Aix-Marseille Université, Faculté de Médecine de Marseille

Ecole Doctorale des Sciences de la Vie et de la Santé

THÈSE DE DOCTORAT

Présentée par

Charbel ABOU-KHATER

Date et lieu de naissance: 08-Juilllet-1990, Zahlé, LIBAN

En vue de l'obtention du grade de Docteur de l'Université d'Aix-Marseille
Mention: **Biologie**, Spécialité: **Microbiologie**

Caractérisation de nouveaux gènes et polymorphismes potentiellement impliqués dans les interactions hôtes-pathogènes

Publiquement soutenue le 5 Juillet 2017 devant le jury composé de :

Pr. Daniel OLIVE	Directeur de Thèse
Pr. Brigitte CROUAU-ROY	Rapporteur
Dr. Benoît FAVIER	Rapporteur
Dr. Pierre PONTAROTTI	Examinateur

Thèse codirigée par Pr. Daniel OLIVE et Dr Laurent ABI-RACHED

Laboratoires d'accueil

URMITE Research Unit on Emerging Infectious and Tropical Diseases, UMR 6236, Faculty of Medicine, 27, Boulevard Jean Moulin, 13385 Marseille, France

CRCM, Centre de Recherche en Cancérologie de Marseille,Inserm 1068, 27 Boulevard Leï Roure, BP 30059, 13273 Marseille Cedex 09, France

Acknowledgements

First and foremost, praises and thanks to God, Holy Mighty, Holy Immortal, All-Holy Trinity, for His showers of blessings throughout my whole life and to whom I owe my very existence. Glory to the Father, and to the Son, and to the Holy Spirit: now and ever and unto ages of ages.

I would like to express my sincere gratitude to my advisors Prof. Daniel Olive and Dr. Laurent Abi-Rached, for the continuous support, for their patience, motivation, and immense knowledge. Someday, I hope to be just like you.

A special thanks to my "Godfather" who perfectly fulfilled his role, Dr. Pierre Pontarotti for his continuous daily support. Dr. Pontarotti was always there to listen and to give advice. He showed me different ways in research problem and the need to be persistent to accomplish any goal.

I want to thanks Prof. Didier Raoult and Mediterranean Infection for the funding and making my PhD experience productive and stimulating.

I am heartily thankful to Marie-Hélène, for being an excellent "mother substitute" during all these years. Without her encouragement and constant support I could not have finished this work.

I would also like to thank all of my Indian friends in the lab: Vivek, Vikas, Jai, Dhamo, Sourabh, Arup, Sweta for all their support in all situations. You were such great ambassadors of your beloved mother India. I thank my other fellow lab mates in both EBM and CRCM labs for the stimulating discussions, motivation and laughs during all these years. Special thanks to Sandrine, Marine, Olivier, Julien, Hassnae, Louis! Thanks for all your encouragement!

I would like to thank my Lebanese friends and compatriots for accepting nothing less than excellence from me.

I am also grateful to my family: Randa, Jean, Hoda, Chantal & Chris, because I owe it all to you. Many Thanks especially for your love, prayers and encouragement!

Last but not the least, I would like to thank Carole, for supporting me throughout all these days... Distance never separates two hearts that really care! Thanks for your presence, your love and your ongoing prayers… Thank you for your patience on my moody days. Thank you for being my number one! I love you!

Charbel

Table of Contents

List of Abbreviations

1000G: 1000 Genomes project
AIDS: Acquired immune deficiency syndrome
APC: Antigen-presenting cell
BTN: Butyrophilin
CD: cluster of differentiation
CDS: Coding DNA sequence
CEACAM: Carcinoembryonic antigen-related cell adhesion molecules
CFTR: Cystic fibrosis transmembrane conductance regulator
CLEC: C-type lectin domain containing
CNV: Copy number variation
DEF: Defensin
DNA: Deoxyribonucleic acid
HbS: Hemoglobin subunit
HGNC: HUGO Gene Nomenclature Committee
HGPD: Human Gene and Protein Database
HIV: Human immunodeficiency virus
HLA: Human leukocyte antigen
HPA: Human platelet antigens
HSV: Herpes simplex virus
HWE: Hardy–Weinberg equilibrium
IFN: Interferon
Ig: Immunoglobulin
IgC: Constant immunoglobulin domain
IgV: Variable immunoglobulin domain
IL: Interleukine
IPD: Immuno Polymorphism Database
IMGT: International ImMunoGeneTics project
KIR: Killer-cell immunoglobulin-like receptors
LDL: Low-density lipoprotein
LILR: Leukocyte immunoglobulin-like receptors
LRP: Low density lipoprotein receptor-related protein
MERS: Middle East Respiratory Syndrome
MHC: major histocompatibility complex
mRNA: Messenger RNA
NGS: Next-generation sequencing
NGS: Next-generation sequencing
NK: Natural killer cells
NS: Non-synonymous
PSG: Pregnancy specific beta-1-glycoprotein
RNA: Ribonucleic acid
RNAseq: RNA sequencing
SARS: Severe acute respiratory syndrome
SIGLEC: sialic acid binding Ig like lectin
SLC: Solute carrier
SNP: Single nucleotide polymorphism
TB: Tuberculosis
TCR: T cell receptor
TGF: Transforming growth factor
TLR: Toll like receptor
TM: Transmembrane
TNF: Tumor necrosis factor
TREM: Triggering receptor expressed on myeloid cells
TRIM: Tripartite motif
UCSC: University of California Santa Cruz
UTR: Untranslated region
VDR: Vitamin D receptor
WBC: White blood cells
WHO: World Health Organization

Chapter 1: Introduction

Introduction

I- Genetic susceptibility to infectious diseases

1- Host-pathogen relationships after the Neolithic Revolution

During our recent evolutionary history, several migrations and cultural changes occurred. About 12,000 years ago, the Neolithic period commenced and had a lot of impact on the lifestyle and overall health of humans. A progression of changes in behavior and cultures was due to a demographic transition from a nomadic lifestyle of hunting and gathering to dependence on agriculture and settlement. Drastic effects were observed, with changes in their natural environment, dietary intake and the way of procuring food and other daily necessities.

This transition and settlement also affected the physical activity of human beings who gradually adopted a sedentary way of life. Increasingly larger populations settled in closer proximity to mammals and different other animals. These settled communities domesticated animals (and also plants) which were kept inside or near the house. Due to this proximity, some microbes were transferred from animals to humans. Thus, the adoption of this new lifestyle increased zoonotic infectious diseases. Parasitic infections were facilitated by the lack of good hygiene practices in these densely populated settlements along with close contact with waste water sources and affected both animals and humans. However, adopting agricultural non-nomadic lifestyle did not necessarily affect the mortality, suggesting that healthy people having healthy diets, started to develop resistance to some infections.[1]

Several devastating pandemics occurred during the human history, resulting in large number of deaths. The “Black Death” killed one third of the European population in the 14th century, due to the *Yersinia pestis* bacterium[2], causing different forms of plague. An unusual deadly influenza pandemic also known as “Spanish flu” involving H1N1 influenza viruses resulted in the death of 50 to 100 million people in the early 20th century.

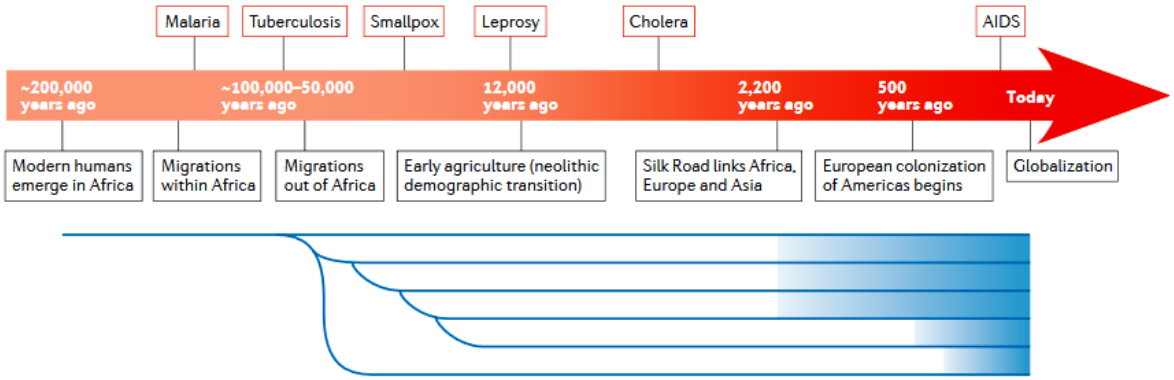


Figure 1. Recent human evolution and emergence infectious disease. Blue lines represent the fragmentation of the human lineage into genetically and geographically distinct populations. After the migration out of Africa, this fragmentation was accelerated. These populations started mixing more (represented by blue shaded regions between the populations)[3]

2- The host-pathogen co-evolution

Host-pathogen co-evolution and interactions affect the genetic diversity of both organisms. Pathogens contribute as one of the main selective pressures in shaping human genetic evolution.[4] Adaptive genetic changes can occur between any interacting populations. The term "co-evolving" means that both pathogen and host have evolutionary effects on one another, due to the intimate nature of the host-pathogen interactions. Pathogens need to survive within hosts and have to find a way to be transmitted from one host to another to ensure their species' survival and achieve their ecologic success. On the other side, the host immune system seeks to eliminate or restrict the pathogen's replication. Both the host immune responses and the pathogens exert selective forces on each other (Figure 2). Thus, these ongoing cycles of adaptation and counter-adaptation result in enhanced host defense mechanisms and/or novel attack techniques, especially when the interaction persists over time. [5,6] Figure 2 shows a schematic representation of co-evolution, emphasizing reciprocity in that changes in allele frequencies due to selection in one species impose selection resulting in changes in allele frequencies in the other.

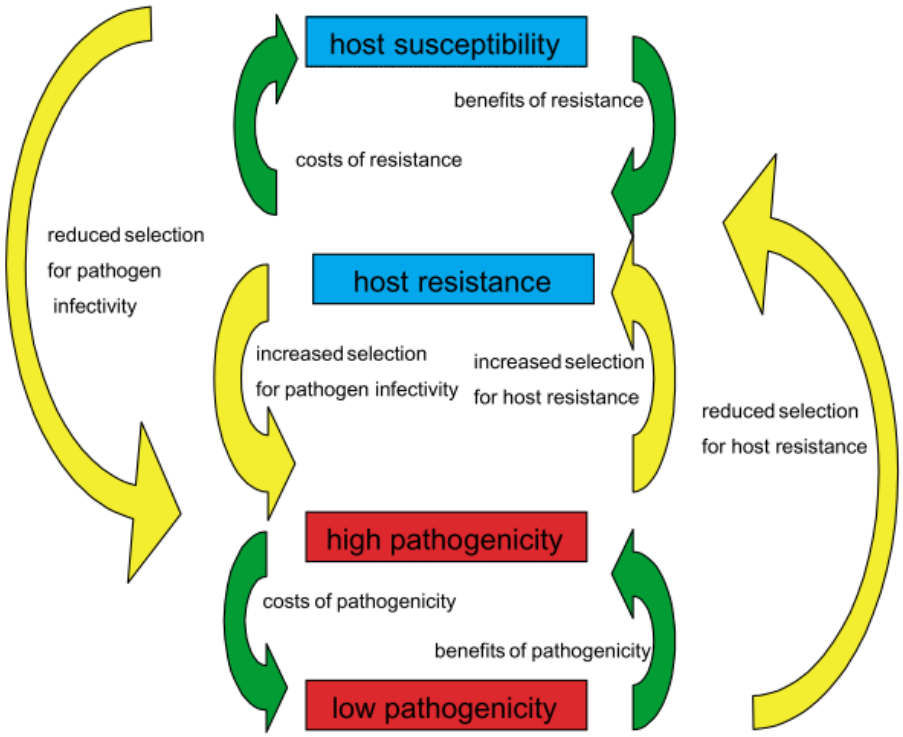


Figure 2. Schematic representation of host pathogen co-evolution and selective pressures.[6]

3- Pathogens adapt to their host

Most of the microbes that interact with humans are "friendly" and don't cause any illness, e.g. the microbiota or the normal flora, which provides several benefits for their host. On the other hand, several protective defense mechanisms help the host to overcome microbes that can cause any damage by invading and/or producing toxins. Pathogens can infect and cause diseases in a healthy host (primary pathogens) or in immunocompromised individuals (opportunistic pathogens).

Starting by the invasion of the host through primary barriers, pathogens could then evade the local and tissue host defenses. They can replicate and spread in the host's body. To eliminate or control these pathogens, the host relies on its immunological abilities. Several virulence factors provide to these microbes the opportunity to infect, such as the ability to adhere to the host, to colonize it and to evade the immune defenses.

Pathogens adapt to their host and try to escape the immune system, which seeks to eliminate or restrict the pathogen's replication. Successful pathogens are those who can avoid or adapt to evolving host defenses, by overcoming the sophisticated innate and adaptive immune responses (skin and mucosal secretions, non-specific local and/or inflammatory responses, specific immune responses). Microbes evolve greater virulence and use different mechanisms to evade the host defenses, such as secreting toxins or modulators (ligand or receptor mimics), inhibiting phagocytosis, latency, killing or subverting immune cells, inhibiting complement/cytokines/interferon/chemokines, downregulating some genes etc. [7]

For example, HIV-1 (human immunodeficiency virus- 1) adapts to human leukocyte antigen (HLA) class I. Successful control of HIV infection is mediated by HLA molecules such as HLA-B*57, HLA-B*27 and HLA-B*51. Immune recognition and killing by $CD8^+$ T cells is enabled once the viral epitopes are presented on the surface of infected cells. Kawashima *et al.* documented an evidence of HIV adaptation to HLA at a population level[8].

4- Factors involved in host-pathogen interactions

Different factors are implicated in host-pathogen interactions, like for example the environment, the biology of the pathogen, as well as the biology of the host (figure 3).

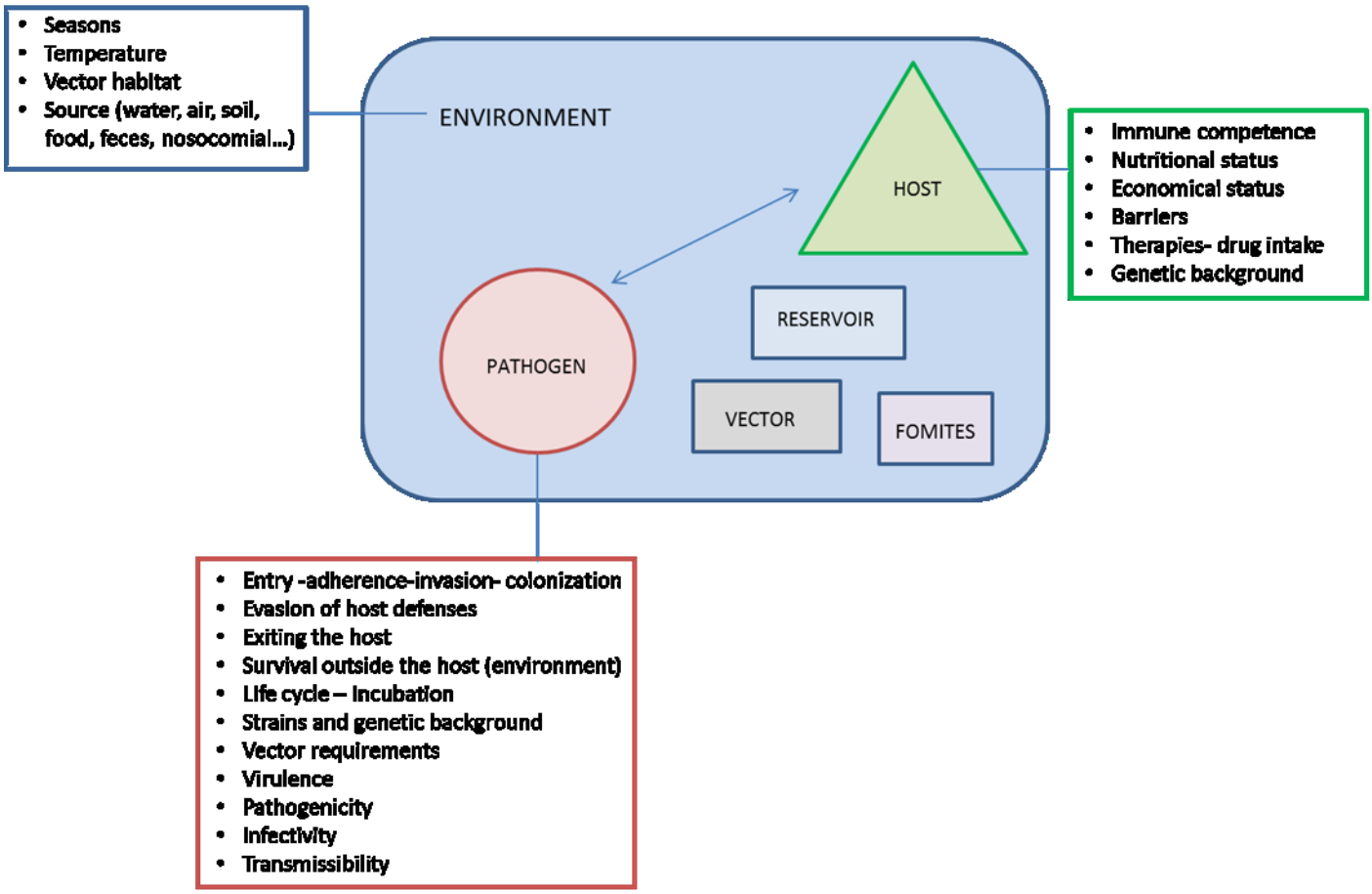


Figure 3. Major factors implicated in host-pathogen relationships

The main objective of this thesis project is to focus on the considerable role of the host's genetic profile by finding novel gene candidates and polymorphisms that can explain the resistance/susceptibility to infectious diseases.

5- Resistance and susceptibility to infectious diseases

In many human infections, only some exposed individuals develop clinical illness and respond differently. Different factors related to the pathogen and/or the host, are considered to have an important role in explaining these inter-individual differences in susceptibility.

The behavior of microbes can vary so greatly between strains. As for example, in the case of HIV that causes HIV infection and over time acquired immunodeficiency syndrome (AIDS), HIV-2 is characterized by lower transmissibility and reduced likelihood of progression to disease, and remains largely confined to West Africa, whereas HIV-1 extends worldwide. [9]

6- Genetic basis of infectious disease susceptibility in humans

Accidents occurred in history after deliberate inoculation of the same strain of an organism simultaneously to previously unexposed individuals. Effects of individual variation are therefore seen. One of the major examples is the accidental administration of *Mycobacterium tuberculosis* to a population in a small town in Northern Germany called Lübeck. In 1929, the Lübeck hospital did not respect the instructions given by the French researchers who were distributing the vaccine against tuberculosis (TB) around the world (*M.bovis* BCG, Bacillus Calmette-Guérin). A cross contamination of the vaccine with fully virulent clinical isolates of *Mycobacterium tuberculosis* occurred, because they were both kept in the same incubator. Unfortunately, on three occasions, 251 babies were fed with this contaminated vaccine. Interestingly, the infants were not given the same amount of the virulent bacteria, resulting in different levels of virulence in each vaccine preparation. The outcome shown in Table 1 depends on the exposure intensity (virulence level 1 with no virulent bacteria, while high levels of *Mycobacterium tuberculosis* in level 4) [10].

Virulence level	Number	Disease severity			
		Death	Serious disease	Mild symptoms	No symptoms
1	1	—	—	—	1
2	93	6 (6.5%)	9 (9.7%)	78 (83.8%)	—
3	83	18 (21.7%)	34 (41.0%)	31 (37.3%)	—
4	74	53 (71.6%)	18 (24.3%)	3 (4.1%)	—
Totals	251	77	61	112	1

Table 1 The Lübeck accident

These infants responded in a different manner, it seems evident that some of them resist more to infection by *Mycobacterium tuberculosis.* Around 31% died, 24% had serious disease, while 45% only showed mild symptoms. All of these babies are newborn. Therefore, the resistance could not be due to social factors. Genetic factors probably made some of them more susceptible than others.

Other incidents also showed that individuals respond differently to infectious diseases. Early experiences of therapeutic infection with microbes were conducted by the Austrian psychiatrist Julius Wagner-Jaureggi (Nobel Prize in Medicine in 1927) to treat mental diseases particularly

neurosyphilis with the “fever therapy”. Several artificial methods such as typhoid vaccines, the *Streptococci* of the erysipelas disease and tuberculin were utilized to produce fever. He concluded that malaria was the most satisfactory to treat neurosyphilis. These experiences led to the observation that individuals differed in their response to the *Plasmodium* parasite that causes malaria. Similar differences can now be seen in volunteers infected during malarial vaccine trials.[11,12]

In the 1970s and 1980s, contaminated haemophilia blood products were a serious public health problem. Individuals having this inherited genetic disorder that impairs the body's ability to make blood clots (needed to stop bleeding) were unknowingly given HIV-infected blood. Interestingly, the rate at which these patients progressed to develop an immunodeficiency differed between individuals[12].

7- Twin studies demonstrate the importance of host genetic factors

Human heterogeneities in the response to infections are well established in twin studies. In monozygous twins (sharing 100% identical genetic constitution), the concordance rate of a disease having genetic causes is higher than among twins sharing an average of 50% of genetic material (dizygous twins).

Results from selected twin studies of infectious disease

		Concordance*	
Disease	**Country**	**MZ (%)**	**DZ (%)**
Tuberculosis	Germany	65	25
	USA	62	18
	UK	32	14
Leprosy	India	52	22
Poliomyelitis	USA	36	6
Hepatitis B	Taiwan	35	4

*Concordance refers to the probability that one twin develops disease if the other is affected. Identical twins are monozygotic (MZ) and non-identical twins are dizygotic (DZ).

Figure 4 –Twin studies highlight the genetically determined components of the response to infectious diseases.[12] (adapted from the cited paper)

The main question remains, what are the factors in the host–pathogen interplay that dictate whether infection progresses to overt disease or not? Infection outcome is determined by the interaction

between bacterial and host genetic components. Several studies show that susceptibility genes and genetic background have, at least in part, a significant contribution in this variability of response against infectious diseases.

8- Examples of genetic susceptibilities to human infectious diseases

Genetic factors modulate the resistance (some of them are known) and may explain why some people resist infection more successfully than others. Here we present some examples available in the literature (Table below). Associations identified by case-control studies are more convincing and proposing biologically plausible gene candidates for further studies.

Infectious agent or disease	**Genetically associated**	**Association**	**References**
Severe malaria *Plasmodium falciparum*	Sickle-cell disease Heterozygosity for the HbS variant of haemoglobin is strongly protective	Resistance	[13]
Malaria	α- thalassaemias α^{+}-thalassemias result in protection against malaria	Resistance	[14] [15]
Severe malaria in East and West Africa	G6PD deficiency (X-linked enzymatic disorder) Hemizygous males are protected	Resistance	[16]
Severe cases of cholera	Blood group O (of the ABO system) O group is found at a higher frequency in these severe cases than in the general population	risk of infection with *V. cholera* O1 is lower in individuals with blood group O, but increased severity of disease once these individuals are infected	[17]
Plasmodium vivax	Duffy blood group (genetic variant)	DARC-mediated entry of *P.vivax* into erythrocytes prevented; Duffy-blood-group-negative human erythrocytes, *FyFy* →*resistance*	[18]

Pseudomonas aeruginosa	Cystic fibrosis Human airway epithelial cells expressing the ΔF508 allele of the *CFTR*: defective in uptake of *P. aeruginosa* → increased amounts of bacteria in the lungs	Susceptibility	[19]
Salmonella typhi	Cystic fibrosis High incidence of cystic fibrosis might be an evolutionary consequence of some protection from typhoid or another gut infection.	Role of CFTR in the internalization of both bacteria. ΔF508 allele *CFTR* (phenylalanine deleted at residue 508) →Resistance to typhoid	[20]
Pneumocystis carinii and *Cryptosporidium parvum*	CD40 ligand mutations (X-linked immunodeficiency with hyper-IgM)	Susceptibility	[21,22]
Environmental mycobacteria	*IFNγR1* recessive mutation	Homozygotes susceptibility to severe infections by these normally innocuous microbes	[23]
Opportunistic non-tuberculous mycobacterial and salmonella infections	Deletions in *IFNGR1* gene	Dominant susceptibility	[24]
Opportunistic non-tuberculous mycobacterial and salmonella infections	IL12 deficiency Or mutation in IL12 receptor	Susceptibility	[25,26]
Streptococcus pneumoniae	MyD88 mutations (loss-of-function) (TLR –IL1R signaling)	Susceptibility	[27]
Encapsulated bacteria (*S.pneumoniae*, *N.meningitidis*, *Haemophilus influenza*)	gene–disease association studies C5, C6, C7, C8, C9, CFP, CFD *(N. meningitides)*	Susceptibility	[28] [29–32] [33,34]
Mycobacteria	Proteins type-1 cytokine pathway	Susceptibility	[35]

HIV-1	CCR5-Δ32 allele of *CCR5*; homozygous are resistant to HIV-1 infection	Resistance	[36,37]
HSV-1	TLR3 deficiency (TLR3 required to control HSV-1)	Susceptibility	[38]
Human herpesvirus-8	STIM1 T cell deficiency → development of lethal Kaposi Sarcoma	Susceptibility	[39]
Human papillomaviruses	nonsense mutations in *EVER1* and *EVER2* associated with Epidermodysplasia verruciformis (skin disorder)	Susceptibility	[40]
Norovirus	*FUT2* Gene Homozygous Nonsense Mutation (428G→A)	Resistance	[41]
Candida albicans	chronic mucocutaneous candidiasis IL-17RA deficiency	Susceptibility	[42]
Bovine spongiform encephalopathy prion	Homozygous for methionine at codon 129 of the prion protein (PrP) gene.	Susceptibility	[43]

9- Host adaptation, signatures of selection

Pathogens became one of the major selective forces shaping human genetic diversity. Unusual patterns of allele frequencies for a certain locus are called “signatures of selection”, when compared with neutral distribution in the genome. We distinguish different types of natural selection: purifying selection (selection against deleterious or detrimental variations, also known as negative selection), positive selection (if a new advantageous variation arises), and balancing selection (when selection favors and maintains two or more intermediate states or alleles of a certain variation). (Figure 5)

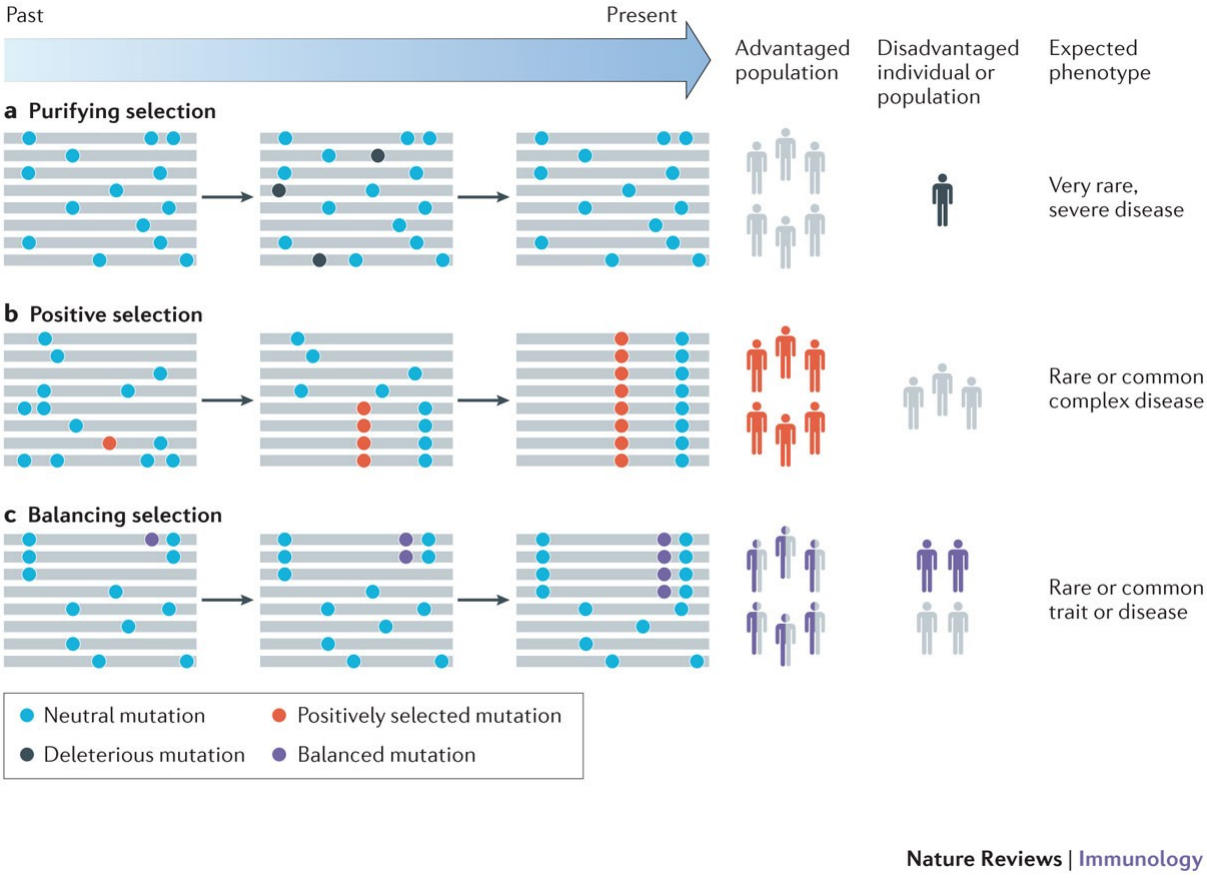


Figure 5-Types of selection and their legacy on the human genome.[44]

Prevalence of naturally selected alleles that were more advantageous increased. This is more evident in case of pathogens that emerged earlier during human history, having with *Homo sapiens* long-standing relationships, e.g. tuberculosis, leprosy, malaria, smallpox, cholera etc. New threats continue to appear, such as the Acquired Immune Deficiency Syndrome *(*AIDS), the Middle East Respiratory Syndrome (MERS), and the Severe Acute Respiratory Syndrome (SARS). Some pathogens induce severe illnesses (like the cholera infection caused by *Vibrio cholera*) while others cause chronic infections. Individuals display variable immune responses to fight infections.

Statistical tools detect signatures of selection, helping us to elucidate immune adaptations against pathogens; accurate detection is facilitated by next-generation sequencing generating whole genome datasets to analyze. Detection of natural selection could be altered by different other population genetic events such as admixture, genetic drift, bottlenecks, gene flow etc.

II- Human genetic variations

1- Major categories of human genetic variation

We distinguish different major categories of human genetic variation. In our work, we focused on Single-base-pair changes, a specific type of variation, to find novel gene candidates implicated in determining differential susceptibility to infectious diseases of humans. Here we briefly describe these different categories:

- Structural rearrangements

Structural rearrangements are a type of chromosome abnormality that reshuffles the DNA nucleotide sequence. Events like deletions, duplications, inversions, and translocations produce new chromosomal arrangement usually caused by a breakage in the double stranded DNA at two different locations, followed by a reattachment of the broken ends, resulting in an abnormal configuration. These changes have no functional consequences unless they affect a gene and thus could give rise to a disease. For example, the chromosome 9 and 22 translocation resulting on the BCR–ABL gene fusion is associated with chronic myelogenous leukemia[45]. These chromosomal rearrangements could affect a whole chromosome, like in the case of the Klinefelter syndrome in males and where one extra X chromosome is added to a Y chromosome such that there is XXY instead of the normal XY genotype[46].

- Insertion and deletion events

Cystic fibrosis is a genetic disorder caused by mutations in the human *CFTR* gene (that encodes the cystic fibrosis transmembrane conductance regulator). The predominant allele is a phenylalanine deletion termed ΔF508. It concerns a deletion of 3 base pairs, CTT, in the *CFTR* gene.[47,48]

Huntington's disease and the fragile X syndrome are examples of insertion mutation wherein trinucleotide repeats are inserted into the DNA sequence leading to these diseases. HTT is a gene known to be associated with Huntington disease. The exon 1 HTT mutation involves a DNA segment known as a CAG trinucleotide repeat. According to the CAG repeat expansion, the HTT alleles are

classified as follow: normal alleles (CAG segment repeated less than 26 times within the gene), intermediate alleles (27-35 CAG repeats) and Huntington's disease-causing alleles with 36 or more CAG repeats (risk of developing the disease in their lifetime). People with 36 to 39 CAG repeats may or may not develop the signs and symptoms of Huntington disease (Reduced-penetrance HD-causing alleles), while people with 40 or more repeats develop the disorder with great certainty. [49]

- Copy number variation

Copy number variation (CNV) is a type of structural variations in the human genome. It is a type of duplication, insertion or deletion event that affects a considerable number of base pairs (1 kb or larger)[50]. Thus, sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population.

- Transposons and retrotransposons

Transposable and retrotransposable elements provide a source of human genetic variation. These DNA sequences (around 44% of the human genome)[51] can change their position within a genome, sometimes creating mutations that lead to disease[52].

- Single-base-pair changes

Single nucleotide substitution is the most common type of mutational change. The replacement of one nucleotide by another is called a transition when the change occurs from one purine base to another (e.g. A to G) or between C and T (pyrimidines). The change is called transversion when it occurs from a purine to a pyrimidine base (e.g. G to) or vice versa. A common single nucleotide substitution in the population is called Single Nucleotide Polymorphism (SNP). A locus is commonly defined as "polymorphic" when the frequency of the most common allele is less than 99%. 3–10 million SNP variants in the human genome with a frequency higher than 1% are estimated.[53]

Any missing or added nucleotide causes a frame shift mutation and modifies entire reading frame of the sequence. Substitutions that occur in coding regions (within exons) can affect the protein product. The mutation is called silent (or synonymous) if it does not alter amino acid sequences. A

nonsynonymous mutation affects the protein product resulting in a modified codon that specifies a different encoded amino acid (missense mutation) or resulting in the production of a truncated protein when the mutation changes the codon into a stop-codon (nonsense mutation). According to the genetic code, all changes in the second amino acid of the codon are nonsynonymous.

Genetic polymorphisms describe the inter-individual DNA differences that make each human genome sequence unique. Several genetic disorders are the direct result of a mutation in one gene.

Sickle Cell Anemia is an example of a disease caused by a substitution mutation. This monogenic disorder caused by homozygosity for a single human β-hemoglobin gene mutation where the GAG codon mutates to GTG and leads to a Glutamic acid-Valine (Glu - Val) change[54].

These different genetic variations can have serious health consequences (like for example β-hemoglobin gene mutations) or may have no direct effect for the affected person (like balanced translocation evens). It depends mainly on when and where such variations occur.

2- NGS data: continuously improving technologies

New available technologies allow us to sequence DNA and RNA. This process revolutionized the study of genomics and enabled researchers to study more precisely any biological system.

Over the past few years, the development of Next Generation Sequencing (NGS) approaches and other technologies have continuously been improving. Since the sequencing of one human genome was declared complete in April 2003, we passed nowadays to thousands of human genomes and beyond especially after substantial reductions in sequencing costs. By using faster software systems, sequencing is taking less time and is much easier to perform. Thus, huge amounts of data are generated. Complete genomes and exomes (coding regions) sequences made available and published in open resources and databases freely accessible to the worldwide scientific community, represent a powerful source to define the human genetic diversity worldwide. These human genomics resources have made it possible to gain increasing key insights into the nature of genetic diversity, population

history and evolution as well as more advances in human population genetics and comparative genomics.

Recent technologies known as high-throughput sequencing such as Illumina, Roche 454, Ion torrent and SOLiD sequencing are much more quick and cheap comparing to other old methods.

Nowadays, research is more focused on finding –characterizing- new polymorphisms across the human genes to find out the functional consequences of such variations. Understanding this variability could find some interesting applications in risk assessment of disease and in predicting responses to treatments.

Thus, several comprehensive resources on genetic variations in humans are available, allowing further downstream analyses such as detection of signals of natural selection. Thanks to the development of high-throughput sequencing and genotyping, researchers are able to find stronger associations between genetic variations and disease (various autoimmune disorders, resistance or susceptibility to infectious diseases etc.).

SNP call sets generated using different methods from different sequencing centers and platforms are publicly accessible. The main problematic idea concerns the accuracy of SNPs' calling, especially when using low-coverage data, not only in population genetic studies but also in RNA-sequencing and expression level, *de novo* sequencing or other mapping studies.

Short reads result from NGS by digesting randomly the whole genome (or targeted regions such as exons in the Exome sequencing). These reads are then assembled or mapped to a reference genome[55]. After aligning these fragments to the reference sequence, variable sites are identified when a single nucleotide (A, T, C, or G) is altered. These single nucleotide polymorphisms (SNPs) create a unique pattern of each individual. The identification of newly SNPs is followed by their association with different phenotypes, complex diseases or genetic disorders. Once the allele frequency is determined, we can know the ratio of the less common variant comparing to the more common variant, within the population or between different human populations.

Alignment errors, SNP-calling errors, high nucleotide identity with paralogues or repeats, present, among others, different technical challenges when dealing with NGS data.

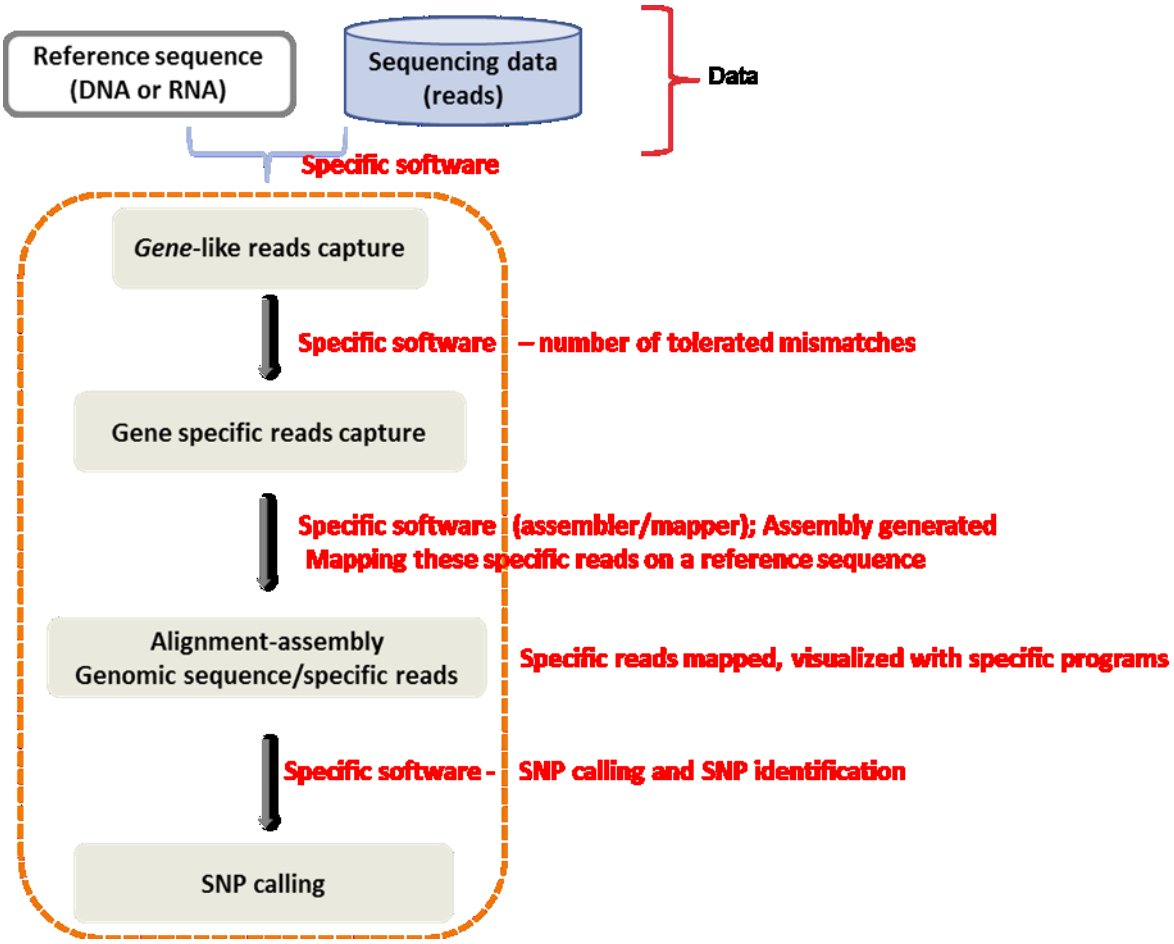


Figure 6. Steps for converting raw next-generation sequencing data into a final set of SNP or genotype calls.

Sequence reads represent short pieces of DNA (or RNA for RNAseq) obtained after sequencing. In order to be useful, these reads need to be mapped. Aligning short reads to a reference sequence is called “mapping”. Numerous mapping programs have been developed and vary in their algorithms and therefore speed. Coverage is the average number of reads that “cover” or map to a specific region on the reference sequence. In next-generation sequencing, coverage level helps us to determine the degree of confidence in the case of variant discovery at particular nucleotides. When a great number of sequence reads are aligned (high level of coverage), the variant calls can be made with a higher degree of confidence.

Several parameters define the alignment process, including for example the number of differences allowed between the reference sequence and the read (number of mismatches allowed), the number allowed and penalty for gaps etc. Optimizing these parameters in the pipeline will help us to obtain high quality mapped reads and thus more accurate downstream results (also with other criteria like sequencing quality scores that measure the probability that a base is correctly called or not).

III- Features of immunity genes

The field of human population genetics has been accelerating since the advent of next-generation sequencing technologies, making it possible to evaluate and figure out how genetics contribute to the variability of immune responses between infected individuals. Thanks to genome-wide association studies at the population level, researchers were able to identify several genes involved in the increase of resistance or susceptibility to infectious diseases.

