Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA

Biodiversità ed Evoluzione

Ciclo XXI

Settore scientifico disciplinare di afferenza: BIO/08 - Antropologia

TNFRSF13B GENETIC VARIABILITY: AN ANTHROPOLOGICAL - EVOLUTIONARY APPROACH TO BIOMEDICAL RESEARCH

Presentata da: Dott. Marco Sazzini

Coordinatore Dottorato — **Prof. Giovanni Cristofolini**

Relatore — **Prof.ssa Donata Luiselli**

Esame finale anno 2009

Table of contents

1. Introduction

4. Results

1. Introduction

1.1 Why Medicine needs Evolution

Biomedical research could be actually improved by an evolutionary perspective that looks at our species as the result of evolutionary processes occurred in extremely variable environmental and socio-cultural contexts.

Accordingly, such a viewpoint also regards our genome as a biological reality shaped by natural selection under the constraints of several tradeoffs that inevitably produce specific compromises and vulnerabilities.

Classical medical research has long tried to provide mechanistic explanations of disease conditions, mainly on the basis of consideration of the simple study of body anatomic and physiological mechanisms, as they currently exist.

In contrast, an evolutionary approach to medicine pursues the aim of exploring the reasons for the human genome to be designed in a way that makes us vulnerable to diseases, offering a broader context in which to conduct research.

As already stated by the distinguished geneticist Theodosius Dobzhansky, it seems to be undeniable that "*nothing in biology makes sense except in the light of evolution*" (Dobzhansky 1973) and nowadays, almost 40 years later this claim, we are even more aware about the fact that evolutionary biology represents the scientific foundation for all biology and that, at the same time, biology is turned out to be the foundation for all medicine.

Whereas variability is a fundamental concept at the core of evolution theory, and *H. sapiens* is notably variable in all of its cultural and biological manifestations, medical research tends to focus on what differs from a perceived "normal" condition and, furthermore, this medical "normal" is often based on health characteristics of Western people (Trevathan 2007).

That being so, a population-evolutionary approach to biomedical research reveals its significance by cautioning that a single "normal" genome is non-existent and by suggesting that the achievement of a genomic region global picture of nucleotide and haplotype diversity, through the study of populations with different ancestry, could facilitate the distinction between variants falling into the standard degree of intra-specific variation and changes which are potentially related to diseases.

1.1.1 Natural selection and complex diseases

Complex genetic diseases, which are due to the interplay of several genes and environmental factors, represent the major source of morbidity and mortality in developed countries, resulting much more common respect to Mendelian-inherited simple diseases.

A possible explanation for this condition comes from evolutionary genetics and states that natural selection against complex diseases-causative alleles is presumably weak, since they individually have a small effect on the disease phenotype, whereas alleles underlying simple diseases, and hence with a greater pathological impact, tend to be rapidly removed from the population by a stronger natural selection (Smith and Lusis 2002).

Such a remark emphasizes how evolutionary genetics can play a substantial role in dissecting the origin, causes and diffusion of human diseases, according to current opinions for which many aspects of human health are strongly influenced by the individual genotype and this genotype is anything else that an awesome result of our species evolutionary history.

At the theoretical basis of this discipline there is a neutral model of molecular evolution for which most of genetic variation within and between species (or populations) has accumulated as a result of neutral processes (Kimura 1983). As a consequence, fixation or loss of most alleles is determined by genetic drift, so that species (or populations) may become genetically and phenotypically differentiated over time simply due to random fluctuations of their allele frequencies.

Nevertheless, many genetic and phenotypic differences among human populations may be also due to adaptative processes, which were historically favored by natural selection. That being so, present-day deeper and deeper survey of human genetic variation in many different populations represents a turning point in the study of natural selection effects on the *H. sapiens* genome.

Nowadays, it is also possible to identify new candidate targets of selection and to reevaluate previous claims by comparison with empirical distributions of DNA sequence variation across the genome and among populations (Sabeti et al. 2007). In particular, identifying regions of the human genome that have been subjected to such selection events might turn out to be extremely important to understand their potential role in the different diseases susceptibility of human populations.

Despite that, it is necessary to keep in mind that a sharp distinction between “normal” and disease-associated genetic variation is hardly achievable, since phenotypic consequences of genetic variants strongly depend also on the environment. As regards this issue, a clear example is depicted by the *“thrifty gene”* hypothesis proposing that in response to scarcity of food in ancient environments alleles causing more efficient food assimilation had increased fitness, while in modern environments they actually increase susceptibility to obesity and Type 2 diabetes (Neel 1962).

As a general rule, looking for genetic and phenotypic variation patterns consistent with adaptive or neutral processes needs the assumption that it is possible to separate functional and neutral genetic variants. Changes that do not alter protein function, such as silent mutations in exons or introns, are usually expected to reflect neutral variation, whereas protein sequence, transcription, translation, or expression levels altering mutations are expected to be under stronger evolutionary constraint and selection.

If a functional mutation is adaptive, positive selection increases its frequency in the population, driving it to a faster fixation respect to neutral expectations. On the contrary, if different functional mutations are favored, or if there is a heterozygote selective advantage, balancing selection can maintain variation in the population longer than expected under a neutral model of evolution (Figure 1.1.1.1).

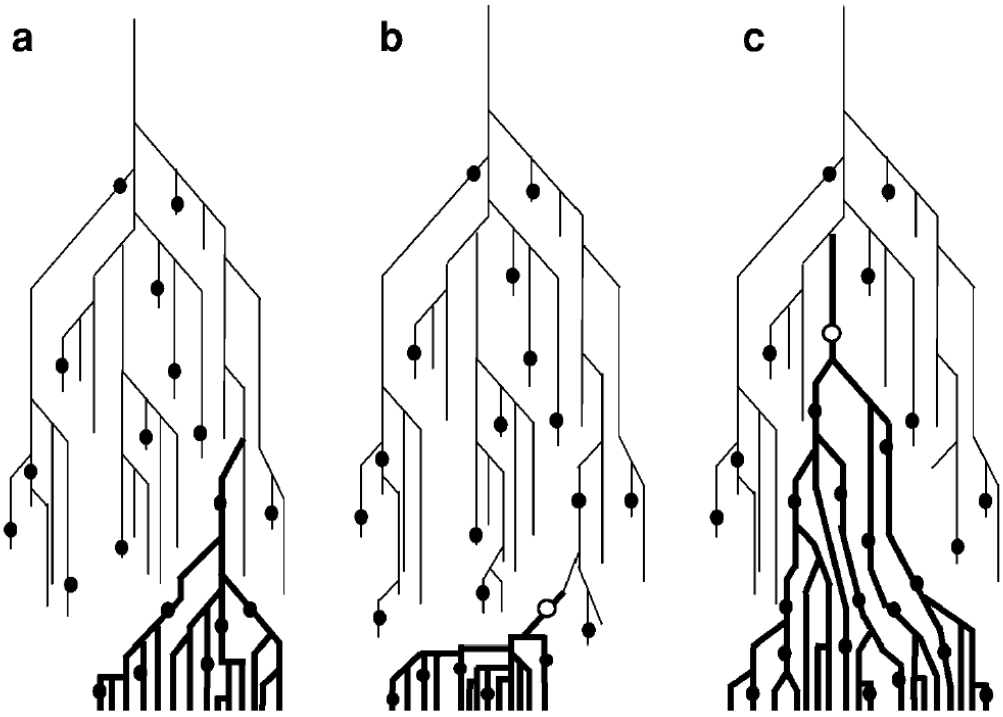

Figure 1.1.1.1 Gene genealogies and lineages coalescence expected under different models of molecular evolution or demographic processes.
a) Neutral model; b) positive selection or population expansion model; c) balancing selection or population subdivision model. Black circles represent neutral mutations, while open circles represent adaptive mutation.

However, it may be very difficult to distinguish between neutral and adaptive genetic variation patterns, because of the almost always simultaneous action of genetic drift and selection on human

populations and since they have had a complex history of both size reductions and expansions which can strongly influence patterns of population variation.

For example, when an adaptive mutation rapidly increases in frequency due to positive selection it can sweep out pre-existing population variation (Payseur et al. 2002). After this selective sweep, new polymorphisms will arise, but they will be rarely shared among individuals.

In this way, a star-like phylogeny, which has many external branches that coalesce back rapidly to the recent common ancestor at the time of the selective sweep, can be observed as a consequence of an excess of rare polymorphisms in the population.

Unfortunately, the same effect can occur when population size rapidly increases. In such a case, genetic drift has less effect and an increase in the length of genealogy external branches, as well as a rapid coalescence prior to the time of population expansion, is observed.

On the contrary, balancing selection can maintain polymorphisms longer than expected under a neutral model of evolution, resulting in significantly long coalescence times and a largely distributed variation on phylogeny internal branches (Figure 1.1.1.1) (Navarro and Barton 2002).

Since population size fluctuations are expected to equally impact the entire genome, whereas selection only targets DNA regions or populations in which adaptive mutations have arisen, a useful approach for distinguishing between neutral and adaptive evolution can be the comparison of nucleotide sequence variation from different genomic regions and populations.

To this end, several studies on the *G6PD* gene, mutations of which may provide protection against malarial infection and therefore are maintained by balancing selection (Tishkoff and Verrelli 2003a), on the *CCR5* gene, positively associated with HIV-1 resistance (Stephens et al. 1998) and on the *MC1R* gene, commonly associated with variation in skin pigmentation (Harding et al. 2000), are representative of a bright use of samples from different geographic areas, as well as of both silent and functional variation data.

Genomic regions with long stretches of nucleotide sites in high linkage disequilibrium (LD), or for which high differentiation among populations or patterns of unusual low diversity are observed, can be generally considered as good candidates for selection studies (Hamblin et al. 2002; Enard et al. 2002). One of the main classical examples is that of the MHC-HLA gene family, for which it has been demonstrated the impact of demographic forces, in addition to both positive and balancing selection (Dean et al. 2002).

In conclusion, investigation on the footprints left by natural selection in human populations genetic diversity represents an extremely precious chance to explore the genetic basis of adaptation and its crucial medical implications. However, this requires a deeper and deeper understanding of

genotype/phenotype relationships, as well as the development of even more powerful tests of selection.

Thankfully, further developments in proteomics, functional genomics, and chip-based technology for a simultaneous screening of the expression of thousands of genes, have recently improved our capability to dissect the roles that selection and demography have played in shaping our genome.

1.2 The Human Genome Variation

1.2.1 Genomic and post-genomic eras

Since the development of the first DNA sequencing technology by Sanger et al. (1977), an unrestrained race to achieve the knowledge of the ins and outs of an organism complete genome has begun, opening the way to the *genomic era*.

Genomic research formally came into existence in 1986 as the study of a living being primary genetic makeup focused on both genome sequence structure and its functional annotation (Groisman and Ehrlich 2003).

The first whole-genome sequencing (WGS) technology was successfully applied in 1995 and led to the achievement of the first entire genomic sequence of an organism, the bacterium *Haemophilus influenzae* (Fleischmann et al. 1995).

This strongly accelerated the progress of another ambitious project, the Human Genome Project (HGP), which has been started in 1992 (Little 1992) and, at first, produced the complete human genome sequence draft in 2001 (International Human Genome Sequencing Consortium 2001), then it completed the final version on April 14, 2003 (International Human Genome Sequencing Consortium 2004).

This success was mainly due to the advent of high-throughput (HTP) sequencing technology platforms, as well as of high-speed bioinformatics platforms able to manage their huge amount of data and to perform sequence assembly and gene annotation. From that moment, full many studies focused on the nature and amount of human genetic polymorphisms have been published (Goldstein and Cavalleri 2005; Hinds et al. 2005; Conrad et al. 2006; Redon et al. 2006; Lao et al. 2007; Myers et al. 2008; Jakkula et al. 2008; Tian et al. 2008; Keinan et al. 2009).

To date, many mammalian genomes, including mouse (*Mus musculus*), dog (*Canis familiaris*) and especially chimpanzee (*Pan troglodytes*) (The Chimpanzee Sequencing and Analysis Consortium 2005) and rhesus macaque (*Macaca mulatta*) (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), have also been completely sequenced, leading to the birth of comparative genomics. This makes the attempt to identify DNA regions that actually contributed to our evolution possible; potentially shedding light on the role that natural selection has played during this process (Sabeti et al. 2006).

At the same time, sequencing of Neanderthal mitochondrial genome (Krings et al. 1997; Serre et al. 2004; Orlando et al. 2006), as well as of some shares of its nuclear genome (Green et al. 2006;

Noonan et al. 2006), has attempted to elucidate evolutionary relationships between Neanderthals and us.

To sum up, a deep characterization of individuals and populations genetic variation has been carried out and has led to reconstruction of our evolutionary history and phylogeny, providing a direct genetic witness of the origin of *H. sapiens* (Garrigan and Hammer 2006).

Moreover, it has also laid the foundation for functional genomics, the assessment of the function of genome regions, and transcriptional genomics, leading to a better understanding of complex interactions between genetic and environmental factors in producing phenotypes and hence of differential susceptibilities to disease and responses to pharmacological agents.

In particular, these data have revealed a hierarchical organization of the human genome as a DNA modules system (Shapiro 2005), whereas genome-wide RNA expression profiles highlighted the clustering of highly expressed genes in specific chromosomal regions, suggesting that genes with similar or linked expression are often grouped together (Hurst et al. 2004).

That being so, a crucial role for the phenotypic outcome is plausibly played by genetic variation also through its effect on gene expression, with single nucleotide polymorphisms (SNPs) accounting for about 84% and copy number variants (CNVs) accounting for about 18% of gene expression variation among and across human populations (Stranger et al. 2005).

Comparative and transcriptional genomics have also confirmed that anatomical differences between humans and chimpanzees are mainly due to differences in the regulation of genes function, as well as that human inter-individual variation in gene expression is in part governed by regulatory genetic determinants, which may be trans- or cis-acting, and which may harbor common haplotypes which affect a gene total expression (Pastinen et al. 2006).

At present, these results have thrown disciplines such as Biology, Molecular Anthropology and Medicine into the *post-genomic era*, inducing a radical shift in theoretical and methodological approaches to the study of human origin, evolution and susceptibility to diseases.

However, the field evolves very rapidly, and our comprehension of evolution of the human genome, which seems to be actually depicted by a colorful mosaic of a multitude of pieces, with different age and telling different stories, is gradually emerging (Paabo 2003).

Undoubtedly, understanding heritable variation in the human genome, as well as genetic basis of physical and behavioral traits that distinguish human beings from each other and from other primates, will be one of the great challenges of science in immediate future.

1.2.2 Nucleotide variation

Genomics and post-genomics data suggest that SNPs would be the main source of genetic and phenotypic human variation, so that their amount seems to be one the most precise measure of the general extent of human genetic polymorphism. A total number of over 10 million SNPs was observed by The International HapMap Consortium (see section *1.3.5)*, most of which are rare, with a minor allele frequency (MAF) lower than 5%, and located within non-coding regions such as gene introns or intergenic regions (Figure 1.2.2.1) (The International HapMap Consortium 2005).

