N° d'ordre : 2012-36 Année 2012

ÉCOLE CENTRALE DE LYON

THÈSE

pour obtenir le grade de

DOCTEUR

Spécialité : « Microbiologie environnementale »

préparée dans le laboratoire Ampère

dans le cadre de l'Ecole Doctorale « Électronique, Électrotechnique et Automatique »

par

Samuel Jacquiod

Métagénomique et approches alternatives pour l'étude fondamentale et l'exploitation de la microflore tellurique

soutenue publiquement le 12 novembre 2012 devant la commission d'examen

JURY

Rapporteur	Gabrielle Potocki-Veronese	Chargée de Recherche INRA - INSA de Toulouse
Rapporteur	Ines Mandic Mulec	Professeur - Université de Ljubljana (Slovénie)
Examinateur	Elizabeth Wellington	Professeur - Université de Warwick (Royaume-Uni)
Examinateur	Mohamed Jebbar	Professeur - Université de Brest
Examinateur	Laurent Philippot	Directeur de Recherche INRA - INRA de Dijon
Encadrant	Laure Franqueville	Ingénieur de Recherche CNRS - Ecole Centrale de Lyon
Directeur de thèse	Pascal Simonet	Directeur de Recherche CNRS - Ecole Centrale de Lyon
Président du jury	Timothy M. Vogel	Professeur - Université Claude Bernard Lyon 1

"No! Try not! Do or do not... there is no try."

Master Yoda (Star Wars)

« Non! N'essaie pas ! Fais le ou ne le fais pas… il n'y a pas d'essai. »

Maitre Yoda (Star Wars)

Remerciements

Avant toute chose, je tiens à remercier sincèrement toutes les personnes, qui de près ou de loin ont participé, contribué, aidé et soutenu ce projet de recherche dans lequel je me suis investi durant ces années.

Je remercie en premier lieu mon directeur de thèse, Pascal Simonet, pour m'avoir donné ma chance et permis de réaliser cette grande expérience, qu'est la préparation du doctorat, dans de bonnes conditions. Je le remercie aussi pour m'avoir attribué sa confiance et son soutien tout au long du parcours, dans les bons comme dans les mauvais moments. Son expérience et sa vision de la science ont été pour moi autant une source d'inspiration que de motivation et d'encouragement tout au long de cette expérience.

Je remercie le comité de thèse, les rapporteurs : Gabrielle Véronèse et Ines Mandic Mulec, et les examinateurs: Elizabeth Wellington, Mohamed Jebbar et Laurent Philippot, pour avoir bien voulu accorder de leurs temps et de leur personne à la lecture ainsi qu'a l'évaluation de ce travail.

Je remercie aussi Timothy Vogel, qui m'a offert l'opportunité d'intégrer l'équipe après mon cursus universitaire. Je tiens également à remercier toutes les personnes qui se sont investie personnellement dans le projet Genefish et que j'ai eu la chance de pouvoir côtoyer : Laurent Philippot, Jean-Claude Lazzaroni, Nathalie Lombard, Jun Yuan et Aurélie Faugier. Un remerciement tout particulier à Laure Franqueville pour son indéniable soutien, sa prévenance, sa gentillesse et son investissement, même dans les moments difficiles, et sans qui cette thèse n'aurait très certainement pas aboutie de la même manière. Un grand merci également à Sébastien Cécillon, pour son aide et ses conseils lors des diverses discussions sur l'analyse des données, mais également pour son fort dévouement (un peu trop peut être) dans la vie du laboratoire. Je tiens également à remercier mes deux stagiaires Claire Laligant et Elif Oz pour m'avoir aidé au cours de ma thèse. Par la même occasion, je tiens à remercier également toute l'équipe de GME, d'une part pour l'excellente ambiance et cohésion qui y règne, mais également pour l'aspect synergique autour des discussions scientifiques et des réunions, et qui s'est avéré extrêmement riche en terme d'apprentissage. Remerciements donc à l'équipe actuelle Jérémy Pivetal, Alban Mathieu, Joseph Nesme, Lorrie Macario, Laurine Blanchard, Jun Yuan, Jérémy Reboulet, Catherine Larose, Frédéric Lehembre, Sandrine Demanèche, Jean-Michel Monier, Isabelle Navarro ; Aux anciens membres : Tom Delmont, Monique Lacroix, Emmanuel Prestat, Aurélie Faugier, Maude David, Délina Lyon, Barbara Pivato, Marina Hery… ainsi qu'à l'équipe ENOVEO : Cédric Malandain, Céline Baguelin, Sandra Entresangles. Un grand remerciement plus large à toutes les personnes du laboratoire Ampère, avec un accent particulier pour : Osman Osman, Samia Menad, Yoan Jaffre, Julie Laforêt-Ast, Riccardo Scorretti, Abdellah Ogbi, Amine Drissaoui, Janvier Sylvestre N'cho, Viet Hung, Kevin Berger, Mathieu Brun, Amine Slama…

Je tiens également à remercier sincèrement toutes les personnes impliquées dans le projet Européen « METAEXPLORE » avec lesquelles j'ai pu collaborer et échanger durant les réunions de travail à travers l'Europe. Remerciements en particulier à Sara Sjölin, Elisabeth Wellington, Ines Mandic Mulec, Flavia Marinelli, Dick van Elsas, Andreas Schluter, Helena Wright, Lucas Ausec, Silvia Cretiou, Anna Kielac, Martha Zakrzweski, Ilaria Presti, Detmer Sipkema ainsi qu'a toutes les personnes avec qui j'ai pu échanger et collaborer durant ces années.

Plus personnellement, je tiens à remercier ma famille pour m'avoir moralement soutenu durant mon parcours. Remerciements à ma mère et à mon père, Annette et Jean-Luc, à ma sœur Enimie et à mes frères Nicolas et Kévin, et à mon grand-père René pour les moments réconfortants passés tous ensemble pendant les weekends et les vacances dans le Jura. Remerciement également à tous mes amis et à la famille éloignée pour les bons moments de décompression passés ensemble. Ces petites pauses ont su me ressourcer et m'apaiser, surtout dans les moments les plus difficiles.

Enfin, je tiens à remercier profondément une personne qui m'est extrêmement chère, et qui a partagé fidèlement mon quotidien de jeune chercheur durant ces 6 dernières années. Elle a su être forte pour rester à mes cotés et me soutenir même durant les passages les plus difficiles. Elle a été pour moi autant une source d'inspiration que de réconfort. Sans elle et son optimisme, je n'aurais pas pu accomplir tout ce que j'ai entrepris. Un grand merci donc à Christelle Dallut, que j'admire profondément, à qui je serais éternellement reconnaissant.

Pour finir, j'aimerais aborder d'un point de vue plus personnel le parcours que j'ai réalisé dans le cadre de cette préparation pour le diplôme de docteur en microbiologie. Cette expérience fut extrêmement enrichissante pour moi, d'une part du point de vue de l'apprentissage, grâce à la multitude d'outils que j'ai pu utiliser afin de répondre à la thématique de mon doctorat, mais également de par les rencontres que j'ai été amené à effectuer, au sein de mon équipe d'accueil, tout comme à l'étranger lors des différentes collaborations, réunions de travail et conférences.
Bien entendu, tout n'a pas été idyllique durant mon parcours. Il y a eu des hauts et des bas, des moments de grande joie mais aussi de profonde tristesse, mais en aucun cas je n'ai renoncé à poursuivre l'expérience, par respect pour les gens qui ont cru en moi, mais également pour croire en moi-même.

Toutes les expériences que j'ai eu la chance de vivre durant ces quatre années m'ont beaucoup apporté et changé, mais elles ont également conforté et confirmé des convictions personnelles, en m'incitant définitivement à poursuive dans le monde de la recherche académique.

Acknowledgements

First and foremost, I would like to thank sincerely all the people that participated, contributed, helped, and supported this work directly or indirectly, in which I was involved during these years.

I would like to acknowledge first my PhD supervisor: Pascal Simonet, for giving me the chance to realize in good conditions the preparation of my thesis. I thank him for giving me his trust and his support along my way, during good and hard moments. His experience and his vision of science were for me a source of both inspiration and motivation, and still today.

I would like to thank the people from the scientific comity, the evaluators: Gabrielle Veronese and Ines Mandic Mulec, and the reviewers: Elizabeth Wellington, Mohamed Jebbar and Laurent Philippot, for having kindly provided some of their time to read and evaluate my work.

I would like also to acknowledge Timothy Vogel, who gave me the opportunity to integrate the Environmental Microbial Genomics group right after my master degree. I would like to thank all the people involved in the "Genefish project" that I personally spend time with: Laurent Philippot, Jean-Claude Lazzaroni, Nathalie Lombard, Jun Yuan and Aurélie Faugier. I would like to address a very special thank to laure Franqueville, for her undeniable support, her attentiveness, her kindness and her personal investment, even in the hardest times, and without whom this thesis would certainly not accomplished in the same way. A special thank to Sebastien Cécillon, for his help and his advices during scientific discussions about data processing, and also for his strong dedication in the functioning of the lab. I would like to thank my two interns Claire Laligant and Elif Oz for the quality of their work, and for helping me out in the achievment of my thesis. At the same time, I would like to acknowledge all the people from the group, on the one hand for the great ambiance and cohesion within the group, and also for the synergic aspect toward the scientific discussions during the meetings, which was extremely rich in terms of learning, and also open-mindedness. Therefore, I thank the actual crew: Jérémy Pivetal, Alban Mathieu, Joseph Nesme, Lorrie Macario, Laurine Blanchard, Jun Yuan, Jérémy Reboulet, Catherine Larose, Yoan Le Digabel Sandrine Demanèche, Jean-Michel Monier, Isabelle Navarro, and also former members: Frédéric Lehembre, Tom Delmont, Monique Lacroix, Emmanuel Prestat, Aurélie Faugier, Maude David, Délina Lyon, Barbara Pivato, Marina Hery, Sibel Berger; the people from ENOVEO: Cédric Malandain, Céline Baguelin, Sandra Entresangles. I would like to thank all the people from the laboratory "AMPERE", with a special emphasis for Osman Osman, Samia Menad, Yoan Jaffre, Julie Laforêt-Ast, Riccardo Scorretti, Abdellah Ogbi, Amine Drissaoui, Janvier Sylvestre N'cho, Viet Hung, Kevin Berger, Mathieu Brun, Amine Slama…

I would like to address my gratitude toward all the persons involved in the European project METAEXPLORE, with whom I could collaborate and exchange during the workshop meetings all across Europe: Sara Sjölin, Elisabeth Wellington, Ines Mandic Mulec, Flavia Marinelli, Dick van Elsas, Andreas Schluter, Helena Wright, Lucas Ausec, Silvia Cretiou, Anna Kielac, Martha Zakrzweski, Ilaria Presti, Detmer Sipkema and also all the persons with whom I could collaborate and exchange during these years.

More personally, I would like to thank my family for constantly supporting me all along my thesis. I thank my mother and my father: Annette and Jean-Luc, my sister and my brothers: Enimie, Nicolas and Kévin, and my grand-father René for all the comforting moments we spend together during weekend and holydays in the family house in the mountains. I would like to thank as well all my friends and all the members of my family for the good times that we spend together. These short breaks helped me a lot, especially during difficult times.

I would like to address my thankfulness to someone dear to me, who share faithfully my daily life as a young researcher during the last 6 years. She was strong enough to stay beside me, and to support me even the hardest times. She was both a source of comfort and ispiration for me, and without her optimism, I would not be able to accomplish all the things that I have started. A sincere and kind “thank you so much” to Christelle Dallut, whom I deeply admire, and whom I would be eternally grateful.

Finally, I would like to end this section by giving a more personal point of view on my experience during the preparation of this thesis. It was extremely rewarding for me, as I had the chance to use many tools and technologies to address the topics of my PhD, but also from the many encounters I had in my host team, but also during the collaborations and the conferences all around the world. Of course, it wasn’t perfect. I had lots of ups and downs, moments of great joy but also deep sadness. But in any case, I didn’t give up, and I pursued the experience, for respecting the people who believed in me, and also for believing in myself.

All the experiences I lived during these four years brought me a lot, and probably changed me a lot as well, but they have also strengthened and confirmed personal beliefs, by definitely inciting me to continue in the world of academic research.

Glossaire et abréviations

- **ADN** : Acide Désoxyribonucléotique, matrice moléculaire contenant l'information génétique. Chez les bactéries, l'ADN se trouve sous forme d'un chromosome unique formant le génome, mais aussi sous une plus petite forme appelée le plasmide.
- **Ara** : L-arabinose, glucide utilisé dans l'induction du gène toxique *colE3*.
- **BLAST** : Basic Local Alignment Search Tool, algorithme de recherche de similarité entre des séquences. Il existe différentes sortes de BLAST en fonction du type de séquences utilisées ou recherchées. Souvent utilisé pour rechercher une séquence non identifiée dans les bases de données.
- **Cb** : Carbénicilline, antibiotique permettant la culture sélective des bactéries possédant le gène de résistance associé.
- **Cm** : Chloramphénicol, antibiotique permettant la culture sélective des bactéries possédant le gène de résistance associé.
- **Ecosystème** : Entité désignant l'ensemble des organismes (la biocénose) vivant dans un milieu donné (le biotope) ainsi que leurs intéractions.
- **IPTG** : Isopropyl β-D-1-thiogalactopyranoside, réactif de biologie moléculaire utilisé dans l'induction du gène toxique *relF*.
- **Kan** : Kanamycine, antibiotique permettant la culture sélective des bactéries possédant le gène de résistance associé.
- **Mobilome** : Ensemble des gènes et séquences nucléiques faisant partie des éléments génétiques mobiles, et donc suceptibles de se déplacer dans un génome ou d'être transféré vers un autre organisme.
- **Ligation** : Technique de manipulation génétique visant à l'insertion d'une séquence d'ADN dans un réceptacle pouvant être de nature variée (plasmide, chromosome…)
- **Oligotrophe** : Terme qui qualifie un milieu pauvre en nutriments.
- **PCR** : Polymérase Chain Reaction, technique permettant l'amplification de l'ADN à l'aide d'amorces spécifiques. Elle est utilisée pour préparer de grande quantité d'ADN ou pour vérifier la présence ou non du matériel génétique recherché.
- **Phénotype recD-** : Souche bactérienne mutée sur le gène recD pour le rendre inopérant. La perte de RecD se traduit par une augmentation des évènements de Recombinaison Homologue.
- **pb/kb** : Abréviation de « Paire de Base » et « Kilo Base ». Les 4 bases, ou nucléotides, ATGC sont les constituants de l'ADN.
- **recBCD** : Ensemble de 3 gènes codant un complexe protéique RecBCD constitué de 3 sous unités B, C et D possédant toutes une activité enzymatique intervenant dans la recombinaison homologue.
- **Recombinaison Homologue**: mécanisme biologique par lequel deux brins d'ADN présentant des similarités sur leurs séquences nucléotidiques vont se réorganiser afin de former une nouvelle séquence dite recombinante. On distingue le brin donneur et le brin receveur. En l'occurrence dans notre système, le plasmide pB35 est le receveur, et l'ADN ciblé devient le donneur.
- **Taq Polymérase** : Enzyme utilisée pour la technique de PCR.
- **Taxon/taxa** : Terme désignant un ensemble d'organismes partageant des similarités selon des critères taxonomiques, souvent d'ordre génétique.

ARTICLES SOUMIS EN REVISION :

S. Jacquiod, L. Franqueville, P. Simonet (2012) Novel trends in soil metagenomics. *Res Mic*. Submitted and under revision. ***Page 59***

S. Jacquiod, L. Franqueville, S. Cécillon, T.M. Vogel, P. Simonet (2012) Soil bacterial community shifts after chitin enrichment in microcosm. *ISMEj*. Submitted and under revision. ***Page 169***

Plan - Summary

Remerciements - Acknowledgements 3

Glossaire et abréviations 7

Figures et tableaux -Tables and figures 14

Préambule 18

Introduction (French version) 19

1. La cellule et ces principaux constituants 20
2. Les bactéries : Définition et caractéristiques 25
3. Rôle des bactéries dans l'environnement 31
4. Contexte de la thèse 32

Introduction (English version) 35

References 44

CHAPTER 1 - French summary 49

CHAPTER 1 - BIBLIOGRAPHICAL REVIEW OF THE SCIENTIFIC CONTEXT 59

Abstract 60

Introduction 61

1. Soil as a microbial habitat 62

2. Metagenomic investigation of soil 63

i. Metagenomic DNA and metagenomic approaches 63

ii. Nucleic acid extraction 65

iii. Metagenomic clone libraries 67

3. Metagenomics and next generation sequencing 71

i. High-throughput sequencing technologies 71

ii. Metagenomic data and bioinformatics 72

4. The single-cell approach 76

i. Introducing single-cell concept 76

ii. Single-cell biases and alternatives 77

iii. A complementary approach for metagenomics 78

Conclusion 79

Acknowledgements 80

References 80

CHAPTER 2 - French summary 91

CHAPTER 2 - THE EUROPEAN PROJECT METAEXPLORE:
METAGENOMICS FOR BIOEXPLORATION 103

CHAPTER 2 - Part.1 - PRESENTATION OF THE EUROPEAN PROJECT METAEXPLORE 104

1. Introduction to the METAEXPLORE project 104

2. Organization of the project 105

i. Metagenomics and targeted enzymes 105

ii. Work packages management 109

iii. Partner #7 implications 112

References 115

Appendixes 116

CHAPTER 2 - Part.2 - THE METAGENOMIC DATABASE FROM THE ROTHAMSTED SOIL:
A BIO-INFORMATIC STRATEGY TO SCREEN FOR CHITINASE GENES 121

1. Introduction 121

2. Material and methods 122

i. Soil sampling and DNA extraction 122

ii. Chitinase assays 123

iii. Chitinase BLASTx assays 125

3. Results and discussion 128

i. Chitinase assays 128

ii. BLASTx: Abundance in metagenomes 128

iii. BLASTx: Diversity in metagenomes 130

iv. BLASTx: Taxonomical affiliation 132

Conclusion 134

References 136

Supplemental informations 138

CHAPTER 2 - Part.3 - THE METAGENOMIC CLONE LIBRARY FROM ROTHAMSTED SOIL: A SEQUENCE-BASED SCREENING 141

Introduction 141

1. Presentation of the sequence-based strategy 141
2. Material and methods 142

i. Clone library construction 142

ii. Nylon membrane settings 142

iii. chiA probes design 143

iv. Preliminary biotin labeling and hybridization attempts 143

v. ^{33}P labeling and hybridization 147

vi. Positive clones' analysis 147

3. Results 148

i. Assembly results and BLASTn verification 148

ii. Annotation through the RAST server 148

iii. *BLASTx against homemade chitin degradation database* 154

4. Discussion 156

References 160

Appendixe 2-3a 163

CHAPTER 2 - Part.4 - SOIL BACTERIAL COMMUNITY SHIFTS AFTER CHITIN ENRICHMENT: AN INTERGRATIVE METAGENOMIC APPROACH 169
Abstract 169
Introduction 170

1. Material and methods 171
   - *i. Soil sampling* 171
   - *ii. Microcosm's settings, chitin enrichment and sampling strategy* 171
   - *iii. Chitin degradation and chitinase assays* 172
   - *iv. Metagenomic DNA extraction* 172
   - *v. rrs qPCR* 172
   - *vi. Phylochip analyses* 173
   - *vii. Microarray scanning and data processing* 173
   - *viii. Pyrosequencing and data analysis* 174

2. Results 174
   - *i. Chitinase activity* 174
   - *ii. rrs qPCR* 176
   - *iii. Phylochip results* 176
   - iv. *Pyrosequencing analysis* 179

3. Discussion 181
   - *i. Factors structuring metagenomic studies of soil bacterial communities* 181
   - *ii. Global impact of two chitin concentrations on soil bacterial communities* 184
   - *iii. Metagenomic analysis at sensu stricto with 0x20 incubation control* 184
   - *iv. Metagenomic analysis at sensu medio with 0x0 non-treated control* 184
   - *v. Metagenomic analysis at sensu lato with Rothamsted database* 185
   - *vi. Taxonomic changes driven by chitin in soil bacterial communities* 187
   - *vii. New potential chitin degraders in the enriched soil microcosms* 187
   - *viii. Impact of enrichment at the functional level* 188

4. Conclusion 189

Acknowledgements 189
References 190
Supplemental informations 194

CHAPTER 3 – French summary 201

CHAPTER 3 - GENEFISH: A NEW TOOL TO INVESTIGATE TARGETED BACETRIAL DIVERSITY 209

Introduction 210

1. The Genefish tool functioning 212
   - *i. Presentation* 212
   - *ii. Recipient strain and targeted DNA* 212
   - *iii. Optimizing the capture step* 213
   - *iv. The counter-selection syste*m 217
   - *v. Recombination sites: design and cloning* 221

2. Results and discussion 224
   - *i.* Preface to results 224
   - *ii.* The λ phage recombinase strategy 224
   - *iii.* Lethal-selection system: efficiency and stability 225
   - *iv.* Testing the Genefish tool: the tandem startegy 227
   - *v.* Testing the Genefish tool: the co-transformation strategy 231
   - *vi.* Testing the Genefish tool: the segregation strategy 234

3. Conclusions and perspectives 236

Acknowledgments 237

References 238

Appendixes 243

Conclusion et persepectives - Version Française 273

Conclusion an perspectives - English version 277

Figures et tableaux - Figures and tables

INTRODUCTION

- Figure 1: The Tree of life 21
- Figure 2: Schematic representation of a Bacterium 22
- Figure 3: Schematic representation of protein's synthesis 24
- Figure 4: Schematic representation of antibiotic resistance 28, 37
- Figure 5: Schematic representation of horizontal gene transferts 30, 40

CHAPTER 1

- Figure 1: Overview of metagenomic and single-cell approaches 53, 69
- Table 1: Frequent bioinformatic tools 56, 73

CHAPTER 2 - Part.1

- Figure 1: Principal use of chitin derivatives 106
- Figure 2: Schematic representation of METAEXPLORE workflow 94, 110
- Figure 3: Schematic representation of Partner 7 implication 113
- Appendixe 2-1b: Metaexplore partners 119

CHAPTER 2 - Part.2

- Figure 1: Schematic representation of chitin biodegradation 96, 124
- Table 1: Description of the 7 metagenomes from Rothamsted soil 124
- Table 2: Enzymes involved in chitin degradation 126
- Table 3: Description of all metagenomes used for comparison 127
- Figure 2: Endo- and exo-chitinase activities 129
- Figure 3: Abundance of chitin degradation related reads in metagenomes 129
- Figure 4: Diversity of chitin degradation related reads in metagenomes 98, 131
- Figure 5: Cluster dendrogram of metagenomes based on BLASTx 131
- Figure 6: Environmental pattern of chitin degradation enzymes 98, 133
- Figure 7: Taxonomical affiliation of reads related to chitin degradation 133
- Supplemental information: dominant hits in annotations 138

CHAPTER 2 - Part.3

- Table 1: Final set of probes 144

- Figure 1: Biotin hybridization assays 146

- Figure 2: ^{33}P hybridization assays 146

- Figure 3: Annotation of the contigs on SEED 149

- Table 2: Affiliation of the two CE4 CDS 149

- Figure 4: Annotation of the contig 55 151

- Figure 5: Annotation of the contig 92 151

- Figure 6: Annotation of the contig 52 152

- Figure 7: Annotation of the contig 72 152

- Figure 8: Annotation of the contig 50 153

- Figure 9: Annotation of the contig 173 155

- Figure 10: Annotation of the contig 164 155

- Table 3: Description of CDS of interest 161

- Table 4: BLASTx results of contigs against chitin degradation enzyme database 162

- Supplemental information: chiA probes design - amplicons 165

- Supplemental information: chiA probes design - tree 166

- Supplemental information: chiA probes design - primers 167

CHAPTER 2 - Part.4

- Figure 1: Enrichment kinetic - endo- and exo-chitinase activities 101, 175

- Figure 2: Enrichment kinetic - qPCR assays 101, 175

- Table 1: Relative abundance of bacterial groups based on *rrs* phylochips 177

- Table 2: Description of metagenomes used in this study 177

- Figure 3: Hybridization signal on phylochips 178

- Figure 4: Cluster dendrogram of metagenomes based on taxonomical and functional affiliations 101,180

- Figure 5: Major taxonomical changes detected 182

- Figure 6: Major functional changes detected 183

- Table 3: Summarized table of genera selected by enrichment 186

- Supplemental information: S1 - Taxonomical affiliation at the phylum level 194

- Supplemental information: S2 - Functional affiliation on SEED level 1 195

- Supplemental information: S3 - Summarized table of genera reduced by enrichment 196

- Supplemental information: S4 - Cross-comparison of taxonomical profils 197

- Supplemental information: S5 - Cross-comparison of functional profils 198

- Supplemental information: S6 - RISA profils 199

CHAPTER 3

- Figure 1: Genefish principle 204, 215

- Figure 2: RecBCD model 215

- Figure 3: Genefish plasmids 204, 218

- Figure 4: colE3 regulation 206, 220

- Figure 5: *relF* regulation 206, 200

- Figure 6: nar operon among Bacteria 222

- Figure 7: *Pseudomonas fluorescens* nar operon 222

- Figure 8: Final capture plasmid pB35G1G2 223

- Figure 9: pSIM6 induction assays 226

- Figure 10: Representation of the escape rate 226

- Figure 11: The tandem strategy 228

- Figure 12: Copy number limitation 228

- Table 1: Summary of pB35G1G2 construction steps 229

- Table 2: Tandem strategy results 230

- Figure 13: Co-transformation strategy 232

- Table 3: Co-transformation strategy results 233

- Figure 14: Segregation strategy 234

- Table 4: Segregation strategy results 235

- Table 5: Summary of successful recombination attempts 235

- Appendixe 3-1: List of strains, plasmids and primers 245, 246

- Appendixe 3-2: Plasmid preparation - CsCl gradient technique 248

- Appendixe 3-3: pSIM6 and FRT-FLP induction assays 252, 254

- Appendixe 3-5: Design of G1 and G2 regions 257

- Appendixe 10: Toxic cassette deletion and recombinant verification 270, 272

- *Préambule*

Cette partie a pour vocation d'introduire le sujet dans son contexte général, et d'apporter des éléments de compréhension essentiels qui permettront au plus grand nombre d'appréhender les bases de ce travail de thèse. Des notions importantes telles que la synthèse des protéines et des enzymes, ainsi que les bactéries, leur fonctionnement et leur rôle seront abordés afin de présenter les bases de ces travaux.

Le manuscrit étant principalement rédigé en langue anglaise, un résumé concis écrit en français a été prévu au début de chacun des trois chapitres qui constituent cette étude. Le choix de la langue anglaise pour l'écriture a été motivé principalement par la présence de membres anglophones dans la composition du jury qui va évaluer ce travail, mais aussi de par la forte connotation internationale de ces travaux de recherche, qui ont été effectués dans le cadre d'un projet Européen.

La partie qui va suivre permet de rappeler quelques points fondamentaux en microbiologie, afin de faciliter l'accessibilité et la compréhension générale de mes travaux de thèse.

INTRODUCTION GENERALE

-

LA CELLULE, LA MICROBILOGIE ET LA METAGENOMIQUE

-

VERSION FRANÇAISE

1. La cellule et ces principaux constituants

La microbiologie environnementale est une science qui vise à étudier les micro-organismes qui vivent sur notre planète, et de mieux comprendre leur rôle, ainsi que les mécanismes auxquels ils se prêtent au sein des écosystèmes. Un écosystème est une entité faisant référence à l'ensemble des organismes (la biocénose) vivant dans un milieu donné (le biotope), ainsi que différentes relations d'intéraction entre organismes, mais aussi entre biocénose et biotope.

Le terme micro-organisme englobe tous les êtres vivants uni ou pluricellulaires microscopiques, c'est-à-dire des formes de vie basées sur une ou plusieurs cellules et dont la taille se situe généralement en dessous des 50 micromètres (0,05 millimètres). Cette appellation englobe les procaryotes, organismes unicellulaires dépourvus de noyau comprenant respectivement les bactéries et les archaebactéries, mais aussi une partie des eucaryotes, organismes uni- ou pluricellulaires pourvus de noyau tels que les champignons, les levures et les protistes (algues et protozoaires). Une représentation de l'arbre du vivant est proposée en figure 1 afin d'illustrer ces propos. Les archaebactéries sont des organismes inféodés généralement aux environnements extrêmes, présentant à la fois des similitudes avec les bactéries et les eucaryotes. Plus discutable est le cas des virus, dont le statut et l'appartenance au domaine du vivant sont un sujet controversé au sein de la communauté des microbiologistes.

La cellule est considérée comme l'entité de base du vivant. Elle est constituée d'une membrane plasmique qui contient le cytoplasme, une sorte d'émulsion granuleuse dans laquelle baignent les constituants cellulaires. Dans le cas des eucaryotes, le cytoplasme contient des organites spécialisés tels que des mitochondries qui fournissent l'énergie à la cellule, et l'appareil de Golgi qui est responsable de la synthèse des protéines à l'aide des ribosomes. La cellule eucaryote contient également une structure particulière : le noyau, dans lequel est contenue l'information génétique : l'ADN, qui se présente sous la forme de chromosomes organisés d'une manière remarquable.

A la différence des eucaryotes, les procaryotes (bactéries et archaebactéries) ne possèdent pas de structure protectrice comme le noyau. L'ADN est directement situé dans le cytoplasme et n'est composé que d'un seul et unique chromosome (Fig.2). En revanche, les procaryotes possèdent une structure génétique supplémentaire nommée « plasmide », beaucoup plus restreinte en termes de taille que le chromosome (environ 8000 paires de base pour un plasmide contre 4 millions pour un chromosome en moyenne chez les bactéries), mais pouvant exister en plusieurs copies au sein d'une même cellule. Ces plasmides ont une importance capitale, car souvent porteurs de gènes conférant un véritable potentiel adaptatif pour les procaryotes tels que la résistance aux antibiotiques, ou encore des enzymes de dégradation permettant l'accès à une nouvelle ressource nutritive.

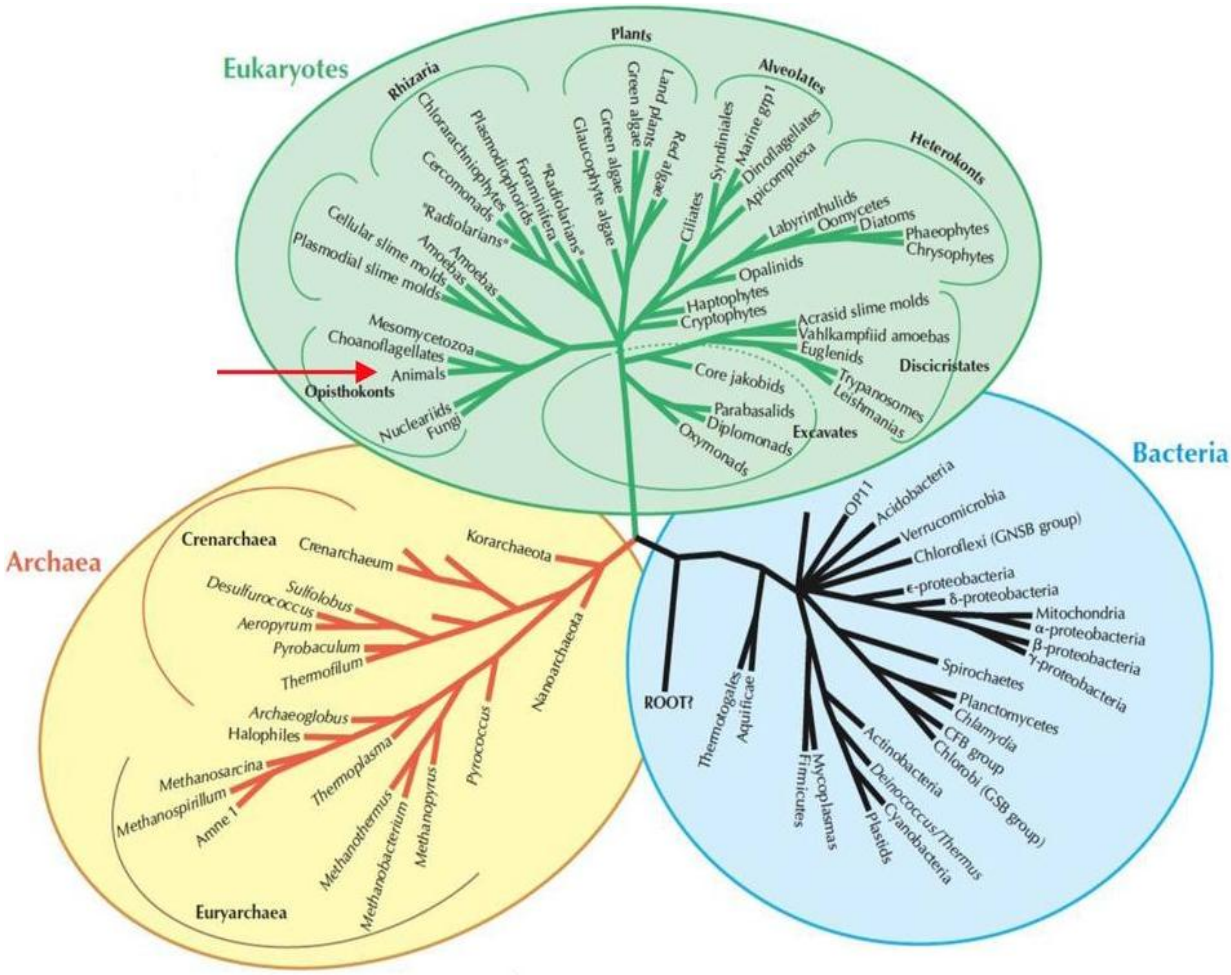


Fig.1 : Représentation schématique de l'arbre du vivant proposée par Baldauf et al. 2004. Sur cette figure sont présentés les trois grands domaines du vivant à savoir les Bactérie (bleu), les Archaebactéries (Jaune) et les Eucaryotes (vert) auxquels les mammifères et les Humains appartiennent (flèche rouge) au même titre que les plantes et les champignons. La racine de l'arbre (ROOT) se situe dans le domaine des Bactéries et représente l'organisme originel : LUCA (Last Universal Common Ancestor) qui ne reste à ce jour qu'une hypothèse sur l'énigme de l'origine de la vie sur Terre.

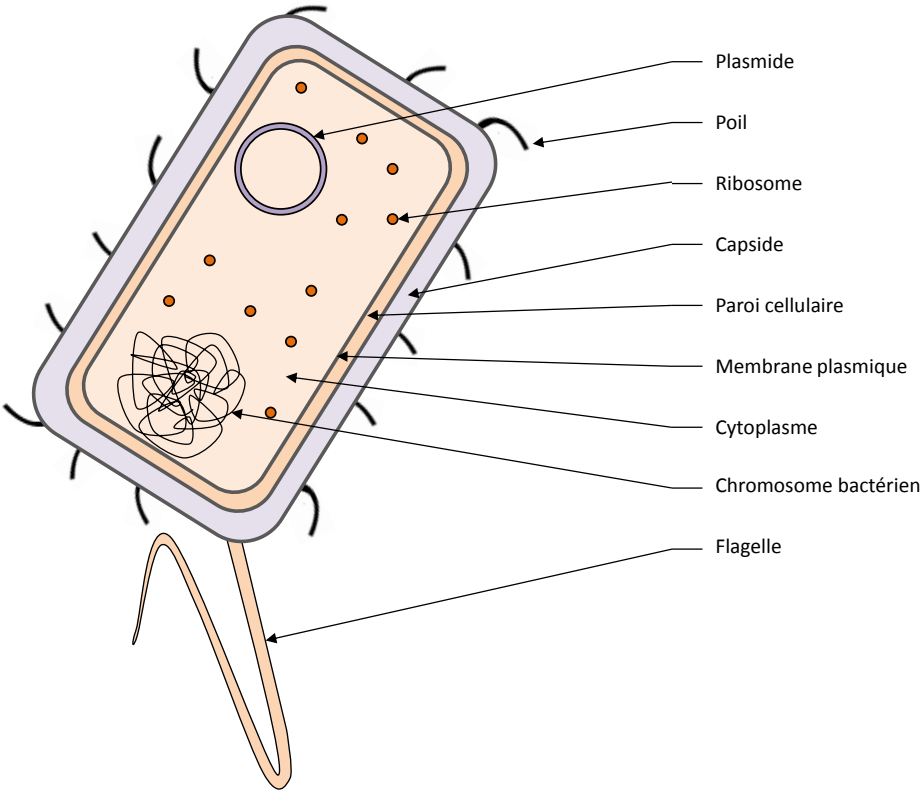


Fig.2 : Représentation schématique d'une bactérie. La bactérie est un organisme unicellulaire possédant un seul et unique chromosome (chromosome bactérien) baignant directement dans le cytoplasme, contrairement à la cellule eucaryote ou les chromosomes sont isolés dans un noyau. Le génome d'une bactérie comprend également les plasmides, molécule d'ADN circulaire à la réplication indépendante du chromosome et portant souvent des caractéristiques essentielles à la survie de la bactérie. Les ribosomes assurant la production des protéines à partir de l'information génétique sont directement localisés dans le cytoplasme, contrairement aux eucaryotes ou ils sont présent sur une structure particulière appelé l'appareil de Golgi. La cellule bactérienne est protégée par une membrane plasmique ainsi qu'une paroi cellulaire et une capside constituée de peptidoglycanes. Cependant certaines bactéries possèdent une deuxième membrane plasmique située au-delà de la paroi. Cette différence permet de distinguer les bactéries à Gram positif (sans deuxième membrane) des Gram négatifs (avec deuxième membrane).

L'information génétique chez les bactéries est donc à la fois portée par les plasmides, mais aussi par le seul et unique chromosome bactérien présent dans chaque cellule (Fig.2). Cette information est contenue dans les molécules d'ADN (Acide Désoxyribo Nucléique), constitué de deux filaments ou brins complémentaires. L'information génétique est véhiculée sous forme d'un codage moléculaire faisant intervenir des nucléotides, ou bases. Il existe au total 4 nucléotides qui rentre dans la composition de l'ADN : L'adénine (A) et la guanine (G) appelés les purines, ainsi que la cytosine (C) et la thymine (T) appelés les pyrimidines (Fig.3). Par jeu de complémentarité sur les deux brins, l'adénosine est toujours associée à la thymine, et la guanine à la cytosine. Les bases ATCG s'organisent sur les brins d'ADN en séquences précises et ordonnées, dont l'ordre détermine ce que l'on appelle les gènes. Un gène est une séquence d'ADN qui, une fois transcrite à l'aide d'enzymes dédiées, aboutit à la synthèse d'une séquence dite d'ARN (Acide Ribonucléique). Cette nouvelle molécule diffère de l'ADN dans le sens ou elle n'est constituée que d'un seul brin, et qu'elle contient de l'uracile (U) à la place de la thymine. Il existe différentes sortes d'ARN qui occupent différentes fonctions au sein de la cellule :

- ARNr : L'ARN ribosomique qui, associé à diverses protéines, forme les ribosomes.
- ARNm : L'ARN messager qui porte l'information d'un gène hors du noyau vers l'appareil de Golgi chez les Eucaryotes, et directement dans le cytoplasme chez les procaryotes où celui-ci sera décrypté par les ribosomes afin de produire la protéine correspondant au gène en question.
- ARNt : L'ARN de transfert, qui transporte les acides aminés au niveau des ribosomes, afin de constituer des protéines à partir de l'ARNm.

L'ARNm est donc le messager porteur de l'information génétique contenue dans l'ADN. Cette molécule a pour vocation d'être décryptée à son tour par des molécules appelées « ribosomes », permettant ainsi la fabrication de ce que l'on nomme « les protéines ».

Les protéines sont des macromolécules organiques composées d'une ou de plusieurs chaines d'acides aminés. L'ordre et la nature de ces acides aminés sont directement liés à la séquence du gène d'origine, et vont directement conditionner les propriétés de la protéine. Véritables ouvrières de la cellule, les protéines occupent une place fondamentale dans le fonctionnement du vivant, notamment au travers de leurs différents rôles :

- rôle structural et protecteur, en participant directement au maintien de la structure cellulaire *via* le cytosquelette (e.g l'actine), mais aussi en intervenant dans la compaction et la protection de l'ADN (e.g les histones).
- rôle dans la mobilité des organismes unicellulaires *via* les flagelles et autres poils cellulaires (*pili*), et bien entendu chez les organismes pluricellulaires *via* les muscles.
- rôle biochimique en exécutant bon nombre de réactions du métabolisme *via* les enzymes.

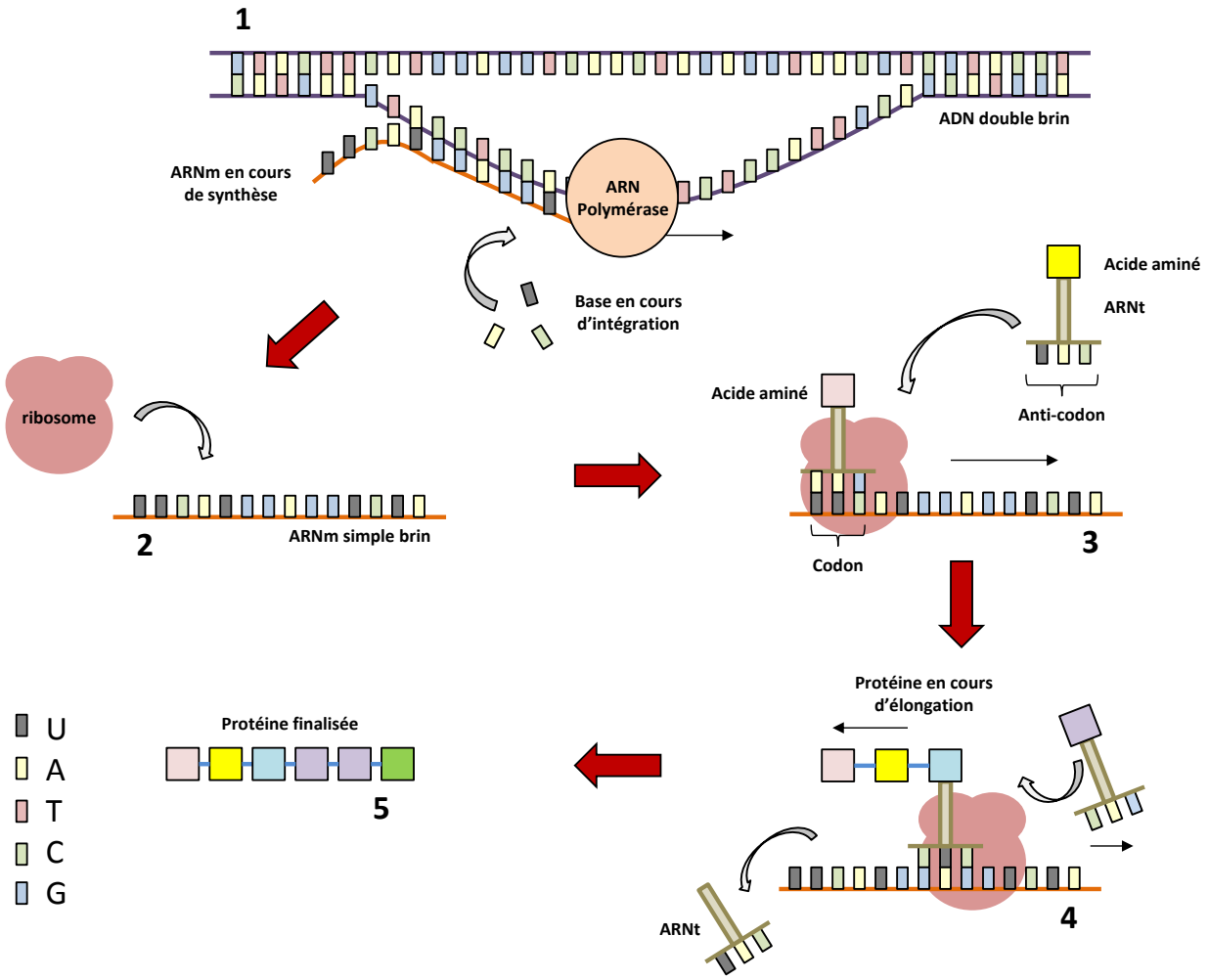


Fig.3 : Représentation schématique des étapes de transcription et de traduction conduisant à la synthèse de protéine dans une cellule. Le schéma (1) représente la transcription d'un gène en ARNm par l'enzyme ARN polymérase. Le brin d'ARNm ainsi formé est ensuite pris en charge par un ribosome (schéma 2) soit au niveau du cytoplasme chez les procaryotes, soit au niveau du réticulum endoplasmique chez les eucaryotes. Le schéma (3) illustre l'étape de traduction de l'ARNm en protéine par le ribosome, qui va intégrer et assembler des acides aminés via les ARNt. Le type d'acide aminé intégré dépend de la séquence génétique de l'ARNm au travers des codons, et des anti-codons présents sur les ARNt. L'élongation de la protéine se poursuit (4) et s'achève par la libération d'une protéine finalisée, qui subira encore quelques modifications post-traductionnelles avant d'être fonctionnelle. La protéine pourra alors remplir son rôle dans la cellule, ou bien en dehors de celle-ci si elle est secrétée.

C'est précisément cette dernière catégorie de protéines : les enzymes, qui seront particulièrement abordées dans cette étude. Ce sont des protéines fonctionnalisées capables d'interagir avec d'autres molécules afin de catalyser une réaction particulière en la facilitant. Par exemple, elles procurent l'énergie en dégradant des molécules comme les sucres et les lipides. Elles sont également impliquées dans la synthèse de certains composés comme l'ADN, la membrane cellulaire, ou également les molécules du métabolisme secondaire comme les antibiotiques, les tanins, les polyphénols... Bon nombre de ces enzymes sont utilisées par l'Homme, ce qui explique leur fort potentiel applicatif dans les secteurs industriels et biotechnologiques. En effet, beaucoup d'enzymes sont employées directement comme médicaments, ou indirectement lors du processus de fabrication de certaines molécules pharmaceutiques, mais aussi lors de l'élaboration de produits cosmétiques. Elles sont employées dans le secteur agroalimenntaire dans certains processus de transformation des aliments, mais également dans d'autres secteurs comme l'environnement, la dépollution, la synthèse de produits détergents, etc…

Bien entendu, la description moléculaire de la cellule ne se limite pas à cette courte présentation. Néanmoins, elle est suffisante pour introduire les bases nécessaires à la compréhension de ces travaux de thèse, qui visent notamment à découvrir de nouvelles enzymes d'intérêt industriel à partir des bactéries. Après cette brève introduction sur le fonctionnement de base de la cellule en général, nous allons maintenant aborder plus précisément l'objet d'étude central de cette étude: la bactérie.

2. Les bactéries : Définition et caractéristiques

Fort de plus de 3,5 milliards d'années d'évolution et du haut de leurs quelques $5x10^{30}$ représentants (Withman et al . 1998), les bactéries ont su s'adapter à un large spectre de conditions environnementales pour coloniser toutes les niches écologiques de notre planète : le sol (Chase *et al.* 1956), l'eau (LeChevallier *et al.* 1980), l'air (Imshenetsky *et al.* 1978), les déserts (Luo *et al.* 2012), la pluie (Morris *et al.* 2008), la neige (Yan *et al.* 2012), la glace (Kim *et al.* 2012), les sédiments des fonds océaniques (Quigley and Colwell, 1968), les grottes (Urzì *et al.* 2010), les pôles (Vincent *et al.* 2004 ; Bian *et al.* 2012), les milieux acides (Tyson *et al.* 2004), et hyper-salins (Sorokin *et al.* 2006), des sources hydrothermales aux températures extrêmes (Stetter, 1999), les milieux pauvres en nutriments (Essoussi *et al.* 2010), et même pollués (Hassan *et al.* 2008). Elles sont également abondamment présentes dans des milieux qui nous sont plus familiers tels que notre propre peau (Grice *et al.* 2009), ainsi que notre tube digestif (O'Hara and Shanahan, 2006). Cette étonnante capacité d'adaptation à des environnements variés voire extrêmes est en partie due à leur forte capacité de colonisation et de diversification. En effet, leur cycle de vie relativement court basé principalement sur la division cellulaire (mitose) leur permet d'accroitre rapidement leurs effectifs jusqu'à l'apparition de larges populations clonales. Par exemple, le temps de division de la bactérie *Escherichia coli* en condition optimale de croissance est d'environ 20 minutes (Cooper and Helmstetter, 1968). Cette capacité de multiplication clonale varie en intensité en fonction des espèces et de leur optimum de croissance. Toutefois, leur cycle de vie cellulaire est très nettement supérieur en termes de vitesse à celui des organismes plus développés, ce qui conduit à l'établissement de population de taille conséquente. De cette forte cinétique de croissance résulte inexorablement l'apparition de mutants au sein des populations, pouvant mener à des phénomènes d'adaptation, voire de « spéciation » *via* l'apparition de nouvelles espèces si les modifications génétiques sont trop importantes.

Les mutations interviennent directement sur le code génétique en modifiant la séquence d'ADN de l'espèce. Plusieurs facteurs peuvent induire des mutations. Elles peuvent d'une part être générées par la propre machinerie enzymatique de la cellule impliquée dans la réplication et la réparation de l'ADN. Ainsi, l'ADN polymérase, l'enzyme responsable du maintien et de la duplication de l'ADN, est connue pour commettre des erreurs sur la séquence nucléique à hauteur de 1 base sur 10 000 000 (Voliotis *et al.* 2012). D'autres facteurs tels que l'exposition au rayonnement UV et le stress oxydatif sont réputés pour en engendrer. Les mutations peuvent se traduire en effets délétères, en touchant un gène important qui ne sera plus exprimé, ou dont la protéine résultante ne sera plus fonctionnelle. La bactérie mutante sera alors moins apte à survivre dans des conditions environnementales données. C'est le phénomène de "sélection naturelle". En revanche, dans certains cas, la mutation peut se traduire par un effet positif, direct ou indirect, et ainsi améliorer la capacité de survie de la bactérie (Matic *et al.* 1997).

Par exemple, des mutations délétères peuvent parfois conférer un avantage sous certaines conditions (changement de température, sécheresse, prédation…). C'est le concept écologique de valeur sélective, ou « fitness », qui reflète la capacité d'un individu à survivre et à se reproduire sous la pression de la sélection naturelle. Cette sélection est conditionnée par des facteurs dits abiotiques (température, humidité, disponibilité en nutriments…) mais aussi par des facteurs biotiques (relations interspécifiques et intraspécifiques). Ainsi, la capacité d'adaptation des bactéries est en partie expliquée par leur cycle vital rapide et l'apparition inexorable de mutants au sein de leurs populations pouvant présenter une "fitness" supérieure à la souche initiale, et donc une meilleure capacité adaptative face aux conditions environnementales fluctuantes (Gordo *et al.* 2011). Toutefois, l'adaptation par mutation est un phénomène lent et n'explique pas à lui seul tout le potentiel adaptatif des bactéries. Il existe un autre phénomène permettant d'expliquer cette étonnante capacité : les transferts horizontaux de gènes (Jain *et al.* 2002).

La diversification et les potentialités d'adaptation des bactéries sont en grande partie expliquées par la présence d'un processus remarquable que l'on nomme « le transfert horizontal de gène » (Aminov *et al.* 2011). Par opposition à la transmission verticale des caractères génétiques, comme celle opérée chez les organismes supérieurs lors de la reproduction sexuée entre parents jusqu'a leur descendance, les procaryotes peuvent transférer ou acquérir une partie de leur patrimoine génétique *via* des mécanismes de transferts et d'échanges directs ou indirects de cellule à cellule. Ce phénomène permet aux bactéries de réaliser ce que l'on nomme le « brassage génétique » au sein de leur population, et ainsi de maintenir un certain degré de diversité. En effet, la division cellulaire étant un mode de reproduction clonal asexué, le brassage génétique reste limité, contrairement au mode de reproduction sexué faisant intervenir le croisement génétique de deux individus. Pourtant, un brassage endogène existe grâce à des éléments tels que les transposons ou les séquences d'insertion, qui permettent de générer de la variabilité au sein d'une population clonale (Fehér *et al.* 2012).

Le transfert horizontal permet ainsi de maintenir un certain niveau de diversité génétique, essentiel pour répondre aux fluctuations environnementales. Cela leur permet notamment une adaptation rapide à des conditions défavorables, en transférant directement les gènes adéquats aux autres membres de la communauté. Les transferts horizontaux permettent également l'accès à de nouvelles niches écologiques, en apportant aux bactéries le matériel génétique nécessaire à la colonisation de nouveaux habitats (Wiedenbeck and Cohan, 2011).

Plusieurs éléments génétiques ont été caractérisés comme étant potentiellement transférables horizontalement. Cependant, les plasmides représentent de loin la catégorie la plus importante, ces éléments portant souvent des gènes impliqués dans bon nombre de processus clefs, pouvant procurer un véritable gain adaptatif à leur hôte (Smillie *et al.* 2010).

Le meilleur exemple est celui de l'adaptation aux antibiotiques, dont les gènes responsables de la résistance sont souvent portés sur des éléments génétiques mobiles comme les plasmides (Fig.4), et donc potentiellement échangeables pour se répandre et se maintenir au sein d'une population bactérienne, parfois même en s'intégrant dans le chromosome (Andersson and Hughes, 2011). L'adaptation à des sites pollués en est un autre exemple ; certaines bactéries possédant les gènes plasmidiques impliqués dans la dégradation de molécules toxiques comme les pesticides peuvent transférer cette capacité à d'autres *via* des transferts horizontaux (Liang *et al.* 2012).

Il existe 3 grands types de mécanismes permettant le transfert horizontal de matériel génétique d'une cellule bactérienne à l'autre (Ochman et al. 2000) (Fig.5):

- ***La transformation*** : phénomène par lequel une bactérie intègre de l'ADN environnemental dans son cytoplasme. Dans l'environnement, l'ADN provient de la lyse de cellules de plantes, champignons, insectes, bactéries ou autres organismes vivants dans le même habitat. Cette réaction peut être soit naturelle, si l'espèce en question possède la capacité génétique liée à l'intégration d'ADN (e.g *Acinetobacter baylyi*, Elliott and Neidle, 2011), ou bien stimulée par des facteurs abiotiques comme les chocs thermiques ou électriques permettant de perméabiliser la membrane bactérienne, facilitant la pénétration d'ADN.

La transformation stimulée artificiellement est une technique utilisée très couramment en laboratoire pour intégrer de l'ADN (linéaire ou plasmidique) dans une souche bactérienne (e.g *E. coli*) (Aune and Aachmann, 2010). A noter que l'intégration d'ADN linéaire au sein du génome ne peut se faire que *via* un mécanisme d'échange appelé « recombinaison homologue » durant lequel le brin d'ADN intégré par transformation pourra éventuellement remplacer une partie du génome déjà existant si une certaine similarité entre les séquences est détectée, sans quoi le brin transformé sera détruit par des systèmes enzymatiques exonucléasiques. Plus la longueur et la similarité sont importantes, plus la recombinaison n'a des chances d'être effectuée.

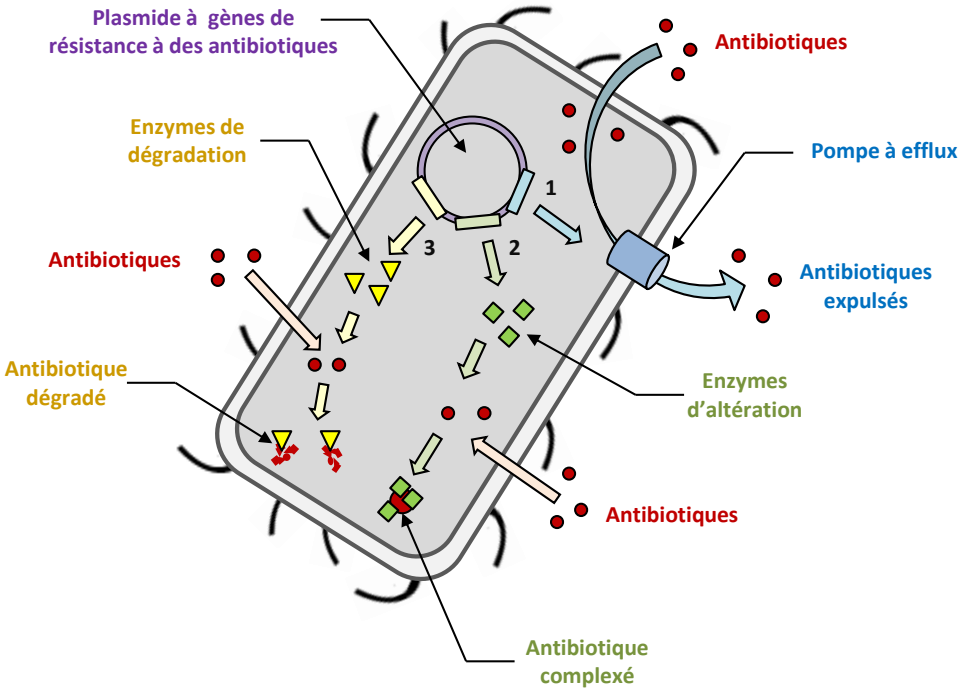


Fig.4 Principe de la résistance antibiotique chez la bactérie. Cette figure illustre schématiquement certaines des différentes possibilités de résistance à un ou plusieurs antibiotiques conférées par des gènes portés par un plasmide. La voie (1) en bleu représente l'expulsion des molécules antibiotiques via une protéine transmembranaire de type pompe à efflux. La voie (2) en vert représente l'inactivation de molécules antibiotiques par des enzymes qui modifient leur conformation chimique. La voie (3) en jaune représente la dégradation de molécules antibiotiques par des enzymes lytiques. Toutes ces enzymes et protéines sont codées par des gènes plasmidiques.

- ***La conjugaison*** : phénomène qui implique un échange direct d'ADN de bactérie à bactérie *via* la création d'un pont cytoplasmique appelé *pilus* (Lederberg and Tatum, 1953). Cette réaction intervient entre deux types de bactéries : (1) les bactéries donnatrices F+ (fertilité positive), équipées d'un plasmide de conjugaison (plasmide F), possédant des gènes conférant un phénotype spécifique se caractérisant par la synthèse du *pilus*. (2) Les bactéries réceptrices F-, non munies du plasmide en question, qui recevront une copie du plasmide F lors de la conjugaison *via* le *pilus*. Les bactéries receptrices F- deviendront à leur tour F+ et synthétiseront le *pilus* pour entamer un nouveau cycle de conjugaison. Cette réaction est souvent assimilée à une reproduction pseudo-sexuée chez les procayotes, la bactérie F+ équipée d'un *pilus* « sexuel » étant appelée « mâle » et la F- « femelle ». Certaines bactéries (Hfr) peuvent également intégrer le facteur F du plasmide dans leur génome, possédant ainsi la capacité d'échanger des gènes chromosomiques avec d'autres partenaires avec une fréquence élevée. Ce processus est très courant chez les bactéries et intervient directement dans les phénomènes d'adaptation, par exemple dans les sols au niveau de la rhizosphère (environnement immédiat des racines) (Ding and Hynes, 2009).

- ***La transduction*** : phénomène qui implique les virus de bactérie (bactériophages) comme vecteurs de transfert de gènes d'une bactérie à l'autre. Les bactériophages sont des virus à ADN ou ARN, protégés par une capsule protéique (la capside), et qui effectue son cycle de vie en infectant les bactéries. Une très grande spécificité existe entre les espèces bactériennes et leur population de phages, se traduisant par une relation étroite de « co-évolution » du couple phage-bactérie, impliquant un modelage réciproque et constant des génomes de chaque partenaire pour s'adapter en permanence à l'autre (Gomez and Buckling, 2011). Au cours du cycle d'infection, le bactériophage intègre son propre matériel génétique au sein du génome bactérien. Cette réaction d'intégration est notamment réalisée par des systèmes enzymatiques puissants, permettant d'initier le phénomène de recombinaison homologue même dans des conditions de similarité faibles. A ce niveau, on distingue deux types de processus d'intégration au chromosome : la transduction « généralisée », dans laquelle le matériel génétique viral est incorporé aléatoirement et, potentiellement, plusieurs fois dans le génome bactérien (cas du phage P22 et de la bactérie *Salmonella typhimurium* ; Ebel-Tsipis et al. 1972) ; et la transduction « spécialisée » ou seule une région précise du chromosome sert à l'intégration (cas du phage λ et de la bactérie *Escherichia coli* ; Campbell, 1962).

A ce stade, on différencie également deux types de bactériophages : (1) les bactériophages virulents, qui entament un cycle lytique au cours duquel la machinerie cellulaire sera détournée afin de produire massivement de nouvelles particules virales à partir de capsides et de brins d'ADN ou d'ARN. La cellule bactérienne finira par littéralement exploser sous la pression générée par les nombreuses particules virales, qui se retrouveront alors expulsées dans l'environnement. Seul les phages qui opèrent la transduction généralisée ont la possibilité d'encapsider par erreur un fragment du genome bactérien, et ce à cause des sites multiples d'insertion. Cette reaction intervient à une faible fréquence (≈1%).

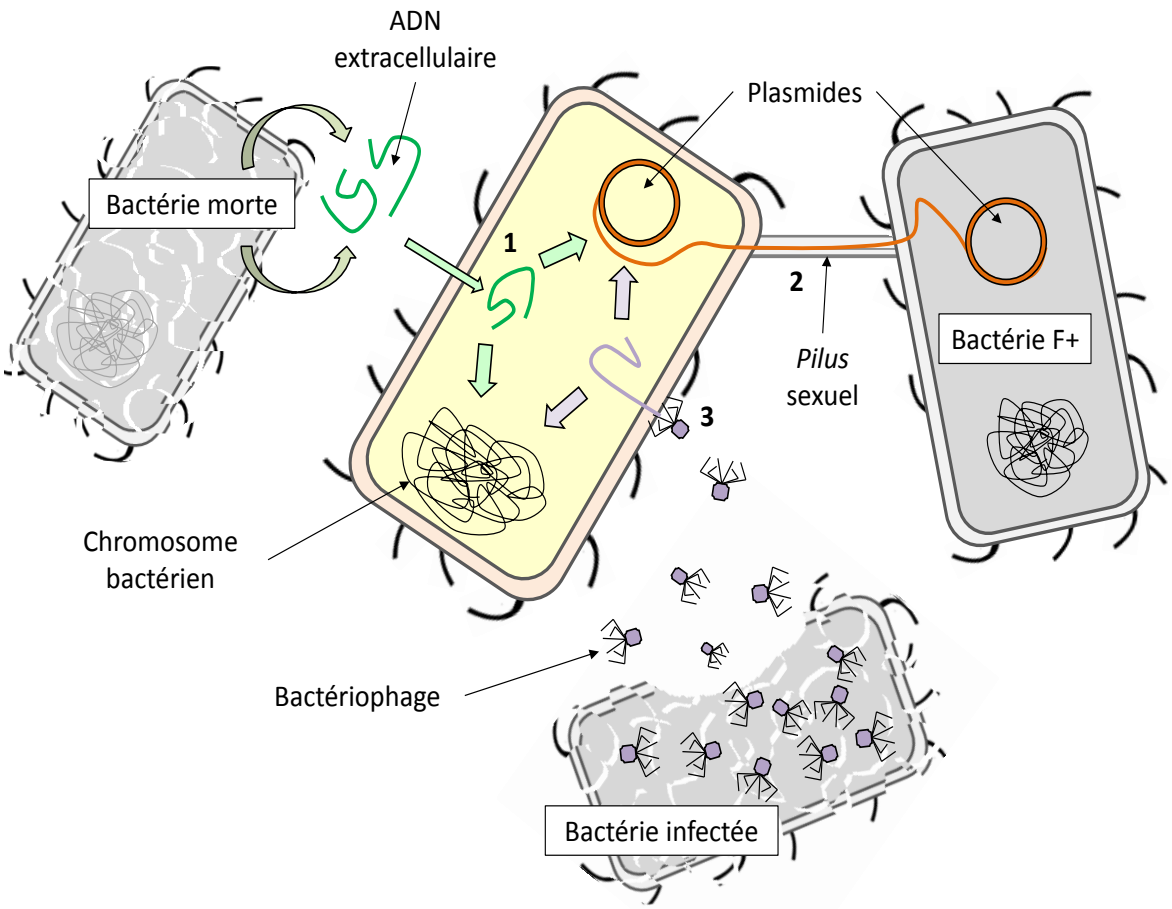


Fig.5 : Représentation schématique des processus de transferts horizontaux de gènes chez les bactéries. La transformation est illustrée en (1) à partir d'ADN environnemental extracellulaire provenant de cellules mortes (bactéries, plantes, champignons, insectes…). L'ADN entrant peut alors être intégré au génome de la bactérie transformée par la voie de la recombinaison homologue, à moins d'être détruit. La conjugaison est illustrée en (2), par le transfert d'un plasmide d'une souche F+ vers une souche F- via le pilus sexuel. La transduction est représentée en (3), via l'infection d'une bactérie par un bactériophage provenant lui-même d'une cellule bactérienne infectée.

(2) Les bactériophages tempérés, qui entament un cycle dit « lysogène » avant la phase lytique, durant lequel la bactérie continue à vivre et à se diviser avec l'ADN du phage dormant dans son génome. C'est durant cette phase que la bactérie infectée peut éventuellement utiliser du matériel génétique provenant d'une autre cellule préalablement infectée. Cette étape durera tant que l'ADN viral restera dormant, à moins que des facteurs environnementaux viennent déclencher le cycle lytique. L'apport d'une valeur sélective supplémentaire par transduction a été démontré dans de nombreux cas de figures comme la survie dans le sol (Schuch and Fischetti, 2009), ou la résistance à des antibiotiques (Mlynarczyk *et al.* 1997). Ces trois phénomènes réunis, ainsi que la variabilité endogène apportée par les mutations et les certains éléments génétiques mobiles (e.g transposons) apportent aux bactéries la possibilité de brasser leurs génomes en permanence et ainsi de générer une variabilité intra-spécifique suffisante pour faire face aux conditions fluctuantes de l'environnement et potentiellement de coloniser de nouvelles niches écologiques.

3. Rôle des bactéries dans l'environnement :

Les micro-organismes détiennent une place importante dans tous les écosystèmes sur Terre. Ils y occupent des fonctions primordiales telles que la dégradation et la minéralisation de la matière organique, évitant l'accumulation de molécules récalcitrantes difficiles à décomposer comme la cellulose (Stursová *et al.* 2012). Ils participent ainsi activement en tant que principaux acteurs des grands cycles biogéochimiques tels celui du carbone (Evan, 1976), de l'azote (Godfrey and Glass, 2011) du phosphore, du soufre... Ils occupent naturellement une place importante dans le secteur de l'agriculture où leurs actions conditionnent la fertilité et le bon fonctionnement des sols (Babalola, 2010), améliorent la croissance des plantes cultivées (Lucy *et al.* 2004) et permet de contrôler certains ravageurs (Raaijmakers *et al.* 2002). On les retrouve également chez les animaux et l'Homme, au travers de la peau et du tube digestif où ils vont jouer un rôle important de barrière ou de « défense naturelle » contre d'éventuels congénères dangereux (Thompson *et al.* 1978) mais aussi participer à la digestion de certains aliments que nous serions bien incapables d'assimiler sans eux (Tappenden and Deutsch, 2007). Ils sont également responsables de bon nombre de maladies infectieuses bénignes et graves, touchant aussi bien l'Homme directement, ou indirectement *via* les animaux domestiques et d'élevage. Outre leur importance indéniable au sein des différents écosystèmes, les micro-organismes sont également utilisés dans des secteurs beaucoup plus appliqués où leurs propriétés sont utilisées au quotidien à des fins industrielles ou environnementales. Ainsi, les micro-organismes sont impliqués dans l'industrie agro-alimentaire et notamment dans la fabrication des produits issus de la fermentation tels que le vin, la bière, le fromage et le cacao. On les emploie aussi indirectement dans les secteurs de la pharmacie et de la cosmétique, où les molécules qu'ils produisent sont utilisées dans les procédés de fabrication de médicaments ou autres formulations. Certains micro-organismes sont même utilisés directement en tant que médicaments connus sous le nom de probiotiques (Heller 2001 ; Gill 2003). Du point de vue environnemental, les micro-organismes sont utilisés dans un procédé que l'on nomme bioremédiation, qui consiste à bénéficier de leurs capacités afin de résoudre des problèmes de pollution sur des sites contaminés (Mikesková *et al.* 2012). Ils peuvent ainsi participer à la transformation et à la dégradation de produits dangereux et rémanents tels que les pesticides ou les hydrocarbures, qui peuvent contaminer sols et cours d'eau (Furukawa, 2003).

4. Contexte de la thèse

C'est dans ce contexte que se situent les objectifs du projet Européen METAEXPLORE, qui a financé mes travaux de thèse. Le but principal de ce projet est d'arriver à découvrir de nouvelles molécules d'intérêt industriel à partir des communautés microbiennes environnementales. Ce projet a été initié dans un contexte de développement rapide des nouvelles technologies utilisées en écologie microbienne et lié à la forte demande du monde industriel en perpétuelle recherche de nouvelles solutions pour répondre à leurs besoins. Ce projet rassemble 18 partenaires, principalement européens, incluant des équipes de recherche universitaires très impliquées dans le domaine des biotechnologies mais également des industriels en vue de pouvoir tester et optimiser les découvertes potentielles.

Les molécules d'intérêt recherchées sont des enzymes bactériennes ou fongiques, capables de dégrader les molécules récalcitrantes naturelles (la cellulose, la lignine, la chitine…) et anthropiques (e.g les pesticides organochlorés). La dégradation de ces molécules est valorisée dans différents secteurs industriels comme la filière des biocarburants, de par l'amélioration de la production de bioéthanol à partir de la matière végétale première : la biomasse ligno-cellulosique (Horn et al. 2006). De la même manière, l'industrie de la pêche va chercher à améliorer sa filière de production en tentant de valoriser ces déchets chitineux (carapaces de crabes et de crevettes), dont les produits de dégradation peuvent être employés en cosmétique et en pharmacologie. La dégradation des molécules récalcitrantes et polluantes d'origine anthropique trouve son utilité dans les secteurs impliqués dans la bioremédiation et la dépollution de sites contaminés. Pour ce faire, le projet mise principalement sur une approche récente et prometteuse appelée « la métagénomique ».

Pendant longtemps et encore aujourd'hui, les microbiologistes ont étudié les micro-organismes à l'aide de techniques culturales basées sur l'isolement et le développement de cellules individualisées sur milieux gélosés sous forme de colonies. Ces approches ont révolutionné la microbiologie en permettant l'accès à ces organismes microscopiques afin de pouvoir mieux les appréhender et les étudier. L'avènement des technologies moléculaires et du séquençage a permis de caractériser en profondeur le fonctionnement des micro-organismes cultivables, aux niveaux protéique et génétique aboutissant ainsi à l'établissement de banques de données. De ces organismes cultivables, on a pu isoler bon nombre de molécules actives d'intérêt tels que les antibiotiques mais aussi beaucoup d'enzymes toujours très utilisées dans l'industrie. A titre d'exemple, plus de 70 % des antibiotiques commercialisés au début des années 2000 sont issus de micro-organismes isolés *via* des techniques culturales (Demain, 1999). Cependant, plusieurs travaux ont apporté des éléments permettant de remettre en question l'accessibilité des micro-organismes par la voie culturale. En effet, on estime que dans certains environnements connus pour abriter une grande diversité microbienne, seulement entre 0,1 et 1% de bactéries sont cultivables (Amann *et al.* 1995). Une large majorité reste donc hors de portée des microbiologistes, remettant en cause notre perception de la diversité bactérienne. A titre d'exemple, de nouveaux phyla procaryotes sont découverts régulièrement, et viennent ainsi constamment enrichir notre connaissance de la taxonomie des procaryotes.

Cette taxonomie à été établie par Woese (Woese, 1987) en se basant sur le gène codant l'ARN ribosomique 16S comme marqueur au sein des représentants procaryotes. En effet, l'étude de ce gène hautement conservé permet de retracer l'évolution des différentes branches en se basant sur les modifications et similitudes de séquence nucléiques. Officiellement, 29 phyla sont acceptés et reconnus dans la classification des bactéries (LPSN, Euzéby 1997). Cependant, avec l'avènement des nouvelles techniques moléculaires, de nombreux nouveaux phyla ont été identifiés sans représentants cultivés. Le nombre est monté à 36 phyla caractérisés en 2002, dont 13 sans représentants cultivés (Hugenholtz, 2002). Ce nombre est passé à 53 en 2004 (Schloss & Handelsman, 2004) et même à l'heure actuelle, d'autres sont encore identifiés (Wylie *et al.* 2012).

De nombreuses stratégies ont été mises en place pour tenter d'améliorer les conditions de cultures et ainsi, tenter de se rapprocher des conditions de développement de ces bactéries non-cultivées. Des représentants des phyla *Verrucomicrobia* et *Acidobacteria* ont ainsi pu être obtenus sur des temps de culture prolongés en conditions oligotrophiques (pauvres en nutriments) (Janssen *et al.* 2002). Une autre technique consiste à isoler les cellules bactériennes dans des blocs d'agar à partir d'échantillons environnementaux, ces blocs sont ensuite exposés à diverses conditions physico-chimiques (température, pH…). Cette technique a permis d'isoler des micro-colonies d'espèces non cultivées (Zengler *et al.* 2002). Malheureusement, ces stratégies restent fastidieuses et peu fructueuses, en raison de la complexité des facteurs intervenant dans la croissance des bactéries environnementales (Ferrer *et al.* 2003; Gabor *et al.* 2004).

Depuis le début des années 1990, une nouvelle approche permettant de contourner cette limitation a vu le jour : « la métagénomique » (Torsvik *et al.* 1990). Son but est d'extraire l'ADN à partir des cellules contenues dans un échantillon environnemental (sol, eau, sédiment…). Cet ADN métagénomique, dans le sens où il contient plusieurs génomes issus d'espèces et d'organismes différents, est le principal objet d'étude de cette approche (Rondon *et al.* 2000). Ainsi, tous les micro-organismes présents dans un échantillon environnemental sont potentiellement accessibles et caractérisables au niveau génétique. D'ailleurs, les premières exploitations de l'ADN métagénomique extrait à partir de l'environnement se focalisaient principalement sur la caractérisation et l'estimation de la diversité bactérienne (Amann *et al.* 1995; Hugenholtz *et al.* 1998).

En revanche, appréhender l'information contenue dans l'ADN métagénomique représente un défi ambitieux et complexe. En effet, une très grande diversité réside au sein de la majorité des écosystèmes, comme par exemple les sols ou les océans. A titre d'exemple, la diversité bactérienne du sol est estimée à environ 6500 taxa différents, pour une abondance totale située entre 10^7 et 10^9 cellules par gramme de sol (Torsvik *et al.* 2002). Diverses stratégies d'extraction et d'utilisation de cet ADN sont alors envisageables pour tenter d'appréhender cette complexité. Les diverses techniques d'obtention et d'exploitation de l'ADN métagénomique seront présentées au cours de cette étude.

J’ai donc participé au projet Européen « METAEXPLORE » en tant que doctorant au sein du groupe de Génomique Microbienne Environnementale (GME) dans le laboratoire Ampère à l’Ecole Centrale de Lyon, sous la direction de Pascal Simonet (Directeur de recherche, CNRS). Le groupe GME intervient dans le projet en tant que partenaire n°7, et apporte son expertise en écologie microbienne du sol. Le groupe est particulièrement impliqué dans l’étude métagénomique du sol de la station expérimentale de Rothamsted en Angleterre, au travers du consortium international Terragenome, financé par l'ANR « METASOIL» (Vogel *et al.* 2009). L’objectif de ce projet est d'étudier en profondeur le métagénome d'un sol particulier afin de constituer une base de données de référence. Le sol qui a été sélectionné pour mener à bien cette étude est celui de la station expérimentale basée à Rothamsted en Angleterre, de par l'importance des données accumulées sur ce sol au travers d'un suivi de plus de 150 ans. Le projet Metasoil, qui a été réalisé par Tom Delmont, ancien doctorant du groupe, s'est déroulé en parallèle à Metaexplore, et a permis de générer :

- une banque de données métagénomiques constituée de plusieurs séquençages directs de l'ADN métagénomique extrait du sol de Rothamsted. Une stratégie spéciale a été mise en place afin d'optimiser l'extraction de l'ADN et de maximiser la diversité obtenue (Delmont *et al.* 2011).

- une banque de clones fosmidiques contenant des inserts métagénomiques provenant du sol de Rothamsted. La banque contient 2 millions de clones et est dupliquée en deux emplacements différents (Ecole Centrale de Lyon, Ecully et Libragen, Toulouse).

C'est dans ce contexte qu'ont été initiés mes travaux de recherche, qui s’articulent autour de trois axes:

- la synthèse bibliographique du contexte scientifique où se situent les thématiques de recherche sur lesquelles j'ai travaillé. Cette synthèse correspond au premier chapitre de cette étude.
- la recherche d’enzymes d’intérêt, en me focalisant principalement sur les chitinases. C'est la première partie de mes travaux développée dans le deuxième chapitre de cette étude. Cette partie comporte 4 points important qui seront discutés séparément, incluant (1) La présentation du projet METAEXPLORE, (2) la recherche de séquence d’intérêt dans les métagénomes de Rothamsted par BLASTx, (3) le cribage génétique d’une partie de la banque de clone de Rothamsted à la recherche de gène d’intérêt, et (4) une approche d’enrichissement en chitine du sol de Rothamsted en microcosme afin d’étudier la réponse de la communauté bactérienne.
- le développement d’un outil biotechnologique innovant et complémentaire des approches métagénomiques et culturales classiques, afin d’améliorer la capture de gènes d’intérêt à partir de l’ADN métagénomique. Cette deuxième partie de mes travaux est développée dans le troisième chapitre de cette étude.

Comme expliqué précédemment, la rédaction des trois chapitres a été réalisée en langue anglaise. Cependant, un résumé en français complet a été prévu en début de chaque partie, synthétisant les démarches ainsi que les résultats les plus significatifs et ce afin de permettre l’accessibilité des mes travaux au plus grand nombre. Une conclusion générale vient clôturer cette étude, au cours de laquelle les perspectives seront évoquées.

GENERAL INTRODUCTION

-

MICROBILOGY AND METAGENOMICS

-

ENGLISH VERSION

With more than 3.5 billion years of evolution, and a world population estimated at $5x10^{30}$ cells, (Withman et al . 1998), bacteria succeed in colonizing most of the ecological niches on Earth thank to their incredible adaptation capacities. They have been identified in soils (Chase *et al.* 1956), water (LeChevallier *et al.* 1980), air (Imshenetsky *et al.* 1978), deserts (Luo *et al.* 2012), rain water (Morris *et al.* 2008), snow (Yan *et al.* 2012), ice (Kim *et al.* 2012), sediments from deep ocean (Quigley and Colwell, 1968), caves (Urzì *et al.* 2010), Earth poles (Vincent *et al.* 2004 ; Bian *et al.* 2012), acidic (Tyson *et al.* 2004), and hyper-salin environments (Sorokin *et al.* 2006), under high temperature (Stetter, 1999), in low nutrient conditions (Essoussi *et al.* 2010), and even in polluted sites (Hassan *et al.* 2008). Bacteria are also abundant in more familiar places for Humans, such as our own skin (Grice *et al.* 2009) and our digestive tract as well (O'Hara and Shanahan, 2006).

This astonishing adaptation capacity to a wide range of environmental conditions, sometimes extreme, is due to their colonization and diversification abilities. In fact, their short life cycle principally based on mitosis divisions is giving them a rapid way to enlarge their population size, by constituting huge clone colonies. For instance, the cellular division time of *Escherichia coli* under optimal growth conditions in about 20 minutes (Cooper and Helmstetter, 1968). This capacity of multiplication varies in intensity according to the species and their respective growth optimums. However, bacteria life cycle is clearly faster than developed organism among Eukaryotes, which often result in apparition of large population in the environment. This strong growth rate also result in apparition of mutations that can contribute to the adaptation potential of bacteria.

Mutations are occurring in the genetic code through modifications in the nucleic sequence. Several factors are known to induce mutations. They can be generated by the enzymatic systems involved in reparation and replication of the DNA. In fact, the DNA polymerase, which is responsible in maintaining and replicating DNA molecules, is actually known for committing errors on the sequence with a rate of 1 wrong base every 10 000 000 bases proceeded (Voliotis *et al.* 2012). Other factors such as UV light exposure and oxidative stress conditions are known to generate mutations by chemical reactions.

Generally, most of these mutations are resulting in deleterious effect on the cell, by modifying key genes that would most likely be knock out, or by producing an altered or non-functional protein. In that case, the bacteria that would harbor such a mutation will be less suitable for passing through the natural selection pressure from the environment. Nevertheless, sometimes mutations can result in a direct or indirect positive effect, by bringing a real selective advantage to the bacteria (Matic *et al.* 1997). This notion is resumed in the ecological concept of "fitness", which reflects the capacity of an individual organism to survive and to reproduce under the natural selection pressure. This selection is expressed through abiotic factors (temperature, moisture, nutrients…) and also biotic factors (inter and intra-specific relations). As a consequence, the adaptation potential of bacteria is partly explained by their rapid life cycle, and the inexorable occurrence of mutations which participate in shaping their genomes. This results in apparition of individual cells with higher fitness than the initial strain under fluctuating environmental conditions (Gordo *et al.* 2011). However, adaptation through mutations is a slow and "uncontrollable" phenomenon, which cannot explain alone the incredible potential of diversification and evolution of bacteria. Another driving force is at work as well: Horizontal Gene Transfers (Jain *et al.* 2002, Aminov *et al.* 2011).

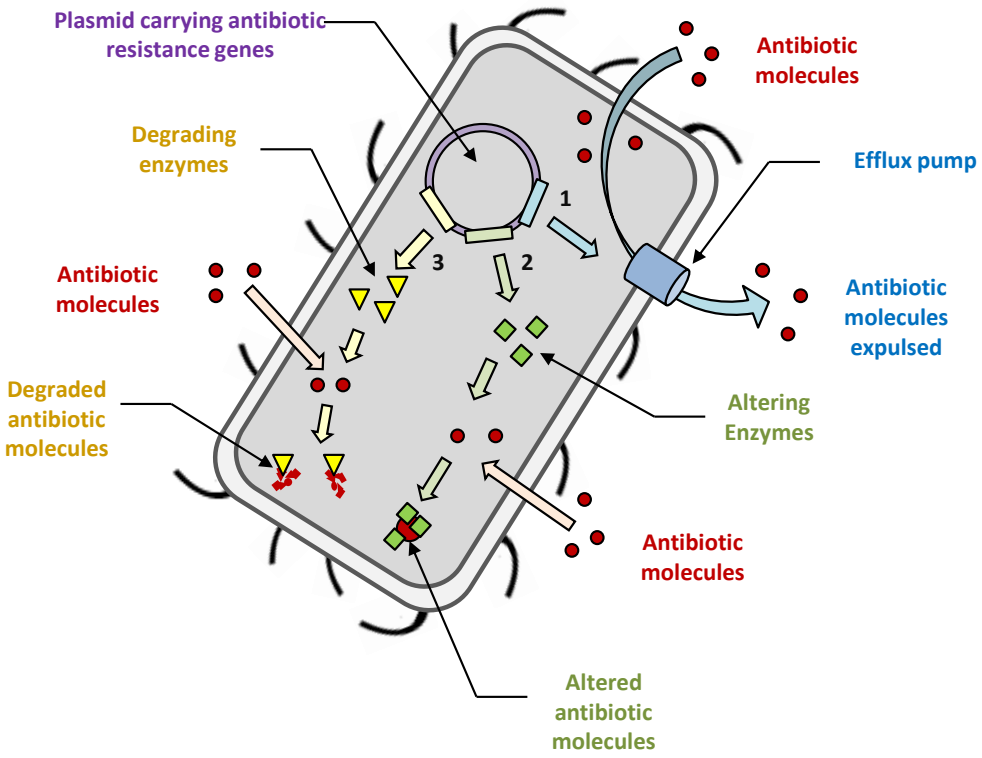


Fig.1: Illustration of several antibiotic resistance strategy brought by plasmid. This picture is representing how specific genes carried on a plasmid can act in the process of antibiotic resistance. The first path (1) in blue is representing the expulsion of antibiotic molecules via an efflux pump encoded by a plasmid gene. The second path (2) in green is displaying the inactivation of antibiotic molecules by binding compounds modifying the chemical structure. The third path (3) is presenting degrading enzymes that destroy the antibiotic molecules.

Contrary to the "traditional" vertical transmission of the genetic characters from parents to the descendants, observed in superior organisms during the sexual reproduction, prokaryotes can transfer and obtain a part of their genetic resources *via* specific direct or indirect exchange mechanisms. This phenomenon allows a certain amount of genetic shuffle among the bacterial population, which results in maintenance of genetic diversity. In fact, mitosis is an asexual reproduction way relying on clones, with only few genetic variations, resulting in a low genetic difference between the cells. This problem does not exist for species that undergo the sexual reproduction, where the descendants got both genetic materials from the two parents, which aims to maintain a certain degree of diversity. However, endogenic processes exist to generate genetic variability inside clonal population through mobile genetic elements such as transposons (e.g insertion sequences). Horizontal gene transfer is occurring rapidly and constantly among bacterial population, between closely and far related species. However, endogene processes that generate genetic variability within a clonal population exist, such as transposable element (e.g insertion sequences) (Fehér *et al.* 2012).

This flux of genetic information is strongly participating into the diversification of bacteria. For instance, these transfers allow a fast adaptation to harsh and fluctuating environmental conditions, by giving and spreading the appropriate genes among populations. They are also involved in colonization processes by giving the ability to access new ecological niches (Wiedenbeck and Cohan, 2011). Several genetic elements were characterized as being potentially transferable horizontally between bacterial cells: The mobile genetic elements, or the mobilome. This category includes numerous DNA sequences, genes, and vectors that will not be presented in this study (transposons, integrons, insertion sequences, phages, plasmids…).

However, plasmids are representing one of the most important categories among mobile genetic elements, as they are often carrying operons and genes involved in key functions responsible in a real selective advantage for the host cell (Smillie *et al.* 2010). The most relevant example is the bacterial adaptation to antibiotics, as the genes implicated in this phenomenon are carried by plasmids and easily transferred among the populations (Fig.1). Sometimes, the genes from plasmids are also known to be able to integrate the bacterial chromosome. (Andersson and Hughes, 2011). The adaptation to polluted sites is also another example, as some enzymes involved in the degradation of toxic compounds are present on plasmids as well (Liang *et al.* 2012).

There are 3 types of mechanisms for acquisition of genetic information through horizontal gene transfer (Fig.2):

- ***Transformation***: which involves integration of environmental DNA in the cell. This DNA ismost likely coming from dead cell (plants, fungi, insects, bacteria…) and requires specific gene that promote the natural uptake of this DNA (e.g *Acinetobacter baylyi*, Elliott and Neidle, 2011). Nevertheless, transformation can be stimulated in other species by diverse factors such as thermal or electrical shocks, resulting in formation of holes into the membrane. Artificially mediate transformation is a well known technique for integration of DNA in host strains (e.g *E. coli*) (Aune and Aachmann, 2010). Furthermore, full integration of the incoming DNA requires homologous recombination in order to integrate existing genetic structures in the cell, like the chromosome or plasmids. Transformation and recombination are important phenomenon that were extensively used in this study for the development of the Genefish strategy (Chapter 3).

- ***Conjugation***: which involves a direct exchange of DNA between two bacteria *via* the constitution of a linking channel called "*pilus*" or "sexual *pilus*" (Lederberg and Tatum, 1953). This kind of reaction occurs between two types of bacteria: (1) male bacteria (F+) that are giving DNA and are equipped with the F plasmid which encodes for the *pilus* synthesis. (2) Female bacteria that receive DNA (F-), and also carry a copy of the F plasmid, making them F+ as well. Some bacteria can integrate the genes from the F plasmid into their chromosome, and are able to exchange a part of their chromosome through conjugation with high frequencies. Conjugation is a frequent process between bacteria, and directly impacts on their adaptation capacities. For instance, this phenomenon is occurring a lot in soils, especially in the rhizosphere (Ding and Hynes, 2009).

- ***Transduction:*** which involve bacterial viruses in the process as vector of the genetic transfer from an infected bacterium to another one. A high specificity exists between bacterial species and their related phage. This is due to a tight relation of "co-evolution" of the phage-bacteria couple, implicating a constant adaptation of each partner, shaping their genomes each other's (Gomez and Buckling, 2011). During the infection cycle, viruses are integrating their own genetic material into the bacterial chromosome through recombination. This requires powerful enzymatic systems in order to promote recombination with high rates, even though the similarity is low. Two transduction phenomena exist: generalized transduction, which involves random and multiple insertions of the viral genetic material within the bacterial chromosome (e.g phage P22 and *Salmonella typhimurium* ; Ebel-Tsipis et al. 1972) ; and the specialized transduction where only one site of insertion exist (e.g phage λ and *Escherichia coli* at the *att* site; Campbell, 1962).

At this stage, two types of phages can be differentiated: (1) virulent phages that undertake lytic stage through the massive production of viral particles thanks to "the molecular machinery" from the infected cell. The bacterial cell will finally explode under the inner pressure generated by the massive synthesis of phage particles. In the case of the generalized transduction, this phenomenon will release some bacteriophage containing pieces of the infected bacterial genome that can be integrated along the next infection in another cell. (2) Tempered phages that rather undertake a "lysogenetic" cycle before the lytic stage. During this period, the bacteria can continue to live and divide normally with the dormant phage DNA inside their chromosome. In fact, it's only during this phase that the infected bacteria can eventually use the genetic material coming from another dead infected cell. This step can go on until the phage DNA is activated by environmental factors. The actual selective advantage that can result from such a phenomenon was demonstrated in several cases, like survival in soil under harsh conditions (Schuch and Fischetti, 2009), and also antibiotic resistance (Mlynarczyk *et al.* 1997). The enzymatic systems from the phage lambda that are involved in the recombination process were applied for the development of the Genefish tool (Chapter 3).

These 3 phenomena are occurring concomitantly into bacterial populations with relatively high frequency rates compared to mutations, which result in generating genetic diversity. Thanks these processes, bacteria succeeded in colonizing most of the ecological niches available on Earth.

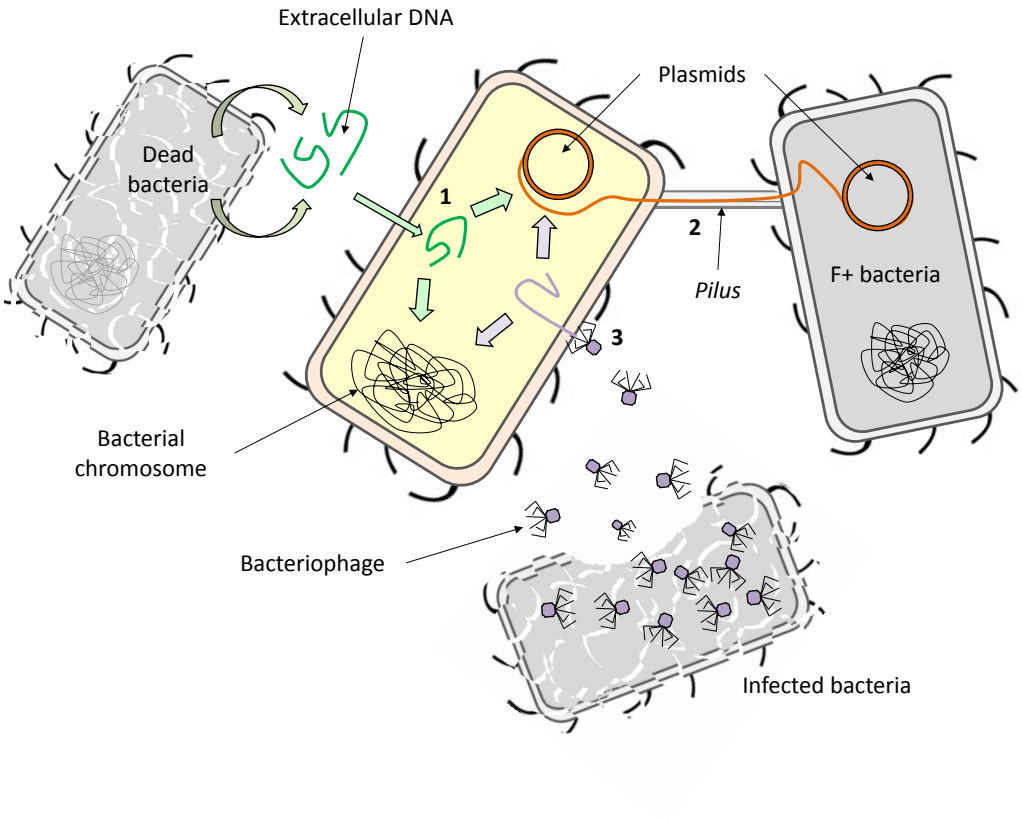


Fig.2: Schematic representation of the 3 different ways to horizontally transfer DNA in bacteria. Transformation is displayed in 1) with internalization and integration of environmental DNA from dead cells. Conjugation is presented in 2) through the transfert of plasmid DNA between to cells. Transduction is represented in 3), with bacterium infection by bacteriophages coming from a dead infected cell.

Furthermore, there are often playing important role in ecosystems. For instance, there are closely related to key functions such as degradation and mineralization of the incoming organic matter in many habitats. Some organic sources are considered as recalcitrant and would most likely be accumulated in environment without the action of microorganisms, such as cellulose (Stursová *et al.* 2012). They are also actively participating in the biogeochemical cycles of many elements such as carbon (Evan, 1976), nitrogen (Godfrey and Glass, 2011), and many others (phosphorous, sulfur, magnesium…). In addition, they are playing huge roles in the Human society, by participating in important economical sectors. For instance, they are strongly involved in agriculture through soil fertilization processes and maintenance (Babalola, 2010), by improving plant growth (Lucy *et al.* 2004), and also in controlling pest populations (Raaijmakers *et al.* 2002). They are also strongly involved in immunity in Human and animal health, by occupying the skin and the digestive track as a "natural defense" against pathogens (Thompson *et al.* 1978). In the same move, they are also participating in digestion of some aliments that we cannot assimilate without their help (Tappenden and Deutsch, 2007). Unfortunately, they are also responsible for many serious and benign infectious diseases, affecting Humans directly or indirectly *via* domestic animals and livestock breeding.

In addition to their undeniable implication in ecosystems, microorganisms are also massively used in applied sectors, where their properties are derived and employed in industrial or environmental processes. Thus, the microorganisms are involved in the food industry, particularly in the manufacture of fermented products such as wine, beer, cheese and cocoa. Some of their specificities such as enzymes are employed in elaboration of medicines and cosmetic products. Some microorganisms are also directly used as medicines known under the name of probiotics (Heller 2001; Gill 2003). In the environmental area, micro-organisms are used in bioremediation and sanitation, where their properties are derived in order to solve pollution issues on contaminated sites (Mikesková *et al.* 2012). Thus, there are participating in the transformation and/ or degradation of toxic and recalcitrant compounds such as pesticides and hydrocarbons that might contaminate soils and water (Furukawa, 2003).

It is within this context that takes place the European Project Metaexplore, which has funded my PhD researches. The main goal of this project is to discover and characterized novel enzymes of interest for industrial applications from environmental microbial communities. This project was initiated in the frame of the strong development of the new technologies in the field of microbial ecology, as well as a strong demand from industrials that are constantly seeking for novel activities in order to answer to their development needs. 18 partners are participating to this project, mostly Europeans, including academicals research groups from universities and industrial collaborators as well in order to test and optimize the potential discoveries. The molecules of interest that are searched in the frame of this project are bacteria and fungi enzymes able to degrade recalcitrant molecules (e.g natural compounds: cellulose, lignin, chitin; man-made compounds: pesticides). Degradation of such molecules can be valorized in different industrial sectors such as biofuel production, by enhancing ethanol production from primary plant biomass. The seafood industry is also concern, as they are producing a lot of chitinous waste (crab and shrimp shells) that can be valorized after being treated by chitinases, as the degradation products can be used in many sector such as cosmetology and pharmaceuticals. Transformation and degradation of man-made polluting molecules is used in bioremediation of polluted sites.

In order to do so, the project is focusing on a relatively novel and promising approach called “metagenomics”. Since the beginning of microbe characterization and still now, microbiologists are studying microorganisms thank culture techniques based on the isolation and growth of a single cell on plates. These approaches have revolutionized microbiology by bringing an access to these microscopic organisms in order to better apprehend and study them. The rises of molecular technologies and sequencing facilities allowed an in depth characterization of microorganisms at the genetic level, leading to the establishment of DNA databases and biological models. Many active molecules were isolated and employed in diverse sectors from cultivable organisms, such as antibiotics and enzymes. For instance, in 2000, more than 70% of commercialized antibiotics have been discovered from microorganisms isolated through culture techniques (Demain, 1999).

However, some studies brought evidences that question the true accessibility of microorganisms’ diversity by cultivable techniques. In fact, based on estimations from environments that are known to harbor a huge diversity, only 0.1-1% of the bacteria were cultivable (Amann *et al.* 1995).

As a consequence, a large majority is still out of the microbiologists reach, staggering our actuakl conception of the environmental bacterial diversity. For instance, new prokaryote phyla are discovered regularly, constantly enriching the tree of life. This tree was initially set up by Woese (Woese, 1987) by analyzing genes encoding for the 16S ribosomal RNA as a genetic marker in the prokaryote domain. In fact, the studies of this highly conserved gene allow retracing the evolution of the different branches based on modifications and similarities in the nucleic sequence. Officially, 29 phyla are accepted and recognized in the classification of bacteria (LPSN, Euzéby 1997). However, with the rise of the novel molecular technologies, numerous new phyla have been identified without actual cultivated representatives. In total, 36 phyla were characterized in 2002, with 13 uncultured ones (Hugenholtz, 2002). This number rose up to 53 in 2004 (Schloss & Handelsman, 2004) and new ones are still being discovered nowadays (Wylie *et al.* 2012). Strategies were developed in order to enhance the culture conditions, trying to match the environmental growth optimums of each strain. Representatives of *Verrucomicrobia* and *Acidobacteria* were obtained thanks to prolonged culture time under poor nutrient conditions (Janssen *et al.* 2002). Another technique consisting in isolating bacterial cells into small agar blocs from environmental samples, followed by an exposure to a wide panel of physic-chemical conditions (temperature, pH, moisture…) succeeded in obtaining micro-colonies of uncultivated bacteria (Zengler *et al.* 2002). However, these strategies still remain tedious and time-consuming, with only detection of few new species, mostly due to the complexity of factors involved in environmental bacterial growth (Ferrer *et al.* 2003; Gabor *et al.* 2004).

Since the early 90’s, a new approach allowing to circumvent the culture limitation has began: “Metagenomics” (Torsvik *et al.* 1990). The principle of this strategy relies on extraction of a maximum of DNA from the cells that are present in environmental samples (soil, water, sediment…). This so-called “metagenomic DNA” is the principal object of study, and contains genetic material from different organisms such as bacteria, fungi, insect, plants… (Rondon *et al.* 2000). Thus, all the microorganisms that are present into a specific habitat are theoretically accessible at the genetic level thanks this approach. For example, the first exploitations of the environmental metagenomic DNA were mostly focused on the characterization and estimation of the bacterial diversity (Amann *et al.* 1995; Hugenholtz *et al.* 1998).