Comparative genome analysis between different species reveals that immune genes, crucial for survival, differ in sequence and number, like for example between human and mice, or between two different insects *Drosophila melanogaster* and *Anopheles gambiae*. For humans, around 5% of the genes are involved in defense[56]. The human immune genes must evolve to counteract the new sophisticated ways developed by pathogens to avoid recognition and/or elimination by the defense system. Trowsdale and Parham listed some exploited features of immune genes in order to increase their potential defense role[56]. They cited for example some of these features: polygeny (e.g. MHC, Ig, KIR, FcR), polymorphism (MHC, KIR, TLR), clustering, linkage disequilibrium, rapid evolution, co-evolution etc.

Duplication of immune genes keeps the useful function of the protein while the other copy could acquire a new function after accumulating mutations. These duplications events along a chromosome generate multiple copies of the genes with different degrees of relatedness. Paralogous clusters of genes on different chromosomes also exist and are generated by bloc duplication. For example, MHC-related genes are found on 4 human chromosomes 1, 6, 9 and 19.

IV- HLA, the extraordinary level of diversity

The Human Leucocyte Antigen (HLA) system plays a crucial role in immunity. These cell-surface proteins present microbe-derived or self-derived antigenic peptides to T cells and thus trigger an immune response. Located, among others, on the short arm of chromosome 6 (6p21.3), the HLA molecules are encoded by 6 class I and 18 class II genes (Figure 7). These 2 classes differ both in structure and function: almost all nucleated cells express class I molecules (HLA-A, B, and C). MHC class I molecules present endogenously peptides from inside the cell, having either self or non-self (pathogen) origin. The structure of these molecules consists of a unique α chain bound to a small β2-microglobulin polypeptide. Presenting pathogenic or non-self intracellular peptides to cytotoxic T cells (CD8+) leads to the lysis of the infected cells.

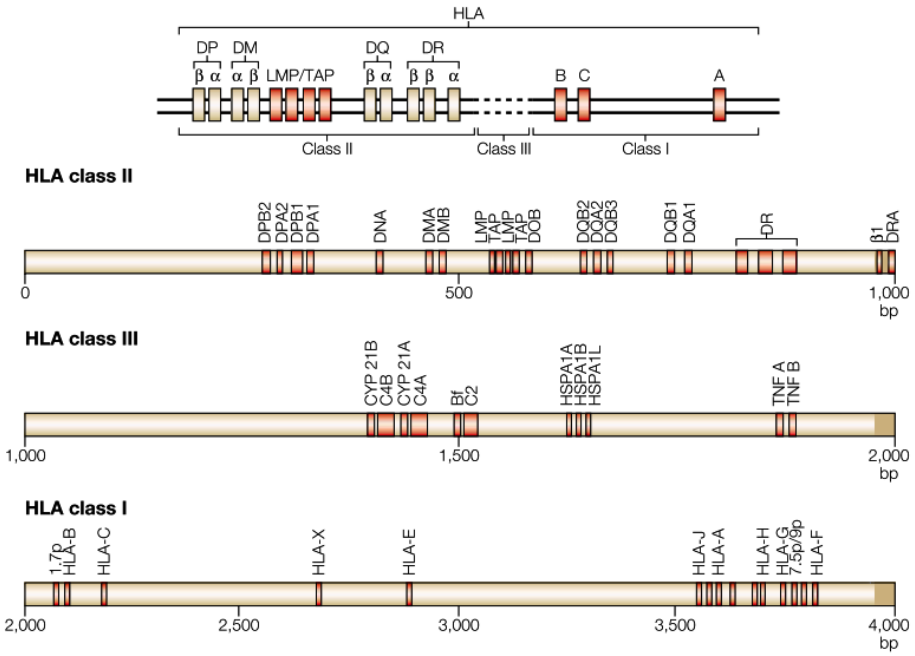


Figure 7- The major histocompatibility complex extended over 4 Mb of DNA on chromosome.

MHC class II molecules present antigens from outside of the cell (through the endocytic route, e.g. from parasites) to T-lymphocytes (CD4+ helper T cells) leading to antibody-mediated immune response. These proteins are expressed by some cells of the immune system such as the Antigen

presenting cells (APC): dendritic cells and macrophages. The structure of these class II molecules consists of two chains, *α* and *β*.

Due to the extraordinary level of *HLA* variation in the human populations, a large set of antigenic peptides could be recognized. In pathogen-rich environments, higher fitness is observed for heterozygotes at the *HLA* loci. This hypothesis was initially suggested in studies of the murine MHC[57]. A very high level of diversity is maintained at the population level (balancing selection) and higher rate of non-synonymous variations in protein-binding sites is observed at the molecular level. Studies indicate that *HLA-DQA1* and *HLA-C* are showing the strongest balancing selection while *DPB1* is compatible with neutrality.[58,59,60,61]

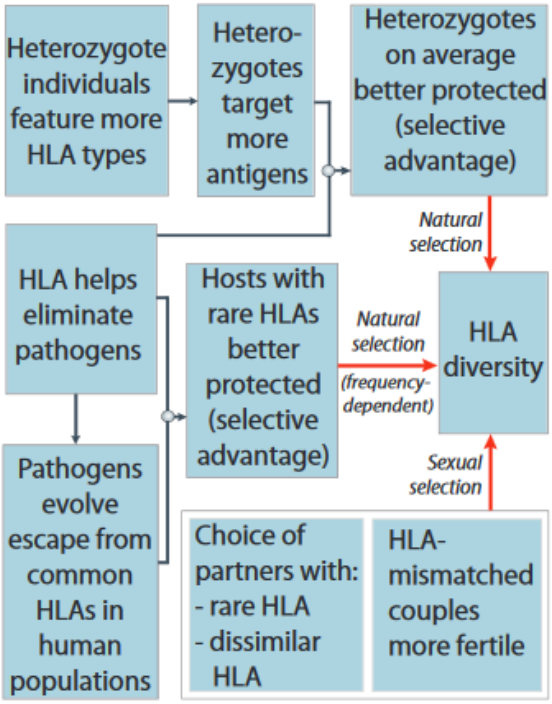


Figure 8- Evolution and diversity of *HLA* genes. [62]

Natural selective forces are involved in the evolution of the *HLA* loci (stochastic forces such as genetic drift are also involved). (Figure 8)

The IPD-IMGT/HLA Database provides a specialist database for sequences of the human major histocompatibility complex (Table 2), and currently contains 15,819 allele sequences (November 2016).

	HLA Class I						HLA Class I - Pseudogenes										Other non-HLA Genes				
Gene	*A*	*B*	*C*	*E*	*F*	*G*	*H*	*J*	*K*	*L*	*P*	*T*	*U*	*V*	*W*	*X*	*Y*	*HFE*	*MICA*	*MICB*	*TAP1*	*TAP2*
Alleles	3399	4242	2950	21	22	53	12	9	6	5	5	0	0	3	0	0	3	6	105	42	12	12
Proteins	2396	3131	2089	8	4	18	0	0	0	0	0	0	0	0	0	0	0	4	82	28	6	5
Nulls	157	134	109	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	0

	HLA Class II											HLA Class II - DRB Alleles								
Gene	*DRA*	*DRB*	*DQA1*	*DQB1*	*DPA1*	*DPB1*	*DPB2*	*DMA*	*DMB*	*DOA*	*DOB*	*DRB1*	*DRB2*	*DRB3*	*DRB4*	*DRB5*	*DRB6*	*DRB7*	*DRB8*	*DRB9*
Alleles	7	2018	69	911	43	644	6	7	13	12	13	1883	1	77	24	26	3	2	1	1
Proteins	2	1470	33	622	21	529	0	4	7	3	5	1369	0	63	16	22	0	0	0	0
Nulls	0	51	1	24	0	16	0	0	0	1	0	44	0	1	4	2	0	0	0	0

Table2. Number of named alleles for each *HLA* gene according to the IPD-IMGT/HLA Database. A total number of 14,473 *HLA* alleles, 177 other non-*HLA* alleles and 7 confidential alleles.

Table 3 shows some examples of available associations found in the literature, correlating between the allelic *HLA* polymorphism and infectious diseases.

	Disease	Effect
Class I		
B8	Pulmonary tuberculosis	Susceptibility
B35	AIDS	Susceptibility
B53	Severe malaria	Resistance
B57	AIDS	Resistance
Class II		
DRB1*1302	Hepatitis B	Resistance
DRB1*1352	Malarial anaemia	Resistance
DRB1*1101	Hepatitis C	Resistance
DRB1*04	Typhoid fever	Resistance
DR2	Pulmonary tuberculosis	Susceptibility
DR2	Leprosy	Susceptibility
DR7	Hepatitis B	Susceptibility

HLA, human leukocyte antigen.

Table 3. Associations between *HLA* alleles and infectious disease.[12] **(adapted from the cited paper)**

Main objective

Host-pathogen co-evolution and interactions contribute to shaping the genetic diversity of both organisms. The main objective of this thesis is to find novel gene candidates and polymorphisms that could define the genetic basis of the variability in resistance/susceptibility to infectious diseases, in particular to tuberculosis through the development of large-scale *in silico* screens and polymorphism analysis. Natural human worldwide diversity for genes is investigated using a specialized *in silico* approach for coding regions.

The most interesting genes will be selected by considering that, as a hypothesis, the most variable genes are more involved in host-pathogen interactions. Immune genes targeted by natural selection will be isolated.

Chapter 2: Materials and Methods

Methods

1- Pipeline description

Analysis of exome sequencing data to generate genotype (polymorphism) data was carried out with a dedicated pipeline. Our strategy was developed specifically for the analysis of multigenic families which may contain highly similar genes and/or pseudogenes. Briefly, it consists of four major steps: a capture of gene-like reads, isolation of reads specific to the gene, mapping of these specific reads to on a reference sequence and SNP calling. An additional "high sensitivity" SNP calling was performed to detect SNPs in low coverage regions. (Figures 1-3).

1. Capture of gene-like reads from the whole exome sequences data using the genomic reference sequence (masked sequence).
2. Isolation of reads specific to the gene. All reads that are identical to the gene as well as reads that contains up to six tolerated mismatches maximum.
3. Mapping of these specific reads on a reference sequence. Specific-reads are mapped to the reference sequence.
4. High sensitive SNP calling, even in low coverage genes.

Standard analysis

In order to perform a specificity analysis in the subsequent step, it is important to reduce the size of the pool of sequences and so this first step will capture all the gene-like sequences from the whole exome sequence data using reference genomic sequences where the repetitive elements in introns and UTRs were masked with RepeatMasker[63] (http://www.repeatmasker.org/). This analysis was performed with Bowtie2[64], a tool for aligning sequencing reads to long reference sequences, by making an index in a set of reference files, and taking all the reads one by one and finding which ones align to the reference.

Step 2. This step is the "specificity step" and will isolate the gene-specific reads: only the reads strictly more closely related to the gene will be kept. This analysis includes all the known alleles for each gene and starts with the reads identical to the gene (and to the alleles if they were added) and slowly increases the numbers of tolerated mismatches (up to six mismatches) to account for novel alleles and sequencing errors. This analysis is also performed with Bowtie2.

Step 3. Gene-specific reads are then mapped on the reference sequence using MIRA (a multi-pass DNA sequence data assembler/mapper)[65]: a SAM assembly is generated for analysis in Step 4 as well as an assembly that can be visualized with the GAP4 program of the STADEN package[66].

Step 4. SNP calling is performed using SAMtools[67] and the SNPs detected in the coding regions and donor and acceptor splicing sites are then isolated. This step is performed using SAMtools. We detect the SNPs in the coding regions, even those represented by a single read. After that, a manual verification of the obtained genotypes, to be sure that that the called SNPs do not correspond to noise especially for rare SNP detected in less than 10 individuals, we manually inspected and verified the sequence assemblies. Verification is necessary to identify false positives generated due to sequencing errors or contamination.

High sensitivity analysis

In order to detect SNPs in low coverage regions that might be missed in the normal analysis, we conducted a 'high sensitivity' analysis. For this analysis steps 2 to 4 are carried out in the same way but with a highly sensitive SNP calling in step 4: in this analysis, all the SNPs are called even those represented by a single read, but only for the positions detected as variable in the normal analysis. Subsequently, a comparison of the high sensitivity and the normal analysis is conducted.

Genotype verification

The 'high sensitivity' analysis will rescue SNP missed in the normal analysis but will also generate many false positives (sequencing errors or contaminations). To identify these false positives we plotted, for each variable position, a distribution of the sequence ratios (reads supporting the reference sequence divided by reads supporting the alternative sequence) and the genotype coverage for all individuals. This graph allowed us to identify typical coverage and the range of sequence ratio for heterozygous and homozygous types. Putative SNP with a sequence ratio out of the heterozygous or homozygous (for the non-reference) range were then discarded.

SNPs with a heterozygote sequence ratio and a low coverage indicate the presence of a low coverage region. In most cases, a low coverage region is the result of a loss of reads in Step 2 due to the presence of a conserved region between two or more genes or of a highly divergent allele. The latter scenario occurred for *SIRPA* gene (chapter 5) due to the presence of divergent allelic lineages: for several positions divergent reads were eliminated due to a number of differences higher than the maximum we set (six differences). These issues were corrected by adding these divergent sequences in the pool of allele sequences in Step 2 so that no reads was discarded.

Allele reconstructions

Genotype phasing was performed using PHASE 2.1[68,69]. The input files were prepared using SNPtools[70].

2- Allotypes nomenclature

Allotypes are forms of the protein. Their nomenclature is based on the HLA convention: the first two digits correspond to amino acid differences. Allotypes were considered as common if present in at least 5 individuals in the panel.

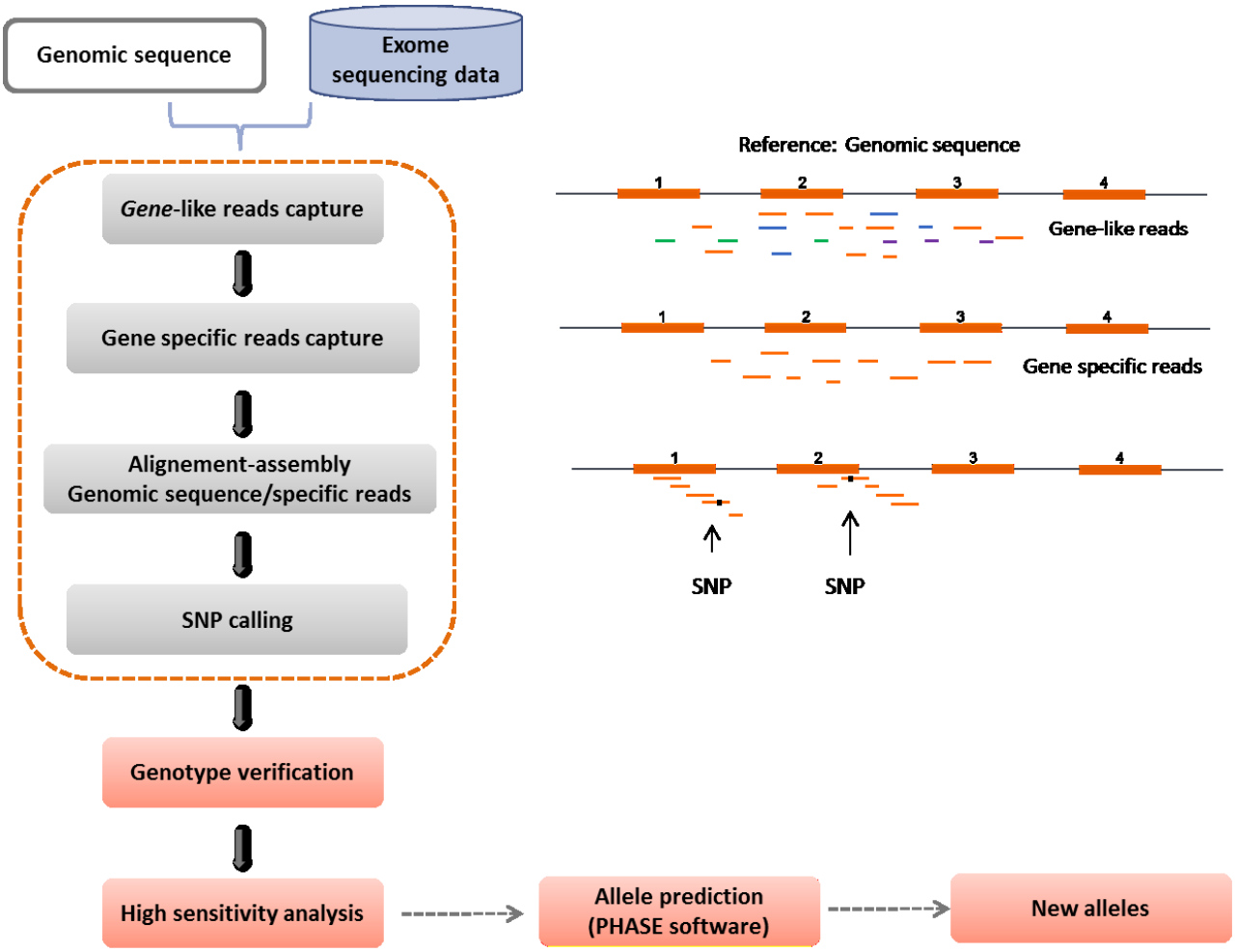


Figure 1. Specialized *in silico* approach for the large scale polymorphism analysis.

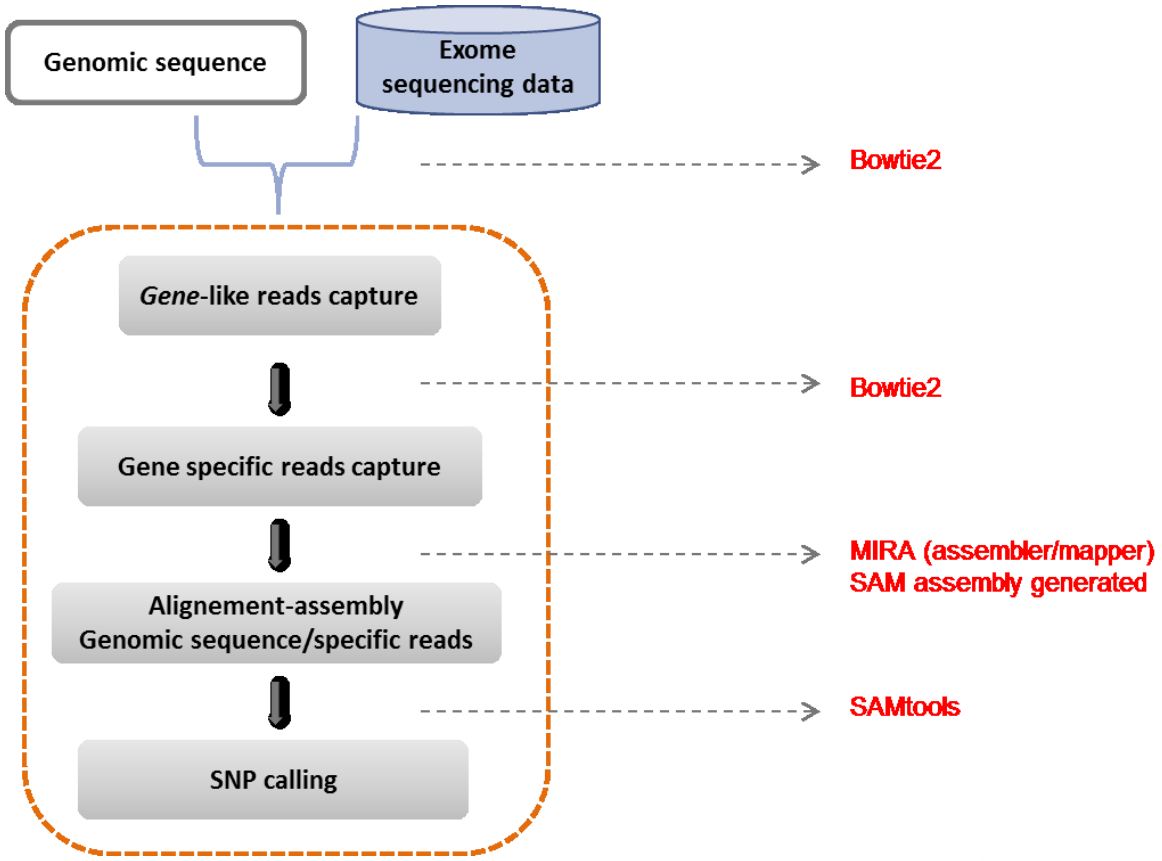


Figure 2. softwares used in this specialized *in silico* approach.

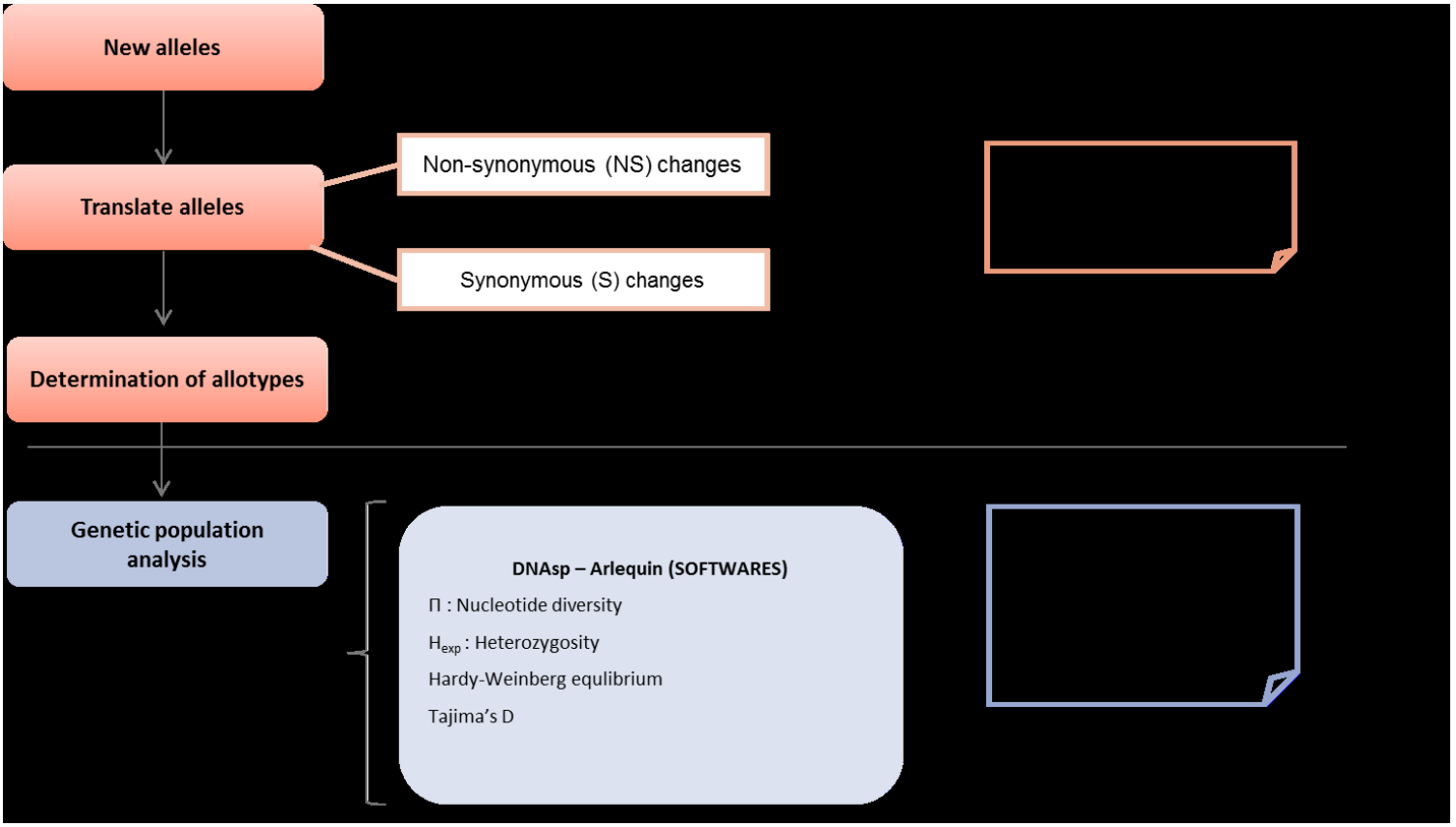


Figure 3. Determining the alleles, the allotypes and their frequencies as well as diversity using specifc softwares.

3- Data sources

1000 Genomes Project

The largest public catalogue of human variation and genotype data was created. The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the studied populations. The project took advantage of developments in sequencing technology, which sharply reduced the cost of sequencing. It was the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation. Data from the 1000 Genomes Project was quickly made available to the worldwide scientific community through freely accessible public databases. The final data set contains data for 2,504 individuals from 26 populations. Low coverage and exome sequence data are present for all of these individuals[71].

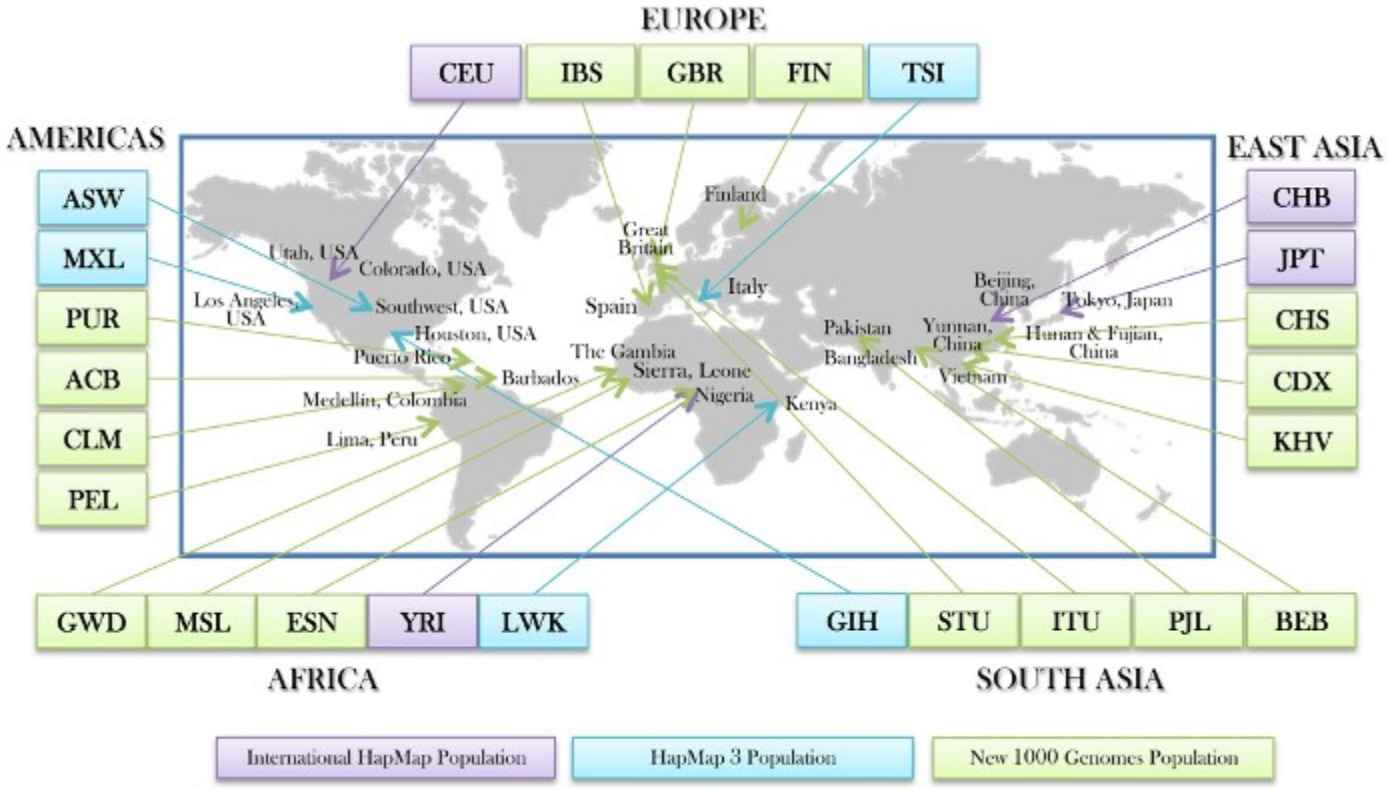


Figure 4. 26 worldwide populations from the 1000 Genomes Project (list of abbreviations for these populations is available in Annex).

Concerning chapter 3 and 4,

The following analyses in these 2 chapters were used to develop, ameliorate and validate our methodology to assess polymorphisms in coding regions.

These pilot studies were used to setup a large scale analysis in terms of number of genes to analyze as well as the number of individuals.

Chapter 3: Investigating the correlation between the SNP CT60 and coding variations in the *CD28/CTLA4/ICOS* gene region

Title: Investigating the correlation between the SNP CT60 and coding variations in the *CD28*/*CTLA4*/*ICOS* gene region

ABOU-KHATER Charbel [1, 2], OLIVE Daniel [1] , ABI-RACHED Laurent [2]

[1]INSERM, Unité 1068, CRCM, Marseille, France
[2]URMITE UM63 CNRS 7278 IRD 198, INSERM U1095, Marseille, France

CD28, CTLA4 (cytotoxic T-lymphocyte–associated antigen 4, also known as CD152), and ICOS (inducible costimulatory) belong to the CD28 family of receptors having an important role in controlling the adaptive immune response. These 3 genes are clustered on the same chromosome (2q33) in close proximity. CD28, CTLA4 and ICOS form covalent homodimers that interact with other molecules.

CD28 is constitutively expressed on the surface of T-cells, and is also present on plasma cells and mature B-cells. These costimulatory receptors are involved in T-cell activation and proliferation, via the interaction with their natural ligands B7-1 (or CD80, inducibly expressed after activation) and B7-2 (or CD86, constitutively expressed) on antigen presenting cells (APCs). T-cell responses, initiated by TCR-MHC signals, are augmented and sustained by the CD28-B7 co-stimulation. This association also provides signals for the production of various interleukins such as IL-2 and IL-6, and promotes T-cell survival.

The inhibitory receptor CTLA-4 transiently expressed and rapidly upregulated following T-cell activation acts as a major negative regulator of T-cell responses. The CTLA-4 signaling pathway is important to maintain an immunologic “homeostasis”. Both molecules CTLA-4 and CD28 bind CD80 and CD86, because they both have the same essential binding motif (MYPPPY) (see supplementary: Figure 1-2 and Table 1). CTLA4 having greater affinity and avidity than CD28, counterbalances by competing with CD28. The signal delivered via CTLA-4 when bound to B7-1 or B7-2 on the surface of APCs, dampens the response by down-regulating T cell functions and inhibiting excessive expansion of these activated cells. Thus, IL-2 and IL-2R productions are reduced, T-cell responses terminated (arrested at G1-phase of cell cycle).

Inducible T-cell costimulatory or ICOS, expressed on activated T-cells after T-cell engagement, is another co-stimulatory molecule that only binds ICOSL (FDPPPF motif of ICOS). The MYPPPY motif is not conserved in ICOS, thus the later does not bind to B7-1/B7-2. The expression of ICOS could be influenced by CD28 signals. ICOS provides some positive co-stimulatory signals, enhances basic T-cell responses, regulates cytokine secretion patterns and plays an important role in cell-cell signaling[72].

I- Investigating the diversity of these immunomodulatory genes in a set of individuals representing different populations

Polymorphisms in this stretch of 300 kb in the human costimulatory locus (on chromosome 2 (2q33) which contains the T-lymphocyte regulatory genes *CD28, CTLA4 and ICOS*) have been associated with autoimmune susceptibility because of the role of these genes in regulating T cell activation. Several autoimmune diseases were reported to have a certain association with a certain variation, such as type-I diabetes, celiac disease, rheumatoid arthritis, Grave's disease and other autoimmune thyroid disorders, Systemic lupus erythematosus etc. A big number of studies focus however on *CTLA4* polymorphisms investigating SNPs in the promoter region, coding region, 3'-UTR and 3' flanking region.

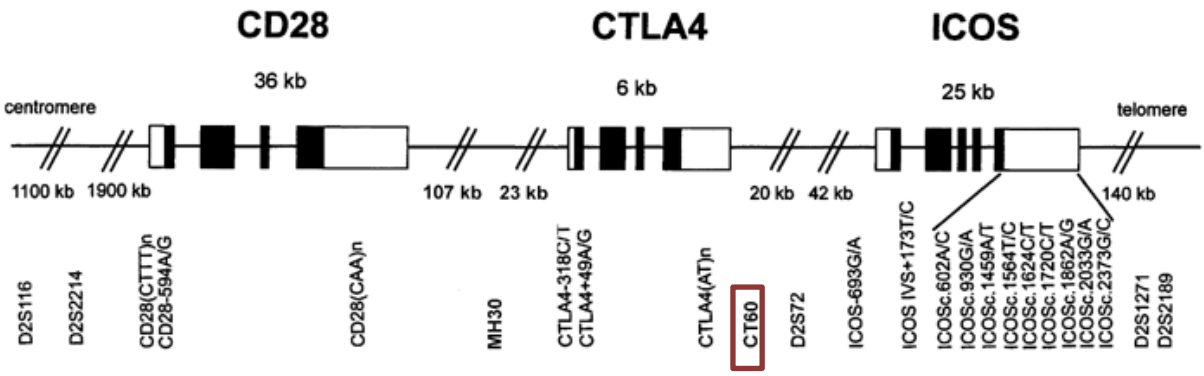


Figure 1. *CD28*, *CTLA4* and *ICOS* genes on human chromosome 2 and some known markers (from[73]). The known SNP called CT60 is marked in red.

CTLA4 is a primary determinant of susceptibility to immune diseases. SNP CT60 (rs3087243) is an already studied SNP in the 3' flanking region of the gene. It is not only associated with autoimmune diseases but also with cancer (AML, renal cell cancer…). For example, CT60 also known as +6230G>A is associated with type 1 diabetes as well as autoimmune thyroid disease and that it is also associated with variation in *CTLA4* gene splicing (study with a large number of Caucasian samples[74]).

sCTLA4 is the soluble form of CTLA4, encoded by exon 1 (leader peptide), exon 2 (ligand-binding domain), and exon 4 (cytoplasmic tail) and which lacks exon 3 (transmembrane domain) (see supplementary figure 2). The level of sCTLA4 mRNA for CT60 AA genotype is higher than in AG and GG genotypes.

Investigating the correlation between the SNP CT60 and coding variations in the *CD28/CTLA4/ICOS* gene region

The main objective of this work is to find if there is any correlation between SNPs found in the coding regions of these genes (*CD28/CTLA4/ICOS*) and the known SNP CT60 (rs3087243). This variation (figure 1) is located in the *CTLA4-ICOS* intergenic region (3' flanking region of the *CTLA4* gene), not covered by Exome sequencing technologies. That's why we searched for available published genotypes for the known SNP CT60 in the HapMap project (a public resource that helps in finding genes associated with human disease). CT60 genotypes were available for 566 individuals out of 1683 of our 1000 Genomes panel.

II- Investigating the coding diversity of 3 immunomodulatory genes (*CD28, CTLA4, ICOS*) in a set of individuals representing different populations

1- Material and methods

To define the diversity in the coding regions of these 3 genes, we used EXOME data from the 1000 Genomes Project to assess the variation worldwide. We analyzed sequencing data generated by the 1000 Genomes project for 566 individuals (healthy) representing 8 populations (Table 2), using the pipeline and strategy explained in Chapter 2 for polymorphism detection and allele reconstruction.

	Population	Individuals
Americas	ASW	56
Europe	CEU	59
	TSI	57
East Asia	CHB	72
	JPT	66
South Asia	GIH	100
Africa	**LWK**	**62**
	YRI	**94**
	Total	566

Table 1. Our data panel from the 1000 Genomes Project. Individuals from 8 different worldwide populations having genotypes for the SNP CT60 were selected.

2- Results

In our data of 566 individuals from 8 different worldwide populations (1000 Genomes project), we were able here to identify 4 SNPs affecting the amino acid sequence in each of the coding region of the human *CD28* and *ICOS.* Other novel SNPs could be identified after analyzing or sequencing more genomes. These nucleotides changes served to define 5 allotypes (forms of the protein) for each of these 2 genes (Tables 2-5).

The percentage of homozygous individuals for *CD28* varies between 89 to 98.6% and for *ICOS* between 74.5 to 100%. (Supplementary table 2)

		CD28*01	CD28*02	CD28*03	CD28*04	CD28*05	
Americas	ASW	97.32	0.89	0	0	1.79	100
Europe	CEU	100	0	0	0	0	100
	TSI	100	0	0	0	0	100
East Asia	CHB	99.31	0	0	0.69	0	100
	JPT	99.24	0	0	0.76	0	100
South Asia	GIH	100.0	0	0	0	0	100
Africa	LWK	97.58	0	2.42	0	0	100
	YRI	98.94	1.06	0	0	0	100
	ALL	99.12	0.27	0.27	0.18	0.18	100

Table 2. Worldwide distribution of *CD28* allotype frequencies (in percentage).

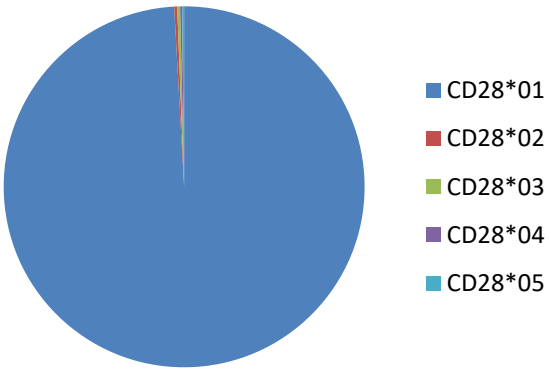


Figure 2. *CD28* allotype frequencies (in percentage).

	Aminoacids			
	8	16	91	198
	signal peptide		extracellular	cytoplasmic
	L	V	G	R
CD28*01	.	.	.	.
CD28*02	.	.	D	.
CD28*03	.	I	.	.
CD28*04	.	.	.	C
CD28*05	F	.	.	.

Table 3. Non-synonymous changes according in each allotype. Only non-synonymous variations are represented

		ICOS*01	ICOS*02	ICOS*03	ICOS*04	ICOS*05
Americas	ASW	98.21	0	0	0	1.79
Europe	CEU	100	0	0	0	0
	TSI	100	0	0	0	0
East Asia	CHB	100	0	0	0	0
	JPT	99.24	0	0	0.76	0
South Asia	GIH	100	0	0	0	0
Africa	LWK	100	0	0	0	0
	YRI	95.74	0.53	1.06	0	2.66
	ALL	99.03	0.09	0.18	0.09	0.62

Table 4. Worldwide distribution of *ICOS* allotype frequencies (in percentage).

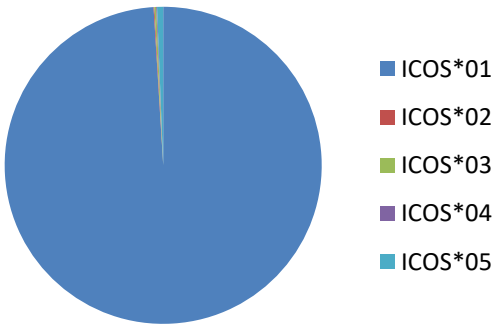


Figure 3. *ICOS* allotype frequencies (in percentage)

	Aminoacids			
	14	61	151	197
	signal peptide	extracellular	transmembrane	cytoplasmic
	R	I	V	V
ICOS*01	.	.	.	.
ICOS*02	.	.	.	M
ICOS*03	.	.	L	.
ICOS*04	.	K	.	.
ICOS*05	C	.	.	.

Table 5. Non-synonymous changes according in each allotype. Only non-synonymous variations are represented

CTLA4 has two common allotypes CTLA4*01 and CTLA4*02 (and one rare allotype CTLA4*03, with 0.18% of frequency in our data). The percentage of homozygous individuals for *CTLA4* varies between 45.76 to 57%. (Supplementary table 2)

		CTLA4*01	CTLA4*02	CTLA4*03	
Americas	ASW	66.07	33.93	0	100
Europe	CEU	61.02	38.98	0	100
	TSI	64.91	35.09	0	100
East Asia	CHB	33.33	66.67	0	100
	JPT	35.61	64.39	0	100
South Asia	GIH	69.50	30.50	0	100
Africa	LWK	53.23	45.16	1.61	100
	YRI	65.43	34.57	0	100
	ALL	56.80	43.02	0.18	100

Table 6. Worldwide distribution of *CTLA4* allotype frequencies (in percentage).

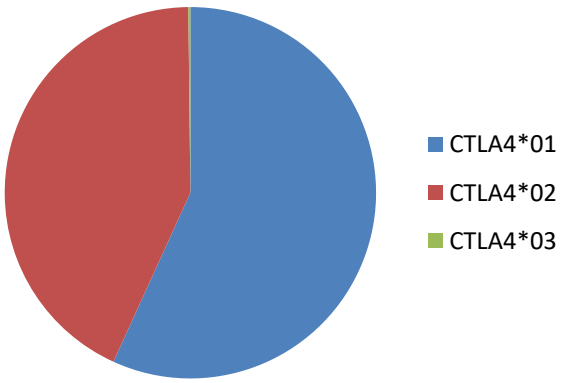


Figure 4. *CTLA4* allotype frequencies (in percentage).

	Aminoacids	
	8	17
	leader	leader
	R	T
CTLA4*01	.	.
CTLA4*02	.	A
CTLA4*03	Q	A

Table 7. Non-synonymous changes according in each allotype.
Only non-synonymous variations are represented

We were able here to identify 2 SNPs affecting the amino acid sequence in the signal peptide of the human *CTLA4* (R8Q and T17A). Other novel SNPs could be identified after sequencing more genomes. These nucleotides changes served to define 3 allotypes.