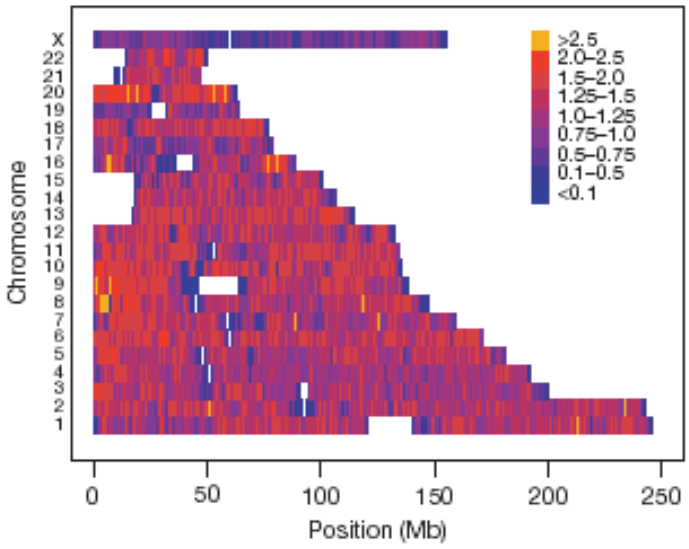

Figure 1.2.2.1 SNP density across the genome in the HapMap Phase II.
Colors indicate the number of polymorphic SNPs every 1,000 bases in the consensus dataset, with white blocks indicating gaps in the assembly.

Since SNPs are characterized by a relatively low mutation rate (10^{-8} substitutions per locus per generation), the majority of nucleotide differences between individuals are not *de novo* mutations, but inherited changes. Thus, two individuals that share the same allele at a given nucleotide position are most likely identical by descent, for that particular DNA segment, rather than carriers of two identical independent mutations.

This is the main reason behind SNPs utility in reconstruction of human evolution, for example through the observation that a greater number of SNPs is found in people of African origin respect to people of European origin (Figure 1.2.2.2), reflecting the common African past of these ethnic groups (Crawford et al. 2005). As a matter of fact, investigation of more than 1.5 million SNPs in Americans of European, African, and Asian ancestry has revealed that 93.5% of them are observed in individuals of African ancestry, 81.1% are found in those of European ancestry, and only 73.6%

in those of Asian ancestry. In addition, African-Americans also showed more private SNPs, nucleotide substitutions which are segregating in one population only, than European-American or Asian-American individuals (Hinds et al. 2005). This pattern of higher level of genetic diversity for African populations respect to non-African ones, the latter consequently showing a subset of the genetic diversity present in Sub-Saharan Africa, was also confirmed by several re-sequencing studies, for example on non-coding regions (Zhao et al. 2006) and on 3,873 genes in European, Latino/Hispanic, Asian, and African-American populations (Guthery et al. 2007).

Nevertheless, DNA sequence similarity among people from around the world is still sensationally high, with any two individuals which are thought to be about 99.9% identical in their DNA sequence (Reich et al. 2002).

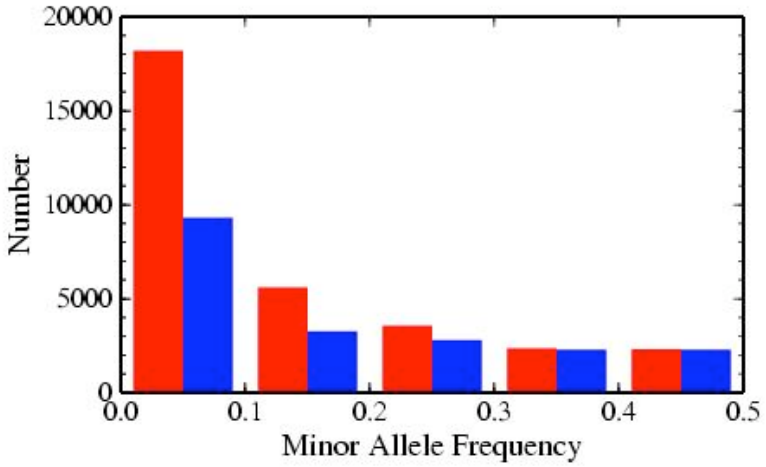

Figure 1.2.2.2 SNP frequency distribution in SeattleSNPs.
The number of SNPs is plotted within each frequency range. African-Americans SNPs are shown in red, Europeans SNPs in blue.

1.2.3 Haplotype variation and linkage disequilibrium (LD)

Haplotypes are described as specific allele combinations for a given set of polymorphic sites. Reciprocal association between these alleles is disrupted only by mutation or recombination events that occurred in some of the nucleotide sites constituting the haplotype. Such a non-random association is known as linkage disequilibrium (LD) and can be measured by indexes such as r^2 and D'. The former is very useful in medical genetics because of its inverse relationship with the power of association studies, while the latter is mainly used to describe historical recombination events in populations (Reich et al. 2001).

Interestingly, although African, European and Asian populations are characterized by different

haplotype frequencies and, to some extent, by different combinations of SNPs inside haplotypes, data from the HapMap project (see section *1.3.5*) have shown that both common and rare haplotypes are often shared across these groups (The International HapMap Consortium 2005). At the same time, it was found that haplotype diversity decreases as distance from Africa increases and, even if the extent of LD varied across the populations, inferred recombination hotspots, genome regions in which historical crossing-over events are clustered and which separate large haplotype blocks, generally match across different continental groups (Conrad et al. 2006).

Moreover, it has been proved that patterns of LD depend on both demographic factors, such as population size and structure, and locus-specific features due to selection, mutation, recombination and gene conversion events, resulting particularly useful for inferences about human evolutionary and demographic processes (Abecasis et al. 2005; Tishkoff and Verrelli 2003b). For example, lower levels of LD are observed in Africans respect to non-Africans and haplotype blocks, regions in which SNPs are in strong LD, extend over greater distances and are more uniform in the latter (Figure 1.2.3.1) (Sawyer et al. 2005).

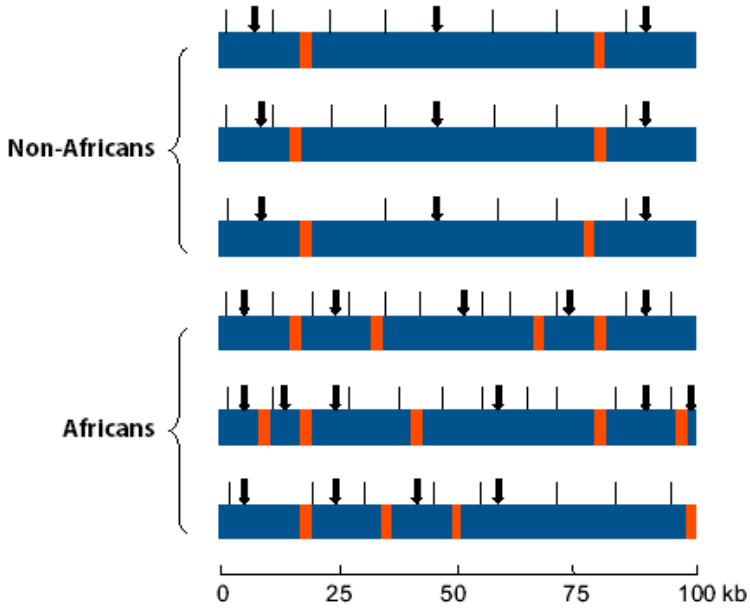

Figure 1.2.3.1 Haplotype blocks in African and non-African populations.
Blue bars represent the haplotype blocks, while orange bars represent the recombination hot spots. Vertical lines indicate SNPs and vertical arrows indicate haplotype tag SNPs.

It has also been shown that Africans have higher population recombination rates (ρ) compared to Europeans and Asians, in accordance to the fact that recombination is a remarkable determinant of LD extent. In particular, recombination hot spots, 1–2 kb DNA segments with a higher

recombination rate respect to surrounding regions, turned out to be not homogeneously distributed across populations and in the genome, covering a very small fraction of it, but accounting for over 80% of all recombination events.

Described divergent LD patterns and recombination levels between African and non-African populations can be explained by different demographic histories of such groups. The former has shorter blocks of LD because of larger effective ancestral population size (N_e) and because there has been more time for recombination to disrupt LD, while the latter shows greater LD values as the result of founding events occurred during the expansion of modern humans out of Africa within the past 100,000 years (Figure 1.2.3.2) (Tishkoff and Verrelli 2003b).

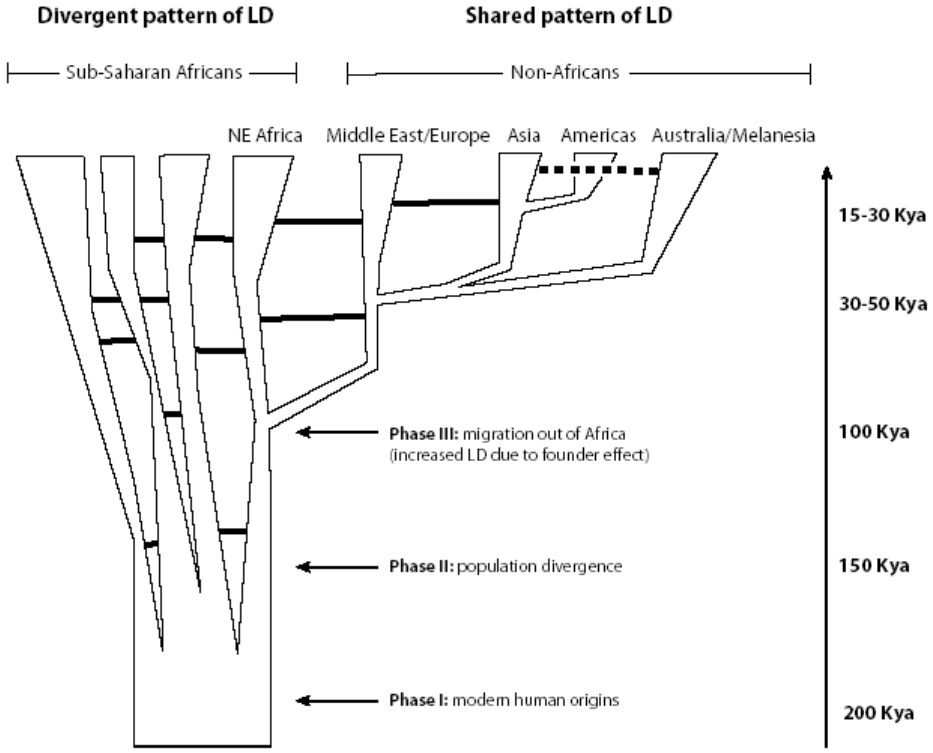

Figure 1.2.3.2 A serial founder model for *H. sapiens* migrations.
Non-Africans geographic expansion occurred in different small steps, each of one involved a sampling of the previous populations variation.
Horizontal lines indicate gene flow between populations.

H. sapiens migration from Africa to Eurasia and the rest of the world is indeed thought to be accompanied by a population bottleneck that produced an inevitable loss of genetic diversity (Liu et al. 2006). Census size of the group/s migrating out of Africa was estimated on the basis of combined mtDNA, Y chromosome, and X chromosome nucleotide diversity analyses and was of about 4,500 individuals, corresponding to 1,500 effective founding males and females (Garrigan et

al. 2007). Such a value implies that Eurasians must have rapidly expanded to a larger size to account for estimates of a long-term N_e of 10,000 individuals and of a census size of 30,000 individuals for their global population (Zhao et al 2006).

In more details, the Out of Africa model described for *H. sapiens* origins assumes that fully modern human traits appeared in East Africa and Southwest Asia around 90,000 years ago, then a rapid spread of modern humans throughout the rest of Africa and Eurasia occurred within the past 40,000-80,000 years (Reed and Tishkoff 2006). Two different routes have been proposed for such diffusion. A southern coastal route around the Indian Ocean, by which modern humans first left Africa via Ethiopia and then rapidly migrated to Southeast Asia and Oceania, is supported by recent mtDNA data (Macaulay et. al 2005), whereas other hypotheses have traditionally favored a second/single northern route via the Sinai Peninsula into the Levant (Figure 1.2.3.3).

Figure 1.2.3.3 A simplified scenario for early human migration routes and dates.
KYA stands for thousands of years ago.

Along these sensational evolutionary inferences, analysis of haplotypes and LD turned out to be incredibly useful also for mapping disease susceptibility loci, as it will be further discussed in section *1.3.2*.

A human genome organization in 10-100 kb haplotype blocks has been indeed proposed and states that the 2-5 most common haplotypes within each block are able to capture the great majority of that DNA region variation, accounting for more than 90% of all analyzed chromosomes (Gabriel et al. 2002).

Thus, a minimal subset of SNPs (*tag SNPs*), which are in strong LD with all the other nucleotide sites within the block and typical of the most common haplotypes, can be used to survey all related SNPs of that entire region. In this way, a full genotyping procedure will be avoided.

1.2.4 Structural variation

The advent of whole-genome scanning technologies has uncovered an unexpectedly large extent of multiple-scale structural variation in our genome (Komura et al. 2006).

Microscopic and submicroscopic variants, such as deletions, duplications and large-scale copy number variants (CNVs), as well as insertions, inversions and translocations have been observed, accounting for millions of nucleotides of heterogeneity within every single genome (Tuzun et al. 2005; Khaja et al. 2006). About 100 CNVs per individual, each over 50 kb in size, a significant number of intermediate sized CNVs and inversions, from 8 to 40 kb, and even smaller variants, from 1 to 8 kb, were indeed found, leading to the conclusion that approximately 3.75×10^6 base pairs of structural polymorphism exist between any two diploid human genomes, accounting for about 12% of the entire genome (Feuk et al. 2006). That being so, structural variants actually represent an important contribution to human genetic diversity (Figure 1.2.4.1).

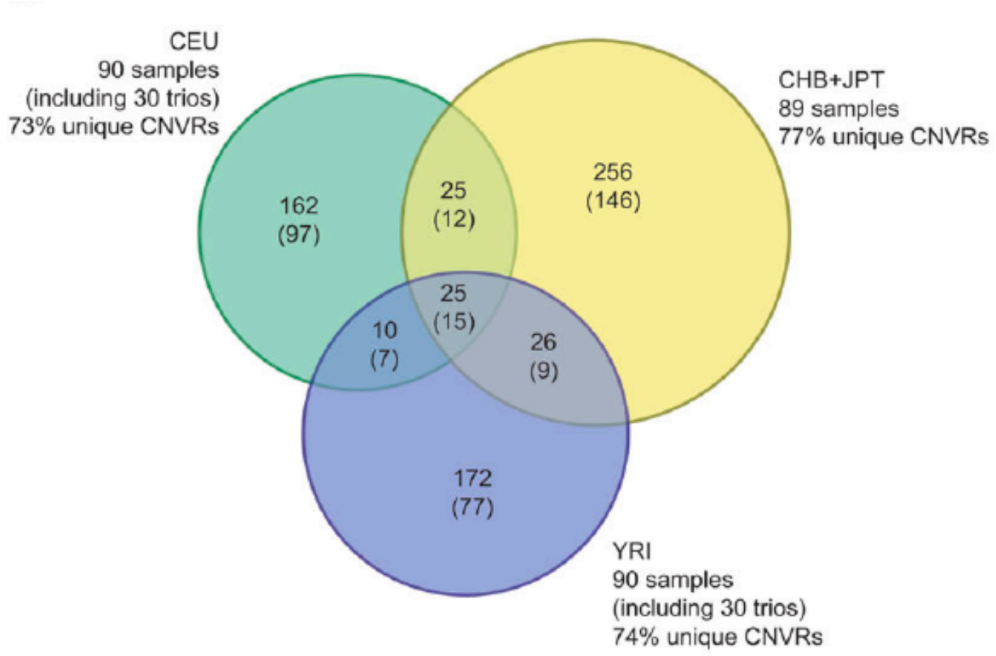

Figure 1.2.4.1 Overlapping degree of CNVs regions in HapMap populations.
Numbers in brackets correspond to the stringent CNV region dataset for each population group.
CEU: European, JPT+CHB: Asian, YRI: African.

In particular, large segmental duplications constitute 2.7% of difference between *H. sapiens* and *P. troglodytes*, in comparison to only 1.2% represented by SNPs (Sharp et al. 2006).

