 2013) in order to shed light on epidemiological and phylogenetical links between strains involved in active transmission.

5. Conclusions

We provided here the first analyses of the genetic population structure of the *M. tuberculosis* strains circulating in the Veneto region, Italy. Despite a very low recent transmission rate and a high proportion of reactivation of latent TB infection in elderly Italian-born patients, the data obtained suggest active TB transmission between foreign-born and Italian patients. This investigation represents a first step towards the control of TB infection in Veneto, Italy by helping to design future control strategies involving an increased TB surveillance among immigrants. Future prospective studies should ideally involve analysis of epidemiologic, demographic and clinical data in conjunction with next generation genome sequencing to quantify how much MTBC isolates vary at the genomic level between epidemiologically linked patient clusters (Rastogi, 2013).

Acknowledgements

The authors wish to thank F. Alborino, L. Bacelle, S. Ciaffoni, G. Da Rin, S. Grandesso, F. Manoni, T. Roncada, D. Rubin, M. Scarin, and. D. Signori for the help in collecting the isolates; C. Olivato, N. Lari and C. Garzelli for assistance in the establishment of the spoligotyping technique in our lab; L. Barzon and F. Russo for support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2014.05.004.

References

Alexandroff, A.B., Jackson, A.M., O'Donnell, M.A., James, K., 1999. BCG immunotherapy of bladder cancer: 20 years on. Lancet 353, 1689–1694.

Alonso-Rodriguez, N., Martinez-Lirola, M., Herranz, M., Sanchez-Benitez, M., Barroso, P., Bouza, E., Garcia de Viedma, D., 2008. Evaluation of the new advanced 15-loci MIRU-VNTR genotyping tool in *Mycobacterium tuberculosis* molecular epidemiology studies. BMC Microbiol. 8, 34.

Boniotti, M.B., Goria, M., Loda, D., Garrone, A., Benedetto, A., Mondo, A., Tisato, E., Zanoni, M., Zoppi, S., Dondo, A., Tagliabue, S., Bonora, S., Zanardi, G., Pacciarini, M.L., 2009. Molecular typing of *Mycobacterium bovis* strains isolated in Italy from 2000 to 2006 and evaluation of variable-number tandem repeats for geographically optimized genotyping. J. Clin. Microbiol. 47, 636–644.

Brandau, S., Suttmann, H., 2007. Thirty years of BCG immunotherapy for non-muscle invasive bladder cancer: a success story with room for improvement. Biomed. Pharmacother. 61, 299–305.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Demay, C., Liens, B., Burguiere, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., Rastogi, N., 2012. SITVITWEB – a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect. Genet. Evol. 12, 755–766.

Ellson, J., Gansner, E., Koutsofios, L., North, S.C., Woodhull, G., 2002. Graphviz – open source graph drawing tools. In: Mutzel, P., Jünger, M., Leipert, S. (Eds.), Graph Drawing. Springer, Berlin Heidelberg, pp. 483–484.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U.S.A. 103, 2869–2873.

Gagneux, S., Small, P.M., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet. Infect. Dis. 7, 328–337.

Garcia de Viedma, D., Mokrousov, I., Rastogi, N., 2011. Innovations in the molecular epidemiology of tuberculosis. Enferm. Infecc. Microbiol. Clin. 29 (Suppl 1), 8–13.

Garzelli, C., Lari, N., Cuccu, B., Tortoli, E., Rindi, L., 2010. Impact of immigration on tuberculosis in a low-incidence area of Italy: a molecular epidemiological approach. Clin. Microbiol. Infect. 16, 1691–1697.

Glaziou, P., Falzon, D., Floyd, K., Raviglione, M., 2013. Global epidemiology of tuberculosis. Semin. Respir. Crit. Care Med. 34, 3–16.

Groenheit, R., Ghebremichael, S., Pennhag, A., Jonsson, J., Hoffner, S., Couvin, D., Koivula, T., Rastogi, N., Kallenius, G., 2012. *Mycobacterium tuberculosis* strains potentially involved in the TB epidemic in Sweden a century ago. PloS one 7, e46848.

Inigo, J., Garcia de Viedma, D., Arce, A., Palenque, E., Herranz, M., Rodriguez, E., Ruiz-Serrano, M.J., Bouza, E., Chaves, F., 2013. Differential findings regarding molecular epidemiology of tuberculosis between two consecutive periods in the context of steady increase of immigration. Clin. Microbiol. Infect. 19, 292–297.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Kaufmann, S.H., 2003. A short history of Robert Koch's fight against tuberculosis: those who do not remember the past are condemned to repeat it. Tuberculosis (Edinb) 83, 86–90.