However, the complete study of the metagenomic DNA remains an ambitious and complex challenge, as most of the environments harbor a huge diversity of microorganisms, such as soils and oceans. For instance, the bacterial diversity in soil is estimated to be approximately containing 6500 different taxa, and a total abundance of 10^7-10^9 bacterial cells per gram of soil (Torsvik *et al.* 2002). Diverse strategies of extraction and exploitation of the metagenomic DNA were attempted in order to apprehend and exploit this reservoir of diversity which can be illustrated by the famous phrase "*like looking for a needle in a haystack*". A presentation of the different strategies and techniques used for analysis of the metagenomic DNA are presented in this study (Chapter 1).

I personally participated to the European project « Metaexplore » in order to prepare my PhD degree under the supervision of Pascal Simonet (Research Director, CNRS, France) at the Environmental Microbial Genomics Group (EMG) in the Ampère laboratory (Ecole Centrale de Lyon, Ecully, France). The EMG group is involved in the project as Partner #7, bringing a strong expertise in soil microbial ecology. The group is particularly implicated in metagenomic investigation of the soil habitat through the study of the reference soil of the Park Grass research station at Rothamsted (UK), in the frame of the "Terragenome" international consortium, funded by the French ANR "Metasoil". The purpose of the consortium is to have an in depth characterization of a reference soil metagenome that can be used by the microbiologist community for comparison purposes. In the frame of the "Metasoil" project, a metagenomic database generated by shotgun pyrosequencing and a fosmid clone library were established from the soil of Rothamsted by Tom Delmont, former PhD student from the EMG group. The knowledge gathered all along the "Metasoil" project has been investigated in the frame of "Metaexplore" in order to seek for new enzymes of interests, with a strong emphasis on chitinases.

My PhD work is organized around tree major research axes:

- Chapter 1: A literature review of the scientific context, with a particular focus on metagenomics.
- Chapter 2: The metagenomic part of this study will be presented in this chapter. It includes 4 important points that will be discussed separately, including (1) the presentation of the METAEXPLORE project, (2) the in silico screening of the metagenomic dataset from the Rothamsted soil using BLASTx for identification of chitinases sequences, (3) the genetic screening of the fosmid clone library from the Rothamsted soil in search for genes of interest, and (4) a chitin enrichement strategy of the Rothamsted soil in microcosm for evaluation of the impact on the bacterial community.
- Chapter 3: The development of a novel biotechnological tool, designed for complementary investigation of the metagenomic DNA aside of the other strategies.

A global conclusion is ending this study, where perspectives are evocated.

References:

Andersson DI, Hughes D (2011) Persistence of antibiotic resistance in bacterial populations. *FEMS Microbiol Rev*. 35:901-911

Amann RI, Ludwig W & Schleifer KH. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*. 59:143-169

Aminov RI (2011) Horizontal gene exchange in environmental microbiota. *Front Microbiol*. 2:158

Aune TE, Aachmann FL (2010) Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed. Appl Microbiol Biotechnol. 85:1301-1313

Babalola OO. (2010) Beneficial bacteria of agricultural importance. *Biotechnol Lett*. 32:1559-1570

Baldauf SL, Bhattacharya D, Cockrill J, Hugenholtz P, Pawlowski J and Simpson AGB. The Tree of Life: A Overview. Craycraft J, Donoghue MJ, editors (2004) Assembling the Tree of Life Oxford University Press; 43-75

Bian F, Xie BB, Qin QL, Shu YL, Zhang XY, Yu Y, Chen B, Chen XL, Zhou BC, Zhang YZ (2012) Genome sequences of six Pseudoalteromonas strains isolated from Arctic sea ice. *J Bacteriol*. 194:908-909

Campbell A (1962) *Ado. Genet*. 11 :101-145
Chase HH, Davis JB and Raymond RL (1956) Mycobacterium paraffinicum n. sp., a bacterium isolated from soil. *Appl Microbiol*. 4:310-315

Cooper S and Helmstetter CE (1968) Chromosome replication and the division cycle of *Escherichia coli* B/r. *J Mol Biol*. 31:519-540

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol*. 77:1315-1324

Demain AL (1999) Pharmaceutically active secondary metabolites of microorganisms. *Appl Microbiol Biotechnol*. 52:455-463

Ding H, Hynes MF (2009) Plasmid transfer systems in the rhizobia. *Can J Microbiol*. 55:917-927

Ebel-Tsipis J, Fox MS, Botstein D (1972) Generalized transduction by bacteriophage P22 in Salmonella typhimurium. II. Mechanism of integration of transducing DNA. *J Mol Biol*. 71:449-469

Elliott KT, Neidle EL (2011) Acinetobacter baylyi ADP1:transforming the choice of model organism. *IUBMB Life*. 63:1075-1080

Essoussi I, Ghodhbane-Gtari F, Amairi H, Sghaier H, Jaouani A, Brusetti L, Daffonchio D, Boudabous A and Gtari M (2010) Esterase as an enzymatic signature of Geodermatophilaceae adaptability to Sahara desert stones and monuments. *J Appl Microbiol*. 108:1723-1732

Euzéby JP (1997) Bacterial phyla entry in LPSN:List of Bacterial Names with Standing in Nomenclature:a folder available on the Internet. *Int J Syst Bacteriol*. 47:590-592

Evans WC (1976) Microbial catabolism:its role in the carbon cycle and environmental significance:a colloquium organized by P. A. Williams (Bangor). *Biochem Soc Trans*. 4:452-455

Fehér T, Bogos B, Méhi O, Fekete G, Csörgo B, Kovács K, Pósfai G, Papp B, Hurst LD, Pál C (2012) Competition between Transposable Elements and Mutator Genes in Bacteria. *Mol Biol Evol*. [Epub ahead of print]

Ferrer M, Chernikova TN, Yakimov MM, Golyshin PN and Timmis KN. (2003) Chaperonins govern growth of Escherichia coli at low temperatures. *Nat Biotechnol*. 21:1266-1267

Furukawa K (2003) 'Super bugs' for bioremediation. *Trends Biotechnol*. 21:187-190

Gabor EM, de Vries EJ and Janssen DB (2004) Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environ Microbiol*. 6:948-958

Gill HS (2003) Probiotics to enhance anti-infective defences in the gastrointestinal tract. *Best Pract Res Clin Gastroenterol*. 17:755-773

Godfrey LV and Glass JB (2011) The geochemical record of the ancient nitrogen cycle, nitrogen isotopes, and metal cofactors. *Methods Enzymol*. 486:483-506

Gomez P, Buckling A (2011) Bacteria-phage antagonistic coevolution in soil. *Science*. 332:106-109

Gordo I, Perfeito L, Sousa A (2011) Fitness effects of mutations in bacteria. *J Mol Microbiol Biotechnol*. 21:20-35

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC; NISC Comparative Sequencing Program, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA (2009) Topographical and temporal diversity of the human skin microbiome. *Science*. 324:1190-1192

Hassan SH, Abskharon RN, El-Rab SM and Shoreit AA (2008) Isolation, characterization of heavy metal resistant strain of Pseudomonas aeruginosa isolated from polluted sites in Assiut city, Egypt. *J Basic Microbiol*. 48:168-176

Heller KJ (2001) Probiotic bacteria in fermented foods:product characteristics and starter organisms. *Am J Clin Nutr*. 73 :374S-379S

Horn SJ, Sikorski P, Cederkvist JB, Vaaje-Kolstad G, Sørlie M, Synstad B, Vriend G, Vårum KM, Eijsink VG (2006) Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides. *Proc Natl Acad Sci USA*. 103:18089-94.

Hugenholtz P, Goebel BM and Pace, NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol*. 180:4765-4774

Hugenholtz P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol*. 3:REVIEWS0003

Imshenetsky AA, Lysenko SV and Kazakov GA (1978) Upper boundary of the biosphere. *Appl Environ Microbiol*. 35 :1-5

Jain R, Rivera MC, Moore JE, Lake JA (2002) Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol*. 61:489-495

Janssen PH, Yates PS, Grinton BE, Taylor PM and Sait M (2002) Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Appl Environ Microbiol*. 68:2391-2396

Kim SJ, Shin SC, Hong SG, Lee YM, Lee H, Lee J, Choi IG and Park H (2012) Genome sequence of Janthinobacterium sp. strain PAMC 25724, isolated from alpine glacier cryoconite. *J Bacteriol*. 194:2096

LeChevallier MW, Seidler RJ and Evans TM. (1980) Enumeration and characterization of standard plate count bacteria in chlorinated and raw water supplies. *Appl Environ Microbiol*. 40:922-930

Lederberg J, Tatum E (1953) Sex in bacteria; genetic studies, 1945-1952. *Science*. 118:169-175

Liang B, Jiang J, Zhang J, Zhao Y, Li S (2012) Horizontal transfer of dehalogenase genes involved in the catalysis of chlorinated compounds:evidence and ecological role. *Crit Rev Microbiol*. 38:95-110

Lucy M, Reed E & Glick BR (2004) Applications of free living plant growth-promoting rhizobacteria. *Antonie Van Leeuwenhoek* 86:1-25

Luo X, Wang J, Zeng XC, Wang Y, Zhou L, Nie Y, Dai J and Fang C (2012) *Mycetocola manganoxydans sp. nov.*, novel actinobacteria isolated from the Taklamakan desert. *Int J Syst Evol Microbiol*. [Epub ahead of print]

Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, Denamur E, Elion J (1997) Highly variable mutation rates in commensal and pathogenic Escherichia coli. *Science*. 277:1833-1834

Mikesková H, Novotný C, and Svobodová K (2012) Interspecific interactions in mixed microbial cultures in a biodegradation perspective. *Appl Microbiol Biotechnol*. [Epub ahead of print]

Mlynarczyk G, Mlynarczyk A, Zabicka D, Jeljaszewicz J (1997) Lysogenic conversion as a factor influencing the vancomycin tolerance phenomenon in Staphylococcus aureus. *J Antimicrob Chemother*. 40:136-137

Morris CE, Sands DC, Vinatzer BA, Glaux C, Guilbaud C, Buffière A, Yan S, Dominguez H and Thompson BM (2008) The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *ISME J*. 2:321-334

O'Hara AM, Shanahan F (2006) The gut flora as a forgotten organ. *EMBO Rep*. 7:688-693

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*. 405:299–304

Quigley MM, Colwell RR (1968) Properties of bacteria isolated from deep-sea sediments. J Bacteriol. 95:211-220

Raaijmakers JM, Vlami M & de Souza JT (2002) Antibiotic production by bacterial biocontrol agents. *Antonie Van Leeuwenhoek*. 81:537-547

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol*. 66:2541-2547

Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EP and de la Cruz F (2010) Mobility of plasmids. *Microbiol Mol Biol Rev*. 74:434-452

Sorokin DY, Tourova TP, Lysenko AM and Muyzer G (2006) Diversity of culturable halophilic sulfur-oxidizing bacteria in hypersaline habitats. *Microbiology*. 152:3013-3023

Stetter KO (1999) Extremophiles and their adaptation to hot environments. *FEBS Lett*. 452 :22-25

Schloss PD and Handelsman J (2004) Status of the microbial census. *Microbiol Mol Biol Rev*. 68:686-691

Schuch R, Fischetti VA (2009) The secret life of the anthrax agent Bacillus anthracis: bacteriophage-mediated ecological adaptations. *PLoS One*. 4:e6532

Stursová M, Zifčáková L, Leigh MB, Burgess R and Baldrian P (2012) Cellulose utilization in forest litter and soil:identification of bacterial and fungal decomposers. *FEMS Microbiol Ecol*. 80:735-746

Tappenden KA, Deutsch AS (2007) The physiological relevance of the intestinal microbiota-contributions to human health. *J Am Coll Nutr.* 26:679S-83S

Thompson GE (1978) Control of intestinal flora in animals and humans:implications for toxicology and health. *J Environ Pathol Toxicol.* 1:113-123

Torsvik V, Goksoyr J and Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol*. 56:782-787

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS and Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 428:37-43

Urzì C, De Leo F, Bruno L and Albertano P (2010) Microbial diversity in paleolithic caves:a study case on the phototrophic biofilms of the Cave of Bats (Zuheros, Spain) Microb Ecol. 60:116-129

Vincent WF, Mueller DR. And Bonilla S (2004) Ecosystems on ice:the microbial ecology of Markham Ice Shelf in the high Arctic. *Cryobiology* 48 :103-112

Voliotis M, Cohen N, Molina-París C and Liverpool TB (2012) Proofreading of misincorporated nucleotides in DNA transcription. *Phys Biol*. 9:036007

Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, Van Elsas JD, Bailey MJ, Nalin R, Philippot L (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol*. 7:252

Whitman WB, Coleman DC and Wiebe WJ (1998) Prokaryotes:the unseen majority. Proc Natl Acad Sci U S A. 95:6578-6583

Wiedenbeck J, Cohan FM. (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev*. 35:957-976

Woese, C. R (1987) Bacterial evolution. *Microbiol Rev*. 51:221-271

Wylie KM, Truty RM, Sharpton TJ, Mihindukulasuriya KA, Zhou Y, Gao H, Sodergren E, Weinstock GM, Pollard KS (2012) Novel Bacterial Taxa in the Human Microbiome. *PLoS ONE*. 7:e35294

Yan P, Hou S, Chen T, Ma X and Zhang S (2012) Culturable bacteria isolated from snow cores along the 1300 km traverse from Zhongshan Station to Dome A, East Antarctica. *Extremophiles*. [Epub ahead of print]

Zengler K, Toledo G, Rappe M, Elkins J, Mathur E J, Short JM and Keller M (2002) Cultivating the uncultured. *Proc Natl Acad Sci U S A*. 99:15681-15686

CHAPITRE 1

-

SYNTHESE BIBLIOGRAPHIQUE DU CONTEXTE SCIENTIFIQUE

-

LA METAGENOMIQUE DU SOL

-

RESUME EN LANGUE FRANÇAISE

Dans cette section, l'état de l'art en matière de métagénomique est discuté. Cette partie est essentiellement basée sur une synthèse bibliographique axée sur la microbiologie du sol et les dernières avancées en matière de métagénomique. Le chapitre est structuré autour de 4 axes principaux, détaillant successivement :

- 1 : le sol en tant qu'habitat pour les microorganismes et en particulier les bactéries. Cette partie se focalisera également sur les protocoles d'échantillonnage du sol ainsi que les principaux facteurs susceptibles d'impacter sur la qualité des prélèvements.
- 2 : l'approche métagénomique, en détaillant les différentes techniques d'extraction de l'ADN à partir des échantillons environnementaux mais aussi en présentant les différentes possibilités d'exploration de l'ADN métagénomique. La technologie des banques de clones métagénomiques ainsi que les différentes possibilités de criblage seront également abordées.
- 3 : la métagénomique associée aux nouvelles technologies de séquençage à haut débit sera discutée dans un troisième temps, avec notamment un accent particulier porté sur la bioinformatique et l'analyse de données.

1. Le sol : un habitat pour les bactéries

Le sol représente l'interface entre la biosphère (espace contenant le « vivant ») et la lithosphère (l'enveloppe terrestre rigide de surface). La matière organique provenant des organismes vivants y est déposée et accumulée, puis mélangée avec la fraction minérale *via* de complexes processus physico-chimiques. Il en résultera la formation d'agrégats constituant le complexe argilo-humique (Paul and David, 1989). Cette conformation très caractéristique du sol offre un très large éventail de micro-habitats pour les microorganismes résidant dans cet environnement. Ainsi, les procaryotes, qui représentent la majeure partie de la biomasse du sol, colonisent et occupent toutes les micro-niches présentes en surface, mais aussi les pores et interstices des agrégats (Hassink *et al.* 1993). Les bactéries s'accrochent à leur substrat en secrétant des exo-polysaccharides. La quantité et la diversité des espèces bactériennes sont inversement corrélées à la taille des agrégats (Sessitsch *et al.* 2001). D'autres facteurs tels que la quantité de nutriments carbonés, le pH, l'eau et la température impactent aussi directement qualitativement et quantitativement les populations bactériennes (Van Veen *et al.* 1997). A cet égard, le sol est un environnement considéré comme étant plutôt fluctuant et changeant contrairement à d'autres écosystèmes plus stables (ex. les océans, lacs…), souvent soumis à de grands écarts thermiques, au gel et au dégel ainsi qu'à la sécheresse et à la saturation en eau. De plus, l'impact humain y est important, notamment par le biais des pratiques agricoles qui vont modifier qualitativement et quantitativement les communautés bactériennes (Van Elsas *et al.* 2002). Cela s'explique en partie par l'utilisation massive des produits phytosanitaires et leur rémanence dans l'environnement, et notamment les sols, avec un impact négatif sur les communautés microbiennes selon certains auteurs (Yen *et al.* 2009). Tous ces paramètres doivent être consciencieusement pris en compte lors de la campagne d'échantillonnage.

La profondeur du sol impacte aussi quantitativement et qualitativement la répartition des microorganismes (Whitman *et al.* 1998, Van Elsas *et al.* 2002, Veluci *et al.* 2006). Plus localement, la pente et les dépressions conditionnent la distribution de l'eau *via* les processus d'écoulement et de drainage. De plus, la végétation de couverture a un impact sur le sol en changeant la composition de la litière. Ces variations verticales et horizontales doivent être prises en considération avant de prélever.

Dans le cadre d'une étude visant à caractériser la communauté microbienne d'un sol homogène, la meilleure stratégie consiste à échantillonner en plusieurs points de la parcelle à une profondeur suffisante pour avoir une bonne représentation (20-25cm) et de mélanger les prélèvements (Kakirde *et al.* 2010). Les outils doivent être désinfectés à l'éthanol 70% au préalable afin d'éviter les contaminations. Après prélèvement, le sol, très sensible à la dessiccation, doit être analysé rapidement ou bien congelé en l'état, sous peine d'impacter drastiquement les microorganismes (Wallenius *et al.* 2010).

2. L'étude métagénomique du sol

Deux techniques d'extraction de l'ADN métagénomique ont été développées à partir des échantillons environnementaux :

- la technique dite « directe », faisant intervenir une étape de destruction cellulaire : « la lyse » directement dans l'échantillon d'intérêt (Ogram *et al.* 1987). L'ADN est ensuite extrait et purifié avant d'être analysé.
- la technique dite « indirecte », incluant une étape supplémentaire de séparation permettant l'isolement des cellules hors de la matrice environnementale (Holben *et al.* 1988). L'ADN est ensuite récupéré après la lyse cellulaire.

Chacune des deux approches présente avantages et inconvénients spécifiques au procédé utilisé. Le choix réside principalement dans la question scientifique adressée, sachant que l'ADN métagénomique obtenu présentera des différences tant qualitatives que quantitatives en fonction de l'approche sélectionnée.

La technique directe est rapide et permet d'obtenir de plus grandes quantités d'ADN (≈10µg/g^{-1} de sol) (Ogram *et al.* 1987), ce qui est largement suffisant pour entreprendre la plupart des techniques analytiques actuelles. En revanche, cette stratégie présente l'inconvénient d'être agressive envers l'ADN, ce qui diminue sa qualité. L'extraction directe est basée sur une lyse mécanique puissante, afin de détruire les membranes des microorganismes les plus résistants afin d'en extraire leur matériel génétique. En revanche, l'ADN s'en retrouve fortement morcelé, avec une longueur moyenne de fragments dépassant rarement les 10kb (Griffiths *et al.* 2000). De plus, cette stratégie ne permet pas de distinguer le matériel issu des cellules procaryotes des eucaryotes. L'ADN ainsi obtenu peut alors être hautement enrichi par du matériel génétique de champignons, plantes et insectes, ce qui induit une limitation potentiellement importante pour les études visant les bactéries (Tsai and Olson, 1991 ; Tebbe and Vahjen, 1993).

La technique indirecte permet de réduire considérablement cette limitation, en incluant une étape de séparation des cellules de la matrice environnementale. La technique la plus utilisée repose sur une séparation en gradient de densité par centrifugation à faible vitesse. Cette stratégie permet aisément de dissocier les eucaryotes, plus volumineux, des procaryotes et ainsi d'enrichir l'ADN métagénomique en matériel bactérien. Cependant, la technique est longue et fastidieuse, avec à la clef de faibles rendements d'ADN (≈0.1-1µg/g^{-1} de sol) ce qui limite considérablement les moyens d'analyses. En revanche, l'ADN peut être extrait avec des techniques moins agressives, comme la lyse enzymatique et alcaline, ce qui permet de préserver l'ADN et d'obtenir des fragments de grande taille (≈400kb) (Bertrand *et al.* 2005).

La technique directe est donc conseillée dans des approches faisant intervenir la PCR pour détecter des gènes particuliers mais aussi le séquençage et dans une certaine mesure, le clonage. A l'inverse, la technique indirecte est recommandée dans le cadre de l'élaboration de banques de clones à large insert, de manière à préserver les opérons et à garder l'information présente dans l'environnement des gènes d'intérêt (Bertrand *et al.* 2005).

La technique d'extraction est présumée être une source de biais lors de l'interprétation des résultats car chaque stratégie semble favoriser l'obtention d'ADN de phyla particulier à l'instar d'autres (Gabor *et al.* 2003). Cette différence est principalement due à la capacité que présentent certaines espèces à résister à l'étape de lyse cellulaire (Frostegård *et al.* 1999 ; Delmont *et al.* 2011a). C'est pourquoi le choix de la stratégie d'extraction de l'ADN métagénomique dépendra des objectifs fixés et des questions scientifiques abordées.

Une fois obtenu, l'ADN métagénomique peut être utilisé de différentes manières en fonction de l'objectif scientifique. Un schéma récapitulatif présente toutes les possibilités d'exploitation de l'ADN métagénomique (Fig.1). Pour la suite de ce résumé, les techniques de construction de banques de clones métagénomiques ainsi que le séquençage haut débit et la bioinformatique seront présentées.

3. Les banques de clones métagénomiques

L'élaboration et l'exploitation des banques de clones à partir d'ADN environnemental restent les approches les plus appropriées pour identifier et caractériser de nouveaux gènes. Cette approche repose sur la ligature de petits ou grands fragments d'ADN dans un vecteur. Ce mélange est ensuite utilisé pour transformer une souche bactérienne appropriée. Le choix des vecteurs dépend de la taille des fragments à cloner, tandis que le choix de la souche est effectué afin d'optimiser l'éventuelle expression des protéines et enzymes associées aux gènes clonés. En revanche, de par l'imposante diversité résidant dans la plupart des environnements, la taille des banques de clones doit être souvent conséquente afin d'augmenter les chances de détecter un gène en particulier.

Dans la plupart des cas, l'hôte des banques est *Escherichia coli* pour des raisons de convenance. Cependant, l'ADN métagénomique provenant de certains environnements, notamment le sol, est connu pour être riche en matériel provenant de micro-organismes possédant un taux de bases GC élevé, notamment les membres du phylum des Actinobactéries. De plus, les espèces inféodées à ce phylum sont reconnues comme étant des producteurs de molécules actives potentiellement exploitables comme les antibiotiques (Gonzáles *et al.* 2005). Malheureusement, les gènes de ces espèces sont faiblement compatibles pour être exprimés chez *E. coli*. D'après certaines estimations, seulement 20% des promoteurs des gènes des Actinobactéries sont potentiellement reconnus par *E. coli* (Strohl, 1992). Ainsi, certains vecteurs ont été développés afin de permettre le transfert depuis *E. coli* vers d'autres hôtes plus compatibles comme *Streptomyces spp* (Sosio *et al.* 2000).

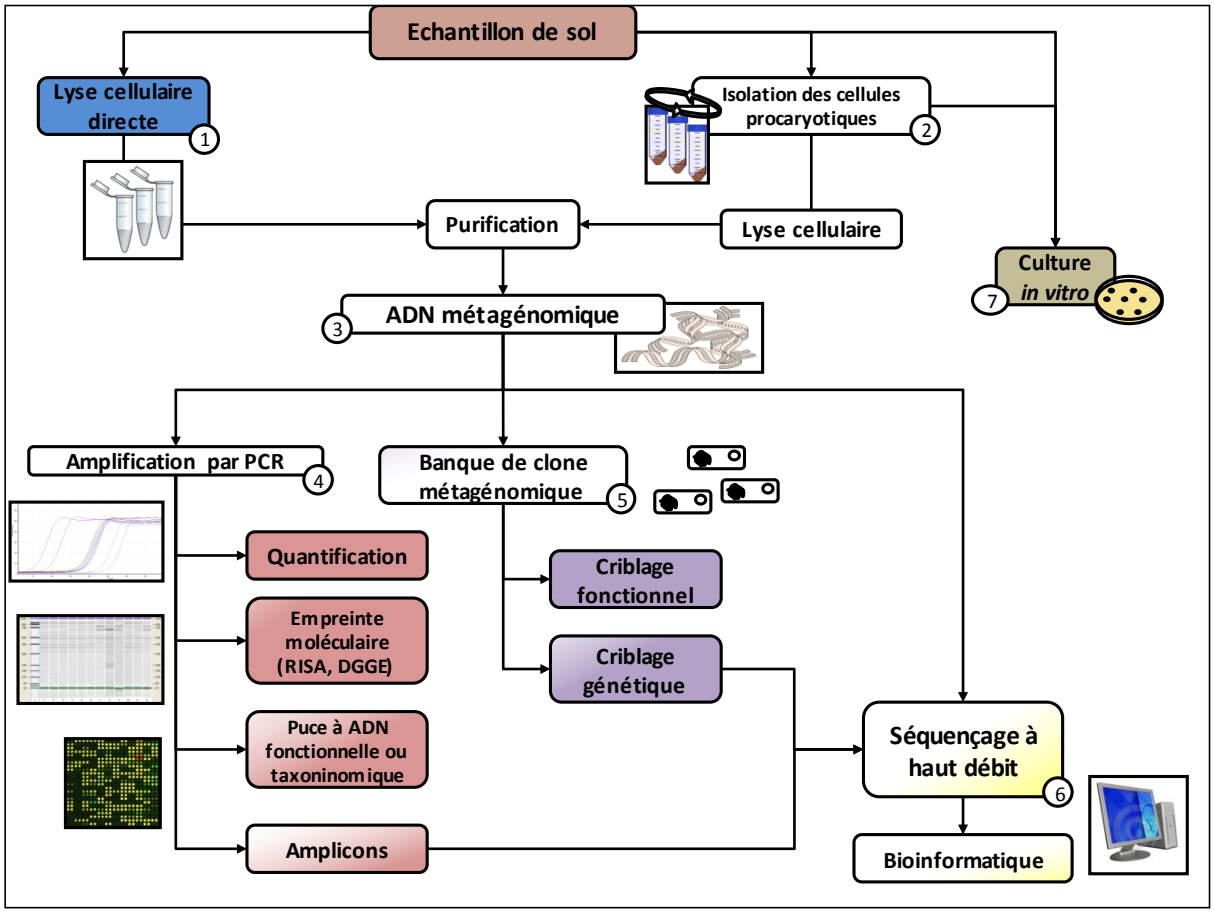


Fig.1 : Représentation schématique des différentes stratégies d'obtention et d'utilisation de l'ADN métagénomique. Les voies (1) et (2) représentent respectivement les techniques d'extractions directes et indirectes de l'ADN métagénomique (3) à partir d'un échantillon de sol. Des techniques d'amplification par PCR (4) peuvent être appliquées pour quantifier la présence d'un gène et/ou d'un taxon spécifique. Des approches préliminaires dites « d'empreinte moléculaire » permettent d'obtenir une image de la communauté microbienne. L'investigation peut être poussée plus loin via les techniques de puces à ADN, basées sur l'hybridation de produits PCR sur une lame présentant des fragments d'ADN connus, ou bien par séquençage direct des « amplicons » afin d'obtenir la composition fonctionnelle ou taxinomique d'une communauté microbienne. L'ADN métagénomique peut être utilisé pour la construction de banques de clones (5) qui pourront être criblées afin de rechercher des gènes (criblage génétique) et/ou des activités d'intérêt (criblage fonctionnel). Les clones d'intérêt ainsi sélectionnés pourront être séquencés intégralement afin d'obtenir la séquence complète de l'insert métagénomique. Enfin, l'ADN métagénomique peut être séquencé directement (« shotgun sequencing », (6), et analysé via les techniques de bioinformatique. En parallèle des approches métagénomiques, l'isolation de la fraction microbienne cultivable reste une approche sûre, complémentaire et nécessaire afin de caractériser au mieux les communautés de microorganismes (7).

Afin de détecter les gènes codant des activités d'intérêt, les banques doivent être criblées pour identifier les candidats positifs parmi la totalité des clones. On distingue deux types de criblage des banques de clones :

- le criblage génétique : basé principalement sur la séquence d'ADN insérée dans les vecteurs *via* des techniques d'amplification à l'aide d'amorces ou d'hybridation grâce à des sondes marquées (Demaneche *et al.* 2008*;* Ginolhac *et al.* 2004). Cette technique présente l'inconvénient majeur de nécessiter des informations préalables pour définir les amorces et les sondes à partir de séquences connues, limitant ainsi la découverte de nouveaux gènes.
- le criblage fonctionnel : consistant à rechercher l'expression du matériel génétique inséré *via* une activité ou un phénotype particulier aisément repérable permettant alors de repérer les clones dont l'insert contient des gènes d'intérêt. Par exemple, une activité enzymatique peut être détectée en ajoutant au milieu de culture un substrat adéquat dont les produits de dégradation présentent une coloration identifiable. En revanche, la faible expression du matériel métagénomique dans les souches classiques comme *E. coli* reste une limitation majeure (Gabor *et al.* 2004).

Même si cette stratégie reste efficace, l'élaboration et le criblage de banques de clones restent un travail couteux et fastidieux, nécessitant de puissants moyens techniques et financiers. Cette approche reste donc malheureusement hors de portée de bon nombre de laboratoires, faute d'accessibilité.

4. La métagénomique et les nouvelles technologies de séquençage :

Depuis le rapide et récent développement des technologies de séquençage à haut débit, l'utilisation de ces approches s'est répandue dans le domaine de la microbiologie environnementale (Sogin *et al.* 2006), aidée par une baisse significative des coûts (Novais *et al.* 2011). Le principe de la technique réside sur la détermination de la séquence nucléotidique des molécules d'ADN contenues dans une solution. Le résultat consiste en une succession de petites séquences appelées lectures ou « reads ». Les techniques de séquençage à haut débit peuvent être utilisées dans le cadre de plusieurs approches métagénomiques comprenant :

- le séquençage direct de l'ADN métagénomique ou « shotgun sequencing », permettant d'avoir une image globale de la communauté microbienne (Delmont *et al.* 2012). Cette approche peut également servir à assembler les séquences entre elles afin de reconstituer des fragments plus longs ou « contigs ».
- le séquençage de produits PCR : « amplicons », amplifiés à partir de l'ADN métagénomique et ciblant un ou plusieurs gènes spécifiques. Cette stratégie permet d'appliquer le séquençage à un ou plusieurs déterminants génétiques soit en vue d'une affiliation taxonomique (amplicon 16S, Bartram *et al.* 2011), soit pour inventorier des gènes fonctionels (Lüke and Frenzel, 2011).
- le séquençage de banque de clones métagénomiques, où les inserts sont séquencés afin d'être assemblés en contigs.

La profondeur du séquençage (volume de séquences) ainsi que la longueur des lectures dépendent grandement des différentes technologies mises au point par les prestataires sur leurs plateformes. Voici les deux plus grands prestataires :

- Illumina Genome Analyser (Solexa : 75-100pb la lecture pour un volume de 18-35Gbp ; HighSeq : 100-125pb la lecture pour un volume total de 250-600Gbp)
- Roche Sequencer 454 pyrosequencing FLX Systems (FLX : 300-500pb la lecture pour un volume de 450Mbp)

Toutes ces techniques génèrent une quantité conséquente de données, offrant l'opportunité aux microbiologistes d'avoir un nouvel accès aux communautés microbiennes par une approche totalement différente de la culture *in vitro*. Cependant, appréhender une telle masse d'information constitue un véritable défi technologique, nécessitant des ressources informatiques et des moyens de calcul conséquents. Le champ disciplinaire de la bioinformatique est un secteur qui se développe rapidement, principalement dû à la forte évolution des capacités de stockage et de calcul en informatique sous la forte demande des biologistes souhaitant pouvoir traiter leur données. L'analyse des données métagénomiques obtenues en séquençage nécessite un protocole très précis, débutant par un tri qualitatif sur filtre pour éliminer les séquences peu fiables (Niu *et al.* 2010, Cox *et al.* 2010) suivi de différentes stratégies d'analyse incluant l'utilisation de nombreux outils et logiciels adaptés en fonction de l'objectif (assemblage, analyse de diversité…). Le tableau 1 récapitule les principaux outils bioinformatiques disponibles pour analyser les données métagénomiques. Pour l'heure, de nombreux métagénomes sont disponibles sur les bases de données publiques ce qui permet de comparer différents environnements entre eux d'après les données issues de l'étude de la métagénomique (Delmont *et al.* 2011).

Cependant, malgré les progrès réalisés à la fois en matière de séquençage et en bioinformatique, les hypothèses et conclusions biologiques qui découlent de l'analyse de ces données doivent être considérées avec précaution. En effet, de nombreux biais accompagnent cette approche comme, par exemple, la distribution inégale des espèces en termes d'abondance dans les métagénomes, les problèmes de représentativité des banques de données publiques qui servent de référence, mais aussi les problèmes d'assemblage générant des chimères (Monier *et al.* 2011). Le fait de considérer des environnements riches en diversité comme le sol complique d'autant plus la problématique. En effet, la répartition des espèces en termes d'abondance est souvent inégale, avec une minorité d'espèces majoritaires, le reste étant faiblement représenté. D'autre part, le manque d'annotation dans les bases de données du fait de la non-cultivabilité de beaucoup de bactéries rend le problème d'autant difficile, limitant l'identification et l'exploitation de l'information métagénomique. En revanche, des habitats pauvres en diversité ont été analysés de manière quasi exhaustive par séquençage haut débit, comme par exemple les biofilms qui se développent dans les sédiments d'une rivière s'écoulant d'une mine acide (Tyson *et al.* 2004). A l'opposé, un environnement tel que le sol possède une grande majorité de séquences métagénomiques non-annotées dans les banques de données (≈10%), témoignant de la grande diversité qui y réside (Vogel *et al.* 2009). Cet aspect est crucial et doit être parfaitement pris en considération avant de tenter d'interpréter des données métagénomiques.

Outil	Référence	Format	Description
Trim	Cox *et al.* 2010	Logiciel	Vérifie la qualité des lectures dans les jeux de données Illumina.
CD-HIT-454	Niu *et al.* 2010	Logiciel	Permet de regrouper les lectures par homologie de séquence. Permet la suppression des dupliquas.
CD-HIT-EST	Huang *et al.* 2010		Permet de regrouper des séquences d'ADN ou de protéines à partir de fichier informatique.
H-CD-HIT	Huang *et al.* 2010		Permet de regrouper les séquences protéiques seulement, avec une efficacité améliorée.
Mothur	Schloss *et al.* 2009	Logiciel	Logiciel multitâche incluant des outils de nettoyage/trie, et des outils d'identification taxonomique. Permet de réaliser des comparaisons statistiques entre jeux de données.
QIIME	Caporaso *et al.* 2010	Logiciel	Idem
RAMMCAP	Li *et al.* 2009	Logiciel	Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline. Permet l'annotation de métagénome, et les comparaisons statistiques.
Megan	Huson *et al.* 2011	Logiciel	Permet l'annotation des séquences sur les bases de données par BLAST. Ne nécessite pas beaucoup de puissance de calcul (idéal pour les ordinateurs portables).
Treephyler	Schreiber *et al.* 2010	Logiciel	Permet de réaliser des affiliations phylogenetiques à partir de prédiction de gène.
CAMERA	Seshadri *et al.* 2007	En ligne	Série d'application en ligne permettant l'analyse métagénomique. Besoin de s'enregistrer en tant qu'utilisateur. Les jeux de données doivent être chargés, avec des temps d'analyse conséquents.
MG-RAST	Meyer *et al.* 2008	En ligne	Idem
WebMGA	Wu *et al.* 2011	En ligne	Idem. Pas d'enregistrement et pas de chargement en ligne des données. Temps d'analyse court.
STAMP	Parks and Beiko, 2010	Logiciel	Logiciel statistique développé pour analyser les fichiers générés par MG-RAST, RITA, CoMET. Permet de réaliser des comparaisons inter-métagénome et inter-groupe.
Galaxy	Goecks *et al.* 2010	En ligne	Source en ligne ouverte contenant une très large gamme d'outils pour de nombreuses applications. Permet de personnaliser son analyse.
Newbler	Roche Applied Science	Logiciel	Outils créé par Roche permettant l'assemblage et l'alignement des lectures obtenues en 454.
Velvet	Zerbino and Birney, 2008	Logiciel	Idem. Polyvalent (Illumina, Roche…)
Soap	Li R *et al.* 2009	Logiciel	Idem, mais uniquement pour le traitement des lectures obtenues en Illumina.

(Tab.1) Tableau récapitulatif présentant quelques-uns des outils bioinformatiques disponibles pour l'analyse des données métagénomiques. Une liste non-exhaustive est présente au lien suivant (http://seqanswers.com/wiki/Software/list; https://wiki.nbic.nl/index.php/NGS_Tools)

Conclusion

Les approches métagénomiques ont ouvert une nouvelle ère dans l'exploration des communautés microbiennes, en donnant accès à des organismes préalablement inaccessibles par la voie culturale, révélant ainsi une part plus importante de la diversité microbienne de nombreux environnements. Le sol est d'ailleurs tout particulièrement reconnu comme un immense réservoir de diversité bactérienne, source quasi inépuisable de nouvelles molécules d'intérêt comme les enzymes et les antibiotiques. La métagénomique a permis de confirmer l'étendue de cette diversité du fait de l'extraordinaire abondance de séquences d'ADN inconnues et de l'extrême difficulté à assembler un génome complet à partir de ces données. Même si des progrès impressionnants ont été réalisés en informatique, liés à l'accroissement des capacités de calcul et de stockage, l'exploitation et la valorisation des données générées par les nouvelles technologies de séquençage restent fastidieuses et complexes, soumises à de nombreux biais. Cependant, des techniques complémentaires existent afin d'aider à la caractérisation des communautés microbiennes, comme l'amélioration des approches culturales, ainsi que l'isolement des cellules environnementales (les approches dites de « single-cell »).

En résumé, l'exploration des communautés microbiennes deumeure un défi loin d'être atteint. Mais le contexte actuel, avec une forte progression des technologies expérimentales et *in silico*, est prometteur pour, à terme, espérer obtenir une vision de plus en plus exhaustive et exacte de la structure taxonomique et fonctionelle des communautés microbiennes environnementales complexes, comme celles des sols.

CHAPITER 1

-

BIBLIOGRAPHYCAL REVIEW OF THE SCIENTIFIC CONTEXT

-

SOIL METAGENOMICS

-

ENGLISH VERSION

CHAPTER 1

BIBLIOGRAPHICAL REVIEW OF THE SCIENTIFIC CONTEXT

SAMUEL JACQUIOD[1], LAURE FRANQUEVILLE[1], PASCAL SIMONET[1]

[1]Environmental Microbial Genomics Group, Laboratoire AMPERE, UMR CNRS5005, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Corresponding author: Samuel Jacquiod (samuel.jacquiod@ec-lyon.fr, phone: +33 (0)4 72 18 60 92, fax: +33 (0)4 78 43 37 17)

Abstract

Soil is commonly recognized as the richest habitat on Earth in terms of microbial resources and is regarded as an almost inexhaustible reservoir of new enzymes and natural chemicals, with potential relevant industrial applications. Furthermore, the interest of scientists toward this environment was raised by the ambitious challenge of understanding its biological functioning. The recent development of metagenomic approaches became one of the most popular strategies to bypass the inherent cultivation step, resulting in new discoveries through metagenomic libraries screening. This trend was also reinforced by fast development of high throughput sequencing technologies and bioinformatics, leading to the democratization of these approaches in the field of environmental microbiology. However, will these efforts be sufficient considering the incredible heterogeneity and complexity of soil matrix? We present and discuss here some of the various conceptual and technological improvements and limitations related to metagenomic approaches. We also indicate that alternatives are in progress to overcome these limitations such as the so-called single-cell approach. We also focus on new issues such as viral metagenomics which has recently received lots of consideration, as it opens a totally new area in the field of metagenomics, and as it brings novel insights in bacterial population at the ecological scale.

Key words: metagenomics, soil, bacteria, bacterial virus, single-cell genomics

Introduction

Because of its heterogeneity and the multiple interactions linking biotic and abiotic components, soil is commonly recognized as the most complex habitat on Earth, and the richest reservoir of microbial diversity as well (Robe *et al.* 2003). Soil provides a wide range of heterogeneous micro-niches and substrates, participating to the genotypic and phenotypic diversification of micro-organisms (Daniel *et al.* 2005). A single gram of forest soil was reported to contain up to 4.10^7 prokaryote cells (Richter *et al.* 1995), and up to 2.10^9 in agricultural grassland soil (Fierer *et al.* 2007a*;* Paul and David, 1989*;* Wommack *et al.* 2008). However, diversity estimations are found to be extremely different depending on the techniques and calculation methods used. For instance, estimations based on the number of different bacterial genomes in a single gram of soil vary by almost three orders of magnitude, from $1.6\ 10^4$ (Torsvik *et al.* 2002; Doolittle, 1999; Torsvik *et al.* 1998*;* Torsvik *et al.* 1996) to $8.3\ 10^6$ (Gans *et al.* 2005). Moreover, due to the uneven distribution of bacterial species, as well as the exclusion of rare and unrecovered species, even higher predictions may still underestimate the actual diversity level (Torsvik *et al.* 2002). Even if clear evidences are suggesting a huge diversity in soils, only a few species have been retrieved through cultivation (Hugenholtz *et al.* 1998; Curtis & Sloan, 2005). Furthermore, phylogenic and functional surveys demonstrate that soil bacterial diversity largely overwhelms the number of species retrieved from cultivation (Dinsdale *et al.* 2008; Breitbart *et al.* 2003; Rondon *et al.* 2000; Hugenholtz *et al.* 1998; Ward *et al.* 1990). In addition, previous comparison with microscopic observations of soil showed significantly more cells than culture based techniques (Richaume *et al.* 1993; Staley & Konopka, 1985). Similarly for viruses, direct counts with electron microscopy revealed at least 10^9 particles per gram of dried soil (Ashelford *et al.* 2003), whereas this represents 350-fold more viral particles than previously retrieved from culture-based techniques. Even if cultivation already proved that soil microbes are a good resource for natural product discovery (Schartz and Waksman, 1944), one of the most significant challenges in the field of microbial ecology is to unravel the actual taxonomic and functional diversity of a reference soil microbial community (Vogel *et al.* 2009). As a consequence, accessing more diversity has become a key step, with direct repercussion in our understanding of bacterial communities, and novel molecular discoveries as well. Furthermore, because of their particular life-style toward bacteria, from lysogenic to virulent forms, bacteriophages play a crucial role onto environmental bacterial communities. For instance, Gomez and Buckling (Gomez and Buckling, 2011) investigated the ecological and evolutionary dynamics between *Pseudomonas fluorescens* and its related lytic phage through soil microcosm experiments. This study highlighted the fast co-evolution of bacteria and bacteriophages in soil, shaping their own genomes each other. This fluctuating selection model, where bacteria hosts become hardy toward their contemporary phages, and inversely phages adapted to local bacteria, suggests strong involvement of phages in microbial community's evolution, even at the sub-specific level. Similarly, Schuch and Fischetti (Schuch and Fischetti, 2009) demonstrated that non-lysogenic bacteriophages infecting *Bacillus anthracis* strain increase host survival in soil. It was mostly due to the co-induction of biofilm formation, rather than undertaking traditional sporulation or starvation states. Maintenance of non-lytic phage DNA in the bacterial genome may contribute to advantageous phenotypic changes for hosts, like antibiotic resistance (Mlynarczyk *et al.* 1997), virulence (Banks *et al.* 2002), and fitness factor acquisition (Canchaya *et al.* 2004). All these aspects underline the non- negligible participation of bacteriophages to bacterial diversity stimulation through population regulation and horizontal gene transfer. Accessing the phage metagenome represents a big challenge for providing new insights on soil microbial population.

Hopefully, metagenomic approaches allow soil microbial investigation, bypassing the inherent cultivation biases, and are also recognized as the best option when combined with high throughput sequencing methodologies to retrieve the soil microbial genetic diversity and their associated chemicals (Vogel *et al.* 2009, Lefevre *et al.* 2008). In the same way, a metagenomic study based on sequences comparison from distinct environmental samples has reported a 10 times higher richness within virus population than bacteria, particularly in the rain forest soil (Fierer *et al.* 2007a). These preliminary metagenomic investigations suggest the prevalence of bacteriophages and their potential impact on bacterial communities in soil habitats. However, soil metagenomic exploration still remains a delicate task to achieve, considering the numerous biases coming along the process. The use of molecular tools in the field of microbiology has been already discussed, with regards to the mycosphere as well (Van Elsas and Boersma, 2011). The purpose of this review is to present the interest and limitations of metagenomic approaches, as well as some of the possible alternative strategies that are currently being developed to complete soil exploration, such as the single-cell approach and viral metagenomics.

1. Soil as a microbial habitat

Soil represents the interface between the biosphere and the lithosphere. Organic matter from plants, insects, animals and microbes is collected, accumulated and transformed through abiotic and biotic processes into humus, composed of stable macro-molecules, highly recalcitrant to decomposition. Humus half-life was evaluated to reach 2,000 years under natural conditions (Paul and David, 1989). Depending on location, humus can rather be accumulated (wetland, peat bog) or mixed to eroded mineral particles with different characteristics (granite, limestone) and size (sand, silt and clay). These interactions between organic and mineral matter result in constitution of aggregates, forming the clay-humus complex.

Formation and stabilization of clay-humus complexes into suitable aggregates is the dominant structural aspect of soil matrix (Daniel *et al.* 2005). Aggregate size can range from more than 2mm for macro-aggregates, down to the micro-scale for colloidal particles and microbial aggregates. Prokaryotes are largely dominant and constitute the major part of soil biomass (Hassink *et al.* 1993). They often adsorb onto particles, or adhere tightly by secreting exopolysaccharides. Soil microbes inhabit a wide range of micro-habitats including the chaotic surface of soil aggregates, as well as the complex cavitations and pore spaces between and inside aggregates (Foster *et al.* 1988*;* Hassink *et al.* 1993). In this context, it was shown that bacteria number and diversity is inversely related to the soil particle size (Sessitsch *et al.* 2001). Furthermore, it was shown that bacterial community composition is correlated to carbon nutrient level in soil (Fierer *et al.* 2007b). Contrarily to stable environments, soil is submitted to drastic changes depending on the geographical location such as dehydration or water saturation, temperature contrasts, froze/thaw cycles and nutrient loss. These random conditions are directly impacting on microbial communities, resulting in rapid responses and high adaptation potential (Kieft *et al.* 1987). As a result of these fluctuations, antagonism between microbes toward nutrition and space is strong and is considered to be a major survival feature (Garbeva *et al.* 2009). Antagonism is noteworthy at the molecular level, as soil bacteria are known to produce and secrete a wide range of secondary metabolites, with potential anti-microbial activities (Schirmer *et al.* 2005; Ginolhac *et al.* 2004*;* Rondon *et al.* 2000).

Moreover, abiotic factors such as soil organic matter, clay level, pH, water holding capacity and temperature, are known to significantly control microbial activities in soil (Van Veen *et al.*. 1997). The impact of human activities on soils is also a key parameter to consider. Agricultural practices are known to directly modify community structures, with potential reduction of the observed diversity (Van Elsas *et al.* 2002). Furthermore, anthropogenic compounds such as pesticides, and their massive use in agriculture resulted in novel environmental perturbations, impacting on soil microbial communities (Yen *et al.* 2009). All these factors are contributing to bacterial stimulation and selection, and should be carefully taken into consideration before undertaking sampling campaigns.

Soil depth will qualitatively and quantitatively affects the collected microbes. The upper region, often colonized with roots, is known to maintain numerous bacteria cells owing to easy degrading plant substrates, such as root exudates and dead cells content (Van Elsas *et al.* 2002). Deep soil shows significantly less biomass due to the low availability of organic carbon nutrients, and also anoxic conditions brought by compaction and water saturation. However, deep soil is known to contain most of the soil bacteria diversity, as it represents the major part of soil in terms of volume (Whitman *et al.* 1998). Cell density is generally greater in upper layers and decrease in subsurface, while phototrophic microorganisms (e.g., from the division Cyanobacteria) will only be located at the very top of soil (Veluci *et al.* 2006). At the field scale, local variations such as depressions and steeps will impact on water distribution and nutrient draining processes. The covering vegetation and fertilizers will change the oncoming organic matter and nutrients. These vertical and horizontal aspects should be carefully considered, especially regarding the scientific questions addressed. In the frame of a global soil microbial diversity survey, the best strategy consists in sampling deep enough in diverse locations and pool soil cores together prior to analysis. Pooling is advisable when a representative sample of the microbial community is needed, but not recommendable if the purpose is to target a specific microbial population (Kakirde *et al.* 2010). Cross contaminations between upper and lower layers may occur while sampling since the equipment will carry and transfer microbes up and down. In this case, equipment can be treated with ethanol and outer part of the soil core should be discarded. After sampling, soil should be handled rapidly, as it is strongly exposed to desiccation and temperature changes, resulting in drastic impact on microbial communities in terms of DNA extraction yields (Wallenius *et al.* 2010).

2. Metagenomic investigaion of soil

i. Metagenomic DNA and metagenomic approaches

As preliminary established in previous studies, metagenomics has been described as the isolation and cloning of environmental microbial DNA into specific vectors and domesticated host strains in order to establish clone libraries subsequently submitted to molecular and/or functional screening (Handelsman *et al.* 2004*;* Stein *et al.* 1996; Healy *et al.* 1995). However, since the recent and fast development of high throughput sequencing technologies (Edwards *et al.* 2006), many studies relied on sequencing of extracted metagenomic DNA to investigate microbial functional genes and taxonomic diversity.

Considering these new trends, metagenomics should be described as the genomic and functional analysis of the collective microbial assemblage found in environmental samples, as defined by Handelsman and collaborators (Handelsman *et al.* 1998). In accordance with this definition, as "*the functional and sequence based analysis of the collective microbial genomes contained in an environmental sample*", bacterial viruses should be enclosed in this important challenge too, as their DNA is often found within bacterial genomes. Generally, undertaking a metagenomic approach will aim for i) identification of functional genes, ii) estimation of microbial diversity, iii) whole community/specific population dynamics insights or iv) genome assembly of uncultured organisms (Rajendhran and Gunasekaran, 2008). Depending on these different purposes, the metagenomic strategy should be adjusted adequately.

However, the so-called metagenomic DNA can be handled through diverse strategies (fig.1) including preliminary whole-community PCR based approaches such as DGGE (Denaturing Gel Gradient Electrophoresis) (Nyberg *et al.* 2011) or RISA (Ribosomal Intergenic Spacer Analysis) (Ranjard *et al.* 2001). These fast and cheap preliminary fingerprint techniques allow detection of major differences between samples. Bacterial diversity can be deeply assessed through 16S rRNA-based taxonomic micro-arrays or direct sequencing of rrs amplicons. Quantification of specific taxa or functional genes can be achieved by qPCR approaches (Fierer *et al.* 2005). Cloning and sequencing of specific amplicons remain a reliable preliminary approach to roughly define who and/or what is there within the sample. However, high throughput sequencing is becoming the prevalent technology to deeply investigate metagenomic DNA, progressively outperforming cloning inventories. Nevertheless, cloning of metagenomic DNA prior to functional screening still remains the most efficient strategy to recover fully operational sequences from uncultured microbes that exhibit no or very low sequence similarity to known genes from databases (Kakirde *et al.* 2010). Furthermore, it is important to mention that only strategies looking genome wide from the metagenomic DNA pool can truly be called "metagenomic approaches", such as metagenomic clone library and/or total metagenomic DNA sequencing.

Even if goals and strategies remain different, a wide range of environments have been investigated through metagenomics, including soil (Delmont *et al.* 2011a*;* Vogel *et al.* 2009*;* Tringe *et al.* 2005; Voget *et al.* 2003; Rondon *et al.* 2000), rumen and guts (Brulc *et al.* 2009), planktonic marine environments (Beja *et al.* 2000*;* Breitbart *et al.* 2002), deep sea (Sogin *et al.* 2006), ocean (Nealson and Venter, 2007), acid mine sites (Tyson *et al.* 2004), arctic sediments (Jeon *et al.* 2009) and the Sargasso Sea (Venter *et al.* 2004). However, among all habitats, soil complexity often leads to major limitations in investigating microbial communities, mostly due to material extraction biases. This situation regarding soil habitat justify that, over the past two decades, viral metagenomic analysis were mostly focused on marine habitats. Likewise, it has been shown that viruses are the most abundant biological entities in oceans, with an estimation of 10^7 viral particles per ml of fresh seawater (Wommack and Colwell, 2000).

ii. Nucleic acid extraction

Extraction techniques can be classified in two categories, (1) direct extraction, where the lysis step is applied onto the soil samples (Ogram *et al.* 1987), or (2) indirect extraction, consisting in pulling out prokaryote cells from samples before lysis (Holben *et al.* 1988). Even if no major differences were observed after pyrosequencing (Delmont *et al.* 2011b), each technique was reported to yield DNA from different phylogenic origins (Gabor *et al.* 2003).

Direct extraction is less time consuming, and yields more DNA compared to other techniques (Ogram *et al.* 1987). However, a significant part of the extracted material is made of eukaryote DNA, which is an inherent bias for all studies focused on bacteria (Tebbe and Vahjen, 1993*;* Tsai and Olson, 1991). Among direct extraction techniques, the bead beating protocol is one of the fastest and most effective strategies, because it yields great amount of sheared DNA (Griffiths *et al.* 2000). Longer bead beating treatment will results in greater yields and proportionally more sheared DNA (Burgmann *et al.* 2001*;* Van Elsas *et al.* 1997). Despite lower DNA yields and higher time cost, indirect extraction techniques can overcome limitations of direct methods by isolating bacterial high molecular weight DNA (HMW) (Bertrand *et al.* 2005). The main strategy for prokaryote cell segregation consists in applying a density gradient through centrifugation. Because of their size and density, bacteria will cluster apart from eukaryote cells into a specific fraction of the gradient which can easily be recovered. However, co-extraction of low density eukaryote cells such as fungi spores and pico-eukaryotes (Moreira and Lopez-Garcia, 2002) is a possible source of contamination. Bacterial suspension can be pre-treated with DNase in order to remove extra cellular DNA from pre-lysed cells. However, due to the gradient, indirect methods are more biased toward unreleased microbes that are tightly associated to the matrix (Demaneche *et al.* 2001).

The major guide-line aspect to consider before selecting the DNA extraction strategy is the downstream purpose. If the aim is to constitute a metagenomic clone library devoted to large genetic determinant screening, the DNA size should be maximized in order to preserve the longest fragments. This will improve the probability of getting full operational sequences with promoters on single fragments. In these conditions, indirect extraction followed by a gentle extraction protocol will be preferred, such as enzymatic lysis in sodium dodecyl sulfate (SDS) and proteinase K (Bertrand *et al.* 2005*;* Courtois *et al.* 2001). On the other hand, in the frame of metagenomic survey aiming for: (i) investigation of short genetic determinants requiring small insert-libraries, (ii) PCR based approaches, (iii) or high throughput sequencing, a direct extraction protocol based on mechanical lysis and providing lots of highly sheared DNA will be more appropriate (Griffiths *et al.* 2000). It will allow disruption of recalcitrant cell forms such as spores, thus increasing the amount of recovered DNA.

However, DNA extraction step is commonly recognized as one of the major source of biases, as many cells will resist the membrane disruption step (Delmont *et al.* 2011a*;* Frostegård *et al.* 1999). In the frame of a diversity survey, it is critical to isolate DNA from as much as possible diverse microorganisms in order to get a good representation of the microbial community. Otherwise, downstream application might be biased against or in favor of specific groups (Feinstein *et al.* 2009*;* Liles *et al.* 2003).

Furthermore, differences in the diversity so-recovered between protocols have been reported, thus introducing uncertainty upon results obtained from a single protocol (Delmont *et al.* 2011a*;* Inceoglu *et al.* 2010). An alternative strategy consisting in using a combination of different existing protocols was already proposed and applied successfully (Delmont *et al.* 2011a; Lakay *et al.* 2007). However, due to the lack of normalization between soil metagenomic studies, comparisons between environments based on their metagenomes should be taken with caution. As a consequence, a DNA extraction protocol has been tested by different European laboratories, and validated by the International Organization of Standardization (ISO) as ISO standard 11063 "Soil quality – Method to directly extract DNA from soil samples" (Petric *et al.* 2011).

DNA extraction steps are subjected to contamination due to co-extraction of soil inherent substances such as humic acid, polyphenols, polysaccharides, nucleases (Harry *et al.* 1999; Frostegård *et al.* 1999*;* Zhou *et al.* 1996; Tebbe and Vahjen, 1993; Picard *et al.* 1992), but also heavy metals or pollutants depending on the soil type (Netto *et al.* 1991). Because of co-extracted soil contaminants, DNA extraction and purification are recognized as the most critical steps, as they may considerably affect the efficiency of downstream applications. For instance, clay particles are known to trap DNA by adsorption, thus strongly interfering with its recovery (Frostegård *et al.* 1999). Humic acids and heavy metals are known to be strong enzyme inhibitors (Wilson and Williamson, 1997) thus limiting both conventional PCR and qPCR assays (Audemard *et al.* 2004*;* Loge *et al.* 2002). However, PCR inhibitors effect can be reduced using serial dilutions (Queipo-Ortuno *et al.* 2008) or specific treatments like polymeric adsorbent Supelite™ DAX-8 (Schriewer *et al.* 2011). These contaminating compounds are also known to reduce cloning efficiency of soil DNA during library construction (Daniel, 2005). Furthermore, cell physiological state in soil samples is known to be an important qualitative parameter for DNA extraction. For instance, inactive states such as dormant or sporulated cells result in extraction of naturally sheared DNA, with smaller fragments than active cells (Bertrand *et al.* 2005). Nevertheless, alternative purification strategies such as nonlinear electrophoresis (Pel *et al.* 2009) were developed to remove contaminants such as humic acids. This technique was reported to currently match up with commercial protocols in terms of DNA quality and quantity (Engel *et al.* 2011).

As described above, DNA recovery is highly dependent on the methods selected according to the investigated environment. An exhaustive study devoted to viral DNA extraction aiming at comparing four different extraction buffers determined that glycine buffer results in better recovery of the inoculated phages into two agricultural soils (Williamson *et al.* 2003). Glycine is an amphiphilic amino acid, which may prevent hydrophobic interactions between viral and soil particles. Moreover, it was confirmed that alkaline pH increases the electrostatic repulsion by deprotonating ionizable sites, thus improving extraction efficiency. As the environment usually harbors higher isoelectric points than capsid proteins, the net charge on the bacteriophage surface is often negative, altering both adsorption and desorption processes onto soil particles. Chattopadhyay and Puls (Chattopadhyay and Puls, 2000) established a mathematical equation taking into consideration repulsive electrostatic forces and attractive van der Walls forces. In this study, the authors assume that the van der Walls forces prevent the process of sorption on clay. In addition, eukaryote and bacterial cells should be suppressed from samples to eliminate all sources of non-viral DNA before nucleic acid extraction step.

Due to the size and density of viral particles, tangential-flow filtration and cesium chloride ultracentrifugation are the most common methods for viral particle extraction (Thurber *et al.* 2009). Preferably, the viral suspension purity should be checked by epifluorescence microscopy observation after SYBR Gold or SYBR Green staining (Patel *et al.* 2007; Chen *et al.* 2001). Finally, a DNaseI treatment before capsid lysis can greatly reduce the remaining traces of linear DNA contamination. To date, detailed protocols for viral DNA extraction from complex environments such as soil are available (Casas and Rohwer, 2007). In addition, PCR screening tests can be performed to detect eventual presence of 18S and 16S, giving information on the contamination rate of the extracted viral genetic material.

iii. Metagenomic clone libraries

Environmental metagenomic library still remains one of the most powerful ways to highlight novelties from diverse habitats. It relies on cloning of rather small or large DNA fragments into domesticated host strain. However, numerous clones are required in order to encompass enough diversity, and to increase the chance in getting the desired activities and their related genes. The following vectors are commonly used depending on the fragment size and the library purpose: (1) plasmids for inserts ranging below 15kb, (2) cosmids and fosmids for inserts ranging from 15kb up to 45kb, (3) and BAC vectors for inserts greater than 50kb. Small insert metagenomic libraries are interesting for the detection of gene products encoded by small genetic loci, such as most enzymes or genetic determinants of antibiotic resistance (Parsley *et al.* 2010*;* Riesenfeld *et al.* 2004). Usually, *E. coli* is the universal host strain, but other strains have been used for functional detection of bioactive compounds such as *Streptomyces lividans*, *Rhizobium leguminosarum* or *Pseudomonas aeruginosa* (Singh *et al.* 2009). Actinobacteria phylum represents a great source of novelties, as their representatives are known to produce a wide range of active compounds (Gonzáles *et al.* 2005). On average, between 10 and 20 secondary metabolites can be synthesized by some Actinobacteria species, mostly from *Streptomyces* genus (Omura *et al.* 2001). Thanks to GC content segregation, metagenomic libraries can be enriched with Actinobacteria DNA in order to increase the fragment proportion related to secondary metabolite production (Kauffmann *et al.* 2004; Schloss and Handelsman, 2003). However, it was reported that *E. coli* as a host can recognize only 20% of Actinobacteria promoters (Strohl, 1992), making most functional screening inefficient. This aspect has motivated the development of an *E. coli–Streptomyces* artificial chromosome vectors (ESAC), allowing efficient transfer of former *E. coli* library contents into *Streptomyces* as a new host for expression attempts (Sosio *et al.* 2000). In addition, novel broad-host range vectors are currently being developed and tested in order to bring new possibilities and improve the current cloning strategies (Bartosik *et al.* 2011). Another strategy for microbial community enrichment toward specific groups consists in environmental pre-treatment with specific compounds (Dumont and Murel, 2005). Alternatively, the SIP (Stable Isotope Probing) can be applied to target the active population feeding on specific ^{13}C enriched substrates (Radajewski *et al.* 2000). The ^{13}C labeled DNA can be specifically recovered by differential centrifugation techniques in order to constitute enriched libraries toward particular microorganisms. This strategy was successfully used to shift libraries toward specific activities when compared to non-enriched ones (Schwarz *et al.* 2005).

Clone libraries screening can be achieved through two main approaches: (1) the functional screening, which relies on clones' metabolic activity coupled to phenotype selection techniques; (2) the molecular screening or sequence-based screening, which can be performed through full-insert sequencing, PCR amplification of specific genetic determinants, or labeled-probe hybridization.

(1) The first approach is commonly used to mine specific genes encoding novel enzymes or drugs (Fierer *et al.* 2007*;* Langer *et al.* 2006*;* Lorenz and Eck, 2005; Courtois *et al.* 2003; Rondon *et al.* 2000). In this context, global sequence analysis is not required and only positive clones will be further sequenced. The main limitation regarding this approach is that a great proportion of environmental genes are displaying low or missing expression into *E. coli* host. For instance, it has been evaluated that only 40% of the metagenomic resource is suitable for an *E. coli* host expression (Gabor *et al.* 2004). This is mostly due to the high divergence in sigma factors specificity between uncultured environmental bacteria and *E. coli*, resulting in significant reduction of mRNA transcription from metagenomic material. Functional screening enhancements are needed such as high-throughput screening processes, novel host strains, as well as novel types of screening. As an example, the SIGEX approach (Substrate Induced Gene Expression) based on a green fluorescent protein as a reporter of gene activity seems to be a really promising strategy (Uchiyama *et al.* 2005). This system is also compatible with the fluorescence activated cell sorting (FACS), which allows a fast library screening. The METREX strategy is an implementation of the SIGEX devoted to screen for quorum-sensing inducers in clone libraries (Williamson *et al.* 2005).

(2) The second approach described here is the sequence-based screening, which encompasses several techniques. One of these possibilities involves complete sequencing of clone inserts followed by taxonomic affiliation thanks phylogenic marker genes (Riesenfeld *et al.* 2004). The use of microarray technology as a tool for metagenomic clone library screening has been reported and can provide insights on complex community composition and predictable associated functions (Sebat *et al.* 2003). However, this approach is limited because it relies on the co-cloning of taxonomic markers, mostly *rrs* genes. In the meantime, strategies such as probe hybridization have been reported to be an efficient way to detect specific genes in clone libraries (Demaneche *et al.* 2009*;* Ginolhac *et al.* 2004). This approach is relying on clone spotting onto high density membranes, and a hybridization-based screening using cloned genes or PCR products as specific radio-labeled probes. In addition, promising complementary strategies such as random mutagenesis or sub-cloning help analyzing positive clones prior to full-insert sequencing by focusing on the genetic determinants of interest among the full-inserts (Glogauer *et al.* 2011; Hu *et al.* 2010; Kim *et al.* 2009).

Furthermore, massive sequencing and direct datasets evaluation is also used in this type of screening (Edwards *et al.* 2006). Unfortunately, complex environments such as soil drastically decrease the chance of getting overlapping sequences for assembling purposes, revealing the limits of computational systems to handle such datasets. For instance, 140Mb sequences were generated from a farmland soil clone library, and still, less than 1% of the sequences shared enough overlaps for contigs assembly (Tringe *et al.* 2005). Based on the assumption that all species are equally abundant in one gram of soil, a minimum of 10^6 BAC clones (100kb inserts) would be necessary to encompass the estimated 1000 Gbp of microbial genome sequences (Handelsman *et al.* 1998).

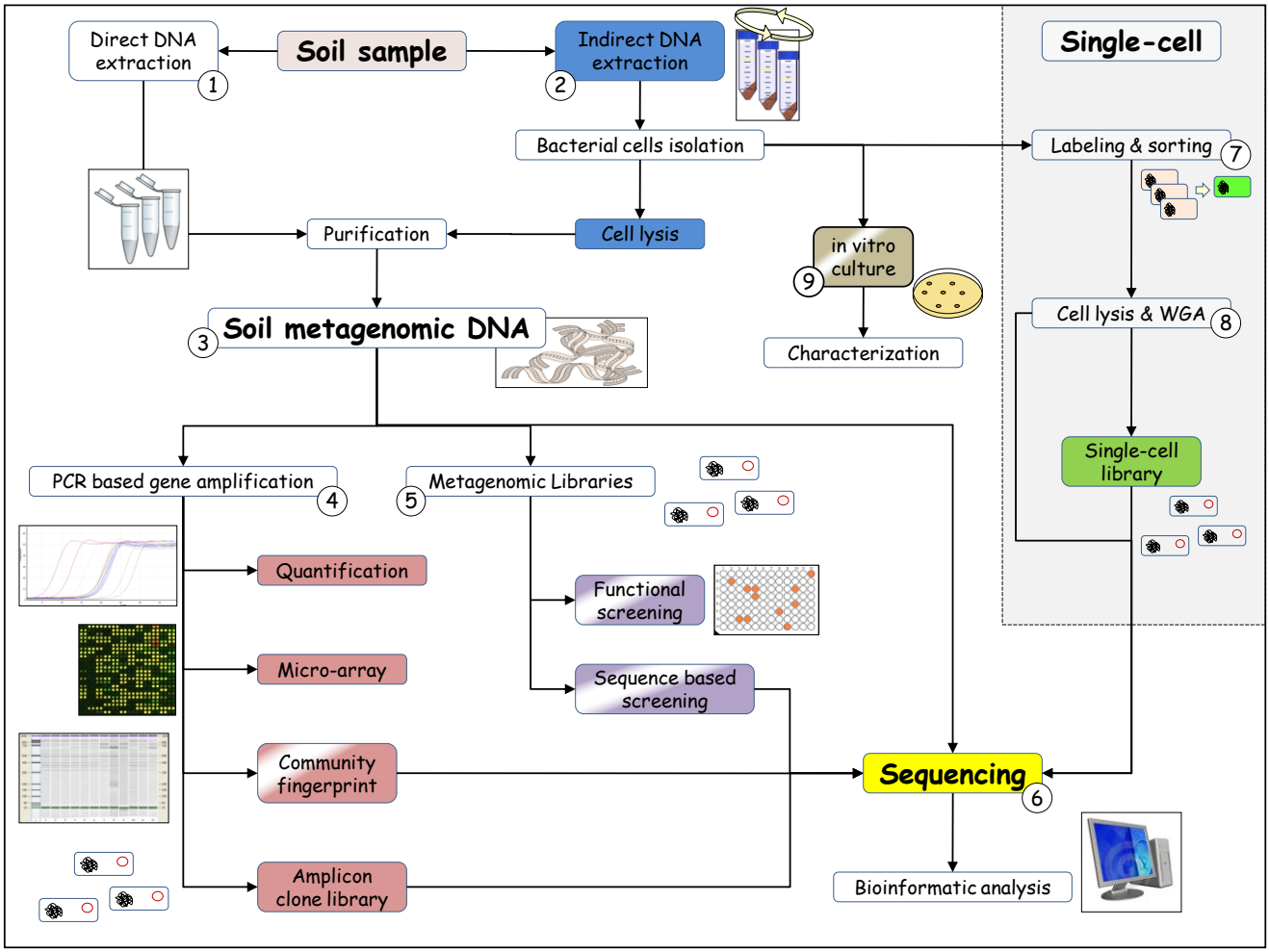


Fig.1: Overview of approaches currently available for investigation of soil microbial communities from metagenomic DNA. Metagenomic DNA can be recovered through (1) direct extraction including cell lysis within the soil samples, or (2) indirectly thanks preliminary isolation of the bacterial cells before lysis. The so-obtained soil metagenomic DNA (3) will vary in quantity and quality regarding the type of extraction and purification protocol previously applied. Metagenomic DNA can be handled thanks different approaches regarding the questions addressed (Van Elsas et al. 2008). (4) PCR based approaches will allow detection and quantification of specific genetic determinant of interest, with further possibilities like amplicon plasmid cloning for preliminary diversity analysis, or fingerprinting for whole-community scheme assessment. Micro-array analysis targeting taxonomic markers such as 16S rRNA genes, or any functional genes can be performed in order to get in-depth diversity profile. PCR product of interest can also be sent to sequencing directly, or after plasmid cloning. (5) Metagenomic clone library remains one of the most promising strategy for the isolation of whether large or small genetic determinant from uncultured bacterium, with potential encoding for new enzymes. Functional screening of metagenomic clone library is a direct and efficient way to isolate clones harboring interesting phenotypes, even though large clone screening is required. Sequence based screening is relying on DNA screening techniques such as PCR, direct sequencing or hybridization. (6) Direct sequencing of the metagenomic DNA will provide huge quantity of data, but unfortunately not enough for genome assembly. However, it will provide enough information for read annotation against databases and assembly into contigs if the read coverage is sufficient. In the case of clone library sequencing, insert sizes are low enough to allow clone pooling with good read coverage after assembly. Another approach consisting in working on isolated cells called single-cell allow full genome assembly with reduced complexity while compared to traditional metagenomics. It relies on cell labeling and sorting (7) for further downstream analysis thanks Whole Genome Amplification (WGA) (8). The so amplified genomic DNA can rather be used for cloning purposes, or directly sent for sequencing. (9) Bacterial isolates can also be obtained thanks traditional culture based approaches, with specific media for growth improvement.

In comparison, other genomic and metagenomic approaches such as the Human Genome Project (www.ncbi.nlm.nih.gov) and the Sargasso Sea projects (Venter *et al.* 2004) respectively yielded 3Gbp and 6Gbp. In the case of a study aiming at specific genetic determinant findings, up to 57500 40kb-fosmid clones would be necessary in order to reach 99% probability in detecting a given gene. This prediction is based on the assumption that a given gene might occur one time in every 100 environmental bacterial genome, with an average size of 5kb (Leveau *et al.* 2007). These observations highlight the huge diversity level in soil, as well as the required efforts to accomplish complete assembly of a single genome out of a soil metagenome. As a matter of fact, this confirms that metagenomic is still an emergent field, requiring major sequencing and bioinformatics improvements, as well as alternative and complementary approaches to help at filling the gaps.

Unfortunately, metagenomic approaches developed to study their bacterial counterpart cannot be directly applied to phages for several reasons: (1) PCR-based methods are not conceivable to investigate viral diversity because phages do not share any conserved genetic element, like *rrs* for bacteria (Edwards and Rohwer, 2005); (2) Standard cloning techniques cannot be directly used because of the presence of modified nucleotide bases in phage DNA (Xu, 2002; Warren, 1980); (3) Isolation of adequate amount of genomic material for large or even small insert library construction is not advisable because of small fragment length and poor extraction yields. For instance, 10^9 viral particles in 1g of soil deliver subnanograms quantities of DNA or RNA ($5.5\ 10^{-17}$g of DNA per viral particle); (4) Phage DNA fragments larger than 2kb cloned into cosmid vectors are known to be lethal for host cells because of toxic products expression, such as holins or lysozymes.

However, some companies are now providing appropriate facilities and new technologies to constitute reliable viral libraries. Linker Amplified Shotgun Libraries "LASL" (http://www.sci.sdsu.edu/PHAGE/LASL/index.htm), developed by Lucigen Corporation (Middleton, WI) is one of the most useful cloning technique and was described for the first time by Rohwer and collaborators (Rohwer *et al.* 2001). This technique is leaning on viral DNA fragmentation in order to avoid the expression of viral lethal genes into the host. Next step relies on viral DNA amplification, with conversion of modified bases into regular ones. Specific adaptors are used to facilitate ligation into low-copy vectors. The LASL method coupled to tBLASTx annotation was wisely used to characterize viral communities from desert, prairie and rain forest soils (Fierer *et al.* 2007). More than 10^{10} OTUs were detected thanks to this approach, with up to 10^3 viral genotypes per gram of rain forest soil. This estimation is 1000-fold higher than predictions made from aquatic environments. Bacteriophage types observed into soil samples were distinct and more diverse than aquatic environments such as marine sediments (Breitbart, 2004), or sea-water (Breitbart, 2002). Actually, the viral assemblages from these three different soil metagenome librairies were almost unique (Srinivasiah *et al.* 2008; Fierer *et al.* 2007). Another method, called Whole Genome Shotgun Library "WGSL" (Henn *et al.* 2010) was established as a standardized clone construction process, with the advantage of not requiring linker ligation and PCR amplification steps. In fact, the WGSL yielded complete assembled viral genomes regardless of a reasonable amount of input DNA (0.5ng). A study comparing LASL and WGSL has demonstrated the benefit of bypassing the cloning step (Henn *et al.* 2010).

Due to technical limitations, most of the studies are focusing only on DNA phages. Isolation of RNA virus genome content requires specific handling precautions, and need to be substantially amplified and reverse transcribed prior to cloning (Thurber *et al.* 2009). Nevertheless two RNA viral metagenomic libraries were established from feces (Zhang T. *et al.* 2006) and marine samples (Culley *et al.* 2006). As far as we know, soil has not been investigated, yet. Likewise, for ssDNA phages, random-priming reverse transcriptase and strand-displacement DNA polymerase represent alternative strategies for cloning and sequencing purposes (Edwards and Rohwer, 2005).

3. Metagenomic and next generation sequencing

i. High-throughput sequencing technologies

High-throughput sequencing, or next generation sequencing (NGS), became a powerful approach to access genetic diversity within environmental microbial communities (Sogin *et al.* 2006). Studies have been investigating microbial diversity in soils (Vogel *et al.* 2009; Roesch *et al.* 2007), deep sea ecosystems (Sogin *et al.* 2006), and also phage populations in diverse environments (Dinsdale *et al.* 2008). Since NGS became affordable, their use has become more and more important in the field of environmental microbiology (Novais *et al.* 2011). Taxonomy surveys based on metagenomic DNA are progressively becoming more important than previous PCR and cloning inventories. For instance, 6 millions *rrs* sequences were obtained from an Arctic toundra soil by assembling 125bp paired-end Illumina reads, bringing access to rare representative bacteria (i.e., GAIIx genome analyser, Hiseq 2000 platform, and MiSeq genome analyser) (Bartram *et al.* 2011). Furthermore, workflows are currently being designed and tested to reduce the impact of PCR and sequencing artifacts (Schloss *et al.* 2011). However, in depth community investigation based on *rrs* amplicon sequencing is also known to be biased depending on the hyper variable region selected (Kumar *et al.* 2011).

NGS provide substantial amount of data, thus increasing the detection sensitivity due to greater chance in getting DNA reads from rare representative species (Ronaghi *et al.* 2001). However, the sequencing depth and read accuracy level provided by NGS strongly differ between platforms: (1) Illumina Genome Analyser (Solexa, 75-100bp reads, 18-35Gbp per run), (2) Applied Biosystems SOLiD system (SOLiD, 50bp reads, 30-50Gbp per run) and (3) Roche Sequencer 454 pyrosequencing FLX System (FLX, 300-500bp reads, 450Mbp per run). Pyrosequencing accuracy is now matching up the quality achieved *via* Sanger based sequencing (Fordyce *et al.* 2011; Huse *et al.* 2007), which is an important parameter as read quality and length directly impact on assembly efficiency (Suzuki *et al.* 2011; Bartram *et al.* 2011). Furthermore, this technology is rapidly upgraded, leading to a gradual increase in read length (Margulies *et al.* 2005). For instance, read length was on average around ≈400bp in 2010, and reached up to 800bp in 2011 (http://454.com). Read length is also very critical as it directly limits annotation possibilities against known genes in databases (Wommack *et al.* 2008). However, it still remains more cost and time consuming per sequenced nucleotide than Illumina technologies (Hugenholtz and Tyson, 2008). For example, the new technology i.e., Hiseq 2000 platform from Illumina is ~50-12,000 fold cheaper per sequenced megabase than pyrosequencing, with improved high standard quality (Bartram *et al.* 2011).

Finally, choosing a NGS technology greatly depends on the addressed scientific goal, as it determines the sequencing depth. Illumina technologies are advisable in the frame of extensive surveys, requiring deep amplicon and/or metagenome sequencing. On the other hand, pyrosequencing is recommended in studies aiming for assembling purposes or whole microbial community overview.

Pyrosequencing relies on PCR, and will be sensitive to the same limitation than traditional PCR-based approach, such as inhibitors. As a consequence, DNA quality and purity should be optimal in order to avoid PCR inhibition. Hopefully, commercial DNA extraction kits are providing suitable quality DNA for downstream sequencing application. The DNA matrix size should also be taken into consideration, as it will vary depending on the extraction protocol applied. Pyrosequencing read length is usually ranging from 300bp up to 500bp, which is appropriate considering the DNA fragment size obtained from any protocol (Metzker, 2005). However, as this technology rapidly upgrades, the matrix DNA size may become an important criterion in sequencing optimization. Another crucial topic concerning NGS is the inherent issues toward AT and GC sequences content. AT-rich regions seem to lack coverage due to lower denaturation temperature (Dohm *et al.* 2008). However, it was shown that reduced temperature during PCR extension step resulted in better coverage of AT-rich regions (López-Barragán *et al.* 2011). On the other hand, divergent observations were reported on high GC content regions, with preferential amplification in Solexa (Hillier *et al.* 2008) and Illumina systems (Dohm *et al.* 2008), in contrast to pyrosequencing with reported inefficient amplification of GC-rich regions (Grossmann *et al.* 2011; Bravo *et al.* 2010). This limitation might become problematic, especially for studies aiming at sequencing genetic determinants with high GC content. However, solutions were reported such as improved emulsion PCR mix during pyrosequencing runs (Grossmann *et al.* 2011) or base substitution (Bravo *et al.* 2010).

ii. Metagenomic data and bioinformatics

NGS technologies are providing massive amount of data from diverse environments, thus giving the opportunity for microbiologists to investigate microbial population at a totally new scale (Delmont *et al.* 2011c). However, dealing with such an incredible quantity of data often represents a challenge by itself, as it requires powerful technical resources. Hopefully, bioinformatics has been developing fast, giving access to software and free-servers allowing rapid handling of these datasets. Table 1 displays some bioinformatic tools currently used for metagenomic data processing. First, raw sequence data files need to be checked and cleaned up to remove artifacts generated during the sequencing runs such as artificial duplicates, low-quality sequences or residual vector sequences. Generally, raw fasta files are treated with sequence quality score systems in order to generate high quality sequence files. The *Trim* tool cleans low-quality sequences from Illumina input files (Cox *et al.* 2010). The CD-HIT 454 software is devoted for pyrosequencing sequence clustering based on homology, thus permitting removal of artificial duplicate (Niu *et al.* 2010). Remaining sequences can be submitted to read assembly into contigs thanks available software such as *Newbler* (Roche Applied Science, 454 Life Science, http://www.454.com/). Working on contigs often results in better taxonomic and functional affiliations, as length is usually greater than initial reads. However, the inherent diversity in most environments often results in short contigs generation, with high risk of chimera formation.

Tool	Reference	Format	Description
Trim	Cox *et al.* 2010	Software	Checks the read quality from Illumina sequencing datasets.
CD-HIT-454	Niu *et al.* 2010	Software package	Performs sequence clustering based on homology. Allow duplicate removal.
CD-HIT-EST	Huang *et al.* 2010		Performs sequence clustering from DNA and protein input files.
H-CD-HIT	Huang *et al.* 2010		Devoted to protein clustering, with greater computational efficiency.
Mothur	Schloss *et al.* 2009	Software package	Workflow including dataset cleaning tools and OTUs identification tools. Allows statistical comparison of community structures from different samples.
QIIME	Caporaso *et al.* 2010	Software package	Workflow including several tools and allowing dataset comparison.
RAMMCAP	Li *et al.* 2009	Software package	Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline. Allows statistical comparison of metagenomes.
Megan	Huson *et al.* 2011	Software package	Assigns NCBI taxonomic groups to query sequences based on BLAST against NR. It does not require huge computational resource.
Treephyler	Schreiber *et al.* 2010	Software	Performs phylogenic affiliations from predicted genes.
CAMERA	Seshadri *et al.* 2007	Online pipeline	Workflow including tools for metagenomic data analysis. Registration required. Datasets need to be uploaded and require extended processing time.
MG-RAST	Meyer *et al.* 2008	Online pipeline	Workflow including useful tools for metagenomic data analysis. Registration required. Datasets need to be uploaded and require extended processing time.
WebMGA	Wu *et al.* 2011	Online pipeline	Workflow including several tools for metagenomic data analysis. No registration required. Datasets need to be uploaded. Shorter processing time.
STAMP	Parks and Beiko, 2010	Software	Statistical tool developed to handle exported tables from MG-RAST, RITA, CoMET. Allows statistical comparison of metagenomes or groups of metagenomes.
Galaxy	Goecks *et al.* 2010	Online workflow	Online open source including lots of useful tools, with very broad range of applications.
Newbler	Roche Applied Science	Software	Assembler algorithm designed by Roche for 454 datasets alignment and assembly.
Velvet	Zerbino and Birney, 2008	Software	Assembler algorithm designed for NGS alignment and assembly (Solexa or 454).
Soap	Li R *et al.* 2009	Software	Assembler algorithm designed for alignment and assembly of Solexa/Illumina datasets

Tab.1: Frequent bioinformatic tools used in metagenomics. The following links are providing non-exhaustive list of reported bioinformatic tools (http://seqanswers.com/wiki/Software/list; https://wiki.nbic.nl/index.php/NGS_Tools)

These limitations are directly impacting on identification of full or partial open reading frame (ORF). Some studies rather prefer focusing directly on reads than contigs, especially if the goal is to consider relative abundances of function and/or taxa (Krause *et al.* 2008). Nevertheless, taxonomic affiliation based on direct read annotation is questionable, as it may result in wrong interpretation mostly because of horizontal gene transfer, and public database redundancy.

After preliminary data treatments, sequences can be analyses through different methods, depending on the question addressed. Datasets can be submitted to in depth diversity analysis by BLAST annotation. This approach allows numbering of hit counts whether at functional or taxonomical level, with an e-value threshold. ORF calling can be performed to retrieve coding sequences. Function can be predicted from hypothetical proteins using specific algorithms for pattern-finding such as Hidden Markov Model (HMM) (Eddy, 2009). HMM is an efficient strategy for novel sequence findings, as it relies on a total different algorithm than traditional techniques such as multiple sequence alignment and pairwise alignment. Several tools based on HMM are already available like Phymm (Brady *et al.* 2009), MetaGeneMark (Zhu *et al.* 2010) and Glimmer (Kelley *et al.* 2011). Furthermore, numerous pipelines were created for metagenomic sequences analysis. For instance, CAMERA pipeline (Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis) was developed for easy storage, handling and analysis of these datasets (Seshadri *et al.* 2007). In addition, the metagenomics RAST server (MG-RAST) proposes a free, rapid and broad annotation system that will allow metagenomic sequence assignments against protein or nucleic databases, coupled to phylogenic summaries (Meyer *et al.* 2008). Other bioinformatics tools such as MEGAN (Metagenome Analyzer) were elaborated for metagenome datasets and/or contigs analysis through BLAST algorithms (Huson *et al.* 2007). After assembly and/or annotation steps, tools such as SEED can be applied to organize data into subsystems for gene function predictions (Overbeek *et al.* 2005). Phylogenic affiliations from predicted genes can be performed thanks *Treephyler* for a fast taxonomic profiling (Schreiber *et al.* 2010). Currently, many metagenomes are publically available for scientists, thus giving the opportunity to cross compare environments based on their metagenomic datasets (Delmont *et al.* 2011c). On the other hand, the diversity in available bioinformatic tools gives the opportunity to cross-compare pipelines and procedures in order to determine the impact on scientific interpretations and conclusions (Barriuso *et al.* 2011).

However, even if great progresses have been made in both sequencing and bioinformatics domains, biological conclusion and hypothesis should be balanced because of inherent biases coming along the whole process. For instance, (1) uneven species representation in metagenomes, (2) sequence annotation based on limited known database and (3) assembling chimera might unfortunately result in inaccurate analysis (Monier *et al.* 2011). Dealing with species-rich environment such as soil is making the problematic even tougher compared to low diversity habitat. For instance, a metagenome from an acid mine drainage biofilm was successfully investigated using high throughput sequencing, mostly due to its low diversity, resulting in almost full-reconstitution of the two dominant bacterial genomes (Tyson *et al.* 2004). Another study has reported better metagenome assembly rates from an *in vitro* insect's gut, harboring lower bacterial diversity than the wild type (Belda *et al.* 2011). As consequence, diversity reduction could be an efficient alternative strategy to increase genome assembly from uncultured bacteria.

For example, a microcosm approach enriched with specific compounds and/or placed under specific conditions can be applied. The selective pressure applied should impact enough onto the overall microbial community, in order to reduce the so-observed diversity, but should also result in specific selection toward adapted species. Such an approach may succeed by increasing the probability to sequence adapted dominant species.

Even though most of the genes extracted from the new viral metagenome datasets cannot be yet correlated to known genes according to BLAST homology, next generation sequencing technologies are becoming prevalent to explore bacteriophage population structure and its genotypic diversity. Whereas only T7 and M13 genomes were available in 1982, approximately two hundred bacteriophage genomes were deposited in NCBI database in 2005. From now on, this number has been increased fourfold (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup; http://www.phantome.org/PhageSeed/Phage.cgi). Since 2005, an online available bioinformatic software called PHACCS (http://biome.sdsu.edu/phaccs/index.htm), enables structure prediction and diversity analysis of environmental phage communities based on contigs spectrum (Angly *et al.* 2005). This tool confirmed the richness of viral communities which are closely related to their host (Fierer *et al.* 2007). An automated method called Phage Finder (http://www.phage-finder.sourceforge/) allows identification of prophage elements in bacterial genome sequences (Fouts, 2006). This tool was successfully used and validated against a control set of manually-defined prophage regions (Canchaya *et al.* 2003*;* Casjens *et al.* 2003). Recently, a couple of publicly available web sites which are specifically dedicated to phages have been developed to supply updated phage and prophage genome sequence databases, as well as links to bioinformatic softwares. These two projects, the "Viral Informatics Ressource for Metagenome Exploration" (http://virome.dbi.udel.edu) and the "Phage Annotation Tools and Methods" (http://www.phantome.org) allow analysis and classification of predicted viral ORF from collected metagenome sequence data, thus resulting in meaningful biological and functional information.

However, a major part of the data generated through NGS still remains non-annotated, especially from complex environment such as soil (Delmont *et al.* 2011b). This aspect is coherent considering the complexity of this habitat, thus bringing new evidences toward the huge diversity dwelling there (Vogel *et al.* 2009). These aspects are crucial and should be carefully taken into consideration before going any further in metagenomes interpretation.

4. The "single-cell" approach

i. Introducing the single-cell concept

A recent alternative to investigate microbial communities is the so-called single-cell approach (Hutchison and Venter, 2006*;* Raghunathan *et al.* 2005), that focuses on isolation and analysis of individual cells rather than broad-range metagenomic approach. The main strategy consists in bypassing the microbial community complexity by preliminary isolation of single candidate cells. Investigations can be performed at the genetic level using whole genome amplification (WGA) and genomic libraries (Rodrigue *et al.* 2009; Kvist *et al.* 2007*;* Marcy *et al.* 2007a*;* Stepanauskas *et al.* 2007*;* Zhang K. *et al.* 2006).

Another way to undertake the single-cell approach is to look at the metabolomic level (Rubakhin *et al.* 2011) thanks recent progress in the mass spectrometry field (Heinemann *et al.* 2011). This review will only focus on the genomic-level of the single cell approach. Different habitats have already been investigated through single-cell approaches such as human mouth biofilm (Marcy *et al.* 2007b), soil (Podar *et al.* 2007), sea water (Rodrigue *et al.* 2009), coastal water (Woyke *et al.* 2010), marine sponges (Siegl *et al.* 2011) and sediment water (Blainey *et al.* 2011a). First requirement is to handle large quantities of cells in order to statistically increase the possibilities of getting rare and/or novel species. Next step is to make sure that only one cell will be processed at the time, relying on powerful calibration system ensuring single-cell sorting. Several strategies were already reported such as serial dilution (Zhang K. *et al.* 2006), microfluidics (Marcy *et al.* 2007a), micromanipulation (Kvist *et al.* 2007; Ishoy *et al.* 2006) and fluorescence-activated cell sorting (FACS) (Stepanauskas *et al.* 2007; Huber *et al.* 1996).

The most efficient strategy was reported by Rodrigue and collaborators in 2009 (Rodrigue *et al.* 2009), using high-throughput single-cell WGA coupled to the FACS technology. This approach allowed fast sorting of isolated individual cell in 96- or 384- well-plates. After cell sorting and isolation, the next step consists in WGA, as bacterial cells possess only few femtograms of DNA (Binga *et al.* 2008*;* Raghunathan *et al.* 2005). Among all the WGA strategies, MDA seems to become a reference since it allows substantial and high fidelity amplification of large fragments, from very low amount of matrix DNA (Dean *et al.* 2002; Nelson *et al.* 2002). The size of generated fragments is 12kb on average and rises up to 100kb, which is suitable for downstream library constructions (Ishoey *et al.* 2008*;* Rodrigue *et al.* 2009). This amplification step relies on random hexamer primers, and the DNA polymerase from bacteriophage *phi29*, which is known to be highly processive and proofreading. The *phi29* polymerase got an efficient displacement activity, allowing simultaneous generation of multiple copies while displacing the previous extended strand (Rajendhran and Gunasekaran, 2008). MDA was already reported as an efficient and successful approach to amplify the genome of a single *E. coli* cell (Jiang *et al.* 2005), and also a single human sperm cell (Raghunathan *et al.* 2005).

ii. MDA biases and alternatives

MDA has raised lots of discussions toward actual minor and major biases occurring during the amplification step such as (1) undesired amplification process inherent to MDA and (2) background amplification due to contamination.

(1) Undesired amplification process such as primer-dimer formation, stochastic binding reaction or complex hyperbranched structures may result in the generation of chimera sequences and a decrease in cloning efficiency. However, these reactions appear to be minimal during the MDA. Moreover, hyperbranched structures can be removed by nuclease treatment prior to cloning (Zhang K. *et al.* 2006). Nevertheless, discussions were also addressed toward major MDA biases such as preferential amplification of high GC content regions (Pinard *et al.* 2006) and random over-amplification of specific locus (Marcy *et al.* 2007a*;* Raghunathan *et al.* 2005*;* Zhang K. *et al.* 2006). These biases will become prevalent as the single-cell approach relies on minute quantities of matrix DNA, resulting in greater sequencing effort to achieve complete genome assembly.
As a consequence, several orders of magnitude can be observed on genetic region abundance, resulting in uneven distributed libraries (Dean *et al.* 2001; Marcy *et al.* 2007a*;* Zhang K. *et al.* 2006). For example, Rodrigue and collaborators (Rodrigue *et al.* 2009) reported that up to 31% of the sequenced bases matched only with 2% of the targeted genome. An alternative to these issues relies on reducing the MDA reaction volume. For instance, microfluidic strategies in nanoliter reactors showed more homogeneous loci representation after MDA (Marcy *et al.* 2007). PCR enhancer such as trehalose was also used in MDA mixes in order to generate more even amplicons distribution (Pan *et al.* 2008). Trehalose is also known to be an effective PCR enhancer, as it helps in stabilizing and activating thermolabile enzymes such as DNA polymerases (Spiess *et al.* 2004; Carninci *et al.* 1998). An alternative approach aiming to normalize libraries was carried out (Rodrigue *et al.* 2009), resulting in more even sequence distribution. This strategy consists in removal of highly represented sequences during re-annealing kinetic by a duplex-specific nuclease (Zhulidov *et al.* 2004).

(2) As MDA is based on random hexamer primers, it is extremely sensitive to extra-cellular DNA contamination. These undesired amplifications are known to introduce incorrect assignments while mapping the genome (Dean *et al.* 2001; Hosono *et al.* 2003; Zhang K. *et al.* 2006). Background contaminations after MDA may reach up to 70% of the final amplified DNA (Raghunathan *et al.* 2005). In fact, extra-cellular DNA from the environment can rather be in solution and/or attached to the cell membranes. Either way, it might be co-extracted during the cell-sorting step. Again here, volume reduction appears as one of the best strategy for limiting extra-cellular DNA contaminations (Hutchison and Venter, 2006; Zhang K. *et al.* 2006). For instance, serial dilutions occurring during the FACS runs result in very low contamination, down to 10^{-2}~10^{-4} fg/reaction (Rodrigue *et al.* 2009).

iii. A complementary approach for metagenomics

The metagenomic quest of new findings can only be achieved when clones of interest are identified from a metagenomic library, and fully assembled through sequencing. Unfortunately, complete genome assembly from complex metagenomic datasets is currently still out of reach since most habitats are too complex regarding the microorganism diversity, and as the financial and technical means required would be unrealistic. Single-cell strategies allow direct investigation of the great uncultured microbe majority, and bring new insights of microbial population heterogeneity at the sub-specific level (Woyke *et al.* 2009; Kvist *et al.* 2007). These approaches might become a good alternative to metagenomic studies, because it helps in linking complex observed diversity to genomic evidences of co-existing cells. Moreover, the fast development of sequencing technologies is making possible the perspective of complete sequencing of single-cell genomes (MacLean *et al.* 2009). However, such a strategy would require consequent high-throughput pipelines in order to isolate cells and generate high-quality sequencing libraries. As a matter of fact, an accurate and in-depth understanding of the real potentialities of the single-cell approach is needed (Rodrigue *et al.* 2009). Even if WGA limitations are well known and partly solved, single-cell genome assembly is still something hard to achieve.

Since the emergence of single cell genomics, only few attempts of genome assembly from isolated uncultured bacterium were reported (De Jager and Siezen, 2011), thus reflecting how difficult cell isolation and analysis can be, especially from complex environments. For instance, genome completion rate do not exceed 50-75% with the MDA approach (Marcy *et al.* 2007a*;* Raghunathan *et al.* 2005*;* Zhang K. *et al.* 2006). This might be attributed to MDA biases, but also to genome integrity and damages, resulting in loss of genomic regions. Despite these inherent limitations, different MDA based strategies have already reported complete genome assembly. An uncultured bacterium genome from protists inhabiting the termite gut was fully assembled by pooling single-cells prior to MDA (Hongoh *et al.* 2008). Another study used partial genome sequences retrieved from MDA and metagenomic data to realize the final genome mapping (Lasken and Stockwell, 2007). Rodrigue and collaborators (Rodrigue *et al.* 2009) yielded up to 99.6% genome coverage thanks FACS high-throughput pipeline coupled to pyrosequencing/Illumina technologies. Another study aiming at discovering new plant biomass degrading enzymes succeeded in assembling 15 uncultured bacterial genomes from a cow rumen (Hess *et al.* 2011). In this study, FACS has been successfully used to isolate single bacterial cells from cow rumen extracts, and a complementary strategy using WGA and massive Illumina sequencing was applied successfully to reconstruct bacterial genomes.

Furthermore, single-virus strategy has been tested in order to provide suitable new viral genomic data, which seems to constitute a promising strategy. For instance, based on the *Phi29* polymerase rolling-circle replication feature (Dean *et al.* 2002), ssDNA virus diversity could be selectively explored by MDA without any denaturation step prior to amplification (Kim *et al.* 2011). Indeed, viral genome sequencing and assembly from individualized virus particle *via* MDA was successfully used (Allen *et al.* 2011). However several biases are remaining, mostly due to DNA contamination traces from phage *phi29* on the commercially available *Phi29* polymerases (Blainey *et al.* 2011b).

As soon as the MDA artifacts will be mastered, the so-called single virus genomics may force the issue in the field of gene identification within viral metagenomes like it has been already shown for non-cultivable bacteria from complex environments. To circumvent artifacts that occur during MDA, such as preferential amplification of circular ssDNA and chimeric sequence synthesis, or to bypass the time consuming LASLibrary construction, an innovative transposon-based method has been recently described (Marine *et al.* 2011). This new tool developed from the Nextera technology (Epicentre Biotechnologies, Madison,WI) allows phage DNA fragmentation and amplification coupled to a 454-adaptor ligation. A 454-library was so-obtained from little amount of phage genomic DNA (50ng) within a short time procedure, and provided high-quality sequence data.

Full genome assembly from individual isolated cells is possible, opening a new area in environmental microbial population study (Rodrigue *et al.* 2009). Single-cell approaches are bringing new insights in microbiology, such as (1) characterization of genetic heterogeneity in a cell population; (2) greater genetic information retrieving; (3) establishment of relationship between sequences, genes and functions previously unreachable due to metagenomic libraries size limitations (40-200kb); (4) information at the cell-to-cell interaction level or at the host/parasite level; (4) a direct access to rare and/or new organisms as well as their genetic and metabolic potentialities. Finally, single-cell is becoming a really powerful complementary approach for metagenomic studies, because it helps in filling the assembly gaps, and because it links diversity to genomic evidences.

Conclusion

Metagenomic approaches opened a new area in soil microbial exploration by giving access to previously unreachable organisms, and bringing a global view of prokaryote diversity dwelling in diverse environments. Furthermore, soil microorganisms' versatility is recognized as the most relevant reservoir of new natural products, including enzymes and antibiotics, harboring high potential value for downstream industrial and pharmaceutical applications. However, metagenomic has also highlighted the considerable and so-expected diversity inhabiting several environments, as well as the current incapacity to achieve complete assembly of bacterial genomes out of complex metagenomic datasets. Even if sequencing technologies are going to evolve in terms of read length and data volume, new limitations coming from informatic pipelines are certainly going to slow down the handling of such huge datasets. Focusing on alternative strategies such as single-cell, coupled with recent advances in sequencing technologies is becoming a really powerful strategy to perform full sequencing and assembly of individual microbial genomes from complex environment. Relying on viral phyla investigation will certainly help to better apprehend metagenomic data, considering that viral DNA signatures can be found with bacterial genetic elements.