As a matter of fact, comparison of structural variation in human, chimpanzee and rhesus macaque genomes has led to the discovery of 130 human specific breakpoints, identifying 58 genes affected by insertions, with 36 gene copies fully contained within insertions, and 22 genes that were either partially duplicated or contained an insertion. The average size of detected rearrangements is of 110,063 bp, with a range spanning from 20 to 1,365,171 bp (Harris et al. 2007).

Moreover, an enrichment of large interspersed segmental duplications with high level of sequence identity was observed comparing human and other primates genomes with those of other mammals (Bailey and Eichler 2006), suggesting that they have been responsible for the creation of novel primate gene families.

Most importantly these segmental duplications might also have influenced human genotypic and phenotypic variation on a previously unappreciated scale, playing a crucial role in separation of lineages leading to humans, on the one hand, and to chimpanzees on the other.

1.3 Genetic Association Studies

Genetic defects causing rare Mendelian recessive or dominant diseases, such as cystic fibrosis and Huntington's disease, were far back recognized by means of classical pedigree-linkage analyses.
Unfortunately, such a typology of studies has failed in identification of susceptibility alleles for complex common diseases, which represent the major source of morbidity and mortality in developed countries.
For this purpose, genetic association studies have been adopted to detect an association between one or more genetic polymorphisms, either individually or as haplotypes, and a trait that might be some quantitative characteristic or even a disease phenotype. The rationale behind such a typology of analyses is the assumption that the same allele is associated with a given trait in a similar manner across the whole population.
In particular, population-based association studies quickly turned out to be more powerful than pedigree-linkage analyses for detecting also weak genetic polymorphisms effects on the development of diseases (Sham et al. 2000), so that they initially played an important role in fine mapping genetic loci previously detected by pedigree analyses.
The power of association studies is mainly due to the fact that linkage disequilibrium extends over shorter distances in distantly related individuals. Nevertheless, the coeval increased possibility for linkage to be destroyed by recombination implies that these studies need a greater density of markers to be performed respect to family-linkage ones.
In the following sections different approaches to deal with candidate genes association studies will be fully discussed.

1.3.1 Direct association studies

Direct association studies are characterized by the fact that putative disease-causing mutations are directly genotyped.
Candidate causal variants are generally non-synonymous mutations, but also synonymous and non-coding changes may induce differential splicing or variation in gene regulation and expression, resulting responsible for heredity of common complex disorders and increasing the difficulty of identifying candidate polymorphisms.
Moreover, SNPs in coding regions are fewer and rarer than non-coding SNPs (Wellcome Trust Case Control Consortium and Australo-Anglo-American Spondylitis Consortium 2007), so that their detection would require twice the sample size actually used by SNP discovery projects, such as

SeattleSNPs, which are calibrated to successfully detect common variants, that is changes with minor allele frequency (MAF) > 5%.

1.3.2 Indirect association studies

Indirect associations are referred as cases in which the plainly associated polymorphism has not a causal role but is simply in strong LD with a nearby causal variant (Figure 1.3.2.1).

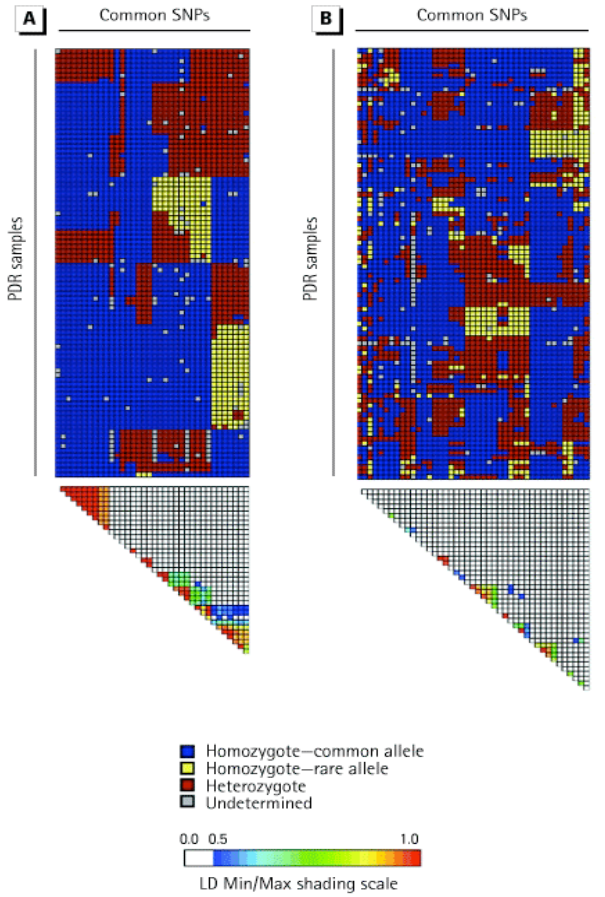

Figure 1.3.2.1 LD examples measured by r^2 for common SNPs (MAF >5%).
At the top of the figure columns represent nucleotide sites, while rows represent genotypes.
At the bottom of the figure LD graphs measured by r^2.
A) Genomic region with average LD and few blocks of correlated SNPs.
B) Genomic region with low LD and less SNPs correlation.

As a consequence, it will be necessary to genotype several surrounding neutral markers to have a high chance of picking up the indirect association. Nevertheless, there cannot be a definitive negative result, since we cannot exclude the possibility that a causal variant exists, but is not picked up by chosen markers. Therefore, indirect association studies should be designed in terms of both

sample size and markers coverage to have sufficient power to detect common disease susceptibility alleles even of modest effect (Byng et al. 2003).

Population-based indirect association studies are focused on candidate genes, identified either on the basis of their known biological function or from animal models.

Exploiting the fact that genotyped SNPs are in LD with disease-causing SNPs, they expect that those variants would be over represented among disease individuals respect to healthy people (Carlson et al. 2004). However, it necessarily implies that investigated diseases have an actual genetic component, but it is not so easy to be proved, since even a strong heredity does not indicate that there is a single major gene underlying the disease. Consequently, these possible limits in capability of estimating the strength of a disease genetic component could negatively affect association studies results (Dahlman et al. 2002).

Another key assumption of indirect association studies that will be subsequently discussed in section *1.3.4*, is that only a few common variants are associated with a common disease (Risch and Merikangas 1996). However it does not take into account that several rare variants at several nucleotide sites might also contribute to the disease phenotype.

In particular, if the causative SNP is more infrequent than genotyped ones, it is likely that only studies searching for genetic determinants of large phenotypic effects will be successful. Moreover, although chromosomal regions with large stretches of LD are ideal for association studies, because of the smaller number of SNPs to be genotyped, they make subsequent isolation of causative SNPs, from SNPs simply in LD with them, extremely difficult (Zondervan and Cardon 2004).

1.3.3 Confounded associations

Unfortunately, confounded associations can arise due to substructures of surveyed populations such as stratification or admixture. For example, in a mixed population, in which strata have different environmental exposures or founder populations entailing different genetic risks, any locus whose allele frequencies differ between strata, or founder populations, will be associated with the examined disease. This raises the possibility of generating false findings or, conversely, of obscuring true causal associations.

Such a confounding effect could be reduced by matching samples by geographical regions or by markers of ethnic origin, so that comparisons can be made, as far as possible, within homogeneous subpopulations (Wacholder et al. 2002).

However, as already discussed, causal variants for complex disease might have small effects that

require large studies to be detect. That being so, even modest confounding by stratification and admixture could have important repercussions on association studies results.
Another possible method for facing this problem is to seek genetic markers for population substructure or ancestry informative markers whose allele frequencies differ between founder populations (Hoggart et al. 2003).
Since confounding is regarded as a random process, potentially affecting all loci, its effect is increasing the false positive rate. Genomic control aim to correct this effect by increasing the threshold required for statistical significance. The extent by which variance is inflated by confounding can further be measured by typing a large number of unselected markers across the genome to empirically estimate the variance of association test statistics (Bacanu et al. 2002).
That being so, taking into account genetic markers for population substructure and using genomic control could turn out to be complementary approaches, with great effects addressed by statistical models and surrogate measures of substructure and more subtle effects, such as those due to cryptic relatedness between cases and controls, left to genomic control.

1.3.4 The Common Disease/Common Variant hypothesis (CD/CV)

The Common Disease/Common Variant (CD/CV) hypothesis, which assumes that much of the genetic variation of common complex diseases is due to a limited number of common variants, present in more than 5% of the population, was suggested in 1996 (Risch and Merikangas 1996; Lander 1996). Ever since, genetic association studies have achieved greater and greater importance, even if their full application should require testing every gene and identifying all common variants in the human genome.
The CD/CV hypothesis implies that more common genetic variants, despite having only moderate disease risk respect to susceptibility variants found by pedigree-linkage analyses, have larger effect on disease risk at a population level, so that they may be far more important in terms of public health simply because they are more common. According to such a hypothesis, common genetic variants were actually found to increase some common diseases risk, as in the case of APOE variants that increase risk for Alzheimer's and heart disease (Lohmueller et al. 2003).
However, there also exist examples of rare variants influencing common diseases (Romeo et al. 2007), suggesting that both rare and common variants may play a role in common diseases manifestations, even if it is not known which of them are more important.
That being so, in spite of recent huge financial and scientific investments in genome-wide association studies, an incontrovertible evidence in support of the CD/CV hypothesis was not found

and, if rare genetic variants were the primary cause of common complex disease, so that for a disease to be common there would be many different causative alleles, association studies would have little power to detect them.

1.3.5 The HapMap project

Regardless of the absence of an incontrovertible evidence in its support, the CD/CV hypothesis gained credence up to the design of the International HapMap Project (http://www.hapmap.org) (International HapMap Consortium 2003), which pursues the ambitious goal of cataloguing all common human genetic variants.

As a matter of fact, strong linkage disequilibrium among SNPs in most chromosomal regions consents that few carefully chosen SNPs, named *tag SNPs*, need to be typed in each region to predict likely alleles at its remaining polymorphic sites, so that a precise map of LD patterns among SNPs can be obtained.

To achieve that, a consortium of researchers from Canada, China, Japan, Nigeria, the United Kingdom, and the United States was launched to produce a human haplotype map by genotyping 270 samples from four populations with different ancestry (Yoruba from Nigeria, Utah residents of Northern and Western European origin, Han Chinese and Japanese individuals).

The HapMap Phase II published more than 3 million SNPs (International HapMap Consortium 2007), proving that *tag SNPs* chosen in this dataset are generally applicable across worldwide populations, even if with limitations for rarer SNPs and for populations with substantial proportions of recent African ancestry (Figure 1.3.5.1) (Conrad et al. 2006; deBakker et al. 2006).

Taking into account such a remark, it seems that the development of different panels of *tag SNPs*, together with a more dense SNPs coverage, will be necessary for populations with low levels of LD, especially for African ones, to achieve the same proportion of variation tagged with fewer SNPs in higher LD populations.

For this reason, 1,301 additional samples were further collected from the four populations mentioned above and from seven additional populations (Luhya and Maasai from Kenya, Tuscans from Italy, individuals of Gujarati Indian, Chinese, Mexican and African ancestry from USA). These additional samples are being genotyped on the Affymetrix 6.0 platform and the Illumina 1 million SNP chip, and promise to be crucial for identification of *tag SNPs* that are more informative across ethnically diverse populations (Manolio et al. 2008).

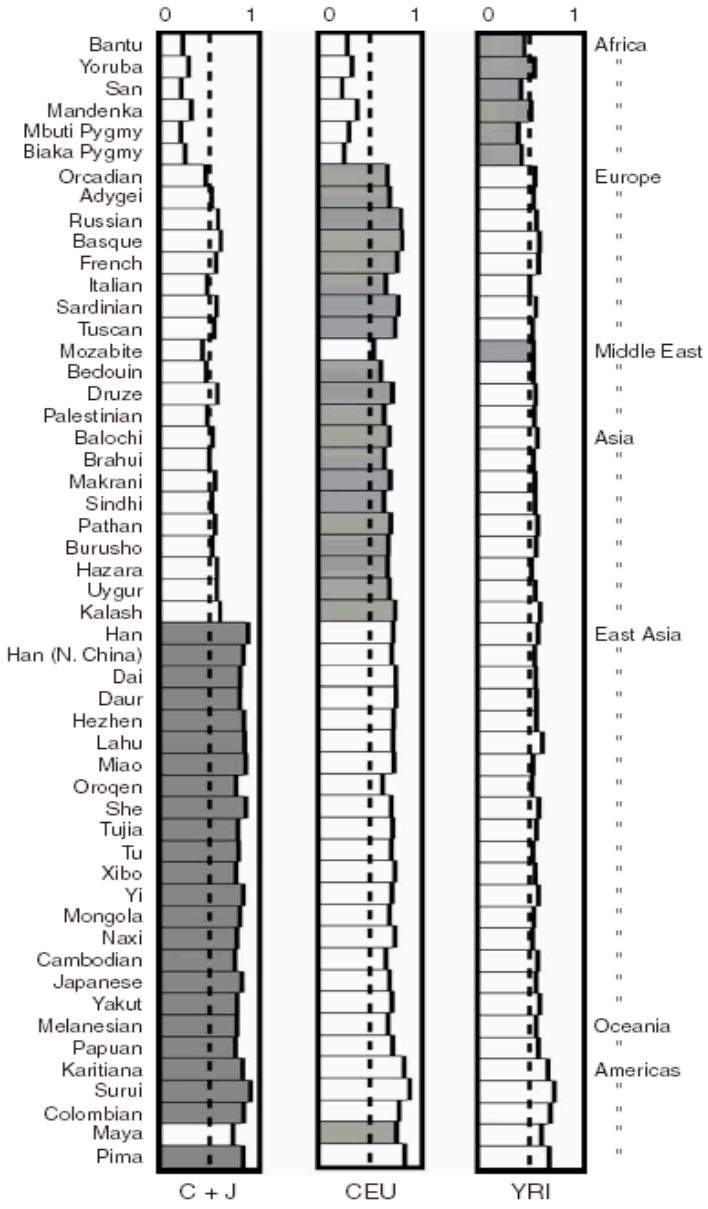

Figure 1.3.5.1 Portability of tag SNPs chosen using the HapMap Phase II 3.1 million SNPs dataset. For each of the 52 populations, columns show the proportion of polymorphic non-tag SNPs that have $r^2 > 0.85$ with at least one tag SNP. For each population, the gray bar indicates which tag SNP set is best. Vertical dashed lines indicate 50% tag portability. Tag SNPs were chosen separately from each of the three HapMap groups (CHB+JPT, CEU, YRI).

1.3.6 Genome-wide association studies (GWA)

The human haplotype and LD map achieved by The International HapMap Project has provided a totally new approach for searching genetic variants related to complex diseases.

As a matter of fact, together with the advent of high-throughput SNP chip genotyping technologies, which simultaneously assay hundreds of thousands of SNPs, it has made genome-wide association studies (GWA) possible, leading to new insights into genomic variation, structure, function and

interaction with environmental factors in diseases causation. However, it is necessary to underline the fact that, in addition to the classical genetic association studies limits described in previous sections, GWA studies have also an enormous potential for generating false-positive, since they simultaneously test hundreds of thousands of statistical hypotheses, one for each allele assessed.

To date, although it is considered a very high conservative method, applying the Bonferroni correction, in which conventional *p*-value is divided by the number of tests performed, is the most used approach to face this problem (Hunter and Kraft 2007).

Despite that, results of first GWA studies have actually highlighted common genetic variants involved in several common diseases (Figure 1.3.6.1), such as coronary heart disease (McPherson et al. 2007; Samani et al. 2007), breast cancer (Easton et al. 2007) and type II diabetes (Salonen et al. 2007; Saxena et al. 2007).