The T17A variation is known (rs231775) and is associated for example with for Graves' disease (increased risk)[75], increases susceptibility to hepatitis B-related hepatocellular carcinoma in some populations[76] etc. (T17A is a non-synonymous A/G Thr/Ala −17 change at position +49 (+49AG) in the CTLA4 leader peptide)

In our data, the CTLA4*01 is less frequent in East Asia compared to the other regions (*p<<0.05* t-student test).

CD28 and *ICOS* are highly conserved genes. As for *CTLA4*, having two allotypes with intermediary frequencies could be indicative of balancing selection. Tajima's D test is done to verify that this is not due to genetic drift (change in the frequency of a gene variant in a population due to chance or random sampling of individuals). Our first results indicate that balancing selection maintains this polymorphism in the leader (p=0.001) (Supplementary figure 3 and Annex for description).

<u>Changes at the protein level</u>

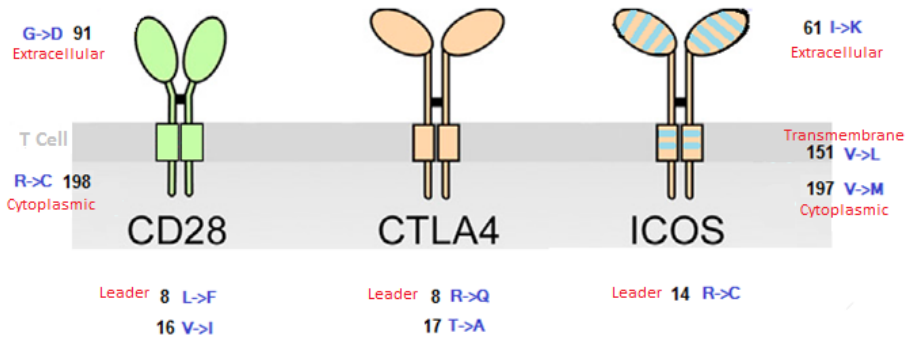


Figure 5. Schematic representation of the non-synonymous changes and localization in each of these three proteins

III- Correlation between SNPs found in the coding regions of these genes and the SNP CT60

After reconstructing the haplotypes in this region, the different combinations regrouping coding variations in *CD28, CTLA4* and *ICOS* are represented in tables 8 and 9. Our results didn't show any significant correlation. Extended analysis regrouping more individuals representing all continents and populations as well as their CT60 genotypes is thus required.

We tried to check if the CT60 SNP affects the expression of these genes. To answer this question, we used RNA-sequencing data for individuals of the 1000 Genomes project. But unfortunately, CD28, CTLA4, ICOS are not expressed in the used cell lines (EBV transformed B cells). In other RNA-sequencing data, no data were available for the SNP CT60 (3' region of *CTLA4*).

Our aim was to check whether there is any correlation between coding SNPs and CT60 variation by defining the haplotype-block structure. It would be interesting to analyze linkage disequilibrium between variants at this locus (linkage disequilibrium means simply a nonrandom association of alleles at two or more loci).

Based on these preliminary results, more analyses with extended data are needed to further dissect the role of these variants or their combinations across the costimulatory locus. This type of analyses shed the light on a central idea: functional variations in a certain gene or region may not be due to polymorphism in a single position or a single gene but to the combination of genetic variants to form alleles (or haplotypes for a certain locus or region).

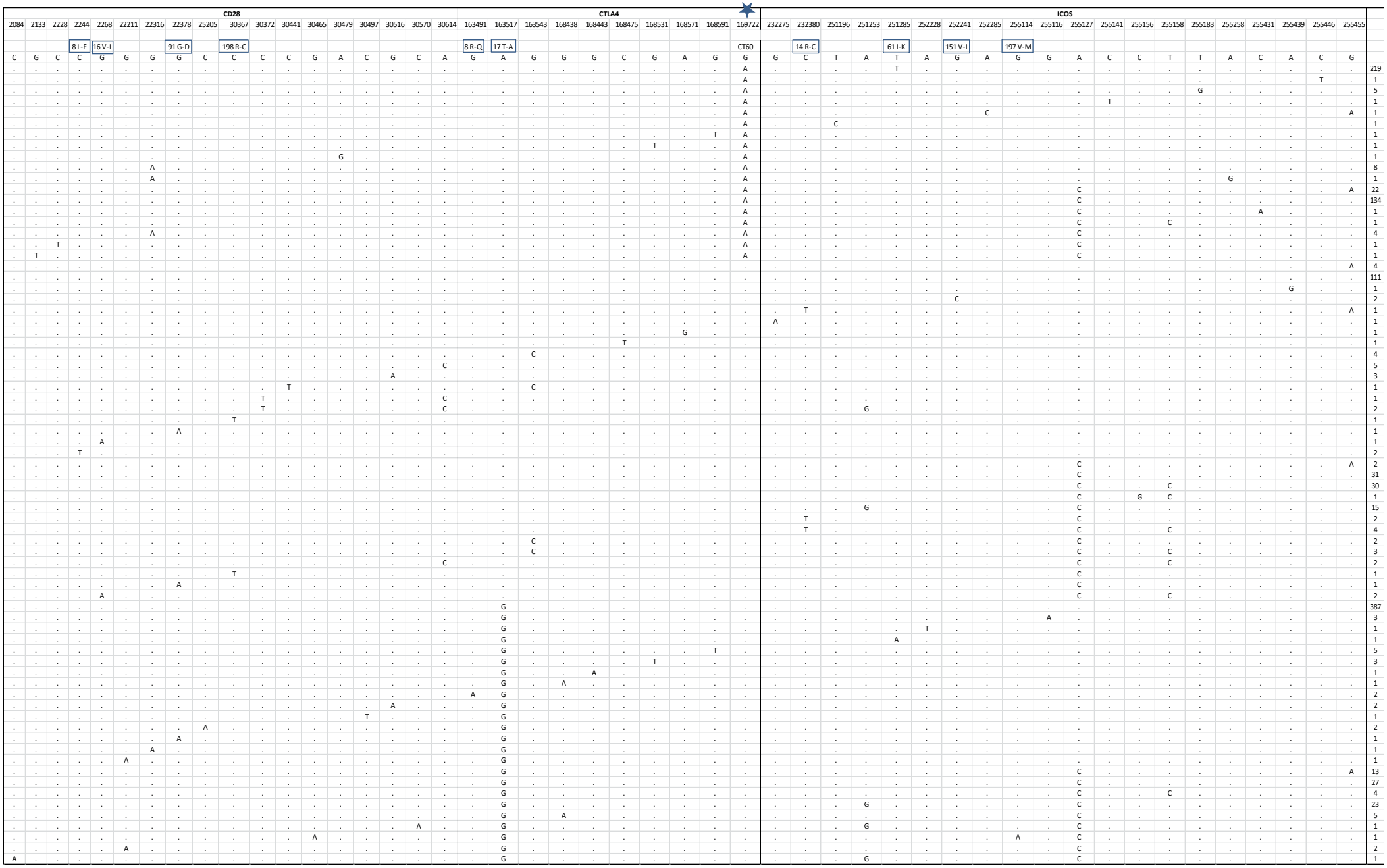


Table 8. Different haplotypes regrouping coding variations in *CD28, CTLA4* and *ICOS*. The last column indicates how many times each haplotype occurred. Positions in rectangles indicate the non-synonymous variations and the CT60 SNP is marked by a star.

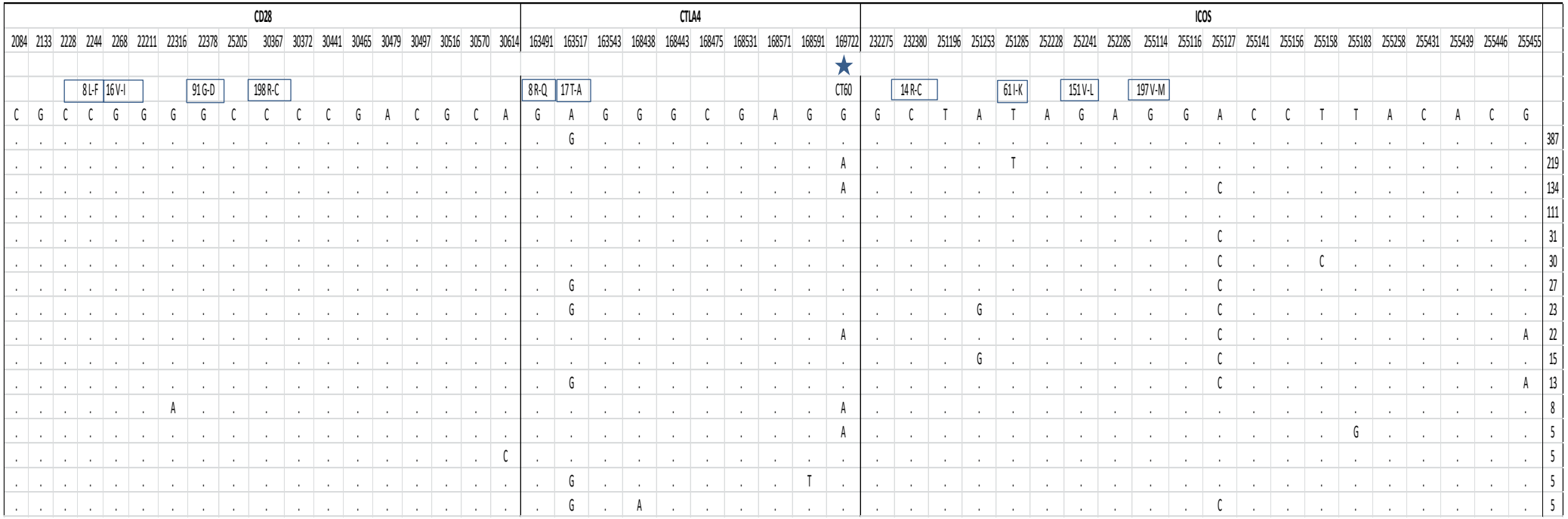


Table 9. Different haplotypes regrouping coding variations in *CD28, CTLA4* and *ICOS*. Only haplotypes that occurred more or equal than 5 times (≥5) are represented. Positions in rectangles indicate the non-synonymous variations and the CT60 SNP is marked by a star.

IV- Annex

Population genetics analyses (supplementary figure 3 and supplementary table 2)

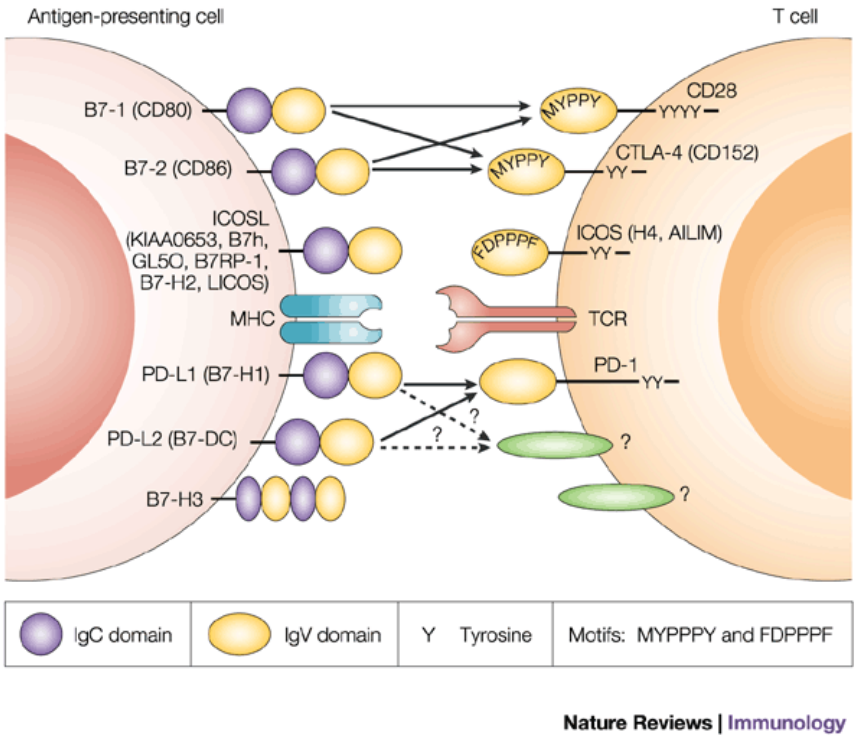
Hardy-Weinberg equilibrium (HWE) for each population was estimated from genotypes using Arlequin version 3.5[77]. For each population, the number of segregating sites (S) and the nucleotide diversity (π) of each population were calculated using DnaSP version 5.10.1[78].

Analysis of Tajima's D

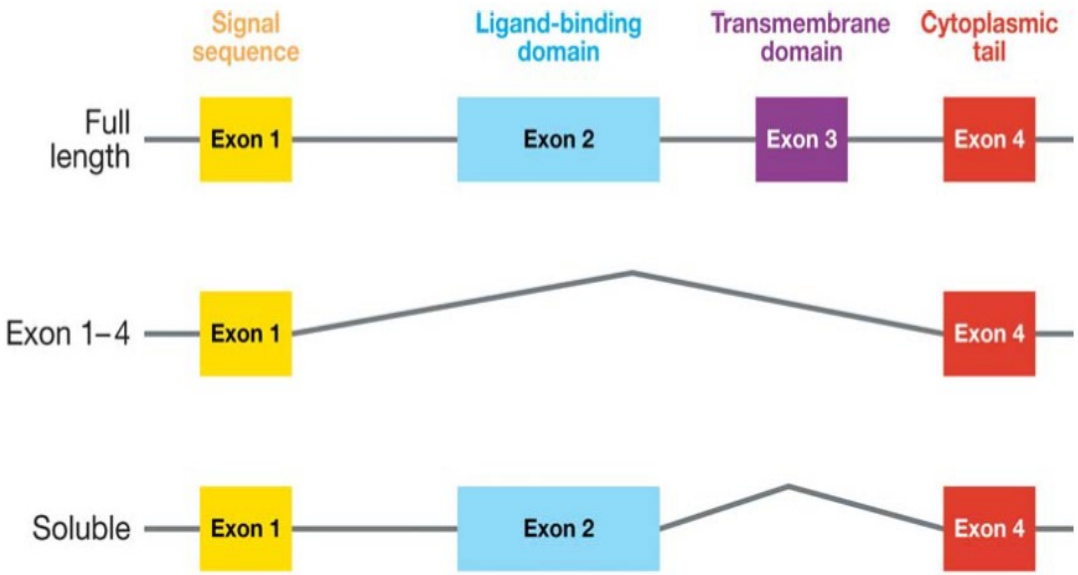
Tajima's D statistics[79] was calculated using DnaSP version 5.10.1. Statistical significance for Tajima's D was assessed by comparing the observed Tajima's D values to those expected under natural-drift equilibrium, which were obtained by coalescent simulations using the programs[80]. For these simulations and for each population the following parameters were specified: the number of chromosomes, the number of segregating sites, the recombination rate and the demographic model. Recombination rates were estimated using PHASE2.1[68,69]. Demographic models used were ancient expansion for African populations[81] and severe bottleneck for European and Asian populations[82], as previously described (Model 2 and Model 4,respectively)[83]. Each model was performed in 100,000 replicates. Sliding window Tajima's D values were calculated using DnaSP version 5.10.1 with a window length of 100 nucleotides and a step size of 3 nucleotides. Coalescent simulations were performed for each window according to the number of segregating sites present in the window with the recombination rate of the entire sequence for each population. Demographic models were used as described previously.

	Structure	Subcellular location	Intracellular domain	Ligand (+structure)	Expressed on (among others)	Motifs in IgV	Other names	Dimerization	Main function
CD28	Ig-like V-type domain	Cell membrane	4 tyrosine residues in the cytoplasmic domain	B7.1 B7.2 (IgC+IgV)	T-Cell	MYPPPY	Tp44	Non-covalent homodimers	Involved in T-cell activation
CTLA4	Ig-like V-type domain	Cell membrane +secreted and soluble form	2 tyrosine residues in the cytoplasmic domain	B7.1 B7.2 (IgC+IgV)	T-Cell	MYPPPY	CD152, CELIAC3, IDDM12	Non-covalent homodimers	downregulation of the immune response
ICOS	Ig-like V-type domain	Cell membrane	2 tyrosine residues in the cytoplasmic domain	ICOSL (IgC+IgV)	T-Cell	FDPPPF	AILIM, CD278	Non-covalent homodimers	cell-cell signaling, immune responses, and regulation of cell proliferation

Supplementary Table 1. Main characteristics of the CD28, CTLA4 and ICOS proteins. (source Uniprot and [72])

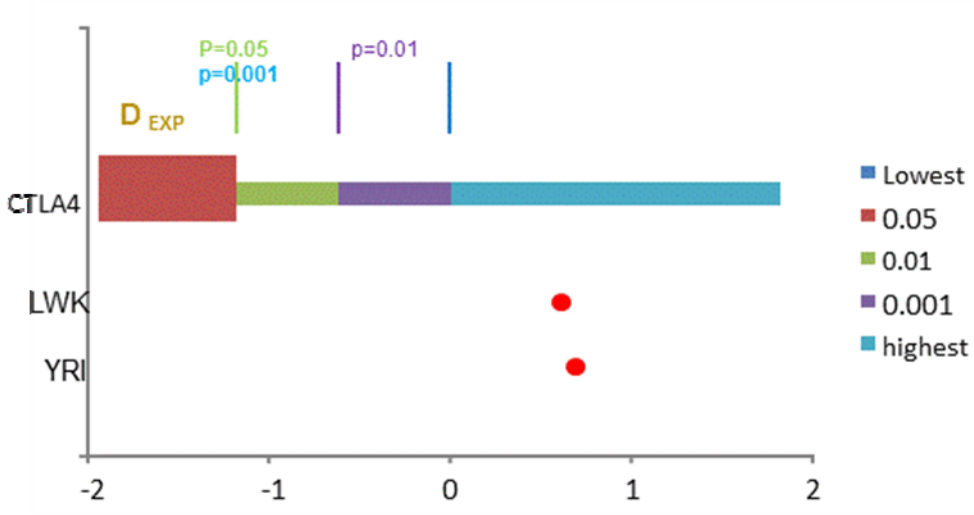


Supplementary Figure 1. Structures of CD28, CTLA4 and B7 family members. (from[72])

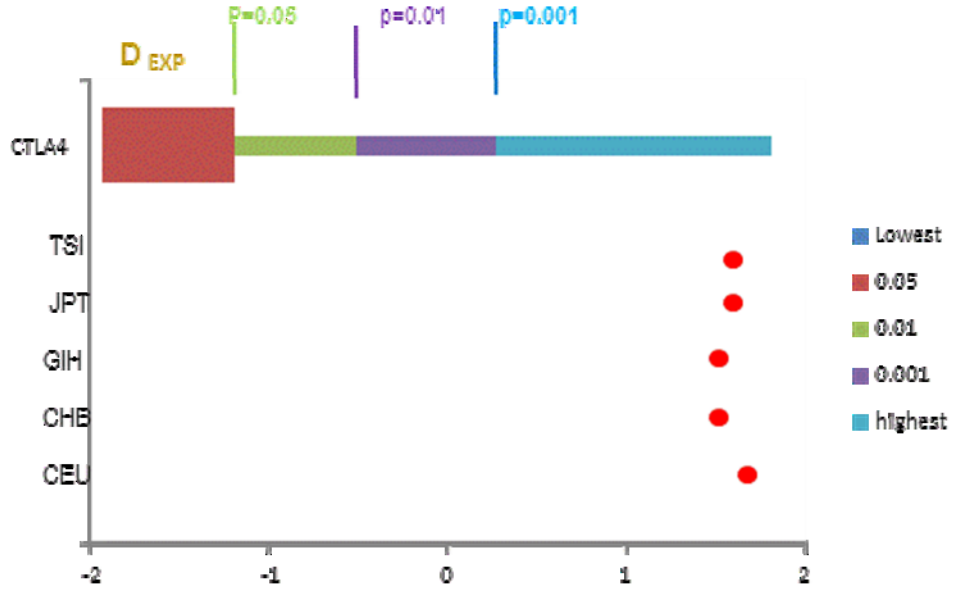


Supplementary Figure 2. *CTLA4* and its various splice variants. (from[84])

Model 2



Model 4



Supplementary Figure 3. Tajima's D statistics

		No. of Hetero	No. of Homo	Obs.Heter.	Exp.Heter.	Tajima's D	π
CD28							
	ASW	5	51	0.0893	0.0875	-1.71617	0.13
	CEU	3	56	0.0508	0.05	-0.81353	0.08
	CHB	1	71	0.0139	0.0139	-0.98747	0.02
	GIH	11	89	0.11	0.1228	-0.87073	0.19
	JPT	3	63	0.0455	0.045	-1.2813	0.07
	LWK	4	58	0.0645	0.0633	-1.22382	0.1
	TSI	1	56	0.0175	0.0175	-1.01327	0.03
	YRI	3	91	0.0319	0.0317	-1.2512	0.05
CTLA4							
	ASW	30	26	0.5357	0.4755	0.4372	0.73
	CEU	32	27	0.5424	0.4798	1.73582	0.71
	CHB	32	40	0.4444	0.4476	1.58447	0.67
	GIH	43	57	0.43	0.4261	1.51969	0.63
	JPT	33	33	0.5	0.4621	1.65329	0.69
	LWK	27	35	0.4355	0.5167	0.65016	0.79
	TSI	30	27	0.5263	0.4596	1.60874	0.68
	YRI	47	47	0.5	0.5072	0.77135	0.08
ICOS							
	ASW	9	47	0.1607	0.1514	-0.89338	0.26
	CEU	0	59	0	0	0	0
	CHB	2	70	0.0278	0.0276	-0.90625	0.05
	GIH	1	99	0.01	0.01	-0.95211	0.02
	JPT	2	64	0.0303	0.0302	-1.34079	0.05
	LWK	13	49	0.2097	0.2616	0.45189	0.44
	TSI	2	55	0.0351	0.0349	-1.36058	0.06
	YRI	24	70	0.2553	0.2322	-1.16796	0.04

Supplementary Table 2 resuming some genetic population analyses

Chapter 4: Polymorphism and functional analysis of some immune genes controlling infection by *Mycobacterium tuberculosis*

Polymorphism and functional analysis of some immune genes controlling infection by *Mycobacterium tuberculosis*

ABOU-KHATER Charbel [1, 2], OLIVE Daniel [1], ABI-RACHED Laurent [2]

[1]INSERM, Unité 1068, CRCM, Marseille, France
[2]URMITE UM63 CNRS 7278 IRD 198, INSERM U1095, Marseille, France

Tuberculosis (TB) was widely believed to be an inherited disease before the discovery of the tubercle bacillus. In 1882, Robert Koch identified the causative agent of the disease to be the slow-growing *Mycobacterium tuberculosis*. Since then, researchers gave more attention to the importance of the pathogen. Interestingly, and by observing that individuals respond differently to that particular infection, they started giving more consideration to the genetic component in such a complex disease.

I- Introduction

Tuberculosis, an ancient disease of mankind, is an ongoing major global cause of death and disease. It causes ill-health among millions of people each year and ranks as the second leading cause of death from an infectious disease worldwide after HIV. The latest estimates included in the World Health Organization (WHO) report are that there were, in 2015, 10.4 million new TB cases (including 1.2 million among HIV-positive people), 1.4 million TB deaths, and an additional 0.4 million deaths resulting from TB disease among HIV-positive people[85]. In France, the total number of TB cases is 4845, according to the WHO's TB surveillance and monitoring in Europe (Annex 3).

Geopolitical events affect also the incidence of TB (and infectious diseases in general). For example, the trend of TB incidence in Lebanon had been declining till the year 2011. In 2012, 48% of all notified cases were among non-Lebanese. A 27% increase in TB incidence in 2012 is noticed. This dramatic increase of TB patients is due to the influx of Syrian refugees, affecting the overall number of reported cases. 681, 664 and 679 cases were reported respectively in 2014, 2015 and 2016. In 2014, 345 non-Lebanese patients including 109 Syrian were reported, while 2015 figures, registered 354 non-Lebanese including 139 Syrian. These numbers continued to increase in 2016 to reach a number of 358 among non-Lebanese TB patients, including 147 Syrian[86] (Annex 3).

One of the striking features of the disease is that in 90% of infected immunocompetent individuals, the infection does not give rise to disease, while only an estimated 10% infected by the pathogen will develop clinical signs and illness. One pertinent question to this well-established epidemiological observation is why the majority of humans are naturally resistant to tuberculosis? In other terms, what are the factors in this host-pathogen interplay that dictate whether the infection progresses to overt disease or not?

It is known that genetic and non-genetic factors (such as environmental, socio-economic etc.) of both the bacterium and the host have impact on the host response and may explain why some people resist infection more successfully than others. Until now, little is known about the interaction of these factors and the resulting impact on disease development.

Some genetic factors that modulate resistance are known and some works suggest that host susceptibility genes or loci exist. It also depends on the individual's clinical and exposure history.

Based on available data in the literature, we selected a set of 10 genes. Some of them are relevant for the immune response against *M. tuberculosis* and had been previously associated with susceptibility to TB.

II- Materials and methods

	Population	Individuals
South Asia	BEB	60
	PJL	51
	STU	63
East Asia	CDX	78
	CHB	72
Africa	ESN	47
	GWD	88
	LWK	62
	YRI	94
Europe	CEU	59
	TSI	57
Americas	CLM	57
	PUR	64
Total	13	852

Table 1. Data panel from the 1000 Genomes Project. 852 healthy individuals from 13 different worldwide populations

Here we present our catalogue results concerning the polymorphism of these ten genes. For accurate and reliable results, alleles that occurred less than 5 times were not considered in the following analysis (represented by "rare alleles" in the following tables). We used the same pipeline for polymorphism detection and allele reconstruction described in Chapter 2.

III- RESULTS

1- TLR2 and Vitamin D pathway

VDR or Vitamin D3 receptor is a nuclear hormone receptor. This transcription factor mediates the action of vitamin D3 by controlling the expression of some genes such as cathelicidin. VDR plays also an essential role in calcium homeostasis.

Role of Vitamin D receptor in innate immunity: fighting infections with Vitamin D

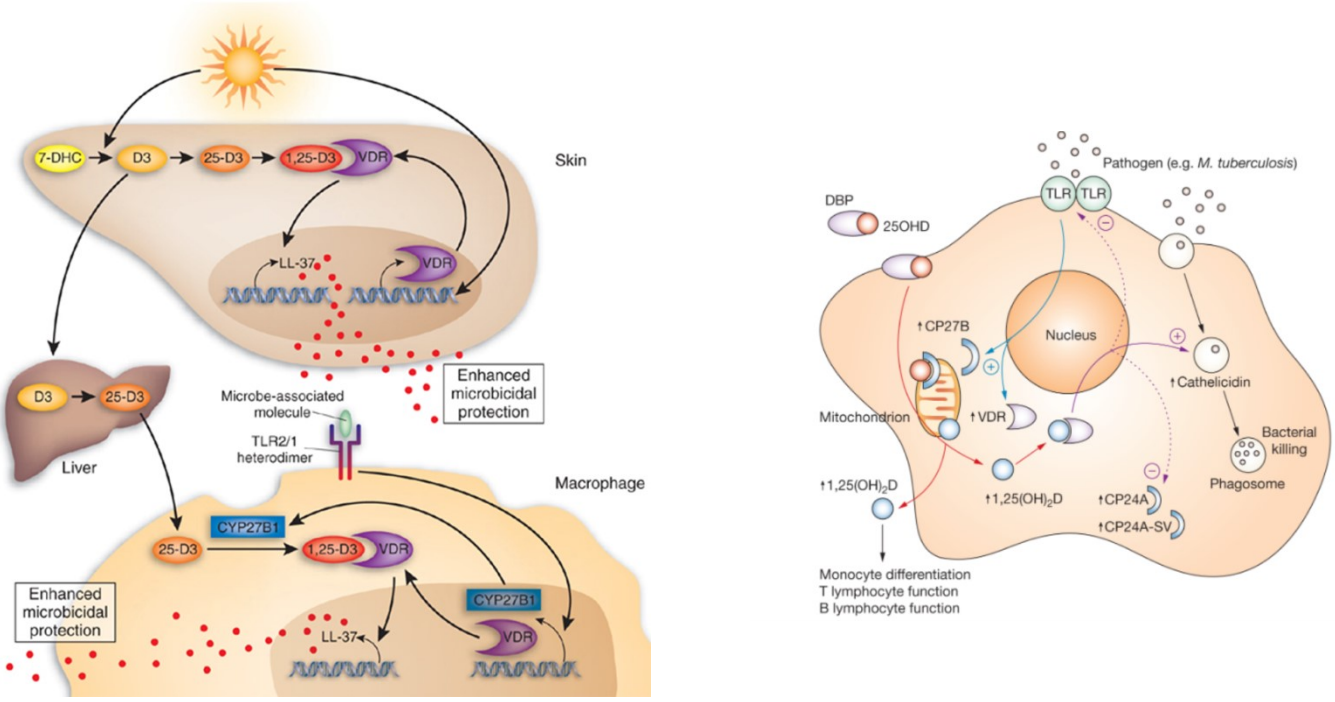


Figure 1. Fighting infections with Vitamin D, important immunoregulatory hormone (from[87] and [88])

The macrophage TLR (i.e. TLR1/2 heterodimer) signaling pathway is activated by pathogens like *M. tuberculosis*: circulating monocytes are activated by TLR2/1 which senses molecules derived from pathogens (the so-called pathogen-activated molecular patterns (PAMPs) present on bacterial surfaces). The genes encoding VDR and CYP27B1 are induced. Circulating 25OHD (25- hydroxyvitamin D) bound to plasma DBP (vitamin-D-binding protein), enters macrophages and is converted to $1,25(OH)_2D$ (1,25-dihydroxyvitamin D) by mitochondrial CYP27B1. $1,25(OH)_2D$ can bind to the VDR in the cell, and thus is able to act as a transcriptional factor leading to the induction of LL-37 expression (cathelicidin). This antibacterial agent functions once it is incorporated into phagosomes that contain internalized pathogens, leading to an enhanced microbicidal activity of the phagocyte (by permeabilizing the cell wall and membrane of bacteria)[89].

Polymorphisms in the *VDR* gene can influence the function of this immunoregulatory hormone and thus contribute to the susceptibility to TB infection.

Investigating the coding diversity of *VDR* and *TLR2* in a set of individuals representing different populations

VDR, Vitamin D3 receptor

		VDR*01	VDR*02	rare alleles
South Asia	BEB	65.00	33.33	
	PJL	79.41	19.61	
	STU	81.75	18.25	
East Asia	CHB	54.86	42.36	
	CDX	45.51	52.56	
Africa	ESN	78.72	18.09	
	GWD	72.73	23.30	
	LWK	80.65	16.13	
	YRI	80.32	17.02	
Americas	PUR	61.72	38.28	
	CLM	58.77	41.23	
Europe	CEU	58.47	41.53	
	TSI	65.79	34.21	
	ALL	67.78	30.52	1.70

Table 2. Worldwide distribution of *VDR* allotype frequencies (in percentage).

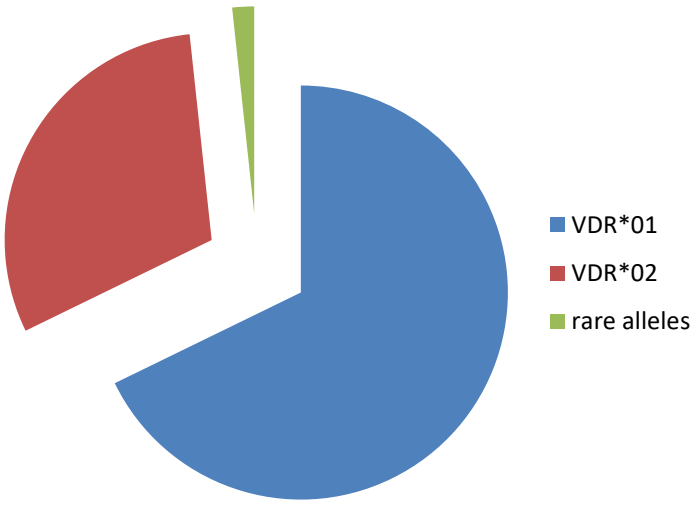


Figure 2. *VDR* allotype frequencies (in percentage).

Amino acids		
	1	
VDR*01	Thr	1155
VDR*02	Met	520

VDR*01	hydrophilic
VDR*02	hydrophobic

Table 3. Non-synonymous changes according in each allotype (*VDR* gene).
For changes at the protein level, only non-synonymous variations are represented.

The polymorphism that we found is the only known protein polymorphism in the *VDR* gene and it was previously described. T to C polymorphism exists (ATG to ACG) and affects the start codon (translation initiation start site ATG). Also known as FokI polymorphism (defined using the FokI restriction enzyme in an RFLP test). Two protein variants can exist:

- Shorter VDR: the C-allele (ACG) or ‘‘F’’ allele: the protein is shortened by three amino acids. Referred to as the M4 form because the Methionine (starting site) is at the fourth position.
  In our data, the shorter form is more frequent than the other one (67.78%)
- Long version of the VDR: the T-allele (ATG) or the ‘‘f’’ allele: referred to as the M1 form because the Methionine is at the first position.

This variation affects the immune cell behavior, with more active immune system for the short F-VDR.[90] Polymorphisms in the *VDR* gene can influence the function of vitamin D and thus contribute to the susceptibility to TB infection.

An increasing risk of TB in HIV-negative individuals, is found (especially in the Asian region with 1.82-fold increased risk): ff genotype individuals have about 1.60-fold increased susceptibility compared with FF or fF genotype and in the Asian group[91].

In our data, the VDR*01 allotype that corresponds to the shorter form of the protein is more frequent in South Asia and Africa compared to the other regions (*p=0.0003* t-student test).

TLR2, Toll-like receptor 2

		TLR2*01	TLR2*02	TLR2*03	TLR2*04	TLR2*05	rare alleles
South Asia	BEB	99.17	0	0	0	0	
	PJL	98.04	0	0.98	0	0	
	STU	97.62	0	0	0	0	
Americas	PUR	91.41	3.91	0.78	0	0	
	CLM	94.74	0.88	1.75	0	0	
East Asia	CDX	96.15	0	0	0	0	
	CHB	97.92	0.69	0	0	0	
Europe	TSI	92.98	1.75	3.51	0	0	
	CEU	93.22	2.54	3.39	0	0	
Africa	ESN	95.74	0	0	0	0	
	GWD	94.32	0	0	0.57	1.14	
	LWK	91.13	0	0	1.61	0	
	YRI	96.28	0	0	1.06	1.60	
	ALL	95.31	0.70	0.70	0.29	0.29	2.70

Table 4. Worldwide distribution of *TLR2* allotype frequencies (in percentage).

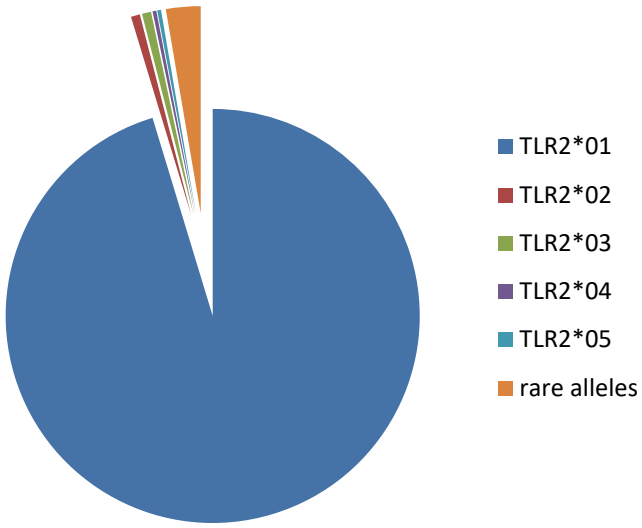


Figure 3. *TLR2* allotype frequencies (in percentage).

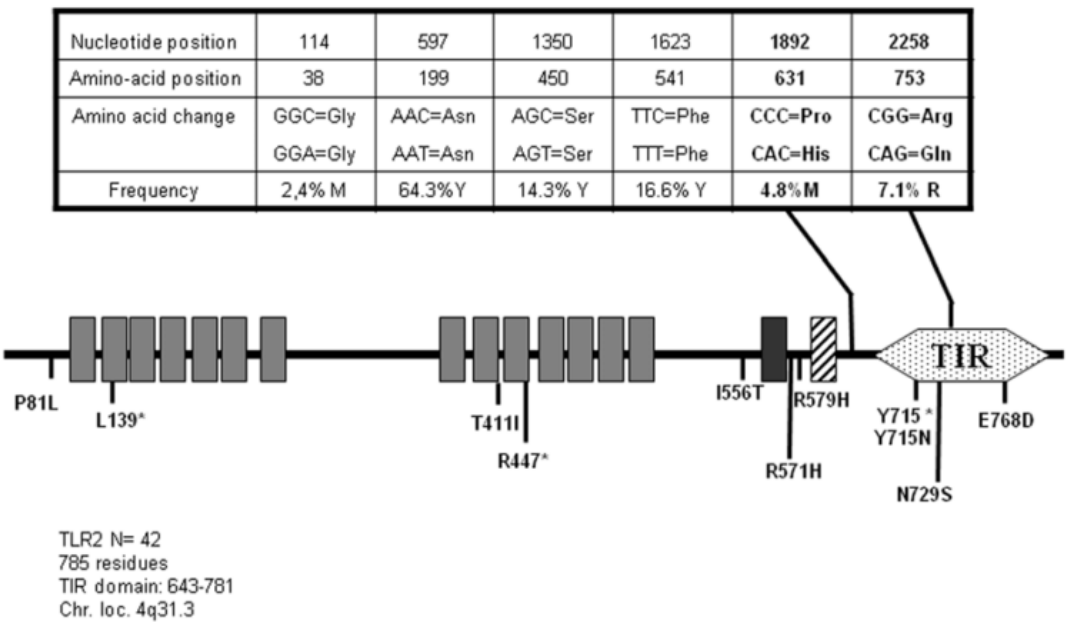
	Amino acids				
	91	571	631	753	
REF	Ile	Arg	Pro	Arg	
TLR2*01	.	.	.	.	1624
TLR2*02	.	.	.	Gln	12
TLR2*03	.	.	His	.	12
TLR2*04	.	His	.	.	5
TLR2*05	Val	.	.	.	5

domain			
Ile	Arg	Pro	Arg
Extracellular		Cytoplasmic	

change			
Ile	Arg	Pro	Arg
Val	His	His	Gln

hydrophobic	(+) charged	hydrophobic	(+) charged
hydrophobic	(+) charged	(+) charged	hydrophilic

Table 5. Non-synonymous changes according in each allotype (*TLR2)*.
For changes at the protein level, only non-synonymous variations are represented.



Schematic representation of TLR2 and position and frequency of Single Nucleotide Polymorphisms (SNPs).

Figure 4. adapted from [92] and showing the position of the two aminoacids 631 and 753.

TLR2 variants are associated to a wide panel of infectious diseases. Georgel *et al.* (2009) mentioned that currently 24 polymorphisms within the human *TLR2* gene had been described, among which 13 are non-synonymous variations[92].

In our data of 852 individuals from 13 different worldwide populations (1000 Genomes project), we were able here to identify 4 SNPs in the coding region of the human *TLR2* and affecting the amino acid sequence (2 of them are well known P681H and R753Q[92]) in addition to 7 synonymous SNPs (not represented here; the unchanged amino acid positions are the following: 199 Asn, 450 Ser, 541 Phe, 542 Leu, 563 Leu, 707 Phe, 781 Ala). Other novel SNPs could be identified after sequencing more genomes. These nucleotides changes served to define 5 allotypes (forms of the TLR2 protein) (Tables 4-5), those expressing P631H and R753Q representing each one 0.7% of our population.

The non-synonymous changes at position 91 and 571 (table 5) are called conservative mutations because the properties of the amino acid remain the same, while those at position 631 and 753 are described as non-conservative mutations.

The protein may lose its function in case of non-conservative mutations, which can result in a disease (for example Sickle Cell Anemia as mentioned in chapter 1).
The V599E mutation (changing a valine –V- to glutamic acid –E-) in the *BRAF* gene (oncogene in human cancer) is another example of missense non-conservative mutation. In this case, the protein exhibit 10-fold greater basal kinase activity than wild-type *BRAF*[93].

2- <u>BTLA (CD272) and TNFRSF14 (HVEM)</u>

BTLA is a member of the Ig super family of proteins, known also to be a receptor which inhibits lymphocytes during immune response. In its cytoplasmic tail, BTLA has an ITIM inhibitory signaling domain. When it interacts with HVEM (a TNFR super family receptor known as TNFRSF14), the tyrosine of the dual ITIMs is phosphorylated, allowing the recruitment of SHP-1 and SHP-2 (phosphatases), which attenuate signaling.

BTLA inhibits the proliferation of human Vγ9Vδ2 T cells[94]. γδ T cells may play a role in the early response to *Mycobacterium tuberculosis*. These cells are present in the alveoli, and tuberculosis generally affects the lungs. Mycobacterial phosphoantigens are recognized by γδ T cells. Once activated, they kill infected macrophages expressing these phosphoantigens on the surface, through cytotoxic granules, IFN-γ and TNF-α signaling etc[95].

BTLA, B- and T-lymphocyte attenuator

		BTLA*01	BTLA*02	BTLA*03	BTLA*04	BTLA*05	BTLA*06	BTLA*07
South Asia	BEB	80.83	17.50	1.67	0	0	0	0
	PJL	81.37	18.63	0	0	0	0	0
	STU	73.02	26.98	0	0	0	0	0
East Asia	CDX	62.82	26.28	8.33	1.28	0	1.28	0
	CHB	75.00	20.14	4.86	0	0	0	0
Americas	CLM	84.21	14.04	1.75	0	0	0	0
	PUR	78.91	15.63	4.69	0.78	0	0	0
Africa	ESN	10.64	73.40	6.38	8.51	1.06	0	0
	GWD	22.16	72.16	0.57	3.98	1.14	0	0
	LWK	9.68	78.23	8.87	2.42	0.81	0	0
	YRI	10.11	75.53	6.91	4.79	2.66	0	0
Europe	CEU	98.31	1.69	0	0	0	0	0
	TSI	94.74	4.39	0	0	0	0	0.88
	ALL	57.45	36.50	3.58	1.76	0.53	0.12	0.06

Table 6. Worldwide distribution of *BTLA* allotype frequencies (in percentage).