Finally, unraveling of the "metagenomic black box" still remains difficult, but the current context seems to be favorable, both in terms of technology improvements and alternative developments, thus allowing optimistic forecasts about the future of metagenomics.

Acknowledgements

This review was written and funded in the frame of the European Union Metaexplore project METAEXPLORE (KBBE-222625), the French National Research Agency (Agence National de Recherche) ANR Genomique programme: METASOIL project and the ADEME (project "Generique", convention de financement no. 0975C0007). We acknowledge Joseph NESME for reviewing the manuscript and valuable discussions.

References

Allen LZ, Ishoey T, Novotny MA, McLean JS, Lasken RS, Williamson SJ (2011) Single virus genomics: a new tool for virus discovery. *PLoS One*. 6:e17722.

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *Bmc Bioinformatics*. 6:41

Ashelford KE, Day MJ, Fry JC (2003) Elevated Abundance of Bacteriophage Infecting Bacteria in Soil. *Appl Environ Microbiol* 69:285-289

Audemard C, Reece KS, Burreson EM (2004) Real-time PCR for detection and quantification of the protistan parasite *Perkinsus marinus* in environmental waters. *Appl Environ Microbiol.* 70:6611-6618

Banks DJ, Beres SB, Musser JM (2002) The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends Microbiol*. 10:515-521

Barriuso J, Valverde JR, Mellado RP (2011) Estimation of bacterial diversity using Next Generation Sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics.* 12:473

Bartosik AA, Markowska A, Szarlak J, Kulińska A, Jagura-Burdzy G (2011) Novel broad-host-range vehicles for cloning and shuffling of gene cassettes. *J Microbiol Methods*. 88:53-62

Bartram AK, Lynch MD, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. *Appl Environ Microbiol.* 77:3846-3852

Beja O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP, Villacorta R, Amjadi M (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol.* 2:516-529

Belda E, Pedrola L, Pereto J, Martinez-Blanch JF, Montagud A, Navarro E, Urchueguia J, Ramon D, Moya A, Porcar M (2011) Microbial Diversity in the Midguts of Field and Lab-Reared Populations of the European Corn Borer Ostrinia nubilalis. *PLoS One.* 6:e21751

Bertrand H, Poly F, Van VT, Lombard N, Nalin R, Vogel TM, Simonet P (2005) High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J Microbiol Methods.* 62:1-11

Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J*. 2:233-241

Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR (2011a) Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS One.* 6:e16626

Blainey PC, Quake SR (2011b) Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.* 39:e19

Brady A and Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*. 6, 673–676

Bravo LT, Tuohy MJ, Shrestha NK. (2010) Successful pyrosequencing of GC-rich DNA sequences by partial substitution of deoxyguanosine with deoxyinosine. *Diagn Mol Pathol*. 19:123-125

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003) Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol.* 185:6220-6223

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A.* 99:14250-14255

Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, Coutinho PM, Henrissat B, Nelson KE, White BA (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A.* 106:1948-1953

Burgmann H, Pesaro M, Widmer F, Zeyer J (2001) A strategy for optimizing quality and quantity of DNA extracted from soil. *J Microbiol Methods*. 45:7-20

Canchaya C, Fournous G, Brussow H (2004) The impact of prophages on bacterial chromosomes. *Mol Microbiol*. 53:9-18

Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H (2003) Prophage genomics. *Microbiol Mol Biol Rev*. 67:238-276

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 7:335-336

Carninci P, Nishiyama Y, Westover A, Itoh M, Nagaoka S, Sasaki N, Okazaki Y, Muramatsu M, Hayashizaki Y (1998) Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc Natl Acad Sci U S A*. 95:520-524

Casas V, Rohwer F (2007) Phage metagenomics. *Methods Enzymol.* 421:259-268

Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol.* 49:277-300

Chattopadhyay S, Puls RW (2000) Forces dictating colloidal interactions between viruses and soil. *Chemosphere*. 41:1279-1286

Chen F, Lu JR, Binder BJ, Liu YC, Hodson RE (2001) Application of digital image analysis and flow cytometry to enumerate marine viruses stained with SYBR gold. *Appl Environ Microbiol*. 67:539-545

Courtois S, Cappellano CM, Ball M, Francou FX, Normand P, Helynck G, Martinez A, Kolvek SJ, Hopke J, Osburne MS, August PR, Nalin R, Guerineau M, Jeannin P, Simonet P, Pernodet JL (2003) Recombinant Environmental Libraries Provide Access to Microbial Diversity for Drug Discovery from Natural Products. *Appl Environ Microbiol*. 69:49-55

Courtois S, Frostegård A, Goransson P, Depret G, Jeannin P, Simonet P (2001) Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environ Microbiol*. 3:431-439

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics.* 11:485

Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312:1795-179å8
Curtis TP, Sloan WT (2005) Microbiology. Exploring microbial diversity-a vast below. *Science*. 309:1331-1333

Daniel R (2005) The metagenomics of soil. *Nat Rev Microbiol.* 3:470-478

De Jager V, Siezen RJ (2011) Single-cell genomics: unravelling the genomes of unculturable microorganisms. *Microb Biotechnol.* 4:431-437

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A.* 99:5261-5266

Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res*. 11:1095-1099

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011a) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol.* 77:1315-1324

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011b) Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods*. 86:397-400

Delmont TO, Malandain C, Prestat E, Larose C, Monier JM, Simonet P, Vogel TM (2011c) Metagenomic mining for microbiologists. *ISME J*. 5:1-7

Demaneche S, Jocteur-Monrozier L, Quiquampoix H, Simonet P (2001) Evaluation of Biological and Physical Protection against Nuclease Degradation of Clay-Bound Plasmid DNA. *Appl Environ Microbiol*. 67:293-299

Demanèche S, David MM, Navarro E, Simonet P, Vogel TM. (2009). Evaluation of functional gene enrichment in a soil metagenomic clone library. *J Microbiol Methods*. 76:105-107

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. *Nature*. 452:629-632

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 36:e105

Doolittle WF (1999) Phylogenic classification and the universal tree. *Science*. 284:2124-2129

Dumont MG, Murrell JC (2005) Community-level analysis: key genes of aerobic methane oxidation. *Methods Enzymol. 397*:413-427

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 23:205-211

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander Ec JR, Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *Bmc Genomics*. 7:57

Edwards RA, Rohwer F (2005) Viral metagenomics. *Nat Rev Microbiol*. 3:504-510

Engel K, Pinnell L, Cheng J, Charles TC, Neufeld JD (2012) Nonlinear electrophoresis for purification of soil DNA for metagenomics. *J Microbiol Methods*. 88:35-40

Feinstein LM, Sul WJ, Blackwood CB (2009) Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Appl Environ Microbiol*. 75:5428-5433

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S, Knight R, Rohwer F, Jackson RB (2007a) Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. *Appl Environ Microbiol*. 73:7059-7066

Fierer N, Bradford MA, Jackson RB (2007b) Toward an ecological classification of soil bacteria. Ecology 88:1354-1364

Fierer N, Jackson JA, Vilgalys R, Jackson RB (2005) Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. *Appl Environ Microbiol*. 71:4117-4120

Fordyce SL, Ávila-Arcos MC, Rockenbauer E, Børsting C, Frank-Hansen R, Petersen FT, Willerslev E, Hansen AJ, Morling N, Gilbert MT (2011) High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *Biotechniques*. 51:127-133

Foster RC (1988) Microenvironments of soil microorganisms. *Biol Fertil soils*. 6:189-203

Fouts DE (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*. 34:5839-5851

Frostegård A, Courtois S, Ramisse V, Clerc S, Bernillon D, Le Gall F, Jeannin P, Nesme X, Simonet P (1999) Quantification of bias related to the extraction of DNA directly from soils. *Appl Environ Microbiol*. 65:5409-5420

Gabor EM, de Vries EJ, Janssen DB (2003) Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiol Ecol*. 44:153-163

Gabor EM, de Vries EJ, Janssen DB (2004) Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environ Microbiol*. 6:948-958

Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*. 309:1387-1390

Garbeva P, de Boer W (2009) Inter-specific interactions between carbon-limited soil bacteria affect behavior and gene expression. *Microb Ecol.* 58:36-46

Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tuphile K, Bertrand H, Vogel TM, Perriere G, Simonet P, Nalin R (2004) Phylogenic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl Environ Microbiol*. 70:5522-5527

Glogauer A, Martini VP, Faoro H, Couto GH, Müller-Santos M, Monteiro RA, Mitchell DA, de Souza EM, Pedrosa FO, Krieger N (2011) Identification and characterization of a new true lipase isolated through metagenomic approach. *Microb Cell Fact*. 10:54

Goecks J, Nekrutenko A, Taylor J and The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 25;11:R86

Gomez P, Buckling A (2011) Bacteria-phage antagonistic coevolution in soil. *Science*. 332:106-109

Gonzalez I, Ayuso-Sacido A, Anderson A, Genilloud O (2005) Actinomycetes isolated from lichens: Evaluation of their diversity and detection of biosynthetic gene sequences. *FEMS Microbiol Ecol.* 54:401- 415

Griffiths RI, Whiteley AS, O'Donnell AG, Bailey MJ (2000) Rapid Method for Coextraction of DNA and RNA from Natural Environments for Analysis of Ribosomal DNA- and rRNA-Based Microbial Community Composition. *Appl Environ Microbiol*. 66:5488-5491

Grossmann V, Schnittger S, Schindela S, Klein HU, Eder C, Dugas M, Kern W, Haferlach T, Haferlach C, Kohlmann A (2011) Strategy for robust detection of insertions, deletions, and point mutations in CEBPA, a GC-rich content gene, using 454 next-generation deep-sequencing technology. *J Mol Diagn*. 13:129-36

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 68:669-685

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 5:R245-R249

Harry M, Gambier B, Bourezgui Y, Garnier-Sillam E (1999) Evaluation of purification procedures for DNA extracted from organic rich samples: Interference with humic substances. *ANALUSIS*. 27:439-442

Hassink J, Bouwman LA, Zwart KB, Bloem J, Brussard L (1993) Relationships between soil texture, physical protection of organic matter, soil biota, and c and n mineralization in grassland soils. *Geoderma*. 57:105-128

Healy FG, Ray RM, Aldrich HC, Wilkie AC, Ingram LO, Shanmugam KT (1995) Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Appl Microbiol Biotechnol*. 43:667-674

Heinemann M, Zenobi R (2011) Single cell metabolomics. *Curr Opin Biotechnol*. 22:26-31

Henn MR, Sullivan MB, Stange-Thomann N, Osburne MS, Berlin AM, Kelly L, Yandava C, Kodira C, Zeng Q, Weiand M, Sparrow T, Saif S, Giannoukos G, Young SK, Nusbaum C, Birren BW, Chisholm SW (2010) Analysis of High-Throughput Sequencing and Annotation Strategies for Phage Genomes. *PLoS One.* 5:e9083

Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*. 331:463-467

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods.* 5:183-188

Holben WE, Jansson JK, Chelm BK, Tiedje JM (1988) DNA probe method for the detection of specific microorganisms in the soil bacterial community. Appl Environ Microbiol 54:703-711

Hongoh Y, Sharma VK, Prakash T, Noda S, Toh H, Taylor TD, Kudo T, Sakaki Y, Toyoda A, Hattori M, Ohkuma M (2008) Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut. *Science*. 322:1108-1109

Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M, Lasken RS (2003) Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* 13:954-964

Hu Y, Fu C, Huang Y, Yin Y, Cheng G, Lei F, Lu N, Li J, Ashforth EJ, Zhang L, Zhu B (2010) Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment. *FEMS Microbiol Ecol*. 72:228-237

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics.* 26:680-682

Huber K, Wolf H, Van Lindern M, Worofka B, Rosen A, Hafner E, Beug H, Philipp K, Bauer K (1996) Development of a rapid means of estimating the haemoglobin F content of candidate fetal cells isolated from maternal blood using HPLC. *Prenat Diagn.* 16:1011-1019

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenic view of bacterial diversity. *J Bacteriol.* 180:4765-4774

Hugenholtz P, Tyson GW (2008) Microbiology: metagenomics. *Nature*. 455:481-483

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8:R143

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res*. 17:377-386

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res*. 21:1552-1560

Hutchison Ca 3RD, Venter JC (2006) Single-cell genomics. *Nat Biotechnol.* 24:657-658
Inceoglu O, Salles JF, Van Overbeek L, Van Elsas JD (2010) Effects of plant genotype and growth stage on the betaproteobacterial communities associated with different potato cultivars in two fields. *Appl Environ Microbiol*. 76:3675-3684

Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS (2008) Genomic sequencing of single microbial cells from environmental samples. *Curr Opin Microbiol.* 11:198-204

Ishoey T, Kvist T, Westermann P, Ahring BK (2006) An improved method for single cell isolation of prokaryotes from meso-, thermo- and hyperthermophilic environments using micromanipulation. *Appl Microbiol Biotechnol.* 69:510-514

Jeon JH, Kim JT, Kang SG, Lee JH, Kim SJ (2009) Characterization and its potential application of two esterases derived from the arctic sediment metagenome. *Mar Biotechnol.* 11:307-316

Jiang Z, Zhang X, Deka R, Jin L (2005) Genome amplification of single sperm using multiple displacement amplification. *Nucleic Acids Res*. 33:e91

Kakirde KS, Parsley LC, Liles MR (2010) Size does matter: Application-driven approaches for soil metagenomics. *Soil Biol Biochem*. 42:1911-1923

Kauffmann IM, Schmitt J, Schmid RD (2004) DNA isolation from soil samples for cloning in different hosts. *Appl Microbiol Biotechnol*. 64:665-670

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2011) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 40:e9

Kieft TL, Soroker E, Firestone MK (1987) Microbial biomass response to a rapid increase in water potential when dry soil is wetted. *Soil Biol Biochem*. 19:119-126

Kim KH, Bae JW (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol*. 77:7663-7668

Kim JS, Lim HK, Lee MH, Park JH, Hwang EC, Moon BJ, Lee SW (2009) Production of porphyrin intermediates in Escherichia coli carrying soil metagenomic genes. *FEMS Microbiol Lett.* 295:42-49

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* 36:2230-2229

Kumar PS, Brooker MR, Dowd SE, Camerlengo T (2011) Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. *PLoS One*. 6:e20956

Kvist T, Ahring BK, Lasken RS, Westermann P (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl Microbiol Biotechnol*. 74:926-935

Lakay FM, Botha A, Prior BA (2007) Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils. *J Appl Microbiol*. 102:265-273

Langer M, Gabor EM, Liebeton K, Meurer G, Niehaus F, Schulze R, Eck J, Lorenz P (2006) Metagenomics: an inexhaustible access to nature's diversity. Biotechnol J 1:815-821

Lasken RS, Stockwell TB (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. Bmc Biotechnol 7:19

Lefevre F, Robe P, Jarrin C, Ginolhac A, Zago C, Auriol D, Vogel TM, Simonet P, Nalin R (2008) Drugs from hidden bugs: their discovery via untapped resources. *Res Microbiol*. 159:153-161

Leveau JHJ (2007) The magic and menace of metagenomics: prospects for the study of plant growth-promoting rhizobacteria. *Eur J Plant Pathology*. 11:279-300

Li W, Wooley JC, Godzik A (2008) Probing metagenomics by rapid cluster analysis of very large datasets. PLoS One. 3:e3375
Li WZ (2009) Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*.10:359

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 25:1966-1967
Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM (2003) A Census of rRNA Genes and Linked Genomic Sequences within a Soil Metagenomic Library. *Appl Environ Microbiol*. 69:2684-291

Loge FJ, Thompson de, Call DR (2002) PCR detection of specific pathogens in water: a risk-based analysis. *Environ Sci Technol*. 36:2754-2759

López-Barragán MJ, Quiñones M, Cui K, Lemieux J, Zhao K, Su XZ (2011) Effect of PCR extension temperature on high-throughput sequencing. *Mol Biochem Parasitol*. 176:64-67

Lorenz P, Eck J (2005) Metagenomics and industrial applications. Nat Rev Microbiol 3:510-516

Lüke C, Frenzel P (2011) Potential of pmoA amplicon pyrosequencing for methanotroph diversity studies. *Appl Environ Microbiol.* 77:6305-6309

MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol*. 7:287-296

Marcy Y, Ishoey T, Lasken RS, Stockwell TB, Walenz BP, Halpern AL, Beeson KY, Goldberg SM, Quake SR (2007a) Nanoliter reactors improve multiple displacement amplification of genomes from single cells. Plos Genet 3:1702-1708

Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR (2007b) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A*. 104:11889-11894

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 437:376-380

Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol*. 77:8071-8079

Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Res 15:1767-1776
Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server - a public resource for the automatic phylogenic and functional analysis of metagenomes. *Bmc Bioinformatics*. 9:386

Mlynarczyk G, Mlynarczyk A, Zabicka D, Jeljaszewicz J (1997) Lysogenic conversion as a factor influencing the vancomycin tolerance phenomenon in *Staphylococcus aureus*. *J Antimicrob Chemother*. 40:136-137

Monier JM, Demaneche S, Delmont TO, Mathieu A, Vogel TM, Simonet P (2011) Metagenomic exploration of antibiotic resistance in soil. *Curr Opin Microbiol*. 14:229-235

Moreira D, Lopez-Garcia P (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* 10:31-38

Mutter GL and Boynton KA (1995) PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic Acids Res*. 23:1411-1418

Nealson KH, Venter JC (2007) Metagenomics and the global ocean survey: what's in it for us, and why should we care? *ISME J*. 1:185-187

Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, Martins Dos Santos VA, Fouts DE, Gill SR, Pop M, Holmes M, Brinkac L, Beanan M, Deboy RT, Daugherty S, Kolonay J, Madupu R, Nelson W, White O, Peterson J, Khouri H, Hance I, Chris Lee P, Holtzapple E, Scanlan D, Tran K, Moazzez A, Utterback T, Rizzo M, Lee K, Kosack D, Moestl D, Wedler H, Lauber J, Stjepandic D, Hoheisel J, Straetz M, Heim S, Kiewitz C, Eisen J, Timmis KN, Dusterhoft A, Tummler B, Fraser CM (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol*. 4:799-808

Netto LE, Ferreira AM, Augusto O (1991) Iron(III) binding in DNA solutions: complex formation and catalytic activity in the oxidation of hydrazine derivatives. *Chem Biol Interact*. 79:1-14

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*. 11:187

Novais RC, Thorstenson YR (2011) The evolution of Pyrosequencing for microbiology: From genes to genomes. *J Microbiol Methods*. 86:1-7

Nyberg KA, Enwall K, Schnürer A, Sundh I, Hallin S (2011) Response of Induced Perturbation on Replicating β-Proteobacterial Ammonia-Oxidizing Populations in Soil. *Microb Ecol*. 63:701-709

Ogram A, Sayler GS, Barkay T (1987) The extraction and purification of microbial DNA from sediments. *J Microbiol Methods*. 7:57-66

Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, Kikuchi H, Shiba T, Sakaki Y, Hattori M (2001) Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A*. 98:12215-12220

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. 33:5691-5702

Pan X, Urban AE, Palejev D, Schulz V, Grubert F, Hu Y, Snyder M, Weissman SM (2008) A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proc Natl Acad Sci U S A*. 105:15499-15504

Parks DH and Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*. 26:715-721

Parsley LC, Consuegra EJ, Kakirde KS, Land AM, Harper Wf JR, Liles MR (2010) Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl Environ Microbiol*. 76:3753-3757

Patel A, Noble RT, Steele JA, Schwalbach MS, Hewson I, Fuhrman JA (2007) Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat Protoc*. 2:269-276

Paul JH, David AW (1989) Production of extracellular nucleic acids by genetically altered bacteria in aquatic-environment microcosms. *Appl Environ Microbiol*. 55:1865-1869

Pel J, Broemeling D, Mai L, Poon HL, Tropini G,Warren RL, Holt RA, Marziali A (2009) Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. *Proc Natl Acad Sci U S A*. 106, 14796–14801

Petric I, Philippot L, Abbate C, Bispo A, Chesnot T, Hallin S, Laval K, Lebeau T, Lemanceau P, Leyval C, Lindstrom K, Pandard P, Romero E, Sarr A, Schloter M, Simonet P, Smalla K, Wilke BM, Martin-Laurent F (2011) Inter-laboratory evaluation of the ISO standard 11063 "Soil quality - Method to directly extract DNA from soil samples". *J Microbiol Methods.* 84:454-460

Picard C, Ponsonnet C, Paget E, Nesme X, Simonet P (1992) Detection and enumeration of bacteria in soil by direct DNA extraction and polymerase chain reaction. *Appl Environ Microbiol*. 58:2717-2722

Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *Bmc Genomics*. 7:216

Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, Holland T, Cotton D, Hauser L, Keller M (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol*. 73:3205-3214

Queipo-Ortuno MI, de Dios Colmenero J, MacIas M, Bravo MJ, Morata P (2008) Preparation of bacterial DNA template by boiling and effect of immunoglobulin G as an inhibitor in real-time PCR for serum samples from patients with brucellosis. *Clin Vaccine Immunol*. 15:293-296

Radajewski S, Ineson P, Parekh NR, Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature*. 403:646-649

Raghunathan A, Ferguson Hr JR, Bornarth CJ, Song W, Driscoll M, Lasken RS (2005) Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol*. 71:3342-3347

Rajendhran J, Gunasekaran P (2008) Strategies for accessing soil metagenome for desired applications. *Biotechnol Adv.* 26:576-590

Ranjard L, Poly F, Lata JC, Mougel C, Thioulouse J, Nazaret S (2001) Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. *Appl Environ Microbiol*. 67:4479-4487

Richaume A, Steinberg C, Jocteur Monrozier L, Faurie G (1993) Difference between direct and indirect enumeration of soil bacteria: influence of soil structure and cell location. *Soil Biol Biochem*. 25:641-643

Richter DD, Markewitz D (1995) How Deep Is Soil? *BioScience*. 45:600-609

Riesenfeld CS, Schloss PD, Handelsman J (2004) METAGENOMICS: Genomic Analysis of Microbial Communities. *Annu Rev Genet*. 38:525-552

Robe P, Nalin R, Capellano C, Vogel TA, Simonet P (2003) Extraction of DNA from soil. Eur J Soil Biol 39:183-190
Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW (2009) Whole genome amplification and *de novo* assembly of single bacterial cells. *PLoS One.* 4:e6864

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*. 1:283-290

Rohwer F, Seguritan V, Choi DH, Segall AM, Azam F (2001) Production of shotgun libraries using random amplification. *BioTechniques*. 31:108-112

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res*. 11:3-11

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM (2000) Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Appl Environ Microbiol*. 66:2541-2547

Rubakhin SS, Romanova EV, Nemes P, Sweedler JV (2011) Profiling metabolites and peptides in single cells. *Nat Methods.* 8:S20-S29

Schartz A, Waksman SA (1944) Effect of streptomycin and other antibiotic substances upon *Mycobacterium tuberculosis* and related organisms. In (eds) Proceedings of the society for Experimental Biology and Medicine, New York, N.Y., pp 244-248

Schirmer A, Gadkari R, Reeves CD, Ibrahim F, Delong EF, Hutchinson CR (2005) Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. *Appl Environ Microbiol*. 71:4840-4849

Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol*. 14:303-310

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 75:7537-7541

Schloss PD, Gevers D, Westcott SL (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS One.* 6:e27310

Schreiber F, Gumrich P, Daniel R, Meinicke P (2010) Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics.* 26:960-961

Schriewer A, Wehlmann A, Wuertz S (2011) Improving qPCR efficiency in environmental samples by selective removal of humic acids with DAX-8. *J Microbiol Methods*. 85:16-21

Schuch R, Fischetti VA (2009) The secret life of the anthrax agent *Bacillus anthracis*: bacteriophage-mediated ecological adaptations. *PLoS One*. 4:e6532

Schwarz S, Waschkowitz T, Daniel R (2005) Enhancement of gene detection frequencies by combining DNA-based stable-isotope probing with the construction of metagenomic DNA libraries. *World Journal of Microbiology and Biotechnology*. DOI 10.1007

Sebat JL, Colwell FS, Crawford RL (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl Environ Microbiol*. 69:4927-4934

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A Community Resource for Metagenomics. *PLoS Biol*. 5:e75

Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, Kandeler E (2001) Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. *Appl Environ Microbiol*. 67:4215-4224

Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang C, Dandekar T, Hentschel U (2011) Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *ISME J*. 5:61-70

Singh BK, Campbell CD, Sorenson SJ, Zhou J (2009) Soil genomics. *Nat Rev Microbiol*. 7:756-757

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A.* 103:12115-12120

Sosio M, Bossi E, Bianchi A, Donadio S (2000) Multiple peptide synthetase gene clusters in Actinomycetes. *Mol Gen Genet*. 264:213-221

Spiess AN, Mueller N, Ivell R (2004) Trehalose is a potent PCR enhancer: lowering of DNA melting temperature and thermal stabilization of taq polymerase by the disaccharide trehalose. *Clin Chem*. 50:1256-1259

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T, Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res Microbiol*. 159:349-357

Staley JT, Konopka A (1985) Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol.* 39:321-346

Stein JL, Marsh TL, Wu KY, Shizuya H, Delong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol*. 178:591-599

Stepanauskas R, Sieracki ME (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci U S A*. 104:9052-9057

Strohl WR (1992) Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res*. 20:961-974

Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS One*. 6:e19534

Tebbe CC, Vahjen W (1993) Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl Environ Microbiol*. 59:2657-2665

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc*. 4:470-483

Torsvik V, Ovreas L (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol*. 5:240-245

Torsvik V, Daae FL, Sandaa RA, Ovreas L (1998) Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol*. 64:53-62

Torsvik V, Sorheim R Goksoyr J (1996) Total bacterial diversity in soil and sediment communities - a review. *J Ind Microbiol*. 17:170-178

Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. *Science*. 308:554-557

Tsai YL, Olson BH (1991) Rapid method for direct extraction of DNA from soil and sediments. *Appl Environ Microbiol*. 57:1070-1074

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 428:37-43

Uchiyama T, Abe T, Ikemura T, Watanabe K (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol*. 23:88-93

Van Elsas JD, Boersma FGH (2011) A review of molecular methods to study the microbiota of soil and the mycosphere. European Journal of Soil Biology 47:77e87

Van Elsas JD, Costa R, Jansson J, Sjöling S, Bailey M, Nalin R, Vogel TM, van Overbeek L (2008) The metagenomics of disease-suppressive soils - experiences from the METACONTROL project. *Trends Biotechnol*. 26:591-601

Van Elsas JD, Garbeva P, Salles J (2002) Effects of agronomical measures on the microbial diversity of soils as related to the suppression of soil-borne plant pathogens. *Biodegradation*. 13:29-40

Van Elsas JD, Mantynen V, Wolters AC (1997) Soil DNA extraction and assessment of the fate of *Mycobacterium chlorophenolicum* strain PCP-1 in different soils by 16S ribosomal RNA gene sequence based most-probable-number PCR and immunofluorescence. *Biol Fertil soils*. 24:188-195

Van Veen JA, van Overbeek LS & van Elsas JD (1997) Fate and activity of microorganisms following release into soil. *Microb Mol Biol Rev*. 61:121-135

Veluci RM, Neher DA, Weicht TR (2006) Nitrogen fixation and leaching of biological soil crust communities in mesic temperate soils. *Microb Ecol*. 51:189-196

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*. 304:66-74

Vogel TM, Simonet P, Jansson JK, Hirsh PR, Tiedje JM, Van Elsas JD, Bailey MJ, Nalin R, Philippot L (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Microbiol*. 7:252

Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, Streit WR (2003) Prospecting for Novel Biocatalysts in a Soil Metagenome. *Appl Environ Microbiol.* 69:6235-6242

Wallenius K, Rita H, Simpanen S, Mikkonen A, Niemi RM (2010) Sample storage for soil enzyme activity and bacterial community profiles. *J Microbiol Methods*. 81:48-55

Ward DM, Weller R, Bateson MM (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*. 345:63-65

Warren RA (1980) Modified bases in bacteriophage DNAs. *Annu Rev Microbiol*. 34:137-158

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*. 95:6578-6583

Williamson KE, Wommack KE, Radosevich M (2003) Sampling natural viral communities from soil for culture-independent analyses. *Appl Environ Microbiol*. 69:6628-6633

Williamson LL, Borlee BR, Schloss PD, Guan C, Allen HK, Handelsman J (2005) Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl Environ Microbiol*. 71:6335-6344

Wilson RJ, Williamson DH (1997) Extrachromosomal DNA in the Apicomplexa. *Microbiol Mol Biol Rev*. 61:1-16

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. *Appl Environ Microbiol*. 74:1453-1463

Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev*. 64:69-114

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonald BR, Moran NA, Bristow J, Cheng JF (2010) One bacterial cell, one complete genome. *PLoS One.* 5:e10314

Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng JF, Eisen JA, Sieracki ME, Stepanauskas R (2009) Assembling the marine metagenome, one cell at a time. *PLoS One*. 4:e5299

Wu S, Zhu Z, Fu L, Niu B, Li W (2011) WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics.* 2011 12:444

Xu X, Kong R, de Bruijn FJ, He SY, Murry MA, Newman T, Wolk CP (2002) DNA sequence and genetic characterization of plasmid pFQ11 from *Frankia alni* strain CpI1. *FEMS Microbiol Lett*. 207:103-107

Yen JH, Chang JS, Huang PJ, Wang YS (2009) Effects of fungicides triadimefon and propiconazole on soil bacterial communities. *J Environ Sci Health*. B44:681-689

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18:821-829

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol*. 24:680-686

Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* 4:e3

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. Appl Environ Microbiol 62:316-322
Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res*. 38:e132

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res*. 32:e37

CHAPITRE 2

-

ETUDE METAGENOMIQUE DU SOL DE ROTHAMSTED

-

APPLICATION A LA RECHERCHE DE NOUVELLES CHITINASES

-

RESUME EN LANGUE FRANÇAISE

Dans ce chapitre, la partie "métagénomique" de mes travaux sera abordée. Au cours du déroulement du projet Metaexplore, plusieurs stratégies ont été mises en place afin d'étudier et d'exploiter la ressource métagénomique du sol de Rothamsted.

Le chapitre se structure de la manière suivante:

- 1: une présentation du projet Metaexplore incluant les objectifs et la stratégie mise en place.

- 2: l'analyse *in silico* du sol de Rothamsted afin de détecter la présence des gènes impliqués dans la dégradation de la chitine ainsi que les activités chitinases. Cette partie inclut des tests enzymatiques mais aussi des criblages de séquence par BLASTx afin de caractériser la diversité génétique dans les jeux de données métagénomiques du sol de Rothamsted.

- 3: le criblage génétique d'une partie de la banque de clones de Rothamsted qui a été élaboré dans le cadre de Metasoil. Ce criblage a été réalisé afin de détecter la présence de gènes d'intérêt au sein des inserts métagénomiques.

- 4: l'approche dite "d'enrichissement", où le sol de Rothamsted a été amendé en chitine dans des microcosmes afin d'établir un suivi de dégradation et d'évaluer l'impact sur la communauté bactérienne.

1. Présentation du projet Metaexplore:

Le projet Metaexplore à été initié sous l'impulsion de deux facteurs:

- une forte demande du secteur industriel, toujours à la recherche de nouvelles solutions pour améliorer leurs procédés de fabrication.

- une amélioration technologique rapide dans le secteur de la microbiologie, principalement due à l'avènement de la métagénomique et des progrès réalisés en matière de séquençage à haut débit.

C'est à la croisée de ces deux tendances qu'est né le projet européen Metaexplore, qui réunit aussi bien des partenaires académiques que des entreprises, tous impliqués dans le domaine des biotechnologies. En tout, 18 partenaires européens sont impliqués dans ce projet, à l'exception d'un groupe de recherche en Argentine. Le but du projet: utiliser les approches métagénomiques afin de découvrir, caractériser et valoriser de nouvelles enzymes d'intérêt industriel issues des communautés microbiennes environnementales.

Initialement, une large gamme d'enzymes d'intérêt avait été établie, mais le projet s'est rapidement centré sur certaines, dont notamment les enzymes intervenant dans la dégradation des composés récalcitrants naturels (Horn *et al.* 2006) ou anthropiques comme:

- les ligninases et les cellulases: enzymes qui interviennent ensemble dans le processus de dégradation de la biomasse végétale ou biomasse ligno-cellulosique. Ces enzymes trouvent des débouchés dans les processus industriels de production de biocarburant par exemple, en améliorant le rendement d'éthanol obtenu à partir de composés difficiles à convertir comme les fibres ligneuses. Elles sont aussi utilisées en papeterie où leur application lors du procédé de fabrication permet d'apporter des propriétés spécifiques au produit fini.

- les chitinases: enzymes qui interviennent dans la minéralisation de la chitine, composant principal de la paroi des cellules fongiques, ainsi que de l'exosquelette des insectes et des crustacés. Les produits générés lors de la dégradation de la chitine sont utilisés pour leurs propriétés dans des secteurs de l'industrie cosmétique, pharmaceutique, ainsi que de l'environnement.

- les déhalogénases: qui interviennent dans la dégradation des composés halogénés comme les pesticides organo-chlorés. Ces enzymes trouvent notamment des débouchés dans le secteur de la bioremédiation, comme par exemple la dépollution de sites contaminés.

La stratégie métagénomique mise en place dans le cadre du projet Metaexplore est expliquée par la figure 1. Trés rapidement, le travail c'est centré sur la recherche de nouvelles chitinases, car les premiers résultats obtenus indiquaient une grande diversité et abondance de séquences affiliées à ces gènes dans les métagénomes du sol de Rothamsted.

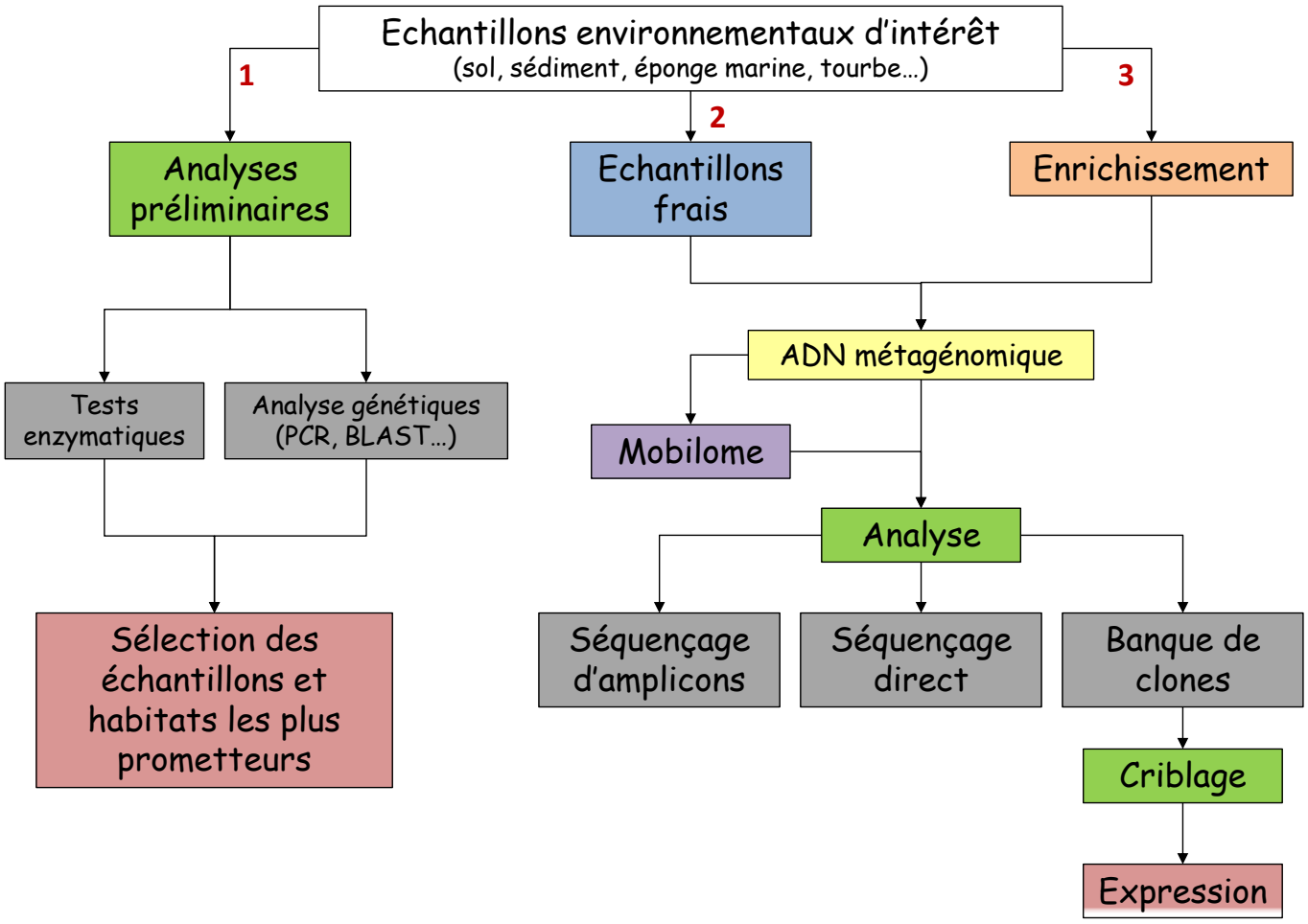


Fig.1: Représentation schématique de la stratégie mise en place dans le cadre du projet Metaexplore. Les environnements et échantillons les plus prometteurs ont tout d'abord été sélectionnés suite à une série d'analyses préliminaires permettant de justifier et d'attester de leur intérêt pour la recherche de nouvelles enzymes (1). Ces analyses préliminaires consistent souvent en des mesures de l'activité enzymatique d'intérêt, ainsi que par la recherche des déterminants génétiques porteurs de ces activités (gènes, opérons, plasmides...) par PCR, ou bien dans des jeux de données par BLAST si disponibles. Ensuite, les échantillons sont analysés par la voie de la métagénomique directement (2) ou bien après une étape d'enrichissement à l'aide des substrats relatifs aux enzymes recherchées afin de stimuler et d'augmenter leur proportion (3). L'ADN métagénomique ainsi obtenu peut être utilisé de différentes manières, incluant le séquençage de l'ADN total ou bien celui de produits PCR spécifiques (amplicons). Enfin, l'ADN métagénomique peut être utilisé pour la constitution de banques de clones métagénomiques, qui pourront être criblés par la suite afin de détecter les gènes ou les activités d'intérêt. Le projet Metaexplore porte également une attention toute particulière sur le mobilome, représentant l'ensemble des éléments génétiques mobiles susceptibles d'être transférés d'une cellule procaryote à une autre par les voies du transfert horizontal de gène. Cette fraction est en effet prometteuse pour le projet, car bon nombre d'enzymes recherchées sont souvent associées aux éléments génétiques mobiles.

La stratégie suivie par le groupe de Génomique Microbienne Environnementale est très semblable à celle du projet en lui-même, à l'exeption du « mobilome » et de l'approche de « séquençage d'amplicon » qui n'on pas été utilisé (Fig.1). Dans le cas de l'analyse préliminaire, les données métagénomiques générées dans le cadre de Metasoil ont été analysées afin d'étudier la diversité de gènes correspondant aux différentes familles d'enzymes recherchées. Dans cette étude, mon travail s'est orienté sur les chitinases pour les raisons suivantes:

- les analyses préliminaires des jeux de données ont révélé une grande quantité et diversité de séquences affiliées aux chitinases dans les métagénomes du sol de Rothamsted.

- les chitinases sont les seules enzymes qui sont suffisamment conservées pour éventuellement être exploitées par l'outil Genefish développé également dans le cadre de Metaexplore (cf. Chapitre 3). En effet, le site responsable de la réaction d'hydrolyse du substrat présente une forte conservation au niveau de la séquence protéique. Cette conservation se retrouve également dans la séquence nucléique, ce qui est un paramètre important pour le bon fonctionnement de l'outil. Pour plus de détails, se reporter au Chapitre 3 de cette étude.

Néanmoins, les autres catégories ont également été recherchées, notamment lors du criblage de la banque de clones. Enfin, l'analyse métagénomique par séquençage direct a été faite en comparant des échantillons de sol frais avec ceux enrichis en chitine dans des microcosmes.

2. Analyses préliminaires du sol de Rothamsted

Le sol de Rothamsted a été analysé de deux manières afin d'obtenir des informations: (1) l'activité chitinase dans les échantillons de sol frais et (2) la diversité de séquences affiliées aux gènes de dégradation de la chitine dans la base de données métagénomiques de Rothamsted.

Les mesures enzymatiques ont été réalisées sur des extraits aqueux de sol frais et ont permis d'évaluer l'importance des activités endochitinases et exochitinases. Ces deux activités se succèdent dans le temps au cours de la dégradation de la chitine, la première générant le substrat de la deuxième.

L'analyse des données métagénomiques a été effectuée par la technique du BLASTx, qui permet d'identifier des séquences protéiques dans une banque d'ADN traduite en protéines. Pour réaliser cette étape, une banque de données protéiques recouvrant la majeure partie des enzymes potentiellement impliquées dans la dégradation de la chitine a été constituée. La figure 2 récapitule les voies de dégradation de la chitine ainsi que les enzymes impliquées. Afin de pouvoir comparer et interpréter les résultats, le sol de Rothamsted a été confronté d'une part à un autre sol pour les mesures d'activités mais également à plusieurs métagénomes issus d'autres environnements (des océans avec 3 métagénomes d'eaux côtières et 3 d'eaux benthiques, un métagénome de sol témoin provenant de Puerto Rico, 7 métagénomes de tractus digestif provenant de prélèvements d'humains et d'animaux, cf. Chapitre 2 - Part.2).

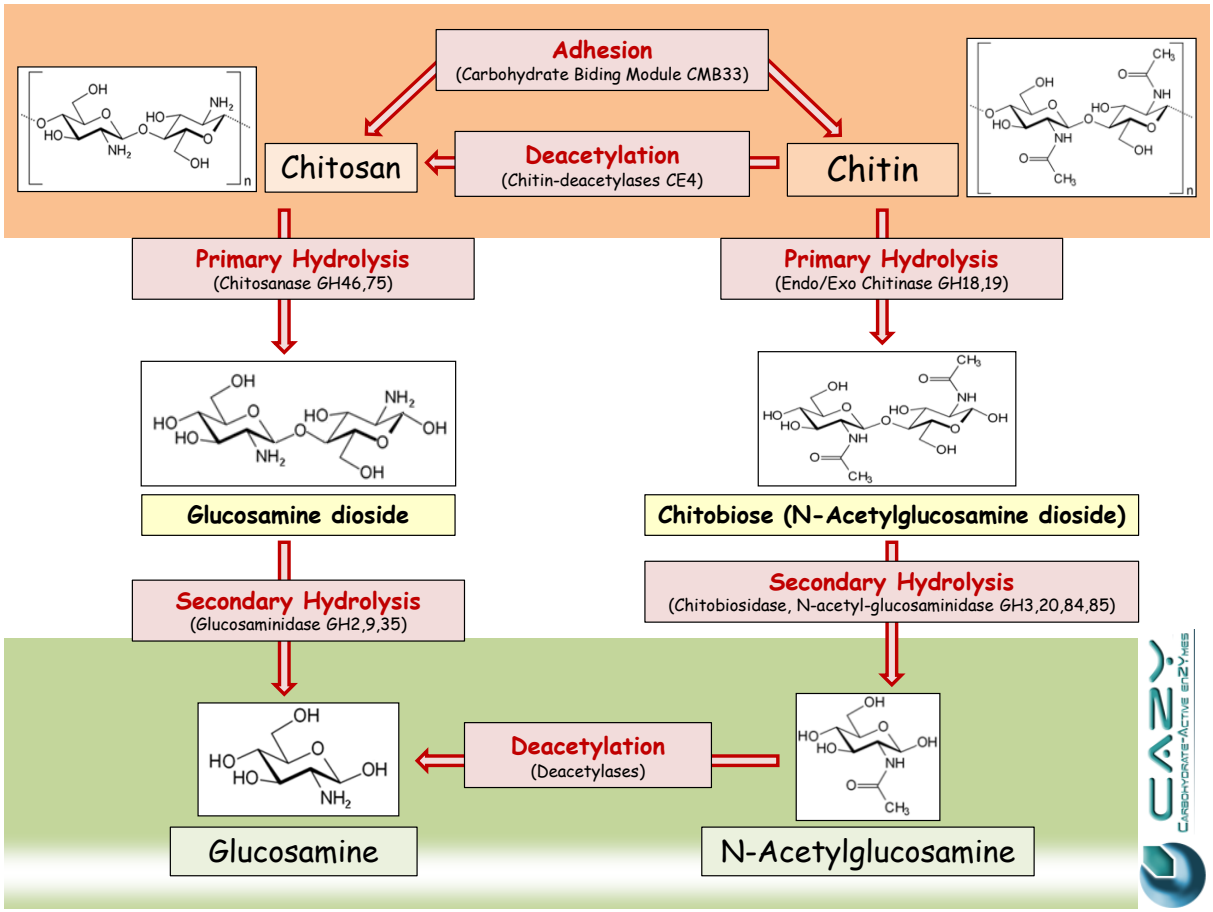


Fig.2 : Représentation schématique de la biodégradation de la chitine. La quasi-totalité des enzymes impliquées dans la dégradation de la chitine sont référencées dans la banque de données CAZY (Carbohydrate Active Enzymes, Cantarel et al. 2009). La chitine est un bio-polymère constitué d'unité N-acétylglucosamine, principal constituant des exosquelettes des insectes et des crustacées, ainsi que de la paroi des cellules fongiques et végétales. Sa dégradation est principalement effectuée via des processus enzymatiques microbiens dans la nature, car cette macromolécule représente une source non-négligeable de carbone et d'azote pour les organismes capables de l'exploiter. La chitine est même le deuxième bio-polymère en termes de production annuelle de biomasse derrière la cellulose. Parmi les microorganismes capables de dégrader la chitine, les bactéries jouent un rôle prépondérant de par leur nombre et leur diversité, surtout dans les environnements complexes tels que le sol. La chitine peut être dégradée selon deux voies : (1) l'hydrolyse et la déacétylation ou bien (2) la déacétylation puis l'hydrolyse, les produits finaux étant le N-acétylglucosamine et le glucosamine. La première voie (à droite sur le schéma) fait intervenir les chitinases afin de dépolymériser les macromolécules (GH18 et GH19). Cette étape est sous-divisée en deux activités distinctes qui se succèdent : l'activité endochitinase qui coupe dans les chaines aléatoirement, puis l'activité exochitinase qui utilise comme substrat les diosides ou triosides N-acétylés générés par les endochitinases. Les diosides résiduels, ou chitobiose, peuvent alors êtres utilisés par les chitobiosidases ou N-acétyl-glucosaminidases (GH3, 20, 84, 85) afin de libéré les monomères de N-acétylglucosamine. La deuxième voie (à gauche) fait intervenir la déacétylation en premier via les chitines déacétylases (CE4), suivit de l'hydrolyse par les chitosanases (GH46, 75) et les glucosaminidases (GH2, 9, 35). Certaines enzymes sont équipées de domaine de reconnaissance du substrat chitine (CBM33) pouvant également avoir une activité lytique.

Les mesures enzymatiques ont permis de démontrer que le sol de Rothamsted présentait les deux types d'activités à un niveau relativement comparable à celui d'un autre sol utilisé dans le projet (station expérimentale agricole De Vredepeel, Pays-Bas). L'analyse des jeux de données métagénomiques a révélé la présence de séquences potentiellement affiliées à des enzymes impliquées dans la dégradation de la chitine dans tous les environnements testés, avec une proportion plus grande dans les océans et les sols que dans les tubes digestifs. Cependant, c'est dans le sol que la diversité est la plus importante, en se basant sur la richesse de ces séquences par rapport à celle répertoriée dans les banques de données publiques. En effet, pour un volume de données métagénomiques équivalent en termes de nombre de séquences, le sol présente le plus fort niveau de diversité par rapport à la base publique (Fig.3). Le sol témoin (PR) suit également cette tendance.

Une analyse plus approfondie a permis de mettre en évidence le profil enzymatique du sol par rapport aux autres environnements pour les différentes familles d'enzymes considérées dans cette étude (Fig.3). Il ressort que le sol s'avère être plus riche en enzymes intervenant dans les premières étapes de la dégradation de la chitine par rapport aux autres environnements (Fig.4). Ces enzymes sont précisément celles recherchées dans le cadre de Metaexplore.

Ces résultats indiquent que le sol est un environnement particulièrement prometteur pour la recherche de nouvelles chitinases, du au fait de leur grande diversité taxonomique et fonctionelle, mais également d'un point de vue quantitatif par rapport à d'autres environnements. En revanche, peu de différences ont été observées entre le sol de Rothamsted et les autres sols témoins, tant au niveau de l'activité que de la diversité. Ces résultats ont toutefois conforté le choix de l'environnement sol comme cible pour la recherche de nouvelles chitinases.

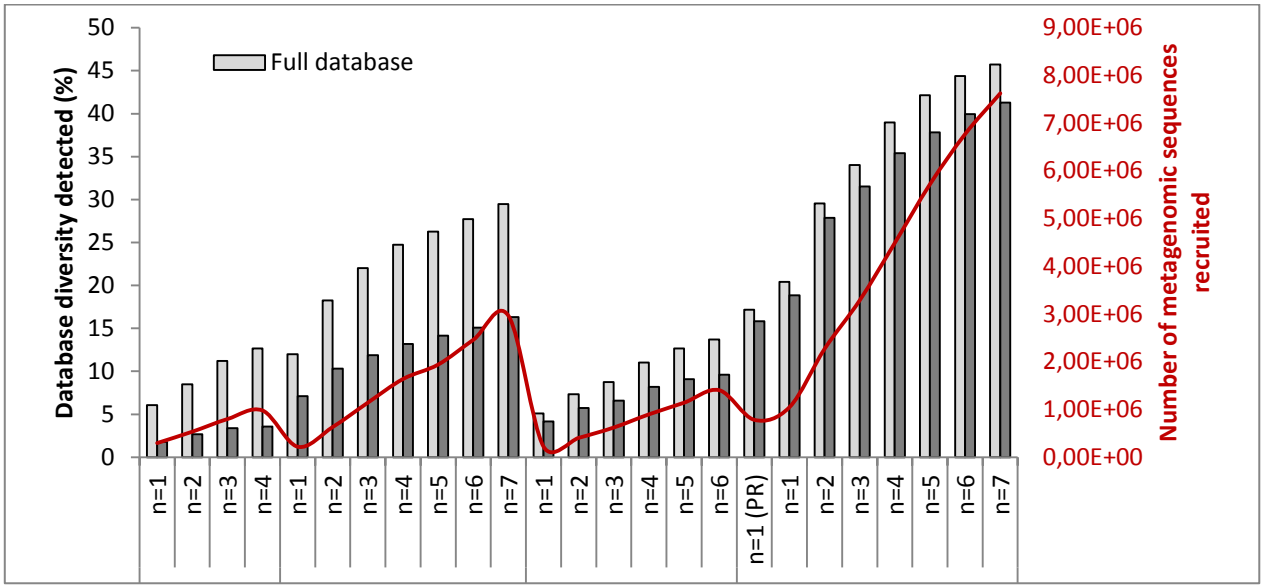


Fig.3: Représentation des niveaux de diversité de séquences affiliées aux chitinases dans les différents métagénomes utilisés. En gris clair sont représentés les pourcentages de diversité de toutes les enzymes impliquées dans la dégradation totale de la chitine. En gris foncé, seules les enzymes impliquées dans les premières étapes sont représentées (CBM33, CE4, GH18 et GH19). La courbe noire représente le volume de séquences métagénomiques recrutées lors du BLASTx à chaque fois qu'un métagénome a été ajouté à l'analyse (n). Le sol témoin est annoté PR (Puerto Rico).

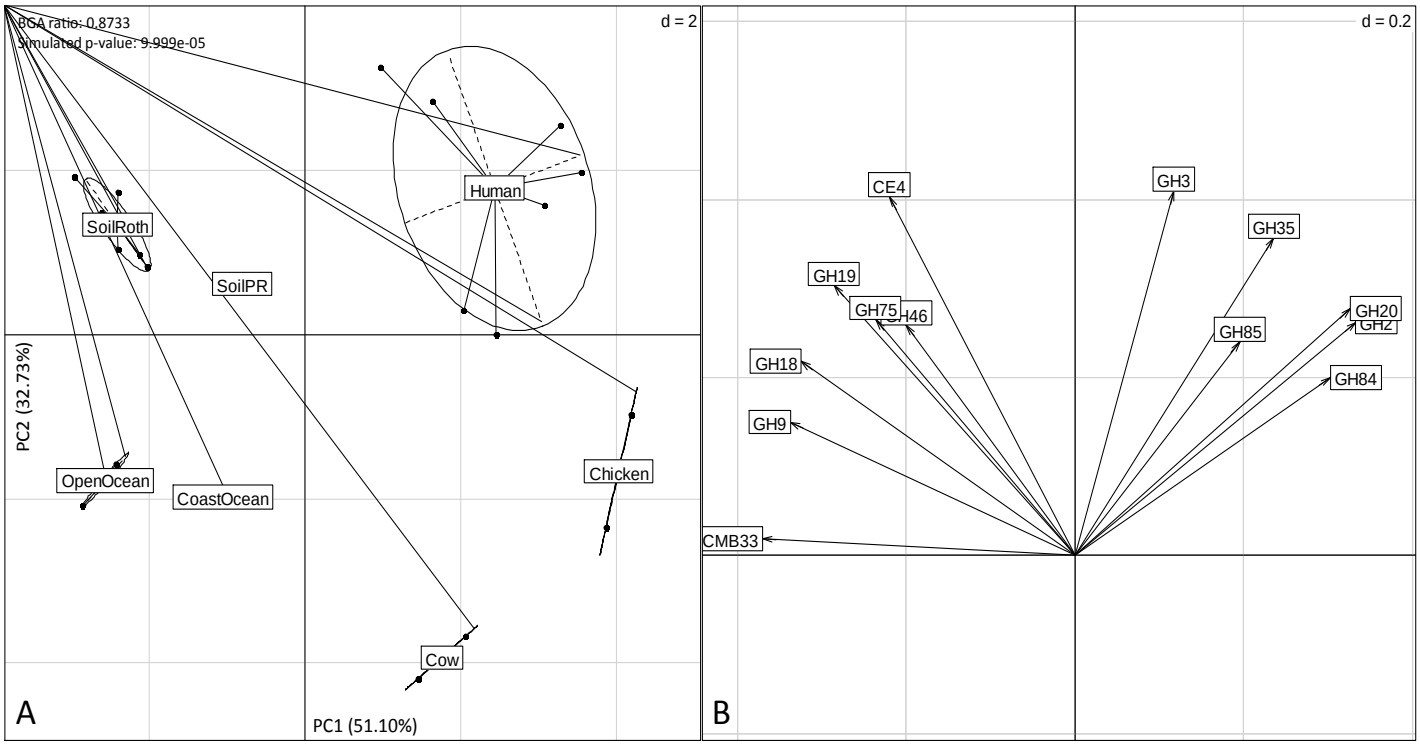


Fig.4: Analyse en composante principale (ACP) des profils enzymatiques de chaque environnement d'après les résultats de BLASTx. L'image A montre l'ACP après groupage des métagénomes selon leur environnement d'origine. Le ratio de groupage est élevé (0,8733) et la p-value du test de Monte-Carlo est hautement significative (9,999E-05) ce qui indique une grande pertinence du facteur de groupage et donc la forte reproductibilité ou similitude entre les profils métagénomiques issus d'un même environnement. L'image B révèle comment les différentes familles d'enzymes se répartissent et influent sur la disposition des échantillons dans l'ACP. Ainsi, les familles à gauche CE4, GH19, GH75, GH46, GH18 et CBM33 sont plus représentées dans les sols et dans les océans, mais avec des profils différents. Tandis que les familles GH3, GH35, GH85, GH20, GH2 et GH84 sont plus présentes dans les tubes digestifs. Or, ce sont précisément les familles de gauche qui participent le plus à la dégradation primaire de la chitine, et qui sont d'intérêt dans le cadre de Metaexplore.

3. Criblage génétique de la banque métagénomique de Rothamsted

Cribler une banque de clones à la recherche d'une catégorie précise de gènes ou d'une activité spécifique reste une tâche difficile à accomplir, du fait que les cibles recherchées sont extrêmement "diluées" dans l'immense diversité du monde microbien, en particulier celui du sol. De ce fait, la probabilité de cloner un gène d'intérêt dans une banque reste extrêmement faible et son expression potentielle l'est d'autant plus. La banque de clones de Rothamsted, constituée avec de l'ADN extrait à partir d'échantillons de sol frais, compte 2 millions de clones contenant chacun un fragment de 40kb d'ADN métagénomique. Cette banque est actuellement la plus importante jamais constituée, avec des biais limités étant donné que l'ADN métagénomique à été extrait à l'aide de différentes méthodes. La constitution d'une telle collection a nécessité énormément d'investissement et la cribler entièrement est une tâche immense, nécessitant des moyens technologiques adaptés (robot automatisé, technologie de crible à haut débit...).

Dans cette partie, 405504 clones (20,3% de la banque totale) ont été analysés par une technique d'hybridation ADN-ADN sur membrane à haute densité, à l'aide d'un marquage radioactif au ^{33}P (Demaneche *et al.* 2009*;* Ginolhac *et al.* 2004). Cette technique consiste à disposer les clones sur des membranes de nylon à l'aide de robot spécifique. Chaque clone est déposé en dupliquât, côte à côte, afin d'avoir une information sur la reproductibilité et éventuellement de déjouer les faux positifs. Les clones sont cultivés afin d'obtenir en théorie 1 million de cellules par dépôt, correspondant à environ 1-2 millions de copies de l'insert métagénomique (le fosmide utilisé ne se maintient qu'a hauteur d'une à deux copies dans chaque cellule). S'ensuit une étape de lyse cellulaire, ainsi qu'une fixation de l'ADN génomique et fosmidique aux UV sur la membrane, puis une phase de lavage. Les membranes ainsi constituées sont prêtes à l'emploi, chaque clone étant précisément déposé suivant un code pariculier, de manière à pouvoir aisément le retrouver dans la banque si besoin. D'un autre coté, des sondes ADN correspondant à la séquence du gène recherché dans la banque sont élaborées.

Dans cette étude, le gène codant la chitinase "chiA" a servi de base à la conception de la sonde par le biais d'une amplification PCR à partir de l'ADN métagénomique du sol de Rothamsted. L'intérêt d'une telle approche est que, contrairement à la PCR qui est basée sur des amorces connues, les sondes ont été sélectionnées parmi les amplicons les plus éloignés par rapport aux séquences de références des banques de données publiques. De plus, la technique tolère un certain degré de divergence entre la séquence de la sonde et celle de la cible (≈50% d'aprés nos observations), et ceci est dû aux conditions d'hybridation appliquées. D'autres gènes ont également été cherchés comme les laccases bactériennes, des déhalogénases, des gènes de résistance à des antibiotiques et des éléments génétiques mobiles. Les sondes sont marquées à l'aide du ^{33}P, ce qui permet une visualisation directe des spots d'hybridation avec une haute sensibilité. Le design des sondes doit en revanche être minutieux afin d'éviter d'éventuelles hybridations parasites avec la séquence du fosmide porteur de l'insert mais aussi avec l'intégralité du génome de la souche *E. coli* utilisée pour contenir la banque.

La technique a permis d'identifier 88 clones positifs après vérification des signaux dupliqués. L'ADN fosmidique de ces clones a ensuite été extrait et mélangé afin d'obtenir la concentration idéale pour un séquençage direct en 454. Les lectures ont ensuite été assemblées afin de reconstituer les inserts métagénomiques et de pouvoir identifier les zones d'hybridation ainsi que les gènes présents. Même si les zones d'hybridation étaient faibles, la technique a permis d'identifier 71 gènes d'intérêt dont : 51 éléments génétiques mobiles, 11 gènes de résistance à des antibiotiques, 7 gènes impliqués dans la dégradation de la chitine ainsi que 1 déhalogénase et 1 laccase. Pour l'instant, l'analyse de ces gènes en vue de leur expression est en cours.

4. Impact de l'enrichissement en chitine sur la communauté bactérienne du sol de Rothamsted

Dans cette partie, le sol de Rothamsted conditionné en microcosmes a été enrichi en chitine afin d'évaluer la cinétique de dégradation de ce composé et de mesurer l'impact sur la communauté bactérienne présente, l'hypothèse de départ étant que l'enrichissement stimule la partie de la population bactérienne la mieux adaptée pour utiliser cette ressource nutritive. Deux concentrations différentes de chitine ont été étudiées : 2mg/g de sol (1x) et 20mg/g de sol (10x), ainsi qu'une condition témoin sans chitine. Une approche métagénomique intégrative a été réalisée afin de répondre à cette question. La cinétique de dégradation a été suivie durant 35 jours en mesurant les activités endo- et exochitinase. L'ADN métagénomique des échantillons de sol a été extrait à chaque temps de prélèvement. Des techniques d'analyses préliminaires ont été employées et ont permis d'établir la manière dont la communauté bactérienne a réagi, d'une part en quantifiant le gène marqueur *rrs* par PCR quantitative (indication sur l'effectif bactérien), puis en établissant l'empreinte moléculaire de la communauté bactérienne en RISA (Ribosomal Intergenic Spacer Analysis). Ces analyses ont permis de mettre en évidence l'effet bio-stimulant de la chitine sur les bactéries, de par une augmentation du nombre de copies dans les conditions enrichies jusqu'a stabilisation à 20 jours (Fig.5, A). Cette observation est cohérente avec les résultats enzymatiques qui montrent un important pic d'activité endochitinase à 20 jours, suivi d'une brusque perte d'activité (Fig.1, B). L'empreinte moléculaire quant à elle, indique un effet "microcosme" qui se traduit par l'apparition de bandes dans le témoin uniquement due à l'incubation.

Suite à ces constatations, les échantillons d'ADN métagénomique obtenus à 20 jours d'incubation, après le pic d'activité, ont été analysés de façon plus approfondie. Des puces à ADN 16S ont été hybridées afin d'évaluer l'impact de l'enrichissement sur la composition taxonomique de la communauté bactérienne. Cette approche a permis de confirmer l'effet bio-stimulant de la chitine, de par l'augmentation du nombre de genres bactériens détectés dans les conditions amendées en chitine 1x et 10x. Du point de vue taxonomique, une nette augmentation du signal correspondant aux Gamma- et Bétaprotéobactéries a été constatée. En revanche, les Actinobactéries, pourtant réputées pour être de bons dégradeurs de la chitine, ont présenté un fort signal uniquement dans le témoin d'incubation. Cette observation peut être corrélée avec les résultats de RISA, où le témoin d'incubation présentait des bandes spécifiques confirmant à nouveau l'effet "microcosme". Enfin, un séquençage direct de l'ADN métagénomique a été réalisé sur tous les échantillons après 20 jours d’incubation. Les jeux de séquences métagénomiques ainsi obtenus ont été comparés à la banque de données de Rothamsted, correspondant au sol frais initial non traité. La banque de séquences métagénomique de Rothamsted correspond à l'ADN extrait à différentes saisons et à différentes profondeurs, mais également par différentes méthodes d'extraction.

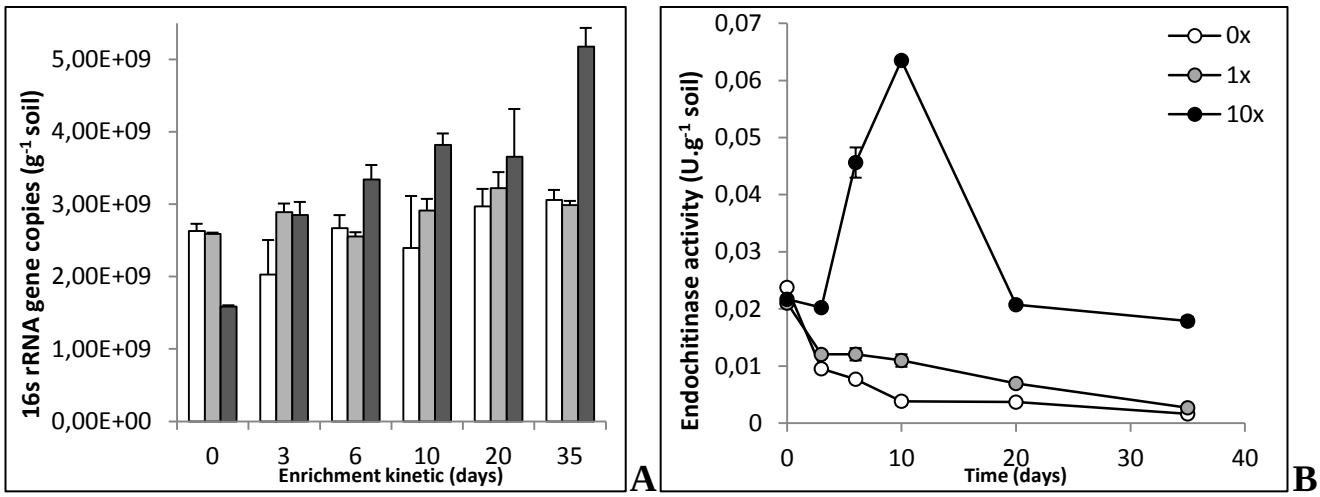


Fig.5: A) Quantification par PCR quantitative du marqueur 16s dans les échantillons d'ADN extraits du sol pendant la cinétique d'enrichissement. B) Mesure de l'activité endochitinase dans les extraits aqueux de sol pendant la cinétique d'enrichissement. La couleur blanche représente le témoin d'incubation, tandis que les deux nuances de gris représentent respectivement la faible (gris clair, 1x) et la forte concentration de chitine (gris foncé, 10x). La quantification du marqueur 16S est exprimée en copie de gène détecté par gramme de sol (A). L'activité endochitinase est mesurée par la quantification de fluorescence émise par le 4-Methylumbelliferon (4MU), après clivage enzymatique. L'activité est exprimée en unité par gramme de sol sur la durée d'incubation (30min), une unité correspondant à 1µmole de 4MU relaché par minute.

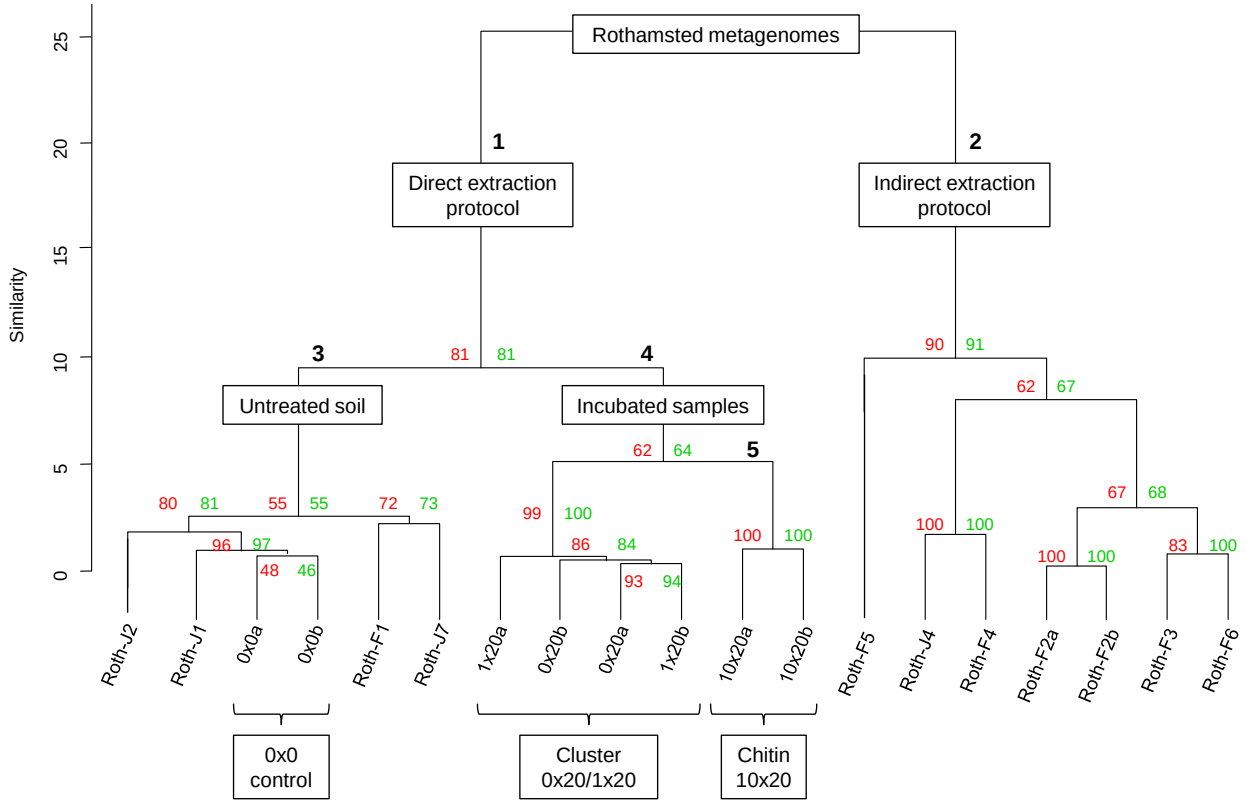


Fig.6 : Classification hiérarchique des métagénomes utilisés dans le cadre de cette étude. Le dendrogramme à été établi sur la base des profils taxonomiques et fonctionels après annotation des métagénomes contre M5NR via MG-RAST. Les groupes ont été soumis à une simulation bootstrap (n=10000) et les AU p-value (Approximated Unbisased) ont été calculé par ré-échantillonage (en rouge pour l'annotation taxonomique, et en vert pour l'annotation fonctionelle). La technique de groupage est basée sur une analyse de variance (Ward 1963) et les distances sont Euclidienne.

Suite à cette comparaison, une hiérarchisation des facteurs influant sur la communauté bactérienne a pu être établie (Fig.6), sur la base des différences observées lors de l'annotation des métagénomes contre les bases de données publiques :

- la technique d'extraction de l'ADN (directe ou indirecte) est le premier facteur structurant la communauté bactérienne. Cette différence est connue et est principalement due à la capacité qu'ont certaines bactéries à résister à la lyse cellulaire dont l'efficacité dépend de la technique utilisée.

- l'effet de l'incubation en microcosmes est le deuxième facteur qui structure la communauté, notamment par une forte augmentation des séquences métagénomiques affiliées aux Actinobactéries dans les conditions incubées. Cette observation vient conforter les précédentes, et l'explication tient au fait que la chitine a été amendée sous forme colloïdale avec de l'eau. Ainsi, pour se placer dans les mêmes conditions, de l'eau a été également ajoutée dans le sol témoin, ce qui a eu pour effet une stimulation des formes dormantes d'Actinobactéries.

- l'effet de l'enrichissement en chitine intervient en troisième position, avec une séparation claire des métagénomes correspondant aux concentrations les plus fortes (10x), tandis que le témoin d'incubation et la faible concentration 1x présentaient de fortes similarités. A nouveau, la proportion de Gamma- et Béta-protéobactéries était plus forte dans la condition 10x.

Au final, 66 genres bactériens ont été statistiquement identifiés comme étant stimulés par l'addition de chitine (1x et 10x confondus). Parmi eux, se trouvent des genres bien connus pour leur capacité de dégradation de la chitine, mais également de nombreux genres qui n'ont jamais encore été signalés comme possédant une activité chitinolytique. C'est le cas par exemple des membres du sous-ordre des *Micrococcineae* (Actinobactéries). Toutefois, cette découverte est à prendre avec précaution, car il est important de souligner que certaines espèces bactériennes opportunistes appelées "tricheuses" auraient pu être stimulées en bénéficiant directement de l'activité de dégradation de leurs voisines.

Si l'on s'intéresse à l'aspect fonctionnel de l'annotation des séquences métagénomiques, très peu de différences ont été observées entre les différentes conditions et la banque de Rothamsted au plus haut niveau de classification. Cet aspect est important, car il démontre que même si des genres particuliers ont été stimulés par l'amendement, l'image fonctionnelle globale de la communauté bactérienne ne s'en trouve pas significativement modifiée. En revanche, certaines fonctions ont été stimulées comme par exemple la dégradation et l'utilisation de la chitine, le métabolisme du fer, certaines fonctions du métabolisme du carbone...

Au final, l'enrichissement peut s'avérer être une stratégie prometteuse pour augmenter la proportion relative de certains genres mieux adaptés pour dégrader la chitine. Toutefois, il est important de rappeler que le potentiel de dégradation de la chitine est déjà bien représenté dans le sol initial, autant du point de vue de l'activité enzymatique que de la diversité génétique.

CHAPTER 2

THE EUROPEAN PROJECT METAEXPLORE: METAGENOMICS FOR BIO-EXPLORATION

-

TOOLS AND APPLICATIONS

-

ENGLISH VERSION

CHAPTER 2 - Part.1

Presentation of the European project METAEXPLORE

SAMUEL JACQUIOD[1], LAURE FRANQUEVILLE[1], TIMOTHY VOGEL[1] and PASCAL SIMONET[1]

[1] Ecole Centrale de Lyon, Laboratoire Ampère, Environmental Genomics Group, 36 avenue Guy de Collongue - 69134 Ecully, France

1. Introduction to the METAEXPLORE project

METAEXPLORE was initiated in 2009, in the frame of the fast development of the environmental metagenomic technologies, with regards to industrial downstream outlets. The purpose of this project is to gather European partners strongly involved in metagenomics and biotechnology, as well as industrial partners, in order to respond to the ongoing demand for new active enzymes. The desired activities are directly derived from industrial needs, with a particular focus on improved enzymes implied in the biodegradation of recalcitrant molecules such as natural biopolymers (e.g cellulose, lignin and chitin), but also xenobiotic compounds (e.g aromatic chemicals). In order to do so, the project is exclusively focusing on metagenomic resources, considering that the major part of the microbiota and their related genes and chemicals remain inaccessible through traditional cultivation-based approaches. The main strategy consists in a pre-screening of habitats of interest in order to identify the desired activities within the microbial communities. The best candidate habitats will be selected for further metagenomic investigations, consisting in clone libraries constitution from the extracted DNA, and also high-throughput sequencing. Finally, libraries will be handled through a deeper screening step, in order to identify positive clones that carry the desired genes and activities. Further characterization and practical tests will be done based on the identified enzymes, before up scaling attempts in pilot fermenters. This strategy is devoted to open up the way for potential downstream industrial applications. The project has been established on a 5-years basis as a schedule for complete achievement of the different goals.

Aside of this guide line, a particular attention will be addressed to the so-called “mobilome”, defined as the collective pool of mobile genetic element (MGE) carried by a microbial community. This consideration is justified by the relative higher occurrence of genes encoding the desired activities into the mobilome. Accessing this genetic pool will be one of the main challenge of this project, as it can considerably enriched clone libraries with relevant genetic determinants, thus increasing the probability to get the desired activities. In addition, the project aims to the development of new biotechnology tools and activity screening assays in order to improve our current level of detection, as well as complementary alternative strategies for metagenomic DNA analysis. Development of bioinformatic tools is also considered, in order to facilitate the handling of the generated sequence data. Finally, the collective work achieved by all partners will result in generation of a database encompassing all sequences and functional genes/operons retrieved along the project.

The project gathers several European partners with strong expertise in environmental microbiology, biotechnology and metagenomic, such as research groups from reputed European university, but also industrial partners implicated in biotechnologies. The work plan is intelligibly organized into seven work packages (WP) which clearly address to specific parts and goals of the project, and specifying the partners involved in its achievement. I did my PhD at l'Ecole Centrale de Lyon, in the Environmental Microbial Genomics Group (EMGG), as partner number #7 in METAEXPLORE. Our participation was to provide expertise in soil microbiology, with a specific focus on the soil used in the frame of the "Park Grass Experiment", from the research station based at Rothamsted (UK). We also brought expertise in metagenomic libraries construction and screening in collaboration with the company LibraGen as a subcontractor (Toulouse, France). Furthermore, we are strongly implied in the development of a new biotechnology tool in collaboration with partner #18 (INRA de Dijon). This tool is called Genefish, and aims to target specific genetic determinants thanks to a new approach based on recombination, which is devoted to be an alternative and complementary strategy to metagenomics. My work consisted in developing and finalizing the Genefish tool in the frame of METAEXPLORE, in collaboration with partner #18. I was also involved in the metagenomic investigation of the Rothamsted Park Grass soil, through the clone library constitution and screening. Finally, I've achieved the development of a strategy relying on soil microcosms amended with chitin in order to evaluate the impact of chitin enrichment on the bacterial community, both in terms of diversity and activity.

2. METAEXPLORE organization

i. Metagenomics and targeted enzymes

As previously presented in chapter one, microorganisms are considered to be the major reservoir of genetic and chemical resources on Earth. This is mostly due to their long term evolution and their impressive adaptation potential, resulting in their wide taxonomical diversification. Thanks to this astounding potential, microorganisms have colonized all kind of habitats, such as soil, water, ice, caves, air and animal guts. They occupy very important niches and are involved in relevant functions within ecosystems. For instance, microorganisms are directly participating into geochemical cycles as major actors through extracellular enzymatic processes (Speir and Ross 2002; Sinsabaugh *et al.* 2008). For instance, organic matter degradation and mineralization is mostly achieved by microorganisms and their related enzymes. However, organic macro-molecules are considered as "recalcitrant", as they are often accumulated in the environment due to slow and complex degradation processes (Horn *et al.* 2006). Complete mineralization of these compounds is implicating many actors at different scales in order to progressively fragment the matter. This long term degradation is finalized by microorganisms through enzymatic processes, resulting in total depolymerization and mineralization of the molecules. Plant biomass is the largest class of macro-molecule in terms of production per year, encompassing cellulose and starch (glucose polymers), and also lignin (monolignols polymer). For instance, the annual yield of worlwide lignocellulosic biomass has been evaluated to exceed 220 billion tons (Ren *et al.* 2009). The second most abundant biopolymer is chitin, a β-1,4-linked N-acetylglucosamine polymer, which is constituting insects and crustacean exoskeleton, and also fungi cell walls.

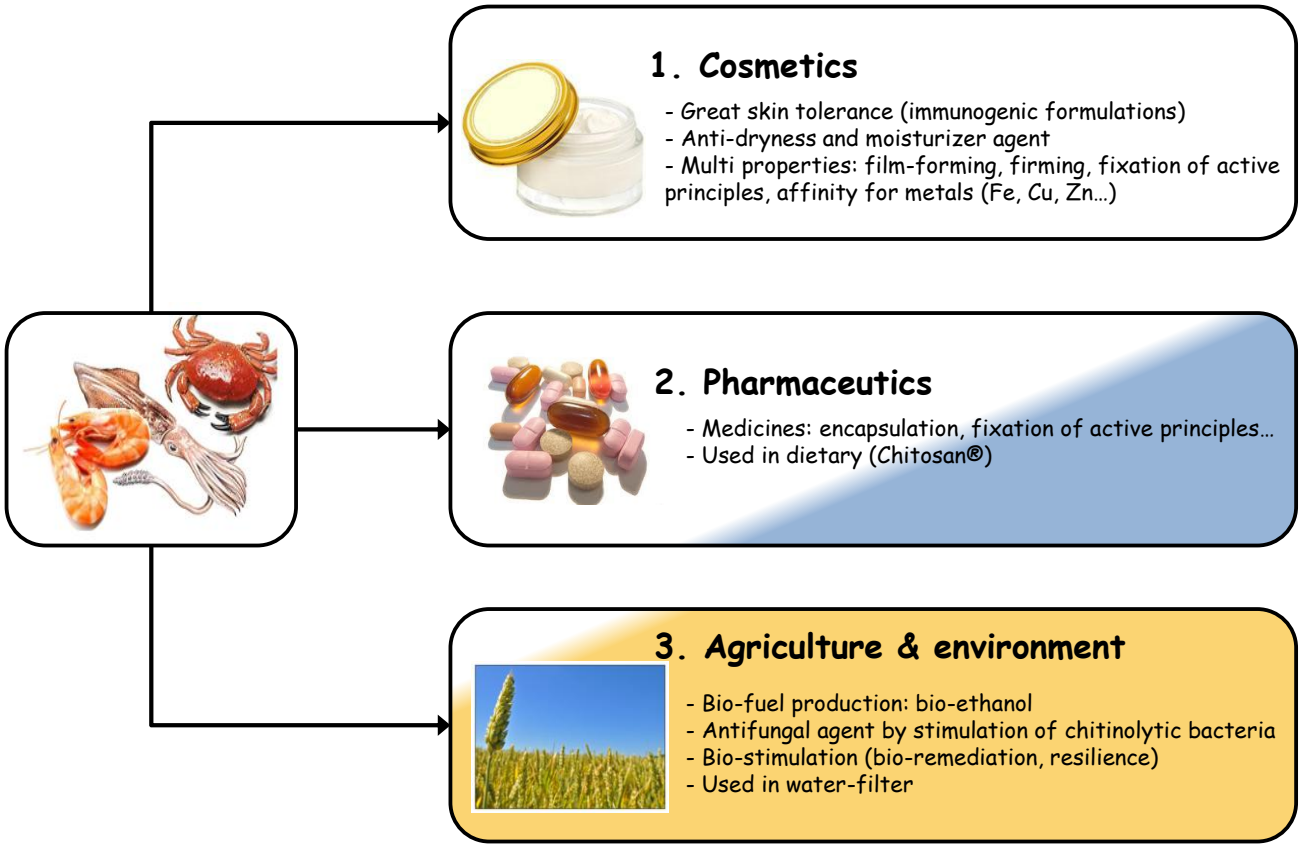


Fig.1: Principal use of chitin derivatives in the industry. The figure is representing the 3 major sectors using products from chitin degradation, as well as their properties of interest and detail concerning their use.

These biopolymers are very abundant in the environment and represent the major part of the on-coming organic matter in the ecosystems. As a consequence, degradation and mineralization of these molecules is a key step in ecosystems functioning, because they represent an important source of nutrients for all organisms. Furthermore, "recalcitrant" compound also encompasses xenobiotic/human made chemicals (e.g pesticides) that were massively used and introduced in the environment during the past decades. These molecules represent a new type of perturbation in the ecosystems, impacting on bacterial communities (Yen *et al.* 2009). However, aside of their known toxicity, these chemicals constitute a novel source of carbon for the ones which are able to degrade them. Thus, in a number of cases, compounds that were previously considered as non-degradable were finally used by microorganisms that were selected through the toxic pressure. This fast adaptation resulted in apparition of novel enzymatic activities for partial and/or total degradation of these recalcitrant xenobiotics. Nowadays, the demand for novel enzymes active toward recalcitrant compounds is becoming urgent in order to valorize the biomass, to solve environmental issues, but also to generate new outlets in the industry. In fact, the biotechnology industry involved in food and waste management is continuously in search of novel enzymes able to degrade recalcitrant natural polymers (e.g lignin, cellulose and chitin), and also polluting man-made compounds (e.g halogenated aliphatic and aromatic chemicals).

The first classes of novel enzymes targeted in this project are responsible in chitin depolymerization and degradation: chitinases and chitin deacetylases, already known and used by the food industry. For instance, the downstream processing of seafood waste (e.g shrimp and crab shells), needs novel enzymatic activities to handle and valorize this biomass. The products generated by chitin degradation are of interest in many sectors like cosmetology (moisture agent, film forming…), pharmaceutic (encapsulation), energy (biofuel) and environment (biostimulation and bioremediation) (Fig.1)

The second class of targeted enzymes is responsible in the intricate organic decay of lignin. Lignin is one of the major components of the lignocellulosic biomass, which harbor an important structural role in plants. Lignin is composed of phenolic molecules assembled through random radical reactions. This particular structure is making lignin really complex, stable, and hardly degradable. Contrarily to other bio-polymer, lignin has no specific organization like chitin or cellulose, which means no specific sites for enzyme module recognition. As a consequence, lignin degradation relies only on oxidation reactions performed by the so-called ligninases, which encompass laccases, peroxidases and hydrogen peroxide producing enzymes. These enzymes are known to be massively produced by fungi species such as white-rot fungi, a major lignin degrader in the environment. Bacteria are also implicated in this degradation, even if not much is known about their enzyme resources (Ausec *et al.* 2011). Therefore, bacteria are seen as a relevant reservoir for ligninases, with potential high activities and turn-over. This would be of interest for many industrial processes involved in plant pulp processing (e.g paper industry and biofuel). Lignin-rich products such as wood and industrial pellets are often used as a source of heat, but enzymatic treatment of lignocellulosic material can open many possibilities and applications (e.g coating, surfactant, adhesive, nanocomposites, fibre products…).

In addition, enzymes involved in environmental sanitation like dehalogenases are also focused in this project, with special regards to marine sponges. As they are known to produce organobromine compounds (Ahn *et al.* 2009), marine sponges are suspected to host specialized microbes with potential relevant functions in relation to these organobromine molecules. Redox enzymes responsible for anaerobic dehalogenation by halo-respiring bacteria will also be targeted, as they act on polyhalogenated aliphatic and aromatic compounds. At the opposite, halogenases, which bind halogen atoms to organic molecules, are also of interest for new pharmaceuticals production (Gribble et al, 2003). Many other enzymes used in the production of pharmaceutical chiral molecules and function transfert/convertion are targeted, like haloalkane and haloalcohol dehalogenases, ammonia lyases and epoxide hydrolases, epoxide-converting enzymes, amidases and lyases/ ammonia-lyases.

Most of the enzymes implicated in the processes mentioned above were identified and characterized from microorganisms retrieved through traditional cultivation-based techniques (Someya *et al.* 2011; Wilson *et al.* 2011; Shankar *et al.* 2011). However, environmental microbiologists agree that only a few part of microbiota is reachable through standard cultivation approaches, and therefore, metagenomics became the most relevant strategy to access the major part of the genetic and functional resources from a microbial pool (Sjöling et al, 2006). Metagenomics represent one of the most popular way for environmental microbial community's exploration and exploitation at the biotechnological level, as it becomes more affordable and as it allows direct access to the global microbial pool. In the frame of the project, a wide range of natural and/or modified aquatic and terrestrial habitats are targeted for further metagenomic investigations such as: natural soil, enriched soil (e.g chitin treated), aerobic and anaerobic composting chitinous waste in landfill sites, peat bog soils, lignin rich habitats, halo-organics polluted soils, biofilters, wastewater, sediments and sponges. The main strategy for efficient metagenomic DNA analysis and screening relies on fosmid clone libraries construction, already known to be an efficient ways to access full operational sequences from uncultured micro-organisms (Kakirde *et al.* 2010). As presented in chapter one, several techniques are available for library screening. There are two main conceptual approaches: (1) a screen based on activity assays in order to detect the eventual production of the desired enzymes; (2) a genetic screen based on the sequence itself through hybridization or PCR approaches. Both techniques are considered in the project, with also the perspective to develop new strategies to enlarge the possibilities. Novel chemical assays are currently being developed by partners with strong expertise in this domain. Furthermore, biotechnological tools are being prepared, like new host/vector systems (Gram-positives and fungi host for instance), as well as the Genefish strategy.

Aside of the metagenomic approach, the project will also focus on the so-called "mobilome", defined as the collective set of genetic mobile elements found in a given microbial community. This genetic resource is known to be strongly participating in bacterial adaptation, as it carries many active genes and operon involved in processes such as pollutant degradation and antibiotic resistance (Ken *et al.* 2011). In order to access this genetic resource, a set of complementary strategies will be used to target the mobilome. For instance, DNA-DNA hybridization has been used to screen libraries for insertion sequences (IS) and integron cassettes, as well as differential centrifugation for environmental plasmid isolation. Finally, a set of bioinformatic tools are also being developed by partner #10 (CeBiTech, University of Bielefeld, Germany) in order to help handling and analysis of the metagenomic data.

ii. Work packages management

As previously described above, METAEXPLORE is an important and ambitious project, with the implication of a significant consortium of 18 European partners from many countries. In order to coordinate the program, and to respond to the scientific goals and tasks, METAEXPLORE has been structured in seven different work packages (WP) which clearly address to specific parts of the project. Each WP is managed by a WP leader, and concern different partners in its achievement. For more information and details about the different European partners involved in METAEXPLORE, please refer to Appendix 2-1a.

WP1 is entitled “Habitat selection and library production - Metagenome and mobilome isolation" which encompasses all preliminary steps of the projects. Habitats of interest are submitted to pre-screening assays in order to detect the targeted enzymatic activities, and also the genetic determinants implicated in the targeted degradation processes. The aim of this WP is to demonstrate that the desired functions are present in the screened habitats, both at the genetic and activity level. This step is essential and crucial, as it will determine the success in finding novel enzymes by focusing on most promising environments and samples. This pre-screening mostly relies on enzymatic assays carried on extracts from environmental samples (e.g chitinases and laccases), and also genetic techniques such as PCR-based detection, plasmid libraries from targeted amplicons (16S, functional genes…), and BLAST analysis. This integrative approach allows identification of the targeted activities and gives an idea of the microorganism’s diversity present in the samples and susceptible to carry these activities. This WP also concerns the DNA extraction protocols to apply, as well as library constitution and storage. Procedures for isolation of the mobilome and its own libraries constitution are also considered in this WP. In order to do so, the project is focusing mostly on environmental plasmid recovery strategies, such as segregation techniques based on differential centrifugation. This allows plasmids purification from environmental DNA, and an affordable plasmid enriched fraction that can further be used for library construction. Another approach consists in examination of the plasmids replication genes as potential genetic markers for targeting purposes. Complementary approaches such as magnetic capture hybridization can be applied to improve isolation rates (Jacobsen, 1995).

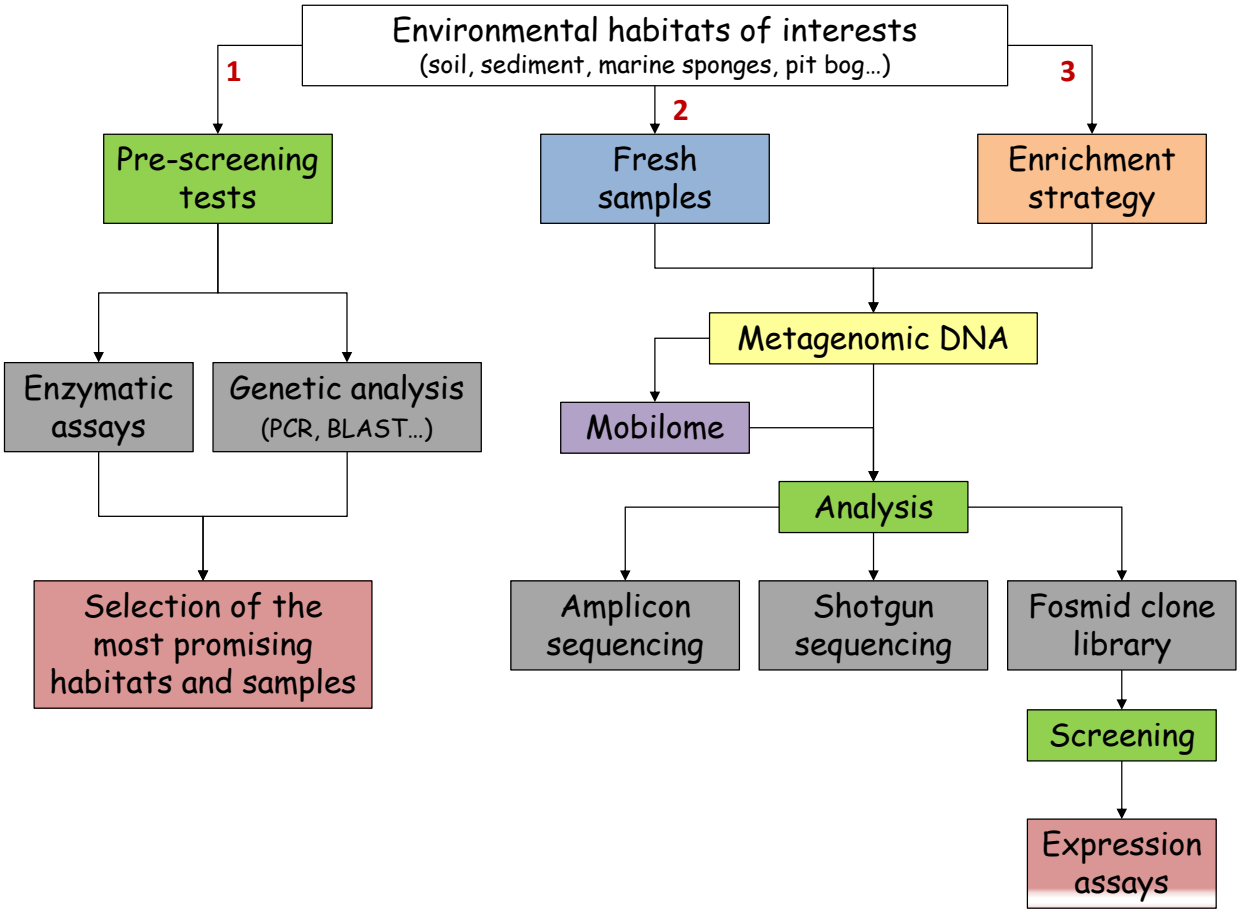


Fig.2: Schematic representation of the metagenomic workflow used in the METAEXPLORE project. The first step (1) consists in having a pre-screening of the environmental samples in order to define whether is the desired activity is present or not. This pre-screen can be achieved by enzymatic assays or by genetic analysis. The next step consist in having a metagenomic approach to seek for these new genes and activities directly from fresh environmental samples (2), or by an enrichment strategy (3) in microcosm for example. After recovering of the metagenomic DNA, several analytic ways are available to investigate for new genes. However, the fosmid clone library strategy is the best way to detect new genes and activities, as the DNA is already cloned into a vector with potential expression.

WP2 “Screening for enzymes – recalcitrant natural compounds” and WP3 “Screening for enzymes – recalcitrant anthropogenic compounds” are major steps in the project, relying on clone libraries screening for the identification of novel activities and/or their related genes. As mentioned, WP2 involves partners with expertise and interest toward degradation of natural recalcitrant compounds (e.g chitin and lignin), whereas WP3 is focusing on xenobiotic degradation enzymes. Metagenomic and mobilome libraries will be screened through high-throughput activity assays, and also a genetic screening thanks to radio-labeled probe hybridization and/or PCR based approaches. Strategies such as clone pooling can be applied in order to increase the screening capacities, and also to minimize the tedious work. Furthermore, the novel biotechnological tools and screening assays developed in WP6 could be used as well in order to enlarge the possibilities for detection.

WP4 “Genetic / functional analysis and production” and WP5 “Sequencing and bioinformatics for environmental metagenomics” are directly linked to WP2 and WP3 as direct following steps in the process of novel enzymes identification. WP4 will focus on characterization of the functions of interest both at the genetic and biochemical levels. The identified genes will be expressed within the clone host strain, and the enzymes characterized in terms of nature (amino acid sequence, folding, active sites…), activity (substrate specificity and kinetic properties…) and expression (genes involved, regulation, optimal host…). Unraveling of the genetic resources contained into the clones will be achieved by complete sequencing of the metagenomic inserts thanks high throughput sequencing technologies. For comprehensive analysis of these data, WP5 is providing the bioinformatic resources necessary in terms of computational power and tools. This will facilitate handling of the high-throughput sequencing data generated. All sequences generated by the project will become available by constitution of a database in the frame of the WP5.

WP6 “Novel tools for environmental metagenomics” is a parallel line in the project, for the development of novel tools and systems for alternative metagenomic exploration. These approaches are devoted to become complementary of the current metagenomic strategies. The purpose of this WP is to design new host, vectors and expression systems that can be used as alternatives to the current *E. coli* – fosmid libraries systems. In addition, attempts to improve the expression rates in *E. coli* will be done. The WP is also focusing on developing new activity assays to improve the current level of detection by enlarging the screening possibilities in terms of substrate specificity and kinetic. The development of the Genefish tool belongs to this WP as well.

WP7 “Management and organization” encompass all consideration regarding the organization and management of the project. Several reporting periods are scheduled in order to report the results achieved to the European Union.

For further information and details about the work package organization, please refer to Appendix 2-1b.

iii. Partner #7 implications

As presented above, my commitment in the METAEXPLORE project was to carry my PhD research within the Environmental Microbial Genomics group (EMG) at l'Ecole Centrale de Lyon (ECL). We are included in the project as Partner #7, represented by Timothy M. Vogel (Professor, University Claude Bernard Lyon 1), Pascal SIMONET (research director, CNRS), and Laure FRANQUEVILLE (research engineer, CNRS). EMG group has noticeable expertise in environmental microbiology and ecology, with strong focus on soil habitats. Indeed, EMGG is already engaged in an international consortium called TERRAGENOME, which aims at unraveling completely the actual microbial diversity from a reference soil thanks to next generation sequencing technologies, and the constitution of a subsequent fosmid clone library of 2 million clones.

The consortium is carried on by the METASOIL project, supported by funding from the French Agency for Research (Agence Nationale de le Recherche ANR). The purpose of this project is to propose to the scientific community a reference soil metagenome with an extensive and deep analysis of the bacterial community. In the frame of the Consortium, the soil from the Park grass station in Rothamsted Research Station (Rothamstedd, UK) has been selected to become the reference soil (Vogel *et al.* 2009). The research station has extensive expertise and data on this park grass, as they carried a long term experiment over 150 years on this park grass, still running nowadays.

The data generated and the experience earned during the on-going of the METASOIL project have been used and applied to METAEXPLORE. For instance, the 2 million clones' library from the Rothamsted soil metagenome has been built in the frame of METASOIL by the company LibraGen. Furthermore, LibraGen is involved in METAEXPLORE as a subcontractor of partner#7 ECL. Thus, one way to use the clone library is to make it available for screening purposes in the frame of METAEXPLORE. In addition, the work realized by the former PhD student of METASOIL: Tom O. Delmont, resulted in the constitution of a huge sequence database generated from a set of different DNA extraction techniques applied on the Rothamsted soil (Delmont *et al.* 2011). These sequences were obtained thanks to 454 pyrosequencing technology, in partnership with the French national sequencing center: Le Genoscope (http://www.genoscope.cns.fr/spip/). This database has been used for sequence-based screening through BLAST search in order to identify genetic determinant related to the desired activities (e.g chitinases, ligninases and dehalogenases).