A promising feature of GWA studies is that they are not limited to known genes or regulatory regions, but they actually represent an "agnostic" approach to identifying common diseases related genetic variants, being unbiased by prior assumptions about DNA changes and unconstrained by current imperfect understanding of genome structure and function (NCI-NHGRI working group on replication in association studies 2007; Altshuler and Daly 2007).

Interestingly, GWA studies successful results suggest that the CD/CV hypothesis is true to an extent, at least for some of the studied diseases. However, even for disorders in which common genetic variants have been found, most genetic variation is still uncovered and it is not possible to rule out the possibility that much genetic variation is due to rare variants. This does not imply that the CD/CV hypothesis is necessarily false; rather that power is low for current study size, unless the allele MAF is high or its effect is large (Iles 2008).

Therefore, while many common diseases variants have been found, there may be many more variants that are of moderate frequency, but that current studies are not large enough to find.

Certainly, sample sizes will increase, leading to greater power to find rare variants. However, as samples become larger and larger, such an increased power may lead also markers in weak LD with disease alleles to reach significance. Thus, as sample sizes increase, rare variants are more easily detected, but the most significantly associated markers may not be rare themselves. Moreover, there may be a limit to how large population-based studies can get, and so there may exist a further class of variants that are too rare to be captured by GWA studies, but that are also not sufficiently high risk to be captured by linkage analyses (Cambien and Tiret 2007).

To date, successes in finding common variants associated with common diseases are encouraging, but it is not yet sure whether observed variants represent only the tip of an undiscovered iceberg. That being so, new approaches will be necessary to find these kind of variants, for example finer

and finer bioinformatics-based methods might become useful to identify both candidate genes and changes. Finally, current high-throughput genotyping platforms are not yet effective at suitably genotyping structural variants. Although these platforms recent improvements are just focused on a deeper CNVs detection (Cooper et al. 2008), sequencing still probably remains the only way to avoid this problem, even if it only slightly improves power and will not, on its own, remove the bias towards finding more common variants.

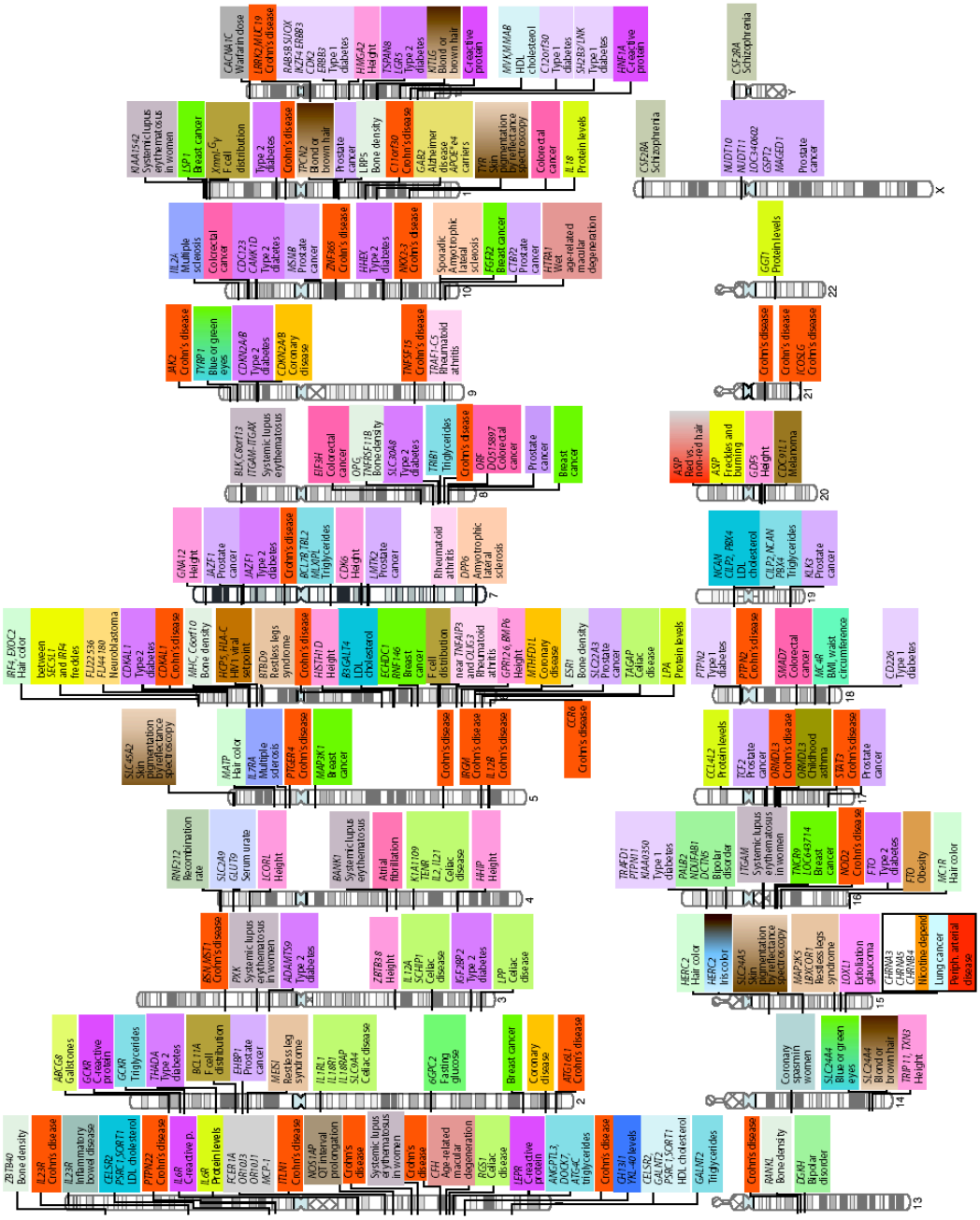

Figure 1.3.6.1 SNP-trait associations detected in genome-wide association studies up to the present. Associations significant at $p<9.9\times10^{-7}$ are shown according to chromosomal location and involved or nearby gene. Colored boxes indicate similar diseases or traits.

1.3.7 Whole-genome sequencing (WGS)

Although SNPs used in the HapMap have been highly informative for association mapping studies, initial identification of SNPs in one, or a few populations, can result in an ascertainment bias toward high frequency changes.

Several studies have shown that such an ascertainment bias can distort estimates of migration (Wakeley et al. 2001), mutation (Nielsen 2000) and recombination rates (Clark et al. 2003), as well as of LD patterns (Akey et al. 2003).

Moreover, as already discussed in the previous section, there exists a limit for GWA studies sample sizes beyond that their reliable associations detection will decrease and they also seem to be not so effective at genotyping structural variants.

These are the reasons why new sequencing methods have been developed to more accurately infer human genetic variation by typing structural variants and by characterizing the entire frequency distribution of nucleotide variants in different populations. This certainly improves the attempt to identify rare potentially causative variants that are now poorly tagged by existing genotyping platforms.

WGS of all samples of the extended HapMap dataset has been suggested to develop a comprehensive catalog of rare variants and will soon begin as part of the international 1000 Genomes Project (http://www.1000genomes.org).

To date, primary data production for sequencing of most genomes has relied on the same type of capillary sequencing instruments used by the HGP and based on bacterial artificial chromosome (BAC) clones. Each BAC clone was usually amplified in bacterial culture, isolated, and sheared to produce size-selected pieces of approximately 2-3 kb. Subsequently, such pieces were sub-cloned into plasmid vectors and then amplified in bacterial culture. Finally, DNA was selectively isolated prior to sequencing.

Nowadays, that scenario is rapidly changing owing to the advent of so-called next-generation sequencing technologies (Bentley 2006).

As a matter of fact, in current WGS approach, genomic DNA is sheared directly into several distinct size classes and placed into plasmid and fosmid sub-clones. Over sampling the ends of these sub-clones, to generate paired-end sequencing reads, provides the necessary linking information to fuel whole genome assembly algorithms (Mardis 2008).

In this way, genomes can be sequenced more rapidly and more readily, even if highly polymorphic or highly repetitive genomes remain quite fragmented after assembly.

1.4 The Immune System

The immune system of vertebrates consists of two main components: the innate and the adaptive immune system. In such a context, many types of proteins, cells, organs, and tissues are responsible for immune responses by interacting in an elaborate and dynamic network.

1.4.1 Innate immunity

Innate immunity is typical of all plant and animal classes as it provides an immediate non-specific defense against infection, recognizing and responding to pathogens without conferring long-lasting immunity (Beck and Habicht 1996).
This kind of response is usually triggered when microbes are identified by pattern recognition receptors that recognize components conserved among broad groups of microorganisms (Medzhitov 2007), or when damaged, injured or stressed cells send out alarm signals, many of which are recognized by the same receptors that recognize pathogens (Matzinger 2002).

1.4.2 Adaptive immunity

Adaptive immunity evolved in early vertebrates to provide both a stronger response respect to innate immunity and an immunological memory to pathogens, since each of them is identified by a specific signature antigen.
The major achievement in adaptive immune system evolution is indeed the capability to generate specific immunoglobulin isotypes, such as IgM, IgG, IgA and IgE, which are directed against invading pathogens, and, subsequently, to memorize this response. As a matter of fact, immunoglobulins represent an important component of the humoral immune system, together with serum complement factors, and are secreted either by plasma cells or by activated memory B cells, being kept at constant concentration in blood.
This antigen-specific response requires recognition of specific “non-self” antigens during the “antigen presentation” process and such specificity enables the generation of responses that are tailored to specific pathogens or pathogen-infected cells.
At the same time, memory cells maintain this ability during the time, so that if a pathogen should infect the body more than once, specific memory cells are used to quickly eliminate it.
Processes such as somatic hypermutation and V(D)J recombination of antigen receptor gene segments confer to humoral response a high adaptability, since a small number of genes is able to

generate a huge number of different antigen receptors that are uniquely expressed on each individual lymphocyte. Moreover, since gene rearrangement leads to an irreversible change in the DNA of each cell, all that cell progeny will inherit genes encoding the same receptor specificity, including Memory B cells and Memory T cells that are keys to long-lived specific immunity.
Although such a high evolutionary level of specialization, systemic cells and processes of the adaptive immune system are still activated by the "non-specific" and evolutionarily older innate immune system (Pancer and Cooper 2006).

1.4.3 B cells development

B cells develop in bone marrow from pluripotent haemopoietic stem cells through rearrangement of Immunoglobulin heavy-chain and light-chain genes and initial selection of the repertoire with selection against auto reactive B cells.
After leaving bone marrow as transitional B cells, they circulate in peripheral blood and become naive B cells, most probably with the help of splenic environment. At that stage they need the B cell activating factor BAFF as a survival signal to prevent apoptosis (see also section *1.7.2*).
Mature B cells expressing both IgM and IgD enter secondary lymphoid organs. Here, affinity maturation takes place through somatic hypermutation of variable region genes. In particular, in the germinal center of secondary lymphoid follicle, B cells receive the help of T cells that provide the correct set of co-stimulatory molecules, such as CD28 and ICOS, to select lymphocytes with the correct receptor. Subsequently, these B cells proliferate and undergo class switch recombination and somatic hypermutation that enable IgG, IgA and IgE isotypes production (Figure 1.4.3.1).
B cells selected through affinity maturation can become either memory B cells, which circulate in blood and engage in secondary immune responses to provide an immunological memory to infections and vaccinations, or long-lived plasma cells, the survival of which depends on the survival factor provided by a proliferating inducing ligand (APRIL) and signaled through B cell maturation antigen (BCMA), for coming back to bone marrow and produce high-affinity antibodies (Figure 1.4.3.1) (Mckay et al. 2003).
Differentiation of mature B cells into effectors capable of specific humoral immunity is strictly regulated and Tumor Necrosis Factor Receptor Superfamily (TNFRSF) members play important and diversified roles in regulation of activation and apoptosis for specific cells of the immune system (see also section *1.7.3*).

As it will be further discussed in following sections, the failure to produce effective immunoglobulins leads to increased host susceptibility to infections and to severe immunological diseases onset.

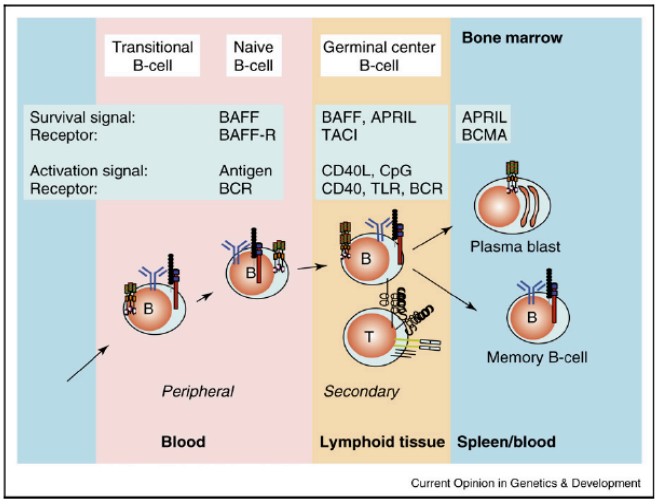

Figure 1.4.3.1 Schematic representation of B cells development.

1.5 Primary Immunodeficiencies Diseases (PIDs)

Human Primary Immunodeficiencies Diseases (PIDs) are a heterogeneous group of disorders in which inherited genetic defects compromise the ability to produce immune responses.

Since the description of first PIDs, such as Bruton's Agammaglobulinaemia and Kostmann's Neutropenia, about 200 different clinical entities have been classified according to detection of immunological abnormalities affecting blood circulating leucocytes or their products, the Immunoglobulins (Ig) (Ochs et al. 2007).

Several mutations associated with these diseases have been identified in about 130 genes, so that a genetic aetiology is known for the majority of PIDs (Table 1.5.1) and functions of many genes involved in immune responses have been elucidated (Geha et al. 2007; Fischer 2007).

In the last 50 years it has been depicted how PIDs plague innate immunity as well as adaptive immunity mechanisms, impairing both cell differentiation and regulatory functions.

Nowadays, antibody-related defects, that is humoral PIDs such as Common Variable Immunodeficiency, Selective IgA Deficit, Hyper-IgM Syndrome and X-Linked Agammaglobulinemia, which are characterized by B cells differentiation and Ig production defects, account for 65% of all primary immunodeficiencies, whereas defects in both cellular and antibody compartments account for another 15% of cases (Yin et al. 2001).

As mentioned above, PIDs wide spectrum of vulnerabilities to microorganisms has offered a precious tool for dissecting the immune system and has enabled in vivo assessment of immune response effectors specific roles.

One of the main conclusions drawn from these studies was that vulnerability to infectious diseases can vary over time, since predisposition to invasive infections by encapsulated bacteria or viruses due to TLR3 deficiencies or defects in the Toll-like receptor (TLR) pathway, in the Interleukin-1 Receptor associated Kinase 4 (IRAK-4) and in the Polytopic Endoplasmic Reticulum CERI-resident membrane protein (UNC93B), result to be limited in time (Casrouge et al. 2006; Ku et al. 2007). This observation suggested that once an adaptive immune response has occurred it is protective.

Moreover, PIDs genetic analysis is not only important for elucidating crucial pathways in the immune system, but also for its notable medical impact by prompting the design of new diagnostic tools and opening up new fields of therapeutic research.

Although the most severe forms of PIDs, such as Severe Combined Immunodeficiency (SCID), in which T lymphocyte development is compromised and associated with disorders of development and functionality of B lymphocytes and natural killer cells, are fatal without treatment,

haematopoietic stem cell transplantation is usually highly successful when a genotypically matched donor is available (Antoine et al. 2003).