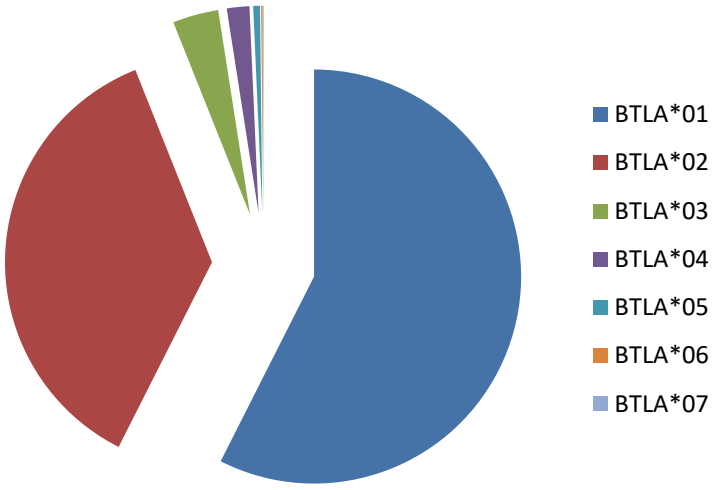


Figure 5. *BTLA* allotype frequencies (in percentage).

	Aminoacids						
REF	124	157	184	197	227	267	
	Ile	Arg	Lys	Asn	Asp	Pro	
BTLA*01	.	Ser	.	.	.	Leu	979
BTLA*02	.	Ser	.	.	.	.	622
BTLA*03	.	Ser	.	Thr	.	.	61
BTLA*04	.	.	.	.	.	.	30
BTLA*05	Val	Ser	.	.	.	.	9
BTLA*06	.	Ser	.	.	Val	Leu	2
BTLA*07	.	Ser	Glu	.	.	Leu	1

Amino acids					
Domain					
124	157	184	197	227	267
Extracellular		Cytoplasmic			
Ig-like V-type					
Change					
Ile	Arg	Lys	Asn	Asp	Pro
Val	Ser	Glu	Thr	Val	Leu
hydrophobic	(+) charged	(+) charged	hydrophilic	(-) charged	hydrophobic
hydrophobic	hydrophilic	(-) charged	hydrophilic	hydrophobic	hydrophobic

Table 7. Non-synonymous changes according in each allotype.
For changes at the protein level, only non-synonymous variations are represented.

In our data of 852 individuals from 13 different worldwide populations (1000 Genomes project), we were able here to identify 6 SNPs in the coding region of the human *BTLA* and affecting the amino acid sequence in addition to 1 synonymous SNP (not represented here; the unchanged amino acid position is 56- Leucine). These nucleotides changes served to define 7 allotypes (Tables 6-7), those expressing Leucine at the position 267 representing 57.63 % of our population.

In our data, the BTLA*01 is less frequent in Africa compared to the other regions (*p=5E-08 <<0.05* t-student test). In Africa, the most frequent allotype corresponds to the protein that has a Proline at the position 267 instead of a Leucine.

The cytoplasmic amino changes that we identified do not belong to any ITIM motif.

3- TNFRSF14, Tumor necrosis factor receptor superfamily, member 14

TNFRSF14 is also known as HVEM (herpesvirus entry mediator). In addition to the binding to BTLA, HVEM is a receptor for several ligands (2 members of the TNF Super Family; LIGHT (TNFSF14) and LTα lymphotoxin α).

TNFRSF14*		*01	*02	*03	*04	*05	*06	rare alleles
South Asia	BEB	50.00	27.50	19.17	0.83	0	0	
	PJL	52.94	28.43	13.73	0.98	0	0	
	STU	46.83	33.33	16.67	0.79	0	0	
East Asia	CHB	43.06	45.83	5.56	4.17	0	0	
	CDX	69.87	21.15	3.85	4.49	0	0	
Africa	ESN	73.40	9.57	3.19	3.19	5.32	0	
	GWD	73.86	9.09	7.39	6.25	0	0	
	LWK	60.48	16.13	15.32	0.81	4.03	1.61	
	YRI	71.81	13.83	6.38	1.06	3.19	1.60	
Americas	PUR	42.97	48.44	6.25	0	0	2.34	
	CLM	44.74	39.47	10.53	1.75	0	0	
Europe	TSI	34.21	48.25	17.54	0	0	0	
	CEU	29.66	55.93	14.41	0	0	0	
	ALL	54.75	29.46	10.33	2.05	0.94	0.47	2

Table 8. Worldwide distribution of *TNFRSF14* allotype frequencies (in percentage).

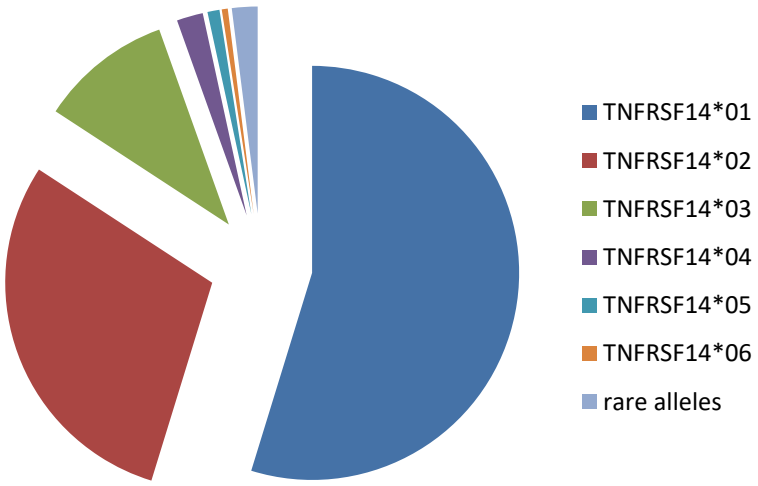


Figure 6. *TNFRSF14* allotype frequencies (in percentage).

	Amino acids				
	17	117	174	241	
REF (Uniprot)	Lys	Ala	Gly	Val	
TNFRSF14*01	Arg	.	.	.	933
TNFRSF14*02	.	.	.	.	502
TNFRSF14*03	Arg	.	.	Ile	176
TNFRSF14*04	.	Thr	.	.	35
TNFRSF14*05	.	.	Glu	.	16
TNFRSF14*06	.	.	.	Ile	8

Aminoacids			
Domain			
17	117	174	241
Signal peptide	Extracellular	Extracellular	Cytoplasmic
	TNFR Cys 2		
Change			
Lys	Ala	Gly	Val
Arg	Thr	Glu	Ile
positively charged	hydrophobic	hydrophobic	hydrophobic
positively charged	hydrophilic	negatively charged	hydrophobic

Table 9. Non-synonymous changes according in each allotype.
For changes at the protein level, only non-synonymous variations are represented.

In our data of 852 individuals from 13 different worldwide populations (1000 Genomes project), we were able here to identify 4 SNPs in the coding region of the human *TNFRSF14* and affecting the amino acid sequence in addition to 2 synonymous SNPs (not represented here; the unchanged amino acid positions are the following: 35 Tyr, 116 Asn). These nucleotides changes served to define 6 allotypes (Tables 8-9).

In our data, the TNFRSF14*01 allotype that corresponds to the protein having Arginine in position 17 is more frequent in East Asia and Africa compared to the other regions (*p<0.005* t-student test).

By checking the crystal structure of the BTLA-HVEM complex[96] visualized with PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.), the concerned amino acids (non-synonymous substitutions) do not belong to the binding sites of the BTLA-HVEM heterodimer (figure in Annex 4).

4- TNFRSF1B, Tumor necrosis factor receptor superfamily, member 1B

TNFRSF1B induces cell survival pathways that can result in cell proliferation.

TNFRSF1B*		*01	*02	*03	*04	*05	rare alleles
South Asia	BEB	73.33	21.67	0	2.50	0	
	PJL	69.61	24.51	0	4.90	0	
	STU	69.05	26.19	0	3.17	0	
East Asia	CDX	80.77	13.46	0	2.56	0	
	CHB	78.47	16.67	0	1.39	0	
Africa	ESN	70.21	18.09	3.19	0	7.45	
	GWD	76.14	16.48	3.98	1.14	1.70	
	LWK	59.68	29.84	7.26	0	0.81	
	YRI	73.40	13.30	6.38	1.06	4.79	
Americas	CLM	78.95	14.91	0	3.51	0	
	PUR	85.94	12.50	0	0	0.78	
Europe	TSI	80.70	17.54	0	1.75	0	
	CEU	78.81	16.95	0	2.54	0	
	ALL	75.23	18.19	1.82	1.82	1.23	1.70

Table 10. Worldwide distribution of *TNFRSF1B* allotype frequencies (in percentage).

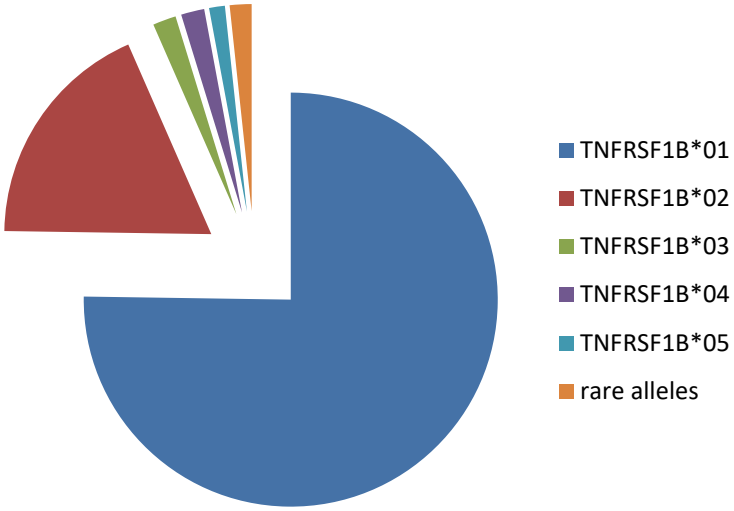


Figure 7. *TNFRSF1B* allotype frequencies (in percentage).

Amino acids			
187	196	232	264

REF	Val	Met	Glu	Leu	
TNFRSF1B*01	.	.	.	.	1282
TNFRSF1B*02	.	Arg	.	.	310
TNFRSF1B*03	.	.	.	Pro	31
TNFRSF1B*04	.	Arg	Lys	.	31
TNFRSF1B*05	Met	.	.	.	21

domain

Val	Met	Glu	Leu
Extracellular			Transmembrane
TNFR Cys 4			

change

Val	Met	Glu	Leu
Met	Arg	Lys	Pro

hydrophobic	hydrophobic	(-) charged	hydrophobic
hydrophobic	(+) charged	(+) charged	hydrophobic

Table 11. Non-synonymous changes according in each allotype.
For changes at the protein level, only non-synonymous variations are represented.

In our data of 852 individuals from 13 different worldwide populations (1000 Genomes project), we were able here to identify 4 SNPs in the coding region of the human *TNFRSF1B* and affecting the amino acid sequence in addition to synonymous 1 SNP (not represented here; the unchanged amino acid position is the following: 56 Lys). These nucleotides changes served to define 5 allotypes (Tables 10-11).

In our data, the TNFRSF1B*01 allotype less frequent in South Asia and Africa compared to the other regions (*p<0.005* t-student test).

5- NRAMP1, Natural resistance-associated macrophage protein 1 (SLC11A1)

Solute carrier family 11, member 1 (SLC11A1) is known as NRAMP1, is a divalent cation transporter (preferentially for iron and manganese) found in the membrane of phagolysosomes, conferring innate resistance to macrophages against the growth of intracellular microorganisms by providing efflux of metal cations such as Fe^{2+}, Zn^{2+} and Mn^{2+} from the phagosomal lumen into cytoplasm. Thus, due to this depletion of metal, these pathogens are not able anymore to activate their defensive enzymes. Several studies tried to confirm the association of *NRAMP1* variability and tuberculosis susceptibility.

	NRAMP1*	*01	*02	*03	*04	*05	rare alleles
South Asia	BEB	80.00	13.33	0.83	0.83	0	
	PJL	94.12	3.92	0	0	0	
	STU	88.10	7.14	0	0	0	
East Asia	CHB	79.17	15.97	0	0	2.08	
	CDX	83.33	8.97	5.77	0.64	1.28	
Africa	ESN	95.74	3.19	0	0	0	
	GWD	91.48	3.98	0	0	0	
	LWK	91.94	6.45	0	0	0	
	YRI	91.49	6.38	0	0	0	
Europe	CEU	98.31	0.00	0	1.69	0	
	TSI	92.11	0.88	0	1.75	0	
Americas	PUR	97.66	1.56	0	0.78	0	
	CLM	94.74	0.88	0	0.88	0	
	ALL	90.26	5.87	0.59	0.47	0.29	2.52

Table 12. Worldwide distribution of *NRAMP1* allotype frequencies (in percentage).

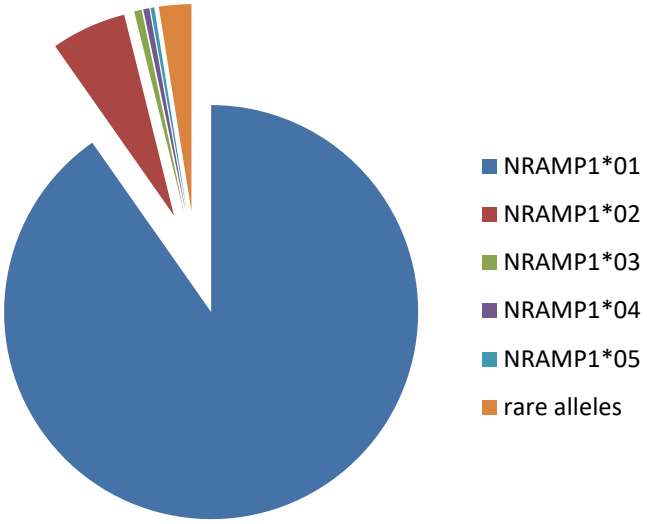


Figure 8. *NRAMP1* allotype frequencies (in percentage).

	Amino acids				
	419	443	498	543	
REF	Arg	Val	Gly	Asp	
NRAMP1*01	.	.	.	.	1538
NRAMP1*02	.	.	.	Asn	100
NRAMP1*03	.	.	Ser	.	10
NRAMP1*04	.	Ala	.	.	8
NRAMP1*05	Gln	.	.	.	5

domain			
Arg	Val	Gly	Asp
Extracellular	Transmembrane		Cytoplasmic

change			
Arg	Val	Gly	Asp
Gln	Ala	Ser	Asn

(+) charged	hydrophobic	hydrophobic	(-) charged
hydrophilic	hydrophobic	hydrophilic	hydrophilic

Table 13. Non-synonymous changes according in each allotype.
For changes at the protein level, only non-synonymous variations are represented.

In our data of 852 individuals from 13 different worldwide populations, we were able here to identify 4 SNPs in the coding region of the human *NRAMP1* and affecting the amino acid sequence in addition to synonymous SNPs (not represented here; the unchanged amino acid

positions are the following: 30 Gln, 39 Leu, 66 Phe, 249 Gly, 275 Ala). These nucleotides changes served to define 5 allotypes (Tables 12-13).

In our data, the NRAMP1*01 allotype is less frequent in East and South Asia compared to the other regions (*p<0.05* t-student test).

Significant associations were observed between *SLC11A1* variation D543N (and other variations in regulatory regions and untranslated regions UTR) and tuberculosis risk (with an odd ratio 1.25 >1). D543N is a non-synonymous variation that alters the amino acid 543 of the protein (Aspartic acid "D" –Asparagine "N")[97]. This variation was identified in our analysis (allotype NRAMP*02). It is more frequent in East and South Asia compared to the other regions (p=0.03).

6- IFNGR1, Interferon γ receptor 1

IFNGR1 (associated with IFNGR2) form a receptor for the soluble cytokine interferon gamma (IFNG). IFN γ (dimer state is biologically active) produced by activated lymphocytes, is essential for immunity against some viral, bacterial and protozoal infections. It has important immunoregulatory and immune functions (potent activator of macrophages, role in phagosome maturation and acidification etc.). IFN γ plays a key role in the control of mycobacterial infections.

	IFNGR1*	*01	*02	*03	*04	rare alleles
South Asia	BEB	71.67	24.17	0	0	
	PJL	64.71	33.33	0	0	
	STU	69.84	26.98	0	0	
East Asia	CDX	92.95	7.05	0	0	
	CHB	96.53	1.39	0	0	
Africa	ESN	94.68	0	2.13	1.06	
	GWD	88.64	1.70	4.55	2.27	
	LWK	99.19	0	0	0	
	YRI	93.09	0	0.53	3.19	
Europe	CEU	99.15	0	0	0	
	TSI	98.25	0	0	0	
Americas	CLM	95.61	2.63	0.88	0	
	PUR	95.31	0.78	0	0	
	ALL	89.61	6.87	0.70	0.65	2.17

Table 14. Worldwide distribution of *IFNGR1* allotype frequencies (in percentage).

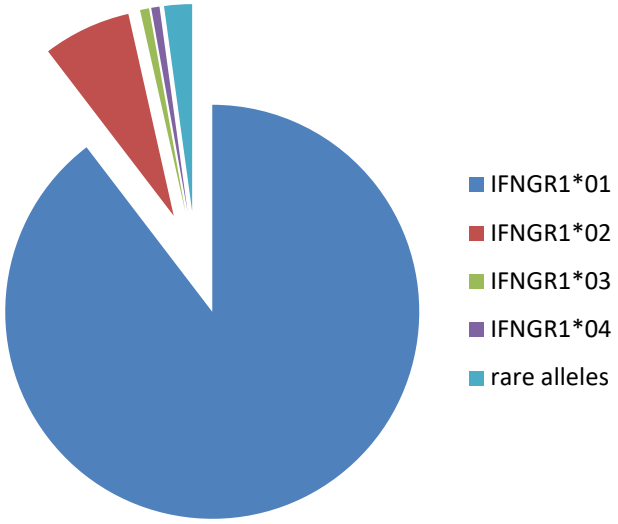


Figure 9. *IFNGR1* allotype frequencies (in percentage).

	Amino acids			
	335	411	467	
REF	His	Gly	Leu	
IFNGR1*01	.	.	.	1527
IFNGR1*02	.	.	Pro	117
IFNGR1*03	Pro	.	.	12
IFNGR1*04	.	Cys	Pro	11

domain		
335	411	467
Cytoplasmic		

change		
His	Gly	Leu
Pro	Cys	Pro

(+) charged	hydrophobic	hydrophobic
hydrophobic	hydrophilic	hydrophobic

Table 15. Non-synonymous changes according in each allotype.
For changes at the protein level, only non-synonymous variations are represented.

In our data of 852 individuals from 13 different worldwide populations (1000 Genomes project), we were able here to identify 3 SNPs in the coding region of the human *IFNGR1* and affecting the amino acid sequence in addition to synonymous 3 SNPs (not represented here; the unchanged amino acid positions are the following: 16 Arg, 112 Ala, 350 Ser). Other novel SNPs could be identified after sequencing or analyzing more genomes. These nucleotides changes served to define 4 allotypes (Tables 14-15).

In our data, the IFNGR1*01 allotype is less frequent in South Asia compared to the other regions ($p<0.005$ t-student test).

7- IFNG, CCL2, CISH

These 3 genes (*IFNG, CCL2, CISH*) are highly conserved (less polymorphic).

IFNG, Interferon γ

IFNG encodes a soluble cytokine that is a member of the type II interferon class. The protein is secreted by cells of both the innate and adaptive immune systems. The active protein is a homodimer that binds to the IFNGR which triggers a cellular response to viral and microbial infections. Mutations in this gene are associated with an increased susceptibility to viral, bacterial and parasitic infections and to several autoimmune diseases (source NCBI Gene ID: 3458).

		IFNG*01	IFNG*02
South Asia	BEB	100	0
	PJL	100	0
	STU	100	0
East Asia	CDX	99.36	0.64
	CHB	100	0
Africa	ESN	100	0
	GWD	100.00	0
	LWK	100	0
	YRI	100	0
Americas	CLM	100	0
	PUR	99.22	0.78
Europe	TSI	100	0
	CEU	100	0
	ALL	99.88	0.117371

Table 16. Worldwide distribution of *IFNG* allotype frequencies (in percentage).

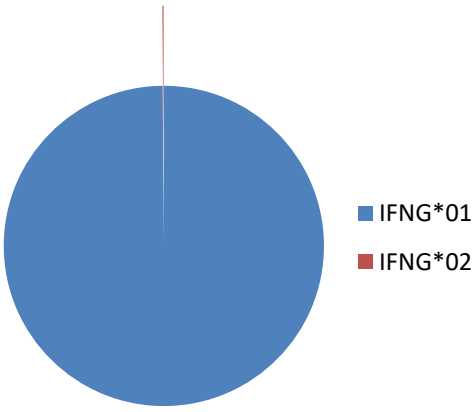


Figure 10. *IFNG* allotype frequencies (in percentage).

	Amino acids		
	77	160	
REF	Phe	Arg	
IFNG*01	Phe	Arg	1702
IFNG*02	Phe	Gln	2
domain			
		chain	
change			
	hydrophobic	(+) charged	
	hydrophobic	hydrophilic	

Table 17. Non-synonymous changes according in each allotype.

CISH, Cytokine-inducible SH2-containing protein

CISH is a negative regulator of cytokine signaling inhibiting inflammatory responses, having also a regulatory role in T-cell activation and differentiation in response to *M.tuberculosis* infections in human.

		CISH*01	CISH*02	rare alleles
South Asia	BEB	99.17	0.83	
	PJL	100	0	
	STU	97.62	0.79	
East Asia	CDX	98.08	0.64	
	CHB	96.53	1.39	
Africa	ESN	97.87	0	
	GWD	100	0	
	LWK	96.77	0	
	YRI	98.94	0	
Americas	CLM	98.25	0	
	PUR	96.88	0	
Europe	CEU	98.31	0	
	TSI	98.25	0	
	ALL	98.24	0.29	1.47

Table 18. Worldwide distribution of *CISH* allotype frequencies (in percentage).

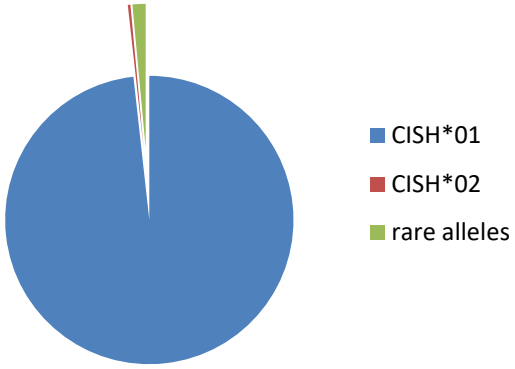


Figure 11. *CISH* allotype frequencies (in percentage).

	Amino acids	
AA	78	
REF	Arg	
CISH*01	Arg	1674
CISH*02	Gln	5

change
(+) charged
hydrophilic

Table 19. Non-synonymous changes according in each allotype.

CCL2, C-C motif chemokine 2

CCL2 is a chemotactic factor involved in the recruitment of monocytes to sites of injury and infection.

		CCL2*01	CCL2*02
South Asia	BEB	100	0
	PJL	100	0
	STU	100	0
East Asia	CDX	100	0
	CHB	100	0
Africa	ESN	100	0
	GWD	100	0
	LWK	98.3871	1.612903
	YRI	100	0
Europe	CEU	100	0
	TSI	100	0
Americas	CLM	100	0
	PUR	100	0
	ALL	99.88263	0.117371

Table 20. Worldwide distribution of *CCL2* allotype frequencies (in percentage).

Figure 12. *CCL2* allotype frequencies (in percentage).

	Amino acids		
	35	69	
REF	Cys	Ile	
CCL2*01	.	.	1702
CCL2*02	.	Val	2
		change	
		hydrophobic	
		hydrophobic	

Table 21. Non-synonymous changes according in each allotype.

IV- Conclusion

67,000–200,000 common non-synonymous (NS) SNPs are estimated in the human population. Each one of us is thought to be heterozygous for 24,000–40,000 NS SNPs.[98] These amino acid point mutations may change the structure (three-dimensional structure), the stability and the function of the protein. In our results, we mentioned the changes in the proprieties of the amino acids (e.g. charge, hydrophobicity, etc.) that could have phenotypic consequences and/or disrupt the function (non-conservative mutations). Further analyses are useful to predict the impact of NS SNPs. Checking whether these amino acids have a specific role in the protein, such as involvement in an active site or in the binding region, remains necessary to elucidate the effect of such modifications.

We developed, ameliorated and validated our methodology to assess polymorphisms in coding regions. This pilot study can be used to setup a large scale analysis in terms of number of genes to analyze as well as the number of individuals or genomes to screen.

Chapter 5: Finding novel gene candidates and polymorphisms potentially involved in host-pathogen interactions

Finding novel gene candidates and polymorphisms potentially involved in host-pathogen interactions

ABOU-KHATER Charbel [1, 2], OLIVE Daniel [1] , PONTAROTTI Pierre [3] , ABI-RACHED Laurent [2]

[1]INSERM, Unité 1068, CRCM, Marseille, France
[2]URMITE UM63 CNRS 7278 IRD 198, INSERM U1095, Marseille, France
[3] Aix Marseille University, CNRS, Centrale Marseille, I2M, UMR 7373 Marseille, France

I- INTRODUCTION

Finding novel gene candidates and polymorphisms involved in host-pathogen interactions

Host-pathogen co-evolution and interactions contribute to shaping the genetic diversity of both organisms.

To find novel gene candidates and polymorphisms that can explain the resistance/susceptibility to infectious diseases, in particular to tuberculosis, we used the Polymorphism Analysis approach.

Polymorphism analysis as a way to identify novel gene candidates involved in host-pathogen interactions

Development of Next Generation Sequencing (NGS) approaches and other technologies have continuously been improving.
Complete genomes and exome (coding regions) sequences represent a powerful source to define the human genetic diversity worldwide. Sequenced data in humans at the population level are available in open resources and databases. In some cases, these available polymorphism data haven't been properly analyzed for complex immune genes and gene families.

Here we developed large-scale *in silico* screens to identify novel candidates. Natural human worldwide diversity for genes is investigated by using our specialized *in silico* approach for coding regions.
The most polymorphic genes will be selected by considering that they will probably be more involved in host-pathogen interactions. Thus, we'll isolate immune genes that were targeted by natural selection, particularly in regions of the World where specific infections such as *tuberculosis* are more common.

1760 candidate genes
CD markers
Interleukins and Interleukin receptors
Immunoglobulin families (C1-set, C2-set, Ig-like domain containing, I-set, V-set)
Chemokines and chemokine receptors
Solute carriers
Melanocortin receptors
Integrins
Scavenger receptors
S100 family = S100 calcium binding proteins
Complement system
Surfactant proteins
Carcinoembryonic antigen family (CEACAM, PSG)
Sialic acid binding Ig-like lectins (SIGLEC)
Blood group antigens
C-type lectin domain family (CLEC)
Collectins
B7 family members
TNF super family and TNF receptors super family
Defensins family/ α and ß
Low density lipoprotein receptors (LRP)
Toll like receptors
HLA ou CMH genes; Major histocompatibility complex MHC
Mannose receptors
TGF family + TGF receptors
Families like TREM, ADGRE, CORO, TAP, ICAM, SLAM, CD300
Lectins
Some multigenic families such as KIR, TRIM
Genes from Host-pathogen interaction papers e.g. some genes implicated in phagocytosis (phagolysosome fusion, acidification, iron restriction)
Genes implicated in Iron homeostasis
Some genes implicated in the Immunity against tuberculosis
Vitamin B12 metabolism related genes
Vitamin D metabolism related genes

Table 1. List of selected candidate genes

We used a specialized approach useful even with 'difficult' genes and multigenic families which may contain highly similar genes and/or pseudogenes.

Our large scale analysis in terms of number of individuals and number of genes analyzed in a reasonable time is often accompanied with manual verifications for more accurate results.

Candidates were selected based on already established relevance for host-pathogen interactions specially in the case of TB (literature searches) and also on evolutionary considerations and domains of interest like multigenic families, structural domains etc.

II- Gene selection

We selected a list of 1760 candidate genes chosen according to some criteria as shown in table 1.

The HUGO Gene Nomenclature was used to facilitate data retrieval in other databases. The candidates' selection was extended to regroup genes encoding some cell surface markers, soluble proteins and families of proteins with key immune functions. Here's a brief description of each group as presented in the HGNC site.

CD molecules (391 genes)

CD stands for "cluster of differentiation". Human CD molecules are a subset of cellular surface markers providing targets for immunophenotyping of cells and with wide functions ranging from receptors or ligands to adhesion molecules. This nomenclature was conceived to classify the surface antigens initially named according to the monoclonal antibodies that bound to them.

ABCB1/ABCG2/ACE/ACKR1/ADAM10/ADAM17/ADAM8/ADGRE2/ADGRE5/ALCAM/ALK/ANPEP/ART1/ART4/ATP1B3/B3GAT1/BCAM/BMPR1A/BMPR1B/BSG/BST1/BST2/BTLA/BTN3A1/C5AR1/CCR1/CCR2/CCR3/CCR4/CCR5/CCR6/CCR7/CCR8/CCR9/CD101/CD109/CD14/CD151/CD160/CD163/CD164/CD177/CD180/CD19/CD1a/CD1b/CD1c/CD1d/CD1e/CD2/CD200/CD207/CD209/CD22/CD226/CD24/CD244/CD247/CD248/CD27/CD274/CD276/CD28/CD300a/CD300C/CD300E/CD300LB/CD300LD/CD300LF/CD300LG/CD302/CD320/CD33/CD34/CD36/CD37/CD38/CD3D/CD3E/CD3G/CD4/CD40/CD40LG/CD44/CD46/CD47/CD48/CD5/CD52/CD53/CD55/CD58/CD59/CD6/CD63/CD68/CD69/CD7/CD70/CD72/CD74/CD79a/CD79b/CD80/CD81/CD82/CD83/CD84/CD86/CD8a/CD8b/CD9/CD93/CD96/CD99/CDCP1/CDH1/CDH2/CDH5/CEACAM1/CEACAM3/CEACAM5/CEACAM6/CEACAM8/CLEC10A/CLEC4C/CLEC4M/CR1/CR2/CRTAM/CSF1R/CSF2RA/CSF2RB/CSF3R/CTLA4/CXCR1/CXCR2/CXCR3/CXCR4/CXCR5/CXCR6/DDR1/DDR2/DDRGK1/DPP4/ENG/ENPEP/ENPP3/ENTPD1/EPCAM/ERBB2/EVI2B/F11R/F3/FAS/FASLG/FCAMR/FCAR/FCER2/FCGR1A/FCGR2A/FCGR2B/FCGR2C/FCGR3A/FCGR3B/FCRL1/FCRL2/FCRL3/FCRL4/FCRL5/FGFR1/FGFR2/FGFR3/FGFR4/FLT3/FUT3/FUT2/FUT4/FZD10/FZD4/FZD9/GGT1/GP1BA/GP1BB/GP5/GP9/GYPA/GYPB/GYPC/HMMR/ICAM1/ICAM2/ICAM3/ICAM4/ICOS/ICOSLG/IFITM1/IFNGR1/IGF1R/IGF2R/IGLL1/IGSF8/IL10RA/IL10RA/IL10RB/IL12RB1/IL13RA1/IL13RA2/IL15RA/IL17RA/IL18R1/IL18RAP/IL1R1/IL1R2/IL21R/IL2RA/IL2RB/IL2RG/IL3RA/IL4R/IL5RA/IL6R/IL6ST/IL7R/IL9R/INSR/ITGA1/ITGA2/ITGA2B/ITGA3/ITGA4/ITGA5/ITGA6/ITGAD/ITGAE/ITGAL/ITGAM/ITGAV/ITGAX/ITGB1/ITGB2/ITGB3/ITGB4/JAG1/JAM2/KDR/KEL/KIR2DL1/KIR2DL2/KIR2DL3/KIR2DL4/KIR2DL5A/KIR2DS1/KIR2DS2/KIR2DS4/KIR2DS5/KIR3DL1/KIR3DL2/KIR3DL3/KIT/KLRB1/KLRC1/KLRC2/KLRD1/KLRK1/L1CAM/LAG3/LAIR1/LAIR2/LAMP1/LAMP2/LAMP3/LEPR/LIFR/LILRA1/LILRA2/LILRA3/LILRA4/LILRA5/LILRA6/LILRB1/LILRB2/LILRB3/LILRB4/LILRB5/LRP1/LY75/LY9/MCAM/MFI2/MME/MPL/MRC1/MRC2/MS4A1/MSR1/MST1R/MUC1/NCAM1/NCR1/NCR2/NCR3/NGFR/NRP1/NT5E/PDCD1/PDCD1LG2/PDGFRA/PDGFRB/PECAM1/PLAUR/PLXNC1/PRNP/PROCR/PROM1/PSG1/PTGDR2/PTGFRN/PTPRC/PTPRJ/PVR/PVRL1/PVRL2/PVRL3/RHAG/RHCE/RHD/S1PR1/SDC1/SDC2/SELE/SELL/SELP/SELPLG/SEMA4D/SEMA7A/SIGLEC1/SIGLEC5/SIGLEC6/SIGLEC7/SIGLEC9/SIRPA/SIRPB1/SIRPG/SLAMF1/SLAMF6/SLAMF7/SLAMF8/SLC44A1/SLC4A1/SLC7A5/SPN/TEK/TFRC/THBD/THY1/TLR1/TLR10/TLR2/TLR3/TLR4/TLR6/TLR8/TLR9/TNF

RSF10A/TNFRSF10B/TNFRSF10C/TNFRSF10D/TNFRSF11A/TNFRSF12A/TNFRSF13B/TNFRSF13C/TNFRSF14/TNFRSF17/TNFRSF18/TNFRSF1A/TNFRSF1B/TNFRSF21/TNFRSF4/TNFRSF8/TNFRSF9/TNFSF10/TNFSF11/TNFSF13/TNFSF13B/TNFSF14/TNFSF4/TNFSF8/TREM1/TSPAN7/VCAM1/VPREB1

Interleukins and Interleukins Receptors (61 genes)

It is a group of cytokines (secreted proteins and signaling molecules). In a large part, the function of the immune system depends on interleukins. An interleukin receptor is a cytokine receptor for interleukins. Rare deficiencies are responsible of some immune diseases. Here only genes that do not belong to the CD markers are represented in the following list.

CXCL8/IL10/IL11/IL11RA/IL12A/IL12B/IL12RB2/IL13/IL15/IL16/IL17A/IL17B/IL17C/IL17D/IL17F/IL17RB/IL17RC/IL17RD/IL17RE/IL18/IL18BP/IL19/IL1A/IL1B/IL1F10/IL1RAP/IL1RAPL1/IL1RAPL2/IL1RL1/IL1RL2/IL1RN/IL2/IL20/IL20RA/IL20RB/IL21/IL22/IL22RA1/IL22RA2/IL23A/IL24/IL25/IL26/IL27/IL27RA/IL3/IL31/IL31RA/IL32/IL33/IL34/IL36A/IL36B/IL36G/IL36RN/IL37/IL4/IL5/IL6/IL7/IL9

Immunoglobulin superfamily domain containing

The immunoglobulin superfamily (IgSF) is a large group of cell surface and soluble proteins that are involved in the recognition, binding, or adhesion processes of cells. Molecules are categorized as members of this superfamily based on shared structural features with immunoglobulins (also known as antibodies); they all possess a domain known as an immunoglobulin domain or fold. Members of the IgSF include cell surface antigen receptors, co-receptors and co-stimulatory molecules of the immune system, molecules involved in antigen presentation to lymphocytes, cell adhesion molecules, certain cytokine receptors and intracellular muscle proteins. They are commonly associated with roles in the immune system.

C1-set (42 genes)

Immunoglobulin C1-set domain: C1-set domains are classical Ig-like domains resembling the antibody constant domain. C1-set domains are found almost exclusively in molecules involved in the immune system.

AZGP1/B2M/CD1A/CD1B/CD1C/CD1D/CD1E/FCGRT/HFE/HHLA2/HLA-A/HLA-B/HLA-C/HLA-DMA/HLA-DMB/HLA-DOA/HLA-DOB/HLA-DPA1/HLA-DPB1/HLA-DQA1/HLA-DQA2/HLA-DQB1/HLA-DQB2/HLA-DRA/HLA-DRB1/HLA-DRB3/HLA-DRB4/HLA-DRB5/HLA-E/HLA-F/HLA-G/IGLL1/IGLL5/MICA/MICB/MR1/NCR3LG1/SIRPA/SIRPB1/SIRPG/TAPBP/TAPBPL

C2-set (40 genes)

Immunoglobulin C2-set domain: C2-set domains, which are Ig-like domains resembling the antibody constant domain. C2-set domains are found primarily in the mammalian T-cell surface antigens CD2 (Cluster of Differentiation 2), CD4 and CD80, as well as in vascular (VCAM) and intercellular (ICAM) cell adhesion molecules.

AGER/ALCAM/BCAM/BTN1A1/BTN2A1/BTN2A2/BTN3A1/BTNL2/CADM1/CADM2/CADM3/CADM4/CD2/CD200R1/CD200R1L/CD22/CD274/CD276/CD4/CD80/CRTAM/ICOSLG/KIRREL/KIRREL2/KIRREL3/MAG/MCAM/NPHS1/PVR/PVRL1/PVRL2/PVRL3/PVRL4/SIGLEC1/SIGLEC10/SIGLEC11/SIGLEC12/SIGLEC16/TMEM25/VCAM1

Immunoglobulin-like domain containing (241 genes)

A1BG/ADAMTSL3/ADGRA2/ADGRF5/AGER/ALCAM/ALPK3/AMIGO2/AMIGO3/AXL/BCAM/BOC/BSG/BTLA/CADM1/CADM3/CADM4/CCDC141/CD101/CD160/CD19/CD200/CD22/CD244/CD33/CD47/CD48/CD84/CD86/CD96/CDON/CEACAM1/CEACAM16/CEACAM18/CEACAM20/CEACAM21/CEACAM5/CEA

CAM6/CEACAM7/CEACAM8/CHL1/CILP/CILP2/CLMP/CNTFR/CNTN1/CNTN2/CNTN3/CNTN4/CNTN5/CSF1R/CXADR/DCC/DSCAM/DSCAML1/EMB/ESAM/F11R/FCAR/FCER1A/FCGR1A/FCGR1B/FCGR2A/FCGR2B/FCGR3A/FCGR3B/FCRL1/FCRL2/FCRL3/FCRL4/FCRL5/FCRL6/FCRLA/FCRLB/FLT1/FLT3/FLT4/FSTL4/GP6/GPA33/HEPACAM/HEPACAM2/HMCN2/HSPG2/ICAM1/ICAM2/ICAM3/ICAM4/ICAM5/IGDCC4/IGLON5/IGSF1/IGSF21/IGSF23/IGSF3/IGSF5/IGSF9/IGSF9B/IL11RA/IL12B/IL18BP/IL18R1/IL18RAP/IL1R1/IL1R2/IL1RAP/IL1RAPL1/IL1RAPL2/IL1RL1/IL1RL2/IL6R/ISLR2/JAM3/KDR/KIR2DL1/KIR2DL2/KIR2DL3/KIR2DL4/KIR2DL5A/KIR2DL5B/KIR2DS1/KIR2DS2/KIR2DS3/KIR2DS4/KIR2DS5/KIR3DL1/KIR3DL2/KIR3DL3/KIR3DS1/KIRREL/KIRREL2/KIRREL3/KIT/LAG3/LAIR1/LAIR2/LEPR/LILRA1/LILRA2/LILRA3/LILRA4/LILRA5/LILRA6/LILRB1/LILRB2/LILRB3/LILRB4/LILRB5/LSR/LY9/MADCAM1/MAG/MALT1/MCAM/MDGA1/MDGA2/MERTK/MILR1/MMP23B/MUSK/MXRA5/MYOM2/MYOM3/NCR1/NEO1/NPHS1/NPTN/NRCAM/NTRK1/OBSL1/OSCAR/PDCD1LG2/PDGFRB/PDGFRL/PECAM1/PSG1/PSG11/PSG2/PSG3/PSG4/PSG5/PSG6/PSG7/PSG8/PSG9/PTGFRN/PTK7/PTPRK/PTPRM/PTPRT/PVR/PVRL1/PVRL2/PVRL3/ROBO3/SCN1B/SDK1/SDK2/SEMA3A/SEMA3B/SEMA3E/SEMA3F/SEMA3G/SEMA4C/SEMA4D/SEMA4G/SEMA7A/SIGIRR/SIGLEC1/SIGLEC10/SIGLEC11/SIGLEC12/SIGLEC14/SIGLEC15/SIGLEC5/SIGLEC6/SIGLEC7/SIGLEC8/SIGLEC9/SLAMF1/SLAMF6/SLAMF8/TARM1/TEK/THY1/TIE1/TMEM81/TMIGD1/TMIGD2/TYRO3/VCAM1/VSIG1/VSIG10/VSIG10L/VSIG4/VSTM1/VTCN1/ZPBP/ZPBP2

I-set (156 genes)

ADAMTSL1/ADAMTSL3/ADGRA3/ADGRF5/ALPK2/ALPK3/BOC/BSG/CADM2/CD101/CDON/CEACAM20/CEACAM5/CHL1/CLMP/CNTN1/CNTN2/CNTN3/CNTN4/CNTN5/CNTN6/DCC/DSCAM/DSCAML1/EMB/FGFR1/FGFR2/FGFR3/FGFR4/FGFRL1/FLT1/FLT4/FSTL4/FSTL5/HEPACAM2/HSPG2/IGDCC3/IGDCC4/IGFBP7/IGFBPL1/IGFN1/IGSF10/IGSF11/IGSF22/IGSF5/IGSF9/IGSF9B/IL1R1/IL1RAPL1/IL1RL1/ISLR/JAM2/KAZALD1/KDR/KIRREL/KIRREL2/KIRREL3/L1CAM/LINGO1/LINGO2/LINGO3/LINGO4/LRFN1/LRFN2/LRFN3/LRFN4/LRFN5/LRIG1/LRIG2/LRIG3/LRIT1/LRIT2/LRIT3/LRRC24/LRRC4/LRRC4B/LRRC4C/LRRN1/LRRN2/LRRN3/LSAMP/MAG/MDGA1/MDGA2/MERTK/MFAP3/MFAP3L/MUSK/MXRA5/MYBPC1/MYBPC2/MYBPC3/MYBPH/MYBPHL/MYLK/MYOM1/MYOM2/MYOM3/MYOT/MYPN/NCAM1/NCAM2/NEGR1/NEO1/NEXN/NFASC/NPTN/NRCAM/NRG1/NRG2/NTM/NTRK2/NTRK3/OBSL1/OPCML/PALLD/PAPLN/PDGFRA/PDGFRB/PDGFRL/PRTG/PTK7/PTPRD/PTPRF/PTPRM/PTPRS/PVRL3/PXDN/PXDNL/ROBO1/ROBO2/ROBO3/ROBO4/ROR1/ROR2/SDK1/SDK2/SEMA3C/SIGLEC1/SIGLEC11/SIGLEC12/SIGLEC16/SIGLEC6/SPEG/TMIGD1/TRIO/TYRO3/UNC5A/UNC5B/UNC5C/UNC5D/VCAM1/VSIG10/VSIG2/WFIKKN1/WFIKKN2

V-set (163 genes)

Immunoglobulin V-set domain: V-set domains are Ig-like domains resembling the antibody variable domain. V-set domains are found in diverse protein families, including immunoglobulin light and heavy chains; in several T-cell receptors such as CD2 (Cluster of Differentiation 2), CD4, CD80, and CD86; in myelin membrane adhesion molecules; in junction adhesion molecules (JAM); in tyrosine-protein kinase receptors; and in the programmed cell death protein 1 (PD1).