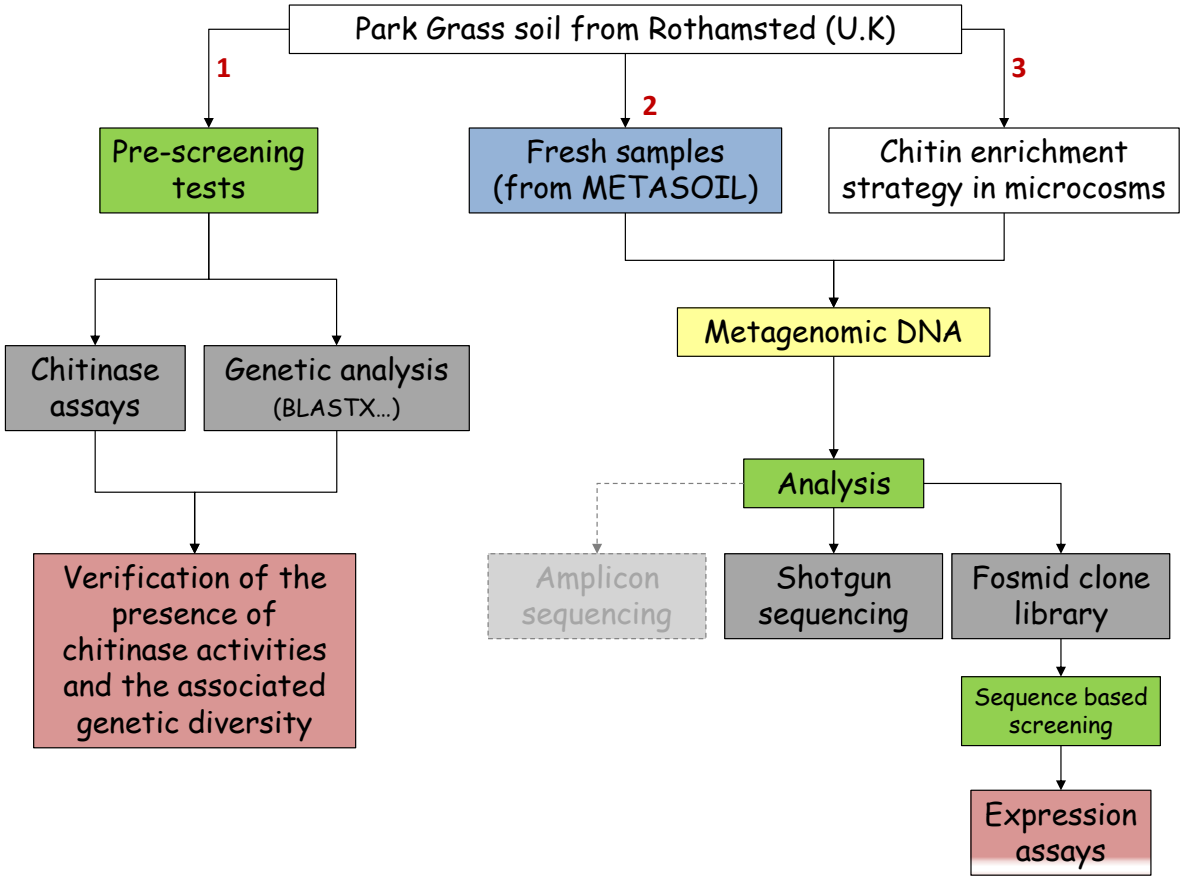


Fig.3: Schematic representation of partner#7 strategy in the METAEXPLORE project. The EMG group is mostly focusing on the soil of the Park Grass experiment at Rothamsted (UK). The pre-screening of this soil was carried on by checking the presence of chitinase-related reads into the metagenomic datasets generated in the frame of the Metasoil project. Enzymatic assays were also performed to quantify the presence of different chitinase activities. Then, a metagenomic analysis was performed on both fresh soil samples and also chitin enriched samples in microcosms. Direct shotgun sequencing (454) of the metagenomic DNA has been performed, as well as a fosmid clone library that was screened through a sequence based strategy.

Besides this preliminary approach, the clone library has been partly analyzed with a sequence-based screening. The technique selected relies on the design of radio-labeled probes from PCR products of genes implicated in the synthesis of the targeted enzymes. These probes were used in hybridization attempts against spotted clones onto high density nylon membranes. In addition, another strategy was used in parallel in order to assess the impact of chitin enrichment onto the microbial community of Rothamsted soil in microcosm's experiments. An integrative approach was used, coupling the data from enzymatic assays carried on soil samples, as well as 16S qPCR assays and RISA profiles in order to follow the community shift along the enrichment kinetic. An in depth analysis of the bacterial diversity has been performed using 16S phylochip and total DNA pyrosequencing. Finally, I've participated in the development of the Genefish tool with Laure Franqueville and in collaboration with Laurent Philippot, partner#18 (INRA de Dijon). The chapter 3 is fully devoted to this part of my work.

The purpose of this chapter is to focus on my participation in METAEXPLORE, with regards to the metagenomic aspects that were covered by the project. The first part presented here described the preliminary results obtained from BLAST screening of the Rothamsted metagenomic database for chitin-related degrading enzymes. Then we will focus on the screening of the fosmid clone library from the Rothamsted soil, which was done through a sequenced-based technique relying on DNA-DNA hybridization. The last part presented in this chapter is exposing another metagenomic strategy consisting in enriching the Rothamsted soil with chitin in microcosm in order to evaluate in impact on the prokaryote community in terms of taxonomy and functions. This chapter will end with a conclusion that summarizes the work and results obtained in the frame of the METAEXPLORE, as well as further perspectives and openings for the project.

References

Ahn YB, Kerkhof LJ, Häggblom MM (2009) Desulfoluna spongiiphila sp. nov., a dehalogenating bacterium in the Desulfobacteraceae from the marine sponge Aplysina aerophoba. *Int J Syst Evol Microbiol*. 59:2133-2139

Ausec L, Zakrzewski M, Goesmann A, Schlüter A, Mandic-Mulec I (2011) Bioinformatic analysis reveals high diversity of bacterial genes for laccase-like enzymes. *PLoS One*. 2011;6(10):e25724

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011a) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol*. 77:1315-1324

Gribble GW (2003) The diversity of naturally produced organohalogens. *Chemosphere*. 52:289-297.

Horn SJ, Sikorski P, Cederkvist JB, Vaaje-Kolstad G, Sørlie M, Synstad B, Vriend G, Vårum KM, Eijsink VG (2006) Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides. *Proc Natl Acad Sci USA*. 103:18089-94.

Jacobsen, CS (1995). Microscale detection of specific bacterial-DNA in soil with a magnetic capture-hybridization and PCR amplification assay. *Appl Environ Microbiol. 61*: 3347-3352.

Kakirde KS, Parsley LC, Liles MR (2010) Size does matter: Application-driven approaches for soil metagenomics. *Soil Biol Biochem*. 42:1911-1923

Ren N, Wang A, Cao G, Xu J, Gao L (2009) Bioconversion of lignocellulosic biomass to hydrogen: Potential and challenges. *Biotechnol Adv*. 27:1051-1060

Sen D, Van der Auwera GA, Rogers LM, Thomas CM, Brown CJ, Top EM (2011) Broad-host-range plasmids from agricultural soils have IncP-1 backbones with diverse accessory genes. *Appl Environ Microbiol*. 77:7975-7983

Shankar S, Shikha. (2011) Laccase Production and Enzymatic Modification of Lignin by a Novel Peniophora sp. *Appl Biochem Biotechnol*. [Epub ahead of print]

Sinsabaugh RL, Lauber CL, Weintraub MN, Ahmed B, Allison SD, Crenshaw C, Contosta AR, Cusack D, Frey S, Gallo ME, Gartner TB, Hobbie SE, Holland K, Keeler BL, Powers JS, Stursova M, Takacs-Vesbach C, Waldrop MP, Wallenstein MD, Zak DR, Zeglin LH. (2008) Stoichiometry of soil enzyme activity at global scale. *Ecol Lett*. 11:1252-1264

Sjöling, S, Cowan, DA. (2008) Metagenomics – microbial community genomes revealed. Ch. 20 in Psychrophiles: from Biodiversity to Biotechnology. Margesin, R.; Schinner, F.; Marx, J.-C.; Gerday, C. (Eds.) 462 pp. ISBN: 978-3-540-74334-7

Someya N, Ikeda S, Morohoshi T, Noguchi Tsujimoto M, Yoshida T, Sawada H, Ikeda T, Tsuchiya K. (2011) Diversity of culturable chitinolytic bacteria from rhizospheres of agronomic plants in Japan. *Microbes Environ*. 26:7-14.

Speir T.W., Ross D.J. (2002). Hydrolytic enzyme activities to assess soil degradation and recovery. In "Enzymes in the Environment: Activity, Ecology and Applications" R.G. Burns, R.P. Dick Eds. Marcel Dekker, Inc., New York, pp. 403-431.

Vogel TM, Simonet P, Jansson JK, Hirsh PR, Tiedje JM, Van Elsas JD, Bailey MJ, Nalin R, Philippot L (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Microbiol*. 7:252

Wilson DB. (2011) Microbial diversity of cellulose hydrolysis. *Curr Opin Microbiol*.14:259-63

Yen JH, Chang JS, Huang PJ, Wang YS (2009) Effects of fungicides triadimefon and propiconazole on soil bacterial communities. *J Environ Sci Health*. 44:681-689

CHAPTER 2 - Part.1

-

APPENDIXE 2-1a

This is an excerpt from the official text constituting the "DOW OF METAEXPLORE (222625)", from 27/01/09, which is summarizing the description of the work, according to the different Work Packages.

Overall strategy and general description

The project will encompass parallel activities taking place in the different laboratories of partners. All activities are grouped into seven work packages (**WP**, numbered 1 through 7), which form the logical units within which coherent work will take place across the different laboratories. A tight organization of the work, in which the work package leaders, the co-ordinator and the project coordination committee will play key roles, will ensure that the different parts of the work (habitat selection, sampling and material distribution, DNA extraction, screening and distribution of results, clone selection) will come together in a timely and organized fashion. Together, the work packages, and the tasks formulated therein, will make up a fully consistent project. The outline below is described in more detail in the seven work package descriptions provided in Tables 1.3a.

WP1 - In a first project stage, we will screen selected natural or pretreated (enriched) environments, on the basis of (bio)chemical analyses that assess the prevalence and degradation of target compounds, and environmental DNA (obtained using elaborate nucleic acid extraction methods), assessing the prevalence and diversity of preselected genes/operons indicative of conversions of interest, mobile genetic elements (MGE) and micro-organisms. In this screening, we will use advanced activity-based (functional) and genetic screening methods, allowing us to assess types and levels of target activities in the respective environment. The habitats to be sampled will include multifold soils and soil-related (chitin and/or lignin - treated soils, peat bogs, biofilters for pesticide removal) as well as aquatic (river bank sediments, wastewater, sponges) environments, to which the Consortium partners have access. The screening work will be executed, in first place, in the laboratories that obtained the samples. If needed, the Consortium may decide that test results are to be confirmed in the laboratory of a collaborating expert partner. The (bio)chemical analysis will provide the key guidance for further stages of the work, whereas the genetic analysis will provide supporting data in respect of presumed function and diversity. From among the prescreened sample habitats, up to seven habitats, which stand out as most promising for mining the target enzymatic activities, will be selected. To this selection, two sample environments with no conspicuous / high levels of apparent target enzymes, will serve as negative controls. This will further validate the screening efforts and assist in enhancing the predictive power of the molecular screening. Thus, a key outcome of the (pre)screening effort, combined with subsequent analyses at the level of the metagenomes, will be the provision of validated screening tools that allow us to indicate the environmental samples that are most likely to lead to success.

Following sample selection, we will isolate high-quality environmental DNA and construct metagenomic libraries from the nine selected habitats. We will add these libraries to the already existing three suppressive soil-derived libraries available within the Consortium, thus giving a total of 12 microbial genetic resources. Whereas all habitats will, thus, have a library prepared on the basis of the full metagenome, for a subset of habitats, libraries will also be constructed on the basis of the mobilome, following separation of this metagenome component. Several Consortium partners possess unique expertise, allowing them to successfully access the mobilome of selected habitats. The mobilome is key to the project, as it allows efficient access to sequences of key novel biodegradative genes which are localized on MGE such as plasmids. To isolate these, an approach consisting of cell detachment and subsequent separation, massive cell lysis, DNA extraction and separation of plasmid from chromosomal DNA, will be used. We will also examine the use of the replication functions of the plasmids as markers and as determinative characters for maintenance in target hosts, and use magnetic capture hybridisation – PCR (Jacobsen 1995) to enhance isolation rates. MGE DNA will be preserved in hosts in microtiter plates at -80ºC and subjected to highthroughput sequencing analysis. As the choice of the proper host and vector will be critical to the success of metagenomics exploration, we will construct the environmental gene libraries in broad-host-range (shuttle) expression plasmids, that is, vectors that can be used for gene expression in hosts beyond *Escherichia coli*, which still constitutes the main lab workhorse. We will also use available vectors of Gram-positive bacteria. From previous environmental metagenomics work, we have access to shuttle vectors that allow work in actinobacteria, bacilli, *Pseudomonas* and *Ralstonia* or *Burkholderia* and, on the basis of existing work, we will in a later stage dispose of *Acidobacterium*, *Verrucomicrobium* and *Desulfitobacterium* vectors. In addition, partner 11 is developing a heterologous expression system based on the fungus *Trichoderma reesei* (see also under WP6). This versatility in terms of host selection is necessary to overcome the problem of possibly poor expression of target genes/operons in *E. coli*. The libraries will, thus, be introduced into different (selected or designed) bacterial hosts and used for high-throughput activity screenings.

WP2 and **WP3** - In a subsequent project phase, screenings for target enzymes will be performed, using activity- (activity screening) and gene sequence (genetic screening) - based approaches. Activity screening will be directed towards gene functions for the degradation of both natural (WP2) and anthropogenic (WP3) compounds. The libraries, both metagenome and mobilome, will thus be screened, addressing the different applications, i.e. chitinases, ligninases, dehalogenases/halogenases and ammonia lyase/epoxide hydrolase activities. This will be done by selection for growth on the target compounds (i.e., chitin, lignin, aliphatic halogenated compounds) or on products of the initial biodegradation, dehalogenation/halogenation or deamination/deamidation reactions. For these tests, high-throughput assays based on microtitre and agar plates are available. Genetic screening will be based on genetic tools (hybridization and/or PCR based) whenever required to examine our libraries for (pre-established) gene sequences. In addition, the aptness of several novel methodologies available with different partners to enhance our success in library screenings in high throughput will be explored, such as “Pooling”, the use of “Molecular traps”, the “GeneFish method” and “Tandem expression vectors” (see work packages).

WP4 and **WP5 -** Following the screenings, a next project phase aims at a detailed characterization of the genes and the functions they encode. Selected functions of interest derived from the screenings will thus be characterized at the genetic (DNA sequence) and biochemical levels (substrate range, kinetics). Genes encoding the target enzymes will be expressed in *E. coli* and/or alternative hosts. The enzymes and the genes that encode them, will then be analyzed so as to understand their nature (amino acid sequence, folding, target active site, etc), functioning (substrate specificity, kinetic properties) and expression (genes involved, regulation thereof, optimal host for expression). Moreover, high-throughput sequencing coupled to expression screening will shed light on genes encoding target functions present in the mobilome. In addition, the original metagenome fragments expressing desired functions will be completely sequenced, including genes flanking the desired gene function. An important asset of this project will be the bioinformatics tools and metagenome sequence database (**WP5**). All sequences generated will become available in this webbased tool.

WP6 - A parallel line in the project aims at the design of novel host/vector/expression systems. Thus, novel vector systems for expression in the aforementioned key alternate hosts will be applied and developed, for which both novel replication and gene transfer functions from the mobilome itself and novel hosts now available with partners will be employed. In addition, several approaches to increase the success rate of expression in *E. coli* will be assessed.

WP7 - This work package concerns the organisation and management of the project.

CHAPTER 2 - Part.1

-

APPENDIXE 2-1b

This is an excerpt from the official text constituting the "DOW OF METAEXPLORE (222625)", from 27/01/09, which is presenting the different European partners involved in this project.

List of Beneficiaries

Beneficiary Number *	Beneficiary name	Beneficiary short name	Country	Date enter project**	Date exit project**
1(coordinator)	RUG-ME	Rijksuniversiteit Groningen	Netherlands	T0	T60
2	RUG2	Rijksuniversiteit Groningen	Netherlands	T0	T60
3	UCPH	University of Copenhagen	Denmark	T0	T60
4	SH	Sodertorn Hogskola Stockholm	Sweden	T0	T60
5	U Warwick	University of Warwick	United Kingdom	T0	T60
6	WU	Wageningen Universiteit	Netherlands	T0	T60
7	ECL	Ecole Centrale de Lyon - subcontractor LBG	France	T0	T60
8	UB	University of Bielefeld	Germany	T0	T60
9	JKI	Julius Kuhn Institut, Braunschweig	Germany	T0	T60
10	KU Leuven	Katholieke Universiteit Leuven	Belgium	T0	T60
11	VTT	Valtion Teknillinen tutkimuskeskus	Finland	T0	T60
12	UCL	Universite Catholique Louvain	Belgium	T0	T60
13	UNINS	University of Insubria, Microbial Technology and Fermentation group - subcontractor Actygea (**SME***)	Italy	T0	T60
14	UL	University of Ljubljana	Slovenia	T0	T60
15	UNLP	University of La Plata	Argentina	T0	T60
16	BDS	Bio Detection Systems (**SME**)	Netherlands	T0	T60
17	WSL	Whitby Seafoods Ltd (**SME**)	United Kingdom	T0	T60
18	INRA	INRA	France	T0	T60

CHAPTER 2 – Part.2

The metagenomic database from Rothamsted:

–

A bioinformatic strategy to screen for chitinase genes

SAMUEL JACQUIOD[1], LAURE FRANQUEVILLE[1], ANNA M. KIELAC[2], MARIANA S. CRETOIU[2], EMMANUEL PRESTAT[1], TOM DELMONT, JAN D. VAN ELSAS[2], TIMOTHY VOGEL[1], PASCAL SIMONET[1]

[1] Ecole Centrale de Lyon, Laboratoire Ampère, Environmental Genomics Group, 36 avenue Guy de Collongue - 69134 Ecully, France

[2]Department of Microbial Ecology, Centre for Ecological and Evolutionary Studies, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands

1. Introduction to the BLAST screening strategy

As previously mentioned, the Environmental Microbial Genomics Group from ECL is implicated in two metagenomic projects aiming to explore the soil from Rothamsted: METASOIL, through the deep analysis of this reference soil, and METAEXPLORE through the screening of the datasets and libraries generated from this soil. During the progress of METASOIL, the soil has been extensively studied in order to define a strategy that would optimize metagenomic DNA extraction. The main purpose was to obtain the best overview of the inherent microbial community, through the recovery of as much diversity as possible. Finally, an efficient strategy has been set up in order to increase the diversity accessible after extraction (Delmont *et al.* 2011) at different depths, and seasons as well. Many DNA extraction protocols were used such as direct techniques, relying on lyses of the cells within the sample, and also indirect techniques based on differential centrifugation for bacterial cells isolation (Delmont *et al.* 2012). The DNA so obtained from these different approaches has been respectively used for deep sequencing purposes by pyrosequencing at the Genoscope (Paris, France).

In parallel, a subsequent fosmid library of 2 000 000 clones (≈40kb insert on average) has been constructed at LibraGen from high molecular weight DNA extracted by the agarose plugs technique (Bertrand *et al.* 2005). The library and data generated from the Rothamsted soil are of interest for the METAEXPLORE project, as it represents an opportunity for screening targeted activities and their related genes. In this part, we present how the knowledge generated during the METASOIL project was used in order to screen for genetic signatures of enzymes of interest. In a first move, the sequences obtained through pyrosequencing were analyzed by BLASTX to find genetic determinants related to chitin degradation. Furthermore, chitinase activity assays were performed on fresh wet soil for assessing the presence of the desired enzymes *in situ*. As the public databases are relatively poor in terms of bacterial laccases and the enzymatic assays did not worked on the Rothamsted soil, we decided to not focus on these enzymes at the time, thus this work was essentially focused on chitin degradation enzymes.

Chitin is an homopolymer of β-1,4-linked N-acetylglucosamine (N-acetyl-D-glucose-2-amine, NAG) with a critical biological role in both terrestrial and aquatic ecosystems as a major constituent of fungi and plant cell walls, insect, krill and shellfish exoskeletons (Gomes *et al.* 2001). Chitin is the second most abundant biopolymer in nature after cellulose, with an estimated natural production of 10^{10} tons year^{-} (LeCleir *et al.* 2007; Muzzarelli *et al.* 1999). The degradation of chitin releases substantial amounts of carbon and nitrogen in terrestrial (Williamson *et al.* 2000; Metcalfe *et al.* 2002; LeCleir *et al.* 2004; Uchiyama & Watanabe, 2006), aquatic (LeCleir *et al.* 2007, Ramaiah *et al.* 2000, Hobel *et al.* 2005) and Antarctic ecosystems (Xiao *et al.* 2005). Its crystallized conformation and heterogenic chemical composition render chitin particularly recalcitrant to degradation (Howard *et al.* 2003). Its complete mineralization is carried out only by a highly specialized microflora through specific microbiological enzymatic processes.

All known enzymes involved in chitin degradation are classified in the CAZy database (Carbohydrate Active Enzymes, http://www.CAZy.org/, Cantarel *et al.* 2009, Fig.1), the chitinases (EC 3.2.1.14) belonging to the glycosyl hydrolases families GH18 and GH19 according to the classification based on amino acid sequence similarity (Henrissat *et al.* 1991; Henrissat 1993). Another chitin degradation pathway relies on chitin deacetylases (EC 3.5.1.41), belonging to the carbohydrate esterase family CE4 (Lombard *et al.* 2010) and chitosanase (EC 3.2.1.132) from family GH46 and GH75. Chitinases are also classified based on their depolymerization activity including endochitinases (EC 3.2.1.14) which randomly cleave chitin molecules and exochitinases such as β-(1,4)-N-acetylglucosaminidases (EC 3.2.1.30) and 1,4-β-chitobiosidases (EC 3.2.1.29) which progressively degrade chitin molecules from non-reducing ends (Graham *et al.* 1994; Harman *et al.* 1993).

2. Material and methods

i. Soil sampling and DNA extraction

Fresh soil cores were collected from Park Grass (lat 51.481481°N, long 0.222231°E), Rothamsted, England (see http://www.rothamsted.ac.uk/ for further information). The Park Grass from Rothamsted is an internationally recognized resource and is targeted to be a reference soil for soil metagenomic studies (Vogel *et al.* 2009). It is classified as chromic luvisol according to FAO guidelines (FAO, 2006) and is silty clay loam overlying clay with flints with a pH of 5.2 (measured in H_2O). Park Grass covers 249 m^2 (13.28 by 18.75 m), and the sampling strategy consisted in taking randomized soil samples in the plot. The soil was sampled at two different time points corresponding to the winter season (February 2009) and the summer season (July 2009). Two depths were sampled: the top 10 centimeters down to 21 centimeters. The soil cores were around 6 cm diameters for 20cm depth, and were collected into plastic bags and rapidly transferred at the Ecole Centrale de Lyon (France). Soil particles were sieved at 2mm and directly used to run microcosm experiments. All tools and materials used were washed and cleaned with 70% ethanol solution.

The DNA has been extracted according to different protocols encompassing direct extraction (lysis within the soil sample) and indirect extraction (lysis applied on extracted bacterial cells). Bacterial cells were segregated from the soil samples through differential centrifugation in a Nycodenz® gradient. The sequences generated by pyrosequencing were pooled in order to constitute a database of 7 627 785 sequences. For further details on soil sampling and DNA extraction procedures, please report to the reference from Delmont *et al.* 2012. Moisture was adjusted around 65% of the water capacity before performing the chitinase assays.The soil from The Netherlands was collected in the experimental farm station De Vredepeel and analyzed by the METAEXPLORE Partner #1 (Anna M. Kielac, Mariana S. Cretiou and Jan D. Van Elsas from the University of Groningen, The Netherlands). The soil was characterized as a sandy soil with pH 5.7 and 2.2 % organic matter. Triplicate 4-kg soil samples were removed from the 10 cm topsoil from subplots. Soil was homogenized by passing through a 2 mm mesh sieve and the moisture was adjusted to 65% of the water capacity (approx. 16%).

ii. Chitinase assays

600mg of Rothamsted soil were collected and 1.2ml phosphate buffer salt pH8 (PBS) has been added. Pulled samples were mixed in a vortex for 2 minutes at maximum speed (Vortex Genie 2, SCIENTIFIC INDUSTRIES). The mixture was clarified by centrifugation at 13'000 rpm and 500µl supernatant was transferred in a clean collection tube and stored on ice. Exochitinase and endochitinase assays were performed on 10 µl of supernatant in duplicate, using a fluorometric chitinase assay kit following provider instructions (Sigma-Aldrich, Germany, Lifesciences). Specific substrate are provided for each activity (4-Methylumbelliferyl N-acetyl-β-D-glucosaminide for exochitinase and 4-Methylumbelliferyl β-D-N',N''-triacetylchitotriose for endochitinase). After chitinases degradation, both substrates release 4-Methylumbelliferone (4MU), a fluorescent compound emitting at 450nm when exited at 360nm. Fluorescence was measured after 45min incubation at 37°C on a microplate reader (Infinite 1000, TECAN) and activities were stated with a 4MU standard curve. Enzymatic activities were normalized and expressed in chitinase units detected per gram of soil during incubation time (1U=1µmole 4MU released per minute). For comparison purposes, activity data obtained in the same way from another soil in another lab were used (De Vredepeel, The Netherlands, University of Groningen, METAEXPLORE partner#1).

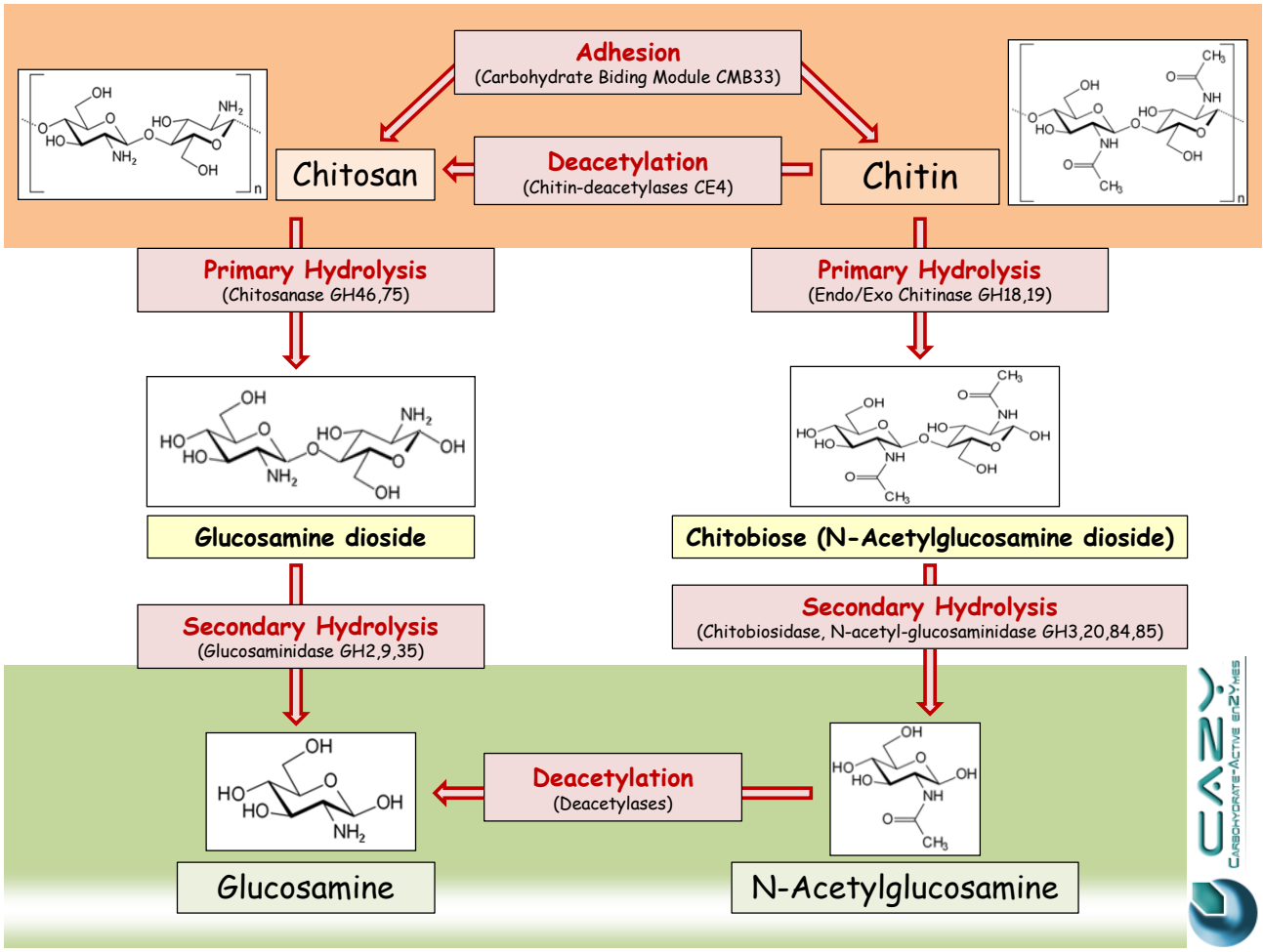


Fig.1: Complete enzymatic degradation of chitin. The figure is displaying all the enzymes families from CAZy involved in total depolymerization of chitin fibers, such as Glycosyl Hydrolase (GH), Carbohydrate Binding Modules (CBM) and Carbohydrate Esterase (CE) families.

Metagenomes	Sampling time	Depth	Extraction protocol	Average sequences length (bp)	Number of sequences	Number of bases
Roth-E1	February 2009	0-21 cm	Direct MP Bio101	365	1047544	3,82E+08
Roth-E41	July 2009	0-21 cm	Direct MP Bio101	409	1223539	5,00E+08
Roth-E4B	February 2009	0-21 cm	Indirect MP Bio101	354	1018152	3,61E+08
Roth-EA4	February 2009	0-21 cm	Indirect MP Bio101	427	1205025	5,15E+08
Roth-F47	July 2009	0-10 cm	Indirect DNA Tissue	443	1229827	5,45E+08
Roth-F48	February 2009	0-10 cm	Indirect DNA Tissue	410	1050181	4,31E+08
Roth-L01	February 2009	0-10 cm	Indirect lysis in plugs	408	853517	3,48E+08

Tab.1 Description of the 7 metagenomes from Rothamsted park grass soil used in this study.

iii. Metagenomes and BLASTX assays

7 metagenomes obtained from Rothamsted Park grass soil and corresponding to different sampling and extraction techniques were collected (Tab.1). In order to seek for genetic determinants related to chitin depolymerization, a protein database encompassing most of the known enzymes and peptides involved in its degradation has been built. The protein sequences were retrieved from the Carbohydrate Active enZyme datase (CAZy, http://www.CAZy.org/, Cantarel *et al.* 2009). The major part of enzymes implicated in chitin depolymerization is represented in Figure 1 (Fig.1).

Chitin is first submitted to a substrate recognition step by carbohydrate binding modules CBM33. It can be handled through two pathways: (1) deacetylation and degradation by chitin deacetylases CE4, chitosanase GH46 and GH75, and glucosaminidase GH2, GH9 and GH35; (2) direct hydrolysis thanks chitinases GH18 and GH19, followed by chitobiosidases and N-acetyl-glucosaminidase GH3, GH20, GH84, GH85. The products generated by these pathways are glucosamine and N-acetyl-glucosamine, which can be further used in the glycolysis process. Most of these enzymes are secreted as the major part of the degradation process occurs outside of the cells and/or in the periplasm. In total, 11 499 protein sequences have been gathered from CAZy (February 23th, 2011), corresponding to the different families mentioned above. A lot of these sequences shared high similarity, as the databases are displaying lots of redundancy.

The protein database so-generated has been cleaned by the software CD-HIT-EST, by performing sequence clusters at 90% similarity (Huang *et al.* 2010). A consensus sequence was generated from each cluster in order to remove highly similar inputs. After cleaning, 59.19% of the sequences were removed and the final protein database harbored 6806 unique sequences. Table 2 shows the different enzyme family used and the related number of sequence available in CAZy (Tab.2). The sequences obtained were used to build a protein database that was screen thanks BLASTX with query metagenomes to find related genetic sequences matching the known proteins (E-value cutoff: E-07). Only the best hits were considered for each read.

For enlarging the screening, several publically available metagenomes were downloaded from MG-RAST server (http://metagenomics.anl.gov/) and used as BLASTX control for comparison purpose. A set of publically available metagenomes from diverse environment has been constructed from MG-RAST (Meyer *et al.* 2008). The selected metagenomes were generated thanks the same sequencing technology. Details about the metagenomes used are given in table 3 (tab.3). Data were analyzed using principal component analysis (PCA) and between group analysis (BGA), with Rgui software (R Development Core Team, 2011) and with the package ADE4TkGUI (http://pbil.univ-lyon1.fr/ade4TkGUI).

Group	CAZy family	E.C number	Description	CAZy (23/02/2011)	CD-HIT-EST (90%)	Percentage after cleaning
Carbohydrate Binding Module	CBM33	None	Binds to chitin fibers, monooxygenase activity	370	221	59,73
Carbohydrate Esterase	CE4	EC 3.5.1.41	Catalyze the deacetylation of chitin ($R\text{-}NH\text{-}COCH_3$ -> $R\text{-}NH_2$)	3024	1674	55,36
Glycosyl Hydrolase	GH2	EC 3.2.1.165	Exo-β-glucosaminidase	1818	979	53,85
	GH3	EC 3.2.1.52	β-N-acetylhexosaminidase	2384	1562	65,52
	GH9	EC 3.2.1.165	Exo-β-glucosaminidase	346	257	74,28
	GH18	EC 3.2.1.14 EC 3.2.1.96	Chitinase (III and V), Endo-β-N-acetylglucosaminidase	2028	1114	54,93
	GH19	EC 3.2.1.14	Chitinase (I, II and IV)	383	244	63,71
	GH20	EC 3.2.1.52	β-hexosaminidase	680	441	64,85
	GH35	EC 3.2.1.165	Exo-β-glucosaminidase	237	167	70,46
	GH46	EC 3.2.1.132	Chitosanase	54	41	75,93
	GH75	EC 3.2.1.132	Chitosanase	12	12	100,00
	GH84	EC 3.2.1.52 EC 3.2.1.169	β-N-acetylhexosaminidase, β-N-acetylglucosaminidase	96	63	65,63
	GH85	EC 3.2.1.96	Endo-β-N-acetylglucosaminidase	67	31	46,27
Number of protein sequences				**11499**	**6806**	**59,19%**

Tab.2: Enzymes involved in chitin degradation. Table B represents enzymes based on CAZy classification. As several enzymatic activities were reported for each family, the respective enzymatic commission numbers (E.C) corresponding to the desired activity was mentioned. Sequences were retrieved from CAZy (23/02/2011) and cleaned thanks CD-HIT-EST software to remove redundancy (90% similarity).

MG-RAST access number	Name	Samples origin	References	Average sequences lenght	Sequence number	Base volume
4440452.7	Human1	Human feces	Turnbaugh and Gordon 2009 Turnbaugh *et al.* 2009 Reyes *et al.* 2010	238	217386	5,17E+07
4440460.5	Human2			229	414754	9,50E+07
4440595.4	Human3			201	510972	1,03E+08
4440611.3	Human4			198	495040	9,81E+07
4440613.3	Human5			335	302780	1,01E+08
4440614.3	Human6			177	519072	9,20E+07
4440616.3	Human7			345	502399	1,73E+08
4440283.3	Chicken1	Chicken caecum	Qu *et al.* 2009 Dinsdale *et al.* 2008	104	294682	3,07E+07
4440284.3	Chicken2			104	237940	2,47E+07
4441679.3	Cow1	Cow rumen	Brulc *et al.* 2009 Dinsdale *et al.* 2008	101	264849	2,66E+07
4441680.3	Cow2			102	178713	1,82E+07
4446892.3	RothE1	Soil (Rothamsted)	Delmont *et al.* 2012	365	1047544	3,82E+08
4446896.3	RothE41			409	1223539	5,00E+08
4446895.3	RothE4B			354	1018152	3,61E+08
4446894.3	RothEA4			427	1205025	5,15E+08
4446902.3	RothF47			443	1229827	5,45E+08
4446903.3	RothF48			410	1050181	4,31E+08
4446904.3	RothL01			408	853517	3,48E+08
4446153.3	SoilPR	Soil (Puerto Rico)	Luquillo Experimental Forest Soil, Puerto Rico**.** Terry Hazen**,** Lawrence Berkeley National Lab	412	780588	3,22E+08
4443697.3	OpenOcean1	Open ocean	David Karl, University of Hawaii, Open Ocean South Pacific Subtropical gyre	232	276234	6,40E+07
4443713.3	CoastOcean1	Coastal ocean	Edward F. Delong, Massachusetts Institute of Technologies, Coastal water, California	239	217549	5,20E+07
4443714.3	CoastOcean2			237	186172	4,42E+07
4443716.3	CoastOcean3			237	216796	5,14E+07
4443725.3	OpenOcean2	Open ocean	Stephen J. Giovannoni, Oregon State University, Ocean Gyre, Sargasso Sea	229	244125	5,58E+07
4443729.3	OpenOcean3			233	261387	6,09E+07

Tab.3: Metagenomes used in this study for BLASTx chitinase screening. Metagenomes are publically available and were downloaded from the MG-RAST server (http://metagenomics.anl.gov). Several environments are represented: Human feces, animal feces, soil and open/coastal ocean.

3. Results and discussion

i. Chitinase assays

The enzymatic assays directly performed on fresh wet samples revealed the presence of both exochitinase and endochitinase activities in the Rothamsted Park grass soil (Fig.2). Both activities were found to be ranging at the same proportion, with a slightly higher level for the exochitinase function. When comparing with the soil from the experimental farm station De Vredepeel (The Nertherlands), both activities are found to range at the same level, but in a leaser proportion for De Vredepeel samples, and with higher error bars as well. This experiment shows that the two chitinases activities are relatively well represented in the fresh soil of Rothamsted when comparing to the other soil. In addition, this experiment shows that the exochitinase activity seems to be stronger than endochitinase in both soils used in this study. This can be explain by the fact that these two activities are occurring successively during chitin degradation, starting with endochitinases which release the subtract for exochitinases. As chitin degradation is known to be important in soil, it is conceivable to infer that most of the free chitin available is rapidly handled by endochitinases, thus explaining why the exochitinase activity is slightly dominant.

ii. BLASTx: Abundance in metagenomes

Over the 6806 unique protein sequences present in the database, 4345 were detected through the BLASTx screening (63.84%). Furthermore, over the 1866 sequences related to enzymes involved in the first step of chitin degradation were retrieved (chitinase GH18 and GH19, chitin deacetylase CE4, chitin binding CBM33), 1671 were detected (89.55%). When looking at the abundance level of these sequences in the different metagenomes used in this study, interesting patterns were detected (Fig.3, light grey). For instance, it is interesting to see that genetic signatures related to chitin degradation are found in all the environments used in this study. This is mostly due to the major role of chitin in the environment, which is known to be an important source of carbon and nitrogen in many ecosystems (Lindahl *et al.* 2006). However, the relative abundance is found to be quite different from an environment to another. As an example, the abundance is lower in animal metagenomes, especially in cow rumen. This can be partly attributed to the size of the metagenomes, but not totally. For example, coastal metagenomes are displaying the same abundance level than cow rumen samples for the complete database (light grey nuance), but the proportion of chitinase is significantly higher in coastal samples (dark grey nuance). Furthermore, differences can be observed within aquatic habitats, with higher abundance in open ocean metagenomes than coastal ones. This difference is not due to the metagenomes size, as they are ranging in the same level in terms of base volume and sequences (Tab.3), while the abundance is found to be increased 3-fold (Fig.3). The higher abundance of chitinase related sequences in open ocean samples can be explained by the fact that chitin is one of the preponderant carbon and nitrogen source in pelagic habitats, while the lower abundance observed in coastal samples is probably due to higher proportion and diversity of other carbonated substrates from the coast. However, when focusing on the enzymes specialized in the first steps of chitin degradation (chitinase GH18 and GH19, chitin deacetylase CE4, chitin binding CM33), the patterns are changing for some habitats (Fig.3, dark grey). For instance, the abundance of chitinase is drastically decreasing in animal, human related metagenomes and coastal samples. Open Ocean and soil metagenomes are displaying the highest values, which is coherent as most of the chitin is known to be produced in aquatic and terrestrial habitats (Uchiyama & Watanabe, 2006; LeCleir *et al.* 2007). The soil from Puerto Rico displays the same patterns than Rothamsted Park grass.

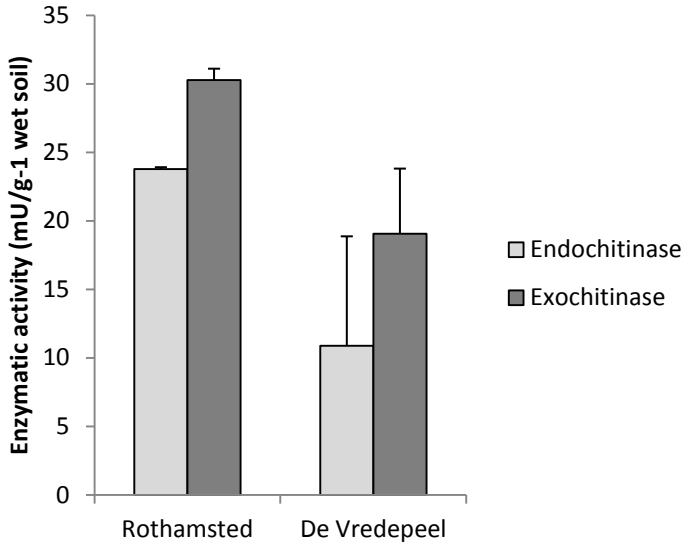


Fig.2: Exo- and endochitinases fluorometric assay on soil sample aqueous extract. Activity was measured by quantification of the fluorescence release by the 4-Methylumbelliferon (4MU) after specific cleavage of exo- and endochitinases in both case. Enzyme activity is expressed in units of chitinase detected per gram of soil during incubation time (1U=1µmole 4MU released per minute) (Jerkovic et al. 2010; Akrem et al. 2011; Hurst et al. 2011)

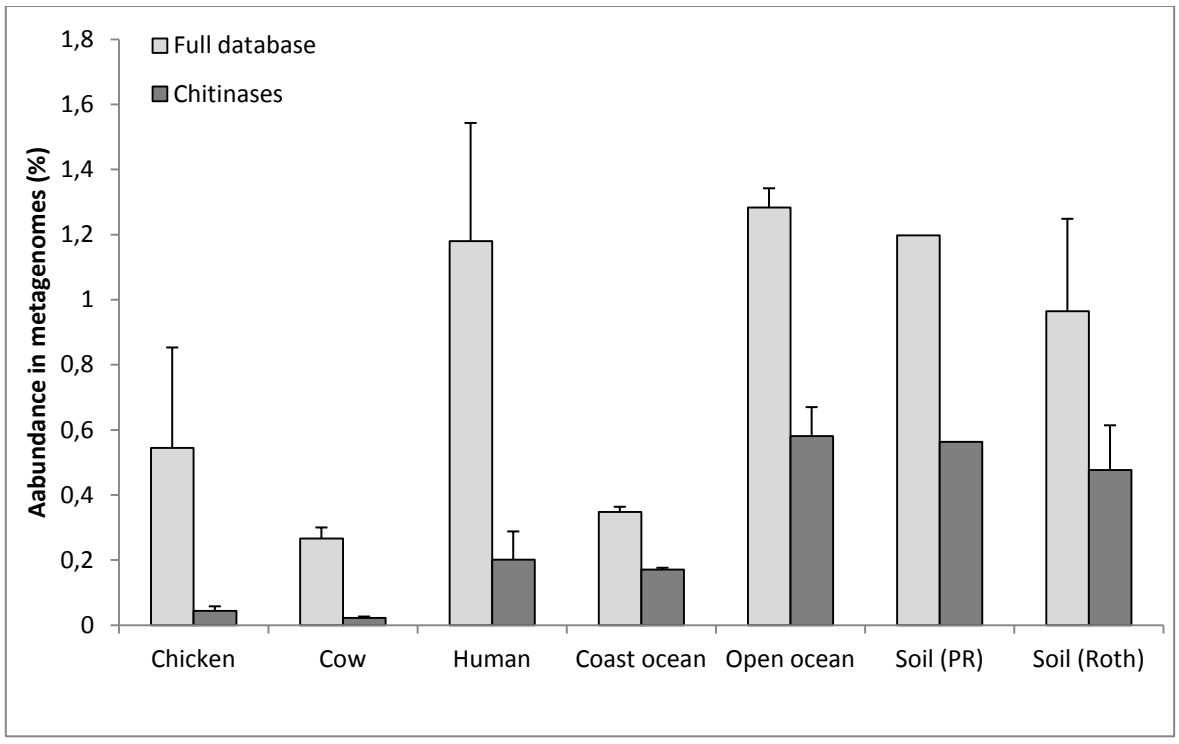


Fig.3: Relative abundance of chitin degrading enzymes in metagenomes after BLAStx annotation (E-value cutoff: E-07). The graphic represents the average percentage of sequence in metagenomes related to a specific habitat with an annotation toward chitin degrading enzymes. Error bars are representing the standard deviation found for each habitat. Results are normalized against total sequences per metagenomes. Metagenomes have been classified in function of their original habitat (Animal: 2 metagenomes from chicken feces and the 2 from cow rumen; Human: 7 metagenomes from human feces; Open ocean: 3 metagenomes; Coastal ocean: 3 metagenomes; Soil Roth: 7 metagenomes from Rothamsted, UK, Soil PR: one metagenomes from Puerto Rico). Light gray nuance represents the percentage against the full database (n=6806 protein sequences) while dark gray represents only the portion related to chitinases GH18, GH19, CBM33 and CE4 (n=3306 protein sequences).

iii. BLASTx: Diversity in metagenomes

When looking at the sequence diversity retrieved from the database that was detected in the different habitats, relevant patterns were observed as well (Fig.4). The sequence diversity retrieved in soil metagenomes is significantly higher than in other environments. This is particularly true when looking at the chitinase diversity (Fig.4, dark grey), with around 45-50% of the database diversity detected in Rothamsted Park grass metagenomes. Here again, the soil from Puerto Rico is also showing the same pattern, with lower diversity values than Park grass. This is partly due to the fact that soil metagenomes are the biggest in terms of sequences in this analysis, thus impacting directly on the probability of getting more diversity (Tab.3).

However, the sequence volume cannot explain this trend alone. For instance, by cumulating 6 ocean metagenomes, which represents more than 1.4 million sequences, the sequence diversity detected is still lower than a single soil metagenomes with around 1 million sequences. Fig.4 clearly shows that the sequence diversity detected is obviously impacted by the sequence volume in metagenomes, but habitat plays a non-negligible role as well when the same amount of sequence is used in the analysis. In aquatic habitats, coastal and open-ocean samples displayed exactly the same level of diversity, with 13.72% of the full database detected and around 13% of the chitinases diversity retrieved (data not shown). Animal metagenomes are displaying higher diversity level toward the full database than chitinases. Surprisingly, human metagenomes are displaying equivalent patterns as soil metagenomes when considering the full database. The diversity level is even found to be higher when considering the same amount of sequence. However, when considering chitinases, this trend is totally reversed, with no differences in terms of representation between full database and chitinases in human samples, while soil displays significantly more diversity for chitinases. It is noteworthy that the pattern is reversed for terrestrial and aquatic environment, with higher diversity level for chitinases than the full database.

This segregation between terrestrial/aquatic metagenomes and human/animal metagenomes is clearly highlighted on the cluster dendrogram based on the BLASTx annotation pattern (Fig.5). The first dichotomy is clearly grouping apart the terrestrial/aquatic metagenomes and human/animal metagenomes. The second dichotomy is clustering aquatic related metagenomes from soil metagenomes while the cow rumen metagenomes are grouping apart on the other side. Then the third dichotomy is clustering the Puerto Rico soil metagenome apart from Rothamsted samples while open ocean and coastal ocean are clustered separately. On the other side, the chicken metagenomes are grouped apart from human feces metagenomes. This highlights the fact that the metagenomic genetic determinant related to chitin degrading enzymes is really specific, with identifiable genetic signatures corresponding to the original habitat.

This tendency can be observed on the protein families' representation in the metagenomes as well, through principal component analysis (PCA, Fig.6). Clear groups corresponding to the original habitat can be formed based on metagenome annotation profiles in the different enzyme families used in this study (Fig.6, panel A). In addition, the segregation observed previously is still present here on the PC1, with and terrestrial/aquatic metagenome clustered on the left, while human/animal related metagenomes are grouped on the right of the PCA.

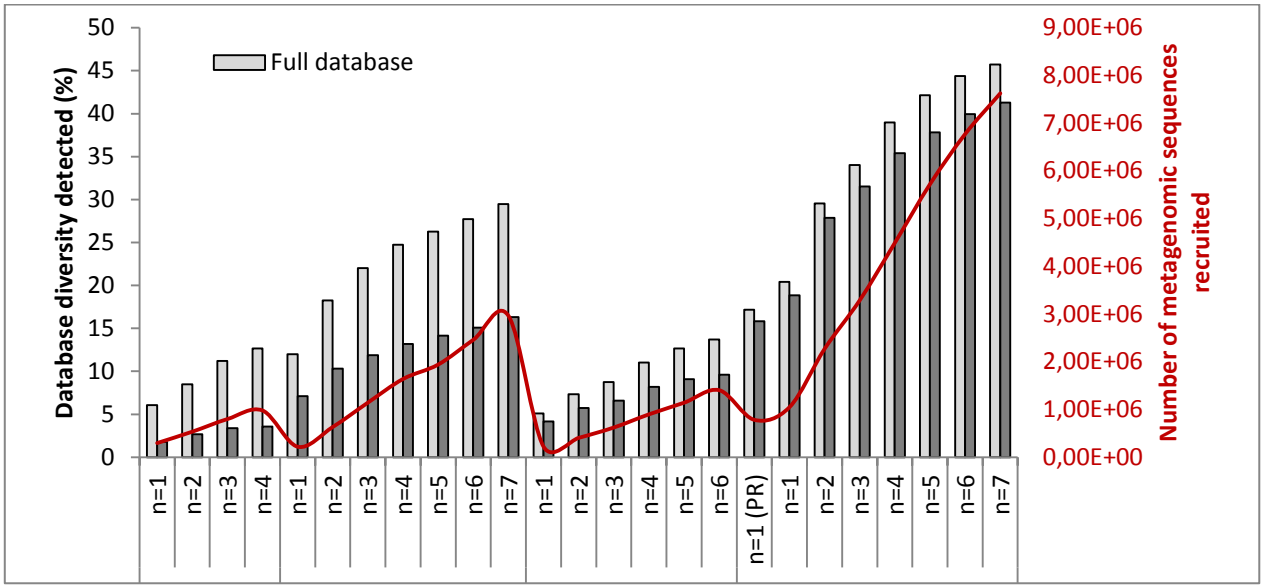


Fig.4: Diversity of chitin degradation enzymes detected in different habitats. The graphic represents the percentage of protein sequences from the database with a significant match in the metagenomes after BLASTx (E-value cut-off: E-07). Metagenomes were recruited progressively for each habitat in order to increase the number of sequences, and to avoid redundancy in the detection. The red curve displays the number of sequences recruited in the analysis, which is directly linked to the number of metagenomes used. Metagenomes have been classified in function of their original habitat (Animal: n=4; Human: n=7; Ocean: n=6; Soil Roth: n=7, Soil PR: n=1). Light gray represents the percentage against the full database (n=6806) while dark gray represents only the portion related to chitinases GH18, GH19, CBM33 and CE4 (n=3306).

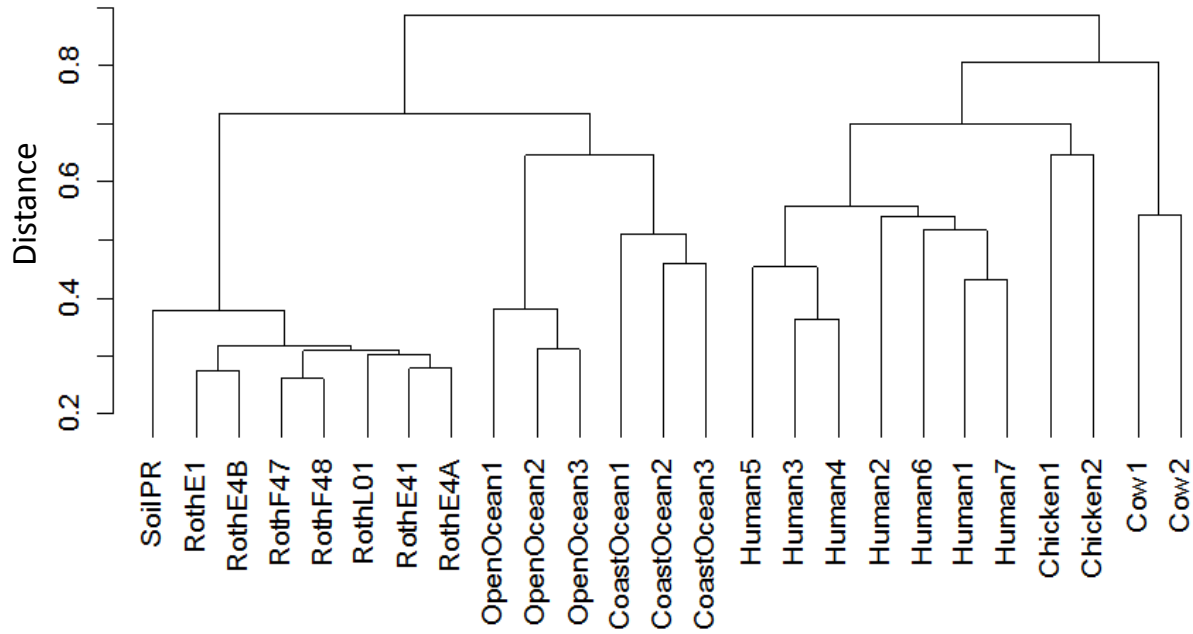


Fig.5 Cluster dendrogram realized on the metagenome patterns after BLASTx. The clusters were calculated using the Bray Curtis dissimilarity distance (Bloom 1981) and the average clustering method.

Furthermore, when looking at the enzymes families as driving vectors in the PCA, the same tendency is detected (Fig.6, panel B). All the families corresponding to the enzymes involved in the first step of chitin degradation are systematically found to be in higher proportion in terrestrial and aquatic metagenomes, whiles families corresponding to downstream degradation steps are more represented in human/animal metagenomes. Going into the details, differences can also be observed inside these two groups when considering the PC2. For instance, animal related metagenomes are clearly clustering apart from human metagenomes while ocean and soil metagenomes are separated as well on the other side. According to panel B, chitin binding CM33, chitinase GH18 and glycosyl hydrolase GH9 are more represented in oceans, with higher abundance in open ocean samples than the coastal ones. Chitinase GH19, chitin deacetylase CE4, and chitosanase GH46 and GH75 are found only in soils, with lower levels in the sample from Puerto Rico. On the other side, families GH2, 3, 20, 35, 84 and 85 are found in higher proportion in human feces metagenomes. Animal related metagenomes are generally displaying lower representation levels than any other environment, probably due to lower metagenome size (Tab.3).

iv. BLASTx: Taxonomical affiliation

When looking at the taxonomical affiliation of the metagenomic sequences against the database, the same segregation can be observed between human/animal metagenomes and soil/ocean metagenomes diversity (Fig.6). The taxonomical patterns are quite similar between human and animal metagenomes, with high level of Bacteroidetes and Firmicutes, while the proportion of Actinobacteria and Proteobacteria is found to be higher in terrestrial and aquatic metagenomes. The higher proportion of Firmicutes and Bacteroidetes has been already shown in the human feces metagenomes (Turnbaugh *et al.* 2009), in chicken cecum metagenomes (Qu *et al.* 2008) and cow rumen metagenomes (Brulc *et al.* 2009). Human feces display higher proportion of Actinobacteria and Proteobacteria, but lower Bacteroidetes than animal metagenomes. On the other hand the profiles from aquatic and terrestrial metagenomes are quite similar, with minor variations. However, the proportion of phyla that are poorly represented is systematically found to be higher in ocean/soil metagenomes. The global picture of the bacterial community involved in chitin depolymerization seems to be more diverse in aquatic and terrestrial environments, which makes sense as chitin is one of the major on-coming sources of organic matter in these habitats (Williamson *et al.* 2000; Souza *et al.* 2011). When comparing the taxonomical affiliation between the full database and chitinases only, differences can be noticed as well (Fig.7, panel A and B). In both human and animal metagenomes, the proportion of Firmicutes is considerably increased when focusing on chitinases only (GH18, GH19, CE4, CBM33). In soil and ocean metagenomes, the proportion of Actinobacteria and Proteobacteria is increased when considering only chitinases. This is coherent with previous work that already showed prevalence of these phyla in chitin degradation process (Manucharova *et al.* 2011).

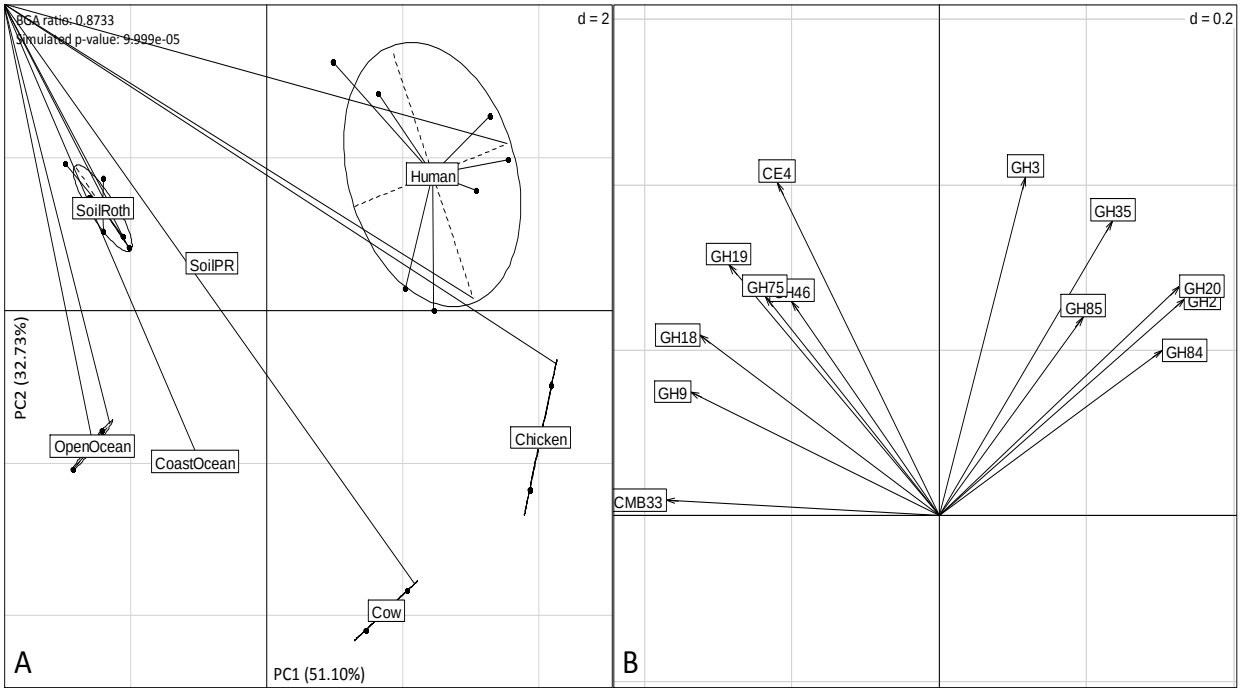


Fig.6: Between group analysis (BGA) of the metagenome annotation against different enzymes families related to chitin degradation. Panel A displays the principal component analysis (PCA) performed on metagenomes patterns with clusters representing the different habitat. BGA ratio indicates the distance modifications between the PCA and the BGA. Simulated p-value was calculated with a Monte Carlo simulation to test the cluster relevance. Principal Component 1 and 2 (PC1 and PC2) are respectively corresponding to 51.10% and 32.73% of the observed data variability. Panel B represents the enzyme families as vectors impacting on the metagenomes patterns distribution.

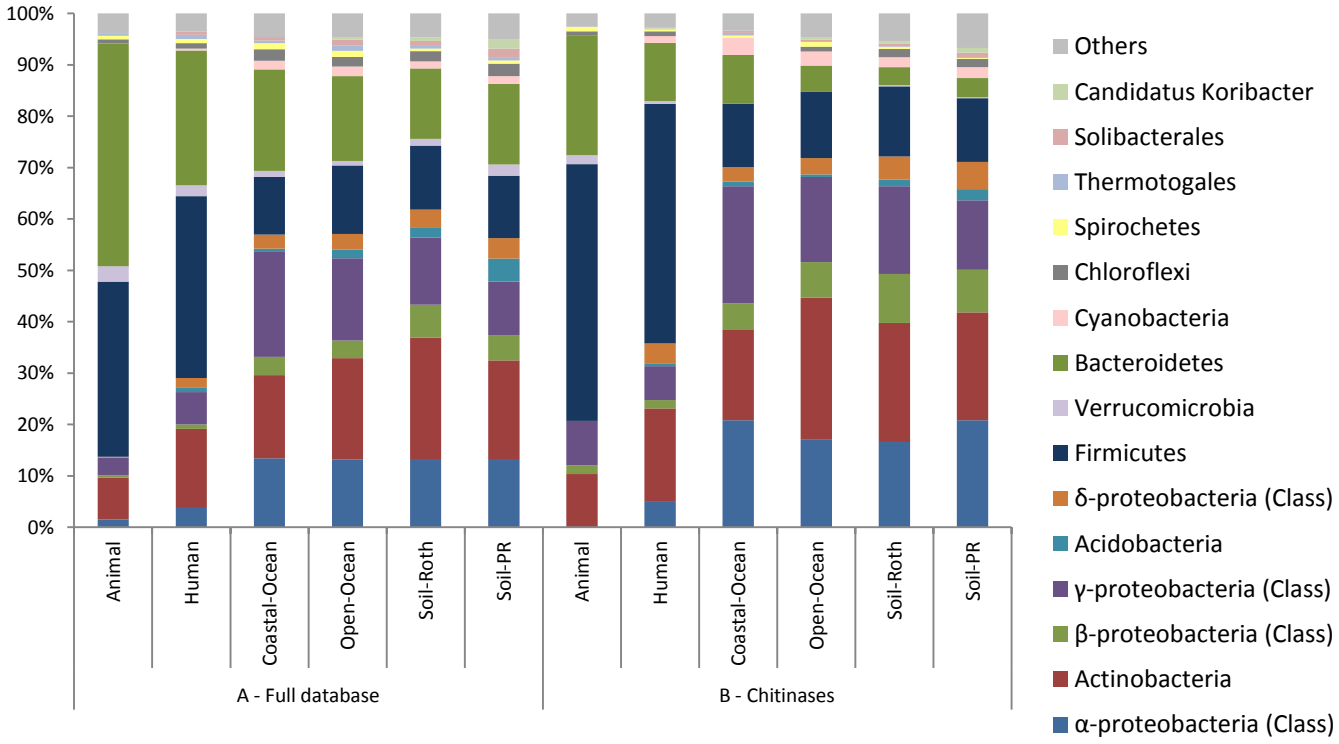


Fig.7: Taxonomical affiliation of the metagenomic sequences against the chitin degrading enzymes database according to BLASTx (E-value cutoff: E-07). The graphic shows in panel A the annotation against the full database encompassing all the enzymatic families used in this study, while panel B displays annotations against chitinases GH18, GH19, CBM33 and CE4 only.

Conclusion

In this section, we tried to investigate the chitin degrading potential of the Rothamsted park grass soil in the frame of METAEXPLORE. A bioinformatic strategy was designed in order to investigate the genetic potential of chitin degradation in several environments, based on their metagenomic resource. The strategy relies on the comparison of different habitat including soils, oceans, human and animal feces, based the genetic patterns found by BLASTx in metagenomic datasets generated by the pyrosequencing technology. In parallel, enzymatic assays were performed to verify the presence of the chitinases activities in the soil of Rothamsted. The enzymatic assays that were performed on fresh wet soil samples revealed the presence of excreted chitinases in the aqueous fraction of the samples. Both endochitinase and exochitinase activities were detected and the comparison with another soil did show that the soil from Rothamsted displays a regular activity level, which was found to be ranging in the same order of magnitude than the soil from The Netherlands.

The data generated in the frame of the project METASOIL were used to screen for relevant genetic determinant involved in chitin degradation. The BLASTx approach has been used successfully to cross-compare the abundance and the diversity of the genetic determinants present in the Park grass metagenomes with other dataset from diverse habitats. Our results clearly shows the preponderance of chitin degrading enzymes in all the environment screened, which is coherent when considering the huge ecological role of chitin in ecosystems (Leclerc *et al.* 2007). However, the abundance of chitinases related sequences is clearly higher in open oceans and soil metagenomes. The coastal ocean samples are showing significantly lower abundance than open oceans. This is probably due to the presence of other source of carbon and nitrogen from coastal water, thus explaining the lower representation of chitin degrading enzymes. The soil of Puerto Rico, which is used as a control, shows the same level of abundance than Park grass.

When looking at the diversity, it is clear that the sequence volume is directly impacting on the detection level in our analysis. However, sequence recruitment tends to indicate that habitats play a crucial role as well. When considering the same sequence volume recruited among the entire environment used in this study, the diversity so-observed is still higher in soil metagenomes and human feces. Moreover, when considering only chitinases, the soil samples stand alone with significantly more diversity observed than the others. The statistical analysis of the BLASTx annotation patterns indicates a clear segregation between human/animal metagenomes and ocean/soil metagenomes. The enzymes families involved in early steps of chitin degradation are retrieved in oceans and soil habitats, while the downstream steps are more present in human/animals digestive tract. This segregation is also observed at the taxonomical affiliation of metagenomic reads, which clearly indicate the prevalence of Firmicutes and Bacteroidetes in human/animal metagenomes, while Proteobacteria and Actinobacteria are found in higher abundance in soil/ocean metagenomes.

At the end, for a given volume of metagenomic sequences, the soil stands out from others environments in terms of sequences diversity affiliated to chitin degradation. It also appears that the soil possess higher proportion of sequences affiliated to the enzymes families involved in the early steps of chitin depolymerization such as chitinases, chitin deacetylases and chitin binding modules. However, this diversity remains relative, as it only reflects what is actually known in the databases for now.

Therefore, this observation is qualifying the aim of our strategy, as the final goal of the project is to find new enzymes, with the underlying will to get away from what is already known. In addition, the panel of environment selected in this study has been driven by what was available in public databases, and is certainly not representative of the variety of habitats potentially eligible as candidates for the search of new chitinases. For instance, marine sediments are known to be hosting microbial communities specialized in chitin degradation (Bhuiyan *et al.* 2011).

Nevertheless, the evaluation of environments based on their metagenomic resource remains relevant to cross-compare environments (Delmont *et al.* 2011). Metagenomic data is a really promising source of novelties, which is clearly highlighted by the fact that a non negligible part of the sequences are not matching anything in reference databases. For instance, the soil of Rothamsted is harboring ≈10% of reads per pyrosequencing run with no affiliations. Furthermore, annotation is often performed with low stringency (e.g BLAST e-value threshold: 1.E-05), which includes lots of sequences that shares only few similarities with the closest matches in databases. Obviously, the risk to obtain false positive with such a lack of stringency is high, but according to what we got from the screening of the clone library (section 2.3), the genes recovered are displaying limited similarities with reference sequences. As a consequence and in spite of all the biased identified along the process, soil remains a really good candidate for carrying the search of novel chitinases. The choice of the soil to be used is still a crucial point as well, but according to our results, the differences observed between the soil of Rothamsted and the control from Puerto Rico are minor in the light of the ones identified with other environments. In addition, their enzymatic patterns were found to be really similar. The main interest of using the soil of Rothamsted rely on the fact that extensive knowledge has been gathered and made available on this environment, which is a key point in its investigation.

This study clearly shows that soil is a good candidate habitat for further chitinase screening. The desired activities are present in the fresh samples, and the deep investigation of the sequence database generated from METASOIL allowed the detection of relatively abundant and diverse genetic signatures related to chitin degradation.

References

Akrem A, Iqbal S, Buck F, Meyer A, Perbandt M, Voelter W, Betzel C. (2011) Isolation, purification, crystallization and preliminary crystallographic studies of a chitinase from Crocus vernus. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 67:340-343

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 215: 403–410

Bertrand H, Poly F, Van VT, Lombard N, Nalin R, Vogel TM, Simonet P (2005) High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J Microbiol Methods*. 62:1-11

Bhuiyan FA, Nagata S, Ohnishi K (2011) Novel chitinase genes from metagenomic DNA prepared from marine sediments in southwest Japan. *Pak J Biol Sci*. 14:204-211.

Bloom SA (1981) Similarity indices in community studies: Potential Pitfalls. *Marine Ecology - Progress Series* 5: 125-128

Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, Coutinho PM, Henrissat B, Nelson KE, White BA. (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A*.106:1948-1953

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. 37(Database issue):D233-238

Delmont TO, Prestat E, Kevin P Keegan, Michael Faubladier, Patrick Robe, Ian M Clark, Eric Pelletier, Penny R Hirsch, Folker Meyer, Jack A Gilbert, Denis Le Paslier, Pascal Simonet and Timothy M Vogel (2012) Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J*. doi: 10.1038/ismej.2011.197

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011a) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol*. 77:1315-1324

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods*. 86:397-400

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. (2008) Functional metagenomic profiling of nine biomes. *Nature*. 452:629-632

FAO. (2006). Guidelines for soil description. FAO, Rome,Italy. ftp://ftp.fao.org/agl/agll/docs/guidel_soil_descr.pdf

Gomes RC, Semedo LT, Soares RM, Linhares LF, Ulhoa CJ, Alviano CS, Coelho RR (2001) Purification of a thermostable endochitinases from Streptomyces RC1071 isolated from a cerrado soil and its antagonism against phytopathogenic fungi. *J Appl Microbiol*. 90:653-661

Gooday GW (1990) The ecology of chitin degradation. *Adv Microb Ecol*. 11:387-430

Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 293: 781-788

Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. *Nature*.457:480-484

Hobel CF, Marteinsson VT, Hreggvidsson GO, Kristjansson JK (2005) Investigation of the microbial ecology of intertidal hot springs by using diversity analysis of 16S rRNA and chitinases genes. *Appl Environ Microbiol* 71:2771-2776

Howard M, Ekborg N, Taylor II L, Weiner R, Hutcheson S (2003) Chitinase B of "Microbulbifer degradans" 2-40 contains two catalytic domains with different chitinolytic activities. *J Bacteriol*. 186:1297-1303

Hjort K, Bergström M, Adesina M, Jansson J, Smalla K, Sjöling S (2009) Chitinases genes revealed and compared in bacterial isolates, DNA extracts and a metagenomic library from a phytopathogen-suppressive soil. *FEMS Microbiol Ecol*. 71: 197-207

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 26:680-682

Hurst MR, Jones SA, Binglin T, Harper LA, Jackson TA, Glare TR (2011) The main virulence determinant of Yersinia entomophaga MH96 is a broad-host-range toxin complex active against insects. *J Bacteriol*. 193:1966-1980

Jerkovic A, Kriegel AM, Bradner JR, Atwell BJ, Roberts TH, Willows RD (2010) Strategic distribution of protective proteins within bran layers of wheat protects the nutrient-rich endosperm. *Plant Physiol*. 152:1459-1470

LeCleir GR, Buchan A, Maurer J, Moran MA, Hollibaugh JT (2007) Comparison of chitinolytic enzymes from an alkaline hypersaline lake and an estuary. *Environ Microb*. 9:197-205

LeCleir GR, Buchan A, Hollibaugh JT (2004) Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment-specific distributions. *Appl Environ Microbiol*. 70: 6977-6983

Lindahl BD, Finlay RD (2006). Activities of chitinolytic enzymes during primary and secondary colonization of wood by basidiomycetous fungi. *New Phytology*. 169: 389-397

Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem J*. 432:437-444

Manucharova NA, Vlasenko AN, Men'ko EV, Zviagintsev DG. (2011) Specificity of the chitinolytic microbial complex of soils incubated at different temperatures. *Mikrobiologiia*. 80:219-229

Metcalfe AC, Krsek M, Gooday GW, Prosser JI, Wellington EM (2002) Molecular analysis of bacterial chitinolytic community in an upland pasture. *Appl Environ Microbiol*. 68: 5042-5050

Muzzarelli RA (1999) Native, Industrial, and fossil chitin. *EXS*. 87:1-6

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server-a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 9:386

Qu A, Brulc JM, Wilson MK, Law BF, Theoret JR, Joens LA, Konkel ME, Angly F, Dinsdale EA, Edwards RA, Nelson KE, White BA. (2009) Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLoS One*. 3:e2945

Ramaiah N, Hill RT, Chun J, Ravel J, Matte MH, Straube WL, Colwell RR (2000) Use of chiA probe for detection of chitinases genes in bacteria from Chesapeake Bay. *FEMS Microbiol Ecol*. 34: 63-71

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. 466:334-338

Souza CP, Almeida BC, Colwell RR, Rivera IN. (2011) The importance of chitin in the marine environment. *Mar Biotechnol*. 13:823-830

Turnbaugh PJ and Gordon JI (2009). The core gut microbiome, energy balance and obesity. *J Physiol*. 587:4153–4158

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. *Nature*. 457:480-484

Uchiyama T, Watanabe K. (2006). Improved inverse PCR scheme for metagenome walking. *Biotechniques*. 41: 183-188

Williamson N, Brian P, Wellington EM. (2000). Molecular detection of bacterial and streptomycete chitinases in the environment. *Antonie van Leeuwenhoek*. 78: 315-321

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: Read Length Matters. *Appl Environ Microbiol*. 74: 1453–1463

Xiao X, Yin X, Lin J, Sun L, You Z, Wang P, Wang F (2005) Chitinase Genes in Lake Sediments of Ardley Island, Antartica. *Appl Environ Microbiol*.71:7904-7909

Rothamsted soil dominant hits (n=7 metagenomes)

NCBI ID	Related species	Related Protein	CAZy Family	Similarity (%)	E-value	Abundance (%)
ACC40460	*Mycobacterium marinum M*	PE-PGRS family protein, GH18 chiA	GH18	64,67	1,00E-59	3,12
ADU46716	*Intrasporangium calvum DSM 43043*	Polysaccharide deacetylase	CE4	52,38	7,00E-37	2,62
ABK53230	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	74,11	3,00E-46	1,67
ABK53232	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	65,65	3,00E-47	1,40
ABM14300	*Mycobacterium vanbaalenii PYR-1*	Chitinase, Cellulase	GH18	70,69	5,00E-80	1,25
ACC42687	Mycobacterium marinum M	PE-PGRS family protein, GH18 chiA	GH18	55,76	4,00E-46	1,23
CAJ68298	*Clostridium difficile 630*	Putative peroxiredoxin/chitinase	GH18	47,09	8,00E-42	1,05
AAL65184	*uncultured bacterium*	Chitinase	GH18	43,4	1,00E-11	0,96
ABP45478	*Mycobacterium gilvum PYR-GCK*	Chitinase	GH18	64,6	5,00E-61	0,86
BAF02587	*uncultured bacterium*	Chitinase	GH18	51,05	9,00E-35	0,81

Puerto Rico soil, dominant hits (n=1 metagenome)

NCBI ID	Related species	Related Protein	CAZy Family	Similarity (%)	E-value	Abundance (%)
ADU46716	*Intrasporangium calvum DSM 43043*	Polysaccharide deacetylase	CE4	36,51	4,00E-11	1,87
CAJ68298	*Clostridium difficile 630*	Putative peroxiredoxin/chitinase	GH18	48,98	4,00E-38	1,84
CAQ48044	*Stenotrophomonas maltophilia K279a*	Putative polysaccharide deacetylase	CE4	61,79	3,00E-36	1,11
AAL65184	*uncultured bacterium*	Chitinase	GH18	42,86	1,00E-10	0,97
ACC40460	*Mycobacterium marinum M*	PE-PGRS family protein	GH18	37,19	3,00E-19	0,83
ABJ85391	*Candidatus Solibacter usitatus Ellin6076*	Polysaccharide deacetylase	CE4	75,94	5,00E-67	0,82
ABK53230	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	45,78	7,00E-15	0,80
ABK53232	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	36,08	2,00E-15	0,75
ABM14300	*Mycobacterium vanbaalenii PYR-1*	Chitinase, Cellulase	GH18	50,92	5,00E-39	0,70
ABY40380	*Pantoea dispersa*	Chitinase I	GH18	57,28	9,00E-30	0,70

Coastal ocean, dominant hits (n=3 metagenomes)

NCBI ID	Related species	Related Protein	CAZy Family	Similarity (%)	E-value	Abundance (%)
ACC42687	*Mycobacterium marinum M*	PE-PGRS family protein, GH18 chiA	GH18	65,62	7,00E-19	4,00
ABY40380	*Pantoea dispersa*	Chitinase I	GH18	65,62	7,00E-19	3,35
ACC40460	*Mycobacterium marinum M*	PE-PGRS family protein, GH18 chiA	GH18	48,11	1,00E-17	2,47
AAZ21201	*Candidatus Pelagibacter ubique HTCC1062*	Polysaccharide deacetylase	CE4	67,01	5,00E-35	1,68
AAV94177	*Ruegeria pomeroyi DSS-3*	Polysaccharide deacetylase	CE4	81,93	3,00E-37	1,58
ADU28907	*Bacillus cellulosilyticus DSM 2522*	Chitin-binding domain 3 protein	CBM33	46,34	2,00E-18	1,44
ABK53230	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	66,28	2,00E-29	1,35
ACV23495	*Slackia heliotrinireducens DSM 20476*	Predicted xylanase/chitin deacetylase	CE4	49,44	3,00E-19	1,35
ABK53805	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	73,21	1,00E-20	1,21
ACR76120	*Eubacterium rectale ATCC 33656*	Peptidoglycan N-acetylglucosamine deacetylase	CE4	53,75	1,00E-19	0,98

Open ocean, dominant hits (n=3 metagenomes)

NCBI ID	Related species	Related Protein	CAZy Family	Similarity (%)	E-value	Abundance (%)
ABK53805	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	92,19	4,00E-31	6,73
ABK53230	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	93,33	1,00E-36	6,71
ABD80964	*Saccharophagus degradans 2-40*	Chitinase	GH18	47,86	6,00E-17	4,03
ACC40460	*Mycobacterium marinum M*	PE-PGRS family protein, GH18 chiA	GH18	56,48	8,00E-28	3,26
ABB76136	*Francisella novicida*	ChiA	GH18	43,59	4,00E-18	3,24
ABK53232	*Acidothermus cellulolyticus 11B*	Chitinase family 18	GH18	82,98	6,00E-17	2,24
ADU21434	*Ruminococcus albus 7*	Endo-1,4-beta-xylanase	CE4	80,36	2,00E-22	1,54
ABI86986	*Burkholderia ambifaria AMMD*	Chitinase family 18	GH18	68,29	5,00E-13	1,51
ACR14100	*Teredinibacter turnerae T7901*	CBM 33 and 10 domain protein	CBM33	91,94	4,00E-21	0,99
ABC31487	*Hahella chejuensis KCTC 2396*	Chitinase	GH18	62,9	5,00E-21	0,84

Dominant hits in the Rothamsted soil metagenomes after BLASTx (E-value cutoff: E-07). The relative abundance is calculated as a percentage of the total hits detected in the BLASTx.

Chapter 2 - Part. 3

The Rothamsted Park grass soil: Screening for activities and related genes - Clone library screening and sequencing: A sequence based strategy

SAMUEL JACQUIOD[1], LAURE FRANQUEVILLE[1], SANDRINE DEMANECHE[1], TOM O. DELMONT, TIMOTHY VOGEL[1], PASCAL SIMONET[1]

[1] Ecole Centrale de Lyon, Laboratoire Ampère, Environmental Genomics Group, 36 avenue Guy de Collongue - 69134 Ecully, France

Introduction

After the sequence-based screening operated on the datasets generated from the soil of the Park grass experiment (Rothamstd, UK), a fraction of the clone library was used for an advanced molecular screening. The fosmid clone library established during the on-going of the METASOIL project harbors a total of 2 million clones, whose complete screening represent a huge challenge by itself, as it requires extensive resources both in terms of funding and facilities. Here we present the molecular-based screening applied onto a part of this clone library, to seek for genetic determinants of interest in the frame of METAEXPLORE. The technique relies on the use of complementary radio-labeled probes designed to target specific genetic signatures within the metagenomic inserts carried on the fosmids. To do so, the clones and their respective fosmids are massively spotted on high density nylon membranes and set for hybridization purposes with the labeled probes. In this section, we present and describe the technique in its molecular designing and functioning. We report the first results obtained from this strategy as well, with a specific emphasis on chitinase, as it represents the major part of my involvement as a PhD in this section.

1. Presentation of the sequence-based screening

Screening on high density nylon membrane is a molecular strategy that was already successfully applied to detect desired genes out of a clone library (Demaneche *et al.* 2009*;* Ginolhac *et al.* 2004). This approach consists in using radio-labeled probes that target genetic sequences of interest through a hybridization process. The probe shall be designed carefully in order to ensure the detection of the desired genetic determinant. However, its design is entirely dependent on the goals to achieve. For instance, the probe sequence shall be extremely accurate if the purpose is to investigate the quantitative aspect related to one specific gene in a clone library, for example. On the other hand, if the purpose is to investigate qualitative aspects of a given gene, such as its diversity in terms of sequence divergence, the probe shall be designed based on a consensus or conserved region in order to broadly screen this genetic resource. This last strategy is the one retained in the frame of METAEXPLORE, as the project is entirely devoted to seek and indentify novelties, especially in terms of novel genetic sequences that diverge from the one known in public databases, and related to the synthesis of desired enzymatic activities. Thus, the probe strategy is perfectly fitting on these perspectives, as it can allow some flexibility by playing on specific aspects such as the probes construction and synthesis, and also the hybridization conditions.

The strategy selected for probe synthesis consisted in PCR amplicons, amplified with specific primers targeting the genes of interest, and from the metagenomic DNA extracted from the soil of Park grass (Rothamsted, UK). Plasmid clone libraries were constructed from the amplicons, and sequenced in order to select the best candidate probes before starting the labeling process. In this experiment, several genes were targeted: (i) transgenic markers (4 probes), (ii) chitinases GH18 chiA (3 probes), (iii) dehalogenases (3 probes), (iv) bacterial laccases (2 probes) and (v) mobile genetic elements (4 probes). Preliminary labeling and hybridization attempts were carried out with non-radioactive chemicals in order to test probes self-hybridization and detect eventual parasite hybridization between probes and fosmid-genome sequences. Then the radio-labeling and the actual hybridization has been performed on 22 nylon membranes harboring a total of 405 504 clones, which represent 20.28% of the library.

2. Material and methods

i. Clone library construction

The bacterial fraction from the soil of Park grass experiment (Rothamsted, UK) was extracted by differential centrifugation in a Nycodenz® gradient. Metagenomic DNA has been recovered thanks to the agarose plug lysis technique (Bertrand et al. 2005). DNA extraction yield has been evaluated around 120µg/kg-1 of fresh soil, and the fragments size was estimated to be >500 kbp based on electrophoresis assays. Metagenomic DNA has been mechanically sheared into approximately 40kb fragments on average, and cloned in the pEpiFOS-5 single-copy fosmid, and transformed in the E. coli strain EPI100 (EpiFOS™ Fosmid Library Production Kit, Epicentre, Illumina). A total of 2 million clones has been constructed, duplicated and stored at -80°C in two different laboratories (Ampère, Ecole Centrale de Lyon, Ecully, France and Libragen SA, Toulouse, France). Each clone is harboring on average 40kb of metagenomic DNA inserts, which correspond to an estimated total of 80 Mbp.

ii. Nylon membrane settings

One part of these clones has been spotted on 22 GeneScreen Plus (NEF988) nylon membranes previously laid onto Luria-Bertani agar plates with 12.5µg/ml chloramphenicol, and then incubated at 37°C for 18 h. 405 504 transformants were spotted (18432 duplicated clones per membrane), corresponding to 20.28% of the total library. Each clone has been spotted in duplicate side to side in order to reinforce and validate the detection of positive signals after hybridization. The clones were grown on the membranes up to a million bacterial cells per colonies, theoretically corresponding to 1~2 million copies of the target genes (1~2 fosmids per cells). Colonies were lysed by incubating the membranes for 15 min on a sheet of 3M paper (Whatman) saturated with 0.5 M NaOH-1.5 M NaCl. The membranes were then neutralized by incubation for 15 min on a sheet of 3M paper (Whatman) saturated with 1.5 M NaCl-1 M Tris (pH 7.5). After drying at room temperature for 20 min, immobilization of DNA on membranes was performed by the UV cross-linking technique (312 nm for 4 min) (Ginolhac et al. 2004). The genetic information contained into the metagenomic inserts was exploited to seek for desired genetic signatures.

iii. chiA probes design

All probes were designed according to the following criterions: (1) size shall be comprised between 200~1000bp, (2) melting point (Tm) shall be ranging around 70°C +/-10°C according the formula:

$$Tm = 49.82 + 0.41a - 600/b$$

"a" represents the GC content of the sequence in percentage, and "b" the sequence length in base pairs. The optimal size and temperature for efficient hybridization were previously established to be 500bp and Tm-20°C. Partial sequence from the gene chiA has been amplified by PCR from the Rothamsted Park Grass metagenomic DNA using the primer sequence 5'-ACG GCG TGG ACA TCG ACT GGG AGT-3'(foward) and 5'- CCC AGG CGC CGT AGA RRT CRT AG-3' (reverse) (Hobel *et al.* 2005). 2µl of metagenomic DNA (≈50ng) was mixed with 1.25µl of reverse and forward primers (10µM) and 20.5 µl of distilled sterile water. 1µl of the Phusion High Fidelity DNA polymerase was used at the concentration of 1U/µl in order to generate sticky end amplicons.

PCR was performed in a Biometra thermocycler, and cycles consisted in 95°C for 10 min and then 30 cycles of 95°C for 30s, annealing at 55°C for 30s, and elongation at 72°C for 25s, followed by 72°C for 15 min. An electrophoresis step was carried out onto the PCR amplified mixture on an agarose gel (1%), and the desired band corresponding to the expected size (≈300pb) has been cut out, and purified on a silica column (GFX PCR DNA gel band purification kit, Amersham Biosciences). The purified chiA sequences were cloned in a plasmid (TOPO® TA Cloning®, Invitrogen) and transformed in *E. coli* strain (TOP10F´®, Invitrogen). 45 clones were sent to sequencing using M13 standard primers (GATC Technologies). Sequences generated were analyzed by BLASTn against *nr* database and alignments were performed by *Bioedit* software (Hall *et al.* 1999), and trees were constructed with *Seaview* software (Gouy *et al.* 2010). Among the 50 sequences retrieved, 3 were selected as probes for further hybridization experiments. For more details about chiA probes design, please report to Appendix 2-3a at the end of chapter this section.

iv. Preliminary biotin labeling and hybridization attempts

All selected sequences were cloned in the single-copy fosmid vector pCC1fos (CopyControl™ cDNA, Gene & PCR Cloning Kit, Epicentre©, Clontech), which can be retro-induced to multiple copies in the E. coli strain EPI300. Sequence alignments were performed between probes, fosmids and genomic DNA to ensure that no parasite hybridization will occur during the process. In the first move, all sequences were labeled with biotin separately (Biotin DecaLabel DNA Labeling kit, Fermentas, 100ng probe/reaction), and a pool of all sequences (50ng each) has been labeled as well as a control, according to the manufacturer instructions. The labeling efficiency has been controlled on a test membrane, by spotting different concentration of each probes and the pool against a control provided by the manufacturer. The labeling efficiency was evaluated to be 100-folds weaker than theoretical expectations.

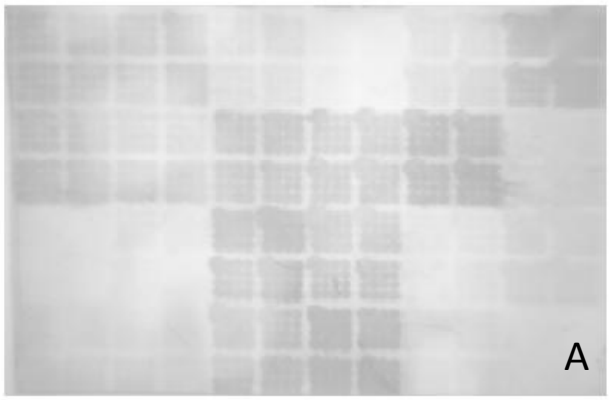
Probe name	Targeted gene	Category	Accession	Size (bp)	GC (%)	Tm (°C)	Primer reference, sequence origin
aadA	aadA	Transgenic marker	CU459141.1	382	54	70.4	Ceccherini *et al.* (2003)
accD	accD		Z00044.2	516	30	51.0	Monier *et al.* (2007)
p24	-		-	556	33	62.3	-
rbcL	rbcL		X04976.1	455	44	66.5	Monier *et al.* (2007)
chiB3	chiA	Chitinase GH18	EU700201.1	303	60.7	72.7	Hobel *et al.* (2005)
chiC9			CP000360.1	246	58.9	71.5	
chiF8			EU700206.1	293	54.6	70.2	
dhaA	dhaA	Dehalogenase	AF060871	891	59.5	73.5	Demanèche *et al.* (2009)
linA	linA		-	696	52.3	70.4	Zhou *et al.* (2008)
linB	linB		AY903216.1	893	62	74.6	Demanèche *et al.* (2009)
FS2-39	-	Laccase	-	580	58.4	72.7	Ausec *et al.* (2011)
FS45	-		-	589	62.1	74.3	
IS1071	Insertion sequence	Mobile Genetic Element	-	179	59.2	70.7	Dunon *et al.* (2010, unpublished)
IA15	Integron		-	809	56.4	72.2	Delmont *et al.* (2010, unpublished)
IA20			-	809	58	72.9	
IA33			-	809	46	67.9	

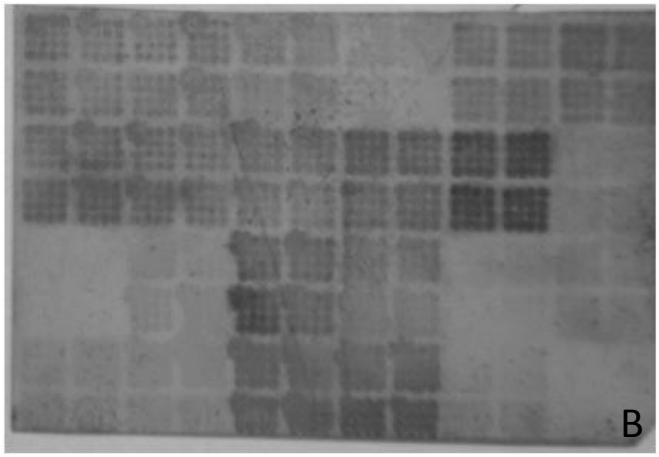
Tab.1: Description of the final set of probes selected for the experiment. Characteristics such as size (bp), GC content, melting point and related references are given.

However, another biotin labeling kit (Mirus) did yield better results, with 10-folds higher signal, but unfortunately with a higher cost and leaser sensitivity. Finally, all experiments were carried out with the biotin labeling kit from Fermentas. Sequences cloned in pCC1 were spotted on a test membrane at different concentrations for self-hybridization attempts against the corresponding probes. The sensitivity threshold for detection is 10ng biotin-labeled DNA. Non specific hybridization was observed on the fosmid at 20ng when exposed to negative control. Same goes for the probe pool, generating a slight non-specific signal. However, the signal observed on probes after self-hybridization was significantly higher than non-specific processes, thus resulting in an unambiguous identification of the positives signals.

Probes Specificity was tested with the same strategy. Sequences cloned in pCC1 were spotted on a test membrane at different concentration (100ng, 10ng and 1ng), with negative control as well (empty fosmid). The pool of labeled sequences has been applied at different dilution in order to test the sensitivity of each probe against its own cloned sequence (1/100, 1/1000, 1/10000). No signal was observed at 1/10000, and the signal intensity was similar to the negative control at 1/100. However, the dilution 1/100 did yield good hybridization results, with weaker signal for the shortest probes. Cross hybridization tests were performed on the grown colonies harboring the fosmid with targeted sequences. Each colony has been cultivated on specific membrane, and the DNA was attached as described above. Each membrane corresponding respectively to one type of sequence has been submitted to hybridization attempts against each probe separately and with the labeled pool in order to detect possible cross-hybridization. No cross-hybridization has been observed, even though some similarities exist between the chiA probes, and also between integron ones. However, the two shortest probes were not detected again (Tab.1, chiF8 and IS1071).

Finally, an attempt has been realized on 24 high density test membranes prepared by Libragen, including 16 positive control membranes (one for each targeted sequence), 2 negative controls corresponding to pEpiFos and pCC1 vectors, 4 unknown clones from the fosmid library (marked as j13, m20, b5, and g10), and 2 empty membranes (Fig.1). All membranes were exposed to hybridization with the diluted probes pool (1/1000). Again, low signal has been observed for the shortest probes (chiB3, chiC9, chiF8 and IS1071), which can barely be differentiate from the negative controls and membrane with unknown fosmids. However, a second hybridization performed on the same membrane only with the shortest probes (1/1000) resulted in better signal detection for chiB3 (303bp) and chiF8 (293bp), but not for the others (Fig.1). No false positives were observed, and the background noise was weak.

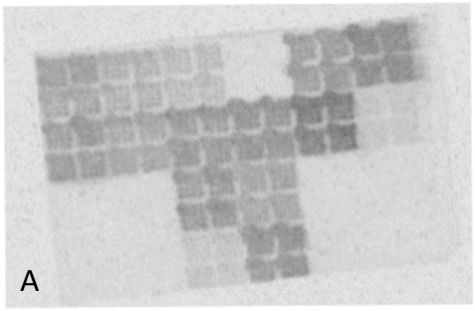




	1 2	3 4	5 6	7 8	9 10	11 12
1 2	aadA	accD	chiB3	chiC9	chiF8	dhaA
3 4	FS2-39	FS-45	IA15	IA20	IA33	IS1071
5 6	empty	Fos	linA	linB	j13	m20
7 8	b5	g10	p24	rbcL	pCC1	empty

C

Fig.1: Hybridization of biotin labeled probes against their own respective sequences spotted on 24 test membranes. Panel A displays hybridization after incubation with the diluted pool of probes (1/1000), while panel B represents a second hybridization of the same membrane at higher probe concentrations for the shortest ones. Table in panel C represents the pattern of the membranes setting (6x4). Each membrane is constituted of 4 blocks of 16 repetitions (dots). Probe names are given in Tab.1. Membranes marked as empty were not spotted, and the ones marked with "Fos" and "pCC1" were spotted with vectors as negative controls. Unknown clones from the fosmid library were spotted on the membrane marked as "j13", "m20", "b5" and "g10".



	1 2	3 4	5 6	7 8	9 10	11 12
1 2	aadA	accD	chiB3	chiC9	chiF8	dhaA
3 4	FS2-39	FS-45	IA15	IA20	IA33	IS1071
5 6	empty	Fos	linA	linB	j13	m20
7 8	b5	g10	p24	rbcL	pCC1	empty

Fig.2: Hybridization of ^{33}P-labeled probes against their own respective sequences spotted on 24 test membranes. Panel A displays the hybridization capture on the photo-sensitive screen, and the table on the right side represents the pattern of the membranes setting. Each membrane is constituted of 4 blocks of 16 repetitions (dots). Probe names are given in Tab.1. Membranes marked as empty were not spotted, and the ones marked with "Fos" and "pCC1" were spotted with vectors as negative controls. Unknown clones from the fosmid library were spotted on the membrane marked as "j13", "m20", "b5" and "g10".

v. ^{33}P labeling and hybridization

All probes were labeled with the DecaLabel DNA Labeling Kit (Fermentas) and 50µCi of (α-^{33}P)-dCTP (Perkin-Elmer) by following the manufacturer instructions. 100ng of PCR amplicons were used for the shortest probes (chiB3, chiC9, chiF8 and IS1051), and 50ng for the others. Unincorporated labels were removed by chromatography on Sephadex® G-50 columns (Amersham Biosciences, Piscataway, New Jersey, USA). Hybridization were carried out overnight at 42°C in a buffer containing 0.5% sodium dodecyl sulfate-5× Denhardt's-6x SSC, 50% deionized formamide and with the denatured (α-^{33}P)-dCTP labeled probes pool. Membranes were washed sequentially in (i) 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate), 0.1% SDS at 20°C for 10 min, (ii) 0.1× SSC-0.1% sodium dodecyl sulfate at 65°C for 20 min. A photo-sensitive screen was used for membrane exposure and visualization of hybridization spots during 90 min. The screen was scanned with Typhoon Trio scanner (GE Healthcare).

A test membrane similar to the one above has been used for the first attempts with the radio-labeling before starting the screening of actual high density membranes. The test membrane revealed no signal for the probe chiC9, as well as very weak intensity for p24 and IS1071 (Fig.2). However, no false positives were observed, and very low background signal was observed, as previously noticed with biotin labeling.

vi. Positive clones' analysis

The clones that were identified as putative positives after hybridization with the radio-labeled probe pool were retrieved from the stored clone library, and grown overnight at 37°C in 96 multi-well plate with Luria-Bertani liquid media under the fosmid selective pressure (12.5µg/ml chloramphenicol). The plate has been duplicated and one copy stored at -80°C, and the remaining one used for fosmid DNA extraction. 10µg of fosmid DNA corresponding to the clones pool were sent to 454 pyrosequencing (GENOSCOPE, France), and the generated reads were assembled using *Newbler* software (Roche). Assembled contigs were used for sequence screening using BLASTn and BLASTx with the former probe sequences and private databases. Full annotation was carried out through the RAST server using SEED subsystems (Aziz et al, 2008).

3. Results

i. Assembly results and BLASTn verification

Among the 22 membranes used in this experiment, 88 putative positive clones were identified after hybridization with the pool of radio-labeled probes. This corresponds to a yield of 0.00022% (88/405 504). From the sequences generated after 454 pyrosequencing, 149 contigs were obtained with an average size of 22.8kb (44.6kb maximum and 2,2kb minimum). 70 contigs are longer than 30kb in size, which correspond to the expected size for a metagenomic insert (≈40kb). As a consequence, the full assembly of some metagenomic inserts among the pool is certainly probable.

The first sequence analysis of the contig pool was done through BLASTn with all the probe sequences used in this experiment. However, this approach was not efficient enough, as no relevant hits were found probably due to a high degree of divergence between the probes and the targets, and also to the possible selection of false positive after membrane visualization. For instance, only the mobile genetic element probes revealed good matches through BLASTn, and only short conserved matches of 15-30bp for all the remaining probes. As a consequence, the contigs were submitted to the RAST server in order to find the coding sequences, and to visualize the genetic context of relevant hits in their respective contigs. Contigs were also screened through BLASTx with the chitinase database (see chapter 2.1) in order to find genetic signature related to chitinases.

ii. Annotation through the RAST server

After SEED annotation of the coding DNA sequences (CDS) found in the 6 reading frames through the RAST server, 3505 coding regions were identified out of the 149 contigs. Among the detected CDS, the great majority is annotated as unidentified protein (2964/3505, 84.56%), as it could not match with anything known from the SEED subsystems classifier. In these unclassified proteins, 65.96% (1955/2964) are referenced as "hypothetical" structures, while 34.04% (1009/2964) are considered as "non hypothetical" protein structures.

On the other hand, only a little part of the CDS (541/3505, 15.44%) has been found to be related to something known from the databases (Fig.3). The major part of the CDS detected is related to fundamental functions such as amino acids and derivatives metabolism, carbohydrate metabolism, respiration, cell wall and capsule, protein and DNA metabolism. Among these CDS, a few are related to "hypothetical" protein structures (22/541, 4.07%), while the others are referenced as "non hypothetical" protein structures.