Table 1.5.1 Primary Immunodeficiencies with known molecular defects.

Disorders	Molecules
Antibody deficiencies	
The agammaglobulinemias	
X-linked	Bruton's tyrosine kinase (BTK)
Autosomal recessive	IgM heavy chain, Ig-α, surrogate light chain, B cell–linker protein (BLNK), LRRC8
Hyper-IgM syndrome, autosomal recessive	AID, UNG
ICF syndrome	DNA methyltransferase 3B
CVID	TACI, ICOS
Cellular deficiencies	
IFN-γ/IL-12 axis	IFN-γ receptor α and β chains, IL-12 p40 subunit, IL-12 receptor α chain, signal transducer and activator of transcription 1 (STAT-1)
Autoimmune polyglandular syndrome type 1	Autoimmune regulator (AIRE)
Defective NK function (CD16 deficiency)	FcγRIII
Combined deficiencies	
SCID	
Defective cytokine signaling	
X-linked	Cytokine receptor common γ chain
Autosomal recessive	IL-2 receptor α chain, IL-7 receptor α chain, Janus kinase 3 (JAK3)
Defective T-cell receptor signaling	CD45, CD3γ, CD3δ, CD3ε
Defective receptor gene recombination	RAG1, RAG2, DNA cross-link repair 1C (DCLRE1C, ARTEMIS)
Defective nucleotide salvage pathway	Adenosine deaminase, purine nucleoside phosphorylase
Defective MHC class I expression	Transporter of antigenic peptides 1 and 2 (TAP1, TAP2), TAP-binding protein
Defective MHC class II transcription complementation groups A-D	Four components of the MHC class II gene transcription complex: CIITA, RFXANK, RFX5, and RFXAP
Other	Winged-helix nude transcription factor
Wiskott-Aldrich syndrome	Wiskott-Aldrich syndrome protein (WASP)
Ataxia-telangiectasia group	Ataxia-telangiectasia mutated (ATM), nibrin
DGS	Chromosome 22q11 deletion, T box-1 transcription factor (TBX1)
Hyper-IgM syndrome	
X-linked	CD40 ligand
Autosomal recessive	CD40
X-linked lymphoproliferative syndrome	SLAM-associated protein (SAP)
Defects of NF-κB regulation	NF-κB essential modulator (NEMO), IκB kinase α chain
Defects of Toll-like receptor signaling	IL-1 receptor associated kinase 4 (IRAK-4)
WHIM syndrome	CXC chemokine receptor 4 (CXCR4)
Caspase 8 deficiency	Caspase 8
Phagocyte defects	
Chronic granulomatous disease	
X-linked	Cytochrome b558 α chain (gp91phox)
Autosomal recessive	Cytochrome b558 β chain (p22phox), neutrophil cytosolic factors 1 (p47phox) and 2 (p67phox), ras-related C3 botulinum toxin substrate 2 (RAC2)
Chediak-Higashi syndrome	Lysosomal trafficking regulator (LYST)
Leukocyte adhesion deficiency 1, 2	β2 integrin (CD18), fucose transporter (solute carrier family 35, member C1)
Neutrophil-specific granule deficiency	CCAAT/enhancer binding protein (C/EBP)-ε
Cyclic neutropenia, Kostmann's syndrome	Elastase 2
X-linked neutropenia	Wiskott-Aldrich syndrome protein
Complement defects	All soluble complement components except factor B

ICF, Immunodeficiency, centromeric instability, and facial anomalies; *WHIM*, warts, hypogammaglobulinemia, infection, myelokathexis.

However, for most individuals this is not the case and survival from mismatched transplants is substantially lower (52%). For this reason, SCID and some other PIDs are particularly attractive targets for gene therapy because of the huge proliferative capacity of haematopoietic system, especially the lymphoid compartment, so that an effective gene transfer to a small proportion of bone marrow precursor cells can result in correction of the immunological deficit (Hirschhorn 2003).

In some instances, gene product supplementation might be envisaged, as is the case with Adenosine Deaminase Deficiency that results in a SCID phenotype. Direct genetic intervention by adding a normal copy of the mutated gene to affected cell lineages is an approach that has been proved to be successful in these conditions (Aiuti et al. 2002).

Different typologies of mutation in a given PID may also prompt new strategies. For example, mutations creating glycosylation sites can cause protein unfolding and degradation. Chemicals able to modify glycosylation may be able to complement these defects by preventing protein degradation, as shown in vitro for several models (Vogt et al. 2007).

Unfortunately, genetics of PIDs has turned out to be very complex, since environment and modifier genes should strongly change the spectrum of infectious diseases encountered in a specific PID.

This is the case of early-in-life infections by Cytomegalovirus (CMV) which trigger a massive expansion of oligoclonal non-Vg9Vd2 gd T cells in association with autoimmune manifestations (Ehl et al. 2005).

Moreover, epigenetic factors and modifier genes can also account for variability of many other PIDs, especially autosomal-dominant inherited ones, such as Common Variable Immunodeficiency and Autoimmune Lymphoproliferative Syndrome, suggesting that there may be a continuum between simple and more complex genetically determined diseases (Antonarakis and Beckmann 2006).

Finally, consequences of PIDs causative mutations can vary from a strong predisposition to infection by a broad range of microorganisms, such as in Severe Combined Immunodeficiencies (SCIDs), to an extremely narrow predisposition, which is for example typical, of Herpes Papilloma Virus (HPV) Disease (Ramoz et al. 2002), Herpes Simplex Virus 1 (HSV-1) Encephalitis (Zhang et al. 2007) and aberrant lymphocyte responses to Epstein-Barr virus infection (Rigaud et al. 2006).

The dominant view in genetics of infectious diseases postulates that rare monogenic PIDs predispose the individual to numerous infections, whereas common infectious diseases are associated with polygenic inheritance of numerous susceptibility genes.

Despite that, it has been proved that novel monogenic PIDs predispose the individual to a principal or single type of infection and major genes exert a nearly Mendelian impact at the population level, largely accounting for common infectious diseases in some individuals.

Recent discovery of such human genes, which confer vulnerability or resistance to a specific infection at the individual level, bridges the gap between the two classical fields of conventional PIDs and polygenic inheritance, providing experimental support for a continuous spectrum of predisposition and an unified theory of human genetics of infectious diseases (Figure 1.5.1).

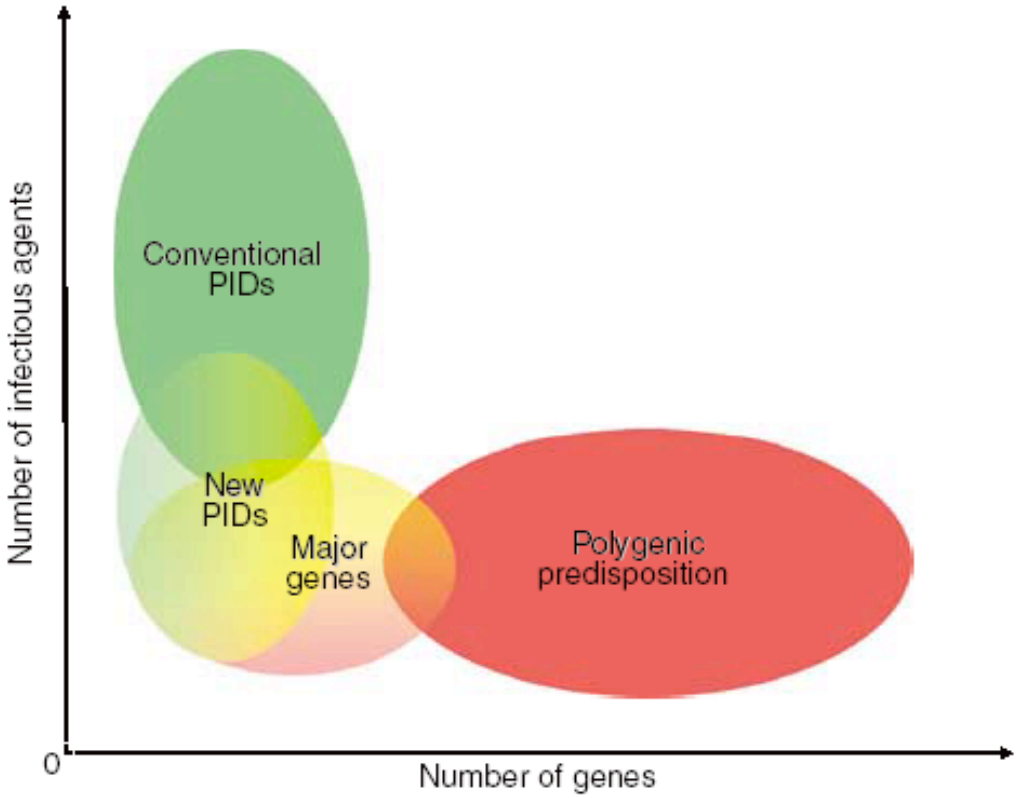

Figure 1.5.1 Spectrum of genetic predisposition to infectious diseases.

In addition, it has been taken into account that multiple paradigm shifts have been witnessed by the field of PIDs, since they were initially thought to be few rare, familial, monogenic, recessive traits impairing development or function of one or several leukocyte subsets and resulting in multiple, recurrent, opportunistic and fatal infections in infancy.

A dozen of epidemiological, clinical and genetic paradigm shifts have indeed occurred over the last decade, expanding the conventional view about Primary Immunodeficiencies and leading to a profound revision of their definition and classification (Casanova and Abel 2007) (Table 1.5.2).

Table 1.5.2 Paradigm shifts in Primary Immunodeficiencies.

Primary immunodeficiencies	Conventional	Novel	Examples
Epidemiological levels			
Frequency	Rare	Common	*FUT2* mutations and norovirus
Occurrence	Familial	Sporadic	*UNC93B1* and *TLR3* mutations and HSE
Age at onset	Childhood	Adulthood	CVID
Prognosis	Spontaneously worsening	Spontaneously improving	IRAK-4 deficiency
Phenotype level			
Disease-defining clinical phenotypes	Opportunistic infections[a]	Other life-threatening infections, other phenotypes[b]	Crohn's disease and impaired inflammation
Number of phenotypes per patient (e.g. infectious agents)	High	Low (even single)	Properdin and *MAC* mutations and Neisseria
Number of episodes per patient (e.g. infectious episodes)	High	Low (even single)	IL-12p40 and IL-12Rβ1 deficiency
Disease-causing cellular phenotypes	Haematopoietic	Nonhaematopoietic	*EVER1* and *EVER2* mutations and HPV
Genotype level			
Mode of Mendelian inheritance	Autosomal and X –recessive	Autosomal dominant	*STAT1* mutations and mycobacteria
Clinical penetrance	Complete	Incomplete	*IFNGR1* mutations and mycobacteria
Disease-causing genes per patient	One (monogenic, Mendelian)	Several (oligogenic, major genes)	*PARK* and *LTA* and Leprosy
Mutations	Inherited from the parental genome	Inherited from the parental germline *de novo*, or somatic	*Fas* mutations and auto-immunity

[a]Infections occurring in patients with overt immunological abnormalities.
[b]Autoimmunity, allergy, virus-induced cancer, angioedema, granulomas, haemophagocytosis, autoinflammation, thrombotic microangiopathy.

1.6 Common Variable Immunodeficiency (CVID)

1.6.1 Clinical manifestations

Although the first case of Common Variable Immunodeficiency (CVID, OMIM #240500) was reported in 1953 (Janeway et al. 1953), a consensual definition for this disease is not yet available and our understanding of it is far from complete. Its clinical manifestations generally present sinopulmonary and systemic bacterial infections, such as recurrent bronchitis, sinusitis, otitis media, pneumonia, as well as gastrointestinal complications (Figure 1.6.1.1), which are consequences of most CVID patients low serum IgG concentrations in spite of detectable levels of circulating B cells.

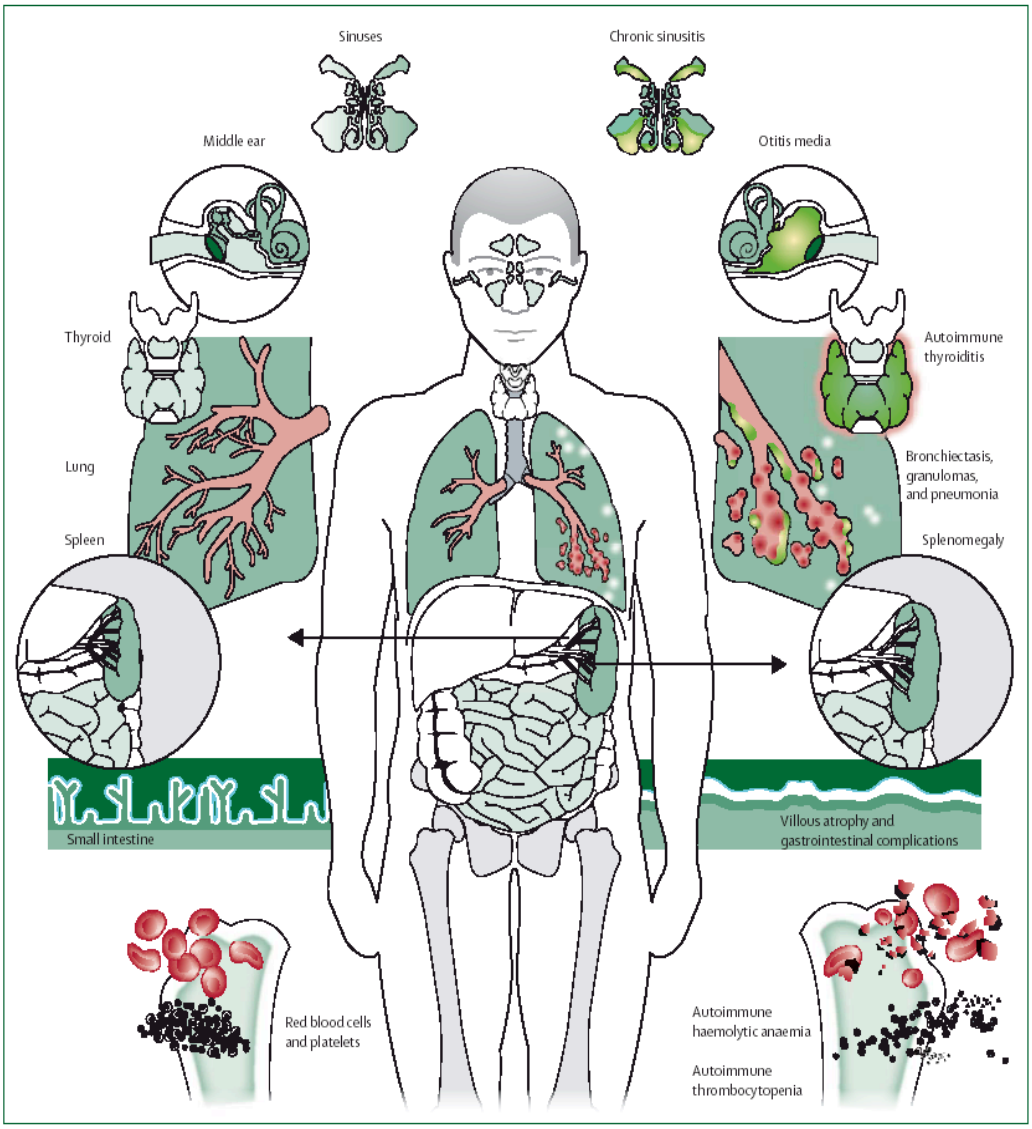

Figure 1.6.1.1 Organ systems involved in CVID pathogenesis.
Left = healthy organs; right = organ system involved in CVID.