ACAN/ADORA3/ALCAM/AMICA1/AMIGO1/AXL/BCAN/BTN1A1/BTN2A1/BTN2A2/BTN3A1/BTN3A2/BTN3A3/BTNL2/BTNL3/BTNL8/BTNL9/C10orf54/CADM1/CADM2/CADM3/CADM4/CD101/CD2/CD226/CD274/CD276/CD28/CD300A/CD300C/CD300E/CD300LB/CD300LD/CD300LF/CD300LG/CD33/CD4/CD48/CD7/CD79A/CD79B/CD80/CD83/CD86/CD8A/CD8B/CEACAM1/CEACAM16/CEACAM18/CEACAM19/CEACAM21/CEACAM3/CEACAM4/CEACAM5/CEACAM6/CEACAM7/CEACAM8/CRTAM/CTLA4/CXADR/ERMAP/ESAM/F11R/FAM187A/FCAMR/GPA33/HAPLN1/HAPLN2/HAPLN3/HAPLN4/HAVCR1/HAVCR2/HEPACAM/HEPACAM2/HHLA2/ICOSLG/IGLON5/IGSF21/IGSF3/IGSF6/IGSF8/IGSF9B/JAM2/JAM3/KDR/LY6G6F/MCAM/MOG/MPZ/MPZL1/MPZL2/MPZL3/MXRA8/NCAN/NCR2/NCR3/NPHS1/PDCD1/PIGR/PILRA/PILRB/PSG1/PSG11/PSG2/PSG3/PSG4/PSG5/PSG6/PSG8/PSG9/PTGFRN/PVR/PVRL1/PVRL2/PVRL3/PVRL4/SCN1B/SCN2B/SCN3B/SCN4B/SEMA3D/SIGLEC1/SIGLEC10/SIGLEC11/SIGLEC12/SIGLEC14/SIGLEC15/SIGLEC16/SIGLEC5/SIGLEC6/SIGLEC7/SIGLEC8/SIGLEC9/SIRPA/SIRPB1/SIRPB2/SIRPD/SIRPG/SLAMF6/SLAMF7/SLAMF9/TIGIT/TIMD4/TREM1/TREM2/TREML1/TREML2/TREML4/VCAN/VPREB1/VPREB3/VSIG1/VSIG10/VSIG2/VSIG4/VSIG8/VSTM2A/VSTM2B/VSTM2L/VSTM4/VSTM5/VTCN1

Chemokine receptors (25 genes)

Chemokine receptors are cytokine receptors found on the surface of certain cells that interact with a type of cytokine called a chemokine. There have been 19 distinct chemokine receptors described in mammals. Each has a 7- transmembrane (7TM) structure and couples to G-protein for signal transduction within a cell, making them members of a large protein family of G protein-coupled receptors. Following interaction with their specific chemokine ligands, chemokine receptors trigger a flux in intracellular calcium (Ca^{2+}) ions (calcium signaling). This causes cell responses, including the onset of a process known as chemotaxis that traffics the cell to a desired location within the organism. Chemokine receptors are divided into different families, CXC chemokine receptors, CC chemokine receptors, CX3C chemokine receptors and XC chemokine receptors that correspond to the 4 distinct subfamilies of chemokines they bind.

ACKR1/ACKR2/ACKR3/ACKR4/CCR1/CCR10/CCR2/CCR3/CCR4/CCR5/CCR6/CCR7/CCR8/CCR9/CCRL2/CCRL2/CX3CR1/CXCR1/CXCR2/CXCR3/CXCR4/CXCR5/CXCR6/PITPNM3/XCR1

CHEMOKINES LIGANDS (45 genes)

Chemokines (Greek -kinos, movement) are a family of small cytokines, or signaling proteins secreted by cells. Their name is derived from their ability to induce directed chemotaxis in nearby responsive cells; they are chemotactic cytokines. Proteins are classified as chemokines according to shared structural characteristics such as small size (they are all approximately 8-10 kilodaltons in size), and the presence of four cysteine residues in conserved locations that are key to forming their 3-dimensional shape. However, these proteins have historically been known under several other names including the SIS family of cytokines, SIG family of cytokines, SCY family of cytokines, Platelet factor-4 superfamily or intercrines. Some chemokines are considered pro- inflammatory and can be induced during an immune response to recruit cells of the immune system to a site of infection, while others are considered homeostatic and are involved in controlling the migration of cells during normal processes of tissue maintenance or development. Chemokines are found in all vertebrates, some viruses and some bacteria, but none have been described for other invertebrates. Chemokines have been classified into four main subfamilies: CXC, CC, CX3C and XC. All of these proteins exert their biological effects by interacting with G protein -linked transmembrane receptors called chemokine receptors that are selectively found on the surfaces of their target cells.

CCL1/CCL11/CCL13/CCL14/CCL15/CCL16/CCL17/CCL18/CCL19/CCL2/CCL20/CCL21/CCL22/CCL23/CCL24/CCL25/CCL26/CCL27/CCL28/CCL3/CCL3L1/CCL3L3/CCL4/CCL4L1/CCL4L2/CCL5/CCL7/CCL8/CX3CL1/CXCL1/CXCL10/CXCL11/CXCL12/CXCL13/CXCL14/CXCL16/CXCL17/CXCL2/CXCL3/CXCL5/CXCL6/CXCL8/CXCL9/XCL1/XCL2

Solute carrier (390 genes)

The solute carrier (SLC) group of membrane transport proteins includes over 300 members organized into 52 families. Most members of the SLC group are located in the cell membrane.

DIRC2/FLVCR1/FLVCR2/MFSD7/MTCH1/MTCH2/RHAG/RHBG/RHCG/SLC10A1/SLC10A2/SLC10A3/SLC10A4/SLC10A5/SLC10A6/SLC10A7/SLC11A1/SLC11A2/SLC12A1/SLC12A2/SLC12A3/SLC12A4/SLC12A5/SLC12A6/SLC12A7/SLC12A8/SLC12A9/SLC13A1/SLC13A2/SLC13A3/SLC13A4/SLC13A5/SLC14A1/SLC14A2/SLC15A1/SLC15A2/SLC15A3/SLC15A4/SLC15A5/SLC16A1/SLC16A10/SLC16A11/SLC16A12/SLC16A13/SLC16A14/SLC16A2/SLC16A3/SLC16A4/SLC16A5/SLC16A6/SLC16A7/SLC16A8/SLC16A9/SLC17A1/SLC17A2/SLC17A3/SLC17A4/SLC17A5/SLC17A6/SLC17A7/SLC17A8/SLC17A9/SLC18A1/SLC18A2/SLC18A3/SLC18B1/SLC19A1/SLC19A2/SLC19A3/SLC1A1/SLC1A2/SLC1A3/SLC1A4/SLC1A5/SLC1A6/SLC1A7/SLC20A1/SLC20A2/SLC22A1/SLC22A10/SLC22A11/SLC22A12/SLC22A13/SLC22A14/SLC22A15/SLC22A16/SLC22A17/SLC22A18/SLC22A2/SLC22A23/SLC22A24/SLC22A25/SLC22A3/SLC22A31/SLC22A4/SLC22A5/SLC22A6/SLC22A7/SLC22A8/SLC22A9/SLC23A1/SLC23A2/SLC23A3/SLC24A1/SLC24A2/SLC24A3/SLC24A4/SLC24A5/SLC25A1/SLC25A10/SLC25A11/SLC25A12/SLC25A13/SLC25A14/SLC25A15/SLC25

A16/SLC25A17/SLC25A18/SLC25A19/SLC25A2/SLC25A20/SLC25A21/SLC25A22/SLC25A23/SLC25A24/SLC25A25/SLC25A26/SLC25A27/SLC25A28/SLC25A29/SLC25A3/SLC25A30/SLC25A31/SLC25A32/SLC25A33/SLC25A34/SLC25A35/SLC25A36/SLC25A37/SLC25A38/SLC25A39/SLC25A4/SLC25A40/SLC25A41/SLC25A42/SLC25A43/SLC25A44/SLC25A45/SLC25A46/SLC25A47/SLC25A48/SLC25A5/SLC25A51/SLC25A52/SLC25A53/SLC25A6/SLC26A1/SLC26A10/SLC26A11/SLC26A2/SLC26A3/SLC26A4/SLC26A5/SLC26A6/SLC26A7/SLC26A8/SLC26A9/SLC27A1/SLC27A2/SLC27A3/SLC27A4/SLC27A5/SLC27A6/SLC28A1/SLC28A2/SLC28A3/SLC29A1/SLC29A2/SLC29A3/SLC29A4/SLC2A1/SLC2A10/SLC2A11/SLC2A12/SLC2A13/SLC2A14/SLC2A2/SLC2A3/SLC2A4/SLC2A5/SLC2A6/SLC2A7/SLC2A8/SLC2A9/SLC30A1/SLC30A10/SLC30A2/SLC30A3/SLC30A4/SLC30A5/SLC30A6/SLC30A7/SLC30A8/SLC30A9/SLC31A1/SLC31A2/SLC32A1/SLC33A1/SLC34A1/SLC34A2/SLC34A3/SLC35A1/SLC35A2/SLC35A3/SLC35A4/SLC35A5/SLC35B1/SLC35B2/SLC35B3/SLC35B4/SLC35C1/SLC35C2/SLC35D1/SLC35D2/SLC35D3/SLC35E1/SLC35E2/SLC35E2B/SLC35E3/SLC35E4/SLC35F1/SLC35F2/SLC35F3/SLC35F4/SLC35F5/SLC35G1/SLC35G2/SLC35G3/SLC35G4/SLC35G5/SLC35G6/SLC36A1/SLC36A2/SLC36A3/SLC36A4/SLC37A1/SLC37A2/SLC37A3/SLC37A4/SLC38A1/SLC38A10/SLC38A11/SLC38A2/SLC38A3/SLC38A4/SLC38A5/SLC38A6/SLC38A7/SLC38A8/SLC38A9/SLC39A1/SLC39A10/SLC39A11/SLC39A12/SLC39A13/SLC39A14/SLC39A2/SLC39A3/SLC39A4/SLC39A5/SLC39A6/SLC39A7/SLC39A8/SLC39A9/SLC3A1/SLC3A2/SLC40A1/SLC41A1/SLC41A2/SLC41A3/SLC43A1/SLC43A2/SLC43A3/SLC44A1/SLC44A2/SLC44A3/SLC44A4/SLC44A5/SLC45A1/SLC45A2/SLC45A3/SLC45A4/SLC46A1/SLC46A2/SLC46A3/SLC47A1/SLC47A2/SLC48A1/SLC4A1/SLC4A10/SLC4A11/SLC4A2/SLC4A3/SLC4A4/SLC4A5/SLC4A7/SLC4A8/SLC4A9/SLC50A1/SLC51A/SLC51B/SLC52A1/SLC52A2/SLC52A3/SLC5A1/SLC5A10/SLC5A11/SLC5A12/SLC5A2/SLC5A3/SLC5A4/SLC5A5/SLC5A6/SLC5A7/SLC5A8/SLC5A9/SLC6A1/SLC6A11/SLC6A12/SLC6A13/SLC6A14/SLC6A15/SLC6A16/SLC6A17/SLC6A18/SLC6A19/SLC6A2/SLC6A20/SLC6A3/SLC6A4/SLC6A5/SLC6A6/SLC6A7/SLC6A8/SLC6A9/SLC7A1/SLC7A10/SLC7A11/SLC7A13/SLC7A14/SLC7A2/SLC7A3/SLC7A4/SLC7A5/SLC7A6/SLC7A7/SLC7A8/SLC7A9/SLC8A1/SLC8A2/SLC8A3/SLC8B1/SLC9A1/SLC9A2/SLC9A3/SLC9A4/SLC9A5/SLC9A6/SLC9A7/SLC9A8/SLC9A9/SLC9B1/SLC9B2/SLC9C1/SLC9C2/SLCO1A2/SLCO1B1/SLCO1B3/SLCO1C1/SLCO2A1/SLCO2B1/SLCO3A1/SLCO4A1/SLCO4C1/SLCO5A1/SLCO6A1/UCP1/UCP2/UCP3

Melanocortin receptor (5 genes)

Melanocortin receptors are members of the rhodopsin family of 7-transmembrane G protein-coupled receptors. There are five known members of the melanocortin receptor system each with differing specificities for melanocortins : MC1R. MC1R is associated with pigmentation genetics. MC2R is also known as the ACTH receptor or corticotropin receptor because it is specific for ACTH alone. Defects in MC4R are a cause of autosomal dominant obesity, accounting for 6% of all cases of early-onset obesity. MC5R These receptors are inhibited by endogenous inverse agonists agouti signalling peptide and agouti-related peptide, and activated by synthetic (i.e. afamelanotide) and endogenous agonist melanocyte-stimulating hormones.

MC1R/MC2R/MC3R/MC4R/MC5R

Integrins (26 genes)

Protein of the integrin family of cell surface heterodimeric receptors that mediates dynamic cell-to-cell as well as cell-to-matrix adhesion. Integrins function as mechanochemical sensors and transducers able to change rapidly and reversibly their adhesive functions by modulating their ligand-binding affinity. Each subunit has a large N-terminal extracellular domain followed by a transmembrane domain and a short C-terminal cytoplasmic region. Some subclasses of integrins share a common beta chain while having different alpha chains.

ITGA1/ITGA10/ITGA11/ITGA2/ITGA2B/ITGA3/ITGA4/ITGA5/ITGA6/ITGA7/ITGA8/ITGA9/ITGAD/ITGAE/ITGAL/ITGAM/ITGAV/ITGAX/ITGB1 /ITGB2/ITGB3/ITGB4/ITGB5/ITGB6/ITGB7/ ITGB8

Scavenger receptors (30 genes)

Scavenger receptors are a group of receptors that recognize modified low-density lipoprotein (LDL) by oxidation or acetylation. This naming is based on a function of cleaning (scavenging): Scavenger receptors widely recognize and take up macromolecules that have a negative charge, like modified LDL.

CD163/CD163L1/CD36/CD5/CD5L/CD6/CD68/COLEC11/COLEC12/COLEC12/CXCL16/LRP1/LRP11/LRP1B/MARCO/MRC1/MRC2/MSR1/OLR1/SCARA3/SCARA5/SCARA5/SCARB1/SCARB2/SCARF1/SCARF2/SORL1/SSC4D/STAB1/STAB2

S100 calcium binding proteins (S100) (26 genes)

The S-100 proteins (often styled without the hyphen, S100) are a family of low-molecular-weight proteins found in vertebrates and characterized by two calcium-binding sites that have helix-loop-helix ("EF-hand type") conformation. There are at least 21 different S-100 proteins. Their name is derived from the fact that these proteins are soluble in 100%, i.e. saturated ammonium sulfate at neutral pH. They are encoded by a family of genes whose symbols use the S100 prefix, for example, S100A1, S100A2, S100A3.

CACYBP/HRNR/S100A1/S100A10/S100A11/S100A12/S100A13/S100A14/S100A16/S100A2/S100A3/S100A4/S100A5/S100A6/S100A7/S100A7A/S100A7L2/S100A8/S100A9/S100B/S100G/S100P/S100PBP/S100Z/SNTN/TCHHL1

Complement system (39 genes)

The complement system helps or "œcomplements" the ability of antibodies and phagocytic cells to clear pathogens from an organism. It is part of the immune system called the innate immune system that is not adaptable and does not change over the course of an individual's lifetime. However, it can be recruited and brought into action by the adaptive immune system. The complement system consists of a number of small proteins found in the blood, in general synthesized by the liver, and normally circulating as inactive precursors (pro-proteins). When stimulated by one of several triggers, proteases in the system cleave specific proteins to release cytokines and initiate an amplifying cascade of further cleavages. The end-result of this activation cascade is massive amplification of the response and activation of the cell-killing membrane attack complex. Over 30 proteins and protein fragments make up the complement system, including serum proteins, serosal proteins, and cell membrane receptors. They account for about 5% of the globulin fraction of blood serum and can serve as opsonins.

C1QA/C1QB/C1QC/C1R/C1S/C2/C3/C3AR1/C3AR1/C4A/C4B/C5/C5AR1/C5AR1/C5AR2/C6/C7/C8A/C8B/C8G/C9/CD46/CD59/CFB/CFD/CFH/CFHR1/CFHR2/CFHR3/CFHR4/CFHR5/CFI/CFP/CR1/CR1L/CR2/ITGAM/ITGAX/ITGB2

Surfactant proteins (5 genes)

Pulmonary surfactant is an essential barrier lowering the surface tension and participating in innate immune defense of the lungs[99].

SFTPA1/SFTPA2/SFTPB/SFTPC/SFTPD

Carcinoembryonic antigen family (CEACAM, PSG) (22 genes)

The carcinoembryonic antigen gene family belongs to the immunoglobulin superfamily and is divided into 2 groups, the carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) and the pregnancy-specific glycoproteins (PSG).

CEACAM1/CEACAM16/CEACAM18/CEACAM19/CEACAM20/CEACAM21/CEACAM3/CEACAM4/CEACAM5/CEACAM6/CEACAM7/CEACAM8/PSG1/PSG11/PSG2/PSG3/PSG4/PSG5/PSG6/PSG7/PSG8/PSG9

Sialic acid binding Ig-like lectins (SIGLECs) (15 genes)

Sialic acid-binding immunoglobulin-type lectins (SIGLEC) are proteins found primarily on the surface of immune cells and bind sialic acid. SIGLECs regulate both innate and adaptive immune responses by interacting with glycans on cells.

CD22/CD33/MAG/SIGLEC1/SIGLEC10/SIGLEC11/SIGLEC12/SIGLEC14/SIGLEC15/SIGLEC5/SIGLEC6/SIGLEC7/SIGLEC8/SIGLEC9

Blood group antigens (35 genes)

Blood group antigens are proteins found mainly on the cell surface of blood cells. In humans, more than fifteen different blood group systems are recognized. In most cases the antigenic determinant resides in the carbohydrate chains of membrane glycoproteins or glycolipids.

ABO/ACHE/ACKR1/AQP1/AQP3/ART4/B3GALNT1/BCAM/BSG/C4A/C4B/CD151/CD44/CD55/CR1/ERMAP/FUT1/FUT3/GCNT2/GYPA/GYPB/GYPC/GYPE/ICAM4/KEL/RHAG/RHCE/RHD/SEMA7A/SLC14A1/SLC4A1/SMIM1/XG/XK

C-type lectin domain family (CLEC) (47 genes)

A C-type lectin (CLEC) is a type of carbohydrate-binding protein domain known as a lectin. The C-type designation is from their requirement for calcium for binding. Proteins that contain C-type lectin domains have a diverse range of functions including cell-cell adhesion, immune response to pathogens and apoptosis

ASGR1/ASGR2/CD207/CD209/CD302/CD69/CLEC10A/CLEC11A/CLEC12A/CLEC12B/CLEC14A/CLEC16A/CLEC17A/CLEC18A/CLEC18B/CLEC18C/CLEC19A/CLEC1A/CLEC1B/CLEC2A/CLEC2B/CLEC2D/CLEC2L/CLEC3A/CLEC3B/CLEC4A/CLEC4C/CLEC4D/CLEC4E/CLEC4F/CLEC4G/CLEC4M/CLEC5A/CLEC6A/CLEC7A/CLEC9A/CLECL1/FCER2/KLRB1/KLRF1/KLRG1/KLRG2/LY75/MRC1/MRC2/OLR1/PLA2R1

Collectins (10 genes)

Collectins (collagen-containing C-type lectins) are a part of the innate immune system. They form a family of collagenous Ca 2+ -dependent defense lectins, which are found in animals. Collectins are soluble pattern recognition receptors (PRRs). Their function is to bind to oligosaccharide structure or lipids which are on the surface of microorganisms. Like other PRRs they bind pathogen-associated molecular patterns (PAMPs) and also danger-associated molecular patterns (DAMPs) of oligosaccharide origin. Binding of collectins to microorganisms may trigger elimination of microorganisms by aggregation, complement activation, opsonization, activation of phagocytosis or inhibition of microbial growth. Other functions of collectins are modulation of inflammatory, allergic responses, adaptive immune system and clearance of apoptotic cells.

COLEC10/COLEC11/COLEC12/MBL1P/MBL2/MBL3P/SFTPA1/SFTPA2/SFTPA3P/SFTPD

B7 Family (16 genes)

BTLA/C10orf54/CD274/CD276/CD28/CD80/CD86/CTLA4/HHLA2/ICOS/ICOSLG/NCR3LG1/PDCD1/PDCD1LG2/PDCD6/VTCN1

TNFSF (18 genes)

Tumor necrosis factors (or the TNF family) refer to a group of cytokines that can cause cell death (apoptosis).

CD40LG/CD70/EDA/FASLG/LTA/LTB/TNF/TNFSF10/TNFSF11/TNFSF12/TNFSF13/TNFSF13B/TNFSF14/TNFSF15/TNFSF18/TNFSF4/TNFSF8/TNFSF9

TNFRSF (29 genes)

A tumor necrosis factor receptor (TNFR), or death receptor, is a trimeric cytokine receptor that binds tumor necrosis factors (TNF).

CD27/CD40/EDA2R/EDAR/FAS/LTBR/NGFR/RELT/TNFRSF10A/TNFRSF10B/TNFRSF10C/TNFRSF10D/TNFRSF11A/TNFRSF11B/TNFRSF12A/TNFRSF13B/TNFRSF13C/TNFRSF14/TNFRSF17/TNFRSF18/TNFRSF19/TNFRSF1A/TNFRSF1B/TNFRSF21/TNFRSF25/TNFRSF4/TNFRSF6B/TNFRSF8/TNFRSF9

Defensins

Defensins are a large family of peptides contributing to host defense and having antibacterial, antifungal and antiviral properties. They are widely expressed in humans (phagocytes such as neutrophils, epithelial secretions (skin, mucosal epithelia…), inflammatory body fluids. Defensin-mediated killing occurs by forming voltage-regulated multimeric channels in the susceptible cell's membrane.

Defensins α alpha (5 genes)

Alpha defensins are a family of mammalian defensin peptides. Defensins are 2-6 kDa, cationic, microbicidal peptides active against many Gram-negative and Gram-positive bacteria, fungi, and enveloped viruses, containing three pairs of intramolecular disulfide bonds. On the basis of their size and pattern of disulfide bonding, mammalian defensins are classified into alpha, beta and theta categories. Alpha-defensins, which have been identified in humans, monkeys and several rodent species, are particularly abundant in neutrophils, certain macrophage populations and Paneth cells of the small intestine.

DEFA1 /DEFA3/DEFA4/DEFA5/DEFA6

Defensins ß beta (37 genes)

Beta defensins are a family of mammalian defensins. The beta defensins are antimicrobial peptides implicated in the resistance of epithelial surfaces to microbial colonization. Defensins are 2-6 kDa, cationic, microbicidal peptides active against many Gram-negative and Gram-positive bacteria, fungi, and enveloped viruses, containing three pairs of intramolecular disulfide bonds. On the basis of their size and pattern of disulfide bonding, mammalian defensins are classified into alpha, beta and theta categories. Every mammalian species explored thus far has beta-defensins. In cows, as many as 13 beta-defensins exist in neutrophils. However, in other species, beta-defensins are more often produced by epithelial cells lining various organs (e.g. the epidermis, bronchial tree and genitourinary tract. Human, rabbit and guinea-pig beta-defensins, as well as human beta-defensin-2 (hBD2), induce the activation and degranulation of mast cells, resulting in the release of histamine and prostaglandin D2.

DEFB1/DEFB103A/DEFB103B/DEFB104A/DEFB104B/DEFB105A/DEFB105B/DEFB106A/DEFB106B/DEFB107A/DEFB107B/DEFB108B/DEFB110/DEFB112/DEFB113/DEFB114/DEFB115/DEFB116/DEFB118/DEFB119/DEFB121/DEFB123/DEFB124/DEFB125/DEFB126/DEFB127/DEFB128/DEFB129/DEFB130/DEFB131/DEFB132/DEFB133/DEFB134/DEFB135/DEFB136/DEFB4A/DEFB4B

LRP (13 genes)

The low-density lipoprotein receptor gene family codes for a class of structurally related cell surface receptors that fulfill diverse biological functions in different organs, tissues, and cell types. The role that is most commonly associated with this evolutionarily ancient family is cholesterol homeostasis

(maintenance of appropriate concentration of cholesterol). In humans, excess cholesterol in the blood is captured by low-density lipoprotein (LDL) and removed by the liver via endocytosis of the LDL receptor.

LDLR/LRP1/LRP10/LRP11/LRP12/LRP1B/LRP2/LRP3/LRP4/LRP5/LRP6/LRP8/VLDLR

TLR (10 genes)

Toll-like receptors (TLRs) are a class of proteins that play a key role in the innate immune system. They are single, membrane-spanning, non-catalytic receptors usually expressed in sentinel cells such as macrophages and dendritic cells, that recognize structurally conserved molecules derived from microbes. Once these microbes have breached physical barriers such as the skin or intestinal tract mucosa, they are recognized by TLRs, which activate immune cell responses. The TLRs include TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TLR10, TLR11, TLR12, and TLR13, though the latter 3 are not found in human. They received their name from their similarity to the protein coded by the toll gene identified in Drosophila in 1985 by Christiane Nüsslein-Volhard. The researchers were so surprised that they spontaneously shouted out in German, " Das ist ja toll! " which translates as "That's great!"

TLR1/TLR10/TLR2/TLR3/TLR4/TLR5/TLR6/TLR7/TLR8/TLR9

HLA (22 genes)

The major histocompatibility complex (MHC) is a set of cell surface molecules encoded by a large gene family in all vertebrates. MHC molecules mediate interactions of leukocytes, also called white blood cells (WBCs), which are immune cells, with other leukocytes or body cells. MHC determines compatibility of donors for organ transplant as well as one's susceptibility to an autoimmune disease via crossreacting immunization.

HLAA/HLAB/HLAC/HLADMA/HLADMB/HLADOA/HLADOB/HLADPA1/HLADPB1/HLADQA1/HLADQA2/HLADQB1/HLADQB2/HLADQB3/HLADRA/HLADRB1/HLADRB3/HLADRB4/HLADRB5/HLAE/HLAF/HLAG

TGF and TGF Receptors(12 genes)

Transforming growth factors

TGFA/TGFB1/TGFB1I1/TGFB2/TGFB3/TGFBI/TGFBR1/TGFBR2/TGFBR3/TGFBR3L/TGFBRAP1/TIAF1

Lectins (15 genes)

Lectins are specific carbohydrate-binding proteins playing important roles in the first-line immune defense against invading microbes.

CLC/LGALS1/LGALS12/LGALS13/LGALS14/LGALS16/LGALS2/LGALS3/LGALS4/LGALS7/LGALS7B/LGALS8/LGALS9/LGALS9B/LGALS9C

KIR (15 genes)

Killer-cell immunoglobulin-like receptors (KIRs), are a family of cell surface proteins found on important cells of the immune system called natural killer (NK) cells. They regulate the killing function of these cells by interacting with MHC class I molecules, which are expressed on all cell types. This interaction allows them to detect virally infected cells or tumor cells that have a characteristic low level of Class I MHC on their surface. Most KIRs are inhibitory, meaning that their recognition of MHC suppresses the cytotoxic activity of their NK cell. Only a limited number of KIRs have the ability to activate cells. KIR molecules are highly polymorphic, meaning their gene

sequences differ greatly between individuals, so that different individuals possess different arrays/repertoires of *KIR* genes.

KIR3DL1/KIR3DS1/KIR3DL2/KIR3DL3/KIR2DS1/KIR2DS2/KIR2DS3/KIR2DS4/KIR2DS5/KIR2DL1/KIR2DL2/KIR2DL3/KIR2DL4/KIR2DL5A/KIR2DL5B

TRIM (77 genes)

Superfamily of Tripartite motif containing proteins, implicated in the restriction of viral growth.

CMYA5/MEFV/MID1/MID2/PML/TRIM10/TRIM11/TRIM13/TRIM14/TRIM15/TRIM16/TRIM17/TRIM2/TRIM21/TRIM22/TRIM23/TRIM24/TRIM25/TRIM26/TRIM27/TRIM28/TRIM29/TRIM3/TRIM31/TRIM32/TRIM33/TRIM34/TRIM35/TRIM36/TRIM37/TRIM38/TRIM39/TRIM4/TRIM40/TRIM41/TRIM42/TRIM43/TRIM44/TRIM45/TRIM46/TRIM47/TRIM48/TRIM49/TRIM49B/TRIM49C/TRIM49D1/TRIM49D2/TRIM5/TRIM50/TRIM51/TRIM52/TRIM54/TRIM55/TRIM56/TRIM58/TRIM59/TRIM6/TRIM60/TRIM61/TRIM62/TRIM63/TRIM64/TRIM64B/TRIM64C/TRIM65/TRIM66/TRIM67/TRIM68/TRIM69/TRIM7/TRIM71/TRIM72/TRIM73/TRIM74/TRIM77/TRIM8/TRIM9

CD300 family (7 genes)

The CD300 family of molecules modulates a broad and diverse array of immune cell processes via their paired activating and inhibitory receptor functions.[100]

CD300LG/CD300LF/CD300LD/CD300LB/CD300E/CD300C/CD300A

TREM (5 genes)

TREM proteins are a family of cell surface receptors that participate in diverse cell processes, including inflammation, bone homeostasis, neurological development and coagulation. TREM-1, the first to be identified, acts to amplify inflammation. Other TREM proteins regulate the differentiation and function of macrophages, microglia, dendritic cells, osteoclasts and platelets.[101]

TREM1/TREM2/TREML1/TREML2/TREML4

ADGRE (4 genes)

Adhesion G protein-coupled receptors, subfamily E, also known as epidermal growth factor (EGF)-TM7 (seven-span trans-membrane) subgroup of G-protein-coupled receptors, are involved in several immune processes.

ADGRE1/ADGRE2/ADGRE3/ADGRE5

CORO (7 genes)

Coronin is an actin binding protein which also interacts with microtubules and in some cell types is associated with phagocytosis. Coronin proteins are expressed in a large number of eukaryotic organisms from yeast to man.

CORO1A/CORO1B/CORO1C/CORO2A/CORO2B/CORO6/CORO7

ICAM(5 genes)

Intercellular adhesion molecule

ICAM1/ICAM2/ICAM3/ICAM4/ICAM5

LAMP (6 genes)

Lysosomal associated membrane protein

CD68 /LAMP1/LAMP2/LAMP3/LAMP5/LSAMP

SLAM (5 genes)

Signaling lymphocytic activation molecule family member

SLAMF1/SLAMF6/SLAMF7/SLAMF8/SLAMF9

Genes implicated in iron homeostasis (22 genes)

CD163/CP/CYBRD1/FTH1/FTL/HAMP/HEPH/HFE/HMOX1/HPX/HPX/ITLN1/LCN2/LRP1/LTF/NFE2L2/PCBP1/SLC11A1/SLC11A2/SLC40A1/TF/TFRC

STAT (8 genes)

Signal transducer and activator of transcription

MYD88/CORO1A/LAMP1/NOS2/TIRAP/TAP2/TAP1/STAT1

ABCB family (5 genes)

ATP binding cassette subfamily B (ABCB)

Also known as : "ATP-binding cassette, sub-family B (MDR/TAP)"

A subset of : ATP binding cassette transporters

TAP1/TAP2/TAPBP/TAPBPL/TAPT1

Genes implicated in host-pathogen interactions (268 genes)

ACBD3/AKAP1/AKAP10/AKAP11/AKAP12/AKAP13/AKAP14/AKAP17A/AKAP2/AKAP3/AKAP4/AKAP5/AKAP6/AKAP7/AKAP8/AKAP9/AMBP/APC2/APOD/APOM/ARFGEF2/BLM/BRCA1/BRCA2/BRIP1/C8G/CBFA2T3/CFTR/CMYA5/DYSF/ELP2/ELP3/ELP4/ELP5/ELP6/ERCC4/EZR/F10/F11/F12/F13A1/F13B/F2/F2R/F2RL1/F2RL2/F2RL3/F3/F5/F7/F7R/F8/F8A1/F8A3/F9/FANCA/FANCB/FANCC/FANCD2/FANCE/FANCF/FANCG/FANCI/FANCL/FANCM/FER1L5/FER1L6/FH/GJB1/GJB2/GJB3/GJB4/GJB5/GJB6/GJB7/IKBKAP/LCN1/LCN10/LCN12/LCN15/LCN2/LCN6/LCN8/LCN9/MAP2/MCOLN1/MCOLN2/MCOLN3/MYO7A/MYOF/MYRIP/NBEA/NF2/OBP2A/OBP2B/ORM1/ORM2/OTOF/PAEP/PALB2/PKD1/PKD2/PKD2L1/PKD2L2/PROC/PTGDS/RAB32/RAD51C/RBP4/SLX4/SPHKAP/SYNM/TFPI/TOR1A/TOR1AIP1/TOR1AIP2/TOR1B/TOR2A/TOR3A/TOR4A/TRPA1/TRPC1/TRPC3/TRPC4/TRPC5/TRPC6/TRPC7/TRPM1/TRPM2/TRPM3/TRPM4/TRPM5/TRPM6/TRPM7/TRPM8/TRPV1/TRPV2/TRPV3/TRPV4/TRPV5/TRPV6/UBE2T/WASF1/ADK/AK3/ALOX15/APOL1/APOL2/APOL3/APOL6/AREG/ATF3/BCL2A1/BIRC3/CA2/CCL13/CCL15/CCL18/CCL19/CCL2/CCL20/CCL23/CCL5/CCR7/CD14/CD209/CD302/CD36/CD40/CERK/CFB/CHI3L2/CHN2/CLEC10A/CLEC7A/CR1/CTSC/CXCL10/CXCL11/CXCL2/CXCL9/CXCR4/EDN1/EGR2/FAS/FGL2/FN1/FOS/GADD45G/C1GALT1/GAS7/HESX1/HEXB/HNMT/HRH1/HS3ST1/HS3ST2/HSD11B1/IDO1/IGF1/IGFBP4/IL10/IL12B/IL15/IL15RA/IL18/IL1B/IL2RA/IL6/IL7R/INHBA/IRF1/IRF7/ITGAM/LIPA/LPAR6/LTA4H/MAF/MBL2/MRC1/MS4A4A/MS4A6A/MSR1/NAMPT/NFKB2/NOD1/NOD2/NOS2/OAS2/OASL/P2RX7/P2RY13/P2RY14/PDGFA/PFKFB3/PFKP/PIK3IP1/PLA1A/PSMA2/PSMB9/PSME2/PTX3/SEPP1/SFTPA1/SFTPA2/SLC11A1/SLC2A6/SLC31A2/SLC38A6/SLC4A7/SLC7A5/SPHK1/SPRY2/TGFB1/TGFBR2/TIRAP/TLR1/TLR2/TLR4/TLR5/TLR8/TLR9/TNF/TNFSF10/TPST2/TYMP/VCAN/VDR/XAF1

Vitamine B12 related genes(7 genes)

A vitamin B12 deficiency can be the result of a variety of health problems.

TCN2/TCN1/MMACHC/FOLR3/FOLR2/FOLR1/CUBN

Vitamine D related genes(4 genes)

Several studies suggested the link between susceptibility to infection and vitamin D deficiency.

GC/DHCR7/VDR/DBP

III- Data sources

To define diversity in current populations, we used data from 3 different sources to assess the variation worldwide.

1- The 1000 Genomes Project set up in 2008, is one of the largest public catalogues of genotype data providing a solid foundation to understand the human genetic variations. This project analyses 2,504 genomes from 26 worldwide populations. In our analysis, we used genomes from 9 different populations from this project as shown in Table 2.
2- Human Genome Diversity Project: we used data from the HGPD to complete our 1000 Genomes panel to regroup populations from around the world. 7 populations from this project were included in our analysis.
3- Other available and published data; we included genomes from 2 other populations: Khoisan and Melanesian.

To reconstruct the history of a certain variation, we used 2 ancient DNA genomes from archaic humans: *Neanderthal Altai* (from the Altai Mountains) and *Denisova*. These are full genomes with relatively high coverage.

	Data	Data type	# indiv	# pop	pop	# indiv	continent
Panels	1000 Genomes	Exome	613	9	CEU	59	Europe
					FIN	71	
					BEB	60	South Asia
					PJL	51	
					CHB	72	East Asia
					CDX	78	
					JPT	66	
					LWK	62	Africa
					YRI	94	
	HGPD	Exome	54	7	Cambodians	8	East Asia
					Maya	8	Native America
					Mbuti	8	Africa
					Mozabite	8	
					Pathan	8	South Asia
					San	6	Africa
					Yakut	8	East Asia
	Khoisan	Exome	15	1	Khoisan	15	South Africa
	Melanesian	Full Genome	31	1	Melanesian	31	**Oceania**
Ancient DNA Archaic	Neanderthal Altai	Full Genome	1	1			
	Denisova	Full Genome	1	1			
		Total	**715**	**20**			

Table 2. Data sources and panels ("#" means number, "indiv": individuals, "pop": populations)

IV- Methods and strategy

We used the standardized symbols of the human genes according to the HGNC-approved nomenclature as well as the latest available release of the genome assembly (Dec. 2013 GRCh38/hg38).

Using the UCSC Genome Browser database, we determined the reference sequences. These sequences (where the isolated specific reads of the gene will be mapped) are required in order to perform the analysis by using our dedicated pipeline (chapter 2). The repetitive elements in introns and UTRs were masked using the RepeatMasker program.

We determined the positions of the exons, the start and stop codon in our reference sequences by retrieving the chromosome locations obtained from the Table browser of the UCSC Genome Browser website (by taking into consideration the differences between sense and antisense strands). Using these positions, we calculated the total size of all the exons of each gene. Once we have the number of variable positions for each gene, we will normalize the value according to the total coding size to be able to compare between our genes.

Blast hits as negative filters

We took all our reference sequences and blasted them (offline blast- Run BLAST software on a local computer) against the whole human genome “hard-masked” using the Dec. 2013 (GRCh38/hg38) assembly of the human genome (hg38, GRCh38 Genome Reference Consortium Human Reference 38 (GCA_000001405.2)) available online (UCSC Genome Bioinformatics Site - hg38.fa.masked.gz - "Hard-masked" assembly sequence in one file).