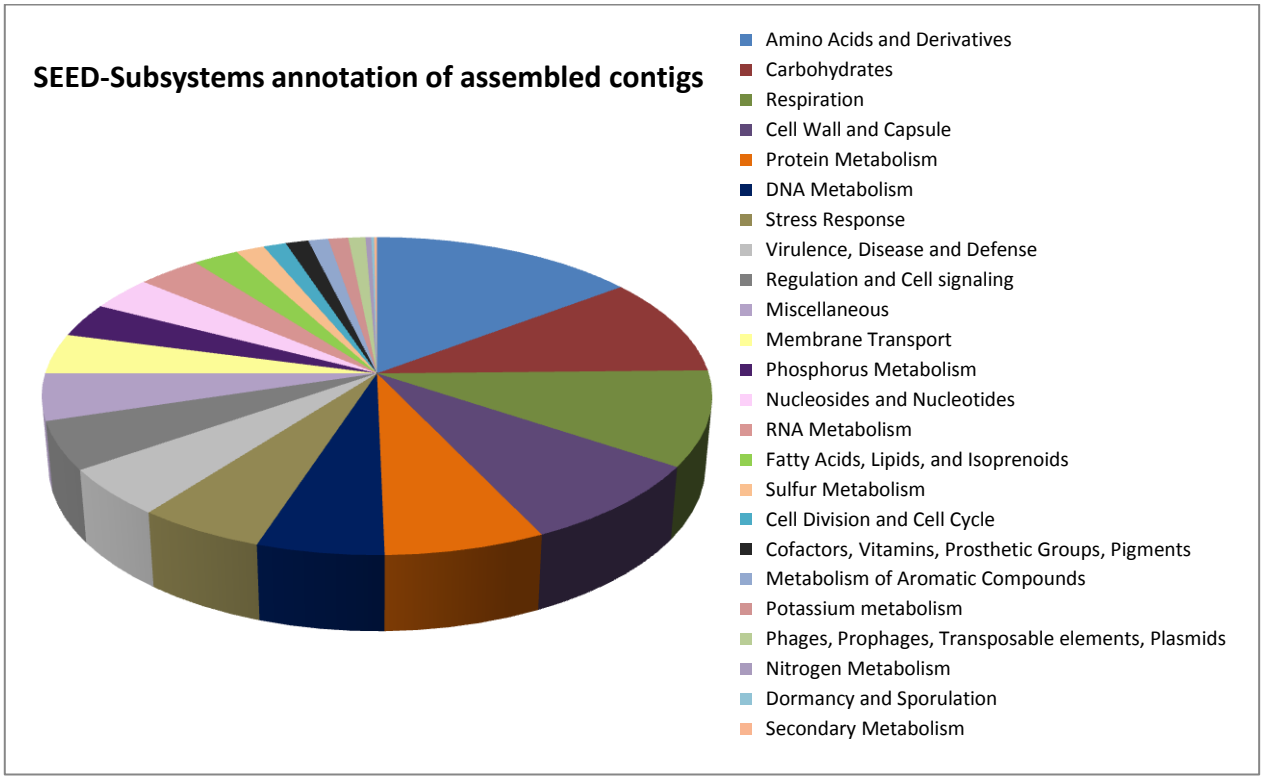


Fig.3: SEED annotation of the contigs through the RAST server. The contigs were submitted to open reading frame (ORF) calling, then the coding DNA sequences (CDS) that were found in the 6 frames were analyzed through BLASTx in order to find the closest known related coding protein in databases. This figure represents only the CDS fraction related to referenced structures in SEED subsystems, which corresponds to 15.44% of the detected CDS (541/3505).

Organism	*Candidatus Solibacter usitatus* Ellin6076
Lineage	Acidobacteria; Solibacteres; Solibacterales; Solibacteraceae; Candidatus Solibacter
Lenght	428 aa
e-value	1.00E-21
Identity	32.57%
Prot name	Acid_5417
GeneBank	ABJ86366.1
UniProt	Q01VE9
Description	Polysaccharide deacetylase from Carbohydrate esterase family 4 (CE4)

Organism	*Natranaerobius thermophilus* JW/NM-WN-LF
Lineage	Firmicutes; Clostridia; Natranaerobiales; Natranaerobiaceae; Natranaerobius
Lenght	320 aa
e-value	5.00E-26
Identity	32.93%
Prot name	Nther_0950
GeneBank	ACB84535.1
UniProt	B2A0H4
Description	Polysaccharide deacetylase from Carbohydrate esterase family 4 (CE4)

Tab.2: Affiliation of the two CDS related to the polysaccharide deacetylase thanks psiBLAST from the RAST server.

When looking closer to the genes of interest, several relevant hits were observed. 51 CDS related to mobile genetic element were retrieved (MGE), encompassing many transposase and insertion sequences, but also 8 CDS related to integrons/integrase. 11 CDS related to beta-lactamase genes were retrieved as well on 11 separated contigs. 5 CDS involved in chitin degradation were also detected in 2 contigs (Fig.4 and 6), and 1 CDS related to a dehalogenase in another one (Fig.7).

No CDS were found for the bacterial laccases, which can be explained by the fact that no SEED subsystem exist for this category of enzymes. The 51 CDS related to mobile genetic elements were found on 26 different contigs, with up to 8 MGE CDS per contigs. In 3 contigs, insertion sequences and transposases were found to be more or less close to beta-lactamase related CDS. This was also observed for the contig carrying the CDS related to the dehalogenase, which also has a transposase. Among the 2 contigs carrying chitin derivative enzymes, 1 is displaying an interesting profile (contig#55, 38.3kb) because it harbors 4 CDS involved in chitin degradation, as well as 2 transposases coupled to many other genes involved in sugar metabolisms and transport (Fig.4). The 4 genes involved in chitin degradation are referenced as followed, from the left to the right side (Fig.4): 2 N-acetylglucosamine-6-phosphate deacetylase, 1 hypothetical oxidoreductase related to N-acetylglucosamine utilization and 1 Beta-hexosaminidase related to glycosyl hydrolase family GH20. The last one is referenced as another Beta-hexosaminidase related to glycosyl hydrolase family GH3, which is located on another contig. However, no chitinase GH18 were found through this approach. The details about the contigs carrying chitin degradation related CDS and the dehalogenase are given in Table 2 and Figures 4, 5, 6, 7, 8 and 9.

In addition, 2 putative CDS related to polysaccharide deacetylases have been identified on the contig#50 and 52 (Fig.6 and 8). They are both affiliated to the Carbohydrate esterase family CE4, which encompasses chitin deacetylase enzymes (Tab.2). The closest match in databases for the first one (contig#50) is related to *Candidatus Solibacter usitatus* Ellin6076 from Acidobacteria, known to be a telluric bacteria adapted to low-nutrient conditions (Ward *et al.* 2009) while the other one (contig#52) is related to a *Natranaerobius thermophilus* from Firmicutes, which is known to be an anaerobic halophilic alkalithermophile bacteria, isolated from a salt lake sediments in Egypt (Zhao *et al.* 2011).

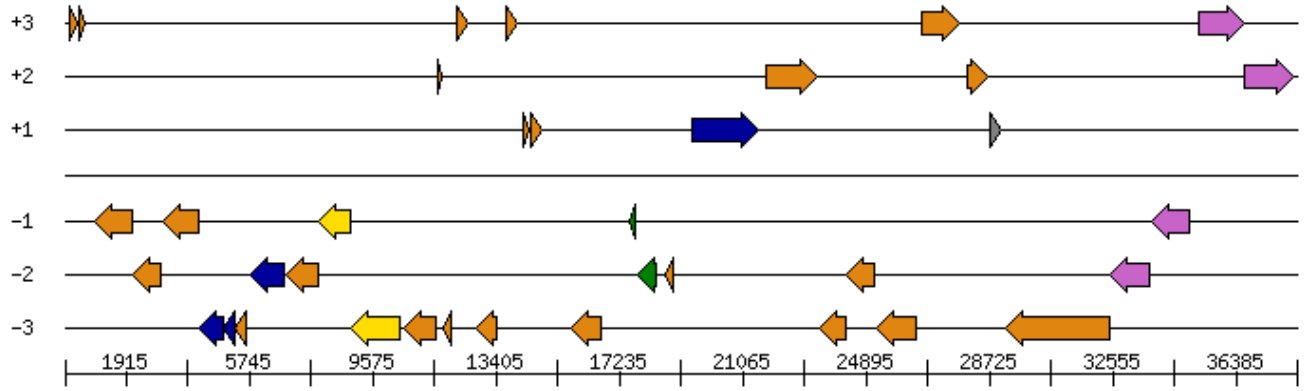


Fig.4: SEED annotation of the CDS detected in contig#55 (38.3kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). The blue arrows are displaying the CDS related to chitin degradation enzymes. The orange and yellow arrows are representing genes involved in ribose and sugar metabolism, as well as other genes involved in carbohydrate metabolism. Green arrows are showing the 2 CDS related to transposase. Purple arrows are displaying CDS involved in DAP (3-diaminopropane) production, and the grey one is representing an endoribonuclease CDS.

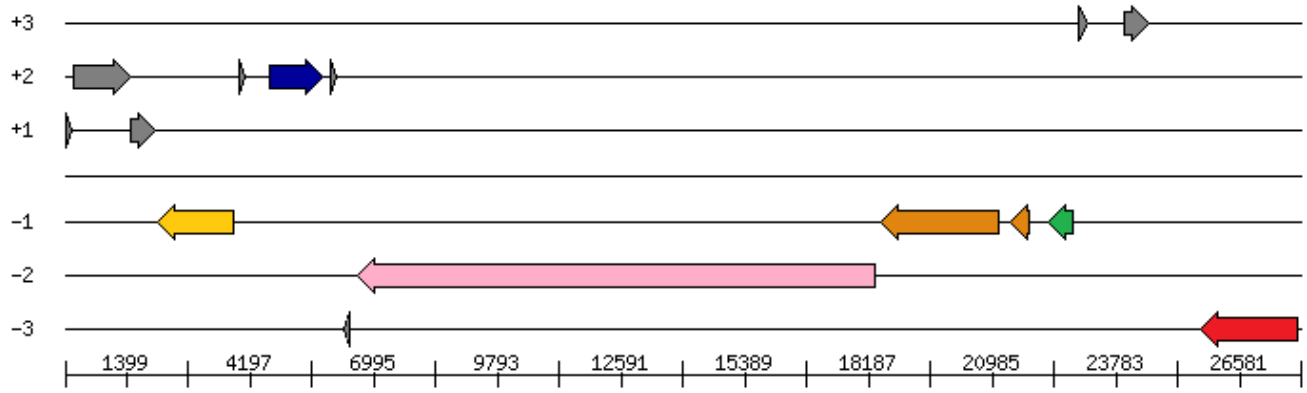


Fig.5: SEED annotation of the CDS detected in contig#92 (29kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). The red arrow displays the CDS related to the putative chitinase detected after confirmation with BLATSx. The yellow arrow represents a putative xylanase from GH43, and the blue one a peptidase. The green arrow shows CDS related to a probable integron group II, and the 2 orange arrows represent probable RNA polymerase and kinase. The pink arrow displays CDS of a TonB-dependant receptor, and the grey ones are representing hypothetical CDS not classified in SEED.

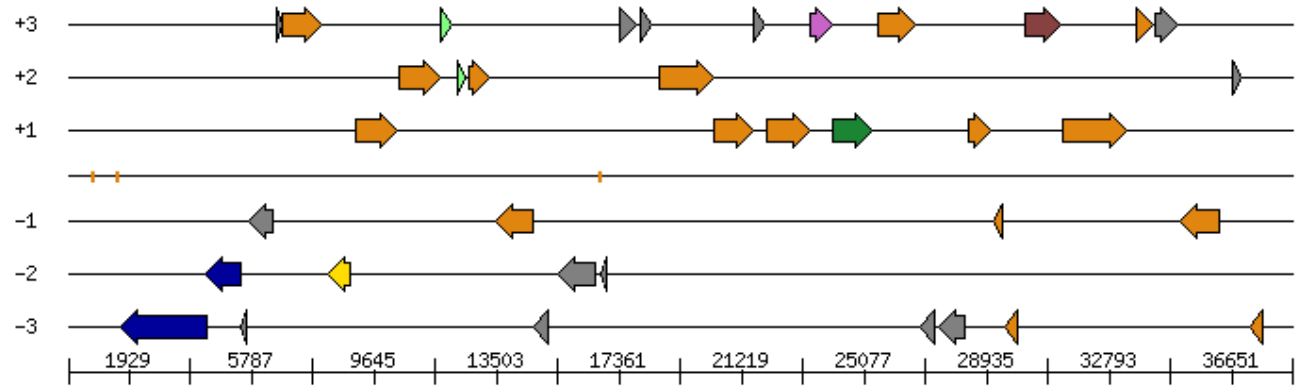


Fig.6: SEED annotation of the CDS detected in contig#52 (38.6kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). Grey arrows are representing hypothetical CDS not classified in SEED. The other genes are annotated on each reading frame from the left on the right as follow: (+3) Permease of the major facilitor superfamilly, SSU ribosomal protein 16S, ATP-dependent clp protease (proteolytic subunit) EC 3.4.21.92, Molybdopterin binding motif, CinA N-terminal domain, Beta-hexosaminidase (E.C 3.2.1.52) and AhpC/Tsa family protein α selenocysteine containing. (+2) Signal recognition particle subunit Ffn SRP54 (TC 3.A.5.1.1), KH domain RNA binding protein YlqC, tRNA (guanine 37-N-1)-methyltransferase (EC 2.1.1.3), Glutamate formiminotransferase (EC 2.1.2.5). (+1) Polysaccharide deacetylase CE4, Imidazolonepropionase, Cell division trigger factor (EC 5.2.1.8), ATP-dependant Clp protease ATP-binding subunit Clpx, FKBP-type peptidyl-prolyl-cis-trans-isomerase fklB (EC 5.2.1.8), oligopeptide transporter. (-1) 3-oxoacyl synthase (EC 2.3.1.41), lojap protein, GTP-binding and nucleasic acid-binding protein YchF. (-2) 5-methyltetrahydrofolate-homocysteine methyltransferase (EC 2.1.1.13), N-acetylmuramoyl-L alanine amidase (EC 3.5.1.20). (-3) 5-methyltetrahydrofolate-homocysteine methyltransferase (EC 2.1.1.13), Nicotinate nucleotide adenyltransferase (EC 2.7.7.18), putative two-component system response regulator.

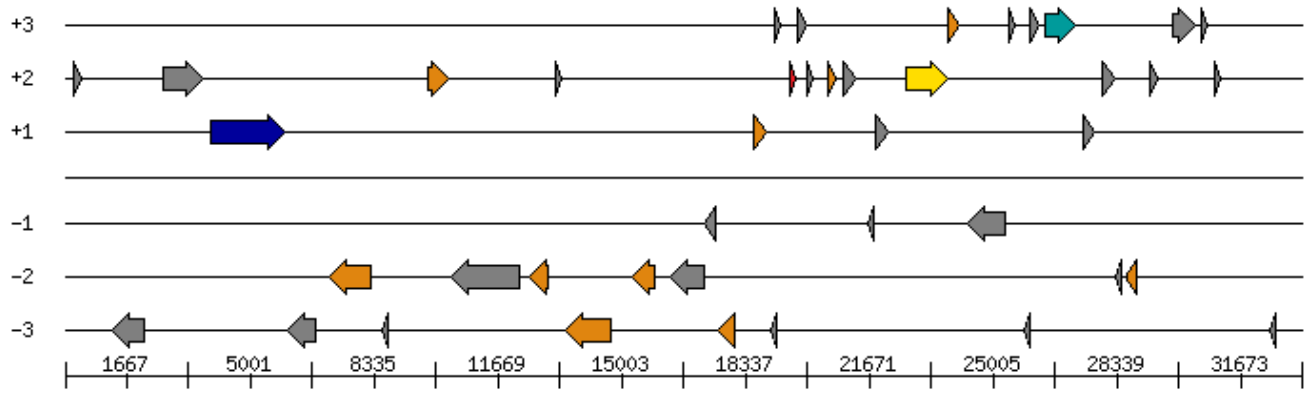


Fig.7: SEED annotation of the CDS detected in contig#72 (33.4kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). Grey arrows are representing hypothetical CDS not classified in SEED. The other genes are annotated on each reading frame from the left on the right as follow: (+3) Short-chain dehydrogenase/reductase SDR, Enoyl-CoA hydratase (EC 4.2.1.17). (+2) Phytoene dehydrogenase and related proteins, Type III restriction-modification enzyme-helicase subunit, Threonine dehydrogenase and related Zn-dependant dehydrogenase. (+1) Multi-sensor hybrid histidine kinase. (-2) Asp-tRNA-Asn/Glu-tRNA-Gln amido transferase A subunit and related amidases, GA16242-PA, 3-oxoacyl reductase (EC 1.1.1.100), Transposase and inactivated derivatives. (-3) DNA topoisomerase IB (poxvirus type) EC 5.99.1.2, , 3-oxoacyl reductase (EC 1.1.1.100).

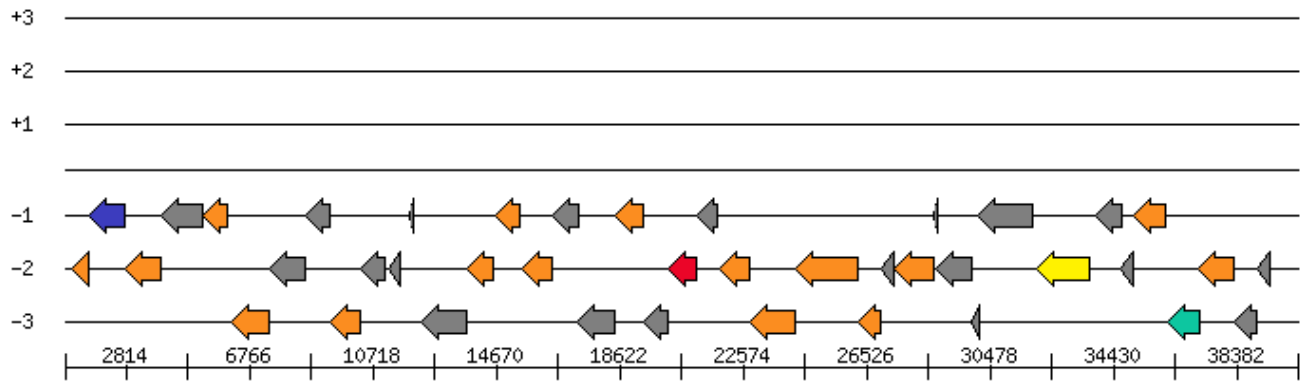


Fig.8: SEED annotation of the CDS detected in contig#50 (39.5kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). Grey arrows are representing hypothetical CDS not classified in SEED. The other genes are annotated on each reading frame from the left on the right as follow: (-1) Glycosyl transferase (blue arrow, COG0451 subsystem), putative methyltransferase, glycosyl transferase fam.2, glycosyl transferase fam.2, Pseudaminic acid biosynthesis protein PseA. (-2) Colanic acid biosynthesis acetyltransferase Wcaf (EC2.3.1.-), Glycosyltransferase fam.1, HAD superfamily hydrolase subfamily IIIA, Heptosyltransferase family protein, polysaccharide deacetylase CE4 (red arrow), N-acetylneuraminate synthase EC2.5.1.56, capsule polysaccharide biosynthesis, alginate biosynthesis protein Algl, Asparagine synthase glutamine-hydrolizing EC6.3.5.4, DegT/DnrJ/Eryc1/STrS aminotransferase family protein. (-3) probable glycosyltransferase, putative glycosyltransferase, sugar transferase, 3-oxoacyl-[acyl-carrier protein reductase] EC1.1.1.100, UDP-glucose-4-epimerase EC5.1.3.2 (green arrow, Rhamnose containing glycans subsystem).

iii. BLASTx against homemade chitin degradation database

As no significant chitinase hits were detected through the RAST annotation, the contig pool has been screened through BLASTx with the database mentioned in section 2.1. This "home-made" database gather all the known enzymes from different families involved in early and late state of chitin degradation. This approach will allow identification of genetic signature related to chitinases at the protein level, but will also help in defining probable relevant CDS among the one classified as hypothetical and unclassified by RAST. Comparison of the two approaches may result in finding putative chitinase out of the 84.56% of unidentified CDS detected (2964/3505).

The BLASTx succeeded in identifying 24 hits related to chitinase GH18 among 15 distinct contigs. Most of these hits were false positives, as they correspond to conserved protein motives among the glycosyl hydrolase and glycosyl transferase families, after confrontation with the RAST annotation. This is mostly due to the relatively high conservation of the protein regions corresponding to lytic sites of glycosyl related enzymes. However, among the 15 hits detected, 3 of them are more relevant than the others (Tab.4). After comparing the results from BLASTx and RAST annotation, a hypothetic CDS that was not classified in subsystems was identified as a "putative Candidate modular chitinase with two GH18 catalytic domains" from *Flavobacterium johnsoniae* UW101 (Fig.5 and Tab.4). The alignment performed in BLASTx covers 79% of the CDS detected with RAST annotation.

Another putative partial GH18 chitinase has been detected through BLASTx at the beginning of the contig#173 resulting in a lack of the start codon, and no CDS detection (Fig.9). The hit is annotated toward *Mycobacterium gilvum PYR-GCK* as the closest known hit for chitinase GH18 in database.

Finally, a third putative partial GH18 chitinase was also detected through BLASTx at the beginning of the contig#164 (Fig.10). Compared to the one above, the sequence is partial and mixed with another CDS related to a putative hemaglutinin protein. The hit is annotated toward *Mycobacterium vanbaalenii PYR-1*as the closest known hit for chitinase GH18 in database.

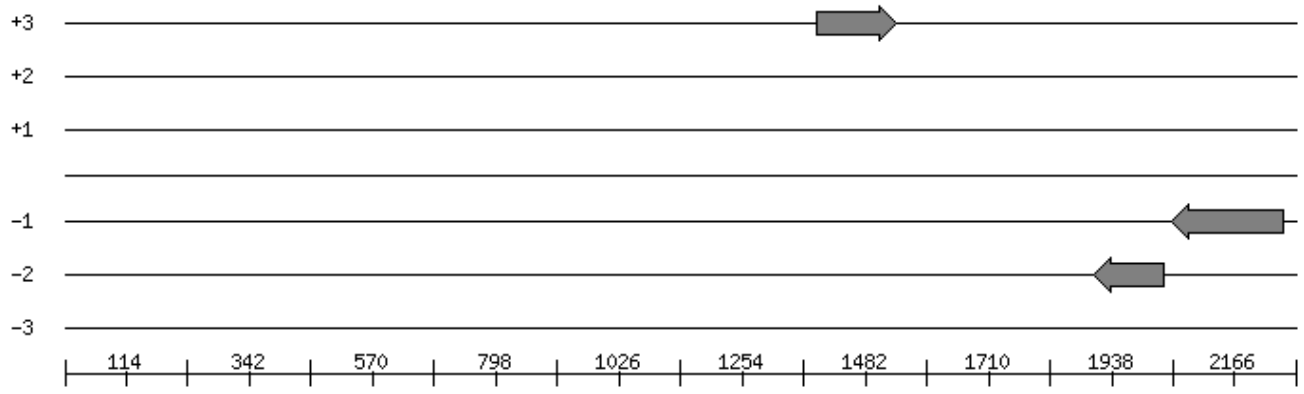


Fig.9: SEED annotation of the CDS detected in contig#173 (2.3kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). Grey arrows are representing hypothetical CDS not classified in SEED. Partial sequence of a putative chitinase has been detected at the very beginning of the contig, from position 3 to 260.

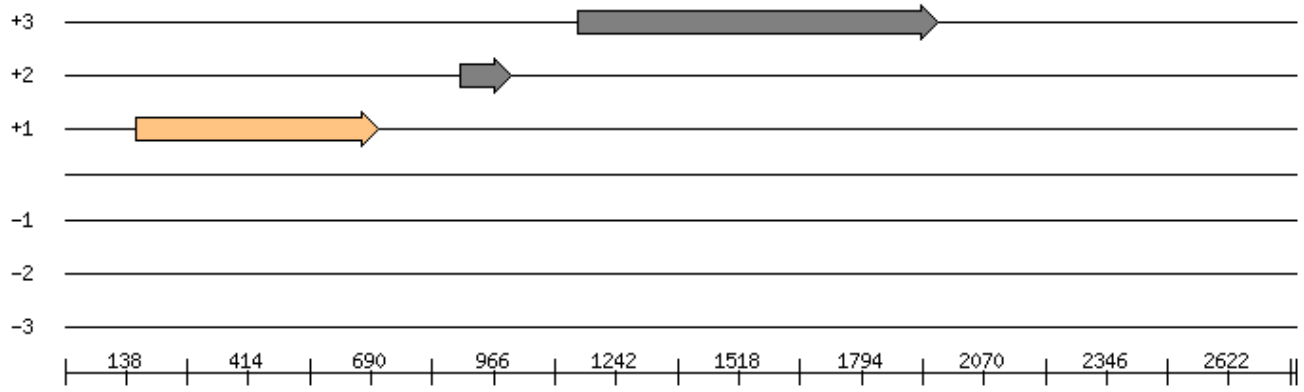


Fig.10: SEED annotation of the CDS detected in contig#164 (2.8kb). Annotations are given on the 6 reading frames (+1, +2, +3, -1, -2, -3). Grey arrows are representing hypothetical CDS not classified in SEED. The beige arrow is displaying the CDS of the hypothetical protein identified as a putative chitinase after comparison with BLASTx results.

4. Discussion

The sequence-based screening of the Rothamsted library was found to be a successful approach to detect putative CDS of interest among the metagenomic inserts carried by the clones. Thanks to this method, full CDS of several gene of interest were detected, encompassing 11 beta-lactamases, 5 enzymes involved in chitin degradation, one dehalogenase and several mobile genetic elements. All the putative sequences of interest identified so far will be used for further characterization during the last stage of the Metaexplore project. At first sight, the technique used in this study showed mitigated results, as poor evidences of an actual hybridization were retrieved from direct BLASTn of the probes sequences against the assembled contigs. Only the probes targeting the mobile genetic elements yielded good matches with this strategy, and only short conserved regions of 15-30bp were detected for the remaining probes. Advanced alignment attempts were realized with the ClustalX tool between the probes and the CDS identified, which yielded medium quality results between 45-55% similarities.

However, even if it wasn't easy to establish a direct link between the probes and the targeted genes with BLASTn and ClustalX tools, good evidences suggesting that the technique is operational were observed. For instance, the simple fact that almost all the genes of interest were retrieved among the 88 positive clones that were identified out of the 405 504 initial ones tend to reinforce the idea that hybridization has operated. Furthermore, the innovative principle of this technique, which makes it stand out from the others, relies on the flexibility of the hybridization process. Temperature, incubation time, and salt content are the main parameters that can be used to modify the pairing stringency. This flexibility should theoretically allow recovery of more or less dissimilar sequences that might diverge from the original probes sequences. This could explain the poor matches found in both BLASTn and ClustalX, and also the low similarity with the closest known reference sequences in the database after annotation through RAST. For instance, after psiBLAST, the genes implicated in chitin degradation that were found in this study showed a coverage of 66-97% with only 33-64% of identity. Furthermore, 84.56% of the coding DNA sequences detected after annotation with RAST were affiliated as unclassified structures (2964/3505), whose 65.96% (1955/2964) are referenced as hypothetical protein structures. This highlights the fact that the major part of the sequences retrieved from soil metagenomes are coding for unknown protein, with hypothetical or unknown functions, which gives us direct evidences of the huge diversity potential laying in the environmental microbial communities. Unfortunately, the RAST strategy did not allow detection of chitinase related CDS in the contigs. This is probably due to the fact that chitinases are sharing some similarities in their protein sequences with other glycosyl hydrolases or glycosyl transferase, which can partly explain the detection of many glycosyl hydrolase and transferase in the contig pool. This was confirmed by the parallel screening through BLASTx approach, which generates many false positives among glycosyl hydrolase and transferase. However, the confrontation of RAST and BLASTx did succeed in identifying a probable putative chitinase from a CDS annotated as unclassified hypothetical structure. The sequence is highly dissimilar to the closest chitinase hit in database (25%) but the alignment in BLASTx covered 79% of all the CDS detected through RAST, which is encouraging. Furthermore, two others partial chitinase GH18 were detected on small contigs, both annotated toward *Mycobacterium sp.* These results show the interest of combining different approaches in order to strengthen observation, and help identifying interesting features.

In addition, it's noticeable and quite surprising to have been able to recover a putative chitin deacetylase whereas no probes were set up to specifically target this kind of genes. The most probable explanation is that we co-select this gene by pure luck, during the hybridization of the dehalogenase probe. However, since the chitin deacetylase are of interest for the Metaexplore project, this "side discovery" is promising and should be investigated deeper. A collaboration is on-going with the partner#13 of the project (*Università* degli Studi dell'Insubria, UNINS, Varese, Italia), since they have a strong back ground in enzymology and functional screening. For instance, the two putative sequence of the chitin deacetylase were investigated by Ilaria Presti (PhD student, *Università* degli Studi dell'Insubria, UNINS, Varese, Italia). The analysis of the chitin deacetylase from contig#52 (Tab.2, Fig.6) at the genetic level reveals a high compatibility of the codon usage with *E. coli* as a host (93%, compatibility threshold = 30%), with a CAI (Codon Adaption Index) of 0.72. The CAI is integrating the codon usage deviation of the gene sequence in respect to a given host (recommended threshold for proper expression = 0.8). In addition, the GC content (54.47%) was found to be more suitable with Gram negative bacteria, such as *E. coli.* These results are suggesting that this gene may be compatible for expression assays in *E. coli* strains. The analysis at the protein level are predicting a stable protein of 427aa (computed index of stability = 34.67) with a size of 46.13 kDa and an isolelectric point of 8.85. In addition, a signal peptide of 25aa was detected at the N-terminal region, with a high probability to be recognized in Gram positive bacteria for secretion (e.g *Streptomyces sp*.). Unfortunately, the condon usage in Streptomyces sp. is only compatible at 61%, with a CAI of only 0.29. The suggested host for expression is *E. coli* Rosetta (DE3, Novagen), which is improved for rare codon translation.

On the other hand, the chitin deacetylase from contig#50 (Tab.2, Fig.10) displays less promising features in terms of putative expression, but still of interest. The analysis at the genetic level reveals a good compatibility of the codon usage with *E. coli* as a host (80%, compatibility threshold = 30%), with a CAI (Codon Adaption Index) of 0.55. The CAI is integrating the codon usage deviation of the gene sequence in respect to a given host (recommended threshold = 0.8). In addition, the GC content (63.88%) was found to be more suitable with Gram positive bacteria, such as *Streptomyces spp.* However, the codon usage is reaching 61% in *Streptomyces spp.*, with a CAI value of 0.28. These results are suggesting that this gene probably needs optimization for an eventual expression in both *E. coli* and *Streptomyces spp.* At this stage, a strategy relying on gene synthesis might be a good alternative. The analysis at the protein level are predicting an unstable protein of 300aa (computed index of stability = 42.43) with a size of 33.09 kDa and an isolelectric point of 5.63. In addition, the N-terminal region as a low probability to be recognized a signal peptide in both Gram positive and negative bacteria. These observations are indicating that sequence arrangement might be necessary in order to optimize the expression and secretion of this protein. The other genes found in this study are currently being analyzed the same way. Expression attempts will be carried on as soon as the clones harboring the genes of interest will be identified

When looking in the literature, the clone recovery yield that was obtained through our approach is found to be in the same order of magnitude than other studies based on soil metagenomic libraries. With 88 positive clones identified out of 405 504, the recovery yield is 2.20E-04%. For instance, a single clone harboring cellulase activity was found in a 10 000 clone library made from forest soil metagenomic DNA, which represents a yield of 1.00E-04% (Wang *et al.* 2009). In the same way, 70 000 clones were necessary to recover a single candidate displaying xylanase activity from another study based on forest soil as well, with a yield of 1,42E-05% (Kim *et al.* 2008).

More recently, 3 clones were identified out of a 13 800 library after a functional screening targeting both cellulase and xylanase activities, corresponding to a yield 2.17E-04 (Nacke *et al.* 2012). All these studies were realized with metagenomic DNA and fosmid vectors. However, if considering other habitats, like the cow rumen, the detection yield can be significantly increased. For instance, 93 clones harboring activities of interest toward plant biomass were detected in a 10 000 metagenomic clone library, which represents a yield of 9.30E-03% (Nguyen *et al.* 2012). These observations are coherent with the results we got for the soil of Rothamsted, in spite of the fact that other environments seem better suited for detecting specific activities. However, the methodology and the technology used in our study is totally different from the ones previously cited. All these studies were relying on functional screening while we undertake a sequence based screening, which doesn't give the same information. In addition, it is important to mention that our strategy was used to target different king of genes, encompassing several and distinct functions. As a consequence, if we consider only one category of gene, the final yield will be lower.

The hybridization technique used in this study relies on ^{33}P, which is known for its sensitivity. This helped us a lot during the detection of the positives clones, as the emitted signal was powerful enough to make clear dots on the screen. However, even if our results are similar to other studies that aimed for metagenomic library screening, the efficiency rate of the technique is still very low. This gives us again direct inquiring about the diversity level in a soil microbial community, as the probability of finding a given sequence is extremely low. However, identifying positive clone out of a nylon membrane after hybridization is not something easy to achieve. Even though the clone spots are often easily remarkable, the background is still really high and non-homogeneous, which results in a high degree of signal variability on a single membrane. Furthermore, lots of parasite signals were observed such as dust or any kind of debris, resulting in generation of disturbing dark dots that can be easily confused with a positive signal if it stands around a clone location onto the membrane. However, it is still conceivable that some positive signal could have been quenched somehow, which would limit the signal intensity barely above the background level. This is something observable when looking closer at the membranes, and it would be interesting to find a process that would allow detection of this probable positives. Collaboration with experts in the field of signal treatment and picture managing would definitively be of interest for a deeper investigation of the potential laying on these membranes. For example, designing a process that would handle the local background and highlight the dots that are locally significantly higher in terms of signal would be something really advantageous for better identification of putative positives. Another point has raised our attention concerning the prevalence of the mobile genetic element in the assembled contigs. Actually, this strong presence coupled with the high matches obtained with their probes through BLASTn might be responsible for almost all the detection process. It would mean that the clones identified by hybridization with the radioactivity spots are mostly due to the presence of mobile genetic elements in every one of them. This would also mean that the detection of the other genes is a "side-discovery" that was obtained by pure luck. This hypothesis might be true for a part of the positive clones, as we actually detected both mobile genetic elements and genes of interest on the same contigs. But the opposite is also true; some of the genes of interest were found alone on some contigs. In addition, the mobile genetic elements were observed on 26 contigs, which is not sufficient to covers all the 88 clones in terms of genetic material. If the effect of the mobile genetic element probes has probably played an important role in the clone recovery process, it doesn't appear to be the major driving force of detection.

Additional work is needed to improve the whole process, but also to clearly identify the clone carrying the detected CDS among the pool. The main strategy to do so would rely on primer design in order to target the CDS of interest, and screen through PCR the 88 clones in order to specifically identify who is carrying what. Further functional screening needs to be done to check whether the protein is expressed, secreted and functional. This can be done in collaboration in the frame of METAEXPLORE with other partners involved in functional screening improvement, or at Libragen SA as they dispose of robotics and facilities for massive screening of the library. Finally, as no specific SEED subsystem exist for the bacterial laccases, they can be screened among the fosmid pool through BLASTx as well, but also with the HMM models developed in the frame of METAEXPLORE (Ausec and Zakrzewski *et al.* 2011).

References

Ausec L, Zakrzewski M, Goesmann A, Schlüter A, Mandic-Mulec I (2011) Bioinformatic analysis reveals high diversity of bacterial genes for laccase-like enzymes. *PLoS One*. 2011;6(10):e25724

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics. 9:75*

Bertrand H, Poly F, Van VT, Lombard N, Nalin R, Vogel TM, Simonet P (2005) High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J Microbiol Methods*. 62:1-11

Ceccherini M, Pote J, Kay E, Van VT, Marechal J, Pietramellara G, Nannipieri P, Vogel TM, Simonet P (2003) Degradation and transformability of DNA from transgenic leaves. *Appl Environ Microbiol.* 69 :673–678

Demanèche S, David MM, Navarro E, Simonet P, Vogel TM (2009) Evaluation of functional gene enrichment in a soil metagenomic clone library. *J Microbiol Methods.* 76:105-107

Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tuphile K, Bertrand H, Vogel TM, Perriere G, Simonet P, Nalin R (2004) Phylogenic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl Environ Microbiol*. 70:5522-5527

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27(2):221-224

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser*. 41:95-98

Hobel CF, Marteinsson VT, Hreggvidsson GO, Kristjansson JK (2005). Investigation of the microbial ecology of intertidal hot springs by using diversity analysis of 16S rRNA and chitinases genes. *Appl Environ Microbiol*. 71:2771-2776

Kim SJ, Lee CM, Han BR, Kim MY, Yeo YS, Yoon SH, Koo BS, Jun HK (2008) Characterization of a gene encoding cellulase from uncultured soil bacteria. *FEMS Microbiol Lett*. 282:44–51

Monier JM, Bernillon D, Kay E, Faugier A, Rybalka O, Dessaux Y, Simonet P, Vogel TM (2007) Detection of potential transgenic plant DNA recipients among soil bacteria. *Environ Biosafety Res*. 6:71-83

Nacke H, Engelhaupt M, Brady S, Fischer C, Tautzt J, Daniel R (2012) Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnol Lett*. 34:663-75

Nguyen NH, Maruset L, Uengwetwanit T, Mhuantong W, Harnpicharnchai P, Champreda V, Tanapongpipat S, Jirajaroenrat K, Rakshit SK, Eurwilaichitr L, Pongpattanakitshote S. (2012) Identification and characterization of a cellulase-encoding gene from the buffalo rumen metagenomic library. *Biosci Biotechnol Biochem.* 76:1075-1084

Wang F, Li F, Chen G, Liu W (2009) Isolation and characterization of novel cellulase genes from uncultured microorganisms in different environmental niches. Microbiol Res 164:650–657

Ward NL, Challacombe JF, Janssen PH, Henrissat B, Coutinho PM, Wu M, Xie G, Haft DH, Sait M, Badger J, Barabote RD, Bradley B, Brettin TS, Brinkac LM, Bruce D, Creasy T, Daugherty SC, Davidsen TM, DeBoy RT, Detter JC, Dodson RJ, Durkin AS, Ganapathy A, Gwinn-Giglio M, Han CS, Khouri H, Kiss H, Kothari SP, Madupu R, Nelson KE, Nelson WC, Paulsen I, Penn K, Ren Q, Rosovitz MJ, Selengut JD, Shrivastava S, Sullivan SA, Tapia R, Thompson LS, Watkins KL, Yang Q, Yu C, Zafar N, Zhou L, Kuske,CR (2009) Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. *Appl. Environ. Microbiol*. 75:2046-2056

Zhao B, Mesbah NM, Dalin E, Goodwin L, Nolan M, Pitluck S, Chertkov O, Brettin TS, Han J, Larimer FW, Land ML, Hauser L, Kyrpides N, Wiegel J (2011) Complete genome sequence of the anaerobic, halophilic alkalithermophile Natranaerobius thermophilus JW/NM-WN-LF. *J Bacteriol*. 193:4023-4024

Zhou F, Badillo-Corona JA, Karcher D, Gonzalez-Rabade N, Piepenburg K, Borchers AM, Maloney AP, Kavanagh TA, Gray JC, Bock R (2008) High-level expression of human immunodeficiency virus antigens from the tobacco and tomato plastid genomes. *Plant Biotechnol J*.6:8

Contig#	Size (bp)	ORF	Length (bp)	Function	Category	Accession	Closest related organism	Coverage	Identity	E-value
72	33352	+2	165	HAD hydrolase, haloacid dehalogenase-like	Dehalogenase	ZP_05031401	*Microcoleus chthonoplastes PCC 7420*	90%	65%	9.00E-15
72	33352	-1	306	Transposase and inactivated derivatives	MGE	NC_010717	*Xanthomonas oryzae pv. oryzae PXO99A*	100%	72%	1.00E-33
50	39529	-2	903	Polysaccharide deacetylase CE4	none	ACB84535.1	*Candidatus Solibacter usitatus Ellin6076*	79%	33%	1.00E-21
52	38596	+3	1134	GH3 Beta-hexosaminidase (EC 3.2.1.52)	Chitin and NAG utilization	ZP_03129014	*Chthoniobacter flavus Ellin428*	97%	64%	1.00E-172
52	38596	+1	1284	Polysaccharide deacetylase CE4	none	ACB84535.1	*Natranaerobius thermophilus JW/NM-WN-LF*	66%	33%	5.00E-26
55	38316	+1	2082	GH20 Beta-hexosaminidase (EC 3.2.1.52)	Chitin and NAG utilization	ZP_06966074	*Ktedonobacter racemifer DSM 44963*	73%	46%	1.00E-149
55	38316	-1	1095	Hypothetical oxidoreductase	Chitin and NAG utilization	NC_008254	*Mesorhizobium sp. BNC1*	99%	54%	3.00E-104
55	38316	-3	768	N-acetylglucosamine-6-P deacetylase (EC 3.5.1.25)	Chitin and NAG utilization	AAKY00000000	*Burkholderia dolosa AUO158*	98%	49%	1.00E-52
55	38316	-3	345	N-acetylglucosamine-6-P deacetylase (EC 3.5.1.25)	Chitin and NAG utilization	NP_809588	*Bacteroides thetaiotaomicron VPI-5482*	88%	57%	3.00E-06
55	38316	-1	621	Rhodopirellula transposase	MGE	NC_008639	*Chlorobium phaeobacteroides DSM 266*	88%	60%	1.00E-58
55	38316	-3	189	Transposase	MGE	NC_008639	*Chlorobium phaeobacteroides DSM 266*	100%	71%	2.00E-19

Tab.3: Description of the contigs and relevant CDS corresponding to chitin and NAG utilization, dehalogenase and associated mobile genetic elements. The table gives the CDS locations on contigs, as well as details toward the closest known hit detected in databases (psiBLAST)

Contig#	Accession	Function	Clostest related organism	Family	Identity	Length	Start	Stop	E-value
164	ABM14300	Chitinase, Cellulase	*Mycobacterium vanbaalenii PYR-1*	GH18	36%	280	4	804	6,00E-28
173	ABP45478	Chitinase	*Mycobacterium gilvum PYR-GCK*	GH18	54%	85	257	3	4,00E-18
60	ACU75354	Glycoside hydrolase family 18	*Catenulispora acidiphila DSM 44928*	GH18	40%	130	4198	4584	1,00E-17
86	BAH40332	Putative glycosyl transferase/polysaccharide deacetylase	*Gemmatimonas aurantiaca T-27*	GH18	43%	105	6572	6264	9,00E-15
86	ADG12229	Polysaccharide deacetylase	*Caulobacter segnis ATCC 21756*	GH18	39%	107	9485	9171	2,00E-13
103	ADG12229	Polysaccharide deacetylase	*Caulobacter segnis ATCC 21756*	GH18	31%	131	10716	10324	5,00E-10
92	ABQ07554	Candidate modular chitinase with two GH18 catalytic domains	*Flavobacterium johnsoniae UW101*	GH18	25%	640	27446	25701	5,00E-10
39	CAJ68298	Putative bifunctional protein: peroxiredoxin/chitinase	*Clostridium difficile 630*	GH18	33%	105	36436	36140	3,00E-09
102	BAH40183	Putative glycosyl transferase/polysaccharide deacetylase	*Gemmatimonas aurantiaca T-27*	GH18	27%	264	3905	3168	6,00E-09
39	ABU72206	Hypothetical protein VIBHAR_03258	*Vibrio harveyi ATCC BAA-1116*	GH18	24%	385	8110	9204	6,00E-09
60	BAI95553	Putative bi-functional transferase/deacetylase	*Sphingobium japonicum UT26S*	GH18	38%	87	278	18	7,00E-09
53	ABZ69365	Polysaccharide deacetylase	*Caulobacter sp. K31*	GH18	35%	106	38231	37926	7,00E-09
97	ACB54367	Hypothetical protein cce_5021	*Cyanothece sp. ATCC 51142*	GH19	37%	71	15601	15389	2,00E-08
60	BAH40332	Putative glycosyl transferase/polysaccharide deacetylase	*Gemmatimonas aurantiaca T-27*	GH18	26%	217	8525	7923	2,00E-08
60	ADP71373	Polysaccharide deacetylase	*Rhodomicrobium vannielii ATCC 17100*	GH18	37%	87	33921	33664	3,00E-08
62	ADP71373	Polysaccharide deacetylase	*Rhodomicrobium vannielii ATCC 17100*	GH18	33%	123	26267	25902	6,00E-08
159	ADW17702	Polysaccharide deacetylase	*Desulfobulbus propionicus DSM 2032*	GH18	33%	118	2007	1657	4,00E-07
49	ADV45176	Chitinase	*Bacteroides helcogenes P 36-108*	GH18	25%	232	27504	26842	5,00E-07
50	BAH40332	Putative glycosyl transferase/polysaccharide deacetylase	*Gemmatimonas aurantiaca T-27*	GH18	29%	232	15459	14791	7,00E-07
60	AAV89278	Polysaccharide deacetylase	*Zymomonas mobilis subsp. mobilis ZM4*	GH18	29%	119	16588	16241	1,00E-06
92	ADL50908	Glycoside hydrolase family 18	*Clostridium cellulovorans 743B*	GH18	27%	230	9628	8972	4,00E-06
39	ADO73661	Chitinase, class II	*Stigmatella aurantiaca DW4/3-1*	GH18	28%	201	4528	5082	5,00E-06
92	ACC42687	PE-PGRS family protein	*Mycobacterium marinum M*	GH18	23%	337	17815	16880	6,00E-06
92	ACU75354	Glycoside hydrolase family 18	*Catenulispora acidiphila DSM 44928*	GH18	27%	273	9214	8417	1,00E-05

Tab.4: Description of the 15 chitinase GH18 hits obtained after BLASTx of the contigs pool against home-made chitin degradative enzyme database. The table gives details on the closest known hit detected in databases. The lines in grey nuances correspond to the most relevant hits after confrontation with the RAST annotation.

Appendixe 2-3a

SUPPLEMENTAL INFORMATION

Design of the chitinase probes for the hybridization experiment

This section is presenting the strategy used to design the probes that were used in the hybridization experiment. As previously mentioned in the "material and methods" section, chapter 2.3, a preliminary plasmid clone library has been constructed based on chiA amplicons. These sequences were obtained after amplification of the conserved region of chiA genes from the metagenomic DNA extracted from the soil of Rothamsted. 50 sequences were obtained after sequencing (GATC), and a tree has been constructed in order to visualize the global pattern of chiA amplicons. Among the 50 sequences, 3 of them were selected as probes for further hybridization attempts. Here we discuss the results obtained after the tree building, and how we selected the sequences.

1. chiA amplicon analysis

The sequences obtained with the chiA primers from Hobel *et al.* (2005) were 270bp on average (142bp minimum, up to 302bp maximum), which was coherent with indication mentioned by the authors (≈250bp). After the BLASTn annotation against non-redundant database (nr), most of the amplicons were found to be affiliated to *Uncultured bacterium* chitinases, with some of them showing poor similarity and coverage with the closest match in known databases (Tab.1). This tends to indicate the prevalence of a good genetic potential for unknown chitinase-related genes inside the metagenomic resource from the soil of Rothamsted.

The sequence analysis based on sequence alignment revealed a high degree of divergence among the amplicons (Fig.1). Furthermore, a clear dichotomy has been detected, which segregates two distinct groups among the chiA amplicons: group A and group B (Fig.1). When looking closer to the sequence alignments, and more specifically, to the primer specificity, major differences were observed between the two chiA amplicon clusters (Fig.2). For instance, the group A is displaying a high specificity with the primers used in this study, while the group B presents a lot of dissimilarity between the primers and the amplified sequences (Fig.2). The chiA amplicon group A has 87.5% similarity with the forward primer (21/24), and 87% with the reverse (20/23), while the group B shares only 33.3% similarity for the forward primer (8/24), and 39% for the reverse (9/23). However, the last 2 nucleotides located at the -3' extremities of each primer seem to have hybridized with the target, resulting in a successful amplification even though the similarity with the target was low. Furthermore, some of the sequences retrieved through this aspecific hybridization had strong matches toward known chiA-related sequences in the databases (Tab.1, Group B).

2. chiA probes selection

After discussion, we decided to select 3 sequences among the pool as chiA probes for further hybridization attempts. As many probes will be used in this experiment, each targeting a very specific gene (e.g chitinases, laccases, dehalogenases), it has been decided to use a maximum of 3 probes per target in order to limit possible interference between sequences.

The main criterion in the chiA probe selection was to optimize the recovery of something "knew" in terms of nucleic acid sequence, which means low matches against known sequence in databases. In this sense, the chiA group B was particularly interesting as it shows relevant affiliation to chitinase-related genes, and also a high degree of divergence comparing to the primer sequences used. In addition, side criterions should be verified before finalizing the probe selection. For instance, the cross-similarity between the 3 probes should not be too high in order to avoid undesired hybridization process. Furthermore, the probes should also display very low similarity with any of the other probes used in the experiment (e.g laccases, dehalogenases…), but also with the vector backbone and the strain genome as well. All these verifications were carried on through BLASTn in order to check the absence/presence of any similitude among all the nucleic acid sequence involved in this experiment. The final 3 chiA sequences selected were carefully checked in order to verify any of these criterions.

One sequence was selected in the group A that shows high specificity with the primers: Hobel-chiA-C9 (Tab.1). The 2 others sequences were picked up in group B, which displays lots of dissimilarity with the primers used: Hobel-chiA-F8 and B3 (Tab.1). The sequences were selected as follow: Group A chiA-C9 was displaying an average pattern in terms of sequence coverage and similarity, and the minimum size required as a probe. Group B chiA-F8 is showing a high coverage (93.58%) with an average similarity (66%), which is perfect for compromising the hybridization strength and the novelty discovering. Group B chiA-B3 is displaying a very low coverage (25.7%) but with a quite high degree of similarity (84%), which is interesting as it will target the surrounding of this "short known" sequence, with regards to potential new genetic structures. Amplicons chiA-F8 and B3 are some of the longest sequences retrieved in this experiment, thus improving the hybridization strength.

The final set of probes corresponding to the chiA gene was respecting all the criterions regarding the possible similarities with other nucleic acid sequences involved in the hybridization experiment. This set was designed in order to limit interferences with other sequences, but also to screen for probable novelties in terms of chitinases-related genes.

Group	Amplicon	NCBI ID	Description	Size (bp)	Coverage	Similarity	p-value
A	Hobel-chiA-D9	AB361766	*Uncultured bacterium chitinase*	302	81,13%	91,43%	2,00E-92
A	Hobel-chiA-F6	AF484817	*Uncultured bacterium clone controlA3S chitinase*	282	93,26%	76,43%	1,00E-44
A	Hobel-chiA-F11	AF484817	*Uncultured bacterium clone controlA3S chitinase*	286	91,96%	75,67%	6,00E-42
A	Hobel-chiA-C5	CP001700	*Catenulispora acidiphila DSM 44928 chitinase*	289	92,04%	72,56%	2,00E-30
A	Hobel-chiA-C2	AB243211	*Uncultured bacterium chitinase*	287	82,93%	72,27%	9,00E-27
A	Hobel-chiA-B10	AB362031	*Uncultured bacterium chitinase*	281	46,98%	78,03%	1,00E-18
A	Hobel-chiA-A8	AB362031	*Uncultured bacterium chitinase*	293	45,05%	77,27%	1,00E-17
A	Hobel-chiA-A10	EU700235	*Uncultured bacterium clone CHI39 chitinase*	259	92,66%	69,58%	2,00E-14
A	Hobel-chiA-C11	EU700245	*Uncultured bacterium clone CHI49 chitinase-like*	290	25,17%	83,56%	2,00E-10
A	Hobel-chiA-D11	AY523558	*Uncultured organism clone L5-24 chitinase*	284	94,01%	66,67%	2,00E-09
A	Hobel-chiA-E9	AB362031	*Uncultured bacterium chitinase*	279	41,22%	72,17%	7,00E-09
A	Hobel-chiA-A6	AY523558	*Uncultured organism clone L5-24 chitinase*	284	93,66%	66,54%	7,00E-08
A	Hobel-chiA-C3	AY523558	*Uncultured organism clone L5-24 chitinase*	284	94,01%	65,92%	3,00E-07
A	Hobel-chiA-B11	AB361818	*Uncultured bacterium chitinase*	286	74,83%	68,69%	1,00E-06
A	Hobel-chiA-C8	AY523558	*Uncultured organism clone L5-24 chitinase*	280	95,00%	66,54%	1,00E-06
A	Hobel-chiA-A12	EU700210	*Uncultured bacterium clone CHI14 chitinase*	226	21,24%	87,50%	4,00E-06
A	Hobel-chiA-A11	AY523558	*Uncultured organism clone L5-24 chitinase*	286	94,06%	66,91%	1,00E-05
A	Hobel-chiA-F12	AY523558	*Uncultured organism clone L5-24 chitinase*	280	95,00%	65,79%	1,00E-05
A	**Hobel-chiA-C9**	**CP000360**	***Candidatus Koribacter versatilis Ellin345 chitinase***	**235**	**54,89%**	**69,77%**	**2,00E-04**
A	Hobel-chiA-D10	EU700200	*Uncultured bacterium clone CHI4 chitinase*	290	32,76%	72,63%	2,00E-04
A	Hobel-chiA-A2	AP009493	*Streptomyces griseus NBRC 13350 chitinase*	223	41,26%	73,91%	0,002
A	Hobel-chiA-F5	EU700200	*Uncultured bacterium clone CHI4 chitinase*	223	39,46%	72,73%	0,002
A	Hobel-chiA-F4	CP001778	*Stackebrandtia nassauensis DSM 44728 chitinase*	186	26,34%	81,63%	0,005
A	Hobel-chiA-B12	AP009493	*Streptomyces griseus NBRC 13350 chitinase*	142	60,56%	73,26%	0,091
B	Hobel-chiA-B2	AB361977	*Uncultured bacterium chitinase*	226	89,38%	87,62%	1,00E-63
B	Hobel-chiA-C7	AF484817	*Uncultured bacterium clone controlA3S chitinase*	287	92,68%	74,81%	7,00E-41
B	Hobel-chiA-F3	EU700232	*Uncultured bacterium clone CHI36 chitinase*	274	90,51%	76,21%	3,00E-40
B	Hobel-chiA-D8	CP000481	*Acidothermus cellulolyticus 11B chitinase*	299	89,97%	73,23%	2,00E-35
B	Hobel-chiA-F1	AB361700	*Uncultured bacterium chitinase*	289	75,09%	74,65%	4,00E-31
B	Hobel-chiA-D3	U81007	*Doohwaniella chitinasigens chitinase Chi67*	221	97,29%	72,09%	3,00E-20
B	Hobel-chiA-E11	AE016825	*Chromobacterium violaceum ATCC 12472, chitinase*	290	93,10%	70,00%	4,00E-19
B	Hobel-chiA-F10	AB362031	*Uncultured bacterium chitinase*	283	41,70%	77,97%	4,00E-13
B	Hobel-chiA-F7	EU700235	*Uncultured bacterium clone CHI39 chitinase*	263	93,54%	68,29%	5,00E-12
B	**Hobel-chiA-B3**	**EU700201**	***Uncultured bacterium clone CHI5 chitinase-like***	**292**	**25,68%**	**84,00%**	**1,00E-10**
B	Hobel-chiA-A9	AY523558	*Uncultured organism clone L5-24 chitinase*	286	93,01%	68,42%	5,00E-10
B	Hobel-chiA-B9	AB361770	*Uncultured bacterium chitinase*	282	81,56%	67,39%	8,00E-09
B	Hobel-chiA-A7	EU700240	*Uncultured bacterium clone CHI44 chitinase*	285	94,39%	68,03%	2,00E-08
B	Hobel-chiA-F2	AY523558	*Uncultured organism clone L5-24 chitinase*	284	95,07%	67,41%	3,00E-08
B	Hobel-chiA-E2	AY523558	*Uncultured organism clone L5-24 chitinase*	290	93,10%	67,41%	3,00E-08
B	Hobel-chiA-E4	AY523558	*Uncultured organism clone L5-24 chitinase*	284	95,07%	67,41%	3,00E-08
B	**Hobel-chiA-F8**	**EU700206**	***Uncultured bacterium clone CHI10 chitinase***	**296**	**93,58%**	**66,06%**	**1,00E-07**
B	Hobel-chiA-B1	AY523558	*Uncultured organism clone L5-24 chitinase*	271	98,52%	65,54%	3,00E-06
B	Hobel-chiA-D2	AB252252	*Uncultured bacterium chi19 chitinase*	290	36,90%	71,03%	0,002
B	Hobel-chiA-D12	AB252252	*Uncultured bacterium chi19 chitinase*	285	37,54%	71,03%	0,002
B	Hobel-chiA-A1	AY699347	*Uncultured organism clone, family 18 chitinase*	258	82,95%	65,89%	0,007
B	Hobel-chiA-E3	EU599185	*Collimonas fungivorans strain Ter331 chi locus A*	288	40,97%	72,88%	0,008
B	Hobel-chiA-B7	EU599185	*Collimonas fungivorans strain Ter331 chi locus A*	216	54,63%	72,88%	0,02
B	Hobel-chiA-D1	CP000113	*Myxococcus xanthus DK 1622 chitinase*	287	35,19%	72,28%	0,026
B	Hobel-chiA-D5	EU700206	*Uncultured bacterium clone CHI10 chitinase*	227	13,22%	96,67%	0,083
B	Hobel-chiA-E5	AF484829	*Uncultured bacterium clone controlA3S chitinase*	284	36,97%	72,38%	0,33

Tab.1: Best BLASTn annotation of the chiA amplicons obtained from the metagenomic DNA of Rothamsted soil. The PCR has been done with the primers from Hobel et al. (2005), and amplicons were cloned in PCR4® TOPO® plasmids (TOPO TA cloning kit for sequencing, Invitrogen). Inserts were sequenced at GATC Biotech, and sequences analyzed through BLASTn against nr. Sequences in dark grey nuance are the ones selected as probes for hybridization attempts. Group column is classifying the amplicons according the 2 clusters found through sequences alignment.

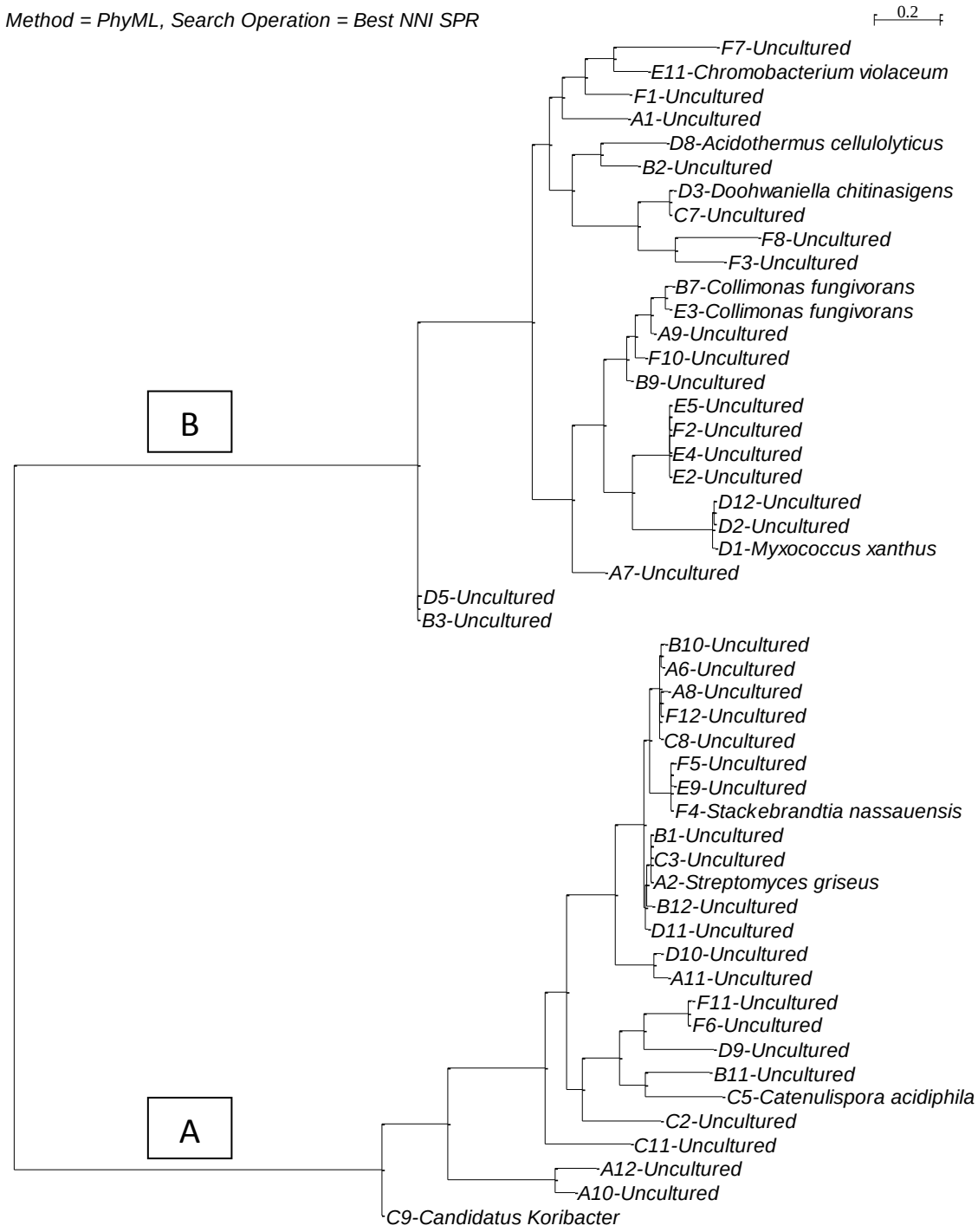


Fig.1: Distance tree of chiA amplicons. The tree was established with the PhyML method, which estimates the maximum likelihood phylogenies from the amplicon alignment. The phylogenetic space has been search with both of the best possibilities in NNI (Nearest neighbor Interchange) and SPR (Subtree Pruning and Regrafting). Letters A and B are representing the two distinct groups of sequenced obtained after alignment. Alignements were performed with Bioedit software (Hall et al. 1999), and trees were constructed with Seaview software (Gouy et al. 2010).

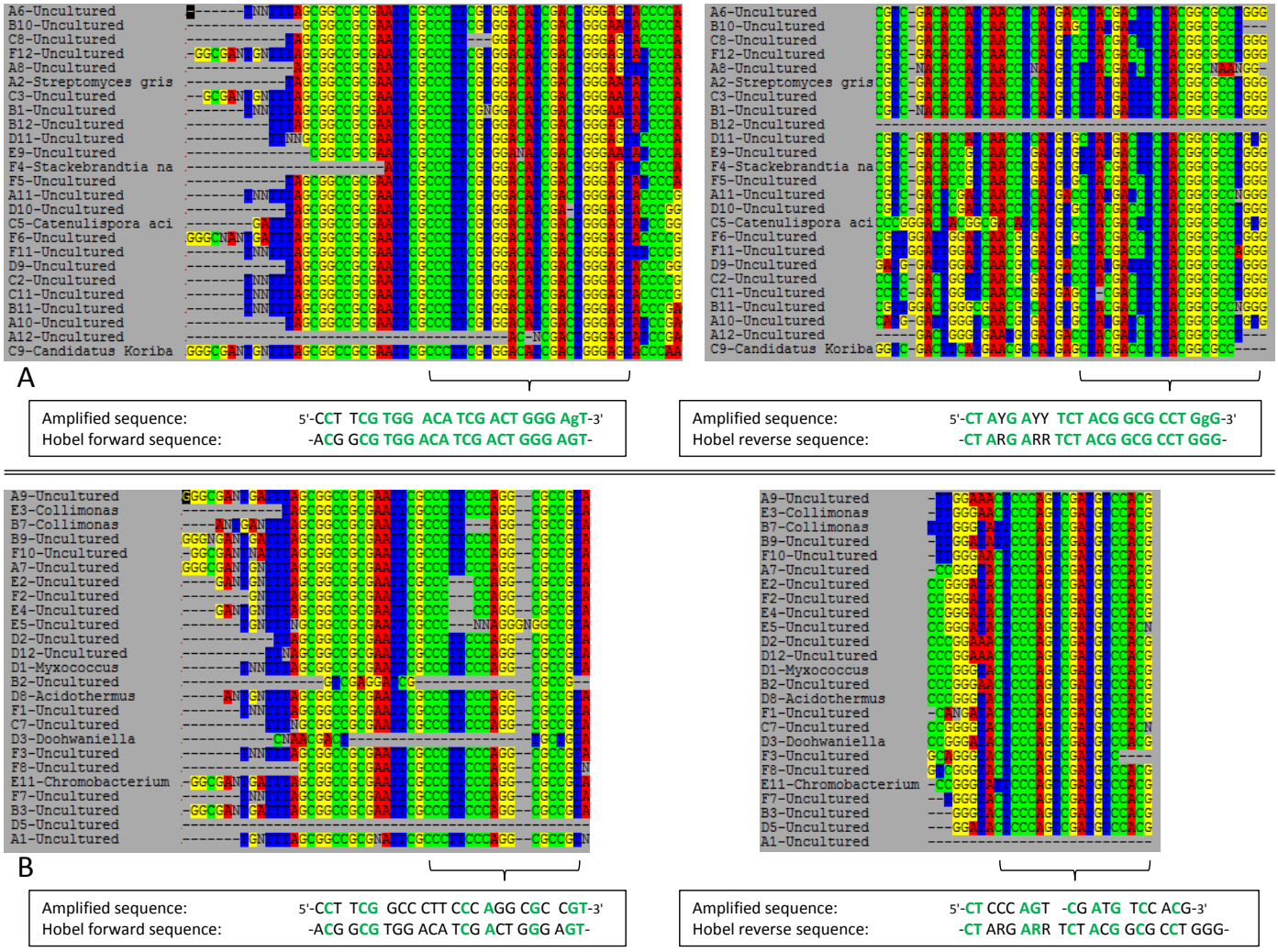


Fig.2: chiA primer specificity with the 2 groups of amplicons detected through alignment. Panel A displays the primer specificity for the group A (cf Fig.1), while panel B shows the specificity for the group B. Alignments were performed with Seaview software (Gouy et al. 2010).

CHAPTER 2 – Part.4

-

Soil bacterial community shifts after chitin enrichment

-

An integrative metagenomic approach

SAMUEL JACQUIOD[1], LAURE FRANQUEVILLE[1], SEBASTIEN CECILLON[1], TIMOTHY M .VOGEL[1] and PASCAL SIMONET[1]

[1] Ecole Centrale de Lyon, Laboratoire Ampère, Environmental Microbial Genomics Group, 36 avenue Guy de Collongue - 69134 Ecully, France

Corresponding author: Samuel Jacquiod (samuel.jacquiod@ec-lyon.fr, phone: +33 (0)4 72 18 60 92, fax: +33 (0)4 78 43 37 17)

Abstract

Chitin is the second most produced bio-polymer annually on Earth after cellulose. Chitinases, which carry out bio-degradation of chitin, are promising but untapped sources for developing novel biocatalysts for industrial applications. Hidden amongst uncultivated micro-organisms, bacterial chitinases can be discovered and exploited by metagenomic approaches, whose efficiency for detecting new enzymes can be increased by chitin enrichment. In this study, the soil from Park Grass at Rothamsted research station (UK) has been enriched with chitin in microcosm in order to investigate chitin degradation kinetics as well as taxonomical shifts of the soil bacterial community. An integrative approach coupling *rrs* phylochip and high throughput shotgun sequencing of the metagenomic DNA has been applied to investigate the shift in bacterial community structure and functions after incubation. Results indicate that chitin enrichment leads to an increase of Actinobacteria, γ-proteobacteria and β-proteobacteria suggesting the specific selection of chitin degrading bacteria belonging to these classes. An increase of the observed bacterial diversity was noticed, with detection of bacteria genera only in chitin treated conditions. This study demonstrates the interest of enrichment for shifting soil bacterial communities toward an enhancement of proportion and diversity of degrading bacteria. In addition, the extensive metagenomic knowledge accumulated from the Park Grass soil experiment reinforced the unique aspect of the chitin added soil analysis by comparison to variations associated with season, depth and DNA extraction protocols.

Chitin is an homopolymer of β-1,4-linked N-acetylglucosamine (N-acetyl-D-glucose-2-amine, NAG) with a critical biological role in both terrestrial and aquatic ecosystems as a major constituent of fungi and plant cell walls, insect, krill and shellfish exoskeletons (Gomes *et al.* 2001). Chitin is the second most abundant biopolymer in nature after cellulose, with an estimated natural production of 10^{10} tons year^{-} (LeCleir *et al.* 2007; Muzzarelli *et al.* 1999). The degradation of chitin releases substantial amounts of carbon and nitrogen in terrestrial (Williamson *et al.* 2000; Metcalfe *et al.* 2002; LeCleir *et al.* 2004; Uchiyama & Watanabe, 2006), aquatic (LeCleir *et al.* 2007, Ramaiah *et al.* 2000, Hobel *et al.* 2005) and Antarctic ecosystems (Xiao *et al.* 2005). Its crystallized conformation and heterogenic chemical composition render chitin particularly recalcitrant to degradation (Howard *et al.* 2003). Its complete mineralization is carried out only by a highly specialized microflora through specific microbiological enzymatic processes. Up to now, the most efficient chitin degraders among bacteria were isolated through culture, including representatives from Actinobacteria (e.g *Streptomyces sp.*) (Hjort *et al.* 2009), β-proteobacteria (e.g *Burkholderia* sp.) (Ogawa *et al.* 2002) and γ-proteobacteria (e.g *Xanthomonas sp.*) (Yamaoka *et al.* 1999). However, other bacteria might also possess the ability to degrade chitin even if they are much more difficult to isolate, and may be also involved as key players in carbon and nitrogen cycles by providing abundant and available nutrients to other microorganisms (Williamson *et al.* 2000; Gooday 1990; Lindahl *et al.* 2006). All known enzymes involved in chitin degradation are classified in the CAZy database (Carbohydrate Active Enzymes, http://www.CAZy.org/) (Cantarel *et al.* 2009), the chitinases (EC 3.2.1.14) belonging to the glycosyl hydrolases families GH18 and GH19 according to the classification based on amino acid sequence similarity (Henrissat *et al.* 1991; Henrissat 1993). Another chitin degradation pathway relies on chitin deacetylases (EC 3.5.1.41), belonging to the carbohydrate esterase family CE4 (Lombard *et al.* 2010) and chitosanase (EC 3.2.1.132) from family GH46 and GH75. Chitinases are also classified based on their depolymerization activity including endochitinases (EC 3.2.1.14) which randomly cleave chitin molecules and exochitinases such as β-(1,4)-N-acetylglucosaminidases (EC 3.2.1.30) and 1,4-β-chitobiosidases (EC 3.2.1.29) which progressively degrade chitin molecules from non-reducing ends (Graham *et al.* 1994; Harman *et al.* 1993). Apart from their fundamental role in ecosystem functioning, chitin degraders and their enzymes have received particular attention during the last decade for numerous applications. Chitin degraders are candidates for *in situ* application as bio-control agents of soil born plant-pathogenic fungi (Hjort *et al.* 2009) while shellfish wastes could be treated in the frame of enzymatic industrial processes involving new and efficient chitinases that could be characterized from environmental microbial communities (Horn *et al.* 2006, Han *et al.* 2010). Environmental bacteria constitute an untapped reservoir of enzymatic diversity considering that the bacterial metagenomic approaches were not yet extensively applied to the detection of new chitinases with the only exceptions of few attempts to clone DNA extracted from ocean water (Cottrell 1999), lake (LeCleir *et al.* 2007), soil (Hjort *et al.* 2009) and marine sediments (Bhuiyan 2011) which did not extend significantly the diversity range of characterized chitinases. These attempts indicate the difficulty of detecting new chitinase genes even by culture independent approaches. Complexity of environmental metagenomes related to the high level of bacterial diversity inhibits attempts for exploiting the hidden enzymatic diversity. The number of clones to be screened must be significantly scaled up to increase the chance of detecting rare genes, an effort that requires robots and screening facilities not necessarily available in an academic laboratory.

An alternative strategy would be to enhance the proportion of the targeted genes in the bacterial community to facilitate their detection. Addition of chitin rich compounds in environmental samples is expected to significantly increase the fitness of the adapted bacteria and consequently the proportion of chitinase genes in the extracted metagenome. We tested here this hypothesis by investigating the impact of chitin enrichment in soil samples on the taxonomic structure of bacterial communities. DNA was extracted from fresh soil microcosms samples incubated for 35 days with 2 different concentrations of colloidal shrimp chitin as well as a control: (1) water control (0x), (2) 2mg/g of soil (1x) and (3) 20mg/g of soil (10x). An extensive analysis of community shifts through the combinatory use of complementary techniques including *rrs* qPCR and phylochips and metagenome DNA shotgun high throughput sequencing in parallel to the survey of chitin degradation by enzymatic chitinase assays. Furthermore, 13 pyrosequencing runs from the same soil were used as control for enhancing detection of the significant changes at the metagenomic level (Delmont *et al.* 2012).

1. Material and methods

i. Soil sampling

Fresh soil cores were collected from Park Grass (lat 51.481481°N, long 0.222231°E), Rothamsted, England (see http://www.rothamsted.ac.uk/ for further information). The Park Grass from Rothamsted is an internationally recognized resource and is selected to be a reference for soil metagenomic studies (Vogel *et al.* 2009). It is classified as chromic luvisol according to FAO guidelines (FAO, 2006) and is silty clay loam overlying clay with flints with a pH of 5.2 (measured in H_2O). Park Grass covers 249 m^2 (13.28 by 18.75 m), and the sampling strategy consisted of taking randomized soil samples in the plot. The soil cores were around 6 cm diameters for 20cm depth, and were collected into plastic bags and rapidly transferred at the Ecole Centrale de Lyon (France). Soil particles were sieved at 2mm and directly used to run microcosm experiments. All tools and materials used were washed and cleaned with 70% ethanol solution.

ii. Microcosm's settings, chitin enrichment and sampling strategy

Three enrichment conditions corresponding to two chitin concentrations and a control were designed. Each condition was set in triplicates consisting in glass bottles containing 50g of sieved soil, and caped with cheese cloth to allow gas exchanges. Colloidal chitin was prepared from shrimp shell (Sigma C7170) as described by Inglis and Kawchuk, 2002. 15g of shrimp shell was dissolved in 50ml HCl (Roth, 37%) added to 300ml water during 4 hours with magnet bar mixing. The solution was neutralized at pH7 with NaOH 0.1M, and colloidal chitin was recovered after centrifugation (10min, 10 000g). A washing step was performed with 200mM phosphate buffer pH7, followed by a second centrifugation (10min, 10 000g). For the first chitin concentration, 100mg of colloidal chitin was added with 5ml water and mixed up with soil using sterile equipment for a final concentration of 2mg/g soil (Chitin 1x). For the second concentration, 1g of colloidal chitin was added in 5ml water, corresponding to 20mg of colloidal chitin per gram of soil (Chitin 10x). Control bottles were amended with 5ml water as well (Control 0x). Final water saturation in all bottles was around 65%. Microcosms were incubated in green-house condition with a constant temperature of 24°C and 60% relative humidity. 1g of soil was sampled for each replicate and condition after respectively 0, 3, 6, 10, 20, and 35 days enrichment. Samples were stored at -20°C until the kinetic end.

iii. Chitin degradation and chitinase assays

200mg of soil were collected from the 3 microcosm replicates and pulled together (3x200mg) with addition of 1.2ml Dulbecco's phosphate buffer saline pH 8.00. Pulled samples were mixed in a vortex for 2 minutes at maximum speed (Vortex Genie 2, SCIENTIFIC INDUSTRIES). The mixture was clarified by centrifugation at 13'000 rpm and 500µl supernatant was transferred in a clean collection tube and stored on ice. Exochitinase and endochitinase assays were performed on 10 µl of supernatant in duplicate, using a fluorometric chitinase assay kit following provider instructions (Sigma-Aldrich, Germany, Lifesciences). Specific substrate are provided for each activity (4-Methylumbelliferyl N-acetyl-β-D-glucosaminide for exochitinase and 4-Methylumbelliferyl β-D-N',N''-triacetylchitotriose for endochitinase). After chitinases degradation, both substrates release 4-Methylumbelliferone (4MU), a fluorescent coumpond emitting at 450nm when exited at 360nm. Fluorescence was measured after 45min incubation at 37°C on a microplate reader (Infinite 1000, TECAN) and activities were stated with a 4MU standard curve. Enzymatic activities were normalized and expressed in chitinase unity detected per gram of soil during incubation time (1U=1µmole 4MU released per minute).

iv. Metagenomic DNA extraction

DNA was extracted using an adapted protocol from (Griffith *et al.* 2000). 500mg of soil were placed in FastPrep lysing matrix. Cell lysis was performed in a FastPrep bead beating system, at 5.5 speed during 30s (MP bead beating; Bio101 Biomedical) with 600µl phenol:chloroform:isoamyl alcohol (Roth, 24:24:1) and 600µl extraction buffer (300µl phosphate buffer pH8, 300µl 10% cetyltrimethylammonium bromide/700mM NaCl). Supernatant aqueous phase was recovered after centrifugation (5min, 16000g, 4°C). A second treatment was performed on the recovered aqueous phase with 500µl chloroform isoamyl alcohol, followed by short vortex mix, and centrifugation (5min, 16000g, 4°C). Total DNA was precipitated over night at 4°C with 2 volumes of absolute ethanol and 1/10 volume NaCl 5M. DNA purification was done onto silica column according to manufacturer instructions (GE Health Care FX). DNA was quantified by fluorometric assays (Qubit fluorometer, Invitrogen, Life technologies).

v. rrs qPCR

Partial *rrs* gene was amplified using eubacterial primers Eub338 5' -ACTCCTACGGGAGGCAGCAG-3' (foward) and Eub 518 5'-ATTACCGCGGCTGCTGG-3' (reverse) (Fierer *et al.* 2005). 2µl of metagenomic DNA (≈15ng) was mixed with 0.4µl of reverse and foward primer (5µM), 10µl of qPCR buffer Sensimix® (Bioline) and 7.2µl of distilled sterile water. qPCR assays were performed in a RotorGene RG-6000 (Corbett Research, QIAGEN) with the following conditions, 10 min at 95°C, followed by 35 cycles at 95°C for 20s and annealing at 53°C for 20s, and elongation at 72°C for 20s. The standard range was performed using purified 16s-rRNA products amplified with the same set of primer from the original soil metagenomic DNA. A 7 points standard curve ranging from 10^4 to 10^8 molecular copies was used to quantify the *rrs* gene in the samples (Efficiency = 0.99, $R^2 = 0.999$). Results were analyzed with the provider software (Rotor-Gene 6000 Series Software 1.7), and detected *rrs* copy numbers were normalized according to DNA extraction yields per gram of soil.

vi. Phylochip analyses

The microarray format used in this study was from Agilent Sureprint Technologies, consisting in 8 blocks of 15,000 spots each, designed on a standard glass slide (25 mm x 75 mm). Each spot holds a 20-mer oligonucleotide probe synthesized *in situ*. Each oligonucleotide probe occurred at least in triplicate within each block. All blocks were identical. This format allows hybridization of eight samples per slide at the time. The use of multiple slides was necessary for hybridization of all samples. Probes target the *rrs* gene covering a wide par of the *Bacteria* and *Archaea* phylogenic tree. Probes were designed with the ARB software package and PhylArray (Martin-Laurent *et al.* 2001). The 20-mer probes have a final melting temperature of 65°C ± 5°C and less than 1.5 weighted mismatches. Our design includes oligonucleotide probes at different taxonomic levels. This microarray covers over 400 genera and 10000 OTUs ("species" or "hits"). The bacterial 16s rRNA gene was amplified from pooled metagenomic DNA, corresponding to microcosm replicates from the original soil at day0, and chitin enriched samples from day20: 0x20, 1x20 and 10x20. The PCR reaction was performed using universal primer *pA* 5'-TAATACGACTCACTATAGAGAGTTTGATCCTGGCTCAG-3' and *pH-T7* 5'-AAGGAGGTGATCCAGCCGCA-3' (Bruce *et al.* 1992). 2µl of diluted metagenomic DNA (≈15ng) was mixed with 1.5µl of reverse and forward primers (10µM), 45 µl of distilled sterile water, and 1µl Taq Polymerase (Invitrogen). PCR was conducted at 94°C for 4 min and then with 35 cycles of 94°C for 45s, annealing at 55°C for 45s, and elongation at 68°C for 95s, followed by 68°C for 5 min. Amplified PCR products were electrophoresed on a 1% agarose gel, and the desired 1500pb band was extracted and purified with GFX PCR DNA and gel band purification kit (GE Healthcare). Purified PCR products were then transcribed into RNA using T7 rRNA polymerase (Invitrogen) with the incorporation of labeled Cy5-UTP. Cy5 is a fluorescent dye, emitting at 670 nm after excitation at 650nm. rRNA purification was performed with the Qiagen RNeasy minikit according to the manufacturer's instructions. Chemical rRNA fragmentation was achieved by the addition of 1.14 µl of Tris-Cl (1 mM) and 4.57 µl of $ZnSO_4$ (100 mM) to 40 µl of labeled RNA sample and incubation for 30 min at 60°C.The fragmented and labeled rRNA was hybridized overnight on the phylochips at 60°C, and washed with the buffer supplied by the manufacturer..

vii. Microarray scanning and data processing

An Innoscan 700 scanner (Carbonne, France) was used for scanning microarray slides according to the manufacturer's instructions. Raw hybridization fluorescence signal for each spot was determined based on the signal-to-noise ratio (SNR), which was calculated by using the following formula: SNR = (signal intensity – background)/ background standard deviation. Total probe fluorescence signal, including negative controls, was transformed by calculating the signal in $\log_2$. Since at least three replicates exist for all oligonucleotide probes, outliers were eliminated when any individual spot was greater than 3 standard deviations from the average of all replicates. Analysis of variance (ANOVA) was used to evaluate positive probes. Since probes have different phylogenetic depths, the genera described here were those for which all relevant probes were positive. While all of the thousands of probes could not be independently verified, many of the probes were validated by the application of DNA from a single bacterium (Sanguin *et al.* 2006).

viii. Pyrosequencing and data analysis

Metagenomic DNA corresponding to the enrichment time day 20 extracted and purified as described in section "*Metagenomic DNA extraction"*. Up to 10µg were extracted for each condition and replicate. Only two replicates were sent for further pyrosequencing investigation. Pyrosequencing replicates are indicated as: “0x20a” “0x20b” “1x20a” “1x20b” “10x20a” “10x20b”. Metagenomic DNA was sequenced using half runs of 454 Titanium pyrosequencing (*Genoscope*, Paris, France) and raw data were cleaned from artificial duplicates using the CD-HIT-454 software (Niu *et al.* 2010). On average, each metagenome yielded 559671 reads (+/- $6.3E^4$), with an average read size of 392bp (+/- 32) and an average GC content of 63% (+/- 0.5).

Metagenomic datasets were analyzed though read annotation against M5NR database and functional subsystems using the MG-RAST platform (Meyer at al. 2008) Rguy (R Development Core Team, 2011) and STAMP software (Parks and Beiko, 2010). Two pyrosequencing datasets corresponding to the original soil at day 0: "0x0a" and "0x0b" were used for comparison (Delmont *et al.* 2012) (Tab.2). Furthermore, the datasets generated from the Park grass experiment corresponding to variation in seasons, depths and extraction techniques were used as a control to cross compare the different conditions, and validate the changes observed after microcosm and chitin treatment (Delmont *et al.* 2012).

2. Results

i. Chitinase activity

Soil samples in the microcosms were enriched with colloidal chitin (0x, 1x and 10x) for up to 35 days. Results presented in Fig. 1, panel A indicate that the main initial impact on the exochitinase activity was due to placing the soil in the microcosms, whatever the amount of chitin added including the control as well. A rapid activity drop in less than 3 days was detected for all conditions, with a slower decrease from day 10 for the control 0x and 1x. In addition, the presence of chitin in soil seems to affect the enzymatic activity as indicated at day 0, with a significant decrease for the highest concentration 10x. This activity decrease is maintained at a slower pace in the control and the lowest chitin concentration. However, the 10x chitin concentration leads to an increase of the exochitinase activity between days 3 and 10 before stabilization and a weak decrease.

A negative microcosm effect was also noticed between day 0 and day 3 on endochitinase activity (Fig.1, panel B). However, the presence of chitin in soil seems to have a lower inhibiting effect than for the exochitinase according to activity values measured at day 0 for all concentrations. Significant changes of the endochitinase activity all along the kinetics were detected in the 10x chitin concentration amended microcosms with a marked activity peak 10 days after amendment before a drop at day 20. Endochitinase activity in the 1x chitin amended and the control samples did not differ significantly, exhibiting both a continual slight decrease after the initial drop reflecting the microcosm effect.

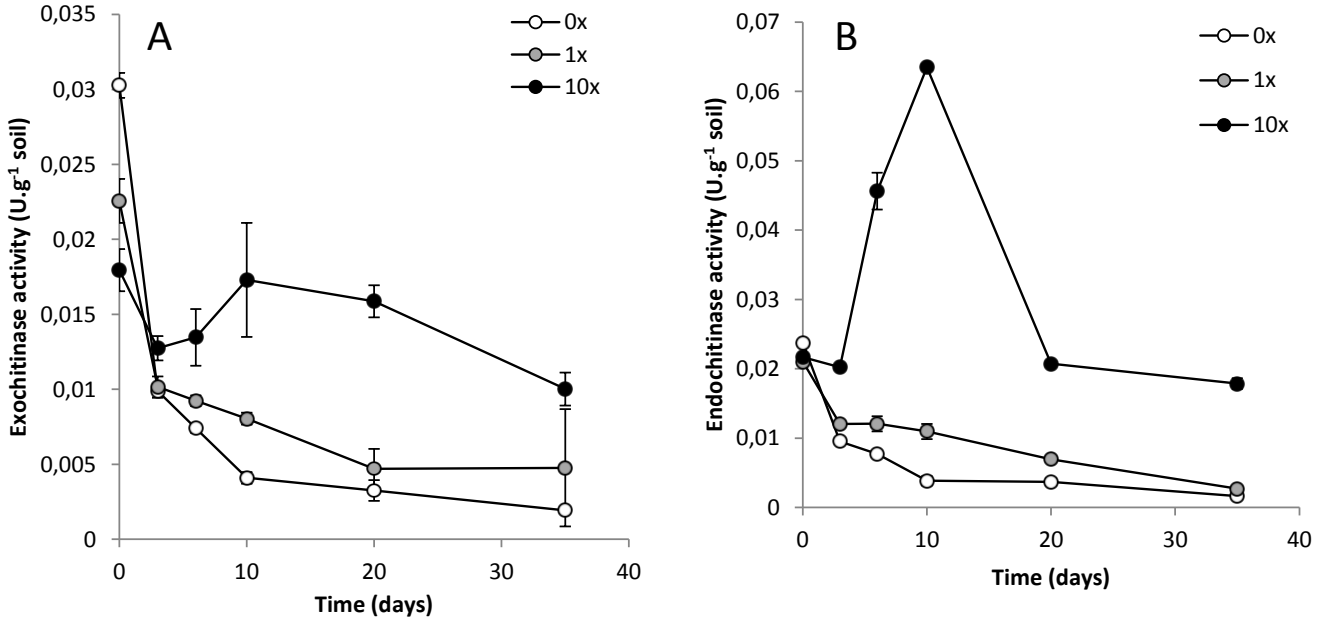


Fig.1: Exo- and endochitinases fluorometric assay on soil aqueous extracts. Panel A shows exochitinases activity and panel B the endochitinases activity. Activity was measured by quantification of the fluorescence release by the 4-Methylumbelliferon (4MU) after specific cleavage of exo- and endochitinases in both case. Enzyme activity is expressed in chitinase unity detected per gram of soil during incubation time (1U=1µmole 4MU released per minute).

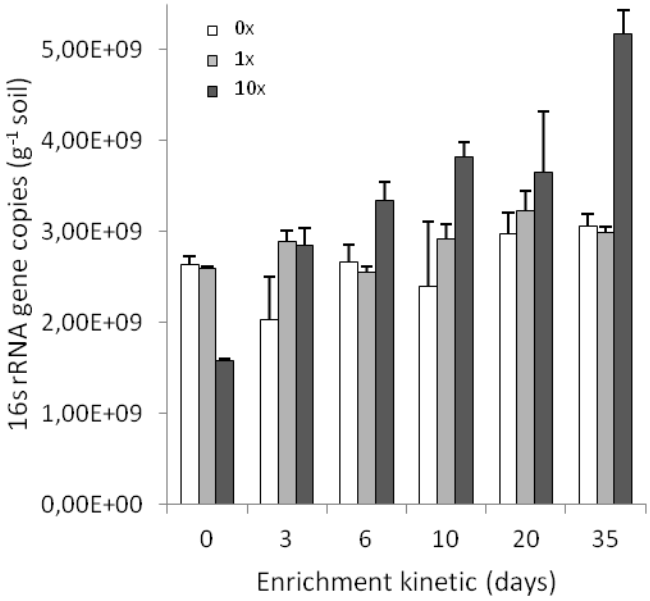


Fig.2: 16s-rRNA gene copies detected per gram of soil during the chitin enrichment kinetic. Chitin concentration is represented by darker gray nuance: light grey (1x), dark grey (10x) and white (control). Results were normalized with obtained DNA yields and expressed per gram of soil. The enrichment kinetic is represented in days. Error bars are corresponding to biological replicates from the 3 microcosms bottles set for each enrichment condition.

ii. rrs qPCR

Quantitative PCR based enumeration of *rrs* gene copy number was used to give an estimate of the bacterial biomass in the microcosms in order to assess the impact of the chitin amendment on this quantitative biological parameter. The microcosm effect was noticed again, as revealed by variations in the control microcosms with an initial drop within the first three days before a slow recovery to reach the initial level between day 6 and 20 days (Fig.2). Presence of chitin is also a source of bias for determining the *rrs* gene copy number as evidenced at t=0 with a significant difference in 10x chitin amended microcosm in comparison to control or 1x amended samples when the theoretical values should be identical. All along the kinetics, 1x chitin treatment displayed almost the same trend as the control, except at day 3 were an increase was noticed. However, a clear bio-stimulating effect of chitin was noticed in the 10x chitin amended microcosms soon after the amendment with a gradual increase all along the 35 days, except at day 20 were all conditions seems to be more or less at the same level. In reference to these previous analyses, only DNA solutions extracted from soil samples incubated for 20 days (0x20, 1x20 and 10x20) as well as the DNA from non incubated control soil (0x0) were selected for additional analyses using more sensitive approaches including hybridization of *rrs* gene amplified PCR products on a phylochip and pyrosequencing of the metagenome DNA.

iii. Phylochip results

Cluster analysis of phylochip results based on hybridization signals at the genus level indicated that the bacterial community from the initial control soil (0x0, not incubated) would be more closely related to the 1x20 sample than to the incubated control sample 0x20 (Fig.3). Only, the 10x chitin amended sample was clearly separated, revealing a chitin effect which was confirmed by phyla occurrence comparison with the detection of Euryarchaeota and Verrucomicrobia related bacteria (Table 1). In addition, significantly more bacterial genera were detected in chitin amended samples, whatever the concentration used (n=104 and n=131 for 1x20 and 10x20 respectively) than in the two control soils (n=81 and n=73 for the 0x0 and 0x20, respectively). The 10x20 chitin concentration also revealed the strongest hybridization signal concomitantly with the highest number of genera detected for the γ-proteobacteria class, mostly related to the Xanthomonadale order. A strong β-proteobacteria class signal specifically detected in the 10x20 chitin amended soil reflects the abundance of *Burkholderia* genus representatives as the specific signal for this genus was also among the strongest on the microarray.

The signal proportion of Firmicutes, Bacteroidetes and Cyanobacteria at the class level did not seem strongly affected by the treatments although at the genus level chitin enrichment led to detect respectively 16 (4+7+5), 3 and 3 new genera in comparison to the controls (Tab.1) All the genera detected in incubated control 0x20 were also detected in the 3 other samples, while only 3 genera were detected only under the initial control 0x0. On the other hand, 8 and 34 genera were only identified in 1x20 and 10x20 respectively, mostly γ-proteobacteria and Firmicutes.

Bacterial groups	A) Number of genera detected				B) Unique genera detected				
	0x0 (n=81)	0x20 (n=73)	1x20 (n=104)	10x20 (n=131)	0x0 (n=3)	0x20 (n=0)	1x20 (n=8)	10x20 (n=34)	1x20-10x20 (n=17)
α-proteobacteria (Class)	22	18	25	25	3		2	3	2
γ-proteobacteria (Class)	9	7	17	29				11	9
Actinobacteria	10	11	12	15			1	4	
Firmicutes	11	9	19	22			4	7	5
Bacteroidetes	5	5	5	8				3	
Acidobacteria	2	2	2	2					
β-proteobacteria (Class)	2	2	4	4			1	1	1
Cyanobacteria	8	7	8	11				3	
Chloroflexi	2	2	2	2				1	
Planctomycetes	2	2	2	2					
Crenarchaeota	3	3	3	3					
Chlorobi	3	3	3	3					
δ-proteobacteria (Class)	1	1	1	2				1	
Aquificae	1	1	1	1					
Euryarchaeota				1				1	
Verrucomicrobia				1				1	

Tab.1: A) Relative abundance of bacterial groups detected on phylochips and B) numbers of unique genera observed. Bacterial groups are given at the phylum level, except for Proteobacteria which are detailed at the Class level. Values in panel A are the direct counts of genera in each bacterial groups, while numbers in panel B represent the direct counts of unique genera. n values in panel A represent the total genera detected per condition.

Code	Extraction protocole	Sampling date	Depht (cm)	Reference
Roth-F2a	Indirect MP Bio1O1	February 2009	0-21	Delmont et al. 2012
Roth-Fb2	Indirect MP Bio1O1	February 2009	0-21	
Roth-F3	Indirect lysis in plug	February 2009	0-10	
Roth-F4	Indirect DNA Tissue	February 2009	0-10	
Roth-F5	Indirect Gram positive	February 2009	0-10	
Roth-F6	Indirect lysis in plug	February 2009	0-10	
Roth-J4	Indirect DNA Tissue	July 2009	0-10	
Roth-F1	Direct MP Bio1O1	February 2009	0-21	
Roth-J1	Direct MP Bio1O1	July 2009		
Roth-J2	Direct MP Bio1O1	July 2010		
Roth-J7	Direct MoBio	July 2009		
0x0a 0x0b	Direct MP Bio1O1	July 2010	0-21	Delmont et al. 2012
0x20a 0x20b 1x20a 1x20b 10x20a 10x20b	Direct MP Bio1O1	July 2010	0-21	This study

Tab.2: Description of all the metagenomes used in this study. The DNA extraction protocols are referring to different strategies including a wide range of approach to extract and lyse cells. Direct approaches are relying on in situ cell lysis within the soil sample, while indirect approaches are relying on bacterial cell separation from the sample matrix before performing the lysis.

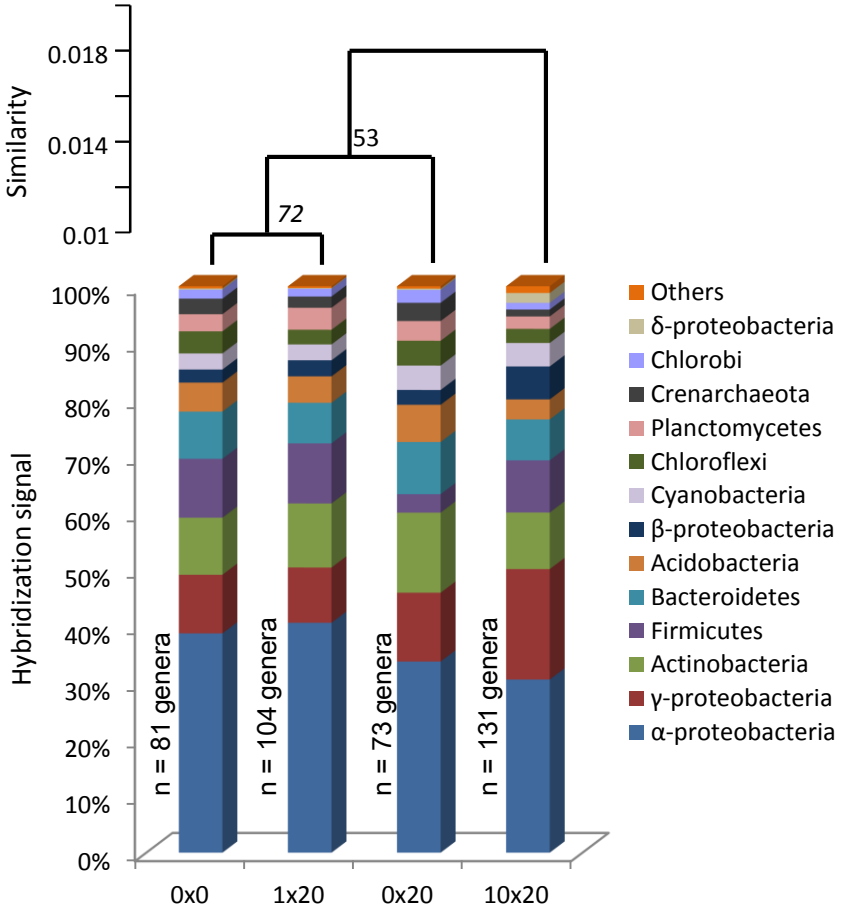


Fig.3: Relative abundance of major bacterial groups detected on rrs phylochips. Stacked columns represent the percentage of total fluorescence signal detected in each bacterial group, based on probe hybridizations. The number of identified genera per condition is indicated on the side of stacked columns. Clusters were calculated at the genus distribution level and exposed to bootstrap simulation (n=10000) and approximately unbiased bootstrap p-value (AU) were computed by multi-scale bootstrap re-sampling. AU are expressed in percentage and indicated at each nod. The grouping method is based on variance analysis (Ward 1963) and distances are calculated according a correlation algorithm (the complement 1-r of Pearson's r correlation). Ladder on the left side of the dendrogram represents similarity based on the Euclidean distance.

iv. Pyrosequencing analysis

For comparison purposes, 13 pyrosequencing runs obtained from the same soil under different conditions were integrated to the analysis to reinforce the statistical comparison (Delmont *et al.* 2012). These metagenomes are representing the metagenomic database from the Park Grass soil at Rothamsted research station in UK (Roth) (Vogel *et al.* 2009), integrating the actual variation imputed to seasonal, depth and DNA extraction protocols (Tab.2). All annotated metagenomes used in this study were submitted to a clusterization-based method in order to hierarchically organize the factors structuring the taxonomic profile of bacterial communities, both at the functional and the taxonomical level. As the two dendrograms from functional and taxonomical affiliations were similar, a summarized version is presented in figure 4.

The method employed for DNA extraction is clearly responsible for the first dichotomy, with on one side metagenomes extracted by the "indirect" approach (bacterial cell separation from the soil matrix before lysis) and on the other side those obtained by a direct technique (*in situ* lysis of bacterial cells prior to DNA purification). This last group includes the various microcosm soil extracted metagenomes (0x20, 1x20 and 10x20), the 2 control metagenomes (0x0) corresponding to untreated former soil samples, and the 4 metagenomes extracted by a direct protocol. The processing of the soil in microcosms is responsible for a second dichotomy level among all the directly extracted metagenomes, highlighting the strong microcosm effect detected earlier. Finally, at a third level, chitin enriched metagenomes 10x20 were separated from the 2 other microcosms conditions (0x20) and (1x20). Duplicated metagenomes clustered adequately in the case of the 10x20 conditions with the 2 replicates as the closest neighbors, while those encompassing 1x20 chitin amended soil and incubated control soil 0x20 displayed a somewhat similar pattern with experimental conditions more closely related than duplicates, indicating that these 2 metagenomes are very similar.