About 25% of affected individuals have also autoimmune manifestations, such as Autoimmune Thrombocytopenic Purpura and Autoimmune Haemolytic Anaemia (5-8% of CVID patients) or Splenomegaly and Autoimmune Thyroiditis (Cunningham-Rundles and Bodian 1999). These manifestations are probably due to the fact that specific checkpoints for autoreactivity during B-cell development either fail or are circumvented (Tsuiji et al. 2006).
Finally, 10-22% of CVID patients are also affected by lungs, liver, skin, spleen and gastrointestinal tracts granulomas, together with some kind of neoplasias such as non-Hodgkin lymphoma and gastric cancers (Cunningham-Rundles and Bodian 1999).
That being so, individuals affected by CVID form a hardly definable group, characterized by extremely high phenotype heterogeneity.

1.6.2 Diagnosis

A well-accepted CVID definition, used for a reliable diagnosis of the disease, generally includes the following key features:

- ✓ the presence of Hypogammaglobulinaemia of two or more Ig isotypes (i.e. low IgG, IgA, or IgM);
- ✓ the presence of recurrent sinopulmonary infections and impaired functional antibody responses.

From a cellular point of view, immune system characteristics of individuals affected by CVID are very complex, with several numerical and functional defects involving B cells, T cells, natural killer cells, macrophages and monocytes (Bayry et al. 2005).
For example, B cells number in peripheral blood can be normal or reduced, whereas T cells abnormalities include reduction in number and function, defects in cytokine production, decreased T-helper cells function and T cells signaling, diminished expression of costimulatory molecule CD40 ligand and increased suppressor T cells function (North et al. 1998).
The number of class-switched memory B cells (CD27+, IgM–, IgD–) is low in 50-75% of CVID patients, even if it can also be low or zero in other humoral PIDs, such as Hyper-IgM syndrome.
Although origin and function of different memory B cells subsets in humoral responses has been ardently debated (Weller et al. 2004), there are enough data to prove a role for them in generating antibodies to both T-dependent and T-independent antigens, since changes in B cells memory

compartment, due to an underlying immunodeficiency, could have substantial effects on quality and quantity of humoral immune responses.

Unfortunately, delays in recognizing CVID are very common because of pervasive misconceptions that PIDs are extremely rare and typical pediatric disorders. Although most of them, like most Mendelian disorders, are observed principally in children, it may be due to their severe nature. As a matter of fact, as medical progress is making it increasingly possible for children with PIDs to survive to adulthood, adults with PIDs are increasingly being encountered (Cunningham-Rundles and Bodian 1999).

1.6.3 Epidemiology

Selective IgA Deficiency (IgAD) is the commonest PID, but most IgAD patients are asymptomatic (Salzer and Grimbacher 2006), so that CVID actually results the commonest clinically relevant primary immunodeficiency, representing about 30% of all PIDs affected individuals in Europe (Eades-Perner et al. 2007).

CVID equally affects both sexes, with prevalence from one per 50,000 to one per 200,000 and a reported incidence of one per 75,000 live births.

Most patients present a sporadic form of the disease, but 10-25% have familial inheritance, generally with an autosomal dominant pattern (Hammarstrom et al. 2000).

CVID age of onset has a bimodal distribution, with few patients presenting the disease in mid childhood, the great majority presenting it in early to mid adulthood and someone that presents it even later, with a reported mean age at the onset of symptoms of 23 years for males and 28 for females (Cunningham-Rundles and Bodian 1999).

1.6.4 Aetiology

Few CVID genetic aetiologies have been identified to date (Table 1.6.4.1), consistent with impairing B cells function autosomal recessive mutations of tumor necrosis factor receptor superfamily member 13B (TACI) (Salzer et al. 2005; Castigli et al. 2005a) and member 13C (BAFF-R) (Warnatz et al. 2005), of inducible T-cell costimulator ICOS (Grimbacher et al. 2003) and of CD19 (van Zelm et al. 2006).

Table 1.6.4.1 Gene defects in CVID and related inheritance patterns.

	Inheritance	B-cell-subset analysis by flow cytometry	Protein expression on cell surface
TNFRSF13C (<1% of CVID cases)[92,93]	Autosomal recessive	Reduced class-switched and non-switched memory B cells with increased transitional B cells	BAFF-R expression is absent on B-cell surface
TNFRSF13B (10–20% of CVID cases[88-91,95*]	Autosomal dominant	Low to absent IgA, autoimmune disease, lymphoproliferative disease, splenomegaly, reduced class-switched memory B cells	95% have normal TACI expression on B-cell surface, <5% have absent TACI expression
ICOS (~2%)[86,87,94,96]	Autosomal recessive	Reduced class-switched memory B cells, nodular lymphoid hyperplasia, autoimmunity, predisposition to neoplastic disease	ICOS expression on the surface of activated T cells is absent
CD19 (<1%)[93]	Autosomal recessive	Decrease in class-switched memory B cells, low CD21 expression on B cells, normal numbers of CD20+ mature B cells in peripheral blood	Low to absent expression of CD19 protein on the surface of CD20+ B cells

BAFF-R

BAFF-R is encoded by the *TNFRSF13C* gene and expressed on B cells surface, whereas its ligands, interactions of which provide crucial survival signals for differentiation of peripheral B cells, are expressed on macrophages, monocytes, and dendritic cells (Mackay and Ambrose 2003).
A homozygous 24 base-pair deletion in *TNFRSF13C* has been described in only one CVID subject with anomalies in peripheral B cells subsets, such as strong reduction of both class-switched (CD27+, M–, D–) and non-switched memory or marginal-zone (CD27+, M+, D+) B cells, increase in the transitional B cells compartment (CD38+++, M++) and decrease in plasmablasts (CD38+++, M–) (Warnatz et al. 2005).

TACI

Since defects on the *TNFRSF13B* gene, which encodes for B cells surface TACI receptor, are the most common nucleotide substitutions in CVID individuals, the great majority of genetic studies on this primary immunodeficiency are concerning such a genomic region.
For this reason, a full description of TACI structure, biological function and known nucleotide substitutions is postponed to a specific section (1.7).

ICOS

About 2% of CVID subjects presents defects in the *ICOS* gene. The protein product of this gene is constitutively expressed in germinal centers and T cells zones of spleen, lymph nodes, and Peyer's patches, whereas its ligand is expressed on lymphoid and non-lymphoid tissue.

ICOS function is that of enhancing natural killers activity and enabling T cells interactions with B cells, monocytes and dendritic cells (Greenwald et al. 2005).

It has been found that ICOS-deficient individuals generally have few peripheral B cells, few or no class-switched memory B cells, and hypogammaglobulinaemia (Salzer et al. 2004). Their T cells also produce very little interleukin-10 and this may be associated with a defective formation of germinal centers that leads to impaired memory B cells (Figure 1.6.4.1).

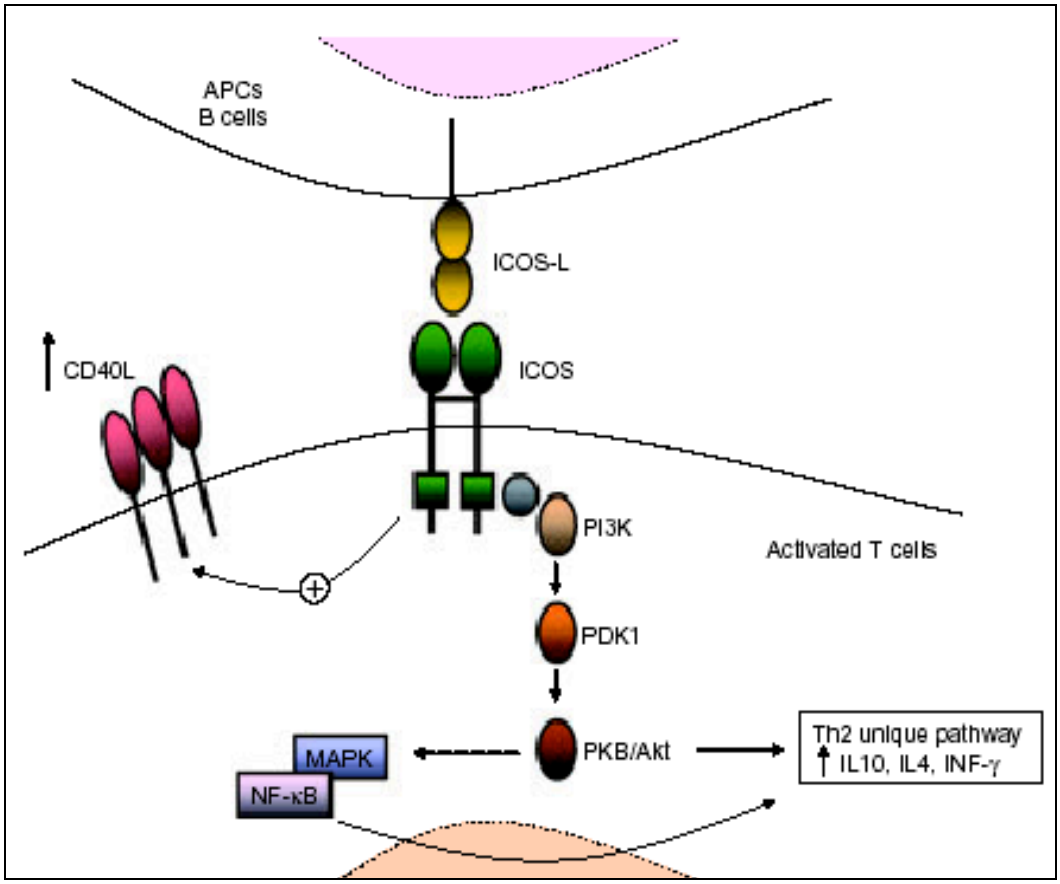

Figure 1.6.4.1 Signaling pathways of ICOS.

CD19

CD19 is expressed on B cells from an early stage of development, as a part of the B cell co-receptor along with CD21 and CD81. Interaction between B cell receptor and this co-receptor complex increases B cells signaling by several thousand times (Figure 1.6.4.2).

Individuals with homozygous mutations in the *CD19* gene show a normal total number of B cells (CD20+), but with low or undetectable surface expression of CD19 and decreased numbers of CD27+ memory B cells and CD5+ B cells (van Zelm et al. 2006).

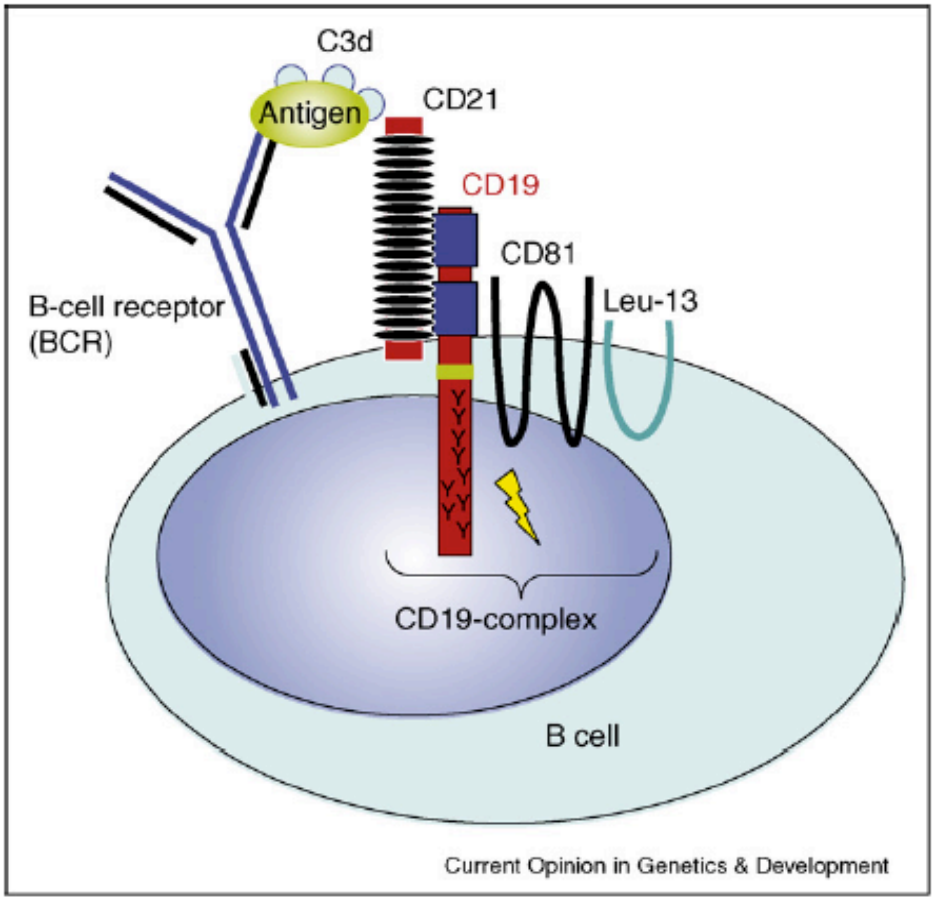

Figure 1.6.4.2 The B cell receptor and co-receptor signaling complex.

Unfortunately, since TACI defects, which are the most common nucleotide changes in CVID individuals, do not seem to be responsible of autosomal dominant forms of the disease (Pan-Hammarstrom et al. 2007), genetic basis of the typical, late-onset, CVID manifestation remains largely unknown.

That being so, the extremely high clinical heterogeneity of CVID subjects is probably a consequence of an equally high genetic heterogeneity of such disease.

For a summary of all receptors/ligands networks that are though to be involved in the CVID onset see Figure 1.6.4.3.