The obtained hits were filtered by eliminating those which:
- correspond to the gene itself
- belong to an alternative haplotype

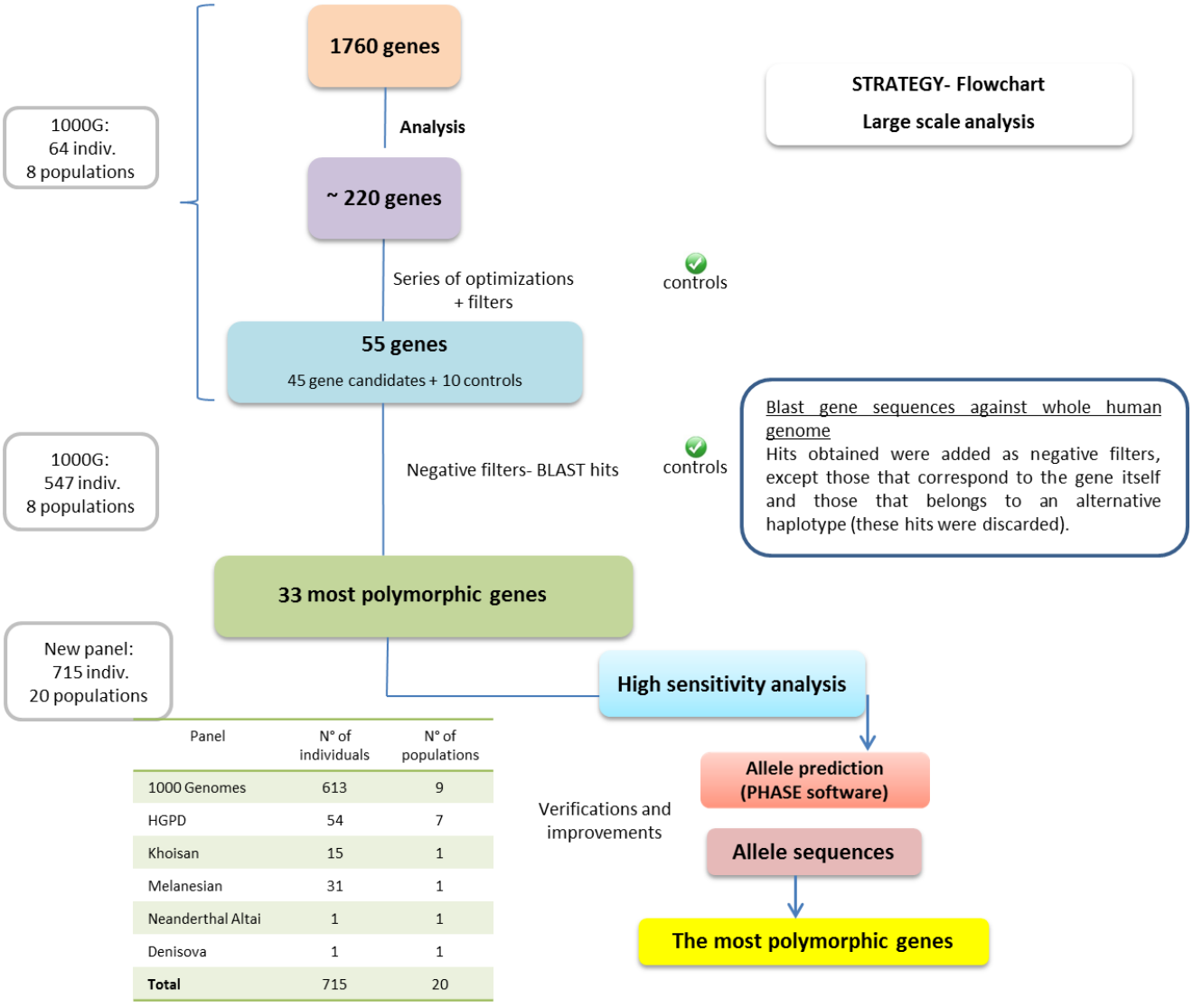
The remaining hits were added as negative filters using this approach developed specifically for the analysis of multigenic families which may contain highly similar genes and/or pseudogenes.

Candidate gene approach and large scale polymorphism affecting mainly protein products

The strategy used is described in Figure 1.

1- We selected a list of 1760 candidate genes (Table 1).
2- We got the DNA sequences, the exon positions and the total coding exons' size.
3- We analyzed the Exome sequencing data generated by the 1000 Genomes project. The first analysis included 64 individuals from 8 populations from the 1000 Genomes project: CEU-FIN-CHB-CDX-BEB-PJL-LWK-YRI (8 individuals per population). We investigated the diversity of the 1760 genes (pipeline described previously). Then we normalized the number of variations by dividing the number of SNPs found by the total exon's size. SNPs found in only one individual were discarded from this calculation.
4- The top 220 genes were selected for further more specific analysis.
5- During all these steps, our controls ranked as expected. We included genes known to be highly polymorphic like HLA and other genes for which the polymorphism was previously defined in our team (KIR, LILR, BTN)
6- After series of optimizations and filters, we took the top 45 genes of our list. 10 control genes were also added.
7- Then we analyzed these 55 genes with a panel of 547 individuals from the same 8 populations 1000G (CEU-FIN-CHB-CDX-BEB-PJL-LWK-YRI) (table 2) by adding negative sequences (obtained after doing the Blast against the whole human genome as explained previously) as filters.
8- The 33 most polymorphic genes (without the controls) were analyzed with a bigger panel of 715 individuals from 20 different populations to assess the worldwide diversity as seen in Table 2 (complete panel).

During all these steps, manual verifications have been done to check whether these filters worked correctly, and also to validate the variations obtained by checking the assembly, the distribution and ratios.

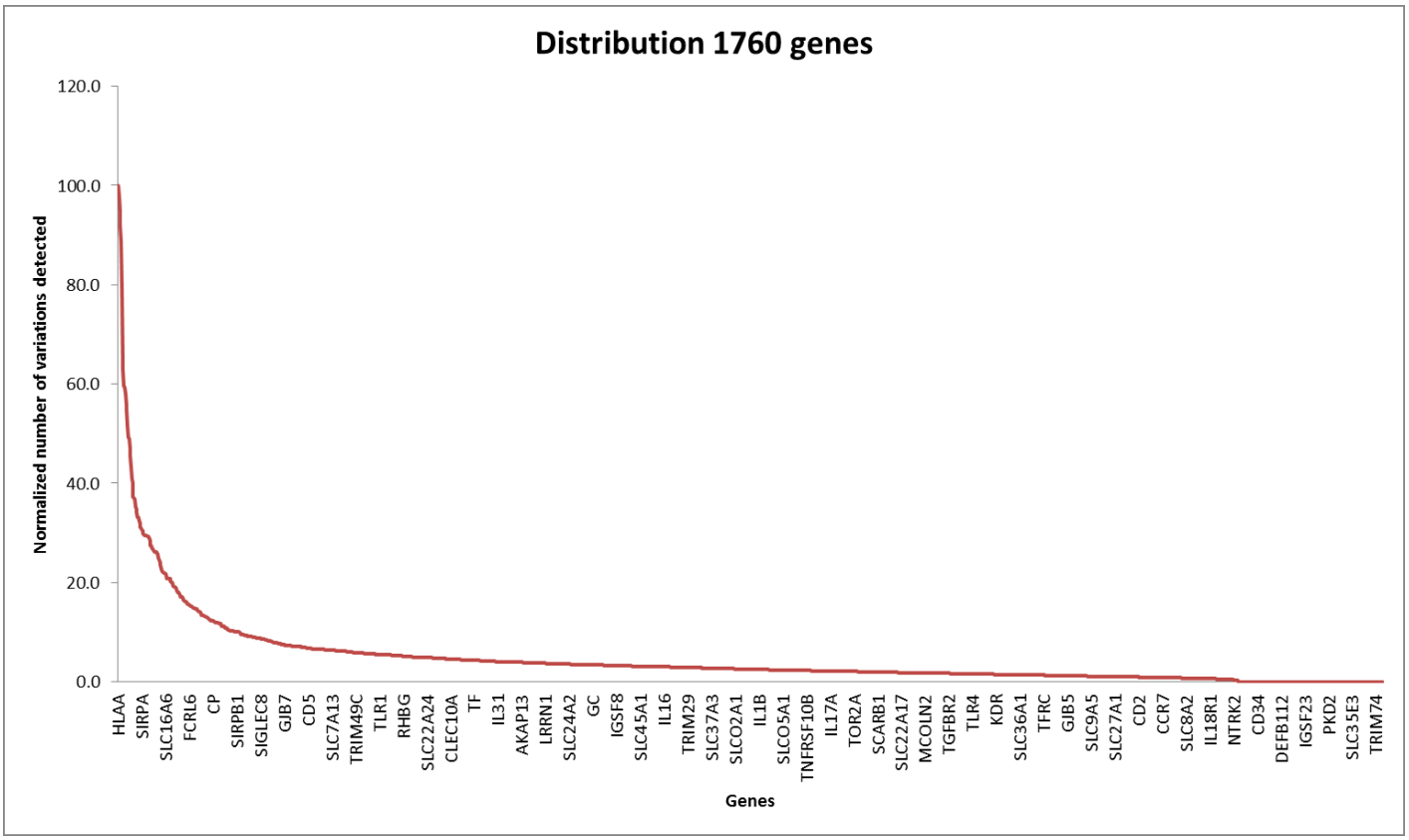


Panel	N° of individuals	N° of populations
1000 Genomes	613	9
HGPD	54	7
Khoisan	15	1
Melanesian	31	1
Neanderthal Altai	1	1
Denisova	1	1
Total	715	20

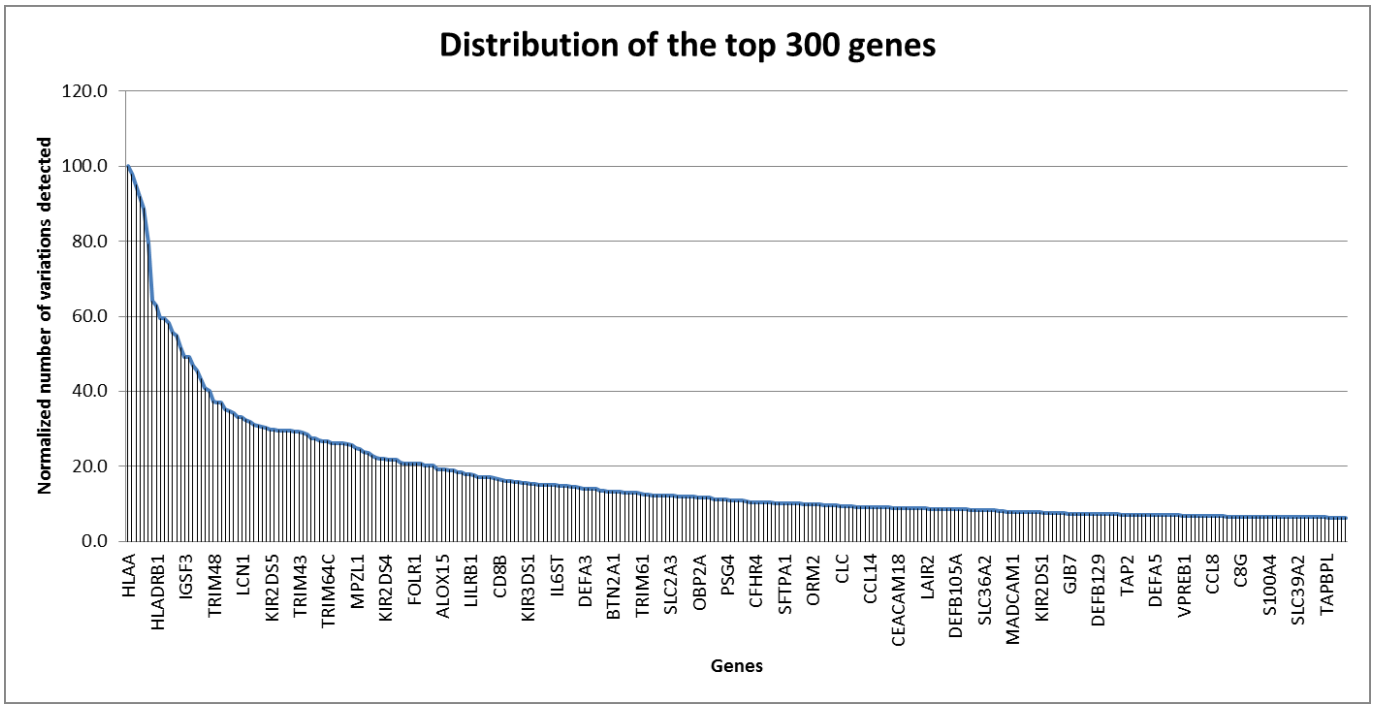
Figure 1. Flowchart explaining the strategy used.

V- Results

Initially, we investigated the diversity in coding regions of the 1760 selected genes in a set of 64 individuals representing 8 modern populations from the 1000 Genomes project. Then we normalized the number of variations by dividing the number of SNPs found by the total exon's size. SNPs found in only one individual were discarded. The complete list of the genes with their normalized number of variations is available in Annex 2. Graphs 1 and 2 show the distribution of the normalized number of variations detected for each gene. This number varies from 1 to 19 for around 84% of these genes, while for 11.5 % of the genes no SNP was detected (Table 3). A minority of genes appear to be highly polymorphic.



Graph 1. Distribution of the coding variations of the 1760 genes.



Graph 2. Distribution of the coding variations of the top 300 genes of our list.

Range of the number of normalized variations	Number of genes in the range	%
100-75	6	0.34
74-40	15	0.85
39-20	55	3.13
19-1	1482	84.20
0	202	11.48
total	1760	

Table 3. Distribution of the genes into ranges

HLAA	*HLAC*	*HLADQB1*	*HLAB*	*HLADRB5*	*HLADQA1*	*HLADRB4*	*HLADRB1*	***SLC35G4***	*HLADPB1*
78.0	75.9	72.4	67.5	61.6	56.3	50.3	49.1	40.4	40.2

HLADRB3	*HLADPA1*	***DEFB108B***	*MICA*	***HRNR***	*BTNL2*	***IGLL1***	*KIR3DL3*	***GYPB***	***DEFB132***
31.4	29.5	27.3	23.1	22.2	19.1	18.8	18.8	18.5	17.5

LILRB3	*LILRB1*	*LILRA6*	*KIR3DL1*	***OBP2A***	***ABO***	***DEFB1***	*LILRA2*	*LILRA1*	*KIR2DS5*
17.0	17.0	15.9	15.8	15.8	15.2	14.6	14.5	14.4	14.3

DEFA3	***PSG1***	*KIR2DL2*	***TRIM48***	*KIR2DS4*	***CLEC1B***	*HLADQA2*	*LILRA3*	*HLAG*	*KIR2DL3*
14.1	14.1	13.5	13.4	13.2	13.2	13.1	12.9	12.9	12.8

SLC25A5	***IL37***	***SLC35G5***	***XCL2***	***LGALS8***	***GYPA***	*LILRB2*	*KIR2DS3*	***PSG8***	***CD52***
12.3	12.3	11.8	11.7	11.2	11.2	11.2	11.0	11.0	10.9

CCL4	***S100A5***	*LILRB4*	***DEFB4B***	*HLADQB2*	***CD207***	***SFTPA1***	***P2RX7***	***LGALS2***	***DEFB127***
10.9	10.8	10.5	10.4	10.3	10.2	10.2	10.1	10.1	10.1

ORM2	***PSG2***	***PSG5***	***DEFA6***	***SIRPA***	*KIR2DS2*	***DEFB114***	***SIGLEC12***
10.0	10.0	10.0	10.0	10.0	9.9	9.6	9.6

Figure 2. Normalized number of variations detected by decreasing order (64 individuals and 8 populations from the 1000 Genomes Project, after series of optimization). Control genes are marked in grey.

After series of optimizations, with manual verifications and validations, the top 220 genes were selected for further more specific analysis. All our controls ranked as expected: the *HLA, BTN, LILR* and *KIR* genes (Figure 2). *HLA* genes are the most polymorphic genes of our list, and very well characterized. The polymorphism of the *BTN* and *LILR* family was defined in our team (Afrache H. *et al.*, to be submitted and Yeh JH. *et al.*, to be submitted). These genes, in addition to *KIR* genes, are also known to be highly polymorphic.

Out of this list, 45 genes were selected in addition to 10 controls, for the further steps. Another analysis with all the individuals of these 8 populations was done, regrouping 547 individuals, and by adding the Blast hits as negative filters to improve the pipeline using this approach developed specifically for the analysis of multigenic families which may contain highly similar genes and/or pseudogenes. The ranking of the top genes is shown in figure 3.

33 most polymorphic genes (marked in red in figure 3, in addition to *KIR3DL3*) are then selected for a complete analysis with 715 individuals from 20 different populations from several sequencing projects (Table 2). Final results are shown in Figure 4.

HLAA	*HLADQB1*	*HLAB*	*HLAC*	*HLADRB1*	***SLC35G4***	***GYPB***	*BTNL2*	*KIR2DL3*
102.75	93.63	89.65	87.83	66.67	54.13	44.28	38.18	34.38

FUT2	***C4B***	***DEFB108B***	***DEFB132***	***KIR3DL3***	***OBP2A***	***CCL4***	***IGLL1***	***PSG1***
34.20	33.25	31.82	31.47	31.02	29.59	28.99	28.17	26.54

PSG2	*KIR3DL1*	***C1GALT1***	***PSG5***	***ABO***	***ORM2***	***FUT3***	*KIR3DL2*	***CD207***
25.92	25.64	25.10	24.93	24.67	23.33	23.04	22.81	21.41

SIGLEC12	***SLC25A5***	***DEFA3***	***SIRPB1***	***SLC35G5***	***PSG8***	***DEFB1***	***SFTPA1***	***PSG11***
21.35	21.28	21.20	20.97	20.67	19.59	19.51	19.06	18.94

SIRPA	***P2RX7***	***DEFB126***	***IL37***	***DEFB127***	***CD52***	*CLEC1B*	*LGALS8*	*CD48*
18.58	18.03	17.96	16.87	16.78	16.30	16.08	15.89	15.87

DEFB4B	*LGALS2*	*FTH1*	*S100A5*	*CLEC4G*	*GYPA*	*DEFA6*	*CFHR1*	*CCL14*
15.54	15.19	14.60	14.44	13.75	13.45	13.29	13.17	12.27

FOLR3	*XCL2*	*DEFB114*	*DEFB124*
12.26	11.70	9.62	9.35

Figure 3. Normalized number of variations detected by decreasing order (547 individuals and 8 populations from the 1000 Genomes Project). Control genes are marked in grey.

SLC35G4	*SIRPA*	*DEFB108B*	*GYPB*	*IGLL1*	*PSG1*	*FUT2*	*SLC25A5*	*KIR3DL3*	*DEFB132*
77.76	76.31	59.09	59.04	45.38	42.94	42.68	42.55	42.45	41.96

PSG2	*PSG5*	*C4B*	*OBP2A*	*CCL4*	*SIRPB1*	*ORM2*	*SFTPA1*	*CD52*	*SIGLEC12*
40.88	39.88	39.47	39.45	36.23	36.07	35.00	33.04	32.61	31.46

ABO	*SLC35G5*	*DEFB126*	*FUT3*	*DEFB1*	*C1GALT1*	*PSG8*	*PSG11*	*P2RX7*	*CD207*
31.31	30.51	29.94	29.49	29.27	28.47	28.21	27.92	26.48	25.48

DEFA3	*IL37*	*DEFB127*
24.73	24.54	23.49

Figure 4. Normalized number of variations detected by decreasing order (715 individuals and 20 populations from the several projects).

By passing from 547 to 715 individuals, we calculated the percentage of variation. Interestingly, an increase of 311% for *SIRPA* gene was observed (supplementary table 1).

In Table 4, we listed the main functions of these 33 genes according to the Uniprot database.

Table 4.	Function (Uniprot)
SLC35G4	**Putative solute carrier family 35 member G4.** Protein uncertain. Drug/metabolite transporter (dmt) superfamily.
SIRPA	**Tyrosine-protein phosphatase non-receptor type substrate 1** Immunoglobulin-like cell surface receptor for CD47. Acts as docking protein and induces translocation of PTPN6, PTPN11 and other binding partners from the cytosol to the plasma membrane. Supports adhesion of cerebellar neurons, neurite outgrowth and glial cell attachment. May play a key role in intracellular signaling during synaptogenesis and in synaptic function (By similarity). Involved in the negative regulation of receptor tyrosine kinase-coupled cellular responses induced by cell adhesion, growth factors or insulin. Mediates negative regulation of phagocytosis, mast cell activation and dendritic cell activation. CD47 binding prevents maturation of immature dendritic cells and inhibits cytokine production by mature dendritic cells
DEFB108B	**Beta-defensin 108B** Has antibacterial activity.
GYPB	**Glycophorin-B** This protein is a minor sialoglycoprotein in erythrocyte membranes.
IGLL1	**Immunoglobulin lambda-like polypeptide 1** Critical for B-cell development.
PSG1	**Pregnancy-specific beta-1-glycoprotein 1** Female pregnancy- leukocyte migration
FUT2	**Galactoside 2-alpha-L-fucosyltransferase 2** Mediates the transfer of fucose to the terminal galactose on glycan chains of cell surface glycoproteins and glycolipids. The resulting epitope plays a role in cell-cell interaction including host-microbe interaction. Mediates interaction with intestinal microbiota influencing its composition. Creates a soluble precursor oligosaccharide FuC-alpha ((1,2)Galbeta-) called the H antigen which is an essential substrate for the final step in the soluble ABO blood group antigen synthesis pathway.
SLC25A5	**ADP/ATP translocase 2** Catalyzes the exchange of cytoplasmic ADP with mitochondrial ATP across the mitochondrial inner membrane. As part of the mitotic spindle-associated MMXD complex it may play a role in chromosome segregation.
KIR3DL3	**Killer cell immunoglobulin-like receptor 3DL3**

	Receptor on natural killer cells. May inhibit the activity of NK cells thus preventing cell lysis
DEFB132	**Beta-defensin 132** Has antibacterial activity.
PSG2	**Pregnancy-specific beta-1-glycoprotein 2** Female pregnancy- leukocyte migration
PSG5	**Pregnancy-specific beta-1-glycoprotein 5** Female pregnancy
C4B	**Complement C4-B** Non-enzymatic component of the C3 and C5 convertases and thus essential for the propagation of the classical complement pathway. Covalently binds to immunoglobulins and immune complexes and enhances the solubilization of immune aggregates and the clearance of IC through CR1 on erythrocytes. C4A isotype is responsible for effective binding to form amide bonds with immune aggregates or protein antigens, while C4B isotype catalyzes the transacylation of the thioester carbonyl group to form ester bonds with carbohydrate antigens. Derived from proteolytic degradation of complement C4, C4a anaphylatoxin is a mediator of local inflammatory process. It induces the contraction of smooth muscle, increases vascular permeability and causes histamine release from mast cells and basophilic leukocytes.
OBP2A	**Odorant-binding protein 2a** Probably binds and transports small hydrophobic volatile molecules with a higher affinity for aldehydes and large fatty acids.
CCL4	**C-C motif chemokine 4** Monokine with inflammatory and chemokinetic properties. Binds to CCR5. One of the major HIV-suppressive factors produced by CD8+ T-cells. Recombinant MIP-1-beta induces a dose-dependent inhibition of different strains of HIV-1, HIV-2, and simian immunodeficiency virus (SIV). The processed form MIP-1-beta(3-69) retains the abilities to induce down-modulation of surface expression of the chemokine receptor CCR5 and to inhibit the CCR5-mediated entry of HIV-1 in T-cells. MIP-1-beta(3-69) is also a ligand for CCR1 and CCR2 isoform B.
SIRPB1	**Signal-regulatory protein beta-1** Immunoglobulin-like cell surface receptor involved in the negative regulation of receptor tyrosine kinase-coupled signaling processes. Participates also in the recruitment of tyrosine kinase SYK.
ORM2	**Alpha-1-acid glycoprotein 2** Functions as transport protein in the blood stream. Binds various hydrophobic ligands in the interior of its beta-barrel domain. Also binds synthetic drugs and influences their distribution and availability. Appears to function in modulating the activity of the immune system during

	the acute-phase reaction.
SFTPA1	**Pulmonary surfactant-associated protein A1** In presence of calcium ions, it binds to surfactant phospholipids and contributes to lower the surface tension at the air-liquid interface in the alveoli of the mammalian lung and is essential for normal respiration. Enhances the expression of MYO18A/SP-R210 on alveolar macrophages (Microbial infection) Recognition of *M.tuberculosis* by dendritic cells may occur partially via this molecule. Can recognize, bind, and opsonize pathogens to enhance their elimination by alveolar macrophages.
CD52	**CAMPATH-1 antigen** May play a role in carrying and orienting carbohydrate, as well as having a more specific role
SIGLEC12	**Sialic acid-binding Ig-like lectin 12** Putative adhesion molecule that mediates sialic-acid dependent binding to cells. The sialic acid recognition site may be masked by cis interactions with sialic acids on the same cell surface.
ABO	**Histo-blood group ABO system transferase** This protein is the basis of the ABO blood group system. The histo-blood group ABO involves three carbohydrate antigens: A, B, and H. A, B, and AB individuals express a glycosyltransferase activity that converts the H antigen to the A antigen (by addition of UDP-GalNAc) or to the B antigen (by addition of UDP-Gal), whereas O individuals lack such activity.
SLC35G5	**Solute carrier family 35 member G5** The gene encoding this protein appears to have arisen by SVA-mediated retrotransposition of the SLC35G6 gene in the primate lineage. Drug/metabolite transporter (dmt) superfamily.
DEFB126	**Beta-defensin 126** Highly glycosylated atypical beta-defensin involved in several aspects of sperm function. Facilitates sperm transport in the female reproductive tract and contributes to sperm protection against immunodetection; both functions are probably implicating the negative surface charge provided by its O-linked oligosaccharides in the sperm glycocalyx. Involved in binding of sperm to oviductal epithelial cells to form a sperm reservoir until ovulation. Release from the sperm surface during capacitation and ovaluation by an elevation of oviductal fluid pH is unmasking other surface components and allows sperm to penetrate the cumulus matrix and bind to the zona pellucida of the oocyte. *In vitro* has antimicrobial activity and may inhibit LPS-mediated inflammation.
FUT3	**Galactoside 3(4)-L-fucosyltransferase** May catalyze alpha-1,3 and alpha-1,4 glycosidic linkages involved in the expression of Vim-2, Lewis A, Lewis B, sialyl Lewis X and Lewis X/SSEA-1 antigens. May be involved in blood group Lewis determination; Lewis-positive (Le^+) individuals have an active enzyme while Lewis-negative (Le^-) individuals have an inactive enzyme. Also acts on the corresponding 1,4-

	galactosyl derivative, forming 1,3-L-fucosyl links.
DEFB1	**Beta-defensin 1** Has bactericidal activity. May act as a ligand for C-C chemokine receptor CCR6. Positively regulates the sperm motility and bactericidal activity in a CCR6-dependent manner. Binds to CCR6 and triggers Ca2+ mobilization in the sperm which is important for its motility
C1GALT1	**Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1** Glycosyltransferase that generates the core 1 O-glycan Gal-beta1-3GalNAc-alpha1-Ser/Thr (T antigen), which is a precursor for many extended O-glycans in glycoproteins. Plays a central role in many processes, such as angiogenesis, thrombopoiesis and kidney homeostasis development.
PSG8	**Pregnancy-specific beta-1-glycoprotein 8** Cell surface interactions at the vascular wall.
PSG11	**Pregnancy-specific beta-1-glycoprotein 11** Female pregnancy
P2RX7	**P2X purinoceptor 7** Receptor for ATP that acts as a ligand-gated ion channel. Responsible for ATP-dependent lysis of macrophages through the formation of membrane pores permeable to large molecules. Could function in both fast synaptic transmission and the ATP-mediated lysis of antigen-presenting cells.
CD207	**C-type lectin domain family 4 member K** Calcium-dependent lectin displaying mannose-binding specificity. Induces the formation of Birbeck granules (BGs); is a potent regulator of membrane superimposition and zippering. Binds to sulfated as well as mannosylated glycans, keratan sulfate (KS) and beta-glucans. Facilitates uptake of antigens and is involved in the routing and/or processing of antigen for presentation to T cells. Major receptor on primary Langerhans cells for Candida species, Saccharomyces species, and Malassezia furfur. Protects against human immunodeficiency virus-1 (HIV-1) infection. Binds to high-mannose structures present on the envelope glycoprotein which is followed by subsequent targeting of the virus to the Birbeck granules leading to its rapid degradation.
DEFA3	**Neutrophil defensin 3** Defensin 2 and defensin 3 have antibiotic, fungicide and antiviral activities. Has antimicrobial activity against Gram-negative and Gram-positive bacteria. Defensins are thought to kill microbes by permeabilizing their plasma membrane.
IL37	**Interleukin-37** Suppressor of innate inflammatory and immune responses involved in curbing excessive

	inflammation. This function requires SMAD3. Suppresses, or reduces, proinflammatory cytokine production, including IL1A and IL6, as well as CCL12, CSF1, CSF2, CXCL13, IL1B, IL23A and IL1RN, but spares anti-inflammatory cytokines. Inhibits dendritic cell activation.
DEFB127	**Beta-defensin 127** Has antibacterial activity.

Interestingly, 6 genes our list of 33 top most polymorphic genes belong to the defensins family. These widely distributed antimicrobial peptides are implicated in antimicrobial activity of phagocytes, inflammatory body fluids and epithelial secretions. They contribute to host defense against bacterial, fungal and viral infections. Defensins are chemoattractants for macrophages and neutrophils. In addition to that, they are effective against Gram-positive and Gram-negative bacteria including mycobacteria[102]. A defensin cluster regrouping at least eight genes encoding α- and ß-defensins and located on chromosome 8p23, is incompletely mapped, due to its polymorphic nature with individuals differing in the number of copies of some specific defensins genes[103]. Here is the list of our 6 defensins: *DEFB108B, DEFB132, DEFB126, DEFB1, DEFA3 and DEFB127.* Two of these genes are present in multiple copies, polymorphically, in the human genome (*DEFB108B* and *DEFA3*), which makes it difficult to assess their polymorphism using our pipeline. For *DEFB132, DEFB126, DEFB1 and DEFB127,* the evidence we have supports them being at single copy in the human genome (Edward J. Hollox, *personal communication*).

VI- Further analysis and perspectives

For our selected genes, and after a 'high sensitivity' analysis in order to detect SNPs in low coverage regions that might be missed in the normal analysis as well as manual genotype verifications, we'll reconstruct the allelic sequences and define their frequencies. We'll define also the diversity and we'll check if selection occurred in some populations. Interesting genes are those isolated targeted by natural selection particularly in regions of the World where infections with a particular pathogen such as *Mycobacterium tuberculosis* are more common.

The polymorphism data can then be tested against phenotypic data of resistance/susceptibility to infections (e.g. by *M.tuberculosis*) to isolate gene variants (alleles) significantly involved in the resistance/susceptibility.

Thus, we will be able to correlate our polymorphism data with clinical data (cohorts of patients) and phenotypic data available (expression using RNAseq data, demographic distribution of a certain disease e.g. tuberculosis…). Functional investigations of these particular variants will then be setup (the experimental system will depend on the gene found).

VII- Conclusion

A big number of immune genes were included in our polymorphism analysis. It's a massive analysis in terms of number of individuals and number of genes in a reasonable time. We developed a specialized approach to assess and select SNPs affecting mainly protein products, useful even with 'difficult' genes e.g. multigenic families. The polymorphism analysis in the case of multigenic families requires specific approaches when using NGS data.

A tighter integration between population genetic studies and immunological phenotype studies is now necessary to reveal the mechanisms that have been crucial for our past and present survival against infection.

	547	715	% variation
SIRPA	18.58	76.31	311
CD52	16.30	32.61	100
SLC25A5	21.28	42.55	100
DEFB108B	31.82	59.09	86
SFTPA1	19.06	33.04	73
SIRPB1	20.97	36.07	72
DEFB126	17.96	29.94	67
PSG1	26.54	42.94	62
IGLL1	28.17	45.38	61
PSG5	24.93	39.88	60
PSG2	25.92	40.88	58
DEFB1	19.51	29.27	50
ORM2	23.33	35.00	50
SLC35G5	20.67	30.51	48
SIGLEC12	21.35	31.46	47
PSG11	18.94	27.92	47
P2RX7	18.03	26.48	47
IL37	16.87	24.54	45
PSG8	19.59	28.21	44
SLC35G4	54.13	77.76	44
DEFB127	16.78	23.49	40
KIR3DL3	31.02	42.45	37
OBP2A	29.59	39.45	33
GYPB	44.28	59.04	33
DEFB132	31.47	41.96	33
FUT3	23.04	29.49	28
ABO	24.67	31.31	27
CCL4	28.99	36.23	25
FUT2	34.20	42.68	25
CD207	21.41	25.48	19
C4B	33.25	39.47	19
DEFA3	21.20	24.73	17
C1GALT1	25.10	28.47	13

Supplementary table 1- Percentage of variation calculated after passing from 547 to 715 individuals (normalized variation).

Chapter 6: Discussion and Conclusion

Major technical hurdles

a- High polymorphism and paralogy of genes

Mapping reads is hampered by major technical problems when the gene is a member of a multigene family. These genes are known to have high similarities. Paralogues (genes that derive from the same ancestral gene and now reside at different locations within the same genome) increase the chance that a read could map to two or more regions and thus leading it to be eliminated. Discarding reads decreases the amount of usable data and occurs also in the case of highly diverged reads (by exceeding the number of mismatches allowed).

Meaningful analysis of NGS data relies crucially on the accurate calling of SNPs and genotypes. Therefore, proposing new adapted methodologies to deal with these problems is a must.

That's why we developed a specialized *in silico* approach useful even with "difficult" genes and multigenic families which may contain highly similar genes and/or pseudogenes.

Why we didn't use the available data generated by the 1000 Genomes project?

Development of Next Generation Sequencing (NGS) approaches and other technologies have continuously been improving. In open resources and databases, we have available, population-level, sequenced data in humans. Complete genomes and exome sequences represent a powerful source to define the human genetic diversity worldwide. The available polymorphism data haven't been properly analyzed for complex immune genes and gene families. Thus, NGS techniques are known to be problematic when applied to highly polymorphic genomic regions, such as the Human Leucocyte antigen *HLA* genes.

Accurate genotype calls and allele frequency estimations are crucial to population genomics analyses. Thus, it is important to assess the reliability of NGS data.

The human leucocyte antigen *HLA* classes I and II loci is a well-documented example of highly polymorphic multigene families. Brandt el al. evaluated the reliability of *HLA* SNP calls (*HLA-A, -B, -C, -DRB1* and *–DQB1*) and frequency estimates published by the phase I of the 1000 Genome project. After comparing the data generated by this project with *HLA* sequence-based typing dataset (PAG2014, Gourraud et al, 2014), they documented a bias toward overestimation and an important error in frequency estimation in the 1000G dataset. 18.6% of SNP calls are incorrect and the allele frequencies are estimated with an error greater than ±0.1 at approximately 25% of the SNPs in *HLA* genes[104].

b- Repeats and ambiguity in read mapping

Repeats present a considerable technical challenge for sequence alignment and assembly, especially when we have increased similarities between the different copies (tandem repeats), producing errors and biases when interpreting the results. Ignoring these repeats is not an option, because important biological information can be missed.[105] For example, we chose to discard the *CD209* gene from our list of genes controlling infection by *M.tuberculosis* (chapter 4). CD209 often referred to as DC-SIGN is a pathogen-recognition receptor expressed on the surface of dendritic cells (DCs) and macrophages. It is involved in the innate immune response and recognizes different pathogens ranging from parasites to viruses, and thought to mediate the endocytosis of pathogens which are subsequently degraded in lysosomal compartments. DC-SIGN is the major *M.tuberculosis* receptor on human DCs[106]. The gene is very closely related to L-SIGN, a neighboring gene (sequences and function). The *CD209* exon 4 corresponding to the neck region of the protein is composed of

seven coding repeats[107]. The high nucleotide identity between these repeats is a major technical problem and constitutes one of the limitations in our study.

Where are the alleles?

An allele is defined as the sequence determined by a combination of SNPs within a given gene. To date, several databases provide huge amounts of information about variants. Sequences variations exist at defined positions within the genome and are available online (mainly SNPs). However, we can easily reconstruct the allele sequences for further functional studies but according to our knowledge, there is no available detailed database providing information at the allele level. Ideally, we should have a database of the genes and their alleles, especially that the distribution of such relevant genetic variations and systematic analyses of polymorphism at the population level for functional studies are becoming increasingly useful.

An interesting polymorphism could be correlated with a certain disease. Major cases reported in the literature consider a specific variation in gene sequence (SNP) and correlate that with a specific disease (genetic, infectious, autoimmune diseases etc.). Correlating these variations at the phenotypic level is much more useful for functional analysis when dealing with alleles rather than a unique SNP. Sequences of different alleles of the gene should be published in a certain database like in the case of some immune genes such as *MHC, LILR, HPA* etc. This could be very much interesting especially in the case of genes shown to be highly polymorphic both at the allelic and haplotypic levels.

Researchers take the reference sequence available in the online databases to do functional tests (e.g. producing recombinant proteins). In the case of polymorphic genes, we should check whether the available sequence corresponds to the most frequent variant among the

different human populations or not, to prevent working with rare forms. With this preselection based on the distribution of these genetic variants, we could prevent such inconvenient. It is very important to reconstruct the allelic sequences especially when we're searching for coding variants (that affects the protein sequence, ligand-receptor recognition, amino acids interactions etc.). This could help in finding whether an interesting variation is correlated with another one, and whether the phenotypic effect requires both variations.

Large scale polymorphism analysis

Concerning the polymorphism analysis at a large scale, we know that our list only regroup 1760 genes and we're sure that we missed some interesting genes that were not initially included. On the other side, we only focused on the coding regions. We should clearly mention that genetic variations could be found also in gene *regulatory elements* (or untranslated regions, or non-coding regions etc.) and contribute to diverse human diseases. Assessing the SNPs affecting mainly protein products was our priority, as well as adapting our approach to large scale analyses.

We also focused on non-synonymous mutations that alter the amino acid sequence of a protein. We should mention that silent mutations (synonymous variations that do not affect the amino acids) could also be associated with diseases. For example, the silent mutation in exon 9 of the *VDR* gene also known as Taq1 polymorphism (rs731236) might be a risk factor for lung cancer[108]. A potential association with multiple sclerosis was reported[109] and also a protection from breast cancer[110].

The very well characterized *HLA* genes ranked as expected on the top of our list. This confirms that the HLA system is known to be the most polymorphic in humans, probably

because of the great pathogenic pressure and the wide-range of presented antigens (either self or non-self).

All the screened individuals from the different sequencing projects were healthy persons. Once we reconstruct the alleles of a particular interesting gene, we will be able to correlate our polymorphism data with clinical data (cohorts of patients infected by a certain pathogen) and phenotypic data available (demographic distribution of that particular pathogen etc.). Further analyses could be also done with a complete panel regrouping all worldwide populations for which full genomes or exomes are available.

We focused on the top most polymorphic genes of our list, but this does not mean that the others are not interesting, even those that are moderately polymorphic. (Each variation could be interesting).

Checking whether a position is conserved across orthologues could be useful because it can be involved in many aspects of structure and function, including specificity.

Other way to identify novel gene candidates involved in host-pathogen interactions

Convergence analysis

The main objective of this work was to identify novel gene candidates involved in host-pathogen interactions that can explain the resistance/susceptibility to infectious diseases. Our first approach was to identify the most variable genes by large-scale in silico screens. Here we discuss about convergence analysis, another way to identify these novel genes.

To describe this strategy, and to give a concrete example that illustrates our idea, it's interesting to check how Parker *et al* showed that echolocation appeared in two different lineages in an independent manner (Convergent evolution in echolocating mammals

published in Nature, 2013)[111]. Coding sequences that were compared across different mammals revealed signatures consistent with convergence in approximately 200 loci. Strong and significant support for convergence among bats and bottlenose dolphin (two echolocating mammals) was seen in numerous genes linked to hearing or deafness, consistent with an involvement in echolocation.

Knowing that tuberculosis is a widespread disease in mammals, and after screening the literature, we distinguished 3 different groups of mammals: susceptible, intermediate and resistant to TB, with a non-classic distribution: different species from distinct taxonomic groups have the same phenotype.

Genomes and proteomes of some mammals are available, published and freely accessible in databases. We propose to build a mammal-wide alignment of orthologous CDSs (Coding DNA Sequence) of genes identified in our *Polymorphism analysis* between these 3 groups, to see whether we can find common genetic bases or signatures consistent with convergence and thus involvement in resistance/susceptibility to TB (specific pattern between groups).

The separation between these groups was based on a screening of available information in the literature[112–114], but this information was incomplete or even contradictory between different papers. For example, we don't know whether the same species of *Mycobacterium tuberculosis* were used (or same strain), or if variability exists between individuals of the same species. Some old papers were only based on observation in veterinary clinics or physiopathological observations without molecular biology techniques, without giving sufficient information about the inoculation test, the infectious dose etc. For example, *Loxodonta spp.* or African elephants are considered to be resistant while Asian elephants (*Elephas maximus*) are considered as susceptible to tuberculosis. This difference is due to

the fact that Asian elephants are more exposed than the African ones because they are more in contact with humans. Animals in captivity are more probably exposed to *Mycobacterium tuberculosis,* the disease can be introduced to the animals from infected persons. TB does not appear naturally in most free-living mammals but it affects captive wildlife species and animals in zoological collections. (An overview of *Mycobacterium tuberculosis* infections in mammals is available in Annex 5)

This powerful tool (convergence analysis) to make phenotype/genotype associations for phenotypes that have unusual distributions requires robust phenotypic data. The exposition to the tubercle bacilli, the methods of infection, investigation and diagnostic should be taken into consideration when assessing susceptibility/resistance to infectious diseases.

General conclusion

Variations in coding regions of immune genes are of great interest especially for their effects on phenotype such as altering protein structure, disease mapping and other applications. Distinguishing SNPs with a high-level of authenticity from artifacts is one of the difficult challenges that polymorphism detection technologies are facing. Accurate genotype calls and allele frequency estimations are crucial to population genomics analyses. Thus, it is important to assess the reliability of NGS data available in databases. We developed a specific pipeline for polymorphism detection with a great ability to sort out highly similar reads from paralogous genes and/or pseudogenes. Thus, our methodology was validated and it can be used to screen all the immune genes (more than 5000 genes) and isolate those that were targeted by natural selection.