All the conditions used in this study, including "Roth", "0x0", "0x20"and "10x20" were statistically compared together through ANOVA, except for the 1x20 metagenomes as only few differences were observed between this condition and the incubation control 0x20 (Fig.4). Three level of analysis were considered in keeping of all the conditions available: (1) analysis at *sensu stricto*, which includes the incubation control 0x20 and the chitin enriched 10x20; (2) analysis at *sensu medio*, including the non-treated samples 0x0 for investigation of the microcosm effect; and (3) analysis at *sensu lato*, considering all the metagenomes available in the Rothamsted database (Roth). Full tables of read annotation are given in supplemental data respectively for: the taxonomical affiliation at the phylum level (S1) and the functional affiliation to SEED subsystems (S2). Statistical comparison of the metagenomes at the different analytic levels (*sensu stricto*, *sensu medio* and *sensu lato)* is presented respectively at: the genus level (S4) and the lowest functional level from SEED (S5). A list of all genera selected by chitin enrichment is given in table 3. A summarized list of some of the genera decreased by chitin enrichment is also proposed (S3).

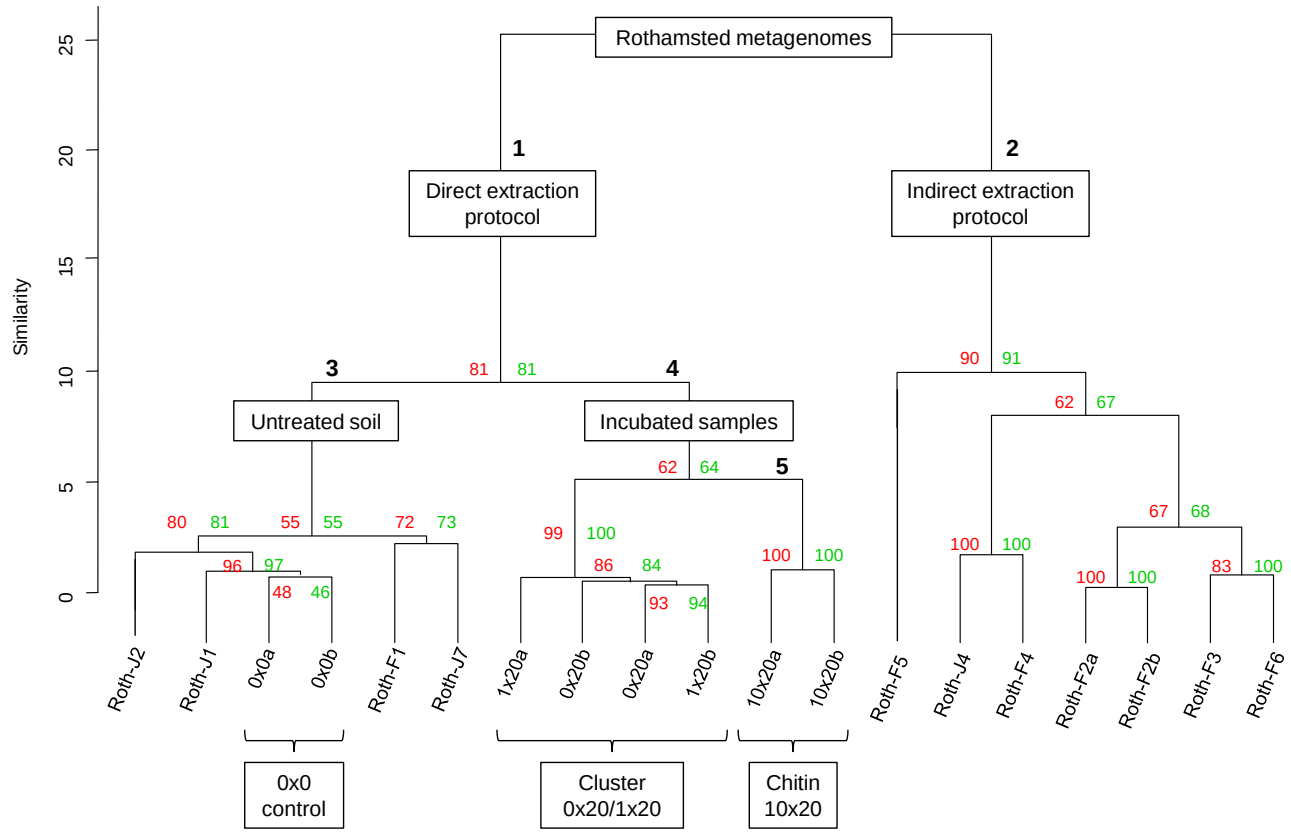


Fig.4: Hierarchical classification of the metagenomes used in this study. The dendrogram was established based on taxonomical and functional annotation of reads on M5NR through MG-RAST. Clusters were exposed to bootstrap simulation (n=10000) and approximately unbiased bootstrap p-value (AU) were computed by multi-scale bootstrap re-sampling. AU are expressed in percentage and indicated at each nod. The red and green values are respectively representing the bootstraps values obtain with the taxonomical and functional affiliation. The grouping method is based on variance analysis (Ward 1963) and distances are calculated according to a correlation algorithm (the complement 1-r of Pearson's r correlation). Ladder on the left side of the dendrogram represents similarity based on the Euclidean distance.

3. Discussion

i. Factors structuring metagenomic studies of soil bacterial communities

The integrative approach that was applied in this study permitted to classify the driving factors impacting the bacterial community structure. Based on the pyrosequencing data, the DNA extraction approach was found to generate the main source of variation between all investigated metagenomes, prior to any other factors, including sampling season and depth, microcosm effect and chitin treatment. These observations are coherent with previous published studies, already pointing out the importance of the DNA extraction procedure in microbial ecology studies (Inceoglu *et al.* 2010). This is mostly due to the ability for some species to resist the membrane disruption treatment, which can differ depending on the type of DNA extraction protocol applied (Frostegård *et al.* 1999). Furthermore, this effect has already been demonstrated for the Rothamsted soil in previous studies (Delmont *et al.* 2011, Delmont *et al.* 2012).

The second strongest bacterial community structuring force in this study was how the soil was processed. All the metagenomes extracted from incubated soil samples (0x20, 1x20 and 10x20) clustered apart from the former untreated soil (0x0), highlighting a "microcosm effect". This is coherent with *rrs* qPCR and enzymatic assays which already pointed out the impact of incubation on the bacterial biomass and chitinase activities. A critical initial decrease of both exo- and endo- chitinase activity was noticed for all conditions directly after incubation at day 3. We hypothesized that incubation temperature and soil moisture changes were responsible, considering that the incubation temperature (24°C) was significantly higher than the average one in the field (15°C). In fact, temperature is known to impact the chitinolytic activity (Vorob'ev *et al.* 2007, Manucharova *et al.* 2011). In addition, chitin amendment resulted in a clear decrease under the 10x condition noticed by *rrs* qPCR. This would rather be due to an interference of colloidal chitin during the DNA extraction step, resulting in lower yields recovery in comparison to the control or the lower chitin concentration. Interestingly, the faster recovery occurred under the strongest chitin concentration and could be due to an increase in bacterial growth favored by chitin degradation. This is consistent with the enzymatic assays that show a clear endochitinase activity peak between day 3 and 20. However, no exo-chitinase activity peak was detected in spite of a higher activity level for the 10x condition. This is in accordance with the usual chitin degradation pathway, first involving endochitinase for primary hydrolysis, progressively releasing the substrate for exochitinase (Ren et al 2000).

The impact of incubation on bacterial community structures has been already reported, and similarly we found that α-proteobacteria were stimulated but we did not confirmed the stimulation of Acidobacteria, as reported previously (Thomson *et al.* 2010). However, our study clearly indicates that the main change in the bacterial community was the increase of Actinobacteria due to microcosm conditions (Fig.5). This is probably due to an activation of dormant cells after addition of water which stimulates spore germination and hyphae growth (Jankins *et al.* 2009).

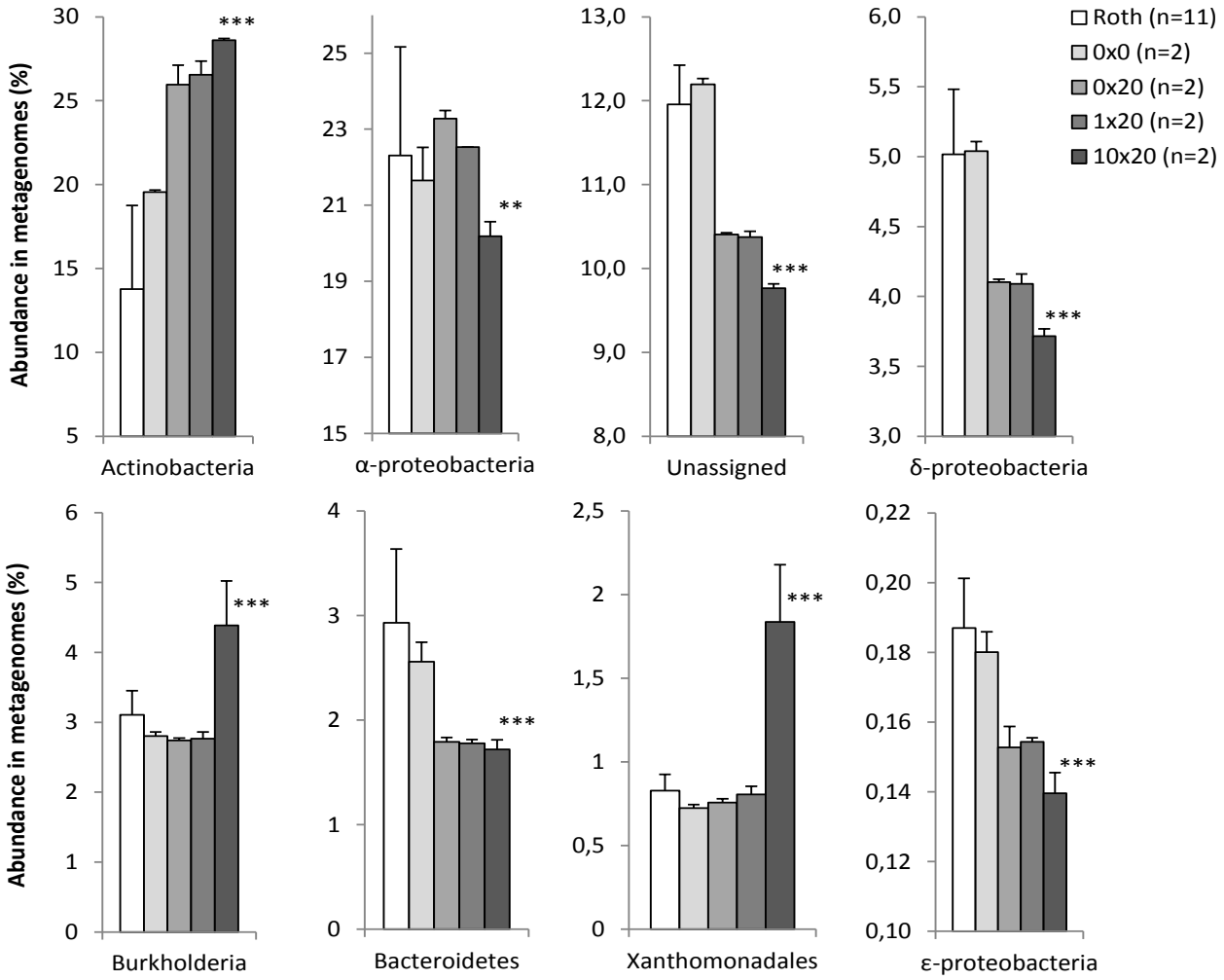


*Fig.5: Major changes in the soil bacterial community after incubation and chitin enrichment. The graphics represent the abundance of different bacterial groups in percentage of sequences in metagenomes. Grey nuances represent the different conditions and the number of metagenome used is given in brackets. ANOVA p-values were calculated from the multiple groups comparison, with *** for significant p-value at sensu lato (against Roth) and ** for significant p-value at sensu medio (against 0x0) (Actinobacteria = 5.64E-04, α-proteobacteria = 0.012; Unassigned = 5.32E-03; δ-proteobacteria = 1.84E-04; Burkholderia = 9.91E-03, Bacteroidetes = 0.029; Xanthomonadales = 1.71E-06 and ε-proteobacteria = 5.30E-04)*

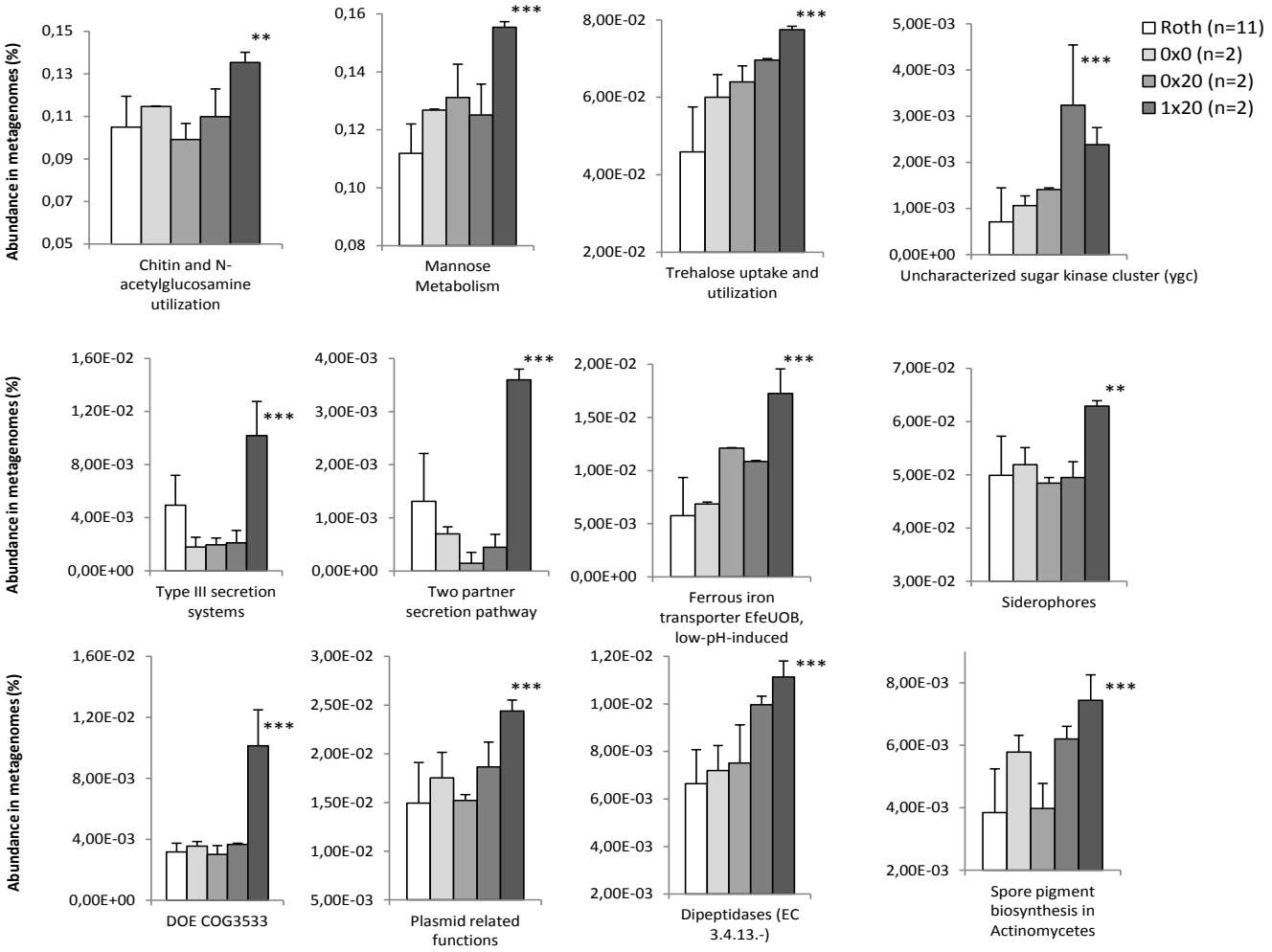


*Fig.6: Major changes in the soil bacterial community at the functional subsystem level. The graphics represent the abundance of different functional subsystems in percentage of sequences in metagenomes. Grey nuances represent the different conditions and the number of metagenome used is given in brackets. ANOVA p-values were calculated from the multiple groups comparison, with *** for significant p-value at sensu lato (against Roth) and ** for significant p-value at sensu medio (against 0x0) (Chitin and N-acetylglucasamine utilization = 0.016; Mannose metabolism = 5.07E-04, Threhalose uptake and utilization = 3.85E-03, Uncharacterized sugar kinase cluster (ygc) = 0.01; Type III secretion systems = 4.30E-03; Two partner secretion patway = 3.53E-03; Ferrous transporter Efe UOB, low pH-induced = 3.80E-03, Siderophore = 0.042, DOE COG 3533 = 6.11E-07; Plasmid related functions = 0.05; Dipeptidase =3.45E-03 and Spore pigment biosynthesis in Actinomycetes = 0.018)*

ii. Global impact of two chitin concentrations on soil bacterial communities

Nevertheless, the "microcosm effect" and the DNA extraction biases did not compromise the interest of the approach for exploring the changes occurring in the bacterial community after chitin enrichment. Indeed, the effect of chitin treatment is classified as the third force impacting on the bacterial community, based on the hierarchical classification (Fig.4). At this stage, the lowest chitin concentration was not found to significantly modify the soil bacterial community structure. This is consistent with the other techniques where only small differences were observed in spite of the biostimulant effect observed at day 3 by *rrs* qPCR and the diversity increase on the *rrs* phylochip. However, this concentration was the one usually tested in previous studies (Inglis and Kawchuk, 2002). The slight differences observed could rather be due to inherent specificities of the soil used than a perturbation of the chitin degradation potential under microcosm conditions as shown by the enzymatic assays. On the other hand, the 10x chitin enrichment resulted in a clear bio-stimulation of the bacterial community, as shown by *rrs* qPCR, and the concomitant detection of an increased number of genera by hybridization on *rrs* phylochips.

iii. Metagenomic analysis at sensu stricto with 0x20 incubation control

The only relevant control for detecting the chitin impact was the metagenomic condition 0x20 corresponding to the soil incubated in the same conditions as the chitin treated 10x20. Analysis of the metagenome sequences indicated that the relative proportion of Actinobacteria, γ-proteobacteria and β-proteobacteria was found to be increased in 10x20, while Planctomycetes, Cyanobacteria, α- and δ-proteobacteria were reduced (S1 and Fig.5). Unfortunately, the changes in the community could not be statistically supported every time because of the low repetition number. Cross comparison revealed 35 genera mostly related to γ -proteobacteria and Actinobacteria, as well as 91 genera (none Actinobacteria) significantly enriched respectively in 10x20 metagenomes and incubation control 0x20 (S4). At this level, these observations rely on changes in the relative distribution of sequences from quantitatively constant DNA solutions and cannot discriminate between a negative biological effect of chitin on some genera or a simple mathematic compensation after the growth of the 35 genera positively affected by chitin. Similar trends were also detected at the functional level with 412 and 244 SEED functions widely distributed in all subsystems whose detection was increased in 10x20 and 0x20 respectively (S5).

iv. Metagenomic analysis at sensu medio with 0x0 non-treated control

When analyzing metagenomes not restricted to the incubated microcosms but including the non incubated control soil (0x0, *sensu medio*) the so-called microcosm effect is strongly highlighted. The strong increase of Actinobacteria and at a lesser extent of α-proteobacteria in incubated control 0x20 in comparison to 0x0 led to a proportional decrease of most of the other phyla (Fig.5). This was also noticed at lower taxonomical level, with a selection of 91 genera mostly related to Actinobacteria and α-proteobacteria in 0x20 (S4). The same trend was also detected at the functional level, with a clear effect of incubation with 458 and 206 functions whose occurrence was decreased and increased in 0x20, respectively (S5). However, addition of this control condition (0x0) to the initial one (0x20) increased the estimated impact of chitin on the community structure with 71 and 182 genera whose occurrence was found to be increased and decreased in 10x20 (S4).

Similarly at the functional level, 430 and 418 functions were respectively augmented and reduced in 10x20 (S5). In addition, cross-comparison between the two controls 0x20 and 0x0 helps in identifying chitin impacted genera and function out of the microcosm effect. Overall, 48.6% (88/181) and 70.4% (50/71) of the genera, as well as 30.7% (132/430) and 38.3% (160/418) of the functions already impacted at *sensu stricto* by incubation, while the remaining part can be specifically attributed to chitin treatment.

v. Metagenomic analysis at sensu lato with Rothamsted database

In a new step, *sensu lato* the statistical analysis was carried out by integrating the 4.88 x 10^9 bp of metagenomic sequence data already available for this soil as a reference control (Delmont *et al.* 2012), considering that the wide range of sampling and methodological conditions they represent could help to specifically identify taxa and functions affected by chitin amendment. Addition of the 11 metagenomes of the Rothamsted database increases significantly the robustness of the statistical analysis (n=11), strengthening some relevant changes and invalidating other ones. For instance, this analytical level did not confirm the previous result obtained at lower analytical levels of an increase of α-proteobacteria related bacteria leaving the Actinobacteria phylum to be the only one whose representative occurrence was increased after chitin treatment (Fig.5).

Interestingly, the number of positively affected genera increased significantly under these last analytical conditions (from 35, 71 and 94 at respectively, *sensu stricto, medio* and *lato* analytical levels) indicating that a drop in the background contributed to reveal a wider range of putative chitin degraders masked when the analysis lacks stringency (S4). The number of bacteria phyla whose occurrence was decreased by the strongest chitin amendments (16, S1) was significantly reduced in comparison to the 2 previous analysis levels including Proteobacteria through δ- and ε-proteobacteria classes, as well as Bacteroidetes, Chlorobi, Spirochaetes, Aquificae, Fusobacteria, Fibrobacteres, Elusimicrobia, Deferribacteres, Tenericutes, Chrysiogenetes, Candidatus Poribacteria, Nanoarchaeota and Chrysiogenetes. In addition, this step by step analytical procedure allowed to dissociate microcosm incubation and chitin effects by specifically identifying in the detected taxa those already positively (63 out of 94 genera, 67%) and negatively (55 out of 116 genera, 47.7%) influenced by incubation conditions. The increase of analytical stringency related to the Rothamsted database also decreased the number of functions whose occurrence was found to be negatively affected by microcosm incubation and chitin amendment (respectively, 361 in comparison to 458 and 343 in comparison to 418 at *sensu medio*, S5). Interestingly, the number of functions whose occurrence is scaled up by microcosm incubation (808) and chitin amendment (1138) is increased at *sensu lato* in comparison to previous ones (*sensu stricto* and *sensu medio*). Again, the specific chitin effect can easily be distinguished considering that the microcosm effect was responsible for only 28.9% (329/1138) and 35% (120/343) of the total number of functions detected.

Bacterial group	Genus level	CAZy related chitinases	References related to chitin/cellulose degradation	Chitin concentration	ANOVA p-value *(sensu lato)*
Actinobacteria (n=46)	*Acidimicrobium*	GH18	NA	1x	1,47E-03
	Conexibacter	GH18 CE4	NA	1x	1,74E-02
	Lechevalieria	NA	NA	1x	1,84E-03
	Okibacterium	NA	NA	1x	4,91E-02
	Pimelobacter	NA	NA	1x	3,41E-03
	Saccharomonospora	GH20 CE4	Steger et al. (2007)	1x	9,80E-04
	Actinoplanes	GH18 GH19 GH20 CE4 CMB33	Nawani and Kapadnis (2003)	10x	1,65E-03
	Actinosynnema	GH18 GH19 GH20 CE4 CMB33	NA	10x	4,38E-04
	Aeromicrobium	NA	Lu et al. (2005)	10x	9,80E-05
	Amycolatopsis	GH18 GH19 GH20 CE4 CMB33	Saito et al. (2009)	10x	8,20E-04
	Arthrobacter	GH18 GH20 CE4	Mavromatis et al. (2003)	10x	4,11E-03
	Beutenbergia	GH20 CE4	NA	10x	2,22E-03
	Catenulispora	GH18 GH19 GH20 CE4 CMB33	NA	10x	2,29E-06
	Cellulomonas	GH18 GH19 GH20 CE4 CMB33	Reguera and Leschine (2003)	10x	1,73E-03
	Cystobacter	NA	NA	10x	5,40E-02
	Dactylosporangium	NA	NA	10x	2,30E-02
	Dermacoccus	NA	NA	10x	7,20E-04
	Gordonia	CE4	NA	10x	1,53E-03
	Intrasporangium	CE4	NA	10x	6,28E-04
	Janibacter	NA	NA	10x	6,45E-04
	Jonesia	GH18 GH20 CE4 CBM33	NA	10x	3,84E-03
	Kineococcus	GH20	NA	10x	1,12E-03
	Kitasatospora	GH19	Hjort et al. 2009	10x	4,44E-05
	Kocuria	NA	NA	10x	4,83E-03
	Kribbella	GH18 GH19 GH20 CE4 CMB33	NA	10x	2,52E-04
	Micrococcus	NA	Cervantes-González et al. 2008	10x	4,82E-03
	Micromonospora	GH18 GH19 GH20 CE4 CBM33	Gacto et al. (2000)	10x	9,75E-04
	Mycobacterium	GH18 GH19 CE4	Varrot et al. (2005)	10x	3,92E-02
	Nakamurella	GH18 CE4	NA	10x	3,88E-03
	Nocardia	CE4	Nanjo et al. (1990)	10x	1,49E-03
	Nocardioides	GH18 CE4	Nawani and Kapadnis (2003)	10x	7,26E-05
	Propionibacterium	GH18 GH20 CE4	NA	10x	1,60E-03
	Renibacterium	GH18 GH20 CE4 CMB33	NA	10x	5,79E-03
	Rhodococcus	GH19 CE4	Steger et al. (2007)	10x	1,80E-03
	Rothia	CE4	NA	10x	3,38E-03
	Saccharopolyspora	GH18 GH20 CE4 CMB33	Nawani and Kapadnis (2003)	10x	1,14E-03
	Salinispora	GH18 GH19 GH20 CE4 CBM33	NA	10x	1,23E-03
	Sanguibacter	GH18 GH20 CMB33	Han et al. 2010	10x	3,73E-03
	Stackebrandtia	GH18 GH19 GH20 CBM33	NA	10x	1,84E-03
	Streptomyces	GH18 GH19 GH20 CE4 CMB33	Hjort et al. 2009	10x	3,23E-04
	Streptosporangium	GH18 GH19 GH20 CE4 CBM33	Manucharova et al. (2004)	10x	7,23E-04
	Thermobispora	GH18 GH20 CE4 CMB33	NA	10x	1,42E-03
	Thermoleophilum	NA	NA	10x	4,91E-02
	Thermomonospora	CE4 CMB33	Anish et al. (2007)	10x	1,02E-03
	Williamsia	NA	NA	10x	4,91E-02
	Xylanimonas	GH18 GH20 CE4 CMB33	NA	10x	2,10E-03
γ-proteobacteria (n=10)	***Thioalkalimicrobium***	NA	NA	1x	4,91E-02
	unclassified	NA	NA	1x	4,91E-02
	Volucribacter	NA	NA	1x	4,91E-02
	Ectothiorhodospira	NA	NA	10x	1,81E-02
	Lysobacter	GH18	Cervantes-González et al. 2008	10x	1,45E-05
	Pseudoxanthomonas	GH20 CE4	Someya et al. 2011	10x	7,73E-05
	Stenotrophomonas	GH18 GH19 GH20 CE4 CMB33	Okeke and Lu, 2011	10x	1,04E-04
	Thiocystis	NA	NA	10x	2,28E-02
	Xanthomonas	GH18 GH19 GH20 CE4	Yamaoka et al. 1999	10x	9,48E-07
	Xylella	GH18 GH20 CE4	Killini et al. 2010	10x	1,89E-08
β-proteobacteria (n=3)	*Tepidimonas*	NA	NA	1x	9,41E-03
	Burkholderia	GH18 GH19 GH20 CE4 CMB33	Ogawa et al. 2002	10x	1,46E-03
	Zoogloea	NA	NA	10x	9,91E-04
Firmicutes (n=3)	***Thermacetogenium***	NA	NA	1x	4,91E-02
	Cohnella	NA	Eida et al. (2012)	10x	4,95E-02
	Seinonella	NA	NA	10x	4,91E-02
α-*proteobacteria*	***Rhodocista***	NA	NA	10x	4,91E-02
δ-*proteobacteria*	***Geopsychrobacter***	NA	NA	1x	4,91E-02
Bacteroidetes	*Terrimonas*	NA	NA	1x	2,98E-02

Tab.3: Summarized table of the bacterial genera selected by chitin enrichment after comparison with all metagenomic conditions. The presence of characterized enzymes related to chitin degradation in CAZy for each genus is given, as well as references that reported chitinase/cellulase activity. We reported cellulase activity as some of the characterized chitinase in CAZy were reported to have a catalytic activity toward cellulose. The chitin concentration where the genus abundance was found to be the highest is also given (1x and 10x). Multi-group ANOVA p-value were obtained from comparison of all the metagenomic condition with the Rothamsted database (sensu lato) (significance threshold ≤0.05)

vi. Taxonomic changes driven by chitin in soil bacterial communities.

This step by step analysis led to identify 53 and 13 bacteria genera whose occurrence was significantly increased by, the strongest and the lowest chitin concentration respectively (Tab.3). The list includes γ-proteobacteria genera, mostly due to the rising of Xanthomonadales related genera known to harbor chitin and cellulose degrading representatives such as *Xanthomonas (*Yamaoka *et al.* 1999), *Xylella* (Killini *et al.* 2010), *Lysobacter* (Cervantes-González *et al.* 2008), *Pseudoxanthomonas* (Someya *et al.* 2011) and *Stenotrophomonas* (Okeke and Lu, 2011). The strong rising of *Burkholderia* related species; also known for their chitin degrading potential (Ogawa *et al.* 2002) explain the increase of β-proteobacteria. In spite of the strong microcosm effect, a selection occurred among Actinobacteria, with only 37.7% (46/122) of the Actinobacteria related genera detected in this study significantly increased in chitin enriched microcosms. This was confirmed by the decrease of the hybridization signals on *rrs* phylochips corresponding to this phylum, concomitantly to the identification of 4 unique genera. The same selection phenomenon was also observed with other taxa containing representatives known to possess a genetic potential for complex biopolymer degradation including the δ-proteobacteria related genera *Geobacter* (GH19, GH20, CE4) (Ren *et al.* 2009), *Pelobacter* (GH19, CE4) (Kiely 2011), *Desulfovibrio* (GH18, CE4) (Kan *et al.* 2011) and *Haliangium* (GH18, GH20, CE4) whose occurrence are lower under chitin rich soil conditions. However, as these genera are usually originated from water and sediment environments, we hypothesized for a habitat related specialization of the chitin degrading genetic potential among members.

The double approach used is this study to determine microbial community changes including phylochips and high throughput sequencing yielded coherent results. However, the taxonomic microarray, relying on PCR amplification of a target gene was more sensitive than pyrosequencing and was the only technique to detect the minor community changes occurring under the lowest chitin concentration with ten unique bacteria genera specifically detected. A stronger impact was detected with the 10x concentration on the soil bacterial community at the taxonomic and functional levels, showing higher taxa specialization.

vii. New potential chitin degraders in the enriched soil microcosms.

The main objective of such an enrichment approach combined with a multi-step metagenome analysis was to determine if chitin degraders other than those already isolated could be enhanced in the soil matrix either for subsequent isolation attempts or direct cloning of their catabolic genes. The study identified several candidate genera with an occurrence in soil significantly increased by the chitin enrichment but never reported in the literature for being related to chitin degradation, with also no or little evidence in CAZy database (Tab.3). Several genera related to the *Micrococcineae* a sub-order of the Actinobacteria were detected under chitin rich conditions including *Renibacterium, Micrococcus, Janibacter, Sanguibacter, Jonesia, Beutenbergia, Kocuria, Cellulomonas, Dermacoccus, Intrasporangium, Xylanimonas, Rothia* and *Arthrobacter*. The list also includes *Cohnella* (Firmicutes), *Ectothiorhodospira* and *Thiocystis* (γ-proteobacteria) while genera *Lechevalieria* and *Pimelobacter* (Actinobacteria), *Terrimonas* (Bacteroidetes) and *Tepidimonas* (β-proteobacteria) belong to the group of microorganisms responding only at the lowest chitin concentration.

The enrichment approach reveals genera that could be classified as members of the soil rare biosphere according to their very low occurrence in all available soil metagenomes including *Seinonella* (Firmicutes), *Williamsia*, *Thermoleophilum* and *Rhodocista* (Actinobacteria) in the 10x20 metagenomes, as well as *Okibacterium* (Actinobacteria), *Geopsychrobacter* (δ-proteobacteria), *Thermacetogenium* (Firmicutes), *Volucribacter*, *Thioalkalimicrobium* and an unclassified Xanthomonadaceae (γ-proteobacteria) in 1x20 metagenomes. Their detection under the enriched conditions indicates that their growth was critically enhanced by chitin, strengthening the hypothesis of an increased fitness related to an expression of the corresponding catabolic genes. They also contribute to the list of the candidates for new enzymes since none of these genera is known to contain chitin degraders.

viii. Impact of enrichment at the functional level

At the functional level, the differences observed are more subtle than the ones detected at the taxonomical level (S2). In fact, the global functional profiles of metagenomes are not strongly impacted by incubation and chitin enrichment, with poor differences observed at the highest SEED level. This indicates that the differences observed at the taxonomical level are not disturbing the overall functional picture of the metagenomes. However, at lower level, specific features are selected by the different chitin concentrations (Fig.6). For instance, the "Chitin and N-acetylglucosamine" subsystem in the chitin amended metagenomes is naturally increased in comparison to the controls indicating that expression of these genes confers an increased fitness to at least some of their hosts. The increase of several carbon metabolism related subsystems also indicates a bio-stimulating effect on the microbial community. The strong increase of Actinobacteria is also detected at this stage, with for instance a significant increase of the subsystem related to “Spore pigment biosynthesis in Actinomycetes”. More surprisingly, some of the secretion systems respond positively to the enrichment, concomitantly to an unknown secreted and conserved protein referenced as DOE COG3533. This protein is apparently affiliated to plant-bacteria interactions, with similarities to glycosyl hydrolase enzymes according to the SEED classification. In addition, strong chitin concentrations leads to an enrichment of bacteria expressing the iron metabolism as revealed by the occurrence increase of the low pH induced iron transporter EfeUOB and siderophore synthesis subsystems. In fact, iron metabolism is known to be stimulated under aerobic condition, as it is a limiting factor for primary metabolic processes (Giessen *et al.* 2012). Plasmid related functions subsystem was also increased, including mostly “Plasmid-encoded T-DNA transfer” related functions.

Are these functions responsible for the selection of their host directly or as side effect considering that full degradation of chitin fibers involves several interacting species (Jagmann *et al.* 2011) or do they only reveal growth of bacterial cheaters developing on chitin degradation products (Allison, 2005). Additional experiments need to be carried out to improve these observations.

4. Conclusion

Enrichment in microcosm is a well known strategy to enhance or reduce the proportion of micro-organisms reacting to the treatment (Brodie *et al.* 2011). Clearly, the selection operated toward chitin degrading organisms confirming the interest of an enrichment strategy for enhancing their proportion and that of their genes in the metagenome. However, these results highlight the importance of chitin concentration for finding the best compromise for an enhanced target organism exploration that would increase the chance of cloning new chitin degradation pathways without reducing the overall diversity including those of the chitin degraders masked by the most competitive strains. Several parameters were already investigating to assess chitin degradation, such as moisture (Vorob'ev *et al.* 2007), temperature and soil types (Manucharova *et al.* 2011). Our results indicate that even at the strongest chitin concentration, the overall functional diversity of the bacterial community is preserved even though some specific taxa are strongly increased.

Acknowledgements

This study was performed and funded in the frame of the European Union Metaexplore project METAEXPLORE (KBBE-222625), the French National Research Agency (Agence National de Recherche) ANR Genomique programme: METASOIL. Material and methods are provided as supplemental information (S6).

References

Allison SD (2005) Cheaters, diffusion and nutrients constraindecomposition by microbial enzymes in spatially structuredenvironments. *Ecol Lett.* 8:626–635

Anish R, Rahman MS, Rao M (2007) Application of cellulases from an alkalothermophilic Thermomonospora sp. in biopolishing of denims. *Biotechnol Bioeng*. 96:48-56

Bhuiyan FA, Nagata S, Ohnishi K (2011) Novel chitinase genes from metagenomic DNA prepared from marine sediments in southwest Japan. *Pak J Biol Sci*. 14:204-211

Brodie EL, Joyner DC, Faybishenko B, Conrad ME, Rios-Velazquez C, Malave J (2011) Microbial community response to addition of polylactate compounds to stimulate hexavalent chromium reduction in groundwater. *Chemosphere*. 85:660-665

Bruce KD, Hiorns WD, Hobman JL, Osborn AM, Strike P, Ritchie DA (1992) Amplification of DNA from native populations of soil bacteria by using the polymerase chain reaction. *Appl Environ Microbiol*. 58:3413-3416

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*. 37(Database issue):D233-238

Cervantes-González E, Rojas-Avelizapa NG, Cruz-Camarillo R, García-Mena J, Rojas-Avelizapa LI (2008) Oil-removal enhancement in media with keratinous or chitinous wastes by hydrocarbon-degrading bacteria isolated from oil-polluted soils. *Environ Technol*. 29:171-182

Cottrell MT, Moore JA, Kirchman DL (1999) Chitinases from uncultured marine microorganisms. *Appl Environ Microbiol*. 65: 2553-2557

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011a) Accessing the soil metagenome for studies of microbial diversity. Appl Environ Microbiol 77:1315-1324

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011b) Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods. 86:397-400

Delmont TO, Prestat E, Kevin P Keegan, Michael Faubladier, Patrick Robe, Ian M Clark, Eric Pelletier, Penny R Hirsch, Folker Meyer, Jack A Gilbert, Denis Le Paslier, Pascal Simonet, Timothy M Vogel (2012) Structure, fluctuation and magnitude of a natural grassland soil metagenome.*The ISME Journal*, doi: 10.1038/ismej.2011.197

Eida MF, Nagaoka T, Wasaki J, Kouno K (2012) Isolation and Characterization of Cellulose-decomposing Bacteria Inhabiting Sawdust and Coffee Residue Composts. *Microbes Environ*. [Epub ahead of print]

FAO (2006) *Guidelines for soil description*. FAO, Rome,Italy. ftp://ftp.fao.org/agl/agll/docs/guidel_soil_descr.pdf
Fierer N, Jackson JA, Vilgalys R, Jackson RB (2005) Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. *Appl Environ Microbiol*. 71:4117-4120

Gacto M, Vicente-Soler J, Cansado J, Villa TG (2000) Characterization of an extracellular enzyme system produced by Micromonospora chalcea with lytic activity on yeast cells. *J Appl Microbiol*. 88:961-967

Giessen TW, Franke KB, Knappe TA, Kraas FI, Bosello M, Xie X, Linne U, Marahiel MA (2012) Isolation, Structure Elucidation, and Biosynthesis of an Unusual Hydroxamic Acid Ester-Containing Siderophore from Actinosynnema mirum. *J Nat Prod*. 75:905-914

Gomes RC, Semedo LT, Soares RM, Linhares LF, Ulhoa CJ, Alviano CS, Coelho RR (2001) Purification of a thermostable endochitinases from Streptomyces RC1071 isolated from a cerrado soil and its antagonism against phytopathogenic fungi. *J Appl Microbiol*. 90:653-661.

Gooday GW (1990) The ecology of chitin degradation. *Adv Microb Ecol*. 11:387-430.

Graham LS, Sticklen MB (1994) Plant chitinases. *Can J Bota*. 72: 1057–1083

Griffiths RI, Whitely AS, O'Donnell AG, Bailey MJ (2000) Rapid method for coextraction of DNA and rRNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl Environ Microbiol*. 66:5488-5491

Harman GE, Hayes CK, Lorito M, Broadway RM, diPietro A, Peterbauer C, Tronsmo A (1993) Chitinolytic enzymes of Trichoderma harzianum: purification of chitobiosidase and endochitinase. *Phytopathol*. 83: 313–318

Han SJ, Park H, Lee SG, Lee HK, Yim JH (2011) Optimization of cold-active chitinase production from the Antarctic bacterium, KOPRI 21702. *Appl Microbiol Biotechnol*. 89:613-621

Horn SJ, Sikorski P, Cederkvist JB, Vaaje-Kolstad G, Sørlie M, Synstad B, Vriend G, Vårum KM, Eijsink VG (2006) Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides. *Proc Natl Acad Sci USA*. 103:18089-94.

Hjort K, Bergström M, Adesina M, Jansson J, Smalla K, Sjöling S (2009) Chitinases genes revealed and compared in bacterial isolates, DNA extracts and a metagenomic library from a phytopathogen-suppressive soil. *FEMS Microbiol Ecol*. 71: 197-207

Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 280: 309-316

Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 293: 781-788

Hobel CF, Marteinsson VT, Hreggvidsson GO, Kristjansson JK (2005) Investigation of the microbial ecology of intertidal hot springs by using diversity analysis of 16S rRNA and chitinases genes. *Appl Environ Microbiol*. 71:2771-2776

Howard M, Ekborg N, Taylor II L, Weiner R, Hutcheson S (2003) Chitinase B of "Microbulbifer degradans" 2-40 contains two catalytic domains with different chitinolytic activities. *J Bacteriol*. 186:1297-1303

Inglis GD, Kawchuk LM (2002) Comparative degradation of oomycete, ascomycete, and basidiomycete cell walls by mycoparasitic and biocontrol fungi. *Can J Microbiol*. 48:60-70

Jagmann N, von Rekowski KS, Philipp B (2011) Interactions of bacteria with different mechanisms for chitin degradation result in the formation of a mixed-species biofilm. *FEMS Microbiol Lett*. 326:69-75

Jenkins SN, Waite IS, Blackburn A, Husband R, Rushton SP, Manning DC, O'Donnell AG (2009) Actinobacterial community dynamics in long term managed grasslands. *Antonie Van Leeuwenhoek*. 95:319-334

Kan J, Wang Y, Obraztsova A, Rosen G, Leather J, Scheckel KG, Nealson KH, Arias-Thode YM (2011) Marine microbial community response to inorganic and organic sediment amendments in laboratory mesocosms. *Ecotoxicol Environ Saf*. 74:1931-1941

Kiely PD, Rader G, Regan JM, Logan BE (2011) Long-term cathode performance and the microbial communities that develop in microbial fuel cells fed different fermentation endproducts. *Bioresour Technol*. 102:361-366

Killiny N, Prado SS, Almeida RP (2010) Chitin utilization by the insect-transmitted bacterium Xylella fastidiosa. *Appl Environ Microbiol*. 76:6134-6140

Lindahl BD, Finlay RD. (2006). Activities of chitinolytic enzymes during primary and secondary colonization of wood by basidiomycetous fungi. *New Phytology*. 169: 389-397

Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem J*. 432:437-444

LeCleir GR, Buchan A, Hollibaugh JT (2004) Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment-specific distributions. *Appl Environ Microbiol*. 70: 6977-6983

LeCleir GR, Buchan A, Maurer J, Moran MA, Hollibaugh JT (2007) Comparison of chitinolytic enzymes from an alkaline hypersaline lake and an estuary. *Environ Microb*. 9:197-205

Lu WJ, Wang HT, Yang SJ, Wang ZC, Nie YF (2005) Isolation and characterization of mesophilic cellulose-degrading bacteria from flower stalks-vegetable waste co-composting system. *J Gen Appl Microbiol*. 51:353-360

Manucharova NA, Belova EV, Polianskaia LM, Zenova GM (2004) A chitinolytic actinomycete complex in black soil. *Mikrobiologiia*. 73:68-72

Manucharova NA, Vlasenko AN, Men'ko EV, Zviagintsev DG (2011) Specificity of the chitinolytic microbial complex of soils incubated at different temperatures. *Mikrobiologiia*. 80:219-229

Martin-Laurent F, Philippot L, Hallet S, Chaussod R, Germon JC, Soulas G Catroux G (2001) DNA extraction from soils: old bias for new microbial diversity analysis methods. *Appl Environ Microbiol*. 67:2354-2359

Mavromatis K, Feller G, Kokkinidis M, Bouriotis V (2003) Cold adaptation of a psychrophilic chitinase: a mutagenesis study. *Protein Eng*. 16:497-503

Metcalfe AC, Krsek M, Gooday GW, Prosser JI, Wellington EM (2002) Molecular analysis of bacterial chitinolytic community in an upland pasture. *Appl Environ Microbiol*. 68: 5042-5050

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server-a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 9:386

Muzzarelli RA (1999) Native, Industrial, and fossil chitin. *EXS*. 87:1-6

Nanjo F, Katsumi R, Sakai K (1990) Purification and characterization of an exo-beta-D-glucosaminidase, a novel type of enzyme, from Nocardia orientalis. *J Biol Chem*. 265:10088-10094

Nawani NN, Kapadnis BP (2003) Chitin degrading potential of bacteria from extreme and moderate environment. *Indian J Exp Biol*. 41:248-254

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*. 11:187

Ogawa K, Yoshida N, Kariya K, Ohnishi C, Ikeda R (2002) Purification and characterization of a novel chitinase from Burkholderia cepacia strain KH2 isolated from the bed log of Lentinus edodes, Shiitake mushroom. *J Gen Appl Microbiol*. 48:25-33

Okeke BC, Lu J (2011) Characterization of a defined cellulolytic and xylanolytic bacterial consortium for bioprocessing of cellulose and hemicelluloses. *Appl Biochem Biotechnol.* 163:869-881
Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*. 26:715-721

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/

Ramaiah N, Hill RT, Chun J, Ravel J, Matte MH, Straube WL, Colwell RR. (2000). Use of chiA probe for detection of chitinases genes in bacteria from Chesapeake Bay. *FEMS Microbiol Ecol*. 34: 63-71

Ranjard L, Brothier E, Nazaret S (2000) Sequencing bands of ribosomal intergenetic spacer analysis fingerprints for characterization and microscale distribution of soil bacterium populations responding to mercury spiking. *Appl Environ Microbiol*. 66:5334-5339

Reguera G, Leschine SB (2003) Biochemical and genetic characterization of ChiA, the major enzyme component for the solubilization of chitin by Cellulomonas uda. *Arch Microbiol*. 180:434-443

Ren Y, Wee KE, Chang FN (2000) Deficiency of current methods in assaying endochitinase activity. *Biochem Biophys Res Commun*. 268:302-305

Ren Z, Steinberg LM, Regan JM (2008) Electricity production and microbial biofilm characterization in cellulose-fed microbial fuel cells. *Water Sci Technol*. 58:617-622

Saito A, Ooya T, Miyatsuchi D, Fuchigami H, Terakado K, Nakayama SY, Watanabe T, Nagata Y, Ando A (2009) Molecular characterization and antifungal activity of a family 46 chitosanase from Amycolatopsis sp. CsO-2. *FEMS Microbiol Lett*. 293:79-84

Sanguin H, Remenant B, Dechesne A, Thioulouse J, Vogel TM, Nesme X Moënne-Loccoz Y, Grundmann GL (2006) Potential of a 16S rRNA-based taxonomic microarray for analyzing the rhizosphere effects of maize on Agrobacterium spp. and bacterial communities. *Appl Environ Microbiol*. 72:4302-4312

Someya N, Ikeda S, Morohoshi T, Noguchi Tsujimoto M, Yoshida T, Sawada H, Ikeda T, Tsuchiya K (2011) Diversity of culturable chitinolytic bacteria from rhizospheres of agronomic plants in Japan. *Microbes Environ*. 26:7-14

Steger K, Jarvis A, Vasara T, Romantschuk M, Sundh I (2007) Effects of differing temperature management on development of Actinobacteria populations during composting. *Res Microbiol*. 158:617-624

Thomson BC, Ostle NJ, McNamara NP, Whiteley AS, Griffiths RI (2010) Effects of sieving, drying and rewetting upon soil bacterial community structure and respiration rates. *J Microbiol Methods*.83:69-73

Uchiyama T, Watanabe K (2006) Improved inverse PCR scheme for metagenome walking. Biotechniques. 41: 183-188

Varrot A, Leydier S, Pell G, Macdonald JM, Stick RV, Henrissat B, Gilbert HJ, Davies GJ (2005) Mycobacterium tuberculosis strains possess functional cellulases. *J Biol Chem*. 280:20181-20184

Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, Van Elsas JD, Bailey MJ, Nalin R, Philippot L (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol*. 7:252

Vorob'ev AV, Manucharova NA, Iaroslavtsev AM, Belova EV, Zviagintsev DG, Sudnitsyn II (2007) The composition of the chitinolytic microbial complex and its effect on chitin decomposition at various humidity levels. *Mikrobiologiia*.76:632-638.

Ward JH (1963) Hierarchical Grouping to Optimize an Objective Function. *J Americ Stat Assoc*. 58:236-244

Williamson N, Brian P, Wellington EM (2000) Molecular detection of bacterial and streptomycete chitinases in the environment. *Antonie van Leeuwenhoek*. 78: 315-321

Xiao X, Yin X, Lin J, Sun L, You Z, Wang P, Wang F (2005) Chitinase Genes in Lake Sediments of Ardley Island, Antartica. *Appl Environ Microbiol*.71:7904-7909

Yamaoka H, Hayashi H, Karita S, Kimura T, Sakka K, Ohmiya K (1999) Purification and some properties of a chitinase from Xanthomonas sp. strain AK. *J Biosci Bioeng*. 88:328-330

Supplemental informations

- S1: Taxonomical affiliation at the phylum level

Prokaryote phyla	Rothamsted (n=11)		0x0 (n=2)		0x20 (n=2)		1x20 (n=2)		10x20 (n=2)		*Sensu lato (Roth)*	*Sensu medio (0x0)*	*Sensu stricto (0x20)*
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD			
Actinobacteria	13.776	4.755	19.555	7.90E-02	25.968	0.407	26.548	0.814	28.616	0.073	***5.69E-04***	***6.02E-04***	0.075
α-proteobacteria (Class)	22.313	2.72	21.65	0.615	23.275	1.55E-01	22.53	2.00E-03	20.183	0.274	0.748	***0.012***	***2.55E-03***
Unassigned reads	11.198	0.71	11.863	0.2	10.097	4.10E-02	10.058	0.212	9.458	0.011	***5.32E-03***	***1.27E-03***	0.061
β-proteobacteria (Class)	8.652	1.514	7.573	0.101	7.192	8.10E-02	7.183	5.50E-02	8.897	0.518	0.398	***0.029***	***0.044***
γ-proteobacteria(Class)	8.396	3.506	6.251	0.14	5.847	4.60E-02	5.895	3.20E-02	7.32	0.369	0.679	***0.019***	***0.027***
Firmicutes	4.936	0.54	4.727	1.10E-02	4.382	8.00E-03	4.467	1.00E-02	4.202	9.90E-02	0.249	***7.86E-03***	0.098
Acidobacteria	4.005	1.025	5.472	0.182	4.103	7.70E-02	4.039	0.172	3.736	0.102	0.345	***3.29E-03***	0.229
δ-proteobacteria (Class)	5.017	0.445	5.04	4.90E-02	4.102	1.50E-02	4.09	5.00E-02	3.716	3.60E-02	***1.84E-03***	***8.24E-05***	***7.98E-03***
Cyanobacteria	2.889	0.43	2.537	4.10E-02	2.289	7.00E-03	2.336	3.30E-02	2.081	3.90E-02	0.055	***2.91E-03***	***0.017***
Verrucomicrobia	2.771	0.445	2.79	4.40E-02	2.196	4.90E-02	2.215	6.50E-02	1.981	8.60E-02	0.067	***3.30E-03***	0.159
Bacteroidetes	2.931	0.671	2.559	0.132	1.793	2.90E-02	1.779	2.40E-02	1.721	6.50E-02	***0.029***	***3.90E-03***	0.53
Planctomycetes	5.42	3.282	2.464	9.30E-02	2.284	6.50E-02	2.279	0.115	1.662	4.70E-02	0.265	***9.28E-03***	***0.019***
Euryarchaeota	0.693	7.30E-02	0.673	7.00E-03	0.646	0	0.659	7.00E-03	0.607	1.40E-02	0.495	***0.022***	0.054
Deinococcus-Thermus	0.567	7.40E-02	0.562	2.00E-03	0.557	2.00E-03	0.548	3.00E-03	0.513	1.00E-03	0.88	***4.29E-04***	***2.15E-03***
Chlorobi	0.464	3.90E-02	0.445	6.00E-03	0.382	2.00E-03	0.377	3.00E-03	0.342	1.60E-02	***1.63E-03***	***4.82E-03***	0.106
Unclassified Bacterial reads	0.384	6.70E-02	0.393	6.00E-03	0.359	2.00E-03	0.348	8.00E-03	0.332	1.20E-02	0.739	***0.023***	0.227
Spirochaetes	0.218	2.20E-02	0.2	4.00E-03	0.165	8.00E-03	0.159	4.00E-03	0.148	6.00E-03	***5.69E-04***	***0.012***	0.296
Nitrospirae	0.175	2.60E-02	0.183	2.00E-03	0.146	1.00E-03	0.152	1.00E-03	0.141	6.00E-03	0.19	***2.98E-03***	0.267
ε-proteobacteria (Class)	0.187	1.40E-02	0.18	4.00E-03	0.153	4.00E-03	0.154	1.00E-03	0.14	4.00E-03	***5.30E-04***	***6.27E-03***	0.102
Gemmatimonadetes	0.208	4.90E-02	0.197	3.00E-03	0.144	1.00E-03	0.137	0	0.119	1.00E-03	0.056	***1.83E-05***	***1.33E-04***
Aquificae	0.157	1.90E-02	0.157	1.00E-03	0.129	3.00E-03	0.13	0	0.117	3.00E-03	***0.028***	***1.12E-03***	0.063
Fusobacteria	9.70E-02	9.00E-03	8.80E-02	0	7.50E-02	4.00E-03	7.20E-02	2.00E-03	7.20E-02	3.00E-03	***1.27E-03***	***0.049***	0.763
Lentisphaerae	8.80E-02	2.30E-02	7.10E-02	1.00E-03	5.90E-02	0	5.90E-02	3.00E-03	4.80E-02	2.00E-03	0.087	***3.59E-03***	***0.038***
Deferribacteres	5.60E-02	7.00E-03	5.50E-02	1.00E-03	4.60E-02	0	4.50E-02	0	4.20E-02	3.00E-03	***0.032***	***0.012***	0.316
Dictyoglomi	4.20E-02	5.00E-03	4.30E-02	1.00E-03	3.70E-02	1.00E-03	4.00E-02	1.00E-03	3.50E-02	1.00E-03	0.225	***0.041***	0.148
Tenericutes	2.10E-02	3.00E-03	1.80E-02	1.00E-03	1.60E-02	1.00E-03	1.50E-02	1.00E-03	1.50E-02	1.00E-03	***0.033***	0.109	0.758
Chrysiogenetes	2.00E-02	2.00E-03	2.10E-02	0	1.60E-02	1.00E-03	1.80E-02	2.00E-03	1.40E-02	1.00E-03	***2.71E-03***	***0.041***	0.271
Candidatus Poribacteria	2.20E-02	4.00E-03	2.10E-02	0.001	1.50E-02	1.00E-03	1.50E-02	1.00E-03	1.30E-02	2.00E-03	***0.031***	***0.022***	0.469
Elusimicrobia	1.70E-02	1.00E-03	1.70E-02	1.00E-03	1.10E-02	1.00E-03	1.20E-02	1.00E-03	1.20E-02	2.00E-03	***2.40E-04***	0.169	0.933
Fibrobacteres	1.50E-02	1.00E-03	1.30E-02	0	1.10E-02	0	1.20E-02	0	9.00E-03	1.00E-03	***1.85E-04***	***0.04***	0.104
Nanoarchaeota	1.00E-03	0	1.00E-03	0	0	0	0	0	0	0	***5.24E-03***	0.076	0.536

S1: Abundance table of the most significant prokaryote groups found in the metagenomes. Relative abundance of each group is given through the means and standard deviations (SD) detected in each metagenomic conditions. The number of metagenomes per conditions is given in brackets (n). Multiple group analysis (ANOVA) has been carried out at different level, which correspond to the successive comparison of the two enriched conditions 1x20 and 10x20 with the 3 controls: Rothamsted (Sensu lato), 0x0 (Sensu medio) and 0x20 (Sensu stricto). The p-values generated from multi-group ANOVA are given for the different analysis level (p-values < 0.05)

- S2: Functional affiliation on SEED level 1

SEED subsystems level 1	Rothamsted (n=11)		0x0 (n=2)		0x20 (n=2)		1x20 (n=2)		10x20 (n=2)		*Sensu lato*	*Sensu medio*	*Sensu stricto*
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD			
Carbohydrates	14.98	0.45	15.48	0.04	15.56	0.06	15.5	0.08	15.38	0.08	0.232	0.493	0.33
Clustering-based subsystems	13.91	0.13	13.98	0.02	14.02	0.1	14.02	0.03	14.01	0.02	0.36	0.48	0.995
Amino Acids and Derivatives	10.28	0.36	10.13	0.02	10.35	0.01	10.41	0.13	10.42	0.07	0.809	0.154	0.822
Miscellaneous	7.74	0.13	7.59	0.05	7.68	0	7.62	0.005	7.61	0.01	0.237	0.834	***6.44E-03***
Protein Metabolism	6.27	0.43	6.21	0.09	5.99	0.08	6.06	0.01	5.99	0.03	0.6	0.14	0.589
Cofactors, Vitamins, Prosthetic Groups, Pigments	6.19	0.11	6.18	0.002	6.25	0.01	6.25	0.04	6.3	0.09	0.379	0.394	0.789
RNA Metabolism	4.68	0.31	4.42	0.08	4.24	0.02	4.31	0.01	4.34	0.08	0.164	0.568	0.439
Fatty Acids, Lipids, and Isoprenoids	3.95	0.13	4.08	0.01	4.27	0	4.16	0.05	4.3	0.01	***0.007***	***0.04***	0.106
Cell Wall and Capsule	3.79	0.1	3.68	0.02	3.49	0.11	3.53	0.05	3.52	0.004	***0.002***	0.058	0.905
DNA Metabolism	3.56	0.13	3.6	0.01	3.55	0	3.53	0.02	3.54	0.004	0.946	0.054	0.394
Virulence, Disease and Defense	3.29	0.23	3.13	0.01	2.87	0.04	2.82	0.03	2.77	0.03	***0.009***	***0.005***	0.27
Respiration	2.91	0.12	3.09	0.05	3.17	0.02	3.17	0.01	3.09	0.01	***0.021***	0.261	0.06
Membrane Transport	2.6	0.14	2.5	0.02	2.38	0.02	2.44	0.01	2.46	0.05	0.244	0.523	0.366
Stress Response	2.4	0.07	2.33	0.01	2.37	0.02	2.33	0.02	2.36	0.06	0.395	0.852	0.787
Nucleosides and Nucleotides	2.28	0.07	2.19	0.03	2.19	0.01	2.19	0.02	2.15	0.002	***0.039***	0.356	0.111
Metabolism of Aromatic Compounds	2.07	0.17	2.15	0.04	2.36	0.04	2.27	0.02	2.28	0.05	0.141	0.161	0.273
Regulation and Cell signaling	1.53	0.12	1.59	0.05	1.68	0.05	1.66	0.03	1.62	0.04	0.293	0.502	0.599
Sulfur Metabolism	1.21	0.05	1.25	0.01	1.3	0.02	1.3	0.004	1.26	0.04	0.088	0.461	0.574
Phages, Prophages, Transposable elements, Plasmids	1.18	0.1	1.27	0.02	1.31	0.02	1.37	0.03	1.35	0.01	***0.023***	0.076	0.252
Phosphorus Metabolism	1.07	0.04	1.07	0.01	1	0.003	1.04	0.004	1.08	0.01	0.499	0.174	***0.014***
Cell Division and Cell Cycle	0.96	0.04	0.98	0.02	0.94	0.02	0.99	0.001	0.98	0.02	0.503	0.901	0.365
Motility and Chemotaxis	0.86	0.1	0.75	0.001	0.72	0.003	0.75	0.02	0.78	0.02	0.235	0.447	0.139
Nitrogen Metabolism	0.74	0.02	0.78	0.01	0.78	0.002	0.78	0.03	0.76	0.001	0.231	0.693	0.705
Iron acquisition and metabolism	0.52	0.08	0.49	0.02	0.45	0.001	0.45	0.01	0.56	0.003	0.39	***0.008***	***3.25E-04***
Secondary Metabolism	0.42	0.02	0.43	0.01	0.46	0.01	0.44	0.01	0.44	0.004	0.383	0.961	0.098
Potassium metabolism	0.39	0.02	0.41	0.001	0.38	0.004	0.39	0.01	0.41	0.02	0.346	0.614	0.371
Dormancy and Sporulation	0.13	0.01	0.14	0.003	0.13	0.001	0.13	0.004	0.14	0.01	0.144	0.199	0.096
Photosynthesis	0.11	0.004	0.11	1.40E-04	0.11	0.003	0.1	0.002	0.1	0.003	0.134	0.208	0.402

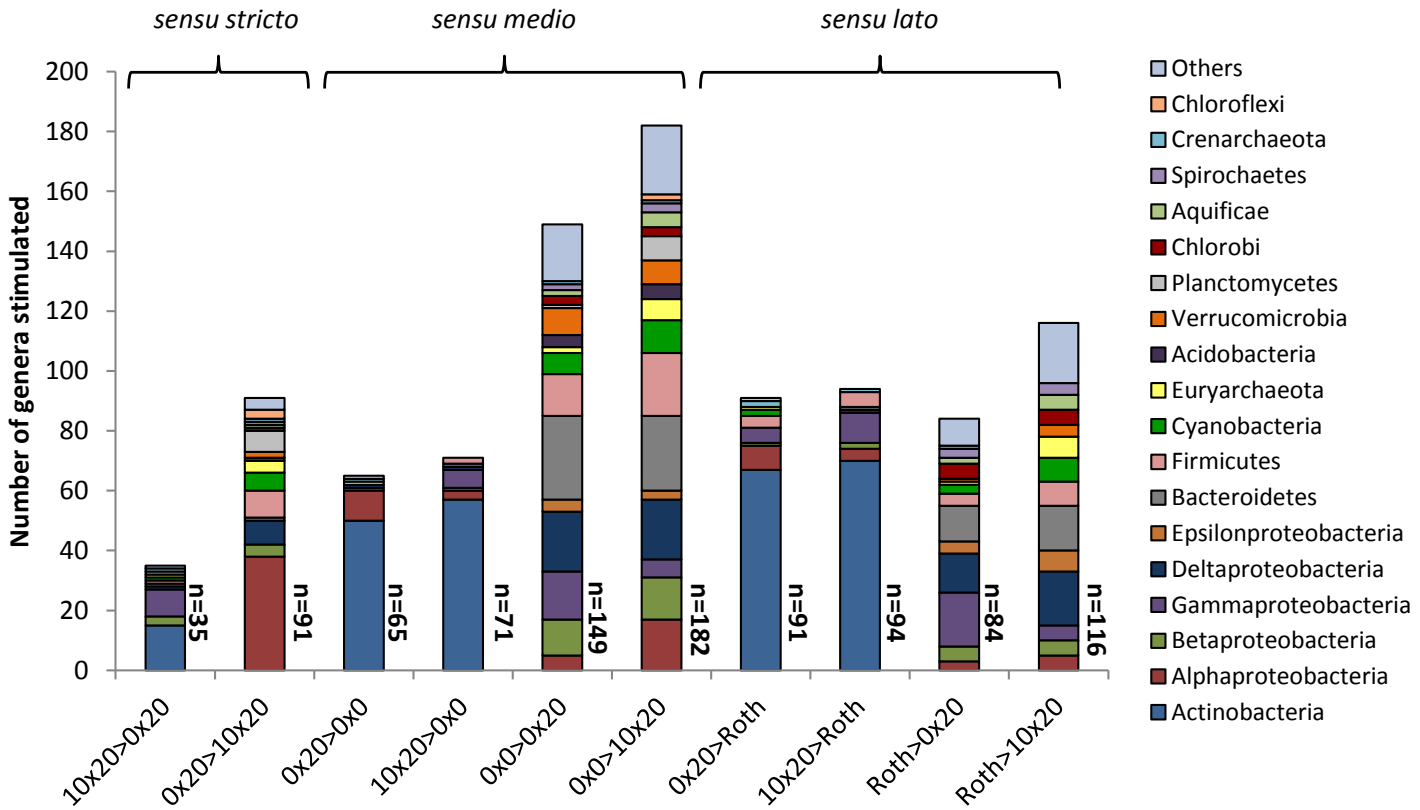
S2: Abundance table of the SEED subsystems (Level 1) in the metagenomes of this study. Relative abundance of each subsystem is given through the means and standard deviations (SD) detected in each metagenomic conditions. The number of metagenomes per conditions is given in brackets (n). Multiple group analysis (ANOVA) has been carried out at different level, which correspond to the successive comparison of the two enriched conditions 1x20 and 10x20 with the 3 controls: Rothamsted (Sensu lato), 0x0 (Sensu medio) and 0x20 (Sensu stricto). The p-values generated from ANOVA are given for the different analysis level (p-values < 0.05)

- S3: Summarized table of genera reduced by enrichment

Bacterial group	Genus level	CAZy related chitinases	ANOVA p-value	Rothamsted (n=11)		Control 0x0 (n=2)		Control 0x20 (n=2)		Chitin 1x20 (n=2)		Chitin 10x20 (n=2)	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
δ-proteobacteria (n=8)	Geobacter	GH19 GH20 CE4	1,93E-05	0,98	6,88E-02	1,00	3,53E-02	0,75	1,12E-04	0,75	1,21E-02	0,70	1,80E-02
	Pelobacter	GH19 CE4	6,49E-05	0,25	1,75E-02	0,25	2,19E-04	0,20	4,21E-03	0,20	9,70E-04	0,18	1,29E-03
	Lawsonia	CE4	1,11E-03	1,52E-02	1,22E-03	1,45E-02	7,09E-04	1,19E-02	1,52E-04	1,22E-02	3,31E-03	1,00E-02	7,70E-05
	Desulfurivibrio	CE4	5,48E-03	0,45	4,63E-02	0,42	5,50E-03	0,38	7,78E-03	0,38	6,25E-03	0,33	1,46E-02
	Syntrophus	CE4	5,23E-03	0,11	9,92E-03	0,11	9,08E-04	9,08E-02	8,04E-04	9,37E-02	1,68E-03	8,12E-02	3,99E-03
	Desulfobacterium	CE4	6,66E-03	4,28E-02	3,52E-03	4,45E-02	4,25E-03	4,09E-02	1,84E-03	3,65E-02	5,09E-04	3,27E-02	2,15E-03
	Desulfovibrio	GH18 CE4	9,04E-03	3,88E-02	3,93E-03	3,75E-02	6,27E-04	3,22E-02	7,46E-04	3,09E-02	2,95E-03	2,87E-02	6,80E-04
	Haliangium	GH18 GH20 CE4	4,20E-02	0,22	3,46E-02	0,22	5,23E-03	0,18	8,06E-03	0,18	3,89E-03	0,16	3,76E-03
ε-proteobacteria (n=4)	Helicobacter	CE4	3,73E-03	3,73E-02	4,96E-03	3,84E-02	3,84E-04	2,88E-02	4,85E-04	2,73E-02	1,88E-03	2,50E-02	1,61E-03
	Arcobacter	GH19 CE4	3,73E-03	1,56E-02	1,91E-03	1,44E-02	5,27E-05	1,17E-02	2,16E-03	1,07E-02	9,29E-04	1,05E-02	1,78E-03
	Campylobacter	CE4	4,60E-02	5,31E-02	8,81E-03	4,81E-02	8,63E-04	3,88E-02	6,52E-04	4,24E-02	1,47E-03	3,80E-02	2,90E-03
	Wolinella	CE4	4,70E-02	1,09E-02	1,59E-03	9,93E-03	1,26E-03	9,58E-03	1,83E-03	8,42E-03	6,30E-04	7,30E-03	2,47E-04
Cyanobacteria (n=3)	Prochlorococcus	GH18	5,47E-03	7,74E-02	7,25E-03	6,99E-02	1,91E-03	6,55E-02	1,22E-03	6,53E-02	2,58E-03	5,71E-02	5,14E-03
	Gloeobacter	CE4	3,00E-02	0,24	3,27E-02	0,23	2,33E-03	0,19	2,38E-04	0,20	3,70E-03	0,18	3,49E-03
	Thermosynechococcus	CE4	6,60E-02	5,50E-02	6,59E-03	5,32E-02	4,92E-04	4,80E-02	1,01E-03	4,71E-02	1,69E-03	4,27E-02	1,08E-03
Bacteroidetes (n=2)	Rhodothermus	GH18 GH20	1,96E-03	0,13	1,29E-02	0,12	8,26E-03	0,10	5,41E-03	0,10	4,03E-03	9,08E-02	4,39E-03
	Spirosoma	GH18 GH20 CE4	2,84E-03	0,23	3,82E-02	0,22	1,17E-02	0,15	3,06E-04	0,15	2,53E-03	0,14	7,40E-03
Aquificae (n=2)	Sulfurihydrogenibium	CE4	6,05E-03	3,27E-02	4,26E-03	3,49E-02	3,30E-03	2,56E-02	2,56E-03	2,50E-02	1,03E-03	2,25E-02	4,58E-04
	Thermocrinis	CE4	6,30E-02	1,24E-02	2,25E-03	1,19E-02	1,10E-03	9,79E-03	5,21E-04	9,73E-03	5,62E-04	8,08E-03	4,78E-04
Spirochaetes (n=2)	Leptospira	GH18 CE4	1,47E-03	0,10	1,40E-02	9,10E-02	5,64E-03	7,28E-02	5,15E-03	7,23E-02	5,07E-03	6,38E-02	3,75E-03
	Spirochaeta	GH18 GH20 CE4	1,10E-02	4,62E-02	4,57E-03	4,63E-02	8,18E-04	4,12E-02	5,82E-04	3,85E-02	1,33E-03	3,51E-02	9,58E-04
Acidobacteria	Fibrobacter	GH18	1,85E-04	1,47E-02	1,36E-03	1,31E-02	9,53E-05	1,11E-02	6,86E-04	1,23E-02	4,38E-04	8,58E-03	1,87E-03
Deferribacteres	Calditerrivibrio	CE4	8,08E-03	1,34E-02	1,78E-03	1,36E-02	4,96E-04	1,07E-02	5,03E-05	9,84E-03	8,56E-04	9,46E-03	9,04E-04
β-proteobacteria	Candidatus accumulibacter	CE4	1,30E-02	8,99E-02	7,94E-03	8,43E-02	2,86E-03	7,71E-02	1,91E-03	7,92E-02	4,41E-03	7,02E-02	1,30E-03
Thermotogae	Fervidobacterium	GH20	8,30E-02	1,71E-02	2,13E-03	1,65E-02	6,19E-04	1,68E-02	1,43E-03	1,55E-02	2,11E-04	1,27E-02	3,12E-04
Verrucomicrobia	Opitutus	GH18 GH20	5,32E-03	0,43	4,94E-02	0,45	4,19E-05	0,35	1,89E-03	0,34	2,85E-02	0,31	1,55E-02

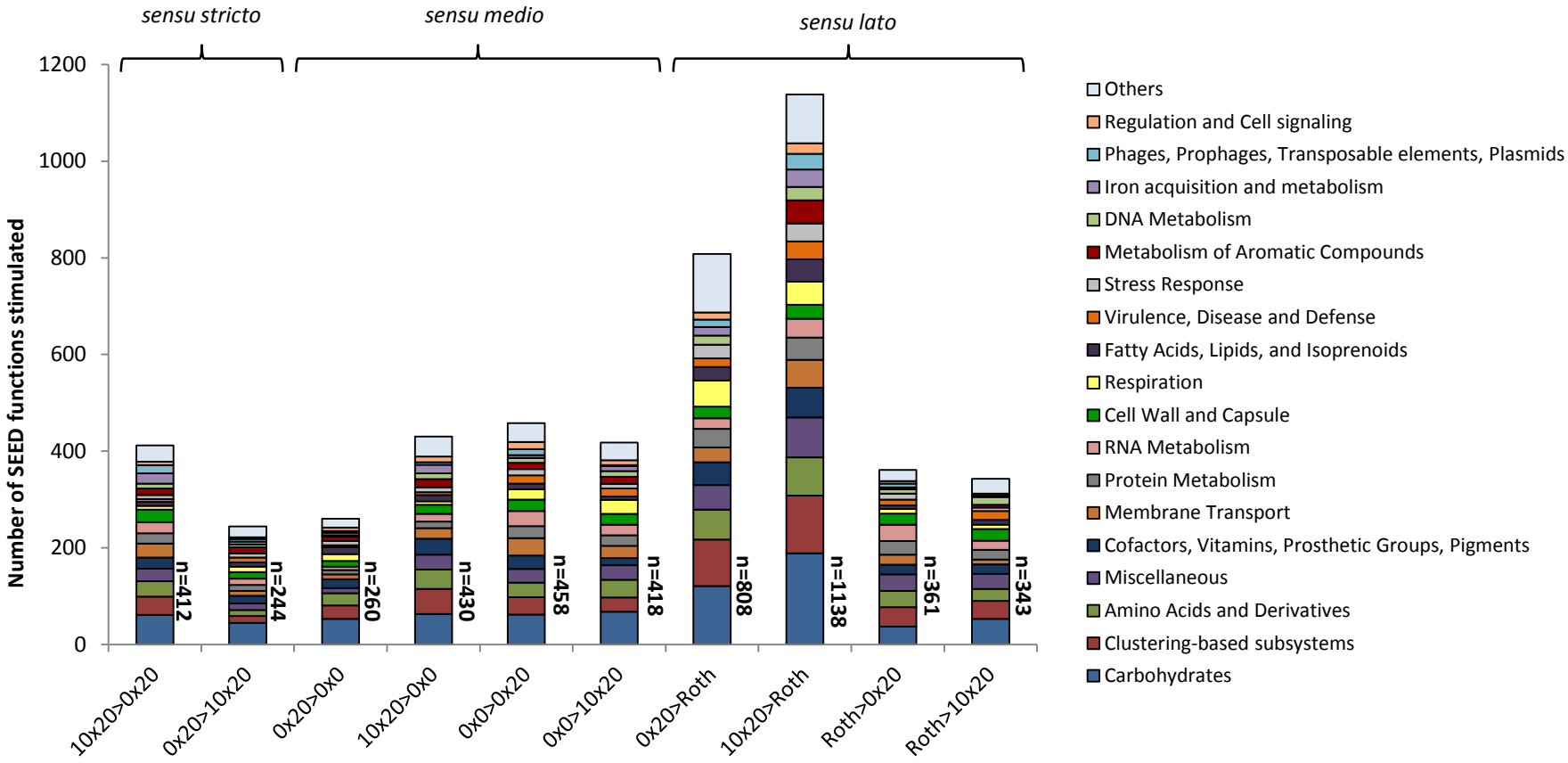
S3: Summarized table of some of the bacterial genera significantly reduced by chitin enrichment after comparison with all metagenomic conditions. We reported here only the genera known to have representatives harboring chitin degradation related genes. The presence of characterized enzymes related to chitin degradation in CAZy for each genus is given, as well as references that reported chitinase/cellulase activity. We reported cellulase activity as some of the characterized chitinase in CAZy were reported to have a catalytic activity toward cellulose. The chitin concentration where the genus abundance was found to be the highest is also given (1x and 10x). Multi-group ANOVA p-value were obtained from comparison of all the metagenomic condition with the Rothamsted database (sensu lato) (significance threshold ≤0.05).

- S4: Cross-comparison of taxonomical profils



S4: Cross-comparison of the taxonomic profiles of all metagenomic conditions. Analysis has been performed on metagenomes at the genus level through ANOVA (p-value cut-off < 0.05). The panel displays the cross-comparison of all conditions, and indicates the number of genera (n) significantly more abundant in metagenomes of one condition against another. The conditions are indicated as follow: The 11 metagenomes from Rothamsted (Roth), the 2 controls before incubation (0x0), the 2 incubation controls (0x20), and chitin treated metagenomes (10x20). The comparison with 1x20 is not presented as it is very similar to the control 0x20. The comparison between 0x0 and Rothamsted is also not presented for the same reasons.

- S5: Cross-comparison of functional profils



S5: Cross-comparison of the functional profiles of all metagenomic conditions. Analysis has been performed on metagenomes at the SEED functional subsystem level through ANOVA (p-value cut-off < 0.05). The panel displays the cross-comparison of all conditions, and indicates the number of functions (n) significantly more abundant in metagenomes of one condition against another. The conditions are indicated as follow: The 11 metagenomes from Rothamsted (Roth), the 2 controls before incubation (0x0), the 2 incubation controls (0x20), and chitin treated metagenomes (10x20)

- S6: RISA profils

Material and methods

Whole bacterial community structure was assessed using Ribosomal Intergenetic Spacer Analysis (RISA). As preliminary experiments demonstrate high reproducibility between biological replicates, metagenomic DNA from replicate microcosm was pooled prior amplification. For more details, report to the supplement data. The intergenetic spacer (IGS) between the small (16S) and the large (23S) subunit of ribosomal sequences were amplified by PCR using primers 5'-TGCGGCTGGATCCCCTCCTT-3' (forward) and 5'-CCGGGTTTCCCCATTCGG-3' (reverse) (Ranjard et al., 2000). 2µl of diluted metagenomic DNA (≈15ng) was mixed with 1.25µl of reverse and forward primers (10µM) and 20.5 µl of distilled sterile water. PCR was performed in a Biometra thermocycler, and cycles consisted in 95°C for 10 min and then 30 cycles of 95°C for 30s, annealing at 55°C for 30s, and elongation at 72°C for 1 min, followed by 72°C for 15 min. 1µl PCR product was loaded on a chip into an Agilent DNA 7500 Lab, and electropherograms were performed and analyzed on an Agilent 2100 Bioanalyzer. Electropherograms were analyzed using principal component analysis (PCA) and between group analysis (BGA), with Rgui software (R Development Core Team, 2011).

RISA profile analysis

Complementary to the quantitative analysis, a direct qualitative comparison between samples was carried out by analysis of RISA patterns obtained with PCR amplified *rrs* gene intergenic spacers. Again, a significant microcosm effect was revealed as soon as day 3 with appearance of a major band in control and chitin amended soil samples whose intensity increased with time. The specific quantitative bias related to addition of the biggest amounts of chitin was not detected at this qualitative level with almost the same RISA patterns for the 3 chitin conditions tested at t=0. Independently of the microcosm effect and regardless of the enrichment conditions tested, electropherograms show a strong similarity level all over the experimentation time indicating only slight changes in the bacterial community structures after chitin amendment (Fig. 1). However, some bands appearing at day 6 and maintaining for the 29 remaining days can be specifically attributed to chitin treatments.

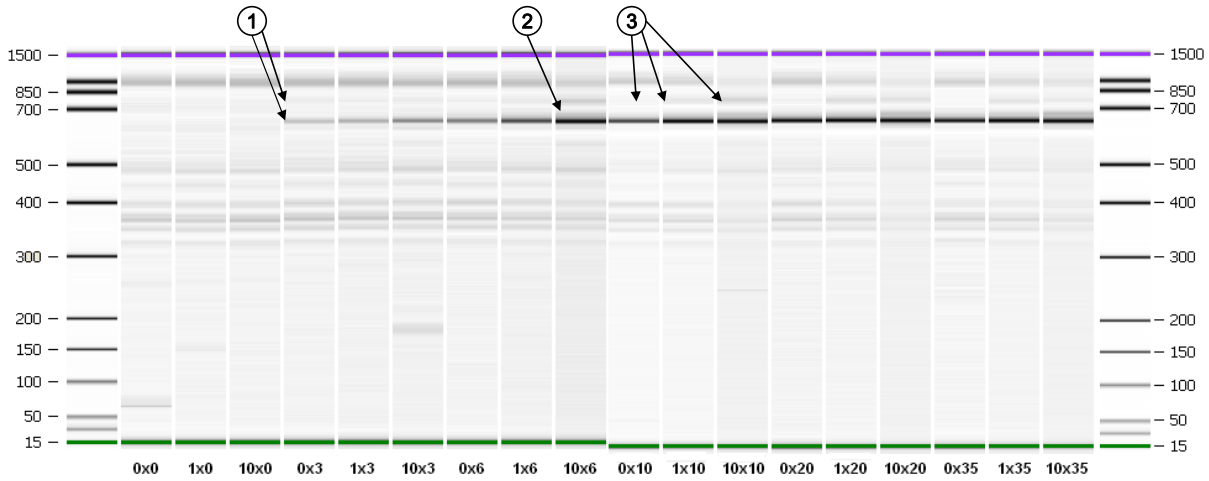


Fig.1. Reconstituted gel picture from RISA electropherograms profiles. Each column represents the gel migration of ribosomal intergenetic spacer PCR products from each enrichment condition and time. The picture is flanked on both sides by ladders representing DNA fragments with known size, expressed in base pair. Numbers and arrows refer to specific indications in the text.

CHAPTER 3

-

GENEFISH: UN NOUVEL OUTIL BIOTECHNOLOGIQUE POUR L'ETUDE CIBLEE DE LA DIVERSITE BACTERIENNE

-

VERSION FRANÇAISE

Introduction

Dans le cadre du projet Metaexplore, une attention particulière a été apportée à l'aspect innovation en matière de techniques permettant une meilleure caractérisation et valorisation de l'ADN métagénomique. Cet aspect fait l'objet d'un axe de recherche intitulé “Novel tools for environmental metagenomics”, incluant le développement de nouveaux tests enzymatiques afin d'améliorer les méthodes de criblage fonctionnel des banques de clones métagénomiques, la mise en place de nouveaux vecteurs et hôtes pour la constitution des banques et les tentatives d'expression. C'est dans ce contexte que se positionne l'outil Genefish, le but de l'approche étant de cloner des fragments d’ADN métagénomique par recombinaison homologue (et homéologue) grâce à une souche *E. coli* équipée d'un plasmide de capture. L’objectif principal qui a justifié le développement de cette approche est de cloner directement l’ADN métagénomique ciblé dans une bactérie avant d’éventuellement appliquer l’outil *in situ* pour capturer des fragments d’ADN environnemental. Cette stratégie se veut innovante et complémentaire des voies traditionnelles d'exploitation de l'ADN métagénomique. Le chapitre est structuré comme suit:

- une première partie vise à présenter l'outil et son principe, en détaillant les aspects moléculaires de son fonctionnement.

- une deuxième partie présentant les premières tentatives d'utilisation de l'outil et les résultats préliminaires.

La construction de l'outil est complexe, faisant intervenir de nombreux réglages moléculaires fins ayant nécessité de nombreuses mises au point. Du fait de la multiplicité des étapes requises pour la construction de l’outil, la rédaction s'appuie sur de nombreuses annexes détaillant exhaustivement les protocoles employés, et ce-ci par souci de clarté afin d'en faciliter la compréhension. Avant toute chose, il est important de resituer ce travail dans un contexte plus large. Le développement de l'outil Genefish a été initié par Nathalie Lombard, ancienne doctorante de l'équipe. Mon rôle en tant que doctorant du projet Metaexplore était de parachever son développement en vue de l'appliquer dans le cadre de la recherche de nouveaux gènes d'intérêt, comme les chitinases par exemple. Nous verrons toutefois que le développement de l’outil a nécessité des étapes supplémentaires, retardant son utilisation. Nous montrerons qu’en dépit de difficultés résiduelles, les résultats préliminaires obtenus sont très encourageants.

1. Présentation de l'outil Genefish:

Le principe de capture repose sur 3 étapes majeures:

- l'internalisation de la cible par une étape d'électro-transformation (Fig.1, A). Cette phase est strictement dépendante de la quantité de cibles présentes dans l'ADN métagénomique considéré, mais va être conditionnée aussi par la capacité de la souche à intégrer l'ADN.

- la phase de capture à proprement parler, qui est réalisée par la voie de la recombinaison homéologue (Fig.1, B). Cette réaction est prise en charge par le plasmide pSIM6 qui porte le système enzymatique λ red gam provenant du phage λ dont les recombinases sont connues pour leur efficacité. Cette phase ne pourrait pas être possible sans zones de similarité entre la cible et le plasmide de capture pB35. C'est pourquoi deux zones présentant des similarités avec la cible ont préalablement été choisies et clonées dans pB35.

- la troisième étape consiste à pouvoir sélectionner les cellules recombinantes (ayant intégré la cible), de celles non transformées. Pour cela, un système de contre-sélection a été ajouté sur le plasmide de capture entre les zones de similarité. Il s'agit d'une cassette comportant deux gènes toxiques réprimés en condition de culture normale, mais aisément inductibles chimiquement sur milieu sélectif. Le but de cette cassette est d'être remplacée lors de la recombinaison par la cible (Fig.1). Ainsi, lors de l'étape de sélection sur milieu gélosé, seules les bactéries présentant un plasmide de capture sans gènes toxiques seront capables de croître et de former des colonies.

L'aspect innovant de cette approche réside d'une part dans le principe de la capture et de la contre-sélection. En effet, la recombinaison homéologue doit pouvoir permettre de recueillir des cibles génétiques différentes de celles déjà connues, voire nouvelles. Même si cette stratégie requiert des zones de similarité conçues à partir de séquences connues, l'efficacité des recombinases du phage λ devrait aboutir au clonage de gènes différents de ceux utilisés pour la construction des zones de similarité, car elles tolèrent un certain degré de divergence de séquences. Ainsi, par opposition à d'autres techniques (ex. la PCR), le fait d'utiliser de larges zones de plusieur scentaines de paires de bases (≈500pb) doit pouvoir permettre la récupération de séquences potentiellement nouvelles, partageant peu de similarités avec les éléments des bases de données. En revanche, la stratégie Genefish présente certaines limitations, notamment liés à l'obligation de trouver deux zones de similarité avec la cible, ce qui limite la portée de l'approche pour découvrir des mécanismes génétiques nouveaux codant les fonctions recherchées.

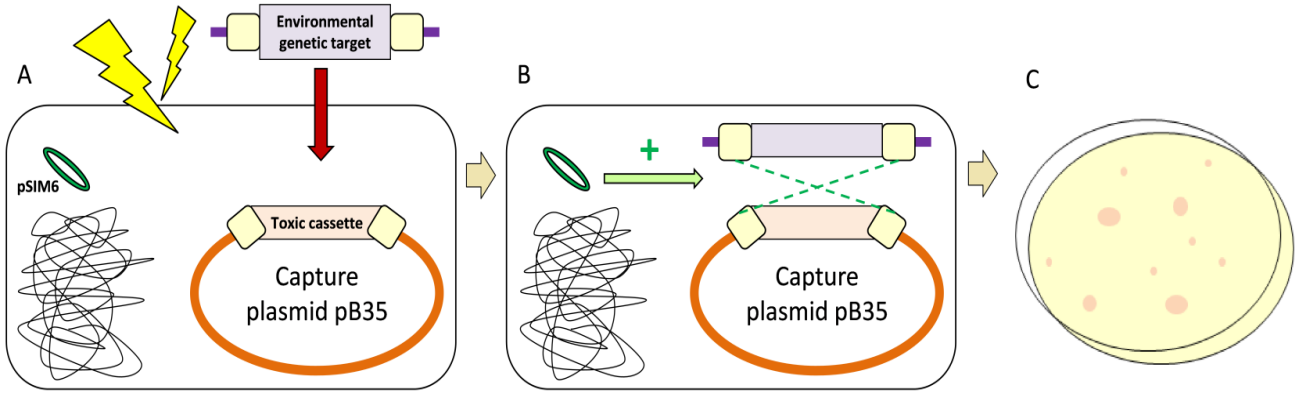


Fig.1: La strategie de capture d'ADN par l'outil Genefish. L'image A montre l'internalisation du fragment d'ADN cible (ADN métagénomique) suite à une étape d'électro-poration. L'image B montre la phase de capture sur le plasmide pB35 par la voie de la recombinaison homologue. Cette étape est réalisée grâce à la présence du plasmide recombinogène pSIM6 (élipse verte) et des zones d'homologie préalablement clonées dans le plasmide (carré jaune clair). L'intégration du matériel cible conduit à la suppression de la cassette toxique présente initialement sur le plasmide. L'image C montre l'étape finale de sélection sur boite, où seules les cellules ayant délété leur cassette toxique lors de la capture sont capables de pousser, en condition d'induction de la toxicité.

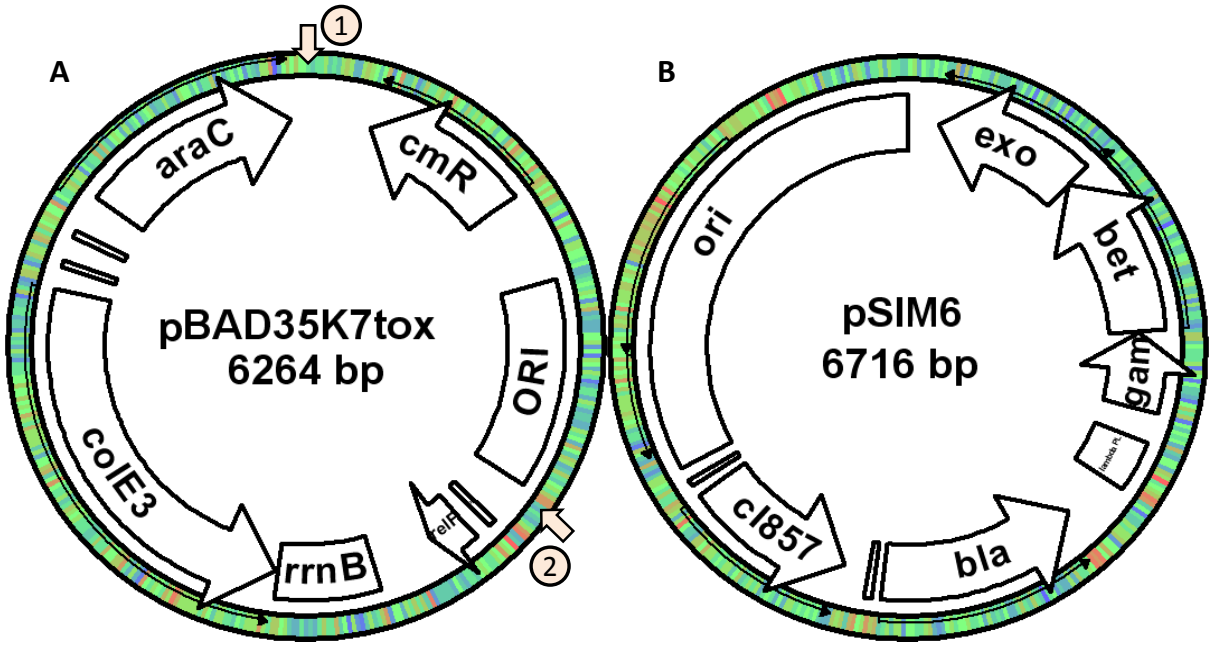


Fig.2: Les plasmides utilisés dans la stratégie Genefish. Le plasmid de capture pB35 est présenté en (A), et le plasmide de recombinaison pSIM6 est présenté en (B). Les flèches en (A) représentent les sites de multiclonages ou sont insérés les zones dédiées à la recombinaison homologues. Les gènes de pBAD35K7tox sont annotés comme suit: "colE3" et "relF"sont les deux gènes toxiques de la cassette, "rrnB" est le terminateur de "colE3", "araC" est son répresseur et "cmR" le marqueur de sélection (résistance au chloramphénicol). "ORI" est l'origine de réplication (oriP15A, 15-30 copies). Les gènes de pSIM6 sont annotés comme suit: "exo", "bet" et "gam" sont les 3 gènes du système λ red gam impliqués dans la recombinaison homologue sous le control du répresseur thermo-sensible "cl857". "bla" est le marqueur de sélection (beta-lactamase: résistance à la carbenicilline). "ori" est l'origine de réplication (pSC101, 5 copies).

2. Présentation des deux plasmides : pB35 et pSIM6

i. Le plasmide de capture pB35 :

Le plasmide sur lequel sera opérée la phase de capture est équipé d'une cassette toxique afin d'assurer la contre-sélection. Cette cassette est composée de deux gènes codant des toxines visant deux fonctions vitales chez la bactérie hôte. Ces gènes sont activables chimiquement par deux composés différents. Le premier gène *relF*, code une protéine de type porine causant un grave dysfonctionnement au niveau de la perméabilité membranaire ainsi que l'arrêt de la respiration. Le deuxième gène toxique *colE3* code une enzyme de type RNase appelée colicine qui inhibe la synthèse des protéines en dégradant les produits de transcription des gènes, l'ARNm. Cette activité est létale pour la cellule, la privant de toutes les fonctions assurées par les protéines et enzymes normalement traduites à partir de l'ARNm.

La régulation de l'expression de *colE3* se fait par un système subtil d'inducteur et d'inhibiteur (Fig.3). En temps normal, le promoteur pBAD de *colE3* est réprimé par la protéine AraC codée par le gène *araC* présent sur le plasmide. L'activation de la toxicité se fait par l'adjonction de L(+) arabinose 1% dans le milieu de culture, qui servira de substrat pour AraC, libérant ainsi le promoteur pBAD de *colE3*. Cependant, ce système de régulation présente une faiblesse. En effet *colE3* est sujet à des fuites d'expression malgré le bloquage d'AraC sur pBAD. Etant donnée la haute toxicité de la colicine, une forte mortalité des cellules étaient observée malgré le blocage, rendant l'outil moins efficace. C'est pourquoi un autre système de contrôle a été mis en place afin de corriger ce problème. Une modification génétique a été apportée à la souche hôte en lui ajoutant le gène *imme3* codant la protéine Imme3, l'anti-toxine de la colicine. Le gène *imme3* a été placé sous le control d'un promoteur constitutif faible de manière à ce que Imme3 soit produit en quantité suffisante pour bloqué la fuite de colicine, mais insuffisante pour l'inhiber lors de l'induction par l'arabinose. Ce système marche correctement, mais induit un ralentissement de croissance de la souche hôte du fait de la fuite.

Pour ce qui est de la régulation du gène *relF*, celui-ci est placé sous le control du promoteur P_{LAC}, bloqué en temps normal par la protéine LacIq produite par le gène chromosomique *lacIq* (Fig.4). En condition d'induction, l'ajout d'IPTG 500µg.ml au milieu de culture permet d'inhiber LacIq et de libérer ainsi l'expression de *relF*.

La co-induction des deux gènes *relF* et *colE3* permet d'obtenir une toxicité suffisante pour éliminer la quasi-totalité des bactéries. L'activation est réalisée directement sur milieu gélosé en présence d'IPTG et d'arabinose en concentration adéquate (1% masse/volume pour l'arabinose et 0,5ng/µl pour l'IPTG). Le système permet une contre-sélection très efficace, avec des taux d'échappement à la toxicité compris entre 10^{-8} et10^{-9}, soit une bactérie sur 100 millions/1 milliard restant capable de se développer en conditions d'induction.

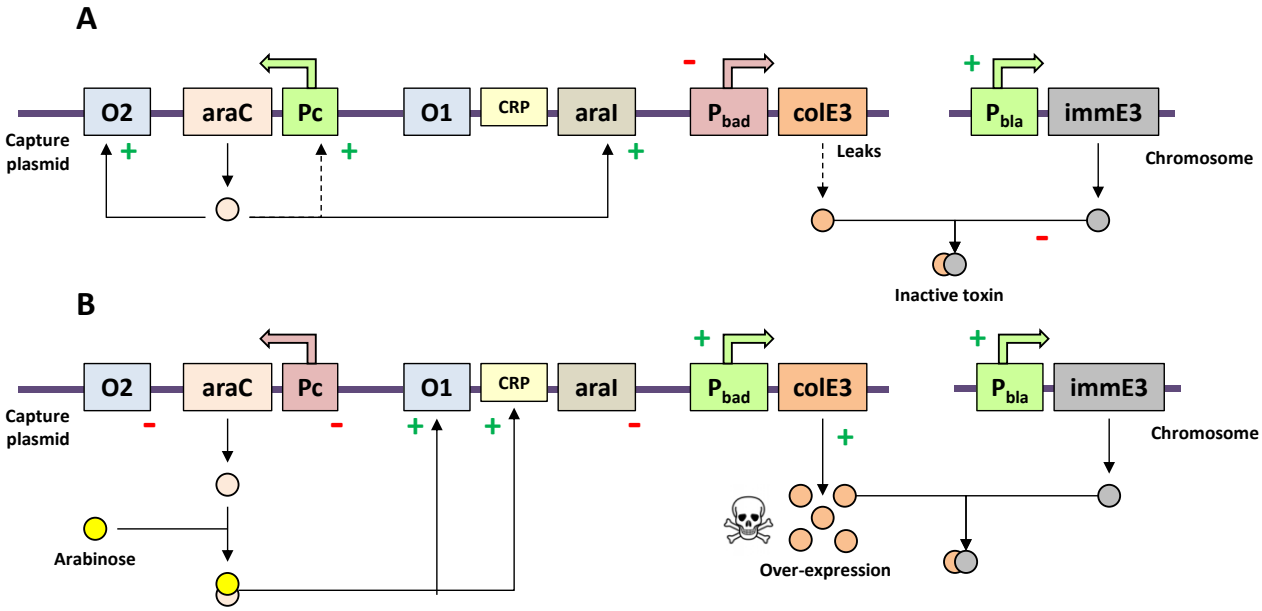


Fig.3: Régulation du gène colE3. La régulation en conditions de croissance normale est présentée en (A). L'activation de la toxicité en conditions d'induction est présentée en (B). Les signes positifs en vert indiquent l'activation de la transcription des gènes, tandis que les signse négatifs en rouge représentent le blocage de la transcription.

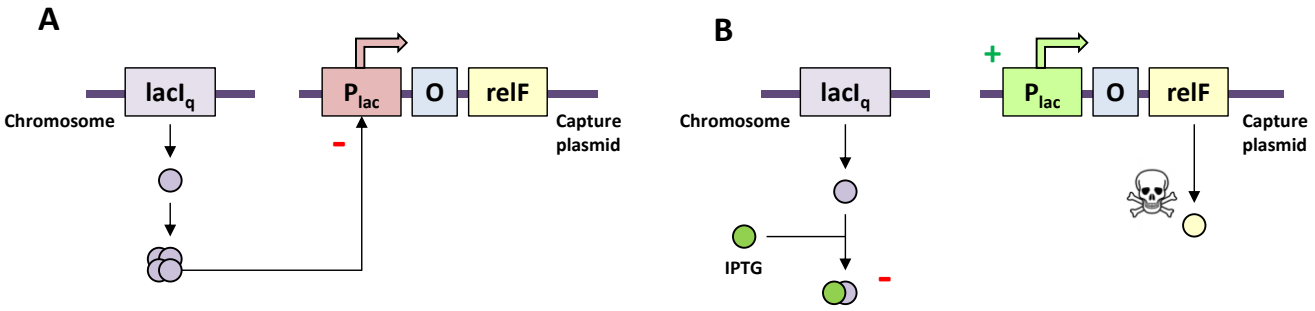


Fig.4: Régulation du gène relF. La régulation en conditions de croissance normale est présentée en (A). L'activation de la toxicité en conditions d'induction est présentée en (B). Les signes positifs en vert indiquent l'activation de la transcription des gènes, tandis que les signse négatifs en rouge représentent le blocage de la transcription.

ii. Le plasmide de recombinaison : pSIM6

Le plasmide pSIM6 est un plasmide thermo-inductible possèdant 3 gènes impliqués dans le processus de recombinaison du phage λ : *exo*, *bet* et *gam* (Datta *et al.* 2006) (Fig.2). Les gènes *exo* et *bet* constituent le système red, responsable de la recombinaison. Le gène *exo* code une exonuclease 3'-5' qui génère de l'ADN simple brin à partir de l'ADN entrant dans la cellule. Le gène *bet* code une protéine possèdant une forte affinité pour l'ADN simple brin et va promouvoir la recombinaison homologue. Le gène *gam* code une protéine possèdant une forte affinité pour les systèmes exo-nucléasiques bactériens (RecBCD, SbcCD) et les inactive (Murphy, 1991; Court *et al.* 2007). Ce système permet d'améliorer les rendements de recombinaison homologue d'un facteur 10 chez *E. coli* (Murphy, 1998).