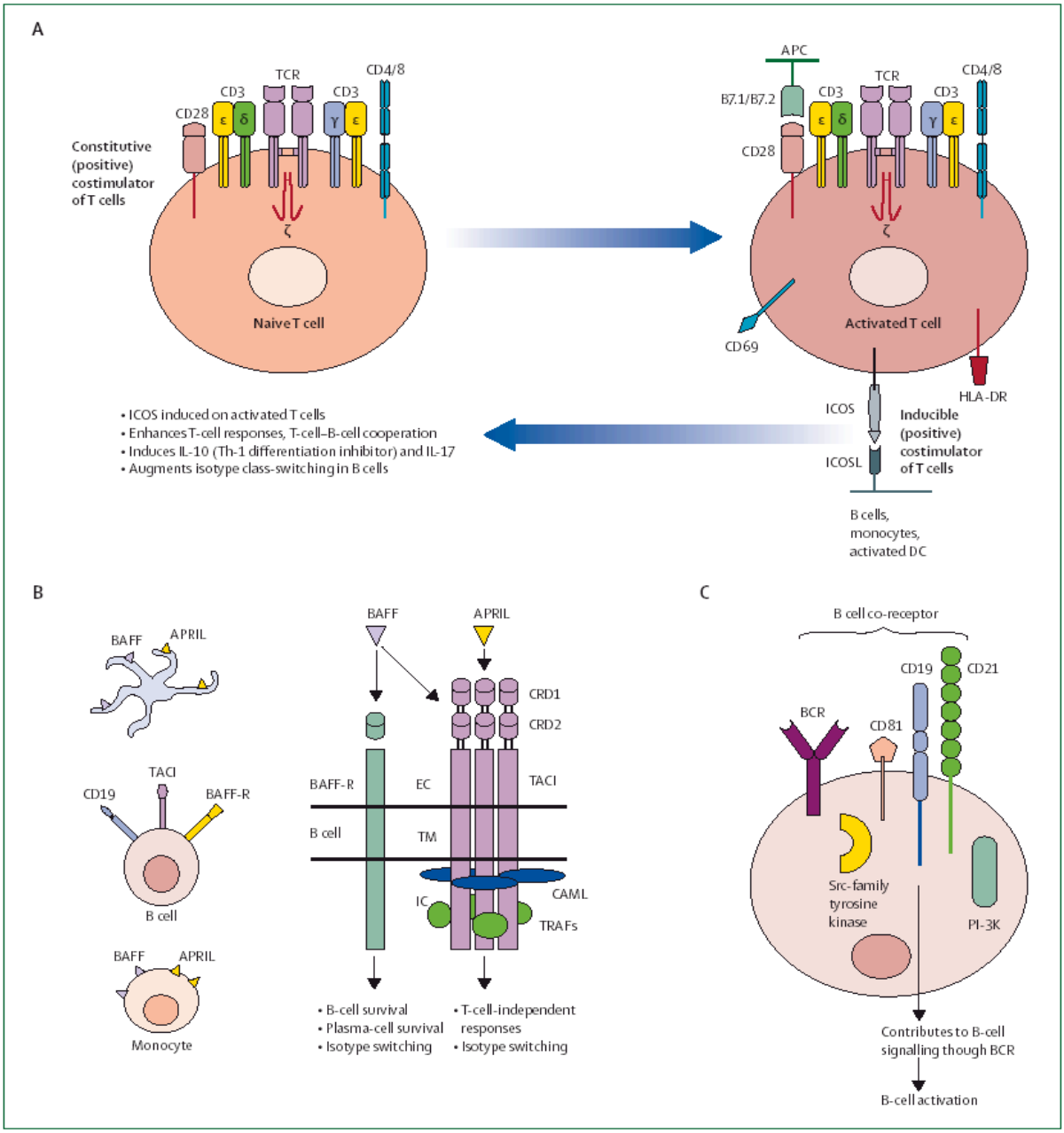

Figure 1.6.4.3 Summary of all molecules implicated in CVID.
A = ICOS
B = BAFF-R and TACI
C = CD19

1.7 Transmembrane Activator and CAML Interactor (TACI)

1.7.1 Structure and signaling

Transmembrane activator and CAML (calcium-modulator and cyclophilin ligand) interactor (TACI) belongs to the tumor necrosis factor receptor (TNF-R) superfamily and is primarily expressed as a 293 aminoacids type III transmembrane protein (Bossen and Schneider 2006) on late transitional B cells and marginal zone B cells surface, in circulation and lymphoid organs (Ng et al. 2004).

TACI extracellular region is mainly constituted by two cysteine-rich domains (CRDs) that are typical of all TNF-Rs. CRD-1 drives the receptor ligand-independent assembly into trimers or more complex multimers (Garibyan et al. 2007), whereas CRD-2 is responsible of ligands binding.

TACI intracellular region is instead structured to recruit signaling proteins such as TNF-R associated factors (TRAFs), which are also bound by BAFF-R and B cell maturation antigen (BCMA), and the calcium modulator and cyclophilin ligand (CAML) expressed on cytoplasmic vesicles surface (Xia et al. 2000) (Figure 1.7.1.1).

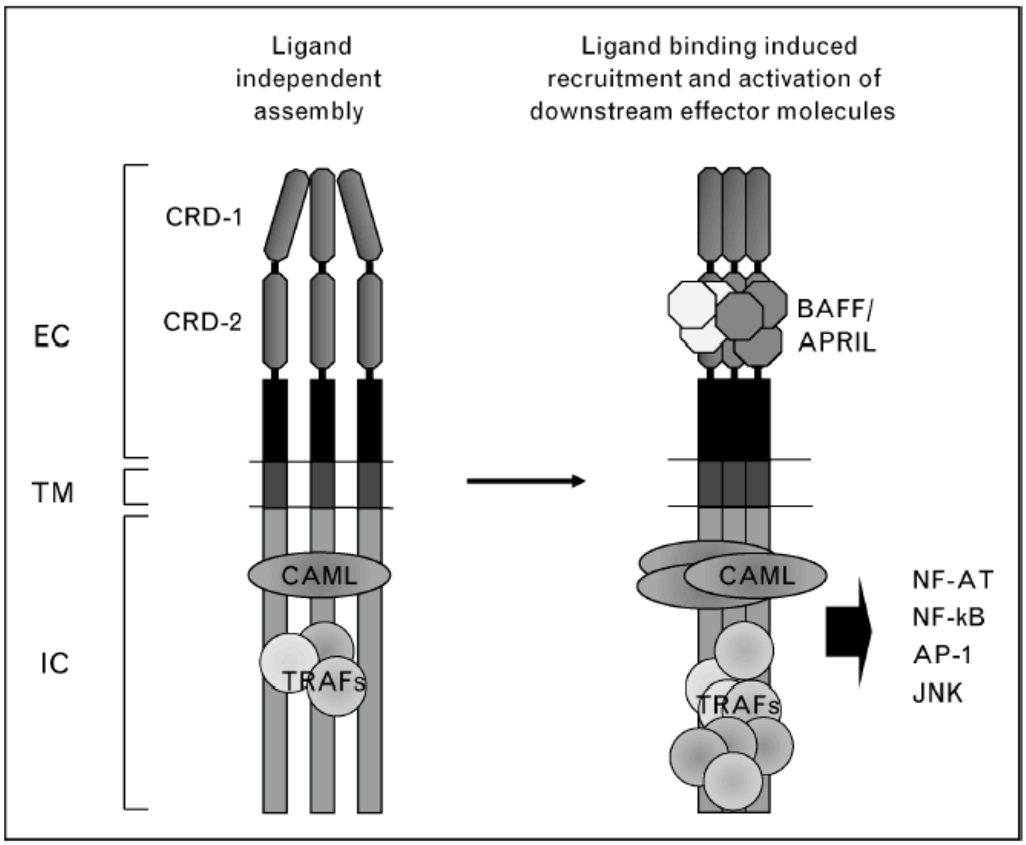

Figure 1.7.1.1 Schematic representation of TACI structure, assembly and signaling.
APRIL, a proliferating inducing ligand; BAFF, B cell activating factor; CAML, calcium modulator & cyclophilin ligand; CRD, cysteine-rich domain; EC, extracelluar domain; IC, intracellular domain; TACI, transmembrane activator and calcium modulator and cyclophilin ligand interactor; TM, transmembrane domain; TRAF, TNF-R-associated factor.

Ligands binding induces a conformational change that strengths the association with CAML and TRAFs and actives downstream effectors. In particular, TRAFs activation induces the nuclear factor κB (NF- κB), while via CAML the nuclear factor of activated T cells (NF-AT) is expressed.
The whole signaling complex also leads to activation of the c-Jun NH2-terminal kinase and of the transcription factor AP-1.
TACI requires a ligand induced oligomerization for such a optimal signaling, since TRAFs bind weakly to a single receptor (Castigli and Geha 2006).
Moreover, two TACI splice variants exist and differ from the already described structure. In particular, one lacks the CRD-1, while the other leads to a soluble form of the protein (Bossen and Schneider 2006).

1.7.2 Ligands/receptors network

TACI plays a role in a very complex network of ligands and receptors with overlapping binding specificities (Fig. 1.7.2.1).

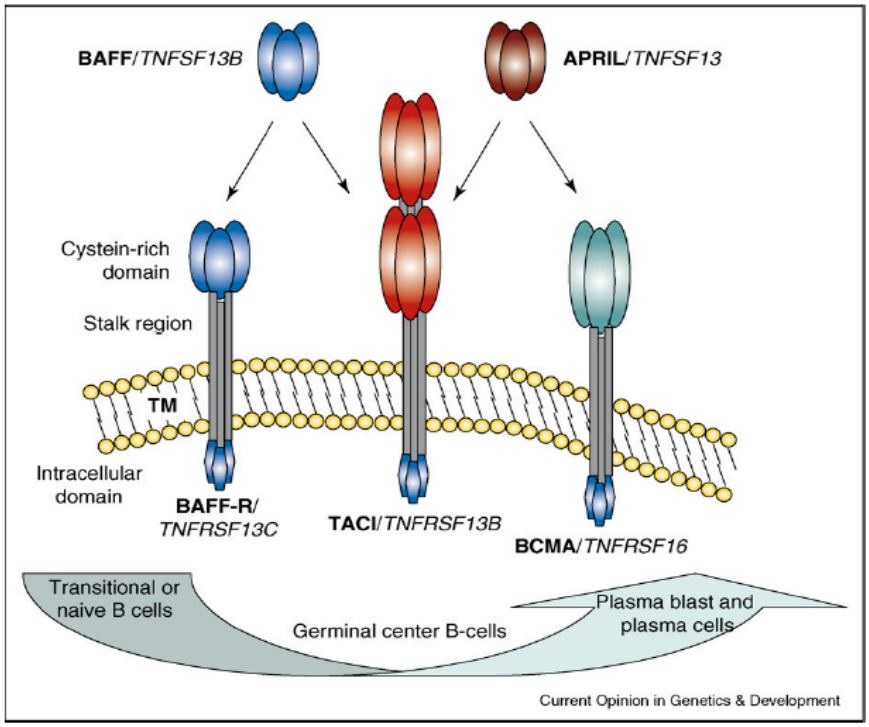

Figure 1.7.2.1 Representation of TACI, BAFF-R and BCMA network.
BAFF-R is universally expressed on transitional and naïve B cells; TACI expression depends on B cells activation both in germinal centers and lymphoid organs; BCMA is expressed in terminally differentiated plasmablasts and plasmacells.
TM, transmembrane domain.

It binds two TNF ligands, B cell activating factor (BAFF) and a proliferating inducing ligand (APRIL), which are expressed as type II transmembrane proteins, so that they can be proteolytically cleaved and secreted (Mackay et al. 2003).
In particular, BAFF is mainly expressed by neutrophils, monocytes, macrophages, and dendritic cells, whereas APRIL is expressed in monocytes, macrophages, dendritic cells and activated T cells. These ligands are also able to bind to other molecules than TACI. For example, APRIL binds to BCMA and to heparan-sulfated proteoglycans, while BAFF binds to BCMA as well as to its own unique receptor BAFF-R.

1.7.3 Biological function

TACI certainly plays a pivotal role in B cells activation and differentiation into plasma cells, since it has been proved that class-switch recombination and isotype switching can occur in a CD40 independent manner in naive B cells in response to APRIL and BAFF (Castigli et al. 2005b).
In vivo studies have also proved that TACI-deficient mice show a normal B cells development, but with contemporaneous IgA deficiency and impaired ability to perform antibody response to thymus-independent type II antigens, so that TACI seems to behave also as a negative regulator in B cells homeostasis.
As a matter of fact, TACI-deficient mice are characterized by splenomegaly and marked increase in circulating B cells (Yan et al. 2001).
An alternative explanation for this finding is that TACI may compete with BAFF-R for binding, limiting BAFF-R mediated B cells survival, so that in absence of TACI, an increase in BAFF-R signaling would result in B cells hyperproliferation.

1.7.4 The TNFRSF13B gene

The *TNFRSF13B* gene (OMIM #604907) is located on chromosome 17 (p11.2), spanning from position 16,783,124 to 16,816,127, for a total of 33 kb and being made up of five exons (Figure 1.7.4.1).
Three different *TNFRSF13B* transcripts exist, of which two are currently reported on databases and are described in the following part of this section. The third transcript is instead not reported on databases, but it has been proved that it is due to a splice variant that leads to a soluble form of the protein (Bossen et al. 2008).

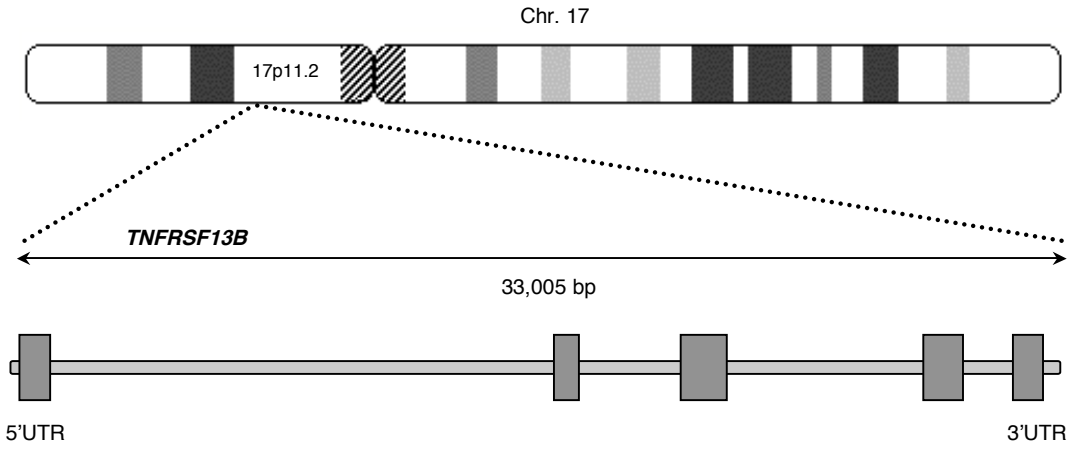

Figure 1.7.4.1 *TNFRSF13B* genomic structure.

The two already described transcripts (Figure 1.7.4.2) are reported on the Ensamble database (http://www.ensembl.org/Homo_sapiens/transcript):

- ✓ *TNFRSF13B*-001 (ENST00000261652).

  This transcript is constituted by 5 exons, for a total of 1,357 bp, and leads to a translated protein of 293 residues.

  Such a protein is that considered in clinical expression and sequencing studies on CVID individuals to date.
- ✓ *TNFRSF13B*-201 (ENST00000343345).

  This transcript is constituted by 4 exons, for a total of 1,219 bp, and leads to a translated protein of 247 residues.

  This form is due to a splice variant in which exon 2, encoding for CRD-1, is replaced by a codon for tryptophan (Hymowitz et al. 2005).

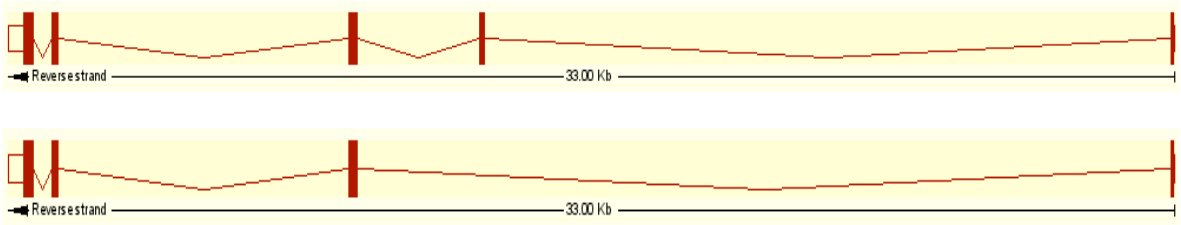

Figure 1.7.4.2 *TNFRSF13B* different transcripts.

1.7.5 TNFRSF13B defects

Several nucleotide substitutions in the *TNFRSF13B* coding region are found in about 10-20% of CVID individuals and are generally associated with a clinical phenotype of lymphoproliferation comprising splenomegaly, tonsillar hyperplasia, IgA Deficiency and autoimmune thyroiditis (Figure 1.7.5.1).

Table 1.7.5.1 The most common *TNFRSF13B* sequence variants.

Sequence variant	a.a. change
Significantly associated with CVID	
310T>C	C104R
542C>A	A181E
Uncertain association with CVID[b]	
277_231del	G76fsZx3
509G>A	C170Y
512T>G	L171R
204insA	L69fsX11
444C>A (−13aa)	S144X
594C>A, 595C>A	S194X
260T>A	I87N
563A>T	K188M
Not significantly associated with CVID	
215G>A	R72H
602G>A	R202H
659T>C	V220A
752C>T	P251L

[b] significance undetermined because of low variant frequency in CVID population and absence in healthy population.