Kisa, O., Tarhan, G., Gunal, S., Albay, A., Durmaz, R., Saribas, Z., Zozio, T., Alp, A., Ceyhan, I., Tombak, A., Rastogi, N., 2012. Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey. PloS One 7, e30331.

Koro Koro, F., Kamdem Simo, Y., Piam, F.F., Noeske, J., Gutierrez, C., Kuaban, C., Eyangoh, S.I., 2013. Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping. J. Clin. Microbiol. 51, 299–302.

Lari, N., Rindi, L., Bonanni, D., Rastogi, N., Sola, C., Tortoli, E., Garzelli, C., 2007. Three-year longitudinal study of genotypes of *Mycobacterium tuberculosis* isolates in Tuscany, Italy. J. Clin. Microbiol. 45, 1851–1857.

Lari, N., Rindi, L., Cristofani, R., Rastogi, N., Tortoli, E., Garzelli, C., 2009. Association of *Mycobacterium tuberculosis* complex isolates of BOVIS and Central Asian (CAS) genotypic lineages with extrapulmonary disease. Clin. Microbiol. Infect. 15, 538–543.

Mathema, B., Kurepina, N.E., Bifani, P.J., Kreiswirth, B.N., 2006. Molecular epidemiology of tuberculosis: current insights. Clin. Microbiol. Rev. 19, 658–685.

Mokrousov, I., 2012. The quiet and controversial: ural family of *Mycobacterium tuberculosis*. Infect. Genet. Evol. 12, 619–629.

Morandi, M., Resi, D., Morsillo, F., Moro, M.L., D'Amato, S., Iannazzo, S., Rizzuto, E., Pompa, M.G., Fattorini, L., Suligoi, B., 2011. La tubercolosi in Italia – Rapporto 2008. Ufficio V Malattie infettive, Direzione generale della prevenzione sanitaria (Ministero della salute), Istituto superiore di sanità, and Agenzia sanitaria e sociale della Regione Emilia-Romagna, Italy. <http://www.salute.gov.it/imgs/c_17_pubblicazioni_1472_allegato.pdf>.

Oelemann, M.C., Diel, R., Vatin, V., Haas, W., Rusch-Gerdes, S., Locht, C., Niemann, S., Supply, P., 2007. Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J. Clin. Microbiol. 45, 691–697.

Rastogi, N., 2013. F1000Prime Recommendations, Dissents and Comments for [Walker, T.M., et al., 2013. Lancet Infect. Dis. 13(2), 137–146], F1000Prime.

Rastogi, N., Sola, C., 2007. Molecular evolution of the *Mycobacterium tuberculosis* complex. In: Palomino, J.C., Leao, S., Ritacco, V. (Eds.), Tuberculosis 2007. Amedeo Online Textbooks.

Reyes, J.F., Francis, A.R., Tanaka, M.M., 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 9, 496.

Roetzer, A., Diel, R., Kohl, T.A., Ruckert, C., Nubel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rusch-Gerdes, S., Supply, P., Kalinowski, J., Niemann, S., 2013. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med. 10, e1001387.

Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., Schoolnik, G.K., 1994. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N. Engl. J. Med. 330, 1703–1709.

Sola, C., Ferdinand, S., Mammina, C., Nastasi, A., Rastogi, N., 2001. Genetic diversity of *Mycobacterium tuberculosis* in Sicily based on spoligotyping and variable number of tandem DNA repeats and comparison with a spoligotyping database for population-based analysis. J. Clin. Microbiol. 39, 1559–1565.

Sola, C., Ferdinand, S., Sechi, L.A., Zanetti, S., Martial, D., Mammina, C., Nastasi, A., Fadda, G., Rastogi, N., 2005. *Mycobacterium tuberculosis* molecular evolution in western Mediterranean Islands of Sicily and Sardinia. Infect. Genet. Evol. 5, 145–156.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

Tang, C., Reyes, J.F., Luciani, F., Francis, A.R., Tanaka, M.M., 2008. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. Bioinformatics 24, 2414–2415.

Vanhomwegen, J., Kwara, A., Martin, M., Gillani, F.S., Fontanet, A., Mutungi, P., Crellin, J., Obaro, S., Gosciminski, M., Carter, E.J., Rastogi, N., 2011. Impact of immigration on the molecular epidemiology of tuberculosis in Rhode Island. J. Clin. Microbiol. 49, 834–844.

Zumla, A., Raviglione, M., Hafner, R., von Reyn, C.F., 2013. Tuberculosis. N. Engl. J. Med. 368, 745–755.