Pour le développement de l'outil Genefish, un système thermo-inductible à été employé pour contrôler le système λ red gam. Le promoteur P_L est induit à 42°C par inactivation de son répresseur CI857. De plus, le plasmide possède une origine de réplication sensible à la température, empêchant le plasmide de se répliquer après une induction à 42°C. La construction sera donc définitivement éliminée après quelques cycles de division cellulaire. Cet aspect est crucial car il garantit un "usage unique" du plasmide et du système λ red gam, permettant aisni d'éviter des problèmes ultérieurs de stabilité dans la souche *E. coli* du fait de la grande efficacité des recombinases du phage λ.

3. Application de l'outil GeneFish

Afin de tester les capacités de l'outil, des premières tentatives ont été réalisées à partir de gènes présentant plusieurs caractéristiques qui en font de bons modèles d'étude. Ces gènes doivent être ubiquitaire, présenter un bon niveau de conservation avec suffisamment de séquences disponibles dans les banques de données, mais également susceptibles d'être abondants dans les sols. Ces aspects permettent, en effet, d'optimiser l'internalisation de la cible et de faciliter la conception des zones de similarités. Ce sont les gènes de l'opéron "*nar*" qui ont été sélectionnés, notamment *narG* et *narH*, codant l'enzyme "nitrate réductase", fortement représentée dans les populations bactériennes de par sa grande importance métabolique. Cette enzyme permet de réaliser la "respiration nitrate" quand l'oxygène vient à manquer. Deux zones de similarité, G1 et G2, ont été créées par nos collaborateurs de l'équipe de Microbiologie des Sols-Géosols, sous la direction de Laurent Philippot (Institut National de la Recherche Agronomique, 21065 Dijon cedex, France). Les tests ont ensuite été réalisés avec le produit PCR des gènes *narG-narH* provenant de la souche *Pseudomonas fluorescens*.

Les premières tentatives d'utilisation du système avec les deux plasmides pB35 et pSIM6 en tandem se sont avérées peu concluantes. Le principal problème identifié est que le plasmide de capture pB35 est porteur d'une origine de réplication qui tolère 10 à 15 copies du plasmide par cellule. Même si l'une d'entre elles parvient à capturer une cible, la probabilité pour qu'il persiste au moins une copie de la cassette toxique lors de l'induction reste très élevée. Il est donc fortement probable qu'une grande partie des recombinants potentiels soient de ce fait simplement éliminés lors de l'étape de sélection sur milieu en présence des inducteurs chimiques.

Ainsi, deux autres approches ont été développées afin de contourner ce problème :

- l'approche dite de "co-transformation" consistant à apporter le plasmide de capture pB35 en même temps que le produit de PCR *narG-H* lors de l'électro-transformation et ceci dans des ratios moléculaires spécifiques afin de contrôler le nombre exact de copies théoriques par cellule. Cette approche a été concluante et a permis d'augmenter les rendements de détection de recombinants. Néanmoins, d'autres problèmes sont venus se greffer notamment ceux liés à la faible stabilité du plasmide de capture pB35 pendant les phases de culture prolongée.

- l'approche dite de "ségrégation": consistant à extraire immédiatement les plasmides de capture pB35 après électro-transformation. Cette étape supplémentaire permet de récupérer toutes les copies recombinantes et non-recombinantes des plasmides. Ce mélange est alors utilisé de nouveau pour transformer la même bactérie hôte, mais dans un ratio moléculaire permettant d'incorporé théoriquement une copie de plasmide par cellule. De ce fait, une ségrégation moléculaire est oppérée, permettant de pourvoir certaines cellules avec des plasmides recombinant seuls, contournant ainsi le biais de la présence résiduelle de copies de plasmides contenant la cassette toxique. Cette stratégie, plus lourde à mettre en place, s'est toutefois avérée payante et a permis d'améliorer les rendements de détection des recombinants.

Une étape complémentaire visant à utiliser un plasmide de capture équipé d'une origine de réplication ne tolérant qu'une copie par cellule est envisagée. Des tentatives de clonage de la cassette toxique dans un tel vecteur sont en cours. De plus, une stratégie visant à cloner la cassette dans le génome est envisagée. De nouvelles tentatives de captures réalisées sous d'autres conditions (taille des zones de similarité, longueur du fragment à recombiner, autres cibles…) seront nécessaires afin d'établir les possibilités et les limites du système Genefish.

CHAPTER 3

-

GENEFISH: A NEW BIOTECHNOLOGICAL TOOL TO INVESTIGATE TARGETED BACTERIAL DIVERSITY

-

PURPOSE, DEVELOPMENT AND APPLICATIONS

-

ENGLISH VERSION

CHAPTER 3

-

Genefish: A new biotechnological tool to investigate targeted bacterial diversity

-

Purpose, development and applications

SAMUEL JACQUIOD[1], NATHALIE LOMBARD[1], JUN HUAN[1], LAURENT PHILIPPOT[2], JEAN CLAUDE LAZZARONI[3], LAURE FRANQUEVILLE[1] and PASCAL SIMONET[1]

[1]Environmental Microbial Genomics Group, AMPERE, UMR CNRS 5005, Ecole Centrale de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France.

[2]Microbiologie des Sols-Géosols, UMR1324 CSGA Centre des Sciences du Goût et de l'Alimentation, *Institut National de la Recherche Agronomique, 21065 Dijon cedex, France.*
[3]Unité de Microbiologie, Adaptatation et Pathogénie, UMR5240, Université de Lyon, 69622 Villeurbanne cedex, France

Introduction

Metagenomic approaches allowed bypassing of traditional bias in microbiology such as cultivation. As a consequence, metagenomics give very promising perspectives to microbiologists in their attempts to characterized the microbial world at both functional and diversity level. However, more and more studies are coming up nowadays to point out the limitations of these approaches, such as the effect of nucleic acid extraction protocols from the same sample (Inceoglu *et al.* 2010; Delmont *et al.* 2011), but also the microbial distribution heterogeneity from environmental samples (Hassink *et al.* 1993). Here we report the development of a new strategy based on a totally different concept than traditional techniques, with the potential to find novelties by targeting specific genes, in environmental microbial populations.

This strategy relies on the development of a highly recombinant *E. coli* strain equipped for internalization and capture of extra-cellular DNA into a hook plasmid. The aim is to target a specific bacterial function by designing and cloning of conserved regions from known genes related to the desired activity into the capture plasmid. The specific recovery of these genes will be achieved through catching of the incoming environmental DNA through a recombination process, mediated by the recombinase system from the lambda phage. The counter-selection of recombinant events is achieved through the specific deletion of a toxic cassette, occuring along the recombination process. This strategy allows positive selection of recombinant bacteria through plate culture under the presence of toxicity inducers. At the end, only the recombinant bacteria that previously deleted their toxic cassette while intergrating the targeted genes will be able to grow on plate under inducing conditions.

The concept was initially developed to be applied directly onto environmental samples, such as soil, in order to be directly in contact to the environmental DNA (*in situ* approach). For instance, the DNA internalization can be mediated in soil samples by lightening-mediated transformation (Demaneche *et al.* 2005). After inoculation, cell recovery can be performed by a whole cells extraction step from the environmental sample (Holben *et al.* 1988), before starting the counter-selection on plates. Furthermore, the system can be applied onto extracted environmental DNA as well, by electro-transformation based techniques (*in vivo* approach).

The Genefish strategy relies on several key points. (1) The first one is the ability for the strain to survive and colonize the environmental sample in order to reach the inherent environmental DNA. (2) The second one is to be able to integrate the targeted DNA. This DNA will most likely come from living cells through horizontal gene transfer, or rather from dead cells and be internalized through natural processes or mediated/stimulated transformation. (3) Third point is to ensure stabilization of the integrated DNA in the genome of the host strain and its perpetuation through cell division. (4) Fourth point is to be able to extract and efficiently select the recombination events from the bacterial pool for downstream analysis. The *E. coli* strain K12 DH5α (Bethesda Research laboratories, BRL) has been selected as a host for developing our system, as it represent a well known model even though its primary natural habitat is the mammalian guts (Mason and Richardson, 1982). In order to increase its survival after inoculation in an environmental sample, some factors such as incubation temperature or cell density could readily be adjusted. *E. coli* has not been reported to undertake natural transformation, but its competence under electro-transformation is quite remarkable, with relative high rate of extra-cellular DNA uptake. For instance, previous studies based on *E. coli* reported that up to 10^{10} transformant can be obtained per µg of overcoided plasmidic DNA (Dower *et al.* 1988; Taketo, 1988; Hengen, 1995). Modifications in the recombination machinery of the *E. coli* host strain were performed in order to ensure integration of the targeted DNA into the genome. Furthermore, a counter-selection system based on the induction of toxic genes has been developed for recovery of the positive recombinational events. Finally, a genetic model has been determined to explore and validate the so-constructed system in order to assess the *in vitro* tool capacities and efficiency.

The development of the Genefish tool was initiated by Nathalie Lombard, a former PhD from the Environmental Microbial Genomics Group. She dedicated most of her PhD work to build the final capture plasmid backbone, as well as selecting and testing the counter-selection system. She also participated to in genetic engineering of the *E. coli* strain, in tight collaboration with Jean-Claude Lazzaroni from the University of Lyon 1 (Unité de Microbiologie, Adaptatation et Pathogénie). My implication in this project was to continue the work initiated by Nathalie, to finalize the tool and apply it through an *in vitro* model experiment in collaboration with Laurent Philippot from The French institute for agronomy research at Dijon (Microbiologie des Sols-Géosols, CSGA Centre des Sciences du Goût et de l'Alimentation, INRA Institut National de la Recherche Agronomique, 21065 Dijon, France). During my work, I was helped by Laure Franqueville, research engineer (CNRS), who is strongly involved in the Genefish project.

In this chapter, we are describing the Genefish strategy, with regards to the molecular tricks employed and also to the first results obtained from the model of application.

1. The Genefish tool functioning

i. Presentation

As previously described, this strategy relies on the integration of targeted genetic elements from environmental DNA and/or extracted metagenomic DNA, into the genome of an engineered recipient strain thanks stimulated transformation. The target will most likely be any cluster of genes or genetic elements that show enough similarity in their 3' and 5' extremities in order to allow homeologous recombination. The capture process is achieved thanks to the recombinase system from the lambda phage, promoting recombination between the targeted DNA and a capture plasmid carrying the specific conserved regions needed for the reaction (Fig.1, Panel A). In this context, each application of the system requires the construction of a specific capture plasmid carrying the adequate regions sharing similarities with the target. The λ phage recombinase system is carried onto the pSIM6 plasmid which harbor a thermo-inducible promoter, thus allowing tight regulation during the induction of the recombinase proteins. Selection of the recombinant clones is achieved thanks activation of a powerful lethal cassette located in the capture plasmid between the 2 cloning sites. Clones harboring the recombinant genotype will be able to grow under toxicity induction, non-recombinant clones will most likely die as they still harbor the toxic cassette. Here we are describing each key step of the system, with molecular and construction details.

ii. Recipient strain and targeted DNA

During the first step, the recipient strain could be applied whether onto extracted metagenomic DNA thanks electro-transformation (*in vitro* approach), or directly into an environmental sample such as soil in order to bypass the extraction bias (*in situ* approach). In this context, the strategy can be improved by (1) pre-lysis of the soil bacterial community in order to release their genomic DNA into the environment, (2) and introduction of a high density of recipient strain to the sample and (3) stimulation of the transformation reaction by applying artificial electric discharge (Demaneche *et al.* 2005). In order to facilitate the workflow and experimental handling, the bacteria *Escherichia coli* K12 was chosen because of its rapid and easy cultivability and well-known genome for eventual modification. This choice was also motivated by the fact that the toxic genes selected by Nathalie Lombard were all tested in *E. coli* (Cf. section 1.iii).
Same goes for the recBCD system, extensively studied in *E. coli*, which was selected and modificate in the early stage of the tool development for increasing recombination frequency (Cf section 1.iv). All these points justificate the use of *E. coli* for being the host strain of the Genefish strategy. A complete table of all the strain used in this project is given in Appendix 3-1.

Recovery of the recipient strain can be achieved thanks a cell extraction step like density segregation at low-speed centrifugation, before undertaking the selection step (Holben *et al.* 1988). However, the probability for the recipient cell to encounter and transform the desired target whether in *situ* or *in vitro* can be relatively low, depending on the target occurrence frequency in the environment and/or in extracted metagenomic DNA. This probability can be even lower if the capture system is displaying low recombination frequency with the target, thus resulting in a considerable decrease of our chance to succeed in cloning and maintaining the targeted DNA into the strain.

As previoulsy mentionned, this problem can be offset by applying more recipient cells in order to improve our chance to encounter the target and catch it. This number can be preliminary estimated using qPCR based approaches in order to quantify the target concentration, and then adjust the appropriate number of recipient cells to be applied. On the other hand, one of the key points of the Genefish strategy is to improve the capture step. This encompass the recombination process, and also the counter selection step, which will allow the screening of positive recombinant clones, and increase our chance to get relevant results. In order to optimize the whole process and validate the concept, the system has to be tested with a highly conserved gene that is commonly found in environmental bacterial genomes. This will increase our chance to succeed in catching the target, and also to determine the frequency of both transformation and recombination steps.

iii. Optimizing the capture step

The capture of the targeted DNA is realized onto a specific plasmid: pBAD35K7tox, which has been designed by Nathalie Lombard during her PhD work (Fig.3, panel A). The backbone of the plasmid is derived from pBAD33 (Guzman *et al.* 1995), and is equipped with the origin of replication oriP15A which allows 15-30 copies per cell. The plasmid was successively mofified for integrating novel element such as the final toxic cassette. Details about the selection systems are further discussed in section 1. iv. Here we describe how the different strategies were tested to optimize the recombination step.

The molecular system responsible for the recombination process has been modified in order to maximize the chance in preserving and integrating the incoming DNA. In the case of *E. coli*, the recombination process is achieved through the RecBCD pathway, involved in the SOS system which aims for DNA reparation and stability. Homologous recombination *in E. coli* is directly relying on the enzymatic complex RecBCD, and the RecA protein (Dabert & Smith, 1997; Delmas & Matic, 2005). The RecA protein is performing the key step of the reaction between the incoming single strand DNA and the genomic DNA (Chaudhury & Smith, 1984a/b; Smith *et al.* 1984). The recombination mediated by RecBCD is strongly dependent toward specific sites, found randomly in the genome called "*chi sites*" (5'-GCTGGTGG-3') (Bianco and Kowalczykowski 1997) (Fig.2a).

The RecBCD complex is composed of three proteins: (1) RecB, which is known to be a DNA-dependant ATPase (Hickson *et al.* 1985) with a weak 3'-5' DNA helicase activity Boehmer and Emmerson, 1992). It's C-terminal part is also displaying independent exo- and endonuclease activities (Dillingham and Kowalczykowski 2008), and is known to be involved in all nuclease reaction of the RecBCD complex (Yu *et al.* 1998; Sun *et al.* 2006). (2) RecC, which is considered to be the subunit that is responsible of recognition and interaction with the *chi* sites (Handa *et al.* 1997; Arnold *et al.* 2000), and guide the ssDNA in a specific tunnel through its own structure (Dillingham and Kowalczykowski 2008). In spite it doesn't possess any active sites, RecC protein sequence shares many similarities with helicases domains and also nuclease at the C-terminal part. This suggests that this protein is a catalytically dead helicase-nuclease that arose from duplication of the recB gene (Kinch *et al.* 2005). (3) Finally, the last subunit RecD is known to be a ssDNA dependant ATPase (Chen at al. 1997) with a 5'-3' helicase activity (Dillingham *et al.* 2003).

The RecBCD complex is responsible for the cleavage of degenerated region within the genome, identifiable by the presence of free 3'-OH ends, thanks its helicase activity and its 3'-5' exonuclease activity. Degradation activity is maintained until the RecBCD complex randomly finds a "*chi"* site, after what the complex is reorganized, resulting in a loss of the 3'-5' exonuclease activity (Amundsen *et al.* 1986). However, the helicase activity of the RecBC complex is still operational, which will allow the formation of a free 3' single strand DNA (ssDNA) (Fig.2b). This ssDNA will be protected against any degrading enzymes by coating proteins such as single strain binding proteins (SSB) and RecA proteins. The ssDNA-SSB-RecA complex becomes active and is able to promote homologous recombination with any double strand DNA if sufficient homology is detected (Matic *et al.* 1996). A previous study already pointed out the necessity of a minimal identity sequence length (Shen & Huang, 1986). Depending on the phylogenetic distance between the two bacterial species, the minimal required size for the homologous regions can range from 25bp up to more than 200bp, and higher recombination rates were observed for the highest sizes (Thomas & Nielsen, 2005).

The first attempt for improving the recombination rate was initiated by Nathalie Lombard and Jean-Claude Lazzaroni, and consisted in shutting down the activity of RecD by mutation of the recD gene. Actually, recD is known to be contributing as an recombination inhibitor, as it was reported to block RecA loading on SSB coated ssDNA (Churchill *et al.* 1999; Amundsen *et al.* 2000), and as it is involved in the exonuclease activity of the RecBCD complex in the cytoplasm toward linear double strand DNA (dsDNA) (Amundsen *et al.* 1986; Palas and Kushner, 1990; Masterson *et al.* 1992). As a consequence, recD mutants were reported to display higher recombination rates (Lovett *et al.* 1988) due to higher persistence of incoming foreign dsDNA in the cell (Amundsen *et al.* 2000), and also the ability of RecBC to directly load RecA and SSB proteins on this same dsDNA during the early stage of the unwinding process (Churchill *et al.* 1999). In addition, recD mutant were identified as unable to recognize *chi* sites and apparently lose any nuclease activity, which seems to promote recombination as well (Amundsen *et al.* 2002; Smith, 2012). Shutting down the degrading activity of RecD will most likely improve the chance for homologous recombination to occur.

Unfortunately, the deletion of RecD activity also led to major instabilities in the recipient strain. Problems were encountered such as random mutations in the capture plasmid, with probable loss of the toxic cassette efficiency, as well as degradation during plasmid DNA extraction steps. We hypothesized that these problems, yet unseen, came from the recD$^-$ phenotype. Actually, the importance of exonuclease systems in *E. coli* viability were already reported (Dermic, 2006) and full inhibition of RecBCD may affect cell growth (Sergueev *et al.* 2001), even though no real evidence for cell instability were reported in RecD mutants. Furthermore, some contradictory data were reported, indicating that in some cases, inactivation of the RecD subunit is not sufficient for complete stimulation of recombination when nuclease activity is absent (Jockovich an Myers, 2001). Finally, this strategy was abandoned at early stage of my PhD work in favor to another one, based on the recombinase system derived from the λ phage.

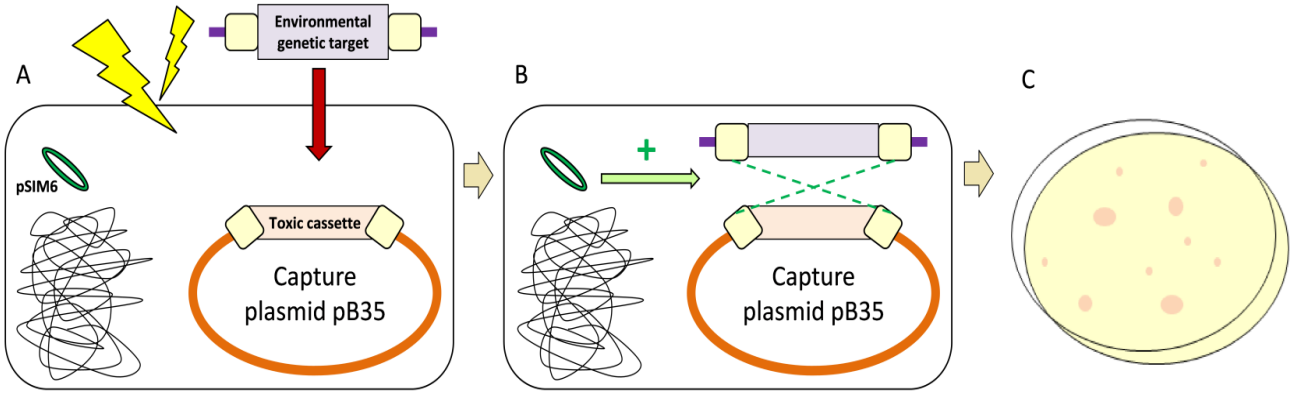


Fig.1: The Genefish strategy. Panel A represents the internalization step mediated by electro-transformation. Panel B shows the capture step through homeologous recombination thanks the λ phage recombinase system carried on the pSIM6 plasmid, and the recombination sites (yellow boxes). Panel C shows the selection step that is performed on plate throught chemical induction of the toxic genes.

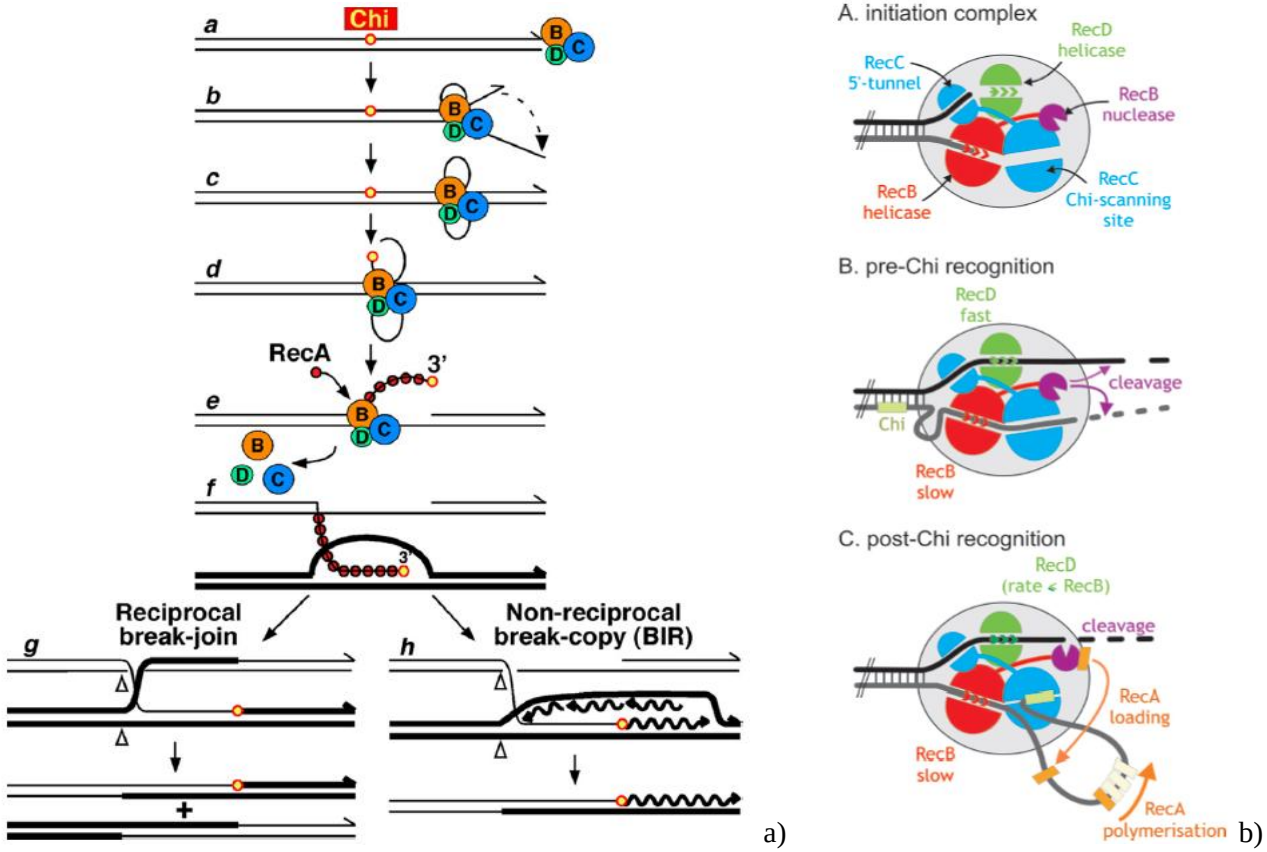


Fig.2. a) Model for recombination based on Chi recognition from Amundsen et al. 2007. Thick lines are representing one parent. The D-loop (f) can be converted into a Holliday junction (g) and resolved into a crossover. Alternatively, the 3' end of the invading Chi tail can prime DNA replication (h); the cutting of strands (open arrowheads), swapping of strands, and ligation produce one crossover-type recombinant but not its reciprocal. This mechanism is also called "break-induced replication" (BIR).

b) Model for RecBCD enzyme mechanism (Dillingham and Kowalczykowski 2008). (A) Cartoon representation of the initiation complex with specific color for each subunit (B) Pre-Chi recognition. RecBCD travels along duplex DNA powered by ATP hydrolysis in the RecB and RecD helicase motors. A loop of ssDNA may form ahead of RecB as RecD as a faster activity. The RecB nuclease domain is positioned to cleave the 3- ssDNA and 5- ssDNA tails. The 3-ssDNA tail is scanned for the Chi sequence as it passes through a tunnel in the RecC protein. (C) The Chi sequence is recognized and remains tightly bound in a tunnel in the RecC protein. The complex pauses before translocation resumes at a reduced rate. A final cleavage event takes place on the 3-ssDNA tail just upstream of the Chi sequence, resulting in generation of preserved 3' end. The 5-ssDNA strand continues to exit from the rear of the enzyme, and cleaved by the nuclease domain more rapidly. A loop of ssDNA accumulates downstream of the Chi sequence. The RecB nuclease domain recruits RecA proteins, and loads them onto the growing ssDNA loop to promote RecA nucleoprotein filament formation. This process is proposed to be triggered by the release of the RecB nuclease domain from RecC following Chi recognition.

Instead of modifying the inner recombination system of the recipient strain, another strategy has been explored consisting in using the recombinase system from the λ phage: *the lambda-red-gam system* (Poteete *et al.* 1988a/b) (Fig.3, panel B). This approach was reported to improved significantly the recombination rates than the "traditional" RecBCD mediated pathway. The gain was estimated to be ranging between 10-10000-folds depending on the lambda-red-gam plasmid constructions tested (Datta *et al.* 2006). Furthermore, the size of the homologous regions that is required for initiating recombination is significantly reduced when using this system. It was reported to be ranging from 20-50bp (Fujitani *et al.* 1995, Poteete, 2001; Court *et al.* 2002), while RecBCD mediated recombination require a minimum 20-200bp, and even more depending on the species (Thomas & Nielsen, 2005).

This strategy relies on the cloning of the 3 genes responsible for the recombination process in the λ phage (*exo, bet & gam*), into a thermo-inducible plasmid pSIM6 (Datta *et al.* 2006). Both *exo* and *bet* genes constitute the *red* system of the bacteriophage, responsible in the recombination activity. The gene *exo* encode for a 3'-5' exonuclease that will generate ssDNA from the incoming DNA. The gene *bet* encode for a coating protein that will bind to ssDNA and promote the recombination process instead of RecA. The third gene *gam* encode for a protein displaying a strong affinity for bacterial exonucleasic complexes such as RecBCD and SbcCD, and inactivate them in order to protect the foreign double strand DNA (Murphy, 1991; Court *et al.* 2007). However, for an efficient recombination, intervention of the RecA protein is needed When comparing the recombination efficiency obtained with traditional RecBCD-mediated in *E. coli*, the *lambda red gam* system shows higher recombination rates, (10-folds) (Murphy, 1998). This system has been also tested successfully with *Salmonella enterica* (Datta *et al.* 2006). Furthermore, the minimal length required to promote recombination is also significantly lower, with a minimum of 27-50bp needed to initiate the reaction (Copeland *et al.* 2001 ; Poteete, 2001). The 3 genes *exo*, *bet* and *gam* were cloned in a vector in order to be used as molecular tool for recombination purposes (Datsenko & Wanner, 2000). In this study, the author used chemically inducible promoters, under the control of specific compounds such as IPTG or L-arabinose for controlling the induction of the *lambda red gam* genes.

For the development of the Genefish tool, the chemically inducible promoters as IPTG and L-arabinose were not advisable because they were already used to regulate the toxic gene cassette (see section 1.iv). However, the thermo-inducible *lambda red gam* plasmid pSIM6 (Datta *et al.* 2006) seemed to be suitable enough and was finally selected, as it could be readily handled within the Genefish strategy. The promoter P_L is induced only at 42°C by inactivation of its repressor protein CI857, which gives a tight control over the production of λ phage recombinase proteins. Furthermore, as it harbors a thermo-sensitive origin of replication, the pSIM6 plasmid can't replicate any more after heating at 42°C, and will most likely be eliminated after a few division cycles. This origin of replication, ORI pSC101, also confers a low copy number phenotype (5 copies) (Hashimoto-Gotoh *et al.* 1981). This aspect is crucial, as it guaranties a "single-use" of the plasmid, thus avoiding potential later genomic instabilities in recipient strain due to the powerful recombinase activity of this system.

iv. The counter-selection system

Retrieving positive transformant-recombinant events out of the pool of cells recovered from the environmental sample or after an electro-transformation run is not easy to perform. The selection of the recipient strain carrying the capture plasmid is achieved by antibiotic selection, thanks to the selective marker. However, this system alone is not enough to clearly make the difference between recombinant event and false positive. Another consideration is that the cell density required should be important, as it maximizes the chance in both internalization and recombination of the target DNA. This makes any screening technique tedious and time consuming, as the amount of clones to be analyzed would be huge. In this context, phenotype-based screening techniques were considered. It can be applied only if the targeted gene product displays a remarkable activity that can be readily used as a criterion for positive clone detection. However, this strategy is not easy to set up, as it requires cloning of the full-operational sequence in the capture plasmid, and an expression in *E. coli* as well. Furthermore, such a strategy would considerably reduce the field of application of our system, by adding a new constraint.

To solve this problem, the use of a lethal cassette has been proposed as an alternative way for selecting only positive events through plate culture. The principle was to clone in the capture plasmid two inducible toxic genes between the homeologous regions in order to replace them along the recombination process with the targeted DNA. These toxic genes are useful only at the final selection step, and should be repressed under normal growing conditions. The recovery of positive clones will be achieved by chemical induction of the lethal genes only during the selection step on plates. The non recombinant recipient cells that still harbor the toxic cassette will most likely be killed under induction conditions, whereas the recombinant ones will grow as they deleted the lethal genes during the recombination process.

However, working with lethal systems is quite challenging , as it imposes many constraints. The toxic genes are generally inducing minor or major perturbations in the host cells, with a potential death issue. Toxic genes are very diverse and found in a wide range of bacterial genome, both in chromosoms and plasmids, with sometimes with a high abundance (Pandey and Gerdes, 2005; Leplae *et al.* 2011). For instance, 36 toxic genes were highlighted in the genome of *Escherichia coli* K12, causing growth inhibition and death under specific conditions (Yamaguchi *et al.* 2011). Under normal growth conditions, these genes are usually co-expressed with their respective cognate anti-toxins in operons structures called toxin-antitoxin systems, thus forming a stable inactive complex when the stoechiometry 1:1 is respected (Yamaguchi and Inouye 2009). However, free anti-toxin proteins are often displaying instability when not complexed, and are also strongly prone to degradation by stress-induced protease systems. As a consequence, an increase of free toxins in the cytoplasm may happen in stressed bacterial cells, potentially leading to death. This was observed on stressed *E. coli* population submited to various types of stress, including amino acid starvation, antibiotic pressure, high temperature and oxydative conditions (Aizenman *et al.* 1996; Kolodkin-Gal *et al.* 2007; Hayes and Van Melderen, 2011).

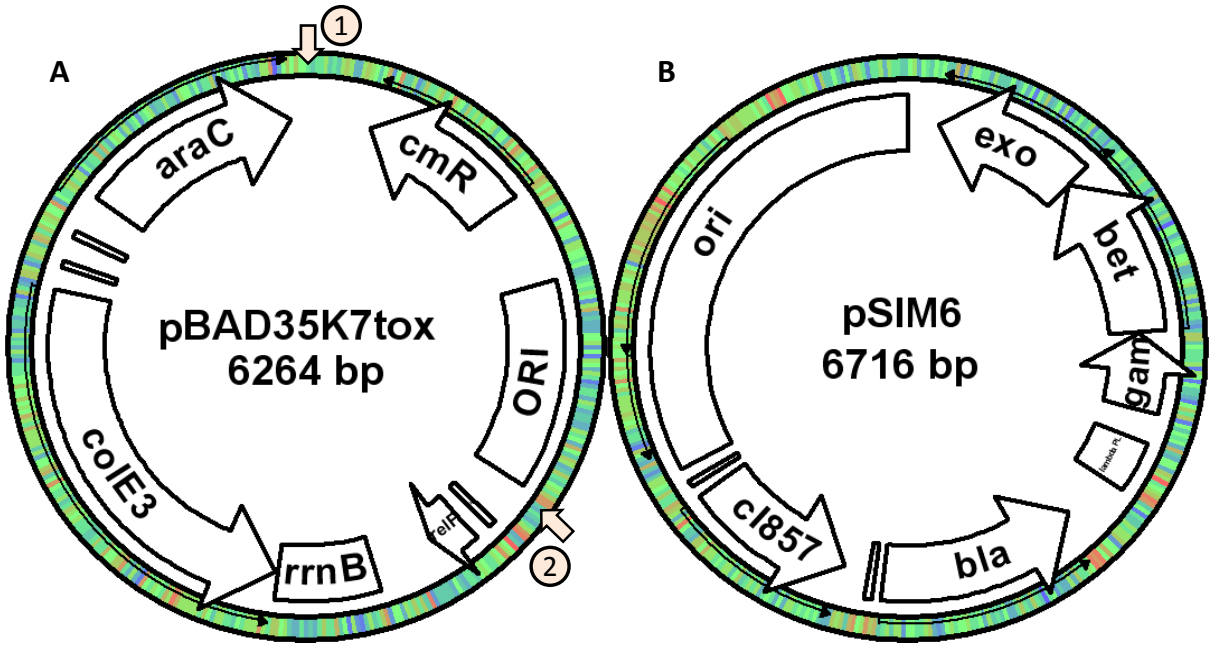


Fig.3: Genefish plasmids. Panel A represents the capture plasmid pB35 and panel B the recombination helper-plasmid pSIM6. Arrows on panel A show the multi-cloning sites devoted to integration of recombinational regions. Multi-cloning site 1 has AvrII (C'CTAG_G), PstI (C_TGCA'G) and HpaI (GTT'AAC). Multi-cloning site 2 has StuI (AGG'CCT), BglII (A'GATC_T) and PvuI (CG_AT'CG). pBAD35K7tox genes are annotated has follow: "colE3" and "relF" are the two toxic genes and "rrnB" is the terminator part of "colE3". "araC" is the repressor of "colE3" and "cmR" the selective marker (chloramphenicol resistance gene). "ORI" is the origin of replication of pB35 oriP15A (15-30 copies). pSIM6 genes are annotated as follow: "exo", "bet" and "gam" are the 3 genes from the λ phage involved in the recombination process under the control of "cI857" as a thermo-sensitive repressor. "bla" is the selective marker (beta-lactamase: carbenicillin reistance gene). "ori" is the origin of replication pSC101 (5 copies).

Toxin-antitoxin systems are a key feature of bacterial genomes, strongly regulated by cell physiology in their natural habitat, and also involved in biofilm formation (Ren *et al.* 2004) and multi-drug resistance (Keren *et al.* 2004). As they act as selfish entities, they are often considered to be part of the mobilome, which also includes phages, plasmids, transpons and integrative/conjugative element (Guglielmini and Van Melderen, 2011). Their role is still debated among the scientif community, with several hypothesis including population regulation through sacrifice under adverse environmental conditions (Engelberg-Kulka *et al.* 2006), and/or maintain of integrative conjugative element in bacterial genomes (Wozniak and Waldor, 2009).

In order to build a powerful selection system, several inducible toxic genes were tested alone or in combination of two for defining the host strain survival rates under different conditions (escape rate). In the case of a single toxic gene cloned into a plasmid under chemical induction systems, the escape rate can range from 10^{-2} down to 10^{-5} (Ahrenholtz *et al.* 1994 ; Aizenman *et al.* 1996). This 1000-fold fluctuation can be attributated to the efficiency of the toxic gene used, and also the capacity of bacteria to face and evade toxicity. Furthermore, the efficiency of inducible toxic systems was reported to be strongly correlated with the expression control. For instance, leaks in inducible systems were reported (Lanzer & Bujard, 1988), and could decrease the toxicity efficiency 100-folds (Szafranski *et al.* 1997). Strategies consisting in the cloning of two toxic genes were successfully tested in order to reduce the escape rate of the host strain down to 10^{-6} (Torres *et al.* 2003). After several attempts in combining toxic genes, Nathalie Lombard selected an association of *colE3* and *relF*, which encode respectively for the colicine E3, with a strong ribonuclease activity which inhibits the protein synthesis (Bowman *et al.* 1971), and the RelF porine which causes membrane disruption (Knudsen *et al.* 1995). These two genes were placed under the control of inducible promoter, thus allowing an accurate handling and activation of their transcription in *E. coli.*

colE3 toxic gene was cloned under the control of the P_{BAD} promoter from the arabinose operon (Lutz & Bujard, 1997). Under normal growth condition, the regulatory protein AraC is produced and bind to operating sequences O2 and araI, thus causes the operon folding and thereby secures P_{BAD} and *colE3* transcription (Fig.4). AraC is also known to be a self-regulator, as it induces its own transcription when the level of its gene product expression is low. However, the control over P_{BAD} promoter is not total, resulting in *colE3* transcription leaks. Colicin E3 is known to be an highly toxic protein, because a single molecule released in the cytoplasm is sufficient to induce severe damage to the cell functioning. Colicin E3 displays a ribonuclease only at the C-terminal domain (CRD) that cuts the 16S ribosomal RNA at a specific site in the 70S ribosome (Ohashi-Kunihiro *et al.* 2006). In order to counter the colicin E3 leaks from P_{BAD}, the related anti toxin gene immE3 (Jakes *et al.* 1974) has been cloned into the chromosome of the recipient strain, under the control of P_{BLA}, a constitutive promoter with a low expression rate (Masaki and Ohta, 1982) (Fig.4, panel A). This modification results in stabilizing the recipient strain, with a slight delayed growth rate when compared to the original strain without the capture plasmid. Under induction of lethal conditions, L-arabinose is introduced into the culture media, resulting in formation of a L-arabinose/AraC complex. This complex has no more affinity for operating sequence O2 and araI, but activates operating sequence O1 and the cAMP-CRP domain, resulting in operon rearrangements which unbind and activate the P_{BAD} promoter.

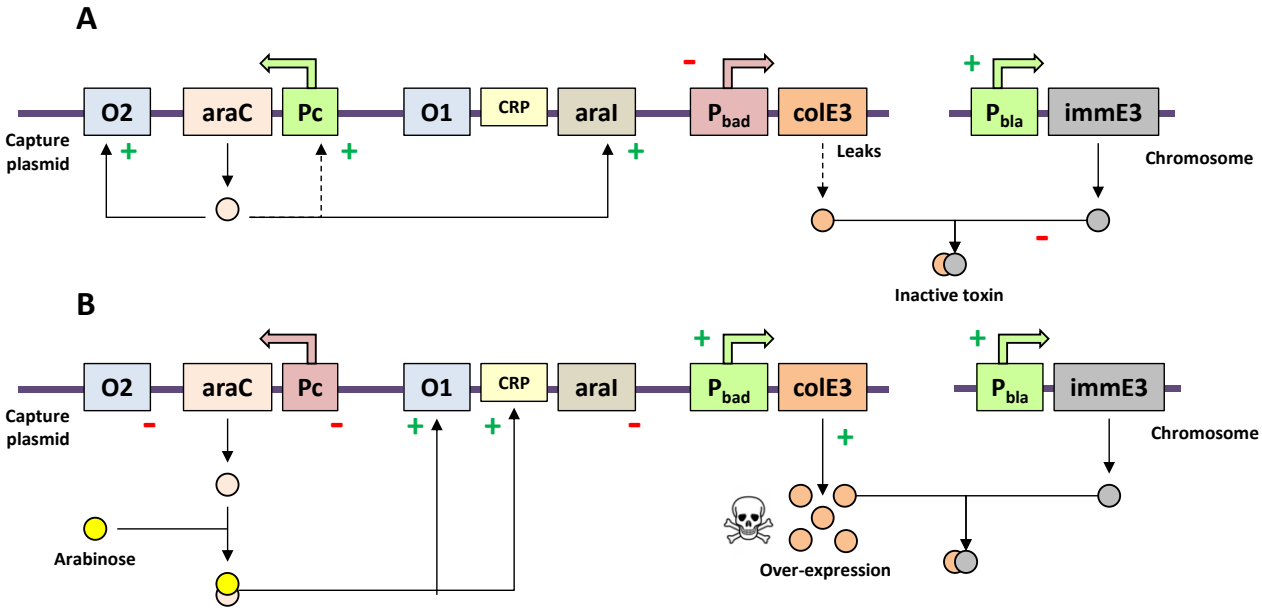


Fig.4: Regulation of colE3 toxic gene expression. Panel A displays regulation system under normal growth conditions while panel B depicts toxic gene expression under induction growth conditions. Green positive signs represent gene induction and positive feedback on operating sequences (O1, O2 and CRP) while red minus signs show promoters and gene inhibition.

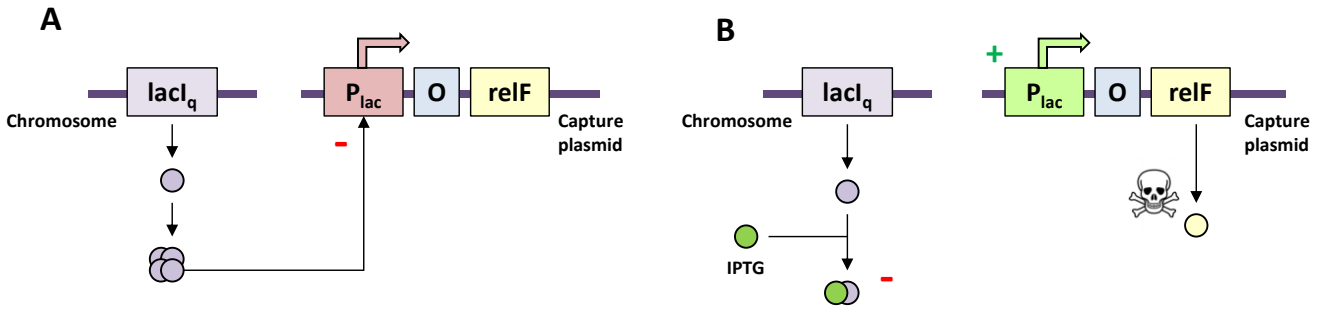


Fig.5: Regulation of relF toxic gene expression. Panel A displays regulation system under normal growth conditions while panel B represents toxicity expression under induction growth conditions. Green positive signs show gene induction while red minus signs represent promoters and gene inhibition.

This induction leads to massive production of colicin E3, overwhelming the ImmE3 antitoxin and leading to cell death (Fig.4, panel B). Unfortunately, the former *E. coli* strain already includes the *ara* operon, which is interfering with our toxic system due to background enzymatic degradation while inducing the lethal conditions. Also, generation of an *E. coli* strain harboring an Ara^- phenotype was required in order to keep a tight control over colicin E3 expression with L-arabinose. Details on employed strains and their chromosome modifications are further described in section 2.i.

The *relF* toxic gene has been cloned under the control of the $P_{A1\text{-}03/04}$ promoter from the lactose operon (P_{lac}) (Lutz & Bujard, 1997). Under normal condition, this promoter is repressed by the tetramer product from $lacI_q$ gene, which was cloned in the chromosome of the recipient strain with a constitutive expression (Fig.5, panel A). Under induction conditions, isopropyl β-D-1-thiogalactopyranoside (IPTG) is added to the growth media, resulting in the formation of an inactive complex $LacI_q$-IPTG (Griffiths *et al.* 1999). This allows the activation of the P_{lac} promoter and expression of the RelF porine, leading to membrane disruption and cell death (Fig.5, panel B).

v. Recombination sites: design and cloning

The purpose of the Genefish tool is to be adjustable to a broad range of applications. In this context, the plasmid has been designed with multi-cloning sites on both sides of the toxic genes cassette for more convenience during the cloning of recombination regions (Fig.3, panel A). The 2 cloned sequences should be conserved enough to initiate homeologous recombination with the environmental target DNA. The required homology level and length between the bait sequences and the target will be discussed later on. Furthermore, these sequences should not share any similarity with the recipient strain genome to avoid any potential undesired background recombination between the capture plasmid and the recipient strain's chromosome. These aspects should be carefully taken into account before undertaking any application with this tool. Considering this, the first model was designed to capture genes from the nitrate reduction operon. These genes are commonly found in bacterial genomes, as nitrate respiration is a key step for survival in anaerobic conditions (Richardson *et al.* 2001). For instance, the *narG* gene, which encode for the α-subunit of the nitrate reductase, is found in approximately 30% of known bacteria according to BLAST searches (Laurent Philippot, personal communication). Furthermore these genes are well conserved and organized into operon, thus making the choice and design of the recombination sites easier (Fig.6).

Operon organization is also a key feature for the model construction, as it will be easier to modify the homology and/or displace the conserved sites in order to increase or decrease the target length, thus giving important information on the size limit tolerated by the recombination machinery. The final sequences were selected in the *narGH* genes, and amplified from *Pseudomonas fluorescens* (Fig.7). The 5'-site was designed at the 5'-end of the *narG* sequence, and the 3'-region in the 5'-side of narH. These regions were amplified with primers flanked with adequate restriction sites at the 5'-end for the cloning into the capture plasmid (Fig.8). This step was carried out at l'INRA (Dijon, France) under the supervision of Laurent Philippot. List of primer used are given in Appendix 3-1.

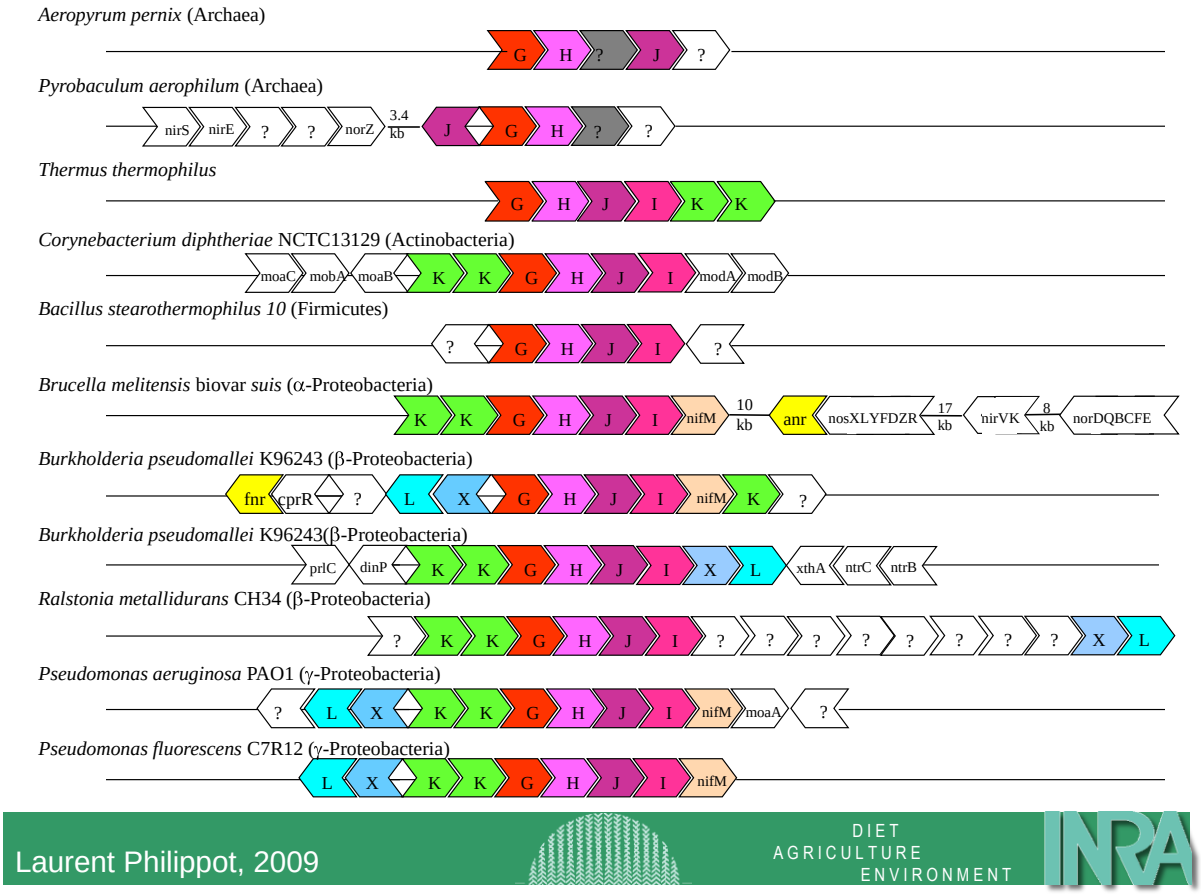


Fig.6. Schematic representation of the nar operon among Bacteria phylums. The figure describes how conserved the nar operon is, especially toward the narGH region. This figure was provided by Laurent Philippot, research director at French institute for diet, agriculture and environment (Microbiologie des Sols-Géosols, UMR1324 CSGA Centre des Sciences du Goût et de l'Alimentation, INRA: Institut National de la Recherche Agronomique, 21065 Dijon cedex, France)

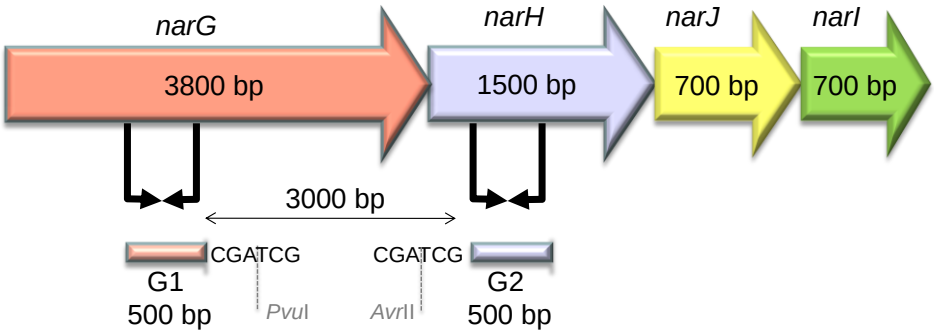


Fig.7. Pseudomonas fluorescens nar operon. The figure displays the operon organization and the genetic regions that were selected for recombination site G1 and G2 desigg. This figure was provided by Laurent Philippot, research director at French institute for diet, agriculture and environment (INRA, Institut National de Recherche Agronomique)

The size of the cross-over regions is approximately 500bp for each domain. Length of homologous regions is known to be an important criterion during the recombinering process, varying a lot: from very short spots (25bp) up to 200bp and more depending on species (Thomas & Nielsen, 2005). Furthermore, it was reported that recombination rates are clearly increased when several homologous regions are getting involved (de Vries & Wackernagel, 2002). This confort our choice for having two homologous regions surrounding the toxic genes, as it would simultaneously improve the recombination rates and disrupt the lethal cassette.

In addition, as previously discussed, the pSIM6 mediated recombination was reported to be initiated with shorter recombination sites (20-50bp) (Fujitani *et al.* 1995, Poteete, 2001; Court *et al.* 2002). This consideration has motivated the development of new shorter recombination regions for testing the limit of our system. The new recombination sites were designed from the former ones *narG1* and *narG2*: *narG3* and *narG4* with a size of 150bp, and also *narG5* and *narG6* with a size of 60bp. These regions were cloned into the capture plasmid as well (Tab.1).

Unfortunately, the nar operon is present in the genome of our *E. coli* recipient strain, thus obliging us to delete *narG* and *narH* before testing. The disruption of these genes was necessary to avoid undesired potential recombination between our capture plasmid and the chromosome. The strategy applied for removal of *narG-H* is relying on a gene disruption technique involving homologous recombination (Duret *et al.* 1999). This deletion step is fully explained in *Chapter 3 protocol appendix 3.2*. In addition, as this dirsuption step involved recombination, the efficiency of the pSIM6 plasmid with our recipient strain was also tested simultaneously. The results of pSIM6 induction with our different strains are presented in section 3.ii.

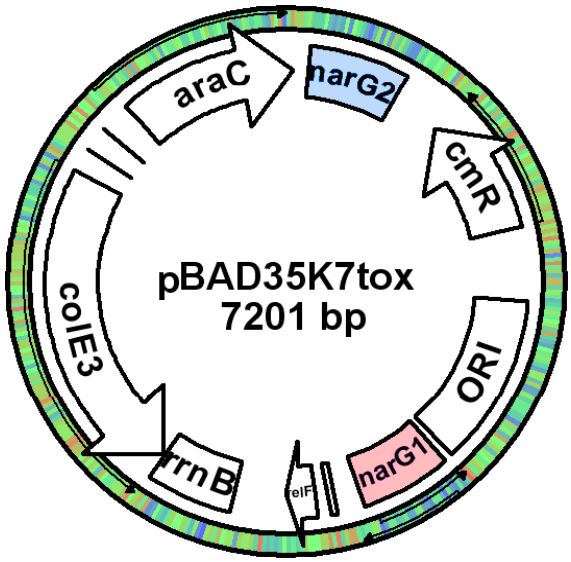


Fig.8. Model of the capture plasmid designed for narG fishing. The recombination regions were cloned isiden the 2 multi-cloning sites, narG2 in site 1 and narG1 in site 2 respectively.

2. Results and discussion

i. Preface to results

In order to test our system, several key aspects have been tested before. First aspect was to test the novel strategy for the recombination frequency improvement within our recipient strain. The recD mutation strategy was abandoned because of genomic instabilities and delayed growth. The replacement solution consisted in using derived recombinase system from phage λ, carried on a thermo-inducible plasmid pSIM6 (Datta *et al.* 2006). Preliminary experiments were realized in order to establish whether this system matched or not with our recipient strain, and with our capture plasmid. The second aspect was to maintain the best selection system along the construction of the pB35 plasmid carrying the recombination sites for *narG* fishing. This was achieved with regards to the fact that the final construction must be as efficient as the original one in terms of toxicity selection. Constant toxic screening at each cloning step is needed in order to select the best clone candidates harboring the better toxic construction. Last step was to merge these two aspects, respectively the recombination plasmid pSIM6 with the so-constructed capture plasmid harboring the best selection system. Several strategies were tested in vitro for defining and optimizing the handling of the Genefish tool, and also evaluate the capacities of our final system.

The first one consisted in a tandem use of pB35G1G2 and pSIM6 for capturing *narGH* PCR product. The second one consisted in maintaining only pSIM6 into the recipient strain, and co-transformed pB35G1G2 and *narGH* PCR product in controlled molecular ratio. The third and last one is based on the same co-transformation principle, but followed by a direct plasmid extraction and re-transformation step in order to spread and segregate molecular copies of the capture plasmid in new recipient strain clones. These different strategies were applied in order to adapt to technical limitations such as toxic cassette integrity and capture plasmid copy numbers. For more details on these Genefish utilization protocols, see the *Chapter 3 protocol appendix 3.6, 3.7 and 3.8.*

ii. The λ phage recombinase strategy

The compatibility of the λ red gam system with our system has been tested in order to verify if recombination is efficiently improved. The pSIM6 plasmid was used in our recipient strains in a gene disruption experiment. As previously mentioned, the chromosomic *narGH* genes needed to be removed in order to avoid potential undesired recombination between the capture plasmid and the chromosome (Appendixe 3-3). This step requires the use of pSIM6-mediated recombination to eliminate *narGH* from the chromosome by integrating a selective marker. This approach is also known as “gene knocking out” as the targeted genomic region will be erased during the recombination process (Duret *et al.* 1999). To do so, the oncoming marker is flanked with region sharing sequence homology with the target to be knocked out. In our case, the targeted gene to be deleted is *narG* from the genome of our recipient strain, as fully explained in Appendixe 3-3. In this experiment, the pSIM6 mediated recombination was tested on the 4 recipient strains available: TELI recD^{+}, TELI recD^{-}, and 1661LI recD^{+} and 1661LI recD^{-}.

As represented in Fig.9, the induction of pSIM6 plasmid resulted in a constitutive increase in the recombination frequency for both TELI recD$^-$ and 1661LI recD$^+$ strains, with a global gain of 2-3 folds. This experiment permitted to verify the efficiency of the pSIM6 plasmid with the 1661LI recD$^+$ strain, confirming the previous results obtained by Datta *et al.* 2006. As previously explained in section *1.iii*, the recD$^-$ phenotype was not selected because of observed unstability problems.

iii. Lethal-selection system: efficiency and stability

After any modification in the capture plasmid, such as insertion of a recombination site, the toxic cassette integrity needs to be tested. Maintaining toxic genes in a living cell is not something easy to achieve, especially with *colE3*, which is extremely toxic. Even with the genomic introduction of the immE3 antitoxin gene and its constitutive expression, the recipient strain harboring the capture plasmid displays a slowered growth comparing to the original one. This is most likely due to the expression leaks from *colE3*, even though the produced proteins are blocked by ImmE3 antitoxins. The final system is a compromise between slowered growth and an efficient suicide system. However, this system is displaying stability problems that can be observed after prolonged liquid culture. When the culture reaches the exponential state, the rapid cell division will most likely result in the apparition of a RelF$^-$ and ColE3$^-$ phenotype. This phenotype has a serious advantage in terms of growth rate, thus resulting in a progressive domination of the ColE3$^-$ among the clonal population of our recipient strain. Surprisingly, the sequencing of the toxic cassette from clones that evade toxicity induction revealed systematic deletion starting from *narG1* and removing *relF*. Sometimes larger deletion may affect partial *colE3* sequence, implicating total deletion of the toxic cassette in the larger ones. This is probably due to plasmic rearrangements that occur during the liquid culture. Details are given in Appendixe 3-10 about escaping phenotypes and deletions.

In order to counter this issue, the toxic construction needs to be regularly tested and validated into the recipient strain. Only the best clones that will display strong toxicity expression and massive cell death on plate will be selected for further experiments. Usually, the best clones we have selected were harboring *ER values* ranging from 1E.10^{-8} down to 1E.10^{-9}, which will minimize the false positive occurrence. The selection could be refined among the best candidates by evaluating the *Er* of each toxic gene separately on plate. Typical *Er* values obtained for each genes are summarized in Table 1. The construction of the plasmid was carried in the *E. coli* 1661LI recD+, as this strain displays the highest recombination frequency with pSIM6 (section 3.ii). However, stability problems were encounter with the final plasmidic construction, because of the genomic presence of the *narG-H genes*. The construction was transfered in the strain *E. coli* TELI 13755b Δ*narG* for the first recombination test until the deletion of *narG* was completeted in 1661 LI. Then the construction was transferered again in the new strain *E. coli* 1661 LI 14102 recD+ Δ*narG* before testing other recombination strategy.

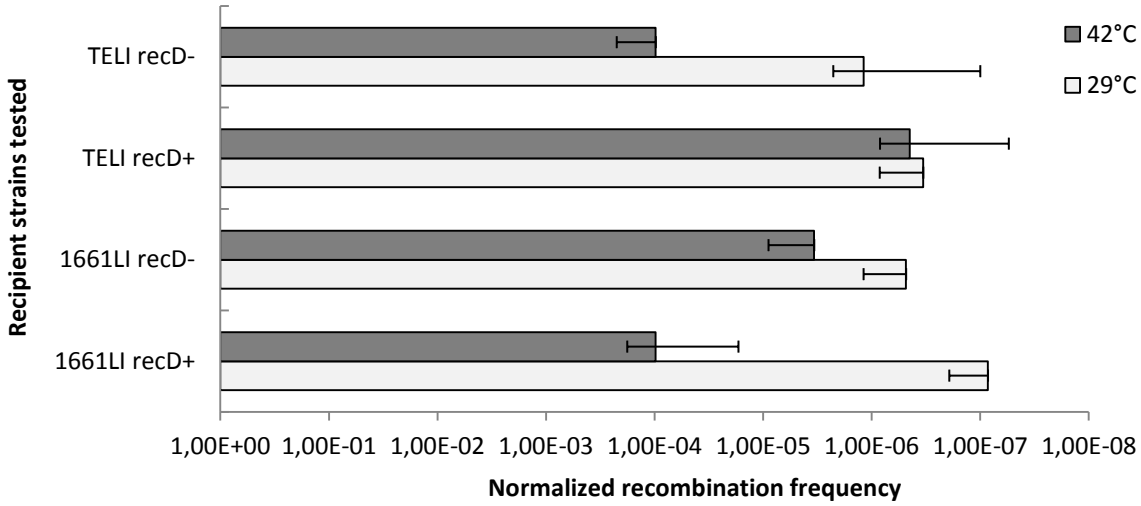


Fig.9. Transformation and recombination frequency observed after pSIM6 induction. Panel displays the λ red gam induced condition (dark grey) versus the non-induced (light grey). Recombination frequencies were calculated from observed grown colonies on chloramphenicol media, and results were normalized against the basal level of transformation observed on control plasmid (pB342). Error bars were calculated from 3 replicate experiments.

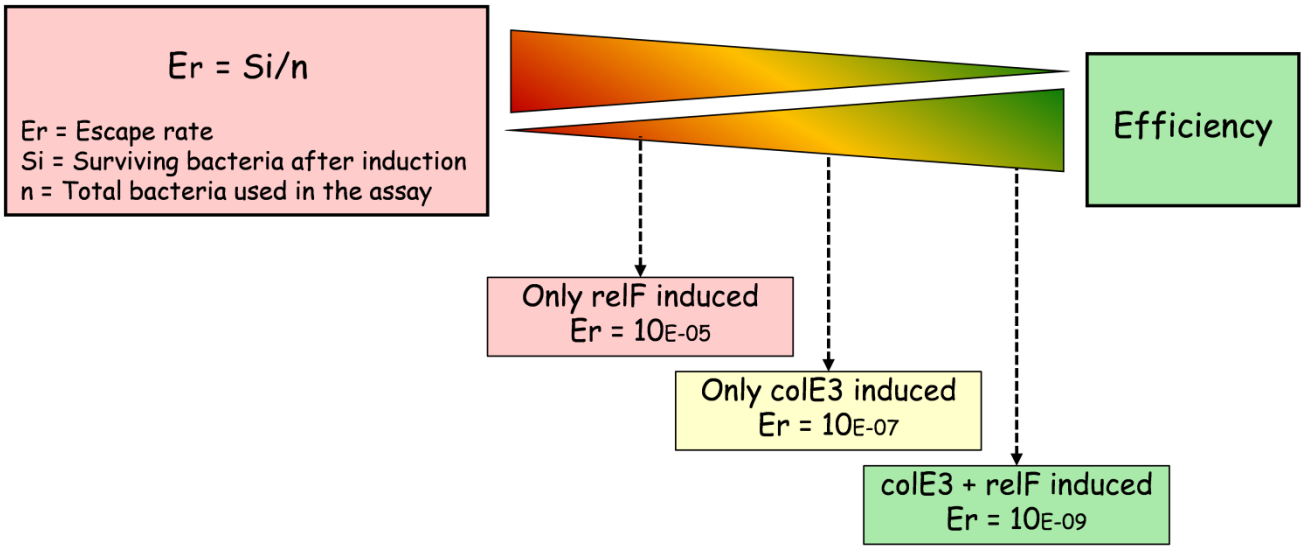


Fig. 9: Representation of the typical escape rate results obtained from toxicity induction experiments. The escape rate (Er) quantifies the number of surviving bacteria after toxicity induction (Si) among the total number of bacteria present in the culture test(n). The smaller Er, the better the efficiency of the system.

iv. Testing the Genefish tool: the tandem approach

The final construction pB35G1G2 was transformed in the *E. coli* TELI 13755b recD^{+} strain and the best clone harboring the lowest Er was selected after induction tests. The plasmid pSIM6 was then transformed into this recipient strain and stored (Tab.1). Both plasmids were co-maintained with their respective antibiotic resistance marker. The tandem strategy protocol is fully-described in the *Chapter 3 protocol appendix 3.6*, and schematically summarized in Fig.11.

The results obtained from this strategy are summarized in Table 2. Several condition were tested by changing the concentration of the *narG* PCR product (5, 12.5, 25, 50, 75, 100, 125ng per assays, corresponding respectively to 1.11E+09, 2.78E+09, 5.55E+09, 1.11E+10, 1,67E+10, 2,22E+10 and 2,78E+10 molecular copies of *narGH* PCR product). On the first attempt, the escape rate was found to be relatively high compared to the values obtained from previous tests. This is probably due to growth problems impacting on the evaluation of the actual number of total bacteria cells used in this assay. No effect of pSIM6 induction were observed in this experiment, as the escape rate were found to be the same between induced condition (42°C) and non induced (29°C). However, it seems that the concentration of *narG* has an effect on the escape rate, with lower frequency detected with the higher concentration (125ng per assay). Recombinants were detected by *narG* PCR only for the condition corresponding to 75ng of *narG* PCR product, with only 1 positive clone over 6 tested (16.7%). Another attempt was realized with only one concentration of *narG* PCR product (125ng) (Tab. 3, repetition 2). In this experiment, the escape rate was found to be closer than previous experiments. However, an effect of pSIM6 was observed, with higher escape rates detected after induction. A third experiment was realized at different concentrations as well. Here again, the highest concentration (100ng per assay) resulted in lower escape rates frequencies. Unfortunately, no induction controls (29°C) were done this time. However, recombinants were also detected by PCR among growing bacteria, only at the condition corresponding to 25ng of *narG* PCR product (2/7, 28.6%). The validation protocol for putative recombinant clone is presented in Appendix 3-10.

The tandem approach was succesfull as it resulted in detection of recombinant clones among the growing bacteria after toxicity induction on plate. These results validate the approach of catching foreign DNA on a designed capture plasmid, in a highly recombinant *E. coli* strain. However, the recombination frequencies observed are extremely low (1.0E-08 to 1.0E-09). The most probable hypothesis that might explain these results is the relatively multiple copy number of the toxic plasmid per host cell. This is due to the presence of the origin of replication ORIp15A which allow 15-30 copies of the plasmid in the host cell. Even if the recombination occurs on one copy of the plasmid, the remaining copies carrying the toxic genes will most likely be induced, thus resulting in potential death of positive recombination events (Fig.12). In order to prevent this problem, a new construction is being developed based on a low copy plasmid backbone. In addition, other startegies were used to limitate the impact of plasmid copy number, such as the co-transformation and the segregation strategies.

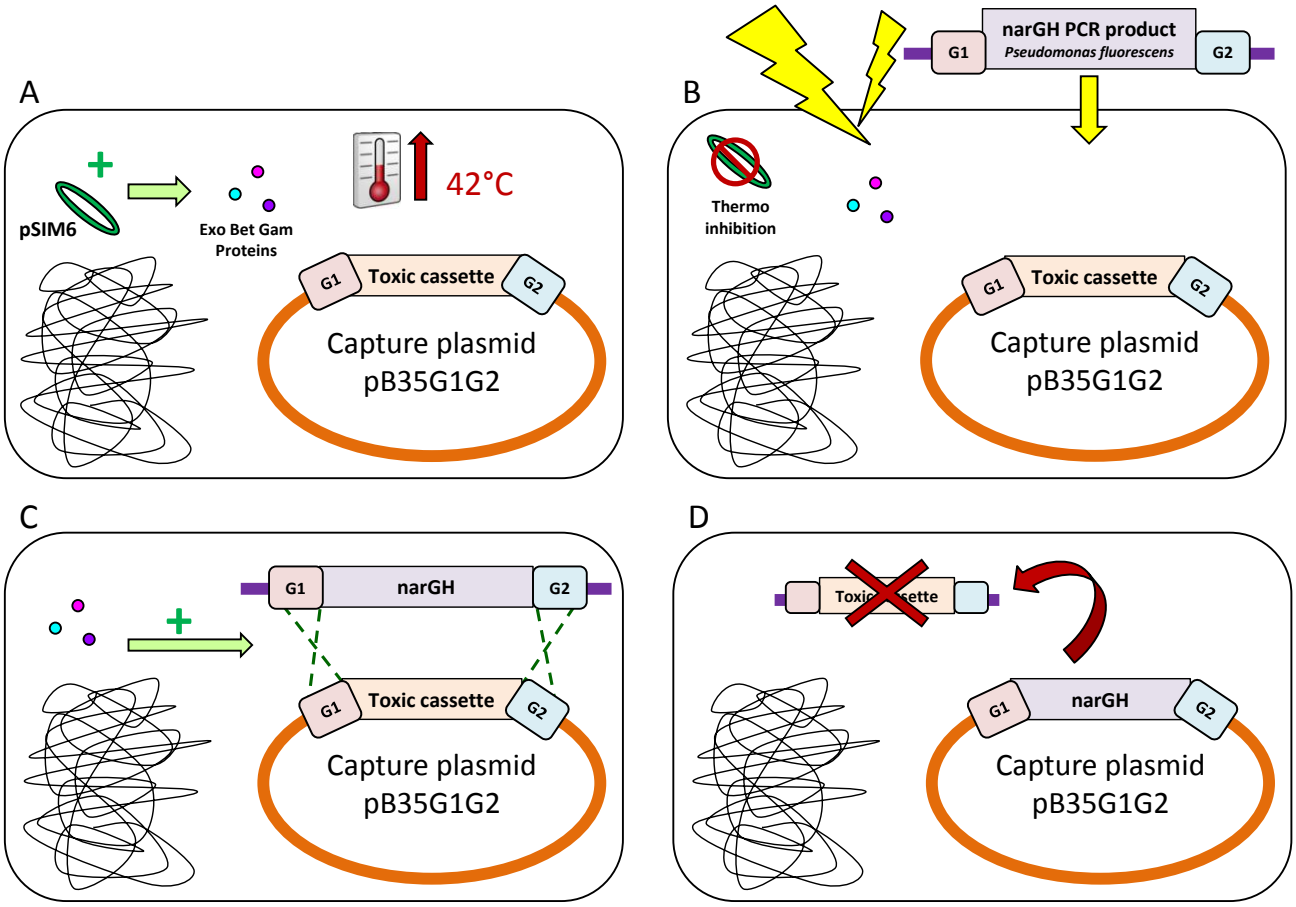


Fig.11: Schematic description of the tandem strategy. Panels chronologically describe the workflow. Panel A displays the thermo-induction of pSIM6 at 42°C, resulting in transcription of the Exo, Bet and Gam proteins. Panel B represents the thermo destruction of pSIM6 and the electro-transformation of the narGH PCR fragment from Pseudomonas fluorescens. Panel C describes the recombination capture step performed by the λ red gam proteins Exo-Bet-Gam at the homologous regions G1 and G2. Panel D shows the toxic cassette removal and final state ready for the selection step.

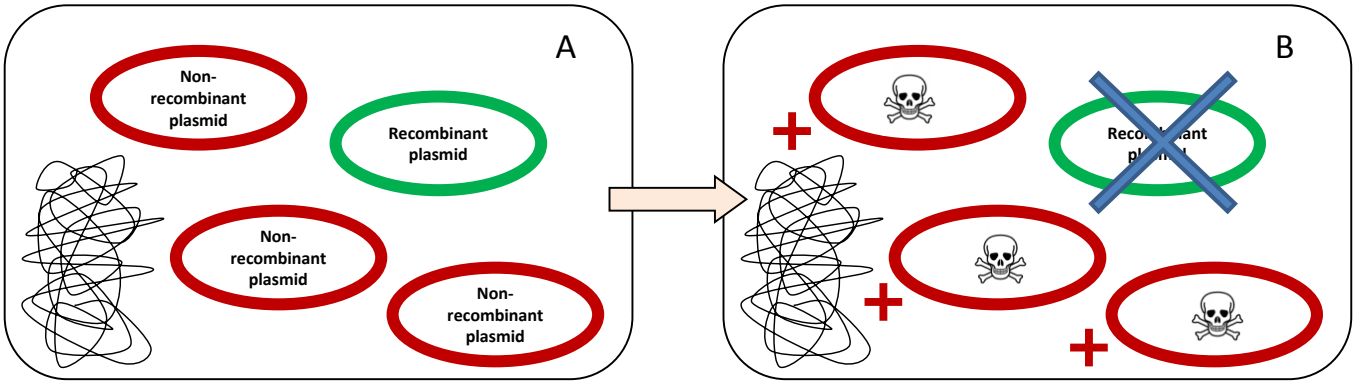


Fig.12: Schematic representation of the capture plasmid copy number limitation in tandem recombination attempts. Panel A shows the probable status of the host cell after recombination, harboring both recombinant events and non-recombinant as well. The non-recombinant plasmids are still carrying the toxic cassette, thus leading to an undesired killing process of the positive events coming along the toxicity induction (Panel B).

Repetition	Date	Manipulation	Strain	Escape rate (Er)			Notes
				colE3	***relF***	***colE3 + relF***	
1	25/06/2009	narG1 ligation	*E. coli* 1661LI 13421 recD+	4,93E-06	2,76E-05	3,18E-09	Ligation succeeded, Er OK
2	30/06/2009			9,69E-06	3,18E-05	3,18E-09	
3	21/08/2009	*narG2* ligation	*E. coli* 1661LI 13421 recD+	8,37E-06	3,53E-04	1,03E-06	Ligation succeeded, problem with total bacterial growth
4	26/08/2009			7,83E-04	1,49E-02	1,12E-05	
5	27/08/2009			-	-	8,17E-08	Growth problems solved
6	03/09/2009			9,62E-05	7,92E-05	8,86E-07	Problem with the cassette due to parasite recombination with genomic *narG*
7	10/09/2009			4,13E-05	4,61E-03	6,99E-08	
8	18/09/2009			3,53E-06	3,53E-03	2,12E-08	
9	08/10/2009	Transfert in TELI	*E. coli* TELI 13755b recD+ Δ*narG*	7,17E-08	3,91E-04	7,17E-08	Transfert succeeded in a Δ*narG* strain, Er OK
10	21/10/2009	pSIM6 Transformation	*E. coli* TELI 13755b recD+ Δ*narG*	1,51E-08	7,40E-07	1,01E-07	Growth problems, selective markers (Cm+ Cb)? pSIM6?
11	05/11/2009	Selection tests	*E. coli* TELI 13755b recD+ Δ*narG*	-	-	9,00E-08	Selection of the best candidate for the tandem strategy
12	12/11/2009			-	-	1,66E-09	
13	17/11/2009			7,94E-06	2,38E-08	1,98E-10	
14	18/01/2010	Transfert in 1661 LI	*E. coli* 1661LI 14101 recD- Δ*narG*	1,14E-09	2,52E-07	<1,14E-09	Transfert of the final construction in 1661 LI Δ*narG*
			E. coli 1661LI 14102 recD+ Δ*narG*	1,45E-07	5,39E-07	1,45E-09	
15	20/01/2010	Selection of the best clones	*E. coli* 1661LI 14101 recD- Δ*narG*	2,49E-07	1,77E-05	1,38E-09	Er too high with the 14102 recD+ strain
			E. coli 1661LI 14102 recD+ Δ*narG*	1,38E-05	1,15E-03	2,76E-07	
16	22/01/2010		*E. coli* 1661LI 14102 recD+ Δ*narG*	1,05E-07	7,35E-06	4,20E-09	Er Ok for the 14102 recD+
17	28/01/2010	pSIM6 Transformation	*E. coli* 1661LI 14101 recD- Δ*narG*	6,52E-06	4,26E-01	<1,63E-07	Er OK for each strain, but high for *relF* in 14101 recD-. To be confirmed
			E. coli 1661LI 14102 recD+ Δ*narG*	1,92E-08	3,84E-06	<1,92-08	
18	10/02/2010	Selection of the best clones	*E. coli* 1661LI 14101 recD- Δ*narG*	3,29E-08	4,65E-04	<3,29E-08	Confirmation of the good Er values for each strain. Confirmation of the higher Er for *relF* in 14101
			E. coli 1661LI 14102 recD+ Δ*narG*	5,91E-08	5,91E-08	<5,91E-08	
19	16/04/2010	narG3 ligation	*E. coli* 1661LI 14102 recD+ Δ*narG*	<6,45E-10	1,41E-06	<6,45E-10	Construction of 2 new capture plasmids: pB35G3G4 and pB35G5G6
		narG5 ligation		<1,12E-09	2,71E-09	<1,12E-09	
20	01/06/2010	narG4 ligation	*E. coli* 1661LI 14102 recD+ Δ*narG*	7,43E-08	3,72E-07	9,29E-10	Ligation of G4 in pB35G3
21	03/06/2010	narG6 ligation	*E. coli* 1661LI 14102 recD+ Δ*narG*	1,93E-07	1,66E-06	3,85E-10	Ligation of G6 in pB35G5
22				5,98E-07	3,02E-05	8,55E-10	Ligation of G6 in pB35G3

Tab.1: Summary of the different steps in constructing the final capture plasmid pB35G1G2.

Repetition	Date	[*narG*] (ng)	Strain	Escape rate (Er)		Er normalized per µg DNA		Notes
				29°C	42°C	29°C	42°C	
1	18/11/2009	0ng	*E. coli* TELI 13755b recD+ Δ*narG*	2,85E-04	2,63E-04	-	-	One recombinant detected only at 75ng (3,7% positives, 1/27)
		12,5ng		5,08E-05	-	4,06E-03	-	
		75ng		7,69E-05	8,13E-04	1,03E-03	1,08E-02	
		125ng		4,31E-05	2,13E-05	3,45E-04	1,70E-04	
2	25/11/2009	125ng	*E. coli* TELI 13755b recD+ Δ*narG*	4,64E-08	8,97E-06	3,59E-06	7,18E-06	No recombinant detected
3	22/12/2009	0ng	*E. coli* TELI 13755b recD+ Δ*narG*	-	2,26E-08	-	-	Growth problems due to extended incubation time. 28,6% positives (2/7, 25ng)
		5ng		-	7,65E-09	-	1,53E-06	
		25ng		-	6,54E-08	-	2,62E-06	
		50ng		-	1,28E-07	-	2,56E-06	
		100ng		-	5,09E-09	-	5,09E-08	
4	11/02/2010	40ng, 7min, 42°C	*E. coli* 1661LI 14102 recD+ Δ*narG*	<7,47E-08	7,13E-08	1,87E-06	1,78E-06	Test 1661 LI and pSIM6 incubation time. No recombinant detected
		40ng, 15min, 42°C		<1,12E-07	2,00E-06	2,80E-06	5,00E-05	
5	17/02/2010	0ng, 7min, 42°C	*E. coli* 1661LI 14102 recD+ Δ*narG*	2,36E-07	7,80E-08	9,44E-06	3,12E-06	No effect of incubation time detected. No recombinant detected
		25ng, 7min, 42°C		3,55E-08	2,20E-09	1,42E-06	8,80E-08	
		25ng, 15min, 42°C		7,82E-08	5,24E-08	3,13E-06	2,10E-06	

Tab.2: Summary of the results obtained from recombination attempts with the tandem strategy. Escape rates are established against total bacteria counts for each condition, and normalized per microgramme of DNA to croos-compare conditions together.

v. Testing the Genefish tool: the co-transformation strategy

This strategy has been set up for better control over the copy number of the capture plasmid during recombination attempts. In this experiment, the *E. coli* strain is grown with pSIM6 only, and the capture plasmid is brought along the transformation of the *narG* PCR product. The schematic description of this approach is presented in Figure 12. The quantity of *narGH* PCR product and pB35G1G2 are adjusted at the appropriate molecular ratio depending on the experiment design. The summary of all co-transformation attempts is given in Table 3. Ideally, one copy of capture plasmid per cell should be transformed during the process. However, after preliminary attemps, the molecular ratio 1:1 wasn't enough to efficiently incorporate the plasmid in the cells. The usual ratio used in experiment corresponds to two or three copies of capture plasmid per bacterial cell, and between 3-7 copies of *narG* PCR products per cell. The number of bacterial cells used in experiment for molecular ratio calculation was estimated thanks optical density at 600nm measures (e.g 8.00E08 bacterial cells per milliliter for an optical density of 0.6). The co-transformation strategy protocol is described into details in the *Chapter 3 protocol appendix 3.7*.

As previously observed with the tandem approach, this strategy was also successful in detecting recombinants. In addition, the frequency of positive event detection was also increased, event though the actual number of potential recombinant cells was lower than in the tandem approach, due to the additional transformation step of the capture plasmid. For instance, in the tandem approach, 1.00E09 bacterial cells carrying the capture plasmid pB35G1G2 were used per assays, resulting in low escape rate between 1.53E-06 down to 5.09E-08 per µg of *narGH* PCR product transformed. On the the other hand, the co-transformation approach results in significantly higher escape rates, from 1.14E-03 down to 9.62E-06 per µg of DNA transformed (pB35G1G2 + *narGH* PCR product). This is mostly due to the additional transformation of the capture plasmid, which considerably reduces the actual number of potential hook cells. However, the co-transformation strategy resulted in significantly higher number of growing colonies after induction while compared to the tandem approach and after all, the actual number of positive event was higher with this approach after PCR verification.

However, this strategy has also its limitations. The workflow requires preliminary preparation of high quality capture plasmid pB35G1G2, with short time growing steps and higher culture volumes in order to respectively limitate emergence of toxicity escaping phenotypes and to get enough material for the low-copy plasmid DNA extraction protocol. Extended culture periods result in the occurrence of mutations into the toxic genes sequences and plasmidic rearrangements that delete part of the toxic cassette, as explained in section 3.*iii*. In this case, higher escape rates are observed after co-transformation which considerably reduce our chance to find positive events among the pool of growing bacteria. In fact, several additional attempts were done with another pB35G1G2 preparation with extended culture times, resulting in a considerable increase of the escape rate, and reducing significantly the detection threshold when searching for recombinants (data not shown). The final protocol established allowed detection of approximately 10% of recombinant per assay (10/100 tested clones), which corresponds to a ratio of 2 copies of pB35G1G2 per bacterial cell and 5 copies of *narGH* PCR product.

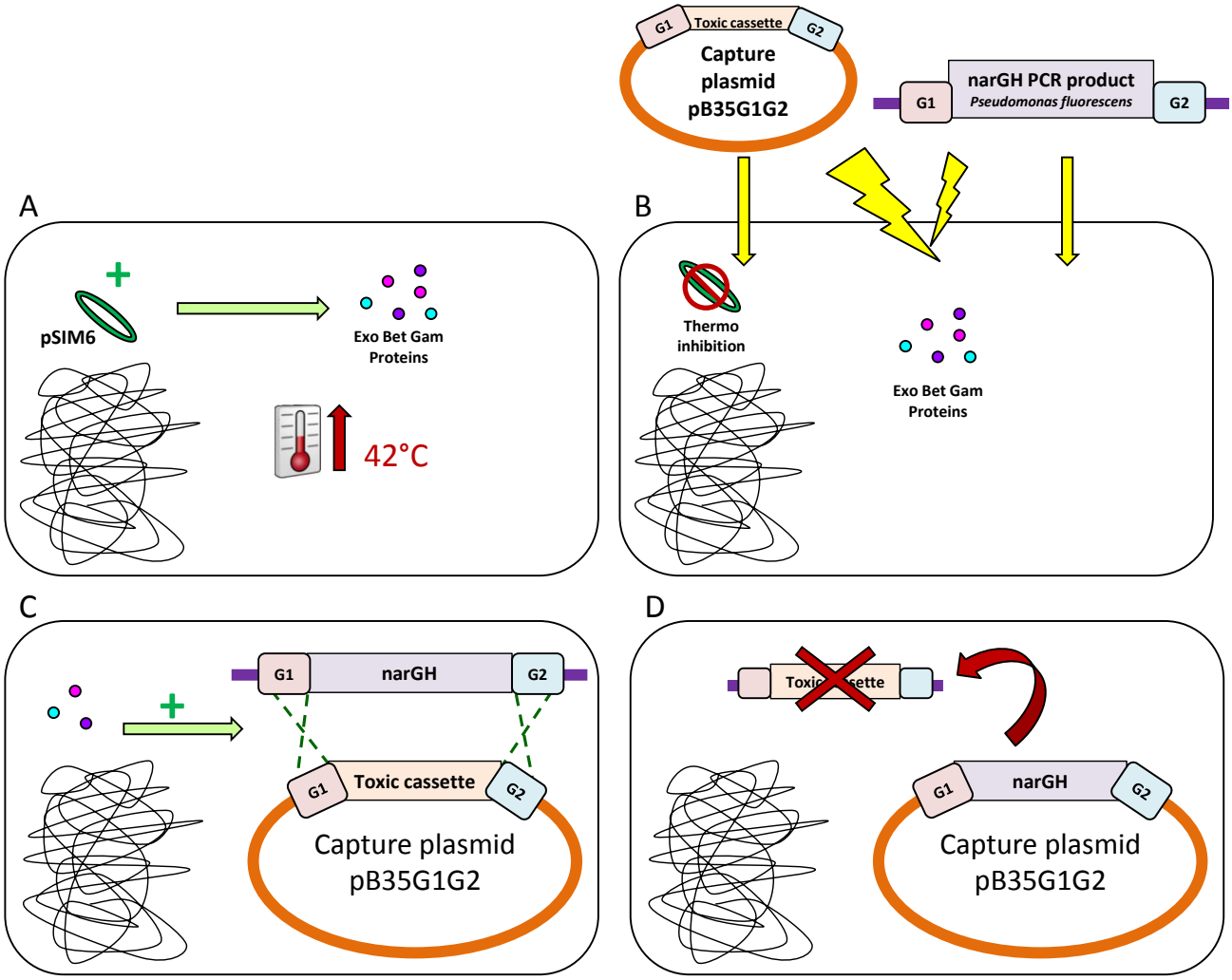


Fig.13: Schematic description of the co-transformation strategy. The 4 panels describe the workflow step by step. Panel A displays the thermo-induction of pSIM6 at 42°C, resulting in transcription of the Exo, Bet and Gam proteins. Panel B represents the thermo inhibition of pSIM6 and the co-electro-transformation of the narGH PCR fragment from Pseudomonas fluorescens, and the capture plasmid pB35G1G2. Panel C describes the recombination step performed by the λ red gam proteins Exo-Bet-Gam at the homologous regions G1 and G2. Panel D shows the toxic cassette removal and final state ready for the selection step.

N°	Date	[narG] (ng)	pB35G1G2 (ng)	Strain	Transformation rates (Tr)		Tr normalized per µg DNA		Escape rate (Er)		Er normalized per µg DNA		Notes
					29°C	42°C	29°C	42°C	29°C	42°C	29°C	42°C	
1	29/01/2010	0	50	*E. coli* 1661LI 14102 recD+ Δ*narG*	1,81E-03	4,67E-02	3,62E-02	9,34E-01	2,11E-05	1,43E-04	4,22E-04	2,86E-03	Only 1 recombinant at 15ng 42°C (1/5, 20%)
		15			8,82E-04	8,35E-04	1,36E-02	1,28E-02	<6,67E-05	1,01E-04	<1,03E-03	1,55E-03	
		50			7,89E-04	3,53E-03	7,89E-03	3,53E-02	1,67E-04	2,00E-05	1,67E-03	2,00E-04	
		0	50	*E. coli* 1661LI 14101 recD- Δ*narG*	2,96E-02	1,46E-03	5,92E-01	2,92E-02	7,23E-05	5,71E-05	1,45E-03	1,14E-03	No recombinant detected
		15			1,96E-03	8,50E-04	3,02E-02	1,31E-02	2,00E-04	<3,92E-05	3,08E-03	<6,03E-04	
		50			2,49E-04	3,91E-04	2,49E-03	3,91E-03	1,43E-04	<7,69E-05	1,43E-03	<7,69E-04	
2	04/02/2010	0	100	*E. coli* 1661LI 14102 recD+ Δ*narG*	2,61E-01	4,61E-02	5,22	9,22E-01	2,48E-06	4,81E-07	4,96E-05	9,62E-06	4 recombinants at 30ng 42°C (4/16, 25%)
		15			6,83E-02	1,57E-02	1,05	2,42E-01	1,17E-05	2,17E-05	1,80E-04	3,34E-04	
		30			3,82E-02	3,01E-02	4,78E-02	3,76E-02	3,75E-05	3,43E-05	4,69E-05	4,29E-05	
3	03/03/2010	1:1:3		*E. coli* 1661LI 14102 recD+ Δ*narG*	4,25E-02	3,41E-02	1,80E-01	1,45E-01	5,33E-06	1,80E-05	2,26E-05	7,63E-05	Molecular copy reasoning (Cell/Plasmid/*narG*). Recombinant detected at 1:1:3 (3/29, 10.4%) and 1:2:3 (10/78; 12.8%)
		1:2:3			6,51E-02	5,00E-02	1,94E-01	1,49E-01	2,68E-05	3,50E-05	7,97E-05	1,04E-04	
		1:3:3			4,97E-02	4,22E-02	1,13E-01	9,56E-02	4,34E-06	5,48E-06	9,85E-06	1,24E-05	
4	24/03/2010	1:2:0		*E. coli* 1661LI 14102 recD+ Δ*narG*	1,62E-02	2,39E-02	1,08E-01	1,59E-01	5,24E-05	8,57E-05	3,49E-04	5,71E-04	Molecular copy reasoning (Cell/Plasmid/*narG*). Recombinant detected at 1:2:5 (4/64, 5.9%) and 1:2:7 (3/39; 7.7%)
		1:2:5			9,70E-03	2,05E-02	2,59E-02	5,48E-02	1,90E-05	1,23E-04	5,08E-05	3,29E-04	
		1:2:7			4,14E-03	4,05E-03	8,88E-03	8,69E-03	5,33E-05	2,15E-04	1,14E-04	4,61E-04	

Tab.3: Summary of the results obtained from recombination attempts with the co-transformation strategy. Transformation rates of pB35G1G2 are established against the total bacteria counts for each condition, and normalized per microgramme of DNA to cross-compare conditions together

.

vi. Testing the Genefish tool: the segregation strategy

The last strategy to test our system was based on the extraction of the plasmids directly after the tandem recombination, followed by their re-transformation in the recipient strain at a molecular ratio of 1:1 followed by toxicity induction on plates. The recombinant events will be selected out of the plasmid pool by the 2d transformation and by the selection. Schematic representation of this strategy is given in Figure 13.

Several attempts were carried out with the tandem approach, starting with the *E. coli* strain equipped with the two plasmids pB35G1G2 and pSIM6 (Tab.4). The plasmidic DNA has been extracted 1hour and 30 minutes after the transformation of the *narGH* PCR product. This strategy was also successfull, as several recombinant were recovered on the plates after segregation of the plasmids. However, we did encounter some problems with the electrotransformation of the putative recombinant plasmids, with several electrical breakdown. The transformation efficency was also lower with the plasmids corresponding to the recombination condition (with *narGH* PCR product), while the efficency of the control (without *narGH* PCR product) is higher. In addition, the escape rate was also found to be higher under recombination conditions. Furthermore, when looking at the plasmids loaded on gel after recombination, some differences can be observed in the band patterns between the control condition and the recombination condition. This might be due to stability problems probably generated by the recombination, resulting in a loss of transformation efficiency, problems while trying to electroporate the bacterial cells, and an increase of the escape rate.

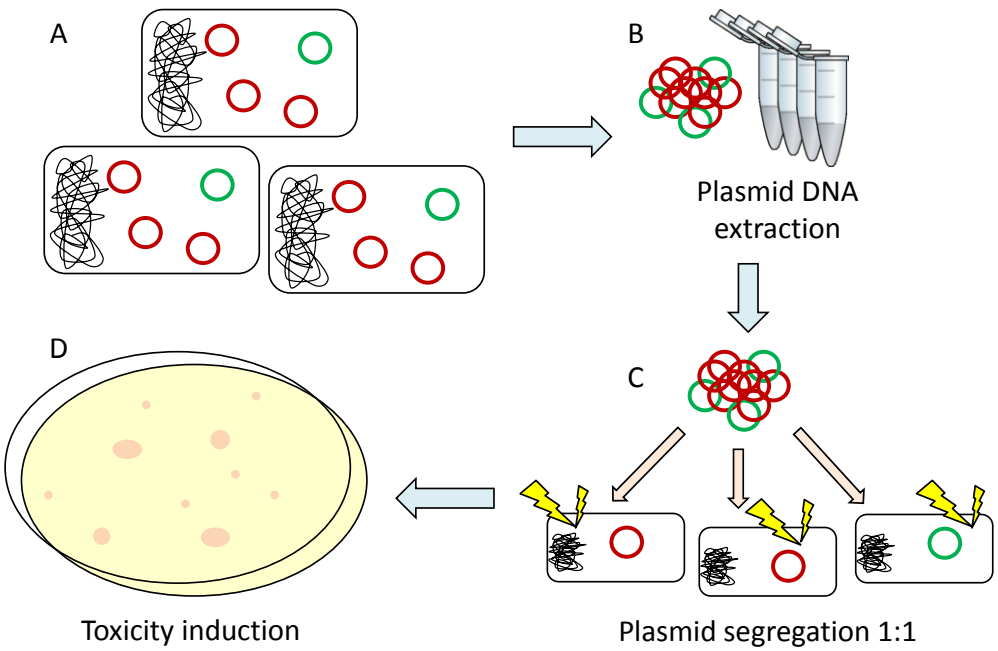


Fig.14: Schematic representation of the segregation strategy. The bacterial cells are recovered directly after recombination, wether from tandem or co-transformation strategies (Panel A), and the plasmidic DNA extracted (Panel B). At this stage, both recombinant and non-recombinant plasmids are pooled. The segregation is performed by a new transformation step, with a molecular ratio that allow 1 copy of plasmid per cell (Panel C). Final selection is performed on plate through toxicity induction in standard conditions.

Repetition	Date	Ratio cell:*narGH*	Strain	Segregation ratio	Transformation rates (Tr)	Tr normalized per µg DNA	Escape Rate (Er)	Er normalized per µg DNA	Notes
1	02/04/2010	1:3	*E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6	1:2	na	na	na	na	Growth problem, results not exploitable
				1:1	na	na	na	na	
2	14/04/2010	1:0	*E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6	1:1	3,86E-04	3,86E-03	3,35E-04	3,35E-03	2 recombinants detected (2/22, 9,01%)
		1:5			8,83E-05	2,73E-04	2,24E-03	6,92E-03	
3	02/05/2010	1:5	*E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6	0,8:1	2,15E-03	6,52E-03	1,15E-03	3,48E-03	7 recombinants at 0,8:1 (7,3%, 7/96) and 4 at 1:1 (4,2%, 4/96)
				1:1	2,38E-03	7,35E-03	9,05E-04	2,80E-03	
4	04/05/2010	1:0	*E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6	1:1	na	na	na	na	Problem with the electrotransformation of putative recombinant plasmids
		1:5			na	na	na	na	
5	10/05/2010	1:3	*E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6	0,5:1	na	na	na	na	1 recombinant detected (100%, 1/1, 0,8:1). Growth problem, no total bacteria
				0,8:1	na	na	na	na	

Tab.4: Summary of the results obtained from recombination attempts with the segregation strategy. Transformation rates of pB35G1G2 and escape rates are established against the total bacteria counts for each condition, and normalized per microgramme of DNA to cross-compare conditions together.

Strategy	N°	% positive	Recombination rates	Transformation + Recombination rates
Tandem	1	3,7% (1/27)	2,50E-06	-
	2	28,6% (2/7)	1,87E-08	-
Cotransformation	1	20% (1/5)	2,02E-05	1,69E-08
	2	25% (4/16)	8,57E-06	2,58E-07
	3	10,4% (3/29)	1,59E-06	4,63E+06
		12,8% (10/78)	3,98E-06	1,99E-07
	4	5,9% (4/64)	6,29E-06	1,29E-07
		7,7% (3/39)	1,46E-05	5,94E-08
Segregation	2	9,01% (2/22)	2,04E-04	-
	3	7,3% (7/96)	2,59E-05	-
		4,2% (4/96)	1,31E-05	-
	5	100% (1/1)	-	-

Tab.5: Recombination frequencies obtained from the different recombination strategies. Recombination and transformation rates are established against total bacteria counts.

3. Conclusions and perspectives

At the end, when looking at all the successfull assays (Tab.5), we can see that whether the strategy used for testing the system, the percentage of positive events growing on plates after toxicity induction is fluctuating around the same average value of approximately 10%. The tandem strategy has permitted to highlight the problem of capture plasmid copy number in host cell, obliging us to find alternatives. The co-transformation was a successfull, but hard to realize with lots of manipulation in a tight schedule, and also required a high quality capture plasmid, unless the escape rate is way to high for convenient detection of positive events. The segregation strategy was also successfull, but this approach highlighted some questionning aspects regarding the stability of the capture plasmid during and after the recombination attempts. All these results are pointing out that the capture of foreign DNA on our system by mediated homeologous recombination is something feasible, but at relatively low frequencies. The tandem approach resulted in recombination rates of 2.5E-06 down to 1.87E-08 depending on the number of total bacteria cells applied. On the other hand, higher recombination rates were observed with the co-transformation strategy, from 1.46E-05 down to 8.57E-06. This is mostly due to the tight control over capture plasmid copy number along the experiment. And finally, the segregation strategy resulted in the highest recombination rates, from 1.31E-05 down to 2.04E-04, which is due to the highest number of cells equipped with the capture plasmid at the begining of the experiment, while the co-transformation is strongly relying on pB35G1G2 transformation rates. When confronting these results with the frequencies obtained by Datta and co-workers (Datta *et al.* 2006), our results are almost reaching the same rates. In their study, the recombination frequencies observed with their set of inducible Red plasmids reached 3.20E-05 up to 4.30E-04 with linear dsDNA of 950bp flanked with ≈35bp homologous regions. However, our homologous regions are 10 times longer than the one they used, which seems to be still a limiting parameter for our Genefish tool.

The use of the λ red gam recombination system was successfull as well, as no recombinant clones were observed under control condition without induction of the recombinase (29°C). However, the hypothesis that λ red gam recombination system might impact on the integrity of the capture plasmid has been proposed. In fact, inducing the pSIM6 plasmid may probably result in undesired recombination processes in genetic sequences of the host strain, as the minimal lenght require is extremely low (35bp), and as the recombinatory power of this system is high. Such unwanted recombination events coming along with lambda Red systems were already reported in the literature (Murphy and Campellone, 2003). But when looking on our results, no significant differences in escape rates were observed between induced (42°C) and non induced (29°C), which demonstrate that pSIM6 does not seem to generate negative effect on our capture plasmid. However, the quality of the plasmid is directly correlated with growth conditions. Our experiments clearly highlighted that extended culture time of *E. coli* strains harboring pB35 is resulting in emergence of a "toxicity escaping" phenotype that is positively selected in culture. Sequencing of the toxic cassette of escaping cells on plates has revealed several deletion patterns that affect the toxic gene *colE3*, and sometimes *relF* as well. As colE3 has expression leaks, the cells that harbor the intact plasmid exhibit growth delay due to the extremely powerfull toxic effect of colicin E3.