ANNEX

Annex 1- populations which are part of the 1000 Genomes Project

There are 26 different populations which are part of the 1000 Genomes Project from many different locations around the globe. The following table lists these populations:

Population Code	Population Description	Super Population Code
CHB	Han Chinese in Bejing, China	EAS
JPT	Japanese in Tokyo, Japan	EAS
CHS	Southern Han Chinese	EAS
CDX	Chinese Dai in Xishuangbanna, China	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian Population in Spain	EUR
YRI	Yoruba in Ibadan, Nigeria	AFR
LWK	Luhya in Webuye, Kenya	AFR
GWD	Gambian in Western Divisions in the Gambia	AFR
MSL	Mende in Sierra Leone	AFR
ESN	Esan in Nigeria	AFR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Caribbeans in Barbados	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
PEL	Peruvians from Lima, Peru	AMR
GIH	Gujarati Indian from Houston, Texas	SAS
PJL	Punjabi from Lahore, Pakistan	SAS
BEB	Bengali from Bangladesh	SAS
STU	Sri Lankan Tamil from the UK	SAS
ITU	Indian Telugu from the UK	SAS

These populations have been divided into 5 super populations:

- **AFR**, African
- **AMR**, Ad Mixed American
- **EAS**, East Asian
- **EUR**, European
- **SAS**, South Asian

Annex 2- Complete list of the genes with their normalized number of variations

HLAA	100.0	*TREML4*	28.4	*MYLK*	17.7	*C4A*	12.3
HLAC	97.9	*FUT2*	28.2	*SIGLEC11*	17.3	*SLC25A6*	12.3
HLADQB1	94.9	*SLC25A38*	27.5	*SLCO1B3*	17.2	*IL37*	12.3
HLADRB5	91.8	*FTH1*	27.4	*PSG1*	17.2	*CP*	12.3
HLAB	88.7	*PSG5*	26.9	*CUBN*	17.0	*SLC2A3*	12.2
HLADRB4	79.2	*ITLN1*	26.7	*DEFB130*	16.8	*LGALS8*	12.1
HLADQA1	64.1	*TRIM64C*	26.7	*CD8B*	16.5	*LILRB4*	12.0
HLADRB1	62.9	*MXRA5*	26.3	*KIR2DL1*	16.4	*DEFB126*	12.0
GGT1	59.5	*MICA*	26.2	*MS4A6A*	16.2	*SLC9A7*	12.0
PKD1	59.5	*KIR3DL3*	26.1	*KIR3DL2*	16.2	*LILRB5*	11.9
SLC29A4	58.1	*HRNR*	26.1	*IGLON5*	16.0	*FCGR2B*	11.9
HLADPB1	55.7	*LILRB3*	26.0	*SLC7A5*	15.9	*OBP2A*	11.8
SLC25A5	54.9	*HLAF*	25.7	*TREML2*	15.6	*TRIM49B*	11.8
SLC6A8	51.7	*MPZL1*	24.9	*KIR3DS1*	15.6	*APOL1*	11.7
IGSF3	49.3	*MR1*	24.5	*FCRL6*	15.4	*XCL2*	11.7
HLADRB3	49.1	*LILRA1*	24.0	*BCL2A1*	15.2	*PSME2*	11.3
IGLL1	46.9	*BTNL2*	23.5	*ABO*	15.2	*P2RX7*	11.3
DEFB108B	45.5	*LILRA3*	22.8	*SLC25A52*	15.2	*GYPA*	11.2
SLC35G4	43.3	*CD177*	22.2	*SLC6A14*	15.1	*PSG4*	11.2
FOLR3	40.9	*KIR2DL2*	22.1	*PSG2*	15.0	*CD99*	10.9
SLC25A53	40.1	*KIR2DS4*	22.1	*IL6ST*	15.0	*CD52*	10.9
TRIM48	37.3	*KIR3DL1*	21.9	*BTN3A3*	14.9	*CCL4*	10.9
SLC9B1	37.1	*NOS2*	21.8	*CXCR2*	14.8	*S100A5*	10.8
SLC31A1	36.9	*C4B*	21.8	*SLC35G5*	14.8	*IL32*	10.7
KIR2DS3	35.3	*SLC16A6*	21.7	*DEFB1*	14.6	*CD300C*	10.4
ADGRF5	34.8	*DEFB132*	21.0	*SLC25A15*	14.4	*CFHR4*	10.4
FUT3	34.1	*XG*	20.8	*SLC2A14*	14.4	*DEFB4B*	10.4
DEFB131	33.2	*BMPR1A*	20.8	*DEFA3*	14.1	*SIRPG*	10.4
LCN1	33.1	*FOLR1*	20.8	*FCRL5*	14.1	*AK3*	10.3
FTL	32.4	*LILRA2*	20.8	*MS4A4A*	14.0	*HLADQB2*	10.3
KLRC2	31.9	*HLAG*	20.8	*MICB*	14.0	*FANCD2*	10.3
IGLL5	31.2	*SLC25A51*	20.2	*LILRB2*	13.5	*CD207*	10.2
KIR2DS2	30.9	*LILRA6*	20.1	*LILRA4*	13.4	*SFTPA1*	10.2
SIRPA	30.5	*C1GALT1*	20.1	*PSG8*	13.3	*ENPEP*	10.2
HAVCR1	30.4	*VTCN1*	20.1	*BTN2A1*	13.3	*LGALS2*	10.1
KIR2DS5	29.8	*CD33*	19.3	*LGALS9*	13.2	*DEFB127*	10.1
TRIM77	29.7	*ALOX15*	19.2	*CLEC1B*	13.2	*SIRPB1*	10.1
TRIM51	29.6	*TYRO3*	19.2	*HLADQA2*	13.1	*BTNL8*	10.0
HLADPA1	29.5	*S100A11*	19.0	*CFHR3*	13.0	*BTN3A2*	10.0
GYPB	29.5	*AZGP1*	19.0	*PSG11*	13.0	*ORM2*	10.0
KIR2DL3	29.5	*EMB*	18.5	*SIGLEC12*	12.9	*DEFA6*	10.0
CXADR	29.3	*CLEC4G*	18.3	*TRIM61*	12.7	*NBEA*	9.8
TRIM43	29.2	*ATP1B3*	18.0	*TRPC6*	12.6	*DEFB114*	9.6
IL9R	28.9	*LILRB1*	18.0	*FH*	12.5	*TRPA1*	9.6

SLC6A18	9.6	*SLC36A3*	7.8	*EDAR*	6.7	*CLEC4D*	6.2
LRP5	9.5	*GYPC*	7.8	*CCL8*	6.7	*CFHR2*	6.2
CLC	9.4	*TNFRSF10D*	7.8	*MBL2*	6.7	*SLC38A8*	6.2
PSG3	9.4	*FCRL4*	7.8	*SFTPA2*	6.7	*CD1A*	6.1
DEFB124	9.3	*ADGRE2*	7.7	*HLADOA*	6.7	*IL17F*	6.1
HAPLN3	9.3	*KIR2DS1*	7.7	*ACKR4*	6.7	*SLC25A26*	6.1
CD48	9.3	*CD200R1*	7.7	*CXCR1*	6.7	*PDCD1LG2*	6.1
DEFB104A	9.2	*IL4R*	7.7	*SLC22A1*	6.7	*CXCL16*	6.1
FCGR3A	9.2	*LGALS9C*	7.5	*C8G*	6.6	*CLEC12B*	6.1
CCL14	9.2	*CD276*	7.5	*AREG*	6.6	*B3GALNT1*	6.0
SLC29A3	9.1	*NAMPT*	7.5	*IL1RL1*	6.6	*CD1E*	6.0
APOL3	9.1	*LAIR1*	7.5	*PLA1A*	6.6	*SLC14A1*	6.0
ADGRA3	9.1	*GJB7*	7.5	*SLC45A4*	6.6	*ESAM*	6.0
CFHR1	9.1	*IGSF5*	7.4	*IL1F10*	6.6	*CRTAM*	6.0
TNFRSF10C	9.0	*TLR10*	7.4	*S100A3*	6.6	*ZPBP2*	5.9
LILRA5	9.0	*ACAN*	7.4	*S100A4*	6.6	*TRIM6*	5.9
CEACAM18	8.9	*HLADOB*	7.4	*S100A7*	6.6	*TMIGD2*	5.9
CCL25	8.9	*CACYBP*	7.3	*HLADRA*	6.6	*CCL15*	5.9
TRIM64B	8.9	*SLC22A14*	7.3	*SLC16A14*	6.5	*OBP2B*	5.9
SIGLEC6	8.9	*DEFB129*	7.3	*TMEM81*	6.5	*TRIM49C*	5.9
LCN8	8.8	*NCR2*	7.3	*DEFB116*	6.5	*SDK1*	5.9
RHD	8.8	*CCL3*	7.2	*SLC25A47*	6.5	*CLEC18B*	5.9
GJB2	8.8	*ALPK2*	7.2	*SLC39A2*	6.5	*PILRB*	5.9
LAIR2	8.8	*LGALS14*	7.2	*CR1L*	6.5	*F5*	5.9
SLC19A3	8.7	*TRPV3*	7.2	*TRIM40*	6.5	*IL7R*	5.8
SIGLEC8	8.7	*CCL22*	7.2	*TRPV6*	6.5	*LRIG1*	5.8
TRIM31	8.7	*TAP2*	7.2	*TRIM38*	6.5	*IL22RA1*	5.8
SIGLEC10	8.6	*CEACAM5*	7.1	*SELPLG*	6.5	*SLCO1B1*	5.8
GP6	8.6	*KIR2DL4*	7.1	*PFKP*	6.4	*TAP1*	5.8
GJB3	8.6	*CD320*	7.1	*TAPBPL*	6.4	*ALK*	5.8
DEFB105A	8.5	*MYOM2*	7.1	*MEFV*	6.4	*TIAF1*	5.8
SLC52A3	8.5	*TIRAP*	7.1	*EPCAM*	6.4	*KLRC1*	5.7
LTF	8.5	*CFH*	7.1	*SLC7A13*	6.4	*BTNL3*	5.7
SIGLEC5	8.5	*DEFA5*	7.1	*HLADMB*	6.4	*C8A*	5.7
PSG6	8.4	*TCN2*	7.1	*IL36A*	6.3	*SIGIRR*	5.7
TRIM17	8.4	*COLEC11*	7.1	*FCGR2A*	6.3	*SLC25A14*	5.7
CCL16	8.3	*AKAP3*	7.0	*ALCAM*	6.3	*MERTK*	5.7
SLC36A2	8.3	*BTN2A2*	7.0	*SLC6A5*	6.3	*VSTM1*	5.7
NCR3	8.3	*CLEC2D*	7.0	*ABCB1*	6.3	*IL17RE*	5.7
CD24	8.3	*S100P*	7.0	*ADORA3*	6.3	*IL1RN*	5.7
SLC28A1	8.3	*VPREB1*	6.9	*VSTM4*	6.3	*HHLA2*	5.7
SLAMF9	8.1	*CCL11*	6.9	*CD82*	6.3	*BTNL9*	5.6
LGALS3	8.0	*SLC39A11*	6.9	*SLC25A33*	6.3	*CEACAM20*	5.6
SLC35G6	7.9	*SLC10A5*	6.8	*CXCL1*	6.3	*CXCL17*	5.6
MADCAM1	7.9	*CCL13*	6.8	*SFTPD*	6.2	*CCL24*	5.6
PSG9	7.8	*CD5*	6.8	*CLEC4F*	6.2	*SLC2A9*	5.6

CCL23	5.6	*CD8A*	5.1	*SLC7A4*	4.7	*TF*	4.3
TNFRSF6B	5.6	*LY75*	5.1	*SLC17A4*	4.7	*SLC24A4*	4.3
SLC25A3	5.5	*TRIM15*	5.0	*SLC16A13*	4.7	*DEFB135*	4.3
SLC22A18	5.5	*DEFB134*	5.0	*RELT*	4.7	*IL36RN*	4.3
FLT1	5.5	*SLC16A3*	5.0	*UNC5C*	4.7	*CHN2*	4.3
SLC7A9	5.5	*CD6*	5.0	*MYOM3*	4.7	*TNFRSF10A*	4.3
TLR1	5.5	*SLC25A23*	5.0	*FCAR*	4.7	*PAPLN*	4.3
SLC22A25	5.5	*FAS*	5.0	*PDCD1*	4.6	*UCP3*	4.3
SELE	5.5	*GJB4*	5.0	*SLC2A6*	4.6	*TREM1*	4.3
LCN15	5.5	*ORM1*	5.0	*UNC5B*	4.6	*GYPE*	4.3
LRIT2	5.5	*SDC2*	5.0	*TRIM45*	4.6	*MPZL3*	4.3
SLC9A4	5.5	*SCN1B*	5.0	*CD93*	4.6	*DEFB125*	4.3
SLC2A1	5.4	*FCGR3B*	5.0	*TRPV5*	4.6	*SLC25A48*	4.3
TNFRSF17	5.4	*ROR2*	5.0	*CLEC10A*	4.6	*SLC6A19*	4.2
SLC5A12	5.4	*SLC16A11*	5.0	*HSPG2*	4.6	*TNFRSF14*	4.2
TRIM5	5.4	*VSIG10*	5.0	*PSMB9*	4.6	*SLC46A2*	4.2
SLC5A9	5.4	*DEFB110*	5.0	*SLC7A7*	4.6	*TOR3A*	4.2
DEFB118	5.4	*DEFB123*	5.0	*IL13*	4.6	*TGFA*	4.2
SCARA5	5.4	*IL2RA*	4.9	*CEACAM21*	4.6	*TLR6*	4.2
VPREB3	5.4	*DHCR7*	4.9	*FCRL3*	4.6	*TNFSF14*	4.2
ASGR2	5.4	*ADGRE1*	4.9	*MYPN*	4.6	*TRIM10*	4.2
SLC1A7	5.4	*LRIT3*	4.9	*ITGAE*	4.6	*CLEC9A*	4.2
C10orf54	5.4	*CD101*	4.9	*SLC6A20*	4.5	*IL34*	4.1
TRIM22	5.4	*CD300E*	4.9	*BOC*	4.5	*IL25*	4.1
OSCAR	5.4	*TRIM16*	4.9	*ANPEP*	4.5	*SLC39A4*	4.1
FCGR1A	5.4	*IL24*	4.9	*PSG7*	4.5	*CD209*	4.1
SEMA3C	5.4	*SLC22A24*	4.9	*SLC52A1*	4.5	*CEACAM4*	4.1
LRIT1	5.4	*CSF2RB*	4.8	*TRIM64*	4.5	*F13A1*	4.1
IL17RB	5.3	*SLC9C1*	4.8	*PDGFRL*	4.5	*SIGLEC1*	4.1
MOG	5.3	*SLC39A12*	4.8	*C3*	4.4	*TRIM34*	4.1
SLC7A1	5.3	*OAS1*	4.8	*SELP*	4.4	*KLRG2*	4.1
CEACAM3	5.3	*MRC1*	4.8	*TRIM49*	4.4	*SIGLEC15*	4.1
ART4	5.3	*CCRL2*	4.8	*TPST2*	4.4	*IL1RL2*	4.1
SLC5A6	5.3	*SCARF1*	4.8	*SLC18B1*	4.4	*IL31*	4.1
BST1	5.3	*TRIM58*	4.8	*JAG1*	4.4	*SLC22A16*	4.1
FCRLA	5.2	*PXDNL*	4.8	*SLC25A34*	4.4	*WFIKKN2*	4.0
SLC24A3	5.2	*SLC10A2*	4.8	*CD300LF*	4.4	*SYNM*	4.0
CDON	5.2	*CEACAM8*	4.8	*ROBO2*	4.4	*SLC15A5*	4.0
DEFB4A	5.2	*SLC6A11*	4.8	*SLC28A3*	4.4	*SLC5A1*	4.0
RHBG	5.1	*FCGR1B*	4.8	*IL4*	4.4	*TNFSF12*	4.0
SLC18A1	5.1	*IGFN1*	4.8	*MC1R*	4.4	*SLC9A3*	4.0
MC5R	5.1	*SLC39A14*	4.8	*ITGB2*	4.4	*TNFSF13*	4.0
ART1	5.1	*IL15RA*	4.7	*SELL*	4.4	*TRIM69*	4.0
SIRPD	5.1	*DEFB107B*	4.7	*HS3ST1*	4.3	*OAS3*	4.0
IL12RB1	5.1	*SLX4*	4.7	*SIGLEC9*	4.3	*SLCO1A2*	4.0
LDLR	5.1	*MYBPHL*	4.7	*SLC22A10*	4.3	*SLC37A2*	4.0

PLAUR	4.0	*HAPLN1*	3.8	*PTGDS*	3.5	*CDH2*	3.3
SLAMF7	4.0	*GP9*	3.8	*SLAMF8*	3.5	*ROBO4*	3.3
IRF7	4.0	*PKD2L1*	3.7	*MYRIP*	3.5	*PILRA*	3.3
MYOM1	4.0	*TRPM5*	3.7	*PDCD6*	3.5	*SLC25A2*	3.3
ADAMTSL3	4.0	*AKAP12*	3.7	*SFTPC*	3.5	*SLC4A1*	3.3
DSCAML1	4.0	*HLAE*	3.7	*TGFBRAP1*	3.5	*SLC16A8*	3.3
ITGA11	4.0	*PCBP1*	3.7	*CCL20*	3.5	*SLC41A3*	3.3
SLC22A31	4.0	*LRRN1*	3.7	*SLC5A11*	3.5	*TRIM59*	3.3
TNFRSF19	4.0	*TRIM26*	3.7	*LRP3*	3.5	*CLEC3B*	3.3
SLC25A37	3.9	*CD200R1L*	3.7	*BTLA*	3.5	*SLC12A9*	3.3
SLC51A	3.9	*LRP4*	3.7	*FANCA*	3.5	*S100A7A*	3.3
TNFRSF9	3.9	*CD3G*	3.7	*IL10RA*	3.5	*SLC9C2*	3.3
FLT4	3.9	*ADAM8*	3.7	*LRP2*	3.5	*PALLD*	3.3
ACKR1	3.9	*ICAM3*	3.7	*CD163L1*	3.5	*CMYA5*	3.3
P2RY14	3.9	*SLC11A1*	3.7	*CD300LD*	3.4	*IGSF8*	3.3
SLC35G3	3.9	*ELP2*	3.7	*PTPRS*	3.4	*CD83*	3.3
SLC2A7	3.9	*SLC4A11*	3.7	*CHI3L2*	3.4	*BTN3A1*	3.3
TRIM7	3.9	*CD40*	3.6	*PTPRM*	3.4	*LTA*	3.3
CNTN4	3.9	*NRG1*	3.6	*EZR*	3.4	*CD3E*	3.2
FCRLB	3.9	*SSC4D*	3.6	*F10*	3.4	*UCP2*	3.2
AKAP13	3.9	*NRP1*	3.6	*DEFA4*	3.4	*SLC6A3*	3.2
FCRL1	3.9	*PROC*	3.6	*S100A10*	3.4	*CLEC3A*	3.2
AKAP5	3.9	*COLEC10*	3.6	*S100A2*	3.4	*CR1*	3.2
SEMA3E	3.9	*HEXB*	3.6	*TNFRSF13B*	3.4	*SDC1*	3.2
CEACAM6	3.9	*S100B*	3.6	*CD200*	3.4	*TRIM65*	3.2
IGSF11	3.9	*CLEC19A*	3.6	*SFTPB*	3.4	*SLC15A2*	3.2
SLC26A11	3.9	*TRPV1*	3.6	*SLC26A9*	3.4	*F11*	3.2
BIRC3	3.9	*SLC12A3*	3.6	*GC*	3.4	*HEPACAM*	3.2
OTOF	3.9	*CLEC1A*	3.6	*ITGB3*	3.4	*TNFRSF25*	3.2
IL17RA	3.9	*CCL3L3*	3.6	*ITGB6*	3.4	*PVR*	3.2
SLC5A8	3.9	*CD37*	3.6	*F3*	3.4	*BCAM*	3.2
HLADMA	3.8	*TEK*	3.6	*SDK2*	3.4	*SLC25A46*	3.2
TLR2	3.8	*DEFB128*	3.6	*LY6G6F*	3.4	*CD163*	3.2
IL17RC	3.8	*S100A8*	3.6	*SLC5A10*	3.4	*IL11RA*	3.2
CEACAM1	3.8	*GP5*	3.6	*CCL2*	3.4	*MFI2*	3.2
BTN1A1	3.8	*TNFSF10*	3.6	*IL11*	3.4	*PECAM1*	3.2
C5AR1	3.8	*XAF1*	3.6	*SLC2A11*	3.4	*STAB2*	3.2
SLC26A1	3.8	*SPHK1*	3.6	*AKAP17A*	3.4	*SLC6A7*	3.2
NTRK1	3.8	*CCL17*	3.5	*SLC7A2*	3.4	*PTPRF*	3.2
SLC37A1	3.8	*TIMD4*	3.5	*S100Z*	3.4	*EDN1*	3.2
NFASC	3.8	*SLC24A2*	3.5	*SLC34A3*	3.4	*IL6*	3.2
KIRREL2	3.8	*KLRG1*	3.5	*CLEC4M*	3.4	*SLC35A5*	3.2
IKBKAP	3.8	*IGSF22*	3.5	*FGFR4*	3.3	*INSR*	3.1
CD68	3.8	*DEFA1*	3.5	*AKAP1*	3.3	*TGFBR3*	3.1
ITGB5	3.8	*STAB1*	3.5	*TNFRSF11B*	3.3	*ICAM1*	3.1
PFKFB3	3.8	*TRPM3*	3.5	*SPN*	3.3	*FCER2*	3.1

SLC7A8	3.1	*SLC30A9*	3.0	*MFAP3*	2.8	*SLC38A6*	2.6
SLC45A1	3.1	*CLEC18C*	3.0	*IL27*	2.8	*CX3CR1*	2.6
CLEC2L	3.1	*CXCL5*	2.9	*TMEM25*	2.7	*CA2*	2.6
CERK	3.1	*IGSF10*	2.9	*SLC6A12*	2.7	*CD27*	2.6
SLC30A4	3.1	*CD4*	2.9	*SLC35F3*	2.7	*NOD2*	2.6
MCAM	3.1	*S100A9*	2.9	*PDGFRB*	2.7	*SLC7A10*	2.6
IL19	3.1	*CORO1A*	2.9	*HS3ST2*	2.7	*SLC2A2*	2.6
SLC25A22	3.1	*TREM2*	2.9	*SLC4A3*	2.7	*VCAN*	2.6
SLC25A27	3.1	*CNTN2*	2.9	*SLC37A3*	2.7	*PTPRK*	2.6
ISLR	3.1	*TRPM2*	2.9	*IL17RD*	2.7	*DDR1*	2.6
KLRK1	3.1	*CD302*	2.9	*TLR5*	2.7	*AMICA1*	2.6
PITPNM3	3.1	*TLR8*	2.9	*CD244*	2.7	*AMIGO2*	2.6
SLC12A7	3.1	*IGDCC3*	2.9	*SLC25A41*	2.7	*SLC14A2*	2.6
SLC23A2	3.1	*ITGAL*	2.9	*CLEC7A*	2.7	*FLVCR2*	2.5
SLC25A29	3.1	*SLC10A1*	2.9	*SLC39A13*	2.7	*IL22RA2*	2.5
CSF2RA	3.1	*SLC44A2*	2.9	*SLC17A5*	2.7	*PIK3IP1*	2.5
SLC1A5	3.1	*SLC35E4*	2.9	*FN1*	2.7	*SIGLEC14*	2.5
ADGRE3	3.1	*ROR1*	2.9	*SLC7A3*	2.7	*PTGDR2*	2.5
SLC39A1	3.1	*SLC15A1*	2.9	*COLEC12*	2.7	*IL1R2*	2.5
SLC2A10	3.1	*CD81*	2.8	*ACHE*	2.7	*CEACAM7*	2.5
SLC17A9	3.1	*ITGA2*	2.8	*SLC17A3*	2.7	*ITGB1*	2.5
SLC22A15	3.1	*TRIM29*	2.8	*TSPAN7*	2.7	*SLC35E2*	2.5
PIGR	3.1	*SLC44A4*	2.8	*SLC38A10*	2.7	*SLC18A3*	2.5
SLC28A2	3.1	*LGALS9B*	2.8	*ATF3*	2.7	*FANCE*	2.5
ENG	3.1	*CCR5*	2.8	*CXCR5*	2.7	*SLC9B2*	2.5
SLC33A1	3.0	*NEGR1*	2.8	*CCR2*	2.7	*IL5*	2.5
F13B	3.0	*SLCO2B1*	2.8	*F2RL2*	2.7	*AGER*	2.5
TRPM8	3.0	*C8B*	2.8	*MAG*	2.7	*GCNT2*	2.5
SLC22A4	3.0	*NTM*	2.8	*TRIM25*	2.7	*IL1B*	2.5
SLC22A12	3.0	*SLC37A4*	2.8	*TAPBP*	2.7	*BRCA1*	2.5
CD180	3.0	*CLEC4A*	2.8	*CFB*	2.7	*SLC25A13*	2.5
CD300LG	3.0	*ACE*	2.8	*TRIM56*	2.6	*AQP1*	2.5
CD109	3.0	*OBSL1*	2.8	*CFD*	2.6	*SLC25A10*	2.5
SEMA7A	3.0	*ERMAP*	2.8	*SLC26A3*	2.6	*SLC25A12*	2.5
IL16	3.0	*TNFRSF8*	2.8	*FOS*	2.6	*IL1A*	2.5
DYSF	3.0	*SLC2A8*	2.8	*C1QB*	2.6	*SLC9A1*	2.5
CD1C	3.0	*SLC44A5*	2.8	*IL31RA*	2.6	*TGFBI*	2.5
CXCL14	3.0	*NOD1*	2.8	*TRPV2*	2.6	*MDGA1*	2.5
CLEC18A	3.0	*LRRN2*	2.8	*SLC16A9*	2.6	*AKAP6*	2.4
PTPRJ	3.0	*ALPK3*	2.8	*FOLR2*	2.6	*TRIM55*	2.4
SLAMF1	3.0	*CDCP1*	2.8	*BSG*	2.6	*CNTN5*	2.4
SLC35D2	3.0	*SLCO6A1*	2.8	*SLCO2A1*	2.6	*LRP8*	2.4
CD226	3.0	*OAS2*	2.8	*C5*	2.6	*CLEC12A*	2.4
CCL27	3.0	*FANCI*	2.8	*FCER1A*	2.6	*ICAM2*	2.4
S100A7L2	3.0	*HMMR*	2.8	*CR2*	2.6	*MYO7A*	2.4
NPHS1	3.0	*TRIM42*	2.8	*MARCO*	2.6	*SLC27A5*	2.4

SLC35G2	2.4	*AQP3*	2.3	*SLC39A3*	2.1	*B3GAT1*	2.0
TNFRSF4	2.4	*MYOF*	2.3	*TRIM41*	2.1	*IGSF1*	2.0
IL20RA	2.4	*TNFRSF10B*	2.3	*LRP1B*	2.1	*TRPM6*	2.0
CSF1R	2.4	*CDH1*	2.3	*TRIM4*	2.1	*LRP1*	2.0
LMBRD2	2.4	*ITGA9*	2.3	*LRFN2*	2.1	*C7*	2.0
ADGRE5	2.4	*PXDN*	2.3	*TRPC1*	2.1	*IGSF9*	2.0
RHCE	2.4	*F7*	2.3	*MPL*	2.1	*FGFRL1*	2.0
LAMP1	2.4	*SLC25A21*	2.2	*IL27RA*	2.1	*SEMA4G*	2.0
KIT	2.4	*TRPM1*	2.2	*TRIM62*	2.1	*IL36G*	2.0
SLC22A5	2.4	*CEACAM19*	2.2	*TNFSF11*	2.1	*SCARB1*	2.0
SLC38A9	2.4	*CD300A*	2.2	*MYBPH*	2.1	*CD22*	2.0
SLC26A10	2.4	*SLC6A1*	2.2	*SLC25A24*	2.1	*PVRL4*	2.0
C1R	2.4	*SLC6A13*	2.2	*VDR*	2.1	*CNTN3*	2.0
SLC1A6	2.4	*SLC13A3*	2.2	*EDA2R*	2.1	*CNTN6*	2.0
MMACHC	2.4	*TNFRSF1A*	2.2	*LRFN4*	2.1	*IL26*	2.0
SLCO5A1	2.4	*IL3*	2.2	*RHCG*	2.1	*SIRPB2*	2.0
SLCO3A1	2.4	*NCR1*	2.2	*SLC16A7*	2.1	*SLC22A23*	2.0
MYBPC2	2.4	*TRIM66*	2.2	*ADAMTSL1*	2.1	*OASL*	1.9
CFHR5	2.4	*LTA4H*	2.2	*SLC25A19*	2.1	*CXCR6*	1.9
IL1R1	2.4	*KAZALD1*	2.2	*DCC*	2.1	*AKAP2*	1.9
LGALS16	2.4	*IL2*	2.2	*TYMP*	2.1	*ROBO3*	1.9
SLCO1C1	2.4	*SLC39A8*	2.2	*SLC9A9*	2.1	*TRIM39*	1.9
PLXNC1	2.4	*UCP1*	2.2	*TOR2A*	2.1	*PTPRD*	1.9
NGFR	2.3	*HPX*	2.2	*LGALS4*	2.1	*VSIG10L*	1.9
SLC12A8	2.3	*SLC38A7*	2.2	*SLC26A8*	2.1	*AKAP7*	1.9
SLC31A2	2.3	*ICAM5*	2.2	*IL21*	2.1	*CLEC2A*	1.9
CYBRD1	2.3	*FER1L6*	2.2	*SLC16A4*	2.1	*HFE*	1.9
IL12RB2	2.3	*ICOSLG*	2.2	*NRCAM*	2.1	*TRIM71*	1.9
ITGAM	2.3	*CTSC*	2.2	*CD247*	2.1	*VLDLR*	1.9
NT5E	2.3	*SLC7A14*	2.2	*LCN6*	2.1	*SLC1A1*	1.9
SEMA4D	2.3	*SLC6A6*	2.2	*SLC35A4*	2.1	*AMBP*	1.9
IL9	2.3	*SLC12A4*	2.2	*CBFA2T3*	2.1	*TRIM63*	1.9
SLC25A45	2.3	*SLC4A2*	2.2	*SLC12A6*	2.0	*IL7*	1.9
CD80	2.3	*SLC27A2*	2.2	*MRC2*	2.0	*SLC45A2*	1.9
TCN1	2.3	*PDGFRA*	2.2	*CFTR*	2.0	*SLC1A4*	1.9
SLC47A2	2.3	*IL17A*	2.2	*SLC5A4*	2.0	*IGF2R*	1.9
SLC15A4	2.3	*LINGO1*	2.1	*CD86*	2.0	*FUT4*	1.9
SLCO4A1	2.3	*IL6R*	2.1	*TRIM35*	2.0	*P2RY13*	1.9
SLC35B2	2.3	*SIGLEC7*	2.1	*FLT3*	2.0	*KLRD1*	1.9
LTBR	2.3	*IGDCC4*	2.1	*SLC35B4*	2.0	*SLC25A39*	1.9
MUSK	2.3	*TRIM11*	2.1	*AKAP10*	2.0	*IL22*	1.9
TRPV4	2.3	*ERBB2*	2.1	*SLC19A2*	2.0	*SLC35B1*	1.9
SLC15A3	2.3	*SLC25A32*	2.1	*LEPR*	2.0	*SLC22A17*	1.9
SLC6A17	2.3	*SLC25A31*	2.1	*LRP11*	2.0	*NFKB2*	1.9
ITGAX	2.3	*LRFN3*	2.1	*MC4R*	2.0	*IL21R*	1.9
TOR1AIP1	2.3	*ELP5*	2.1	*NTRK3*	2.0	*IGSF9B*	1.9

PAEP	1.9	*SLC9A8*	1.7	*CSF3R*	1.6	*C1S*	1.5
BRCA2	1.9	*ACKR2*	1.7	*CLEC16A*	1.6	*PRTG*	1.5
IL17B	1.9	*IL18*	1.7	*SLC3A2*	1.6	*CD79B*	1.5
TLR3	1.8	*TRIO*	1.7	*TLR4*	1.6	*SLC34A2*	1.5
SPEG	1.8	*F2RL3*	1.7	*PDGFA*	1.6	*TGFB1I1*	1.5
SLC22A7	1.8	*CD70*	1.7	*MYBPC3*	1.6	*TNFRSF1B*	1.5
ACKR3	1.8	*MTCH1*	1.7	*FSTL5*	1.6	*APC2*	1.4
FCGRT	1.8	*SLC17A6*	1.7	*ITGA8*	1.6	*KLRF1*	1.4
SLC35A3	1.8	*CFI*	1.7	*AKAP11*	1.6	*CILP2*	1.4
FUT1	1.8	*SLC8B1*	1.7	*CLEC4C*	1.6	*ITGA10*	1.4
IL2RB	1.8	*IGF1R*	1.7	*LRIG2*	1.6	*SLC17A1*	1.4
SORL1	1.8	*SEMA3F*	1.7	*TRIM47*	1.6	*C6*	1.4
SLC30A8	1.8	*ITGA1*	1.7	*AKAP4*	1.6	*MST1R*	1.4
SLC22A2	1.8	*CDH5*	1.7	*MPZL2*	1.6	*MYBPC1*	1.4
FLVCR1	1.8	*TRPC4*	1.7	*SCN3B*	1.6	*CFP*	1.4
SLC45A3	1.8	*UBE2T*	1.7	*CADM3*	1.6	*TNFSF8*	1.4
CD44	1.8	*TGFBR2*	1.7	*LSR*	1.5	*TOR1AIP2*	1.4
VCAM1	1.8	*IL17C*	1.7	*PTPRC*	1.5	*TRIM60*	1.4
CD19	1.8	*ICOS*	1.7	*SCARF2*	1.5	*TRIM9*	1.4
SLC22A3	1.8	*LIPA*	1.7	*AKAP9*	1.5	*CORO2B*	1.4
HESX1	1.8	*F2RL1*	1.7	*SLC4A8*	1.5	*GAS7*	1.4
FANCC	1.8	*VSIG4*	1.7	*ITGA6*	1.5	*CILP*	1.4
SLC26A2	1.8	*LCN10*	1.7	*ENPP3*	1.5	*CD63*	1.4
MFSD7	1.8	*SLC23A1*	1.7	*ABCG2*	1.5	*SLC36A1*	1.4
LRIG3	1.8	*CNTN1*	1.7	*LY9*	1.5	*PROCR*	1.4
KIR2DL5A	1.8	*RBP4*	1.7	*CD53*	1.5	*SLC17A2*	1.4
RAD51C	1.8	*CD300LB*	1.7	*LRRC4*	1.5	*EGR2*	1.4
CCR6	1.8	*LRP6*	1.7	*CLEC4E*	1.5	*CD7*	1.4
CLEC17A	1.8	*TRPM4*	1.7	*LIFR*	1.5	*IGSF6*	1.4
MCOLN2	1.8	*BLM*	1.7	*PTGFRN*	1.5	*TNFRSF18*	1.4
IL3RA	1.8	*SLC35E2B*	1.7	*NCAN*	1.5	*SLC4A7*	1.4
SLC10A6	1.8	*IL17D*	1.6	*CD28*	1.5	*TRIM36*	1.4
SLC13A5	1.8	*PKD2L2*	1.6	*FGL2*	1.5	*NEO1*	1.4
SLC43A2	1.8	*RHAG*	1.6	*CTLA4*	1.5	*KEL*	1.4
SPHKAP	1.8	*VSTM2L*	1.6	*KLRB1*	1.5	*PLA2R1*	1.4
IL23A	1.8	*FGFR2*	1.6	*MSR1*	1.5	*CHL1*	1.4
CD248	1.8	*LAMP2*	1.6	*KDR*	1.5	*IFNGR1*	1.4
UNC5D	1.8	*SLC35E1*	1.6	*TRIM49D1*	1.5	*SLC43A3*	1.4
SLC40A1	1.8	*TNFRSF11A*	1.6	*SLC12A5*	1.5	*C1QA*	1.4
SLC30A5	1.8	*MFAP3L*	1.6	*SLC29A1*	1.5	*C1QC*	1.4
ADGRA2	1.8	*SLC27A6*	1.6	*ITGA7*	1.5	*TRIM3*	1.3
SLC1A2	1.8	*ITGA2B*	1.6	*BCAN*	1.5	*MYOT*	1.3
TRIM23	1.8	*F2*	1.6	*SCN4B*	1.5	*ISLR2*	1.3
PTX3	1.7	*PTPRT*	1.6	*SLC3A1*	1.5	*BRIP1*	1.3
LCN12	1.7	*LAMP3*	1.6	*MAP2*	1.5	*CD58*	1.3
SEMA3A	1.7	*CLEC6A*	1.6	*ERCC4*	1.5	*SLC24A5*	1.3

SLC2A5	1.3	*PVRL3*	1.2	*SLC30A3*	1.1	*ITGA5*	1.0
SLC7A11	1.3	*WFIKKN1*	1.2	*TRPM7*	1.1	*SLC9A6*	1.0
TRIM73	1.3	*TRIM28*	1.2	*SLC25A1*	1.1	*ZPBP*	1.0
DSCAM	1.3	*C9*	1.2	*FANCG*	1.1	*TLR7*	1.0
SLC25A25	1.3	*LAMP5*	1.2	*JAM2*	1.1	*CD2*	1.0
TNFSF15	1.3	*FSTL4*	1.2	*SLC13A4*	1.1	*SLC6A9*	0.9
KIRREL	1.3	*UNC5A*	1.2	*DDRGK1*	1.1	*SLC35C1*	0.9
SLC39A6	1.3	*MUC1*	1.2	*SLC6A2*	1.1	*SLC35D1*	0.9
SLC38A2	1.3	*TRIM33*	1.2	*SLC6A4*	1.1	*CXCR4*	0.9
CD151	1.3	*FASLG*	1.2	*SPRY2*	1.1	*LRRN3*	0.9
TFRC	1.3	*TAPT1*	1.2	*SLC25A42*	1.1	*PTK7*	0.9
TRIM46	1.3	*SLC4A5*	1.2	*MDGA2*	1.1	*CCR1*	0.9
DPP4	1.3	*CD96*	1.2	*GPA33*	1.0	*CCR8*	0.9
CCDC141	1.3	*DDR2*	1.2	*SLC34A1*	1.0	*LRP10*	0.9
TNFSF9	1.3	*SLC41A2*	1.2	*SLC13A2*	1.0	*LRRC4B*	0.9
SLC18A2	1.3	*PROM1*	1.2	*SLC5A5*	1.0	*CD72*	0.9
ITGB8	1.3	*ITGAD*	1.2	*SLC27A4*	1.0	*ADK*	0.9
SLC41A1	1.3	*MCOLN1*	1.2	*SLC27A1*	1.0	*SLCO4C1*	0.9
ITGA4	1.3	*SLC25A30*	1.2	*SLC25A44*	1.0	*CCR4*	0.9
TRIM2	1.3	*SLC5A7*	1.2	*CLEC11A*	1.0	*CCR10*	0.9
LRRC24	1.3	*ASGR1*	1.2	*IRF1*	1.0	*SLC27A3*	0.9
SLC7A6	1.3	*FZD10*	1.1	*IL10RB*	1.0	*SLC24A1*	0.9
ENTPD1	1.3	*HNMT*	1.1	*MC3R*	1.0	*MID2*	0.9
SEMA3D	1.3	*FANCM*	1.1	*SLC44A3*	1.0	*SLC6A16*	0.9
IGFBP4	1.3	*SLC17A8*	1.1	*SLC20A2*	1.0	*SLC26A5*	0.9
KIRREL3	1.3	*ARFGEF2*	1.1	*SLC4A9*	1.0	*SLC30A2*	0.9
CD40LG	1.3	*SLC13A1*	1.1	*ROBO1*	1.0	*FANCL*	0.9
CORO2A	1.3	*PALB2*	1.1	*SLC26A7*	1.0	*CLMP*	0.9
ACBD3	1.3	*MS4A1*	1.1	*SLAMF6*	1.0	*SEMA3B*	0.9
ITGA3	1.3	*LINGO3*	1.1	*TOR1A*	1.0	*SLC35F2*	0.9
ITGB7	1.3	*SLC9A5*	1.1	*CD1B*	1.0	*KIR2DL5B*	0.9
FGFR3	1.2	*TRIM52*	1.1	*MID1*	1.0	*FANCF*	0.9
LMBRD1	1.2	*MC2R*	1.1	*XCR1*	1.0	*CD14*	0.9
IL33	1.2	*CD38*	1.1	*TOR1B*	1.0	*TLR9*	0.9
SLC39A5	1.2	*IL18RAP*	1.1	*LSAMP*	1.0	*CCR3*	0.9
SLC22A8	1.2	*HAVCR2*	1.1	*APOL2*	1.0	*SLC26A6*	0.9
TARM1	1.2	*MTCH2*	1.1	*SLC10A7*	1.0	*TIE1*	0.9
ELP3	1.2	*HEPH*	1.1	*C5AR2*	1.0	*CCR7*	0.9
SLC38A4	1.2	*ITGB4*	1.1	*SLC20A1*	1.0	*SEPP1*	0.9
SLC4A4	1.2	*SCARA3*	1.1	*HAPLN2*	1.0	*S1PR1*	0.9
OLR1	1.2	*SLC25A17*	1.1	*SLC25A43*	1.0	*LRFN1*	0.9
SLC12A1	1.2	*F12*	1.1	*SLC19A1*	1.0	*SLC26A4*	0.9
ADAM17	1.2	*SLC39A9*	1.1	*APOL6*	1.0	*FCMR*	0.9
GJB5	1.2	*TRPC3*	1.1	*AKAP8*	1.0	*MMP23B*	0.9
SLC22A11	1.2	*JAM3*	1.1	*CD5L*	1.0	*TGFB1*	0.9
SLC22A6	1.2	*SLC8A3*	1.1	*ITGAV*	1.0	*SEMA3G*	0.9