As a consequence, the escaping phenotypes that shut down colE3 will display faster growth rates, resulting in a higher fitness than the others. This phenotype will overcome the wild one, and become dominant rapidly during the exponential growth phase. In fact, it was already reported that escaping bacteria display mutations in the coding sequence of the toxic gene (Diaz *et al.* 1994 ; Torres *et al.* 2000) or in the promotor regions of the toxic gene (Knudsen *et al.* 1995 ; Torres *et al.* 2003). These mutations result in shutting down the lethal propreties of gene product by blocking their transcription or making them non-functional.

The system, as it was designed, succeeded in performing a recombination of approximately 4kb length with homeologous regions of 500pb, with a similarity of 100% as both region and PCR products were all generated from *Pseudomonas fluorescens.* However, the presence of genomic *narGH* in *E. coli* 1661 LI 13421 recD+ strain resulted in an undesired recombination process with the capture plasmid. This has permitted to evaluate that the recombination was possible at lower similarity level, as *E. coli narGH* sequence share approximately 80% similarity with *Pseudomonas fluorescens* (Laurent philippot, personal communication) without the presence of pSIM6. New and shorter homeologous regions were designed (G3, G4, G5 and G6) based on the former sequence G1 and G2, and successfuly cloned in the capture plasmid. However, no recombinant were obtained from these constructions in co-transformation (data not shown).

For now on, the system is working, but the performance is still limitated by some technical aspects, such as the capture plasmid copy number, and the "toxicity escaping" phenotype. The tests that were carried on were performed with relatively high quantity of target product (e.g *narGH* PCR product). Usually, *narGH* PCR products were applied at 5.00E+10 copies, corresponding to a ratio of 5*narGH*:1 cell, which is higher than the typical concentration that may be found in environmental sample such as soil. For instance, *narG* gene copy number were evaluated thanks qPCR to be ranging between 3.2E+08 to 2.6E+09 per gram of Chinese paddy soil (Chen *et al.* 2011). The actual efficiency of the system is not enough to reach the target, both in extracted metagenomic DNA and directly into the environmental samples.

The transfert of the toxic cassette onto a low copy backbone is still on going. As soon as it will be done, the true potentialities of the Genefish strategy would be accessible.

Aknowledgement

This study was performed and funded in the frame of the European Union Metaexplore project METAEXPLORE (KBBE-222625). I would like to thank sincerely thank Laure Franqueville how was strongly involved in this study, for her constant help and devotion. I would like to thank my supervisor Pascal Simonet, and all the people involved in the development of the tool, and more specifically the persons that I directly worked with: Nathalie Lombard, Yuan Jun, Aurélie Faugier, Laurent Philippot and Jean Claude Lazzaroni.

References

Ahrenholtz I, Lorenz MG, Wackernagel W (1994) A conditional suicide system in Escherichia coli based on the intracellular degradation of DNA. *Appl Environ Microbiol*. 60:3746-3751

Aizenman E, Engelberg-Kulka H, Glaser G (1996) An Escherichia coli chromosomal "addiction module" regulated by guanosine [corrected] 3',5'-bispyrophosphate: a model for programmed bacterial cell death. *Proc Natl Acad Sci U S A*. 93:6059-6063

Amundsen SK, Taylor AF, Chaudhury AM, Smith GR (1986) recD: the gene for an essential third subunit of exonuclease V. *Proc Natl Acad Sci U S A*. 83:5558-5562

Amundsen SK, Taylor AF, Smith GR. (2000) The RecD subunit of the Escherichia coli RecBCD enzyme inhibits RecA loading, homologous recombination, and DNA repair. *Proc Natl Acad Sci U S A*. 97:7399-7404

Amundsen SK, Taylor AF, Smith GR (2002). A domain of RecC required for assembly of the regulatory RecD subunit into the *Escherichia coli* RecBCD holoenzyme. *Genetics*. 161:483- 492

Amundsen SK, Taylor AF, Reddy M, Smith GR (2007) Intersubunit signaling in RecBCD enzyme, a complex protein machine regulated by Chi hot spots. *Genes Dev*. 21:3296 –3307

Arnold DA, Handa N, Kobayashi I and Kowalczykowski SC (2000) A novel, 11 nucleotide variant of X,X*: one of a class of sequences defining the *Escherichia coli* recombination hotspot. *J Mol Biol*. 300:469–479

Bianco PR and Kowalczykowski SC (1997). The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5_-GCTGGTGG-3_. *Proc Natl Acad Sci U S A*. 94:6706-6711

Boehmer PE and Emmerson PT (1992) The RecB subunit of the Escherichia coli RecBCD enzyme couples ATP hydrolysis to DNA unwinding. *J Biol Chem*. 267:4981-4987

Bowman CM, Dahlberg JE, Ikemura T, Konisky J, Nomura M (1971) Specific inactivation of 16S ribosomal RNA induced by colicin E3 in vivo. *Proc Natl Acad Sci U S A*. 68:964-968

Chaudhury AM, Smith GR (1984a) A new class of Escherichia coli recBC mutants: implications for the role of RecBC enzyme in homologous recombination. *Proc Natl Acad Sci U S A*. 81:7850-7854

Chaudhury AM, Smith GR (1984b) Escherichia coli recBC deletion mutants. *J Bacteriol*. 160:788-91

Chen HW, Ruan B, Yu M, Wang J, Julin DA (1997) The RecD subunit of the RecBCD enzyme from Escherichia coli is a single-stranded DNA-dependent ATPase. *J Biol Chem*. 272:10072–10079

Chen Z, Liu J, Wu M, Xie X, Wu J, Wei W (2011) Differentiated response of denitrifying communities to fertilization regime in paddy soil. *Microb Ecol*.63:446-459

Churchill JJ, Anderson DG, and Kowalczykowski SC (1999) The RecBC enzyme loads RecA protein onto ssDNA asymmetrically and independent of x, resulting in constitutive recombination activation. *Genes Dev*. 13: 901–911

Copeland NG, Jenkins NA, Court DL (2001) Recombineering: a powerful new tool for mouse functional genomics. *Nat Rev Genet*. 2:769-779

Court DL, Sawitzke JA and Thomason LC (2002) Genetic engineering using homologous recombination. *Annu Rev Genet*. 36: 361–388

Court R, Cook N, Saikrishnan K, Wigley D (2007) The crystal structure of lambda-Gam protein suggests a model for RecBCD inhibition. *J Mol Biol*. 371:25-33

Dabert P, Smith GR (1997) Gene replacement with linear DNA fragments in wild-type Escherichia coli: enhancement by Chi sites. *Genetics*. 145:877-889

Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A*. 97:6640-6645

Datta S, Costantino N, Court DL (2006) A set of recombineering plasmids for gram-negative bacteria. *Gene*.379:109-115

Delmas S, Matic I (2005) Cellular response to horizontally transferred DNA in Escherichia coli is tuned by DNA repair systems. *DNA Repair* (Amst). 4:221-229

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol*. 77:1315-1324

Demanèche, S., Cérémonie, H., Buret, F., Auriol, P., Timothy M Vogel, & Simonet, P (2005) Gene Transfer in Soil Induced by Lightning. *Transformation*. *0*:4-8

Dermić D (2006) Functions of multiple exonucleases are essential for cell viability, DNA repair and homologous recombination in recD mutants of Escherichia coli. *Genetics*. 172:2057-2069

de Vries J and Wackernagel W (2002) Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. *Proc Natl Acad Sci U S A*. 99:2094-2099

Díaz E, Munthali M, de Lorenzo V, Timmis KN (1994) Universal barrier to lateral spread of specific genes among microorganisms. *Mol Microbiol*. 13:855-861

Dillingham MS, Spies M and Kowalczykowski SC (2003) RecBCD enzyme is a bipolar DNA helicase. *Nature*. 423:893–897

Dillingham MS, Kowalczykowski SC (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol Mol Biol Rev*. 72:642-671

Dower WJ, Miller JF, Ragsdale CW (1988) High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Res*. 16:6127-6145

Duret S, Danet JL, Garnier M, Renaudin J (1999) Gene disruption through homologous recombination in Spiroplasma citri: an scm1-disrupted motility mutant is pathogenic. *J Bacteriol*.181:7449-7456

Engelberg-Kulka H, Amitai S, Kolodkin-Gal I, Hazan R (2006) Bacterial programmed cell death and multicellular behavior in bacteria. *PLoS Genet*. 2:e135

Fujitani Y, Yamamoto K, Kobayashi I (1995) Dependence of frequency of homologous recombination on the homology length. *Genetics*. 140:797-809

Guglielmini J, Van Melderen L (2011) Bacterial toxin-antitoxin systems: Translation inhibitors everywhere. *Mob Genet Elements*. 1:283-290

Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol*. 177:4121-4130

Griffiths AJF, Gelbart WM, Miller JH and Lewontin RC (1999) Modern genetic analysis. New York: W.H. Freeman and company

Handa N, Ohashi S, Kusano K and Kobayashi I (1997) Chi-star, a chi-related 11-mer sequence partially active in an E. coli recC1004 strain. *Genes Cells*. 2:525–536

Hashimoto-Gotoh, T, Franklin FC, Nordheim A and Timmis KN (1981) Specific-purpose plasmid cloning vectors. I. Low copy number, temperature-sensitive, mobilization-defective pSC101-derived containment vectors. *Gene*. 16:227–235

Hassink J, Bouwman LA, Zwart KB, Bloem J, Brussard L (1993) Relationships between soil texture, physical protection of organic matter, soil biota, and c and n mineralization in grassland soils. *Geoderma*. 57:105-128

Hayes F, Van Melderen L (2011) Toxins-antitoxins: diversity, evolution and function. *Crit Rev Biochem Mol Biol*. 46:386-408

Hengen PN (1995) Methods and reagents. Electro-transformation of Escherichia coli with plasmid DNA. *Trends Biochem Sci*. 20:248-249

Hickson ID, Robson CN, Atkinson KE, Hutton L, Emmerson PT (1985) Reconstitution of RecBC DNase activity from purified *Escherichia coli* RecB and RecC proteins. *J Biol Chem*. 260:1224–1229

Holben WE, Jansson JK, Chelm BK, Tiedje JM (1988) DNA Probe Method for the Detection of Specific Microorganisms in the Soil Bacterial Community. *Appl Environ Microbiol*. 54:703-711

Inceoglu O, Salles JF, Van Overbeek L, Van Elsas JD (2010) Effects of plant genotype and growth stage on the betaproteobacterial communities associated with different potato cultivars in two fields. *Appl Environ Microbiol*. 76:3675-3684

Jakes KS and Zinder ND (1974) Highly purified colicin E3 contains immunity protein. *Proc Natl Acad Sci U S A.* 71:3380-3384

Jockovich ME and Myers RS (2001) Nuclease activity is essential for RecBCD recombination in Escherichia coli. *Mol Microbiol*. 41:949–962

Keren I, Shah D, Spoering A, Kaldalu N, Lewis K (2004) Specialized persister cells and the mechanism of multidrug tolerance in Escherichia coli. *J Bacteriol*. 186:8172-8180

Kinch LN, Ginalski K, Rychlewski L, Grishin NV (2005) Identification of novel restriction endonuclease-like fold families among hypothetical proteins. *Nucleic Acids Res*. 33:3598–3605

Knudsen S, Saadbye P, Hansen LH, Collier A, Jacobsen BL, Schlundt J, Karlstrom OH (1995) Development and testing of improved suicide functions for biological containment of bacteria. *Appl Environ Microbiol*. 61:985-991

Kolodkin-Gal I, Hazan R, Gaathon A, Carmeli S, Engelberg-Kulka H (2007) A linear pentapeptide is a quorum-sensing factor required for mazEF-mediated cell death in Escherichia coli. *Science*. 318:652-655

Lanzer M, Bujard H (1988) Promoters largely determine the efficiency of repressor action. *Proc Natl Acad Sci U S A.* 85:8973-8977

Leplae R, Geeraerts D, Hallez R, Guglielmini J, Drèze P, Van Melderen L (2011) Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. *Nucleic Acids Res*. 39:5513-5525

Lovett ST, Luisi-DeLuca C, Kolodner RD (1988) The genetic dependence of recombination in recD mutants of Escherichia coli. *Genetics*. 120:37-45

Lutz R, Bujard H (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* 25:1203-1210

Masaki H and Ohta T (1982) A plasmid region encoding the active fragment and the inhibitor protein of colicin E3--CA38. *FEBS Lett*. 149:129-132

Mason TG, Richardson G (1982) Observations on the in vivo and in vitro competition between strains of Escherichia coli isolated from the human gut. *J Appl Bacteriol*. 53:19-27

Masterson C, Boehmer PE, McDonald F, Chaudhuri S, Hickson ID, Emmerson PT (1992) Reconstitution of the activities of the RecBCD holoenzyme of Escherichia coli from the purified subunits. *J Biol Chem*. 267: 13564–13572

Matic I, Taddei F, Radman M (1996) Genetic barriers among bacteria. *Trends Microbiol*. 4:69-72

Murphy KC. (1991) Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of Escherichia coli RecBCD enzyme. *J Bacteriol*. 173:5808-5821

Murphy KC (1998) Use of bacteriophage lambda recombination functions to promote gene replacement in Escherichia coli. *J Bacteriol*. 180:2063-2071

Murphy KC and Campellone KG (2003) Lambda Red-mediated recombinogenic engineering of enterohemorrhagic and enteropathogenic E. coli. *BMC Mol Biol*. 4:11

Ohashi-Kunihiro S, Hagiwara H, Yohda M, Masaki H, Machida M (2006) Construction of a positive selection marker by a lethal gene with the amber stop codon(s) regulator. *Biosci Biotechnol Biochem*. 70:119-125

Palas KM, and Kushner SR (1990) Biochemical and physical characterization of exonuclease V from Escherichia coli. Comparison of the catalytic activities of the RecBC and RecBCD enzymes. *J Biol Chem*. 265:3447–3454

Pandey DP and Gerdes K (2005) Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res*. 33:966-976

Poteete AR, Fenton AC, Murphy KC (1988a) Modulation of Escherichia coli RecBCD activity by the bacteriophage lambda Gam and P22 Abc functions. *J Bacteriol*. 170:2012-21

Poteete AR, Volkert MR. (1988b) Activation of recF-dependent recombination in Escherichia coli by bacteriophage lambda- and P22-encoded functions. *J Bacteriol*. 170(9):4379-81

Poteete AR (2001) What makes the bacteriophage lambda Red system useful for genetic engineering: molecular mechanism and biological function. *FEMS Microbiol Lett*. 201:9-14

Ren D, Bedzyk LA, Thomas SM, Ye RW, Wood TK (2004) Gene expression in Escherichia coli biofilms. *Appl Microbiol Biotechnol*. 64:515-524

Richardson DJ, Berks BC, Russell DA, Spiro S, Taylor CJ (2001) Functional, biochemical and genetic diversity of prokaryotic nitrate reductases. *Cellular and Molecular Life Sciences*. 58:165-178

Sergueev K, Yu D, Austin S, Court D (2001) Cell toxicity caused by products of the pL operon of bacteriophage lambda. *Gene*. 272: 227–235

Shen P and Huang HV (1986) Homologous recombination in Escherichia coli: dependence on substrate length and homology. *Genetics*. 112:441-457

Smith GR, Amundsen SK, Chaudhury AM, Cheng KC, Ponticelli AS, Roberts CM, Schultz DW, Taylor AF (1984) Roles of RecBC enzyme and chi sites in homologous recombination. *Cold Spring Harb Symp Quant Biol*. 49:485-495

Smith GR (2012) How RecBCD Enzyme and Chi Promote DNA Break Repair and Recombination: a Molecular Biologist's View. *Microbiol Mol Biol Rev*. 76:217-228

Sun JZ, Julin DA, Hu JS (2006) The nuclease domain of the Escherichia coli RecBCD enzyme catalyzes degradation of linear and circular single-stranded and double-stranded DNA. *Biochemistry*. 45:131–140

Szafranski P, Mello CM, Sano T, Smith CL, Kaplan DL and Cantor CR (1997) A new approach for containment of microorganisms: dual control of streptavidin expression by antisense RNA and the T7 transcription system. *Proc Natl Acad Sci U S A*. 94:1059-1063

Taketo A (1988) DNA transfection of Escherichia coli by electroporation. *Biochim Biophys Acta*. 949:318-324

Thomas CM, Nielsen KM. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol*. 3:711-721

Thomason LC, Ausubel FM, Brent R, Kingston RE, Moore DD, Seidma, JG, Smith JA and Struhl K (2005) Recombineering: genetic engineering in bacteria using homologous recombination. Eds. Curr Protoc Mol Biol. JohnWiley&Sons, Inc., Hoboken NJ, pp. 1–21. Chapter 1,Unit 16

Torres B, Jaenecke S, Timmis KN, García JL, Díaz E (2000) A gene containment strategy based on a restriction-modification system. *Environ Microbiol*. 2:555-563

Torres B, Jaenecke S, Timmis KN, García JL, Díaz E (2003) A dual lethal system to enhance containment of recombinant micro-organisms. *Microbiology*. 149:3595-3601

Wozniak RA and Waldor MK (2009) A toxin-antitoxin system promotes the maintenance of an integrative conjugative element. *PLoS Genet*. 5:e1000439

Yamaguchi Y and Inouye M (2009) mRNA interferases, sequence-specific endoribonucleases from the toxin-antitoxin systems. Prog. Mol. Biol. Transl. Sci. 85:467–500

Yamaguchi Y and Inouye M (2011) Regulation of growth and death in Escherichia coli by toxin–antitoxin systems. *Nat Rev*. 9:779-790

Yu M, Souaya J, Julin DA (1998) The 30-kDa C-terminal domain of the RecB protein is critical for the nuclease activity, but not the helicaseactivity, of the RecBCD enzyme from Escherichia coli. *Proc. Natl. Acad. Sci. U S A.* 95:981–989

CHAPTER 3

-

Protocol appendixes

This section is presenting the detailed protocol used in the GeneFish experiments presented in the Chapter 3. Protocols are listed as follow:

Chapter 3 Appendixes summary:

- Appendix 3-1: *Escherichia coli* strains, plasmids and primers.....244

- Appendix 3-2: Preparation of pB35K7tox through CsCl ultra-centrifugation gradient.....247

- Appendix 3-3: Deletion of genomic *narG* gene in *E. coli* strains.....251

- Appendix 3-4: Electrocompent cell preparation.....255

- Appendix 3-5: Cloning of the recombination sites.....257

- Appendix 3-6: Toxicity induction assay.....261

- Appendix 3-7: The tandem recombination strategy.....263

- Appendix 3-8: The co-transformation recombination strategy.....265

- Appendix 3-9: The segregation recombination strategy.....267

- Appendix 3-10: Escaping phenotypes and recombinant clone validation.....269

Appendix 3-1

-

Escherichia coli strains, plasmids and primers

Developing and optimizing the Genefish tool required major chromosome modifications in the *E. coli* K12 strain. Here we chronologically report these modifications (Tab.1). (1) The first one aimed to knock out *recD* in order to improve recombination efficiency. Four *E. coli* strains harboring respectively a $recD^+$ and a $recD^-$ phenotype due to sequence mutation were employed: 1788 TE2680LI::recD1903 and V355 1661LI::recD1014, both derived from Amundsen et al., 1986. These strains were obtained from our collaborator Jean-Claude Lazzaroni (Unité de Microbiologie, Adaptatation et Pathogénie, UMR5240, Université de Lyon, 69622 Villeurbanne cedex, France) who was in charge of genetic modifications of the recipient strain. (2) Next step consisted in integrating the antitoxin gene immE3 into the chromosome of these 4 recipient strains, (3) followed by the removal of the arabinose degrading capacity by knocking out the *ara* operon. Finally, part of the genomic *nar* operon has been removed, including *narGH* in order to avoid undesired recombination process between the host genome and the capture plasmid.

A simplified workflow presenting the construction of the final capture plasmid is given in Tab.2. This work was done by Nathalie Lombard during here PhD. Several backbones were used, and the different toxic cassettes were amplified from other constructions. Table 3 is presenting all set of primers used in this study.

References:

Amundsen SK, Taylor AF, Chaudhury AM and Smith GR (1986) recD: the gene for an essential third subunit of exonuclease V. *Proc Natl Acad Sci U S A*. 83:5558-5562

Chaudhury AM and Smith GR (1984) A new class of Escherichia coli recBC mutants: implications for the role of RecBC enzyme in homologous recombination. *Proc Natl Acad Sci U S A*. 81:7850-7854

Datsenko KA and Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A*. 97:6640-6645

Elliott T (1992) A method for constructing single-copy lac fusions in Salmonella typhimurium and its application to the hemA-prfA operon. *J Bacteriol*. 174:245-253

Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol*. 177: 4121-4130

Knudsen S, Saadbye P, Hansen LH, Collier A, Jacobsen BL, Schlundt J, Karlstrom OH (1995) Development and testing of improved suicide functions for biological containment of bacteria. *Appl Environ Microbiol*. 61: 985-991

Lutz R and Bujard H (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res*. 25:1203-1210

***E. coli* strain**	**Genotype**	**Resistance and Phenotype**	**Reference**
DH5	F- ϕ80lacZΔM15 Δ(lac-argF)U169 deoR recA1 endA1 hsdR15(rk-, mk+) phoA supE44 thi-1 gyrA96 relA1 λ-	Sp^R	Invitrogen
DH5ZI	F- ϕ80lacZΔM15(lac-argF)U169 deoR recA1 endA1 hsdR15(rk-,mk+) phoA supE44 thi-1 gyrA96 relA1 λ- spr $lacI^q$ tet^R	Sp^R	Lutz et al. 1997
TE2680	F- λ- IN(rrnD-rrnE)1 Δ(lacX)74 rpsL galK2 recD1903::Tn10d-tet trp DC700:: putPA1308::[kan-lacS-chlR$lacI^q$immE3_lac]	Kn^R, Tc^R, $ImmE3^+$, Cm^S	Elliott, 1992
TELI	F- λ- IN(rrnD-rrnE)1 ΔlacX74 rpsL galK2 recD1903::Tn10d-Tet trpDC700::putA1303::[knr- -lacIq-immE3-lac] Kn, Tc	Kn^R, Tc^R, $ImmE3^+$, Cm^S	This study
TELIΔara*ΔnarG*	F- λ- IN(rrnD-rrnE)1 Δ(lacX)74 rpsL galK2 recD1903::Tn10d-tet trp DC700:: putA1303::[knr-$lacI^q$immE3_lac] ΔaraBAD, Δ*narGnarH*	Kn^R, Tc^R, $ImmE3^+$, Cm^S , Ara^-, $NarG^-$	This study
1661 (derived form V355)	F- lac-3350, galK2, galT22, recD1014, IN(rrnD-rrnE)1, rpsL179 Smr, ImmE3+	Sm^R, Kn^R, Trp^-, $ImmE3^+$	Chaudhurry and Smith, 1984
1661LI	F- lac-3350 galK2 galT22 I- recD1014 IN(rrnD-rrnE)1 rpsL179 Δtrp::[Kn^R-immE3_$lacI^q$]	Sm^R, Kn^R, Trp^-, $ImmE3^+$	This study
1661LIΔ*ara*	F- lac-3350, galT22, recD1014, IN(rrnD-rrnE)1, rpsL179 Δtrp::[Kn^R-immE3_$lacI^q$] Δ*ara*BAD	Sm^R, Kn^R, Trp^-, $ImmE3^+$, Ara^-	This study
1661LIΔ*ara*Δ*narG*	F- lac-3350, galT22, recD1014, IN(rrnD-rrnE)1, rpsL179 Δtrp::[Kn^R-immE3_$lacI^q$] Δ*ara*BAD, Δ*narGnarH*	Sm^R, Kn^R, Trp^-, $ImmE3^+$, Ara^-, $NarG^-$	This study

Tab.1: List of all the E. coli strains used for the development of the GeneFish tool. Ara^-: auxotrophe for L+arabinose, , Cm^R: chloramphenicol resistance, Cm^R: sensistive to chloramphenicol, $ImmE3^+$: production of colE3 antitoxine ImmE3, Kn^R: kanamycin resistance, Sm^R: strptomycin resistance, Tc^R: tetracyclin resistance, Trp^-: Auxotrophe for tryptophane, $NarG^-$: deletion of narG and narH genes.

Plasmids	**Description**	**Genotype**	**Phenotype**	**Reference**
pBAD33	Expression vector	cm^R, p15A, *araC*, Pbad Cmr	Cm^R	Guzman et al., 1995
pB341	Backbone cloning vector	cm^R, ori p15A	Cm^R	This study
pSK360	Target for amplification of [$P_{A1\text{-}O3/O4}$::*relF-tcr*] cassette	amp^R, tetR, 2x($P_{A1\text{-}O3/O4}$::*relF*), ori pMB1	Amp^R, Tet^R, $IPTG^S$	Knudsen et al., 1995
pB33C3	Target for amplification of [*araC*-P_{bad}::*colE3-rrnB*] cassette	cm^R, *araC*-P_{bad}::colE3, ori p15A	Cm^R, Ara^S	Lombard, unpublished
pB341C3	Cloning of the [*araC*-P_{bad}::*colE3-rrnB*] cassette from pB33C3 in pB341	cm^R, *araC*-P_{bad}::colE3, ori p15A	Cm^R, Ara^S	This study
pB341RTc	Cloning of the [$P_{A1\text{-}O3/O4}$::*relF-tetR*] from pSK360 cassette in pB341C3	cm^R, ori p15A, $P_{A1\text{-}O3/O4}$::*relF*	Amp^R, Tc^R, $IPTG^S$	This study
pB341C3RTc2	Cloning of the [$P_{A1\text{-}O3/O4}$::*relF-tcr*] cassette in pB341C3	cm^R, ori p15A, *araC*-P_{bad}::colE3, $P_{A1\text{-}O3/O4}$::*relF*-tcr	Cm^R, Tc^R, $IPTG^S$, Ara^S	This study
pBAD35K7toxN	Cloning of the [*araC*-P_{bad}::*colE3-rrnB* ; $P_{A1\text{-}O3/O4}$::*relF*-] cassette from pB341C3RTc2	cm^R, ori p15A, *araC*-P_{bad}::colE3, $P_{A1\text{-}O3/O4}$::*relF*	Cm^R, $IPTG^S$, Ara^S	This study
pBAD35K7G1G2	Cloning of conserved region G1 and G2 for *narG* recombination application	cm^R, ori p15A, *araC*-P_{bad}::colE3, $P_{A1\text{-}O3/O4}$::*relF*	Cm^R, $IPTG^S$, Ara^S	This study

Tab.2: List of all the plasmids construction achieved during the Genefish tool development. Amp^r: ampicillin resistance, Ara^s: sensitive to L+arabinose induction, Cm^r: chloramphenicol resistance, $IPTG^s$: sensitive to IPTG induction, Tc^r: tetracyclin resistance Ara^s: sensitive to L+arabinose induction.

Name	Sequence	Size(bp)	%GC	Tm (°C)	Description	Reference
P1dnarFw1	*ATTCATTCACTTGCCGCCTTCCTGTAAACCGAATTA*GTGTAGGCTGGAGCTGCTTC	56	48.2	58.63	Diruption of genomic *narG* (Appendix 3-3)	derived from Datsenko and Wanner, 2000
P2dnarRv1	*TCTTCGTAGTTAGCAATCGCCAGATAACGGTACATC*CATATGAATATCCTCCTTAG	56	41	45.01		
narG-fw	CCAGCATCGTCGAGAACAA	19	52	59.35	Amplification of the *narG* PCR product for recombination from *Pseudomonas fluorescens* C7R12	This study
narG-rv	CGGGTAGCAGAAGATGCATT	20	50	58.74		
G1-fw	*AGATCTAGATCTAGATCT*AGAATTCCACCTGGACAAGC	20	50	57.61	Amplification of the homologous region *narG*1 from *Pseudomonas fluorescens C7R12*	
G1-rv	*CGATCGCGATCGCGATCG*GGCAAGACGATGTCGGAATA	20	50	59.48		
G2-fw	*CCTAGGCCTAGGCCTAGG*CCACCTGGTACGAAAAAGAC	20	50	55.67	Amplification of the homologous region *narG*2 from *Pseudomonas fluorescens C7R12*	
G2-rv	*CTGCAGCTGCAGCTGCAG*TCGATTCGACGTTGTTGAAC	20	45	58.13		
R1-fw2	TGAATGATGTAGCCGTCAAG	20	45	55.71	Cloning verification primers for the multi-cloning site R1 from pb35	
R1-rv1	CAGGGCAGGGTCGTTAAATA	20	50	58.50		
R2-fw2	AAGAGATTACGCGCAGACCA	20	50	59.43	Cloning verification primers for the multi-cloning site R2 from pb35	
R2-rv1	AGACGTCAGGTGGCAATTTT	20	45	58.09		
narGtot-fw1	AGAATTCCACCTGGACAAGC	20	50	57.61	*narG* sequencing primers for recombinant verification and validation	
narGtot-rv1	GGCAAGACGATGTCGGAATA	20	50	59.48		
narGtot-fw2	CCACCTGGTACGAAAAAGAC	20	50	55.67		
narGtot-rv2	TCGATTCGACGTTGTTGAAC	20	45	58.13		
Red1-fw1	ACCTTGCCGATCAGCCAAA	19	53	62.44	pSIM6 verification primers. Targeted gene: exo	
Red1-rv2	ACCCGTCAGCTGTCTCTTATACAC	24	50	59.56		
Red2-fw1	CCGGAACGATGCCATTCTGCTTAT	24	50	67.69	pSIM6 verification primers. Targeted gene: bet	
Red2-rv2	GTTTTCGTTCCACTGAGCGTCAGA	24	50	65.55		
Red3-fw1	GTCATCGCCATTGCTCCCCAAATA	24	50	68.45	pSIM6 verification primers. Targeted gene: gam	
Red3-rv1	TTGCCAACCAGTACGGCCTTAATC	24	50	65.79		

Tab.3: List of primers used in yhis study and their description. Tm values were calculated with the on-line application Tm calculator (http://www6.appliedbiosystems.com/support/techtools) at 50mM salt and 2µM of each primer. The sequences in green are representing the narG homologous short tags used for genetic disruption of genomic narG (Appendix 3-3). Sequences in red are representing the repeated palindrome sequence (x3) of the restriction enzymes used for cloning of the recombination sites narG1 and narG2 (Appendix 3-5).

Appendix 3-2

-

Preparation of pB35K7tox through CsCl ultra-centrifugation gradient

The preparation of large quantity of plasmid is an important step in many molecular biology studies. There are many existing protocols relying on different principle to isolate huge amount of highly pure plasmid. In a global way, these protocols can be structured in two major steps. The first one consists in preparing crude lysates enriched in plasmid DNA from bacterial cells grown in liquid culture. This encompasses different techniques such as alkaline lysis, boiling, and Triton lysis (Heilig et al. 2001).

The second step consist in isolating, enriching and purifying plasmid DNA from genomic DNA from crude lysates. This step should also prevent contamination from RNA and proteins as well. Different techniques were reported such as: CsCl/ethidium bromide density gradient centrifugation (Good and Nazar, 1995), polyethylene glycol (PEG) precipitation (Nicoletti and Condorelli, 1993), anion-exchange chromatography (Guerrero-Germán et al. 2009), and size-exclusion chromatography (Li et al. 2007).

In this section we describe the protocole that we used for preparation of large quantities of capture plasmid pB35k7tox. This preparation has provided the plasmid DNA material that was required along this study.

1. Prepare the plasmid from 1liter of overnight culture with the appropriate selective marker using the Maxiprep Plasmid DNA Purification Kit, Endotoxin-Free (Macherey-Nagel). Stop the manufacturer protocol after the elution step with 5ml TE-EF buffer (Macherey-Nagel).
2. Determine the exact mass of plasmid solution with a precision scale.
3. Add the CsCl powder (Sigma, C4036) at a rate of 1g per gram of plasmid solution. Use a becher with a magnet bar to smoothly homogenize the solution.
4. Add the ethidium bromide (Euromedex, EU 0070) (10mg/ml) at a rate of 20µl per gram of initial plasmid solution.
5. Transfer the final solution in specific OPTISEAL tubes devoted to ultra-centrifugation. Carefully equilibrate the tubes to the brim using a precision scale. In this study, an Optima^{TM} MAX-XP Ultracentrifuge was used, with a MLN80 rotor (Beckman Coolter).
6. Carefully close the tubes with adequate joins and correctly screw the lid. Dispose the tubes in the rotor and close all the rotor sites with adequate lids. Place the rotor in the vacuum chamber of the ultra-centrifugation.
7. Set the ultra-centrifugation run at 45 000 rpm, 8h, 20°C. Launch the run overnight.
8. Delicately recover the tubes and places them in front of a UV light to visualize the plasmid bands. Make sure to not expose yourself directly to the UV light source.

9. Open the tubes and remove carefully the upper part of the supernatant above the plasmid band with a needle and/or a syringe. Proceed as presented on the picture, by picking the needle right above the plasmid band, and let the supernatant flow down in a collection recipient. Be sure to not dismantle the structure of the plasmid band.

10. Recover the plasmid band in a clean 2 ml eppendorf.

11. Add approximately 1.5 ml of isobutyl alcohol (Sigma, 217905) and invert the tubes 10 times to mix the solution. The ethidium bromide will be separate in the organic phase.

12. Discard the organic phase (on the top with pink coloration) and repeat step 11, 2 or 3 times in order to remove as much as possible the ethidium bromide. Add 3 volumes of sterile water.

13. Eliminate the cesium salts by ethanol precipitation. Add 8 volumes of cooled pure ethanol (Fischer scientifique, A0347B) (4°C). Invert the tubes 5 times and let precipitate at 4°C 3 hours (overnight for maximum recovery yield). Spin the tubes at 20 000 rpm, 50min, 4°C. DO NOT STORE THE TUBES AT -20°C OR CESIUM SALT WILL CO-PRECIPITATE!!!

14. Supernatant can be discarded or used in a second ethanol precipitation to increase the yields.

15. Wash the plasmid pellets with 100µl of 70% ethanol. Spin 20 000 rpm, 30min. Remove the ethanol.

16. Let the pellets dry for 5-10min. Extended drying time can result in problems in pellet dissolution. Add the desired volume of water of TE buffer (Macherey-Nagel), dispatch in small aliquots and store at -20°C.

References:

Good L and Nazar RN (1995) Visualization of CsCl/EtdBr plasmid preparations under visible light. *Biotechniques*. 18:556-558

Guerrero-Germán P, Prazeres DM, Guzmán R, Montesinos-Cisneros RM, Tejeda-Mansir A (2009) Purification of plasmid DNA using tangential flow filtration and tandem anion-exchange membrane chromatography. *Bioprocess Biosyst Eng*. 32:615-623

Heilig JS, Elbing KL, Brent R (2001) Large-scale preparation of plasmid DNA. *Curr Protoc Mol Biol*. Chapter 1:Unit1.7

Li LZ, Liu Y, Sun MS, Shao YM (2007) Effect of salt on purification of plasmid DNA using size-exclusion chromatography. *J Chromatogr A*. 1139:228-235

Nicoletti VG and Condorelli DF (1993) Optimized PEG method for rapid plasmid DNA purification: high yield from "midi-prep". *Biotechniques*. 14:532-536

Appendix 3-3

-

Deletion of genomic *narG* gene in *E. coli* strains

As the *narG* gene has been chosen for testing and validating the Genefish strategy, the presence of any copies of this gene in the host strain genome would be deleterious. In fact, undesired recombination process may occur between the final capture plasmid pB35G1G2 and the genomic *narG-H* genes. This was confirmed, with the observation of parasite recombination between the plasmid and the *narG* gene that was inadvertently still present in the genome of the *E. coli* 1661 Li 13421 recD+ strain. In order to remove it, a strategy has been tested by coupling the pSIM6 mediated recombination system (Datta et al. 2006) with the FRT/FLP disruption system (Datsenko and Wanner, 2000).

The principle relies on two major steps: (1) the disruption of a targeted genomic region through homologous recombination, concomitantly to integration of a selective marker (Fig.1). (2) The removal of this selective marker through activation of the FLP recombinase from *Saccharomyces cerevisea*, and joining of specific regions called FRT (Fig.2).

The selective marker should be carefully prepared for theses two steps. This includes a particular amplification step, with specific primers flanked with homologous regions shared with the target gene that needs to be knocked down (Fig.1, panel A). The selective marker also needs to carry the specific homologous FRT regions (Fig.1, panel B). These regions of 36bp are 100% similar to each other, and should be present at each extremity of the PCR product (Baudin et al. 1993). In our case, the selective marker used was the chloramphenicol resistance gene "cat". The selective marker has been amplified from the pKD3 plasmid (Datsenko and Wanner, 2000), with specific primers of 20bp homologous to *narG* and *narH* genes. The strain needs to be equipped beforehand with the pSIM6 plasmid, in order to facilitate the recombination between the PCR product and genomic *narG-H* genes (Fig.1, panel C). After transformation, the "*narG*-FRT-*cat*" PCR product may recombine with the genomic *narG-H* gene that needs to be eliminated, thanks to the pSIM6 plasmid lambda red gam expression activation. Afterward, the "*cat*" gene can be used as a selective marker to get positive recombination events on plates.

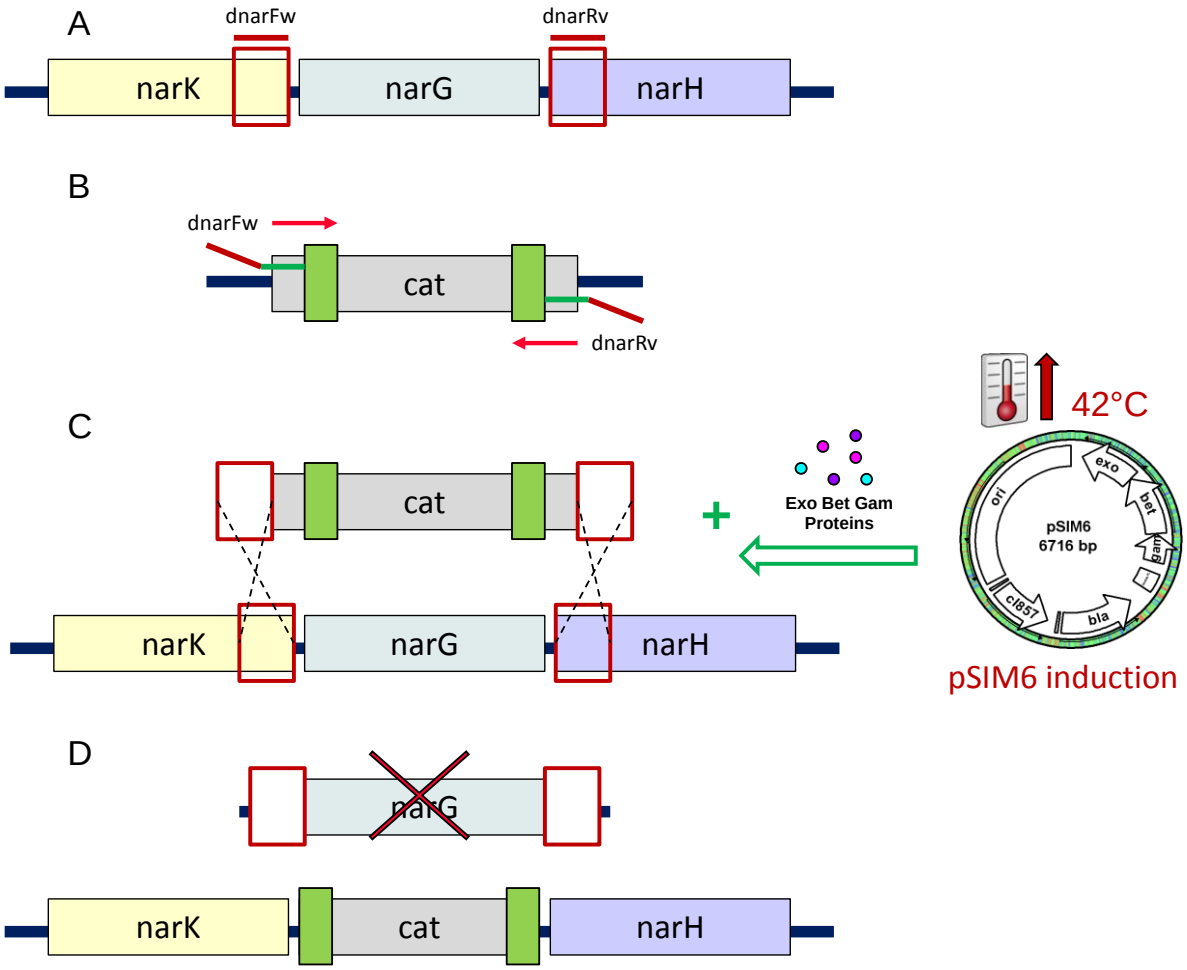


Fig. 1: Representation of genomic narG disruption in E. coli, using a selective marker and homologous recombination mediated with the λ red gam system. Panel A shows the chromosome region that needs to be disrupted in the nar operon of the host strain. Red boxes are representing the homologous sites that will be used for recombination "dnar". Panel B displays the PCR step that led to amplification of the selective marker that will recombine with the chromosome "cat", conferring chloramphenicol resistance. Primers are flanked with the corresponding homologous sequences "dnar". Panel C is describing the recombination step mediated by pSIM6, resulting in disruption of narG. Panel D is showing the intergration of the selective marker in the chromosome while narG is eliminated.

Next step consists in introducing the pCP20 plasmid that carries the FLP gene coding for the FLP recombinase. This enzyme can promote the matching of FRT regions together, followed by a total disruption of the DNA fragment surrounded by these regions. This pCP20 plasmid shows temperature-sensitive replication and thermal induction of FLP synthesis by activation of the ci857 promoter (Cherepanov and Wackernagel, 1995). This FLP recombinase will promote the joining of the two FRT regions, and will cut and remove the loop that contains the selective marker (Fig.2).

This strategy has been actually used for two main purposes:

- The disruption of the genomic copy of *narG* in *E. coli*, due to high risk of undesired recombination process between the finalized capture plasmid and this gene.
- To test the capacities of the λ red gam system and its compatibility with our *E. coli* strains. It was a critical aspect in the development of the Genefish tool, as previous RecD- strategy to increase the recombination frequencies in *E. coli* were not as satisfying as expected. These experiments are further described in Chapter 3 section 3.*ii*

References:

Baudin A, Ozier-Kalogeropoulos O, Denouel A, Lacroute F, Cullin C (1993) A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic Acids Res*. 21:3329-30

Datta S, Costantino N, Court DL (2006) A set of recombineering plasmids for gram-negative bacteria. *Gene*.379:109-15

Datsenko KA and Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A*. 97:6640-6645

Cherepanov PP, Wackernagel W. (1995) Gene disruption in *Escherichia coli*: Tc^R and Km^R cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene*. 158:9-14

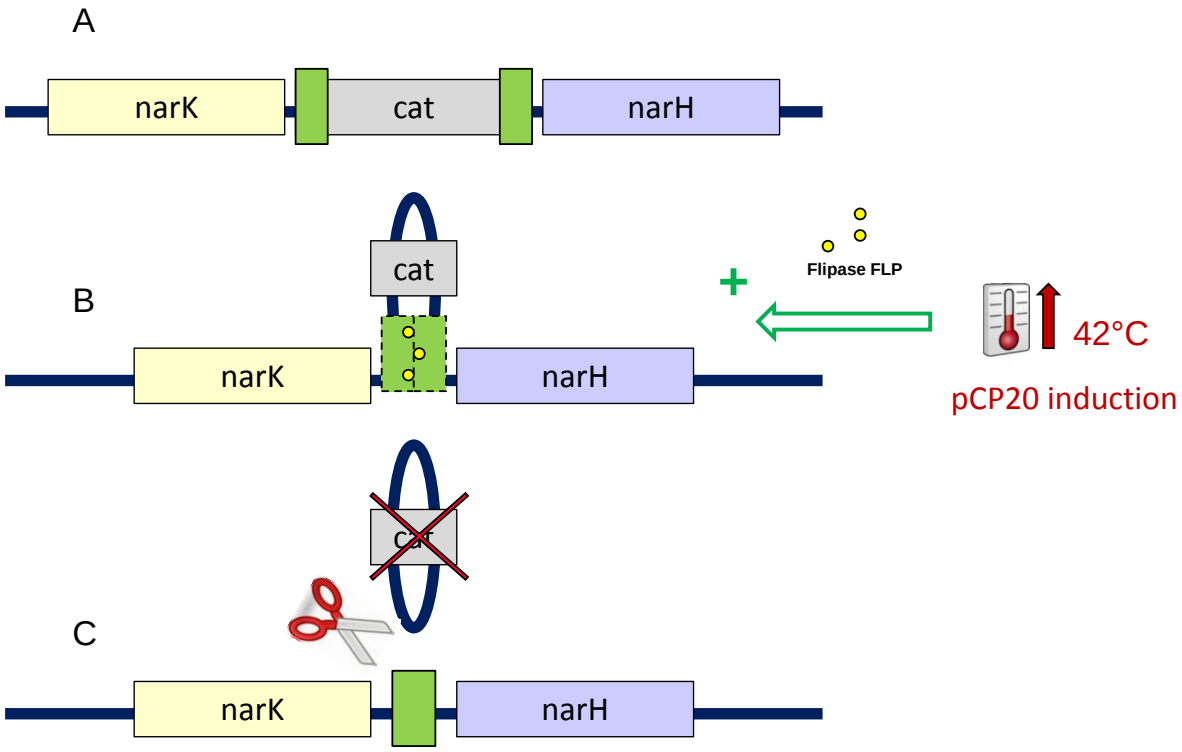


Fig. 2: Representation of the disruption of the selective marker thanks the FRT-FLP system in E. coli. Panel A shows the selective marker before induction of the flipase system. Panel B is describing induction of pCP20 at 4°C and how the flipase act on the homologous FRT regions, resulting in formation of loop. Panel C represent the final step consisting in cutting the loop, while a typical scare will remain on the chromosome

Appendix 3-4

-

Electrocompent cell preparation

1. All the steps shall be performed in a sterile and ventilated area (Bio-safety hood...)

2. Get a fresh plate culture of the suitable strain, grown overnight with the adequate selective markers and at the appropriate temperature. All strains used in this study have a genomic resistance gene to kanamycine, applied at 25µg/ml (Duchefa Biochemie, K0126.0025). Plates are prepared with Luria Bertani agar (LBa, Duchefa Biochemie, L1705.2500).

3. Preparation of inoculum: Pick an isolated colony from the plate, and inoculate it in a 50ml tube containing 5ml Luria Bertani medium (LBm, Duchefa Biochemie, L1703.2500) and the appropriate selective marker. Incubate overnight with shaking (150-200 rpm) at the adequate temperature.

4. Next morning, inoculate the appropriate volume (1/100 dilution) in an erlen containing LBm and the selective marker as well (e.g 1ml of inoculum for 100ml final volume). Incubate the erlen under shaking (150-200 rpm) at the appropriate temperature.

5. Measure the optical density (OD600nm) at regular intervals. Do not forget to establish the control with LBm containing the appropriate concentration of selective marker. Sample more frequently when the OD600 is reaching ≈ 0.15 (beginning of the exponential growth time period).

6. When the OD600 is reaching: $0.45 < OD600 < 0.55$, immediately stop the cellular growth by chilling the erlen on ice during 15-30min.

7. Get all the material cooled down at 4°C beforehand (glycerol solutions, centrifuge rotors, tubes...).

8. Transfer the erlen content in 50ml tubes, cooled down at 4°C beforehand, and spin the culture down at 4000g, 10min, 4°C.

9. First washing: discard carefully the supernatant and re-suspend cellular pellets with 5ml of cooled glycerol 10% (Duchefa Biochemie, G1345-5000) with a 5ml pipette. Pipette up and down slowly and carefully 5 times, and complete the volume to 50ml with cooled glycerol 10%. Invert the tubes 5 times to homogenize. DO NOT VORTEX. Spin again at 4000g, 10min, 4°C.

10. Second washing: repeat the step 8, but reduce the final volume to 20ml. Spin again at 4000g, 10min and 4°C.

11. Third washing: discard carefully the supernatant and re-suspend cellular pellets in 1ml of cooled glycerol 10%. Transfer the cells in a 2ml eppendorf and complete the volume up to 2ml with cooled glycerol 10%. Change the centrifuge rotor (cooled down at 4°C beforehand). Spin again at 4000g, 10min, 4°C.

12. Discard carefully the supernatant with a pipette P1000, and re-suspend the cells in the desired volume of cooled glycerol 10%. DO NOT VORTEX. Dispatch the cells in small aliquots of 50µl and use immediately in electroporation assays, or store the cells at -80°C up to 3 months.

13. Melt a bootle of LBa in a microwave and let it equilibrate in a water bath at 55°C with slight shaking if possible.

14. Put all the required material on ice at 4°C: electroporation tubes (Dominique Dutscher, 038190), electro-competent cells, sterile water, DNA preparations...

15. Prepare eppendorf 1.5-2ml tubes and LBm ready.

16. Set the electroporation apparatus device on the appropriate protocol (e.g E. coli, 2mm and 2.5kV, Gene Pulser Xcell Electroporation System, Biorad)

17. Be sure that the electro-competent cells are defrosted if previously stored at -80°C.

18. Take an electro-competent cells tube and open it in a sterile area. Drop off a droplet of the DNA material to be transformed on a wall of the electroporation tube and get the droplet down to the cells thanks to the pipette tip. Water can be used to help the droplet going down and to dilute salt in case of electrical break down. BE SURE TO ADJUST YOUR DNA CONCENTRATION BEFORE.

19. Pipette up and down smoothly to homogenize the DNA with the cells, and immediately transfer the electro-competent cells to a cooled electroporation tube. Try to avoid bubbles formation.

20. Close immediately the electroporation tube and clean the metal part of the system with fine paper to avoid residual humidity. Rapidly transfer the tube into the electroporation device, close the tip and launch the electrical pulse. THE TRANSFER BETWEEN DNA ADDITION AND THE ELECTRICAL PULSE SHOULD NOT EXCEED 30 SECONDS.

21. If an electrical break down is observed (explosion and flash), the assay is failed. Go back to step 17.

22. If the electrical pulse was correctly delivered, immediately add 950µl of LBm. Homogenize delicately by pipetting up and down. Transfer the cells in an eppendorf 1.5-2ml and incubate in a stove under appropriate temperature (37°C) with shaking (150-200 rpm) during 1h-1h30.

23. Prepare the plates with LBa, P10 for total bacteria and P15 for transformants detection. Prepare the dilution series as well with 900µl NaCl 0.8% (Roth, HN00.3). Add the appropriate selective markers to the medium for total bacteria, and transformants selection.

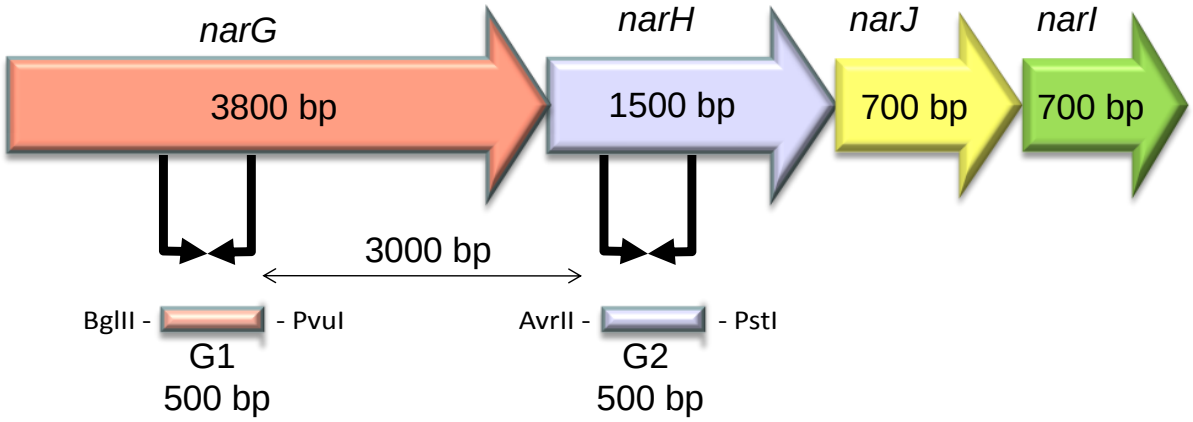
24. Get the cells back from the stove, and use 100µl for serial dilutions (10E-03 down to 10E-06). Use the remaining cells for transformants detection (between 300-500µl per plate, dilution 10E00 or 10E-01 depending on transformation efficiency.

Appendix 3-5

-

Cloning of the recombination sites

1. Amplification of the region *narG1* and *narG2* from the genomic DNA of *Pseudomonas fluorescens* (step realized by Laurent Philippot, INRA, Dijon). PCR primers used in this amplification are flanked with palindromic sequences of one of the restriction enzymes present into the 2 multi-cloning sites of the capture plasmid.



2. Preparation of the region *narG1* with double cohesive restriction enzymes BglII and PvuI (Fermentas, ER0081 and ER0621) in the yellow tango buffer, 2x final concentration (Fermentas).

Digestion mix	**Volume (µl)**	**Quantity (ng)**	**Concentration (ng/µl)**
narG1	5	250	50
Tango Buffer 10x	10	/	/
Sterile water	33	/	/
PvuI (10U/µl)	1	10U	/
BglII (10U/µl)	1	10U	/
Final volume	50	(Incubation 2-3h at 37°C)	

3. Let the mix incubate in a stove at 37°C during 2-3 hours.

4. Inactivate the restriction enzymes at 75°C during 2min. Purification of the digested product *narG1* with GFX PCR DNA & Gel Band Purification Kit (GE Healthcare). Eluate the DNA with 50µl of sterile water or TE buffer.

5. Use the digestion product directly, or store at -20°C.

6. Preparation of the double cohesive region *narG2* with restriction enzymes AvrII and PstI (Fermentas, ER1561 and ER0611) in yellow tango buffer, 1x final concentration (Fermentas).

Digestion mix	Volume (µl)	Quantity (ng)	Concentration (ng/µl)
narG2	10	200	20
Tango Buffer 10x	5	/	/
Sterile water	32	/	/
PstI (10U/µl)	2	20U	/
BglII (10U/µl)	1	10U	/
Final volume	50	(Incubation 2-3h at 37°C)	

7. Let the mix incubate in a stove at 37°C during 2-3 hours.

8. Inactivate the restriction enzymes at 75°C during 2min. Purification of the digested product *narG1* with GFX PCR DNA & Gel Band Purification Kit (GE Healthcare). Eluate the DNA with 50µl of sterile water or TE buffer.

9. Use directly the digested product, or store at -20°C.

10. **Preparation of the capture plasmid pB35K7tox.** The plasmid has been prepared through the CsCl gradient technique (See appendix 3.2).

11. Double digestion of pB35K7tox for *narG1* cloning, with restriction enzymes BglII and PvuI (Fermentas) in tango buffer 2x final (yellow, Fermentas).

Digestion mix	Volume (µl)	Quantity (ng)	Concentration (ng/µl)
pB35K7tox	10	750	75
Tango Buffer 10x	20	/	/
Sterile water	68	/	/
PvuI (10U/µl)	1	10U	/
BglII (10U/µl)	1	10U	/
Final volume	2x50	(Incubation 2-3h at 37°C)	

12. Let the mix incubate in a stove at 37°C for 2-3 hours.

13. Inactivate the restriction enzymes at 75°C during 2min. Purification of opened plasmid with GFX PCR DNA & Gel Band Purification Kit (GE Healthcare). Eluate the DNA with 2x45µl of sterile water or TE buffer to maximize the recovery yield before dephosphorylation step. *Verification of plasmid quality on gel to evaluate the expected size is advisable*.

14. Dephosphorylation step with alkaline phosphatase (Roche, 19758250001).

Dephosphorylation mix	Volume (µl)	Quantity (ng)	Concentration (ng/µl)
pB35K7tox	89	/	/
Alkaline phosphatase Buffer 10x	10	/	/
Sterile water	/	/	/
Alkaline phosphatase (1U/µl)	1	1U	/
Final volume	2x50	(Incubation 15min at 37°C)	

15. Inactivate the alkaline phosphatase at 75°C for 2min. Purification of opened and dephosphorylated plasmid with GFX PCR DNA & Gel Band Purification Kit (GE Healthcare). Eluate the DNA with 35µl of sterile water or TE buffer to concentrate the plasmid solution.

16. Load the plasmid on gel to evaluate the expected size (6500pb) and the concentration. Use it directly for ligation purposes, or store at -20°C up to 5 months.

17. Ligation of *narG1* in pB35k7tox with the T4 DNA ligase (Roche, 10481220001). Different molecular ratio were applied (pB35K7tox : *narG1*): ratio 1:1, 1:2, 1:3, 1:4 and 1:5. The formula used to calculate the molecular ratio is:

$$(nv / vs) \,.\, is$$

Were "*nv*" is the quantity of vector in ng (e.g 50ng of pB35K7tox), "*vs*" the vector size in bp (6500bp for pB35K7tox) and "*is*" the insert size (350bp for *narG1*).

***narG1* ligation mix**	**Volume (µl)**	**Quantity (ng)**	**Concentration (ng/µl)**
pB35K7tox double digest, dephosphorylated	3	qsp 50	17
Insert *narG1* (1:1)	2.7	67.3	25
Insert *narG1* (1:2)	5.4	134.6	25
Insert *narG1* (1:3)	8.1	201.9	25
Insert *narG1* (1:4)	10.8	269.2	25
Insert *narG1* (1:5)	13.5	336.5	25
Self-Ligation vector control	0	0	0
Tango Buffer 10x	10	/	/
Sterile water	qsp 25	/	/
T4 DNA Ligase (1U/µl)	2	2U	/
Final volume	25	(Incubation overnight at 16°C)	

18. Inactivation of the T4 DNA ligase 10min at 65°C. The ligation product can be loaded on gel to verify the size against the control.

19. Electro-transformation of the ligation product in the appropriate E. coli strain (Cf. Appendix 3-4 for more details)

20. Screen the transformant colonies by PCR to identify the ones that harbor the pB35G1 construction.

21. Positive clones detected after PCR that harbor the desired construction pB35G1 can be used for toxicity induction assays in order to select the best candidate for the next step. For more details on toxicity induction assays, please refer to Appendix 3-5.

22. Ligation of *narG2*. The plasmid pB35G1 was extracted through the Plasmid DNA Purification Kit, Endotoxin-Free (Macherey-Nagel). The plasmid has been double digested with restriction enzymes AvrII and PstI (Fermentas) in the yellow tango buffer 1x final (Fermentas).

Digestion mix	**Volume (µl)**	**Quantity (ng)**	**Concentration (ng/µl)**
pB35G1	10	450	45
Tango Buffer 10x	10	/	/
Sterile water	77	/	/
PstI (10U/µl)	2	20U	/
BglII (10U/µl)	1	10U	/
Final volume	2x50	(Incubation 2-3h at 37°C)	

23. Let the mix incubate in a stove at 37°C during 2-3 hours.

24. Inactivate the restriction enzymes at 75°C during 2min. Purification of pB35G1 with GFX PCR DNA & Gel Band Purification Kit (GE Healthcare). Eluate the DNA with 2x45µl of sterile water or TE buffer.

25. The dephosphorylation step and ligation of *narG2* were done exactly in the same conditions has before with *narG1* (Step 14 to 21).

26. The pB35G1G2 construction was transformed in the appropriate strain and the presence of both *narG1* and *narG2* were verified by PCR. The best toxic cassettes clones were selected by induction and the final optimal strains were used in recombination attempts.

Appendix 3-6

-

Toxicity induction assay

The selection of the best clone candidates is operated through the induction of the toxic cassette of the capture plasmid. This step is crucial, as the lethal construction is not stable during culture periods and also because the plasmid has been modified frequently. As a consequence, toxic induction assays were performed regularly, each time the system was modified, in order to maintain the integrity of the construction by selecting only the most efficient lethal plasmids.

The first step consists in screening a large number of clones harboring the plasmid construction that needs to be tested. This experiment is achieved by liquid culture in a multi-well system (MW96) under different conditions in order to activate the toxic genes. The condition tested corresponds to the induction of each toxic gene separately (*relF* with IPTG and colE3 with L-arabinose), and both at the time, and a growth culture control. After incubation, the clones corresponding to negative wells where no bacterial growths were detected in all the tested conditions are selected for further analysis on agar plates. This last step allowed us to evaluate the actual "escape rate" (*Er*) of selected clones by counting the exact number of bacterial cells that are able to grow under toxicity induction conditions on plates. Determination of *Er* value is an important criterion for finally selecting the right clone that would be used in further recombination experiments.

A. Preliminary induction assay on liquid medium

1. The plasmid construction that needs to be tested should be transformed in the appropriate strain. For example, the plasmid should be tested after cloning of the recombination sites. For more details about cloning and transformation protocols, please refer to Appendixes 3-3 and 3-4

2. After the transformation step, recover several clones that grow on plate with the selective marker of the capture plasmid in a multi-well plate with Lubria Bertani medium (LBm, Duchefa Biochemie, L1703.2500). Make sure that the selective marker of the capture plasmid is still present in the culture medium (Chloramphenicol 50µg/ml: Cm50; Duchefa Biochemie, C0113.0025)

3. Let the multi-well plate incubates in a stove at 37°C (or 29°C if pSIM6 is present in the strain) under shaking conditions (150-200 rpm) during 7-8 hours. Verify the clone growth frequently. Do not exceed 8h culture to minimize the probability of escaping phenotypes occurrence.

4. During incubation time, prepare 4 new multi-well plates:

    - Plate 1: Culture control, with LBm and Cm50, in order to verify if all the tested clones are growing normally without any induction of the toxic genes.
    - Plate 2: Induction of *relF* toxic gene, with LBm, Cm50 and IPTG 200µg/ml (isopropyl β-D-1-thiogalactopyranoside, Roche, 10724815001)
    - Plate 3: Induction of colE3 toxic gene, with LBm, Cm50 and L-arabinose (Roth, 5118,2) 1% final w/v (e.g, 10µl of a solution at 10% w/v in a final volume of 100µl into each well).
    - Plate 4: Induction of both genes at the time, with LBm, Cm50, IPTG 200µg/ml and L-arabinose 1% final.

5. As soon as the medium becomes turbid after 1-2 hour of incubation,, replicate the first plate in all the 4 others. Take care to not contaminate and mix the conditions together while replicating. Incubate the 4 new plates under the same conditions as the former one.

6. Save the former multi-well with glycerol 25% final. Store it at -80°C or -20°C.

7. Check the clone growth under each condition frequently and report everything for latter decision. Ideally, the clones that would be selected should not display any growth on all induction plates (plates 2, 3 and 4) after 24h incubation. The clones can be recovered from the former plate that was saved in glycerol, and then tested on plates for *Er* determination.

B. Induction assay on plates

1. Get an isolated colony from a fresh plate culture (Cm50), done with the selected candidates from the multi-well. Inoculate each isolated colony in a tube containing 15ml LBm and Cm50.

2. Let the culture incubate 8h in a stove at 37°C (29°C if pSIM6 is present in the strain), under shaking (150-200 rpm).

3. During the incubation period, get Luria Bertani agar medium (LBa, Duchefa Biochemie, L1705.2500) ready for preparing all the plates necessary for each tested conditions:

    - 4xP10 plates with Cm50 for determination of total bacteria (dilution 1E-04 down to 10E-07)
    - 3xP10 plates with Cm50 and 1% final L-arabinose (w/v) for determination of colE3 *Er* values (dilution 1E00 down to 1E-02)
    - 3xP10 plates with Cm50 and 500µg/ml final IPTG for determination of *relF Er* values (dilution 1E00 down to 1E-02)
    - 2xP15 plates with Cm50 and both IPTG (500µg/ml final) and L-arabinose 1% final w/v for determination of Er values under total induction of the toxic cassette (dilution 1E00 or 1E-01).

These plates should be prepared for testing only one clone, and more plates need to be prepared if several clones are tested at the same time. Use 15ml of LBA for P10 plates, and ideally 35-40ml per P15 plates. Allow the plates to dry off during 10-15min to facilitate the culture spread and liquid absorption. This will limit considerably the amount of moisture in plates, and facilitate identification of growing bacteria, especially for the highest concentration of cells. Do not over dry the plates as it may result in cracks formation in the medium.

4. Prepare 1.5ml eppendorf with 900µl NaCl 0.8% (Roth, HN00.3) for the serial dilutions.

5. Recover the cells through centrifugation at 4000g, 10min, 4°C. Discard the supernatant but keep enough liquid in the tube to resuspend the cells in a final volume of 1ml. Start the serial dilutions with 100µl of culture in 900ul of NaCl 0.8%.

6. Spread the bacterial dilution on all the plates with sterile balls or rakes. Close all the plates with parafilm to avoid contamination with volatile IPTG molecules.

7. Let incubate overnight in a stove at 37°C (or 29°C depending on the conditions), and count the growing bacteria on each plates. The *Er* values are obtained by dividing the number of growing bacteria on a specific induction condition with the number of total bacteria.

Appendix 3-7

-

The tandem recombination strategy

1. All the steps shall be performed in a sterile and ventilated area (Bio-safety hood...)

2. Get a fresh plate culture of the *E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6, grown overnight with the adequate selective markers (Cm50 and Cb50), at the appropriate temperature (29°C)

3. Preparation of liquid culure: Pick an isolated colony from the plate, and inoculate it in an erlen containing 100ml Luria Bertani medium (LBm) and the appropriate selective markers. Incubate during 8 hours with shaking (150-200 rpm) at the adequate temperature (29°C). DO NOT PROLONGE THE CULTURE PERIOD OVER 9 HOURS!

4. Measure the optical density (OD600nm) at regular intervals. Do not forget to establish the control with LBm containing the appropriate concentration of selective markers. Sample more frequently when the OD600 is reaching ≈ 0.15 (beginning of the exponential growth time period). The optical density should not exceed 0.7, which correspond approximately to 8h of culture.

5. During the culture time, prepare a water bath at 42°C for pSIM6 induction.

6. When the culture time is over, separate the 100ml into 2x50ml in smaller erlens. Place one erlen at 42°C in the water bath (induced condition), and the other one at 29°C (non-induce control). Each erlen should be exposed at the temperature treatment during 15min exactly, under shaking at 160rpm.

7. Place immediately the erlen in ice during 30min to stop the cellular growth.

8. Prepare the electro-competent cells from each erlen, as described in Appendix 3-4. Transform the stain with *narG* PCR product for each condition (42°C and 29°C). Several quantities of PCR product can be applied depending on the experiment design.

9. Place the electro-transformed cells at 37°C for the induced condition, and 29°C for the non-induced control. The cells should be exposed 1h-1h30 under 150-200rpm.

10. During the culture period, prepare all the plates required for the toxic induction step. Follow the procedure as described in Appendix 3-6. Make sure to prepare separate plates for each condition (42°C and 29°C). Plate large volume (400-500µl) of dilution 0 on large p15 plates to increase the screening for recombinant detection.

11. Dispose the plates at the appropriate temperature for growing step (37°C and 29°C). Make sure to separate the IPTG plates in a separate bag to avoid volatile contamination.

12. The colonies can be checked after 24-48h growth. Verification of recombinant clones is achieved by PCR verification on extracted plasmids.

Appendix 3-8

-

The co-transformation recombination strategy

1. All the steps shall be performed in a sterile and ventilated area (Bio-safety hood...)

2. Get a fresh plate culture of the *E. coli* 1661LI 14102 recD+ Δ*narG* pSIM6, grown overnight with the adequate selective markers (Cm50), at the appropriate temperature (29°C)

3. Preparation of liquid culure: Pick an isolated colony from the plate, and inoculate it in an erlen containing 100ml Luria Bertani medium (LBm) and the appropriate selective markers. Incubate during 8 hours with shaking (150-200 rpm) at the adequate temperature (29°C). DO NOT PROLONGE THE CULTURE PERIOD OVER 9 HOURS!

4. Measure the optical density (OD600nm) at regular intervals. Do not forget to establish the control with LBm containing the appropriate concentration of selective markers. Sample more frequently when the OD600 is reaching ≈ 0.15 (beginning of the exponential growth time period). The optical density should not exceed 0.7, which correspond approximately to 8h of culture.

5. During the culture time, prepare a water bath at 42°C for pSIM6 induction.

6. When the culture time is over, separate the 100ml into 2x50ml in smaller erlens. Place one erlen at 42°C in the water bath (induced condition), and the other one at 29°C (non induce control). Each erlen should be exposed at the temperature treatment during 15min exactly, under shaking at 160rpm.

7. Place immediately the erlen in ice during 30min to stop the cellular growth.

8. Prepare the electro-competent cells from each erlen, as described in Appendix 3-4. Transform the stain with both *narG* PCR product and pB35G1G2 for each condition (42°C and 29°C). Several quantities of PCR product and plasmid can be applied depending on the experiment design.

9. Place the electro-transformed cells at 37°C for the induced condition, and 29°C for the non-induced control. The cells should be exposed 1h-1h30 under 150-200rpm.

10. During the culture period, prepare all the plates required for the toxic induction step. Follow the procedure as described in Appendix 3-6. Make sure to prepare separate plates for each condition (42°C and 29°C). Plate large volume (400-500µl) of dilution 0 on large p15 plates to increase the screening for recombinant detection.

11. Dispose the plates at the appropriate temperature for growing step (37°C and 29°C). Make sure to separate the IPTG plates in a separate bag to avoid volatile contamination.

12. The colonies can be checked after 24-48h growth. Verification of recombinant clones is achieved by PCR verification on extracted plasmids.

Appendix 3-9

-

The segregation recombination strategy

1. All the steps shall be performed in a sterile and ventilated area (Bio-safety hood...)

2. Get a fresh plate culture of the *E. coli* 1661LI 14102 recD+ Δ*narG* pB35G1G2 pSIM6, grown overnight with the adequate selective markers (Cm50 and Cb50), at the appropriate temperature (29°C)

3. Preparation of liquid culure: Pick an isolated colony from the plate, and inoculate it in an erlen containing 100ml Luria Bertani medium (LBm) and the appropriate selective markers. Incubate during 8 hours with shaking (150-200 rpm) at the adequate temperature (29°C). DO NOT PROLONGE THE CULTURE PERIOD OVER 9 HOURS!

4. Measure the optical density (OD600nm) at regular intervals. Do not forget to establish the control with LBm containing the appropriate concentration of selective markers. Sample more frequently when the OD600 is reaching ≈ 0.15 (beginning of the exponential growth time period). The optical density should not exceed 0.7, which correspond approximately to 8h of culture.

5. During the culture time, prepare a water bath at 42°C for pSIM6 induction.

6. When the culture time is over, separate the 100ml into 2x50ml in smaller erlens. Place one erlen at 42°C in the water bath (induced condition), and the other one at 29°C (non induce control). Each erlen should be exposed at the temperature treatment during 15min exactly, under shaking at 160rpm.

7. Place immediately the erlen in ice during 30min to stop the cellular growth.

8. Prepare the electro-competent cells from each erlen, as described in Appendix 3-4. Transform the stain with *narG* PCR product for each condition (42°C and 29°C). Several quantities of PCR product can be applied depending on the experiment design.

9. Place the electro-transformed cells at 37°C for the induced condition, and 29°C for the non-induced control. The cells should be exposed 1h-1h30 under 150-200rpm.

10. Directly extract the plasmids from the freshly electro-transformed cells. Follow the instruction of the manufacturer (Plasmid DNA purification kit, endotoxine free, Macherey-nagel). Make sure to isolate plasmid from respectively all tested conditions (42°C and 29°C).

11. Transform again the freshly extracted plasmids into new electro-competent *E. coli* 1661LI 14102 recD+ Δ*narG*. Carefully adjust the ratio between plasmid and cells to 1:1 in order to ensure the molecular segregation.

12. Place the electro-transformed cells at 37°C for the induced condition. The cells should be exposed 1h-1h30 under 150-200rpm.

13. During the culture period, prepare all the plates required for the toxic induction step. Follow the procedure as described in Appendix 3-6. Make sure to prepare separate plates for each condition (42°C and 29°C). Plate large volume (400-500µl) of dilution 0 on large p15 plates to increase the screening for recombinant detection.

14. Dispose the plates at the appropriate temperature for growing step (37°C and 29°C). Make sure to separate the IPTG plates in a separate bag to avoid volatile contamination.

15. The colonies can be checked after 24-48h growth. Verification of recombinant clones is achieved by PCR verification on extracted plasmids.

Appendix 3-10

-

Escaping phenotypes and recombinant clone validation

Identification of positive recombinant clones can only be achieved by sequencing of plasmid inserts. The counter selective system based on toxic genes is working fine, as it considerably reduces to number of growing bacteria that need to be screened. However, the inherent background generated by the presence of escaping phenotypes limits our ability to detect and segregate with certainty the recombinant event from false positive. The only way to completely achieve the identification of recombination is to extract the plasmids of the clones that were able to grow under toxicity induction conditions, and screen by PCR the clones that harbor the expected size at the location between the two cloning sites. Normally, the counter selection system is extremely efficient, so no toxic cassettes should be detected in plasmids after induction. The only possibility for having a toxic cassette passing through our lethal system is punctual mutations that may be responsible in synthesis of non-functional proteins, thus resulting in a loss of toxicity.

However, the insert size of recombinant plasmid that fully integrated *narG* is approximately 3900bp while the size of the toxic cassette is approximately 3150bp. This difference can be easily detected by traditional electrophoresis techniques in 1% agarose gel, which is a reliable technique for preliminary screening of putative recombinants.

In addition, the escaping phenotypes are also usually easily identifiable on gel, as they often harbor shorter toxic cassette due to plasmid rearrangements (between 700-1700bp). Most of these rearrangements are occurring during the liquid culture step and are resulting in partial or total removal of the toxic cassette (Fig.1 and 2).

After Sanger sequencing verification, it appears that these deletions are always implicating *narG1*, which seems to be the starting point of the rearrangement. Usually, the shortest deletions are resulting in removal of *relF* (Fig.1 and 2, 1-2). Larger deletions are reaching part of colE3 as well (Fig.1 and 2, 1-3), and sometimes all the toxic cassette for the biggest ones (Fig.1 and 2, 1-4).

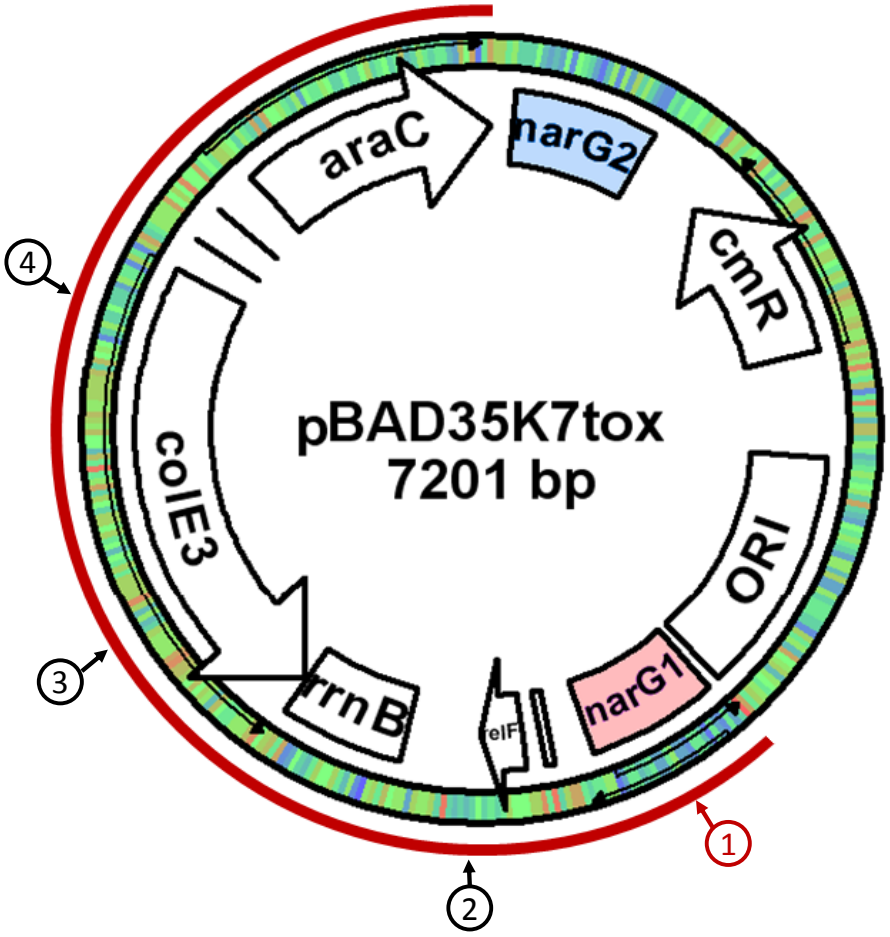


Fig.1: Most common deletion profiles observed on the capture plasmid pB35G1G2 during liquid culture period. Red arrow 1 is locating the start of all deletion profiles observed, corresponding apparetly to random sites into the sequence of narG1. Arrow 2 is showing the shortest deletion, leading to removal of relF toxic gene. Arrow 3 is representing a larger deletion, which is randomly targeting the colE3 sequence. Arrow 4 is displaying the largest deletion observed, including complete removal of both relF and colE3.

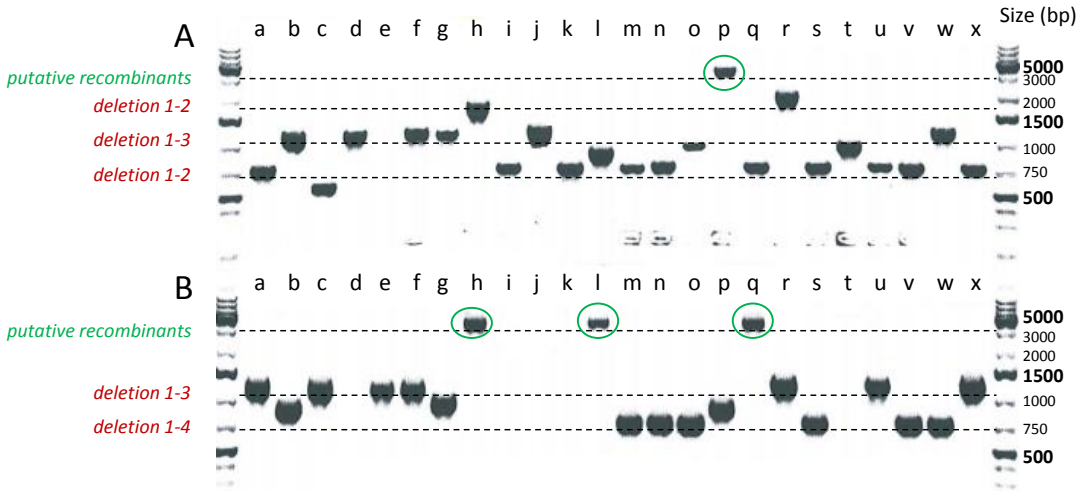


Fig.2: Gel profile of the PCR product obtained on capture plasmids extracted from bacteria after the selection step. The PCR is targeting the location of the toxic cassette with primers surrounding the two multi-cloning sites (R2-fw2 and R1-rv1, fig.3). The line corresponding to deletions (red) are representing the state of the toxic cassette after rearrangement (see fig.1). Deletion 1-2 is the shortest one, and is starting from narG1 and removes relF, resulting in a residual amplification of ≈2000bp. Deletion 1-3 is larger, and results in removing a part of colE3, with a residual cassette of ≈1200bp. Deletion 1-4 is the most consequent, with total removal of the two toxic genes and a residual signature of ≈800bp. Putative recombinants are easily detectable (green), as they are harboring long PCR signatures of 3900bp.

In order to further verify if the bands at 3900bp are corresponding to recombinant clones, internal PCR were carried on with both pB35 primers and *narG* primers (Fig.3). The primers used are described in Appendix 3-1.

If the 4 amplifications are working, a Sanger sequencing can be applied on each fragment to finally confirm the presence of *narG* in the capture plasmid. Each fragment is sequenced on the foward and reverse side, with approximately 800bp lenght which allows complete coverage of the location, between 1-3x (Fig.4). This strategy has permitted to obtain almost the full sequence of *narG* in all the putative clones tested. Only the low coverage regions (1x) are displaying sequence uncertainty. As this strategy was finally successful, identification of clones was mostly based on PCR afterwards, without further sequencing of all putative recombinants.

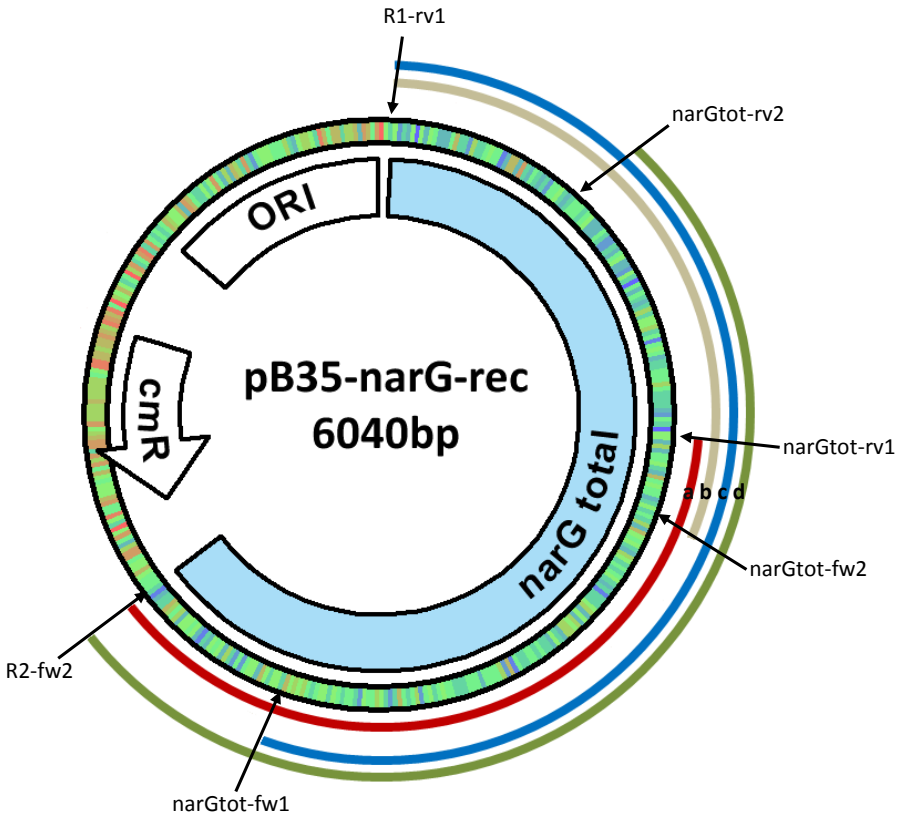


Fig.3: PCR for verification of putative recombinant clones. 4 PCR were carried on to fully cover the narG location. Red fragment (a) was obtained with R2-fw2 and narGtot-rv1 for a size of 1946bp. Grey fragment (b) was amplified with R1-rv1 and narGtot-fw2 for a size of 2174bp. Bleu fragment (c) was obtained with narGtot-fw1 and R1-rv1 for a size of 3607bp. Green fragment (d) was obtained with R2-fw2 and narGtot-rv2 for a size of 3488bp.

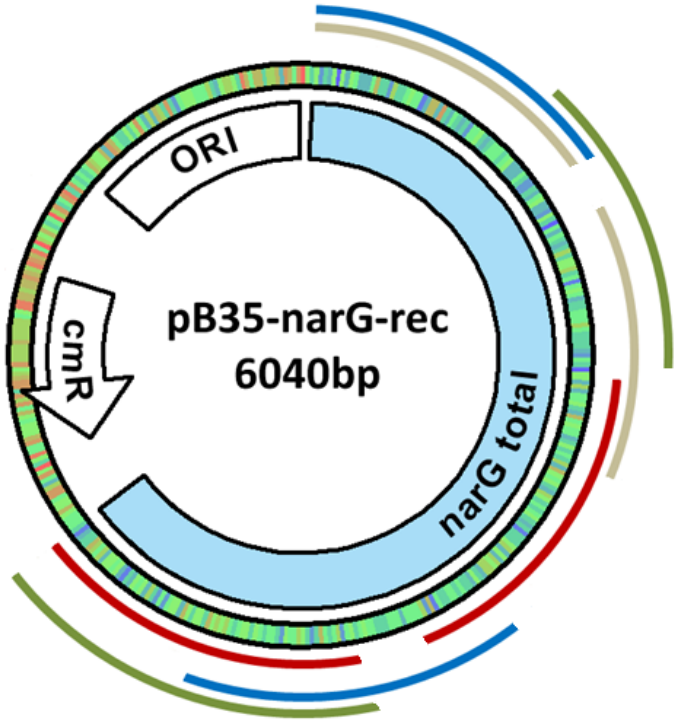


Fig.4: Sanger sequencing of the 4 PCR fragment in foward and reverse. The coverage varies between 1x up to 3x depending on the location.

CONCLUSION ET PERSPECTIVES

-

VERSION FRANÇAISE

Durant ces années passées au sein de l'équipe de Génomique Microbienne Environnementale, j'ai eu la possibilité de pouvoir aborder la thématique de ma thèse sous divers points de vue, à l'aide notamment des outils et technologies mis à notre disposition au sein du laboratoire, mais également de par l'assistance et le soutien indéniable de mes collègues. J'ai donc pu ainsi tenter de répondre à la problématique de ma thèse en l'appréhendant de deux manières distinctes.

L'axe de recherche principal développé dans le cadre de cette thèse est la recherche et la caractérisation de nouveaux gènes codant des enzymes d'intérêt industriel *via* l'exploitation de la ressource métagénomique. Afin de mener à bien cette tâche, deux approches différentes et complémentaires ont été mises en place :

- la première approche développée est basée sur une stratégie métagénomique que l'on pourrait qualifier de « classique », dans laquelle un environnement particulier : le sol de Rothamsted (Angleterre), a été utilisée afin de mener à bien les travaux visant à évaluer le potentiel des bactéries de ce sol comme réservoir de nouvelles enzymes.
- la deuxième approche repose sur une stratégie novatrice visant à développer un nouvel outil biotechnologique : Genefish, permettant d'exploiter la ressource métagénomique d'une façon plus ciblée.

La première approche a été menée en trois étapes :

1. une approche *in silico*, qui a permis d'établir un premier bilan sur le sol de Rothamsted concernant le potentiel de la communauté bactérienne de ce sol à héberger des gènes codant les enzymes impliquées dans la dégradation de la chitine. Cette étape a été principalement réalisée *in silico*, grâce à l'étude des séquences métagénomiques générées dans le cadre du projet Metasoil, *via* des approches BLAST.
2. une approche expérimentale complémentaire de la première, visant à cribler génétiquement la banque de clones réalisée dans le cadre du projet Metasoil *via* une approche d'hybridation utilisant des sondes radioactives visant des gènes d'intérêt.
3. une approche plus fondamentale visant étudier la possibilité d'accroitre la proportion de gènes d'intérêt en évaluant l'impact de l'enrichissement en chitine sur la communauté bactérienne du sol, soumise à des réarangements populationnels majeurs.

Cette approche de métagénomique « intégrative » comprenant plusieurs niveaux d'analyses a permis d'apporter des éléments de réponse et de compréhension sur la thématique de la dégradation de la chitine. Il a ainsi été mis en évidence que le sol présente un intérêt majeur concernant la présence de nouvelles chitinases dans les métagénomes. Cet aspect a été mis en évidence par comparaison de différents environnements sur le critère de leur diversité en termes de séquences métagénomiques affiliées aux enzymes connues pour être impliquées dans la dégradation de la chitine. Il est clairement apparu que, à volume de séquences égal, le sol présente une diversité plus importante que tous les autres environnements étudiés.

D’un autre côté, le criblage de la banque métagénomique du sol de Rothamsted a permis d’identifier de nombreuses séquences d’intérêt rentrant dans le cadre des objectifs du projet Metaexplore. Pour l’instant, le travail va se poursuivre en collaboration avec d’autres partenaires, experts en expression et en enzymologie, afin de mieux caractériser les gènes découverts, ainsi que les produits qu’ils codent. De plus, une collaboration avec l’équipe de l’Université de Ljubljana (Partenaire 14, Slovénie) a permis d’identifier une laccase bactérienne dans les contigs issus du séquençage des fosmides positifs trouvés sur les membranes. Cette détection s’est faite à l’aide du modèle HMM laccase développé par l’équipe de Ljubljana et de Bielefeld (Allemagne), marquant ainsi le fait que d’autres techniques d’identification existent et peuvent être appliquées avec succès pour identifier des séquences d’intérêt.

Enfin, d'autres membranes sont également disponibles afin de réaliser de nouveaux criblages, notamment sur la banque « chitine » réalisée à partir des ADN métagénomiques extraits du sol de Rothamsted après enrichissement en chitine 1x et incubation en microcosmes pendant 30 jours. D'autres sondes pourront alors être créées en se basant sur les résultats obtenus par l’approche d’enrichissement, en ciblant par exemple des gènes spécifiques de certains genres bactériens ayant répondu favorablement à l’amendement. Par exemple, les genres *Burkholderia*, *Xanthomonas* et *Streptomyces* pourraient faire des cibles idéales, avec l’espoir d’augmenter les fréquences d’hybridation et donc de détection de séquences d’intérêt. Toutefois, le risque de voir apparaître des séquences « nouvelles » risque d’être limité par une telle approche. La conception des sondes restera encore une fois le point crucial de cette approche qui devra respecter tous les impératifs à la fois de conservation mais également de divergence afin de garantir le succès de cette stratégie.

D’un autre côté, l’approche alternative Genefish doit pouvoir venir en complément de la métagénomique, en se basant sur une stratégie totalement différente et novatrice afin de pouvoir accéder à une diversité de séquences potentiellement nouvelles. En effet, le principe même de Genefish repose sur la recombinaison homéologue, à l’aide de zones de similarité de grande taille, mais également avec les recombinases du phage λ, tolérant un certain degré de divergence dans le processus recombinogène. Cette stratégie présente des similarités avec la technique d’hybridation présentée précédemment, notamment au niveau de la flexibilité dans le jeu de la recombinaison, ce qui doit pouvoir aboutir à la récupération de séquences divergentes et donc potentiellement différentes de celles des banques de données.

Cependant, même si les résultats obtenus jusqu'à présent sont encourageants, des modifications seront certainement à prévoir sur le système afin de le rendre plus efficace. Par exemple, transférer la cassette toxique sur le chromosome semble être la solution la plus judicieuse pour garantir l’aspect « copie unique » d’une telle construction. La solution plasmidique était de mise au départ car idéale pour la construction et les modifications génétiques. Maintenant, ce choix initial devient malheureusement limitant pour la suite du développement de l’outil. La nécessité d’augmenter le rendement de détection est crucial si l’on veut porter l’outil vers des applications plus concrètes comme la recherche de nouveaux gènes codant des enzymes d’intérêt. Sur ce point, un autre aspect de la technique va se montrer limitant : la nécessité de trouver deux zones de similarité encadrant la cible.

Si l'on prend le cas des chitinases, il sera très difficile de trouver des zones conservées permettant d'atteindre un maximum de diversité. Il serait certainement plus judicieux d'envisager des stratégies d'enrichissement afin d'augmenter les effectifs de certains genres d'intérêt comme évoqué précédemment et de trouver des zones conservées en alignant les génomes connus dans ces genres. De nombreux ajustements seront encore nécessaires afin d'optimiser l'outil Genefish mais les résultats obtenus et les pistes poursuivies actuellement laissent envisager des perspectives encourageantes pour l'avenir de cette approches novatrice.

CONCLUSION AND PERSPECTIVES

-

ENGLISH VERSION

During these years of preparation in the Environmental Microbial Genomics group, I had the chance to answer the problematic of my thesis through diverse point of view. This was possible thanks the wide range of tools and technologies available in my host laboratory, but also thanks the undeniable assistance and support from my collegues. As a consequence, I was able to tackle the questions raised by my thesis and by the objectives of the Metaexplore project by apprenhending it through two distinct approaches.

The main axe of this thesis is the search and the characterisation of novel enzymes of industrial interest *via* the exploitation of the environmental metagenomic ressource. In order to carry out this task, two different and complementary approaches were developped:

- The first one is relying on the « classical » metaegnomic approach, in which a particular environment: the soil form Rothamsted (UK), has been used to find new genes coding for enzymes of interests.
- The second approach relies on a novel strategy aiming at developping a new biotechnological tool: Genefish, allowing exploitation of the metagenomic resource through a totally different and innovative way.

These two strategies are included in the frame of the European project Metaexplore, and aim at exploiting the metagenomic DNA through very different, but complentary ways. The first approach was carried out through 3 steps:

- A preliminary bioinformatic approach, which allowed the establishment of a first statement on the Rothamsted soil by evaluating its potential for discovering new chitinases. This step was realized by analysis of the metagenomic sequences generated in the frame of the Metasoil project *via* BLASTx.

- The sequence-based screening of a fraction of the metagenomic clone library from the Rothamsted soil. The screening was realized through hybridization strategy involving radio-labelled probes.

- A more fundamental approach aiming at comparing the former soil of Rothamsted with itself beforehand enriched with chitin and incubated in microcosms. This strategy aims at investigating the impact of chitin enrichment onto the soil bacterial community, in order to identify the responding taxa.

This « integrative » metagenomic approach including several levels has brought numerous elements of comprehension and discussions toward the thematic of chitin degradation. It was shown that soil is a relevant target for the research of new genes involved in the first steps of chitin degradation. This aspect was highlighted by comparison of several environments based on their metagenomic resource, and their diversity in genetic signatures of known genes involved in the degradation of chitin. It was clear that, apart from the amount of metagenomic sequence used in the analysis, soil is presenting a remarquable pattern, with the highest level of diversity, and also the highest proportion of enzymes involved in the first steps of chitin degradation.

On the other hand, the screening of the metagenomic library of the Rothamsted soil resulted in identification of several sequence of interest, fitting the goals of Metaexplore. For now, the work will go on in collaboration with other partners with more expertise in expression and enzymatic assays, in order to better characterize these new genes, and their coding products. In addition, collaboration with the group from the University of Ljubljana (Partner 14, Slovenia) permitted to identify a bacterial laccase within the contigs from the sequencing of the positive fosmids found on the membranes. This detection was performed with a "laccase specific HMM screening" (Hidden Markov Model), developed by the group from Ljubljana, highlighting the fact that other techniques exist and work for identification of gentic sequences of interest. Furthermore, new membranes are still available in the frame of Metaexplore for new screening purposes. For instance, a chitin enriched fosmid library containing ≈100 000 clones was established from the soil that was incubated in microcosm during 30 days (chitin 1x). New probes can also be designed based on the metagenomic data accumulated on chitin enriched soil. For instance, these probes can be created to target genera that responded to the amendement, like Burkholderia, Xanthomonas and Streptomyces, in order to increase the hybridization yields. However, the risk to fing already known sequences might be strongly increase by such an approach. Probe design will remain as the key step of this approach, as compromise need to be established between conservation and divergence regarding the sequence, in order to guaranty the success of the hybridization strategy.

Finally, the alternative approach Genefish should be coming as a complementary way in metagenomics, based on a totally novel strategy in order to access potentially a different diversity of sequences. In fact, the principle of Genefish relies on homologous recombination, with large similarity regions, and also the use of the pahge λ recombinase system which tolerate a certain amount of sequence divergence during the process. This trategy is presenting some similarities with the hybridization approach described previously, especially concerning the flexibility of the recombination process, which is supposed to help recovering divergente and potentially new sequences. However, even if the results obtained so far are encouraging, the system still does require modifications in order to mecome more efficient. For instance, the transfert of the toxic cassette onto the chromosome seems to be a suitable solution, ensuring the "single copy" aspect of the toxic construction. The plasmidic solution was usefull at the begining, making more confortable the genetic construction and modifications. For now, this initial choice is unfortunately becoming limiting for further application. The necessity to increase the detection yield is crucial, as the tool is devoted to be applied for the recovery of new genes of interest. On this point, another aspect of the technique is going to become limiting: the need to find two regions of homology surounding the target. If we consider the example of chitinases, it is extremely hard to fing these kinds of regions without drastically reducing the taxonomical diversity of sequence involved. A previously exposed for hybridization, an enrichment strategy is certainly more advisable in order to increase the proportion of targets, and by focusing only on specific genera by aligning the known genomes to find such regions. Numerous modifications and adjustments are still needed in order to optimize the Genefih tool. However, the results that were obtained so far, and the solution proposed are suggesting optimistic forecast for the future of this novatrice approach.

Métagénomique et approches alternatives pour l'étude fondamentale et l'exploitation de la microflore tellurique

Résumé

Mes travaux de thèse ont été réalisés dans le cadre du projet Européen METAEXPLORE, visant à découvrir de nouvelles enzymes d'intérêt industriel à partir des communautés microbiennes environnementales *via* les approches dites de « métagénomique ». Je me suis personnellement focalisé sur la recherche de chitinases dans les communautés bactériennes du sol, en particulier celui de la station expérimentale de Rothamsted (Royaume-Uni). Des approches de séquençage en 454 ont été réalisées afin de caractériser les bactéries dégradant la chitine au travers d'une stratégie d'enrichissement du sol en microcosme. J'ai également pu réaliser des criblages génétiques de banque de clones fosmidiques afin d'identifier des gènes d'intérêt potentiel. J'ai également été impliqué dans le développement d'un nouvel outil biotechnologique appelé « Genefish », dont le but est la capture de séquences d'ADN environnementales d'intérêt sur un plasmide à l'aide d'une souche *E. coli* ultra-recombinogène. Mes travaux seront présentés sous la forme de trois chapitres, reprenant successivement :

- Chapitre 1 : Une synthèse bibliographique du contexte scientifique
- Chapitre 2 : La recherche de nouvelles chitinases *via* des approches métagénomiques
- Chapitre 3 : Le développement et l'utilisation de l'outil « Genefish

---

Metagenomics and alternative approaches for the fundamental study and exploitation of soil microbes

Abstract

I realized my PhD in the frame of the European project METAEXPLORE, which aims to discover new enzymes of industrial interest from the environmental microbial communities through the metagenomic approaches. I was personally focused on finding new chitinases within the soil bacterial communities, with a particular emphasis on the Park Grass soil from Rothamsted research station (U.K). Pyrosequencing approaches were applied in order to characterize chitin degrading bacteria through a soil enrichment strategy in microcosm. I was also involved in genetic screening of fosmid clone library for identifying potential genes of interest. Furthermore, I participated to the development of a new biotechnological tool called "Genefish", which aims to capture environmental DNA sequence of interest into a plasmid thanks an hyper-recombineering *E. coli* strain. My PhD work will be presented in 3 chapters:

- Chapter 1: A literature review of the scientific context
- Chapter 2: The search for new chitinases through metagenomic approaches
- Chapter 3: The development and the use of the "Genefish" tool

DOMAINE : Microbiologie Environnementale

MOTS CLEFS : *Microbiologie, Bactérie, Métagénomique, Enzymes*

LABORATOIRE : AMPERE, UMR 5005 CNRS, Ecole Centrale de Lyon, 36 Avenue Guy de Collongue, 69134 Ecully