TRIM67	0.9	*FCRL2*	0.7	*FANCB*	0.4	*CORO1C*	0.0
CX3CL1	0.8	*SLC2A4*	0.7	*TRPC7*	0.4	*CORO6*	0.0
TRIM54	0.8	*SLC16A10*	0.6	*AXL*	0.4	*CXCL10*	0.0
SLC35B3	0.8	*SLC16A12*	0.6	*TRIM37*	0.3	*CXCL11*	0.0
HAPLN4	0.8	*SLC35F5*	0.6	*SLC4A10*	0.3	*CXCL12*	0.0
SLC38A11	0.8	*TRIM24*	0.6	*ADAM10*	0.0	*CXCL13*	0.0
S100PBP	0.8	*LAG3*	0.6	*AKAP14*	0.0	*CXCL2*	0.0
TRIM13	0.8	*SLC32A1*	0.6	*AMIGO3*	0.0	*CXCL3*	0.0
SLC39A10	0.8	*BMPR1B*	0.6	*APOD*	0.0	*CXCL6*	0.0
NCAM2	0.8	*PVRL2*	0.6	*APOM*	0.0	*CXCL8*	0.0
SLC35D3	0.8	*SLC16A2*	0.6	*B2M*	0.0	*CXCL9*	0.0
IL5RA	0.8	*IL18R1*	0.6	*BST2*	0.0	*CXCR3*	0.0
CEACAM16	0.8	*SLC1A3*	0.6	*C2*	0.0	*DBP*	0.0
IL13RA1	0.8	*SLC22A13*	0.6	*CADM1*	0.0	*DEFB103A*	0.0
NCAM1	0.8	*MCOLN3*	0.6	*CADM4*	0.0	*DEFB103B*	0.0
NRG2	0.8	*SLC22A9*	0.6	*CCL1*	0.0	*DEFB105B*	0.0
LRP12	0.8	*SLC43A1*	0.6	*CCL18*	0.0	*DEFB106A*	0.0
CADM2	0.8	*SLC17A7*	0.6	*CCL19*	0.0	*DEFB106B*	0.0
CD55	0.8	*WASF1*	0.6	*CCL21*	0.0	*DEFB112*	0.0
PML	0.8	*SLC47A1*	0.6	*CCL26*	0.0	*DEFB113*	0.0
SLC52A2	0.7	*FCAMR*	0.6	*CCL28*	0.0	*DEFB115*	0.0
TCHHL1	0.7	*THBD*	0.6	*CCL3L1*	0.0	*DEFB119*	0.0
CORO7	0.7	*SLC11A2*	0.6	*CCL4L1*	0.0	*DEFB121*	0.0
F9	0.7	*LINGO4*	0.6	*CCL4L2*	0.0	*DEFB133*	0.0
SLC8A2	0.7	*SLC12A2*	0.6	*CCL5*	0.0	*DEFB136*	0.0
SLC46A3	0.7	*SLC23A3*	0.5	*CCL7*	0.0	*DIRC2*	0.0
IGSF21	0.7	*L1CAM*	0.5	*CCR9*	0.0	*EDA*	0.0
SLC39A7	0.7	*LRRC4C*	0.5	*CD160*	0.0	*ELP6*	0.0
SLC38A5	0.7	*SLC2A13*	0.5	*CD164*	0.0	*EVI2B*	0.0
CD36	0.7	*SLC44A1*	0.5	*CD1D*	0.0	*F11R*	0.0
ELP4	0.7	*TRIM32*	0.5	*CD274*	0.0	*F2R*	0.0
TRIM21	0.7	*TNFRSF21*	0.5	*CD34*	0.0	*F8A1*	0.0
SCARB2	0.7	*SLC5A2*	0.5	*CD3D*	0.0	*F8A3*	0.0
TRIM72	0.7	*NEXN*	0.5	*CD46*	0.0	*FAM187A*	0.0
C3AR1	0.7	*IL1RAPL2*	0.5	*CD47*	0.0	*FZD4*	0.0
SLC38A1	0.7	*IL1RAP*	0.5	*CD59*	0.0	*FZD9*	0.0
TRIM68	0.7	*FER1L5*	0.5	*CD69*	0.0	*GADD45G*	0.0
TRIM50	0.7	*LRFN5*	0.5	*CD74*	0.0	*GJB1*	0.0
HRH1	0.7	*SLC6A15*	0.5	*CD79A*	0.0	*GJB6*	0.0
CLEC14A	0.7	*STAT1*	0.4	*CD84*	0.0	*GP1BA*	0.0
A1BG	0.7	*MME*	0.4	*CD9*	0.0	*GP1BB*	0.0
AMIGO1	0.7	*F8*	0.4	*CLEC2B*	0.0	*HAMP*	0.0
SLC16A1	0.7	*MALT1*	0.4	*CLEC5A*	0.0	*HEPACAM2*	0.0
SLC38A3	0.7	*SEMA4C*	0.4	*CLECL1*	0.0	*HMOX1*	0.0
SLC36A4	0.7	*NTRK2*	0.4	*CNTFR*	0.0	*HSD11B1*	0.0
SLC16A5	0.7	*FGFR1*	0.4	*CORO1B*	0.0	*ICAM4*	0.0

IDO1	0.0	*MILR1*	0.0	*SLC25A35*	0.0	*TGFBR1*	0.0
IFITM1	0.0	*MPZ*	0.0	*SLC25A36*	0.0	*TGFBR3L*	0.0
IGF1	0.0	*MXRA8*	0.0	*SLC25A4*	0.0	*THY1*	0.0
IGFBP7	0.0	*MYD88*	0.0	*SLC25A40*	0.0	*TIGIT*	0.0
IGFBPL1	0.0	*NCR3LG1*	0.0	*SLC29A2*	0.0	*TMIGD1*	0.0
IGSF23	0.0	*NF2*	0.0	*SLC2A12*	0.0	*TNF*	0.0
IL10	0.0	*NFE2L2*	0.0	*SLC30A1*	0.0	*TNFRSF12A*	0.0
IL12A	0.0	*NPTN*	0.0	*SLC30A10*	0.0	*TNFRSF13C*	0.0
IL12B	0.0	*OPCML*	0.0	*SLC30A6*	0.0	*TNFSF13B*	0.0
IL13RA2	0.0	*PKD2*	0.0	*SLC30A7*	0.0	*TNFSF18*	0.0
IL15	0.0	*PRNP*	0.0	*SLC35A1*	0.0	*TNFSF4*	0.0
IL18BP	0.0	*PSMA2*	0.0	*SLC35A2*	0.0	*TOR4A*	0.0
IL1RAPL1	0.0	*PVRL1*	0.0	*SLC35C2*	0.0	*TREML1*	0.0
IL20	0.0	*RAB32*	0.0	*SLC35E3*	0.0	*TRIM14*	0.0
IL20RB	0.0	*S100A1*	0.0	*SLC35F1*	0.0	*TRIM27*	0.0
IL2RG	0.0	*S100A12*	0.0	*SLC35F4*	0.0	*TRIM44*	0.0
IL36B	0.0	*S100A13*	0.0	*SLC35G1*	0.0	*TRIM49D2*	0.0
INHBA	0.0	*S100A14*	0.0	*SLC46A1*	0.0	*TRIM74*	0.0
LCN2	0.0	*S100A16*	0.0	*SLC48A1*	0.0	*TRIM8*	0.0
LCN9	0.0	*S100A6*	0.0	*SLC50A1*	0.0	*TRPC5*	0.0
LGALS1	0.0	*S100G*	0.0	*SLC51B*	0.0	*VSIG1*	0.0
LGALS12	0.0	*SCN2B*	0.0	*SLC5A3*	0.0	*VSIG2*	0.0
LGALS13	0.0	*SLC10A3*	0.0	*SLC8A1*	0.0	*VSIG8*	0.0
LGALS7	0.0	*SLC10A4*	0.0	*SLC9A2*	0.0	*VSTM2A*	0.0
LGALS7B	0.0	*SLC25A11*	0.0	*SMIM1*	0.0	*VSTM2B*	0.0
LINGO2	0.0	*SLC25A16*	0.0	*SNTN*	0.0	*VSTM5*	0.0
LPAR6	0.0	*SLC25A18*	0.0	*TFPI*	0.0	*XCL1*	0.0
LTB	0.0	*SLC25A20*	0.0	*TGFB2*	0.0	*XK*	0.0
MAF	0.0	*SLC25A28*	0.0	*TGFB3*	0.0		

ANNEX 3- TB in France and some data from Lebanon

France

Total population on 1 January 2014 by EUROSTAT: 65 835 579

Tuberculosis case notifications, 2014

Total number of cases	4 845	
Notification rate per 100 000	7.4	
New* & relapses	4 535	
New* & relapses notification rate per 100 000	6.9	
Pulmonary	3 524	(72.7%)
of which microscopy-positive	1 739	(49.3%)
of which laboratory confirmed	1 998	(56.7%)
Laboratory-confirmed TB cases	2 518	(52.0%)
Mean age of new native TB cases	45.8 years	
Mean age of new foreign TB cases	39.9 years	
Foreign origin of all TB cases	2 692	(55.6%)
New (not previously treated)	4 535	(93.6%)

* Cases with unknown previous treatment included in new cases.

Drug resistance surveillance & TB-HIV co-infection, 2014

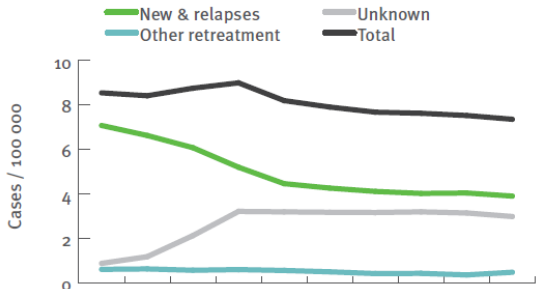
Completeness of DRS data*	Yes	
Completeness of HIV data**	–	
Case-linked data reporting	No	
Cases with DST results	4 235	–
Estimated MDR N, (best–low–high)	56–34–77	
Pulmonary MDR cases	72	(3.6%)
of which XDR cases	–	–
Notified MDR	111	(2.6%)
of which XDR cases	14	(12.6%)
TB cases tested for HIV	–	–
HIV-positive TB cases	–	–
of these on antiretroviral therapy	–	–

* National coverage 100% or culturing 90%, C+/All TB cases 50%, DST done for C+ 75%, EQA 95%.
** More than 50% of TB cases tested for HIV.

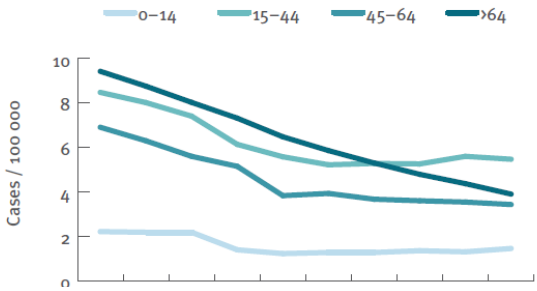
Treatment outcome monitoring

Data not available

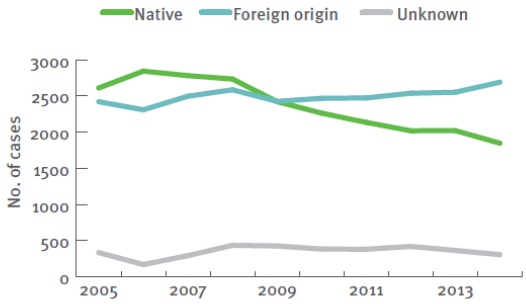
Tuberculosis notification rates by treatment history, 2005–2014



New and relapsed TB cases – notification rates by age group, 2005–2014



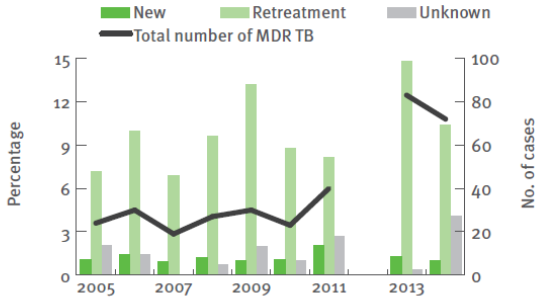
Tuberculosis cases by geographical origin, 2005–2014



TB-HIV co-infection, 2006–2014

Data not available

MDR TB cases by previous treatment history, 2005–2014



Treatment outcome, new culture-confirmed pulmonary TB cases, 2004–2013

Data not available

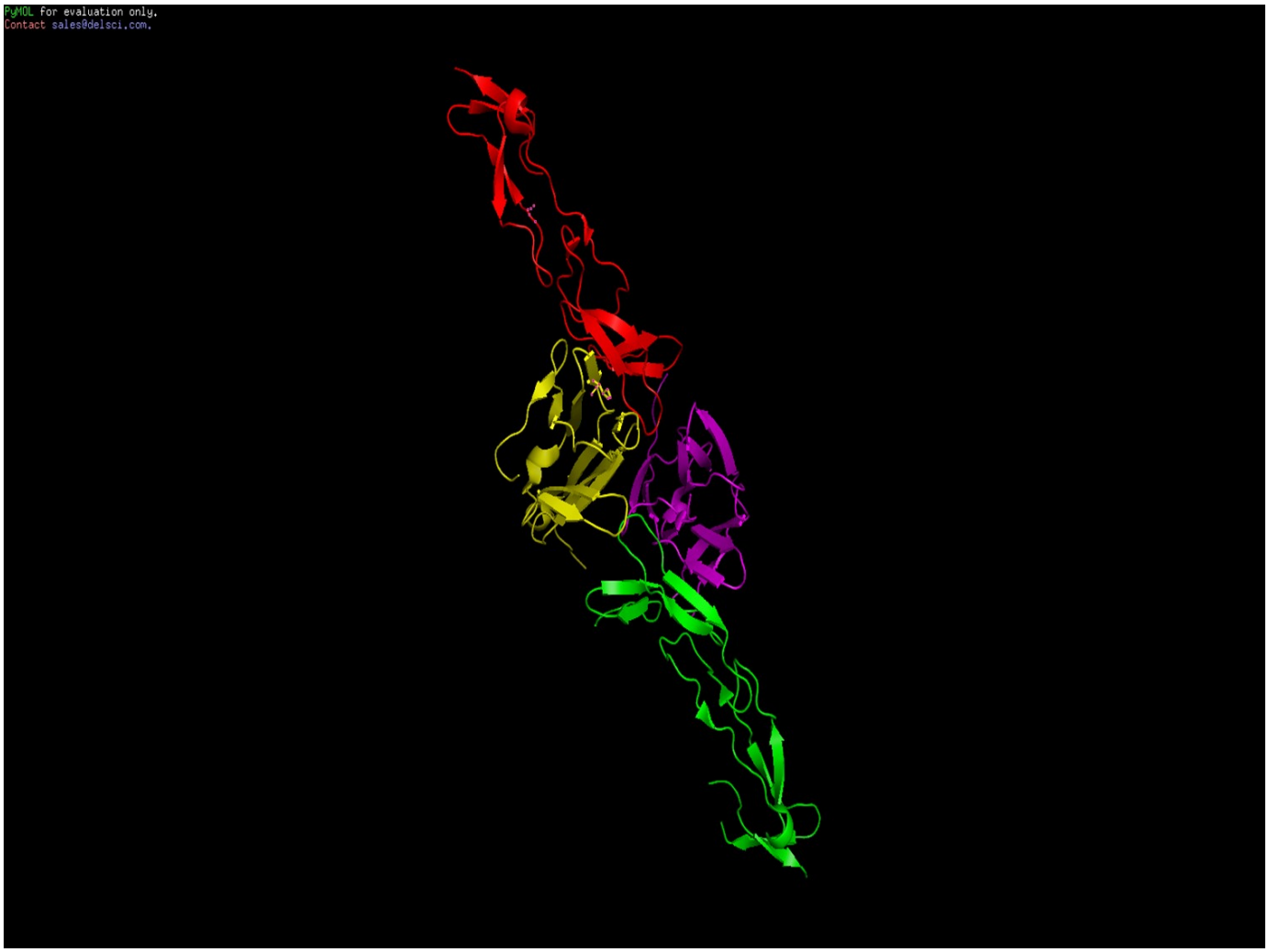
Number of Tuberculosis cases*

Tuberculosis cases	2010	2011	2012	2013
Total number of cases	*515*	*499*	*630*	*689*
Number of Lebanese cases	*344*	*298*	*330*	*341*
Number of Non-Lebanese cases	*171*	*200*	*300*	*348*

*MOPH National Tuberculosis Program

Supplementary Figure. Number of tuberculosis cases in Lebanon (source: Ministry of Public Health and WHO)

Annex 4- Crystal structure of the BTLA-HVEM complex



Crystal structure of the BTLA-HVEM complex

Annex 5- Overview of *Mycobacterium tuberculosis* infections in mammals

References in the following table:

5: Charles O. Thoen. (n.d.). Tuberculosis and other Mycobacterial Infections - Veterinary Manual. Retrieved May 5, 2017, from http://www.merckvetmanual.com/generalized-conditions/tuberculosis-and-other-mycobacterial-infections

4: Cheng, L I, Mikota S.K., M. R. J. (2001). Mycobacterim tuberculosis: in Zoo and Wildlife Species, 20(1), 291–303.

2: Ghodbane, R., & Drancourt, M. (2013). Non-human sources of Mycobacterium tuberculosis. Tuberculosis, 93(6), 589–595. http://doi.org/10.1016/j.tube.2013.09.005

1 and 1b: Isaza Ramiro. (2003). Tuberculosis in All Taxa. Flowler ME, Miller RE, editors. Zoo and wild animal medicine. 5th ed. Pennsylvania: W.B. Saunders Company.

1b corresponds to a table representing relative Mycobacteria susceptibilities in the same textbook.

3: Thoen, C. O., Lobue, P. a, Enarson, D. a, Kaneene, J. B., & de Kantor, I. N. (2009). Tuberculosis: a re-emerging disease in animals and humans. Veterinaria Italiana, 45(1), 135–181.

Resitance Susceptibility		REFERENCES					
	Organism	**1**	**1b**	**2**	**3**	**4**	**5**
Non human primates NHP	Asian Orld World such as rhesus monkeys (*Macaca mulatta*)						
	asian monkeys						
	African Orld World baboons & great apes						
	old word monkeys including rhesus macaques & pig-tailed macaques						
	cynomolgus monkey			0.4% cases		less suscep	
	african monkeys						
	South american NHP	rare	rare				
	great apes		occasional				
	monkeys, large apes						
	a 3 year-old marmoset (*Callithrix jacchus*)						
Elephants	Asian elephants (*Elephas maximus*)	isolated	common			closer to humans	
	African elephants (*Loxodonta* spp.)	rare	not likely				
	elephants						
Psittacines	Psittacines	reported					
	parrots		not likely	new world			
Birds	chickens	not likely					
	waterfowl		not likely				
	hawks and owls		not likely				
	pheasants		not likely				
	passerines		not likely				
	cranes		not likely				
	ratites		not likely				
	flamingos		not likely				
Artiodactylids	cattle		not likely		rarely isolated		
	pigs		rare		isolated swine		
	sheep & goats		not likely	domestic goats	*		
	Rocky mountain goats (*Oreamnos americanus*)				isolated		
Persissodactylids	horses		not likely		relatively		
	rhinoceroses		occasional		isolated		
	black rhinoceros (*Diceros bicornis*)				isolated		
	white rhinoceros (*Ceratotherium simum simum*)						
	great Indian rhinoceros *Rhinoceros unicornis*					*M.tb* ?	
carnivores	dogs		rare		like humans		
	cats		not likely		very Resistant		
	ferrets		not likely				
	lions		rare				
rodents & lagomorphs	rabbits		not likely				
	guinea pigs		classic				
	mice		rare				
	hamsters		classic				
other mammals	tree kangarous		not likely				
	koalas		not likely				
	deers						
	elk						
	oryx				isolated		
	zoo-exhibited oryx (*Oryx gazella beisa*)						
	addax				isolated		
	(*Addax nasomaculatus*)						
	Bovidae from Africa						
	black buck (*Antilope cervicapra*)						
	bongo (*Tragelaphus eurycerus*)						
	a giraffe (*Giraffa camelopardalis*)						
	exotic carnivores					rarely reported	
	european otter (*Lutra lutra*)						
	polar bear (*Ursus maritimus*)						
	Captive and wild sea lions	infected					
	(*Neophoca* spp.)						
	fur seals						
	(*Arctocephalus* spp.)						
	seals and sea lions		occasional				
	water buffalo	not likely					
	cervids	not likely					
	camels	not likely					
	new world camilids	not likely					
	tapirs		rare	*Tapirus terrestris*			

REFERENCES

1. Latham, K. J. Human Health and the Neolithic Revolution: an Overview ofImpacts of the Agricultural Transition on Oral Health, Epidemiology, and the Human Body. *Nebraska Anthropol. Pap.* **187,** (2013).
2. Haensch, S. *et al.* Distinct Clones of Yersinia pestis Caused the Black Death. *PLoS Pathog* **6,** 10011341–2 (2010).
3. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15,** 379–393 (2014).
4. Fumagalli, M. *et al.* Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLoS Genet.* **7,** e1002355 (2011).
5. Brunham, R. C., Plummer, F. a. & Stephens, R. S. Bacterial antigenic variation, host immune response, and pathogen-host coevolution. *Infect. Immun.* **61,** 2273–2276 (1993).
6. Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32,** 569–577 (2002).
7. Finlay, B. B. & McFadden, G. Anti-immunology: Evasion of the host immune system by bacterial and viral pathogens. *Cell* **124,** 767–782 (2006).
8. Kawashima, Y. *et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* **458,** 641–645 (2009).
9. Nyamweya, S. *et al.* Comparing HIV-1 and HIV-2 infection: Lessons for viral immunopathogenesis. *Rev. Med. Virol.* **23,** 221–240 (2013).
10. Schurr, E. The contribution of host genetics to tuberculosis pathogenesis. *Kekkaku* **86,** 17–28 (2011).
11. Tsay, C. J. Julius Wagner-Jauregg and the legacy of malarial therapy for the treatment of general paresis of the insane. *Yale J. Biol. Med.* **86,** 245–54 (2013).
12. Cooke, G. S. & Hill, A. V. S. Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.* **2,** 967–977 (2001).
13. PASVOL, G., WEATHERALL, D. J. & WILSON, R. J. M. Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria. *Nature* **274,** 701–703 (1978).
14. Williams, T. N. *et al.* High incidence of malaria in α-thalassaemic children. *Nature* **383,** 522–525 (1996).
15. Synopses of Research Articles. (2006).
16. Guindo, A. *et al.* X-Linked G6PD Deficiency Protects Hemizygous Males but Not Heterozygous Females against Severe Malaria. *PLoS Med.* **4,** e66 (2007).
17. Harris, J. B. *et al.* Blood group, immunity, and risk of infection with Vibrio cholerae in an area of endemicity. *Infect. Immun.* **73,** 7422–7 (2005).
18. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The Resistance Factor to *Plasmodium vivax* in Blacks. *N. Engl. J. Med.* **295,** 302–304 (1976).
19. Pier, G. B. *et al.* Role of Mutant CFTR in Hypersusceptibility of Cystic Fibrosis Patients to Lung Infections. *Science (80-.).* **271,** (1996).
20. Pier, G. B. *et al.* Salmonella typhi uses CFTR to enter intestinal epithelial cells. *Nature* **393,** 79–82 (1998).

21. Korthäuer, U. *et al.* Defective expression of T-cell CD40 ligand causes X-linked immunodeficiency with hyper-IgM. *Nature* **361,** 539–541 (1993).
22. DiSanto, J. P., Bonnefoy, J. Y., Gauchatt, J. F., Fischer, A. & Saint Basile, G. de. CD40 ligand mutations in X-linked immunodeficiency with hyper-IgM. *Nature* **361,** 541–543 (1993).
23. Newport, M. J. *et al.* A Mutation in the Interferon-γ –Receptor Gene and Susceptibility to Mycobacterial Infection. *N. Engl. J. Med.* **335,** 1941–1949 (1996).
24. Casanova, J.-L. *et al.* A human IFNGR1 small deletion hotspot associated with dominantsusceptibility to mycobacterial infection. *Nat. Genet.* **21,** 370–378 (1999).
25. Altare, F. *et al.* Inherited interleukin 12 deficiency in a child with bacille Calmette-Guérin and Salmonella enteritidis disseminated infection. *J. Clin. Invest.* **102,** 2035–40 (1998).
26. Jong, R. de *et al.* Severe Mycobacterial and Salmonella Infections in Interleukin-12 Receptor-Deficient Patients. *Science (80-.).* **280,** (1998).
27. von Bernuth, H. *et al.* Pyogenic Bacterial Infections in Humans with MyD88 Deficiency. *Science (80-.).* **321,** (2008).
28. Chapman, S. J. & Hill, A. V. S. Human genetic susceptibility to infectious disease. *Nat. Rev. Genet.* (2012). doi:10.1038/nrg3114
29. Brouwer, M. C. *et al.* Host genetic susceptibility to pneumococcal and meningococcal disease: a systematic review and meta-analysis. *Lancet Infect. Dis.* **9,** 31–44 (2009).
30. Degn, S. E., Jensenius, J. C. & Thiel, S. Disease-Causing Mutations in Genes of the Complement System. *Am. J. Hum. Genet.* **88,** 689–705 (2011).
31. Skattum, L., van Deuren, M., van der Poll, T. & Truedsson, L. Complement deficiency states and associated infections. *Mol. Immunol.* **48,** 1643–1655 (2011).
32. Jönsson, G. *et al.* Hereditary C2 deficiency in Sweden: frequent occurrence of invasive infection, atherosclerosis, and rheumatic disease. *Medicine (Baltimore).* **84,** 23–34 (2005).
33. Sprong, T. *et al.* Deficient alternative complement pathway activation due to factor D deficiency by 2 novel mutations in the complement factor D gene in a family with meningococcal infections. *Blood* **107,** (2006).
34. Fijen, C. A., Kuijper, E. J., te Bulte, M. T., Daha, M. R. & Dankert, J. Assessment of complement deficiency in patients with meningococcal disease in The Netherlands. *Clin. Infect. Dis.* **28,** 98–105 (1999).
35. van de Vosse, E., van Dissel, J. T. & Ottenhoff, T. H. Genetic deficiencies of innate immune signalling in human infectious disease. *Lancet Infect. Dis.* **9,** 688–698 (2009).
36. Liu, R. *et al.* Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. *Cell* **86,** 367–377 (1996).
37. Samson, M. *et al.* Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382,** 722–725 (1996).
38. Zhang, S.-Y. *et al.* TLR3 Deficiency in Patients with Herpes Simplex Encephalitis. *Science (80-.).* **317,** (2007).
39. Byun, M. *et al.* Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. *J. Exp. Med.* **207,** (2010).
40. Ramoz, N. *et al.* Mutations in two adjacent novel genes are associated with epidermodysplasia verruciformis. *Nat. Genet.* **32,** 579–581 (2002).
41. Thorven, M. *et al.* A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections.

J. Virol. **79,** 15351–5 (2005).
42. Puel, A. *et al.* Chronic Mucocutaneous Candidiasis in Humans with Inborn Errors of Interleukin-17 Immunity. *Science (80-.).* **332,** (2011).
43. Zeidler, M., Stewart, G., Cousens, S., Estibeiro, K. & Will, R. Codon 129 genotype and new variant CJD. *The Lancet* **350,** (1997).
44. Quintana-Murci, L. & Clark, A. G. Population genetic tools for dissecting innate immunity in humans. *Nat. Rev. Immunol.* **13,** 280–93 (2013).
45. Daley, G. Q. & Ben-Neriaht, Y. IMPLICATING THE bcrlabl GENE IN THE PATHOGENESIS OF HUMAN LEUKEMIA P H I LAD E L P H I A C H RO M 0s 0 M E -POS IT I V E.
46. Visootsak, J. & Graham, J. M. Klinefelter syndrome and other sex chromosomal aneuploidies. *Orphanet J. Rare Dis.* **1,** (2006).
47. Guggino, W. B. & Stanton, B. A. New insights into cystic fibrosis: molecular switches that regulate CFTR. *Nat. Rev. Mol. Cell Biol.* **7,** 426–436 (2006).
48. Drumm, M. *et al.* What happens to deltaF508 in vivo? *J. Clin. Invest.* **103,** 1369–70 (1999).
49. Warby, S. C., Graham, R. K. & Hayden, M. R. *Huntington Disease*. *GeneReviews(®)* (University of Washington, Seattle, 1993). at <http://www.ncbi.nlm.nih.gov/pubmed/20301482>
50. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7,** 85–97 (2006).
51. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural Genetic Variation Caused by Transposable Elements in Humans. doi:10.1534/genetics.104.031757
52. Ray, D. A. *et al.* Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res.* **18,** 717–728 (2008).
53. Wright, A. F. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2005). doi:10.1038/npg.els.0005005
54. Steinberg, M. H. Predicting clinical severity in sickle cell anaemia. *Br. J. Haematol.* **129,** 465–481 (2005).
55. Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* **6,** S6–S12 (2009).
56. Trowsdale, J. & Parham, P. Defense strategies and immunity-related genes. *Eur. J. Immunol.* **34,** 7–17 (2004).
57. DOHERTY, P. A BIOLOGICAL ROLE FOR THE MAJOR HISTOCOMPATIBILITY ANTIGENS. *Lancet* **305,** 1406–1409 (1975).
58. Phane Buhler, S. & Sanchez-Mazas, A. HLA DNA Sequence Variation among Human Populations: Molecular Signatures of Demographic and Selective Events. *PLoS One* **6,** (2011).
59. Solberg, O. D. *et al.* Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum. Immunol.* **69,** 443–64 (2008).
60. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
61. Hughes, A. L. & Nei, M. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 958–62 (1989).
62. Markov, P. V & Pybus, O. G. Evolution and Diversity of the Human Leukocyte

Antigen(HLA). *Evol. Med. Public Heal.* (2015). doi:10.1093/emph/eou033
63. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2010. at <http://www.repeatmasker.org/faq.html#faq3>
64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).
65. Chevreux, B., Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.7465>
66. Staden, R., Beal, K. F. & Bonfield, J. K. The Staden package, 1998. *Methods Mol. Biol.* **132,** 115–30 (2000).
67. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
68. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978–89 (2001).
69. Stephens, M. & Donnelly, P. Report A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *Am. J. Hum. Genet* **73,** 1162–1169 (2003).
70. Wang, Y., Lu, J., Yu, J., Gibbs, R. A. & Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* **23,** 833–42 (2013).
71. About | 1000 Genomes. at <http://www.internationalgenome.org/about>
72. Sharpe, A. H. & Freeman, G. J. THE B7–CD28 SUPERFAMILY. *Nat. Rev. Immunol.* **2,** 116–126 (2002).
73. Haimila, K. *et al.* Genetic association of coeliac disease susceptibility to polymorphisms in the ICOS gene on chromosome 2q33. *Genes Immun.* **5,** 85–92 (2004).
74. Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423,** 506–511 (2003).
75. Chen, P.-L. *et al.* Family-based association study of cytotoxic T-lymphocyte antigen-4 with susceptibility to Graves' disease in Han population of Taiwan. *Genes Immun.* **9,** 87–92 (2008).
76. Gu, X. *et al.* +49G > A polymorphism in the cytotoxic T-lymphocyte antigen-4 gene increases susceptibility to hepatitis B-related hepatocellular carcinoma in a male Chinese population. *Hum. Immunol.* **71,** 83–87 (2010).
77. EXCOFFIER, L. & LISCHER, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10,** 564–567 (2010).
78. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25,** 1451–1452 (2009).
79. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–95 (1989).
80. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–8 (2002).
81. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci.* **102,** 15942–15947 (2005).
82. D. E. Reich S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F.

Farhadian, R. Ward, E. S. Lander, M. C. Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
83. Norman, P. J. *et al.* Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans. *Nat. Genet.* **39,** 1092–1099 (2007).
84. Teft, W. a., Kirchhof, M. G. & Madrenas, J. a Molecular Perspective of Ctla-4 Function. *Annu. Rev. Immunol.* **24,** 65–97 (2006).
85. World Health Organization. *Global Tuberculosis Report 2016*. at <http://apps.who.int/iris/bitstream/10665/250441/1/9789241565394-eng.pdf?ua=1>
86. Ministry of Public Health & Republic of Lebanon. *National Guidelines For Tuberculosis Prevention, Care And Elimination In Lebanon- 2017*. at <http://www.moph.gov.lb/en/Pages/2/11570/national-guidelines-for-tuberculosis-prevention-care-and-elimination-in-lebanon-2017>
87. Zasloff, M. Fighting infections with vitamin D. *Nat. Med.* **12,** 388–390 (2006).
88. Adams, J. S. & Hewison, M. Regulation of Innate and Adaptive Immunity. *Diabetes* **4,** 80–90 (2009).
89. Flannagan, R. S., Cosío, G. & Grinstein, S. Antimicrobial mechanisms of phagocytes and bacterial evasion strategies. *Nat. Rev. Microbiol.* **7,** 355–66 (2009).
90. van Etten, E. *et al.* The vitamin D receptor gene FokI polymorphism: Functional impact on the immune system. *Eur. J. Immunol.* **37,** 395–405 (2007).
91. Xu, C. *et al.* Vitamin D Receptor Gene FOKI Polymorphism Contributes to Increasing the Risk of HIV-Negative Tuberculosis: Evidence from a Meta-Analysis. *PLoS One* **10,** e0140634 (2015).
92. Georgel, P., Macquin, C. & Bahram, S. The heterogeneous allelic repertoire of human toll-like receptor (TLR) genes. *PLoS One* **4,** 1–11 (2009).
93. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).
94. Gertner-Dardenne, J., Fauriat, C. & Olive, D. BTLA, a key regulator of Vγ9Vδ2 T-cell proliferation. *Oncoimmunology* **2,** e25853 (2013).
95. Verrall, A. J., G. Netea, M., Alisjahbana, B., Hill, P. C. & van Crevel, R. Early clearance of Mycobacterium tuberculosis: A new frontier in prevention. *Immunology* **141,** 506–513 (2014).
96. Compaan, D. M. *et al.* Attenuating lymphocyte activity: The crystal structure of the BTLA-HVEM complex. *J. Biol. Chem.* **280,** 39553–39561 (2005).
97. Li, X. *et al.* SLC11A1 (NRAMP1) Polymorphisms and Tuberculosis Susceptibility: Updated Systematic Review and Meta-Analysis. doi:10.1371/journal.pone.0015831
98. Ng, P. C. & Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu. Rev. Genomics Hum. Genet* **7,** 61–80 (2006).
99. Haagsman, H. P. & Diemel, R. V. Surfactant-associated proteins: functions and structural variation. *Comp. Biochem. Physiol. Part A Mol. Integr. Physiol.* **129,** 91–108 (2001).
100. Borrego, F. The CD300 molecules: an emerging family of regulators of the immune system. *Blood* **121,** (2013).
101. Klesney-Tait, J., Turnbull, I. R. & Colonna, M. The TREM receptor family and signal integration. *Nat. Immunol.* (2006). doi:10.1038/ni1411
102. Dong, H., Lv, Y., Zhao, D., Barrow, P. & Zhou, X. Defensins: The Case for Their Use against Mycobacterial Infections. *J. Immunol. Res.* **2016,** 1–9 (2016).
103. Ganz, T. Defensins: antimicrobial peptides of innate immunity. *Nat. Rev. Immunol.* **3,**

710–720 (2003).
104. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3 (Bethesda).* **5,** 931–41 (2015).
105. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* (2011). doi:10.1038/nrg3117
106. Tailleux, L. *et al.* DC-SIGN Is the Major Mycobacterium tuberculosis Receptor on Human Dendritic Cells. *J. Exp. Med.* **197,** 121–127 (2002).
107. Barreiro, L. B. *et al.* The Heritage of Pathogen Pressures and Ancient Demography in the Human Innate-Immunity CD209/CD209L Region. *Am. J. Hum. Genet* **77,** 869–886 (2005).
108. Dogan, I. *et al.* Polymorphisms in the vitamin D receptor gene and risk of lung cancer. *Med. Sci. Monit.* **15,** BR232-42 (2009).
109. Cox, M. B. *et al.* Potential association of vitamin D receptor polymorphism *Taq1* with multiple sclerosis. *Mult. Scler. J.* **18,** 16–22 (2012).
110. Barroso, E. *et al.* Genetic analysis of the vitamin D receptor gene in two epithelial cancers: melanoma and breast cancer case-control studies. *BMC Cancer* **8,** 385 (2008).
111. Parker, J. *et al.* Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502,** 1–9 (2013).
112. Isaza Ramiro. Tuberculosis in All Taxa. 689–96 (2003).
113. Ghodbane, R. & Drancourt, M. Non-human sources of Mycobacterium tuberculosis. *Tuberculosis* **93,** 589–595 (2013).
114. Thoen, C. O., Lobue, P. a, Enarson, D. a, Kaneene, J. B. & de Kantor, I. N. Tuberculosis: a re-emerging disease in animals and humans. *Vet. Ital.* **45,** 135–181 (2009).

Abstract: Finding novel gene candidates and polymorphisms involved in host-pathogen interactions

Host-pathogen co-evolution and interactions contribute in shaping the genetic diversity of both organisms. The objective of this thesis is to define the genetic basis of variability in disease resistance/susceptibility through the development of large-scale *in silico* screens to identify novel gene candidates implicated in host-pathogen interactions (such as tuberculosis).

A pilot study was conducted on *CD28, CTLA4,* and *ICOS* to investigate their polymorphism and determine if there was a correlation between variation in their coding regions and the Single Nucleotide Polymorphism (SNP) CT60 of *CTLA4* that is associated with autoimmune diseases. Our results showed that *CD28* and *ICOS* are highly conserved, while *CTLA4* has two common allotypes. Population genetics analyses indicate that balancing selection maintains this polymorphism in the leader (p=0.001), but, no coding variation correlates with the SNP CT60.

As a first step in our study based on data available in the literature, we selected a set of ten genes relevant for the immune response against *Mycobacterium tuberculosis*: *BTLA*, *TNFRSF14, TNFRSF1B, CCL2, TLR2, INFG, IFNGR, VDR, NRAMP1,* and *CISH*. Seven of these genes were moderately polymorphic, while three of them (*IFNG, CCL2, CISH*) were highly conserved. This analysis was used to prepare and setup the large scale analysis using the same developed pipeline for polymorphism detection and allele reconstruction.

For our *in silico* screens to identify novel candidates at a large-scale, we used sequence data from several projects and consortiums (like the 1000 Genomes Project, HGPD and others) to isolate most polymorphic human genes amongst a list of over 1760 candidates selected based on a) already established relevance for infections (tuberculosis in particular, by literature searches) and b) on evolutionary considerations (multigenic families, structural domains, etc.). A first screen of 64 individuals from eight different populations from several regions of the world (1000 Genomes project) was performed and most variable genes were selected for more further extensive analyses on a larger panel (715 individuals). 30 most polymorphic genes were thus identified *(SLC35G4, SIRPA, DEFB108B, GYPB, IGLL1, PSG1, SLC25A5, KIR3DL3, DEFB132, PSG2, PSG5, OBP2A, CCL4, SIRPB1, ORM2, SFTPA1, CD52, SIGLEC12, ABO, SLC35G5, DEFB126, FUT3, DEFB1, PSG8, PSG11, P2RX7, CD207, DEFA3, IL37,DEFB127)*. The extent of polymorphism and the allelic worldwide variants of each of these 30 genes are ready to be fully characterized. The data generated could be compared against infectious disease resistance/susceptibility data to identify potentially relevant gene variation.

Keywords: polymorphisms, gene candidates, SNP, host-pathogen interactions

Résumé: Caractérisation de nouveaux gènes candidats et polymorphismes impliqués dans les interactions hôtes-pathogènes

La coévolution ainsi que les différentes interactions entre hôte et pathogène contribuent à former la diversité génétique de ces deux organismes. Dans le cadre de cette thèse, nous nous sommes intéressés à l'étude de la variabilité génétique de 1760 gènes immunitaires choisis suivant des critères définis, pour essayer d'expliquer pourquoi il existe une variation individuelle face aux infections. L'objectif principal de ce projet était alors de caractériser et d'analyser de nouveaux gènes et polymorphismes immunitaires pouvant expliquer le contrôle ou la susceptibilité à certaines infections (par exemple la tuberculose).

Deux études pilotes nous ont permis de développer le pipeline de détection de polymorphismes. Pour la première, le polymorphisme de 3 gènes immuno-modulateurs (*CD28, CTLA4,* et *ICOS)* a été caractérisé. On a trouvé que *CD28* et *ICOS* sont des gènes conservés tandis que *CTLA4* présente deux allèles communs sous sélection balancée. Nous avons voulu aussi déterminer s'il y avait une corrélation entre une variation codante chez ces 3 gènes et une variation connue associée à certaines maladies auto-immunes (appelée SNP CT60) dans la région intergénique (en 3') de *CTLA4.* Aucune corrélation n'a été trouvée.
Dans la deuxième étude, nous avons caractérisé le polymorphisme de 10 gènes impliqués dans la réponse immunitaire contre *Mycobacterium tuberculosis : BTLA*, *TNFRSF14, TNFRSF1B, CCL2, TLR2, INFG, IFNGR, VDR, NRAMP1,* et *CISH*. Ces gènes ne sont pas très polymorphes et trois d'entre eux sont très conservés (*IFNG, CCL2, CISH*). Ces deux études nous ont aidés à préparer une analyse à grande échelle avec des séries de mise au point et d'amélioration du pipeline.

Pour l'analyse à grande échelle, nous avons sélectionné 1760 gènes en se basant sur des critères définis (gènes impliqués dans l'immunité contre les infections, gènes appartenant à des familles multigéniques ou présentant des domaines structuraux intéressants etc.). La variabilité génétique a été ainsi étudiée dans les populations humaines par une analyse minutieuse *in silico* de données de séquençage d'exomes générées par différents projets et consortiums (1000 Genomes, HGPD, et d'autres projets) pour plus de 700 individus représentant 20 populations à travers le monde. 30 gènes les plus polymorphes ont été ainsi identifiés *(SLC35G4, SIRPA, DEFB108B, GYPB, IGLL1, PSG1, SLC25A5, KIR3DL3, DEFB132, PSG2, PSG5, OBP2A, CCL4, SIRPB1, ORM2, SFTPA1, CD52, SIGLEC12, ABO, SLC35G5, DEFB126, FUT3, DEFB1, PSG8, PSG11, P2RX7, CD207, DEFA3, IL37,DEFB127)*. Ces gènes pourront être entièrement caractérisés et les données produites pourraient être comparées avec des données de résistance/sensibilité de certaines maladies infectieuses.

Mots-clés : polymorphisme, gènes candidats, SNP, interactions hôtes-pathogènes