The great majority of CVID individuals carries at least one mutated allele, whereas less than 2% carries biallelic mutations (Park et al. 2008; Salzer et al. 2008). Although many variants have been observed also in healthy individuals, homozygous or compound heterozygous *TNFRSF13B* mutations have been found exclusively in individuals with a full-blown antibody deficiency. However, it seems that presence of a heterozygous mutation alone cannot explain a sufficient percentage of multiplex CVID families with dominant inheritance and, in general, does not cause CVID.

C104R and A181E amino acid substitutions represent the most common *TNFRSF13B* mutations, being found in 4-5% of CVID subjects, in both sporadic and familial cases.

However, a critical factor in determining the pathogenicity of such changes is their frequency in healthy population. As a matter of fact, finding of heterozygous C104R and A181E in healthy individuals questions whether they on their own are truly pathogenic or if, indeed, they act as susceptibility or disease-modifying mutations, acting in co-operation with other gene defects.

Despite conflicting data, especially for A181E, the great majority of studies reports a highly significant association of these two mutations with CVID, but not with IgAD (Castigli et al 2007; Pan-Hammarstrom et al. 2007; Zhang et al. 2007; Park et al. 2008). Moreover, in families with dominant inheritance of A181E, some individuals with such a substitution resulted completely asymptomatic, suggesting that the mutation has incomplete penetrance (Salzer et al. 2005; Pan-Hammarstrom et al. 2007).

Many other variants, such as G76fsX3, C170Y and L171R, are instead found to be exclusive of CVID individuals, but they resulted too rare to reach statistical significance.

Finally, amino acid substitutions R72H, R202H, V220A and P251L have been always observed at similar frequency in cases and controls (Castigli et al. 2005a).

1.7.6 TNFRSF13B defects and B cells functionality

A reliable evaluation of *TNFRSF13B* variants involvement in CVID necessarily requires a clear understanding of how they may affect B cells functionality.

In particular, the impact of a mutant allele on wild-type TACI function should be assessed, since the great majority of CVID subjects carrying *TNFRSF13B* defects are heterozygotes.

There are three main hypothesis concerning this issue:

- ✓ dominant negative effect of the mutant allele;
- ✓ haploinsufficiency;
- ✓ CVID individuals heterozygous for *TNFRSF13B* variants preferentially express the dysfunctional allele.

A dominant negative effect was initially though for C104R substitution. As a matter of fact, even if C104R affects the receptor extracellular domain, disrupting a disulphide bond required for ligand biding CRD-2, it has been proved that only homozygous C104R individuals are unable to bind APRIL (Salzer et al. 2005). In heterozygotes, association of the mutant allele with wild-type TACI indeed does not interfere with ligand binding, but inhibits ligand induced NF-kB activation (Garibyan et al. 2007).

Moreover, it has been shown that TACI respond to oligomeric BAFF or APRIL only, implying that at least six wild type receptors are likely required in an active signaling complex, rendering it particularly prone to dominant negative effects of mutated allele (Bossen et al. 2008). However, further in-vivo analyses of transgenic mice co-expressing both wild-type TACI and the mutated allele suggest that C104R exerts its effect via haploinsufficiency (Lee et al. 2008).

As regards A181E substitution, a negatively charged glutamine replaces a neutral alanine in the transmembrane domain. It may prevent downstream signaling by interfering with ligand induced conformational changes of the whole receptor. The mutant-wild-type complex can indeed bind ligands, but is unable to activate NF-kB or NF-AT.

Despite such in-vitro observations, effects of heterozygous C104R and A181E substitutions on B cells function in individuals with CVID are still under debate. Their B cells responses to TACI ligation are actually impaired, but it needs to be further investigated whether these defects are specific to these subjects or are part of a generalized impairment of B cells activation that has been observed in the CVID population in response to TACI, CD40 and TLR9 ligands (Zangh et al. 2007). As already discussed for *TNFRSF13B* variants in general, it seems that presence of a heterozygous C104R and A181E alone does not cause CVID. Whether it is due to incomplete penetrance or delayed onset, or whether additional genetic-environmental factors are required for manifestation of the disease, is not entirely clear. However, C104R and A181E heterozygous mutations are present in up to 1% of healthy population and it is quite unlikely that even a fraction of these individuals will develop CVID later in life, otherwise, its frequency would be much higher. Moreover, incomplete penetrance and familial segregation have been demonstrated for other heterozygous *TNFRSF13B* variants, implying that they increase the risk, but are neither necessary nor sufficient to cause CVID (Salzer et al. 2008).

That being so, other genes and/or environmental factors are probably needed for TACI mutations to result in CVID. In particular, genes along CD40-CD40L or TLR pathways may be potential candidates, as it is known that defects in CD40L expression, CD40 signaling and TLR9 can occur in CVID and that TACI participates in cross-talk with CD40 and TLRs (Lee et al. 2008).

1.8 Selective IgA deficiency (IgAD)

Selective IgA deficiency (IgAD) is the most common primary immunodeficiency disorder (Salzer and Grimbacher 2006) and is characterized by decreased serum IgA concentration (<0.07 g/l) and normal serum IgM and IgG levels.

Many of these individuals have no apparent diseases, whereas others suffer from recurrent mucosal infections, allergies and autoimmune diseases. IgA deficiency is also associated with some autoimmune manifestations such as systemic lupus erythematosus, juvenile onset diabetes mellitus and rheumatoid arthritis (Hammarstrom et al. 2000). IgA deficit is presumed to result from impaired switching to IgA or maturational failure of IgA-producing lymphocytes.

Progression from IgAD to CVID has been also reported in several cases (Carvalho Neves Forte et al. 2000). In addition, fixed haplotypes of MHC genes are frequently associated with both IgAD and CVID. At least two distinct loci, one in the class II region and one in the class III region, confer susceptibility to IgAD and CVID development (Schroeder et al. 2004).

In conclusion, co-occurrence of some autoimmune disorders, IgG subclass deficiency and association of HLA A1, 8, DR3, DQ2 or part of these haplotypes in IgAD individuals, in particular those with affected family members, could be risk factors for CVID induction (Aghamohammadi et al. 2008).

2. Aim of the Study

This study has investigated worldwide genetic variability of the *TNFRSF13B* gene, with the aim of evaluating its variants potential contribution to the development of Common Variable Immunodeficiency (CVID), the most prevalent primary immunodeficiency in individuals of European ancestry.

As a matter of fact, in the recent years this gene was proved to be able to regulate isotype switching, survival and differentiation of B lymphocytes, playing a role in a very complex functional network that is crucial for humoral responses (Mackay and Schneider 2008).

Moreover, some *TNFRSF13B* coding variants have been also implicated in CVID and Selective IgA Deficiency (IgAD) by clinical genetics studies (Castigli et al. 2005a; Salzer et al. 2005), even if they were exclusively based on samples of European ancestry and functional effects of observed mutations in relation to diseases development have not been entirely established. After the initial claim of a strict association of *TNFRSF13B* changes with both diseases, further analyses have indeed shown the existence of some of these variants also in healthy individuals (Castigli et al. 2007; Pan-Hammarstrom et al. 2007; Lee et al. 2008; Salzer et al. 2008).

Given such a complex and still ambiguous scenario, the more comprehensive perspective of an evolutionary approach was applied in this study, as already carried out for other genes with medical implications (Aldea et al. 2004; Sabater-Lleal et al. 2006; Soldevila et al. 2006), offering a broader context in which to conduct research and underling the belief that investigating genetic variation and evolution patterns represents a powerful tool also in human health research. This study would indeed show how evolutionary genetics methods could play a role in dissecting the origin, causes and diffusion of human diseases.

The rationale behind this approach is that if natural selection has acted on the *TNFRSF13B* locus, a specific imprint in its genetic diversity could be observed and this may turn out to be very useful for an exhaustive understanding of its function and potential role in different human populations CVID susceptibilities. In particular, genes involved in adaptive immunity, such as *TNFRSF13B*, are likely to be subjected to geographically localized selective pressures, as they may interact with local pathogen landscapes, resulting in increased inter-population differentiation (Bamshad and Wooding 2003).

In order to verify such a hypothesis, the *TNFRSF13B* coding region was sequenced in 451 healthy individuals belonging to 26 worldwide populations from Sub-Saharan Africa, North Africa, Middle East, Central Asia and South America, in addition to control individuals, CVID and IgAD subjects from Italy. In this way, a global picture of *TNFRSF13B* nucleotide diversity and haplotype structure

was for the first time obtained, making investigation on potential departures from the neutral model of evolution possible.

For this purpose, six different neutrality tests have been applied to infer genetic variability deviations from what is expected under neutrality.

Moreover, pair-wise genetic distances among studied groups were computed and the apportionment of genetic variance among and within large geographically-based groups of populations, among individual populations, and between cases and controls, was investigated by means of the Analysis of the Molecular Variance (AMOVA).

The average mutation rate for such a genomic region was also estimated by computing nucleotide divergence between healthy human samples and the chimpanzee. Subsequently, a phylogenetic analysis was performed to examine evolutionary relationships among inferred haplotypes and to achieve broad age estimates for supposed disease-causing *TNFRSF13B* variants.

Combined results of all these analyses have finally led to the reconstruction of a plausible evolutionary history for the *TNFRSF13B* gene that was achieved in the attempt to elucidate which diversity falls into the standard degree of intra-specific *TNFRSF13B* variation and which instead may be related to CVID and IgAD phenotypes.

3. Materials and Methods

3.1 Population Samples

Two groups of unrelated Italian CVID and IgAD subjects, as well as a panel of 451 unrelated healthy individuals belonging to 26 worldwide populations, were analyzed for a total of 1,132 sequenced chromosomes. A written informed consent was collected from each subject.

3.1.1 CVID and IgAD samples

77 CVID samples were kindly provided by Dr. Isabella Quinti of the "La Sapienza" University Department of Clinical Immunology of Rome. All individuals belong to Center Italy and have been diagnosed for CVID by standard criteria of low levels of serum IgG, IgA, and/or IgM, antibody deficiency with impaired response to tetanus and pneumococcal antigen immunization, more than 2% of peripheral B cells and exclusion of hypogammaglobulinemia due to other primary or secondary immunodeficiencies.
38 IgAD samples were instead provided by Dr. Giampaolo Ricci of University of Bologna Pediatric Department and belong mainly to Northern Italy (Figure 3.1.1.1).
CVID individuals were enrolled from 26 Italian Centers belonging to the Italian Primary Immunodeficiency Network and, together with many other affected individuals, have already been used in a deep analysis of the spectrum of illnesses occurred at the time of the disease onset and over a mean of 11.5 years of follow-up with long-term immunoglobulin replacement therapy.
Such a study, as well as its samples collection, was designed according to the ethical principles for medical research involving human subjects of the World Medical Association Declaration of Helsinki.

3.1.2 Italian samples

96 unrelated Italian blood donors belonging to many different regions of Italy, without any evident immunological manifestations and with an age of over than 55 years, were collected at the Transfusion Center of the Maggiore General Hospital of Bologna, to be used as a control group representative of Italian healthy population.

In addition, 96 unrelated individuals born in the Pre alpine Val di Scalve (BG) were collected thanks to collaboration of the Val di Scalve AVIS group (Volunteer Italian Blood donors Association).

Val di Scalve is located in a mountainous area of Lombardy surrounded by peaks belonging to the Orobie Pre Alps and rising up to 2,700 m a.s.l. Such features make communication with neighboring valleys very difficult and contribute to the high degree of isolation of Val di Scalve villages (Figure 3.1.1.1).

Bio-demographic data about the last three generations were also collected for each subject to check its actual origin.

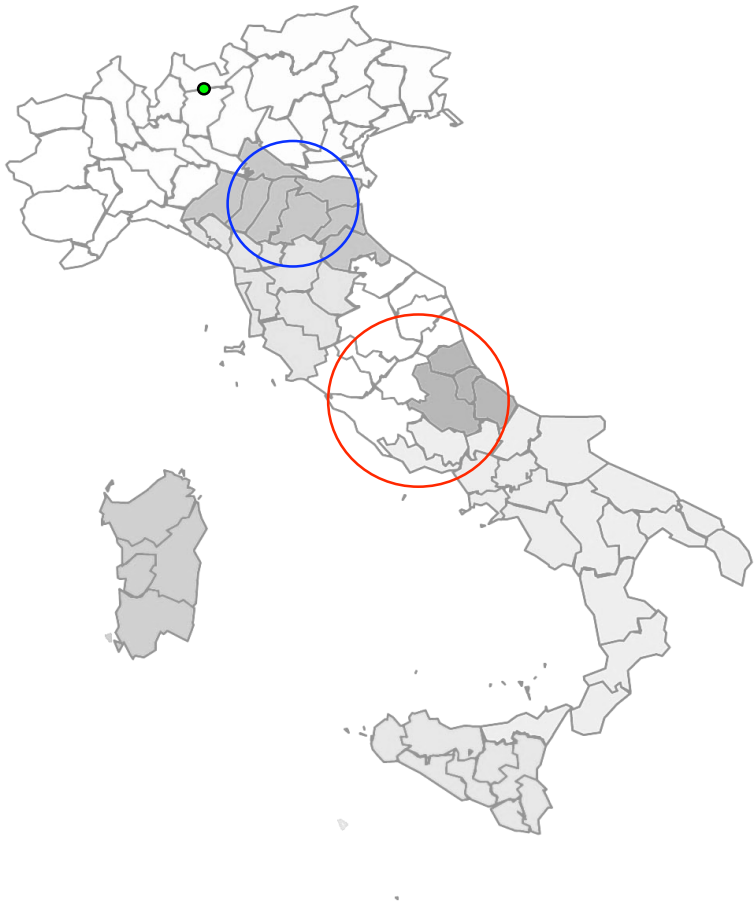

Figure 3.1.1.1 Geographical location of Italian samples.
○ = CVID samples
○ = IgAD samples
● = Val di Scalve samples
Blood donors were representative of Italian healthy population.

3.1.3 Central Asian samples

28 individuals from Central Asia were selected among samples collected in 1993 and 1994 for the CAHAP Project (Central Asia High Altitude People Project).

Individuals from Kazakhstan were collected in the high plain of the Kegen valley (8 Kazakhs) and in the medium-altitude village of Penjim in the East part of the country (7 Uyghurs).

Individuals from Kirghizstan belong to two different groups located in the medium-altitude Talas valley, in the Northern part of Kirghizstan, close to Kazakhstan and Uzbekistan (7), and from the isolated high-altitude village in Pamir mountains in the South of the country (6 Kirgiz).

Figure 3.1.3.1 Geographical location of Central Asian samples.
- = Kazakhs
- = Uyghurs
- = Kirgiz from Talas
- = Kirgiz from Sary-Tash

3.1.4 Middle Eastern samples

Middle Eastern samples were provided by Dr. Shirin Farjadian of the Department of Immunology of the Allergy Reseach Center of Shiraz University of Medical Sciences.

All 96 individuals were collected in Iran, but belong to 6 different ethnic and religious groups: 16 Iranian Arabs from Ahvaz (Western Iran), 16 Iranian Jews from Shiraz (South Western Iran), 16 Balochs from Iranshahr (South Eastern Iran, close to Pakistan), 16 Parsees from Shiraz (South Western Iran), 16 Turkmens from Gonbad (North Eastern Iran) and 16 Zoroastrians from Yazd (Center Iran).

Bio-demographic data were also collected for each individual to check that he was third generation native from the selected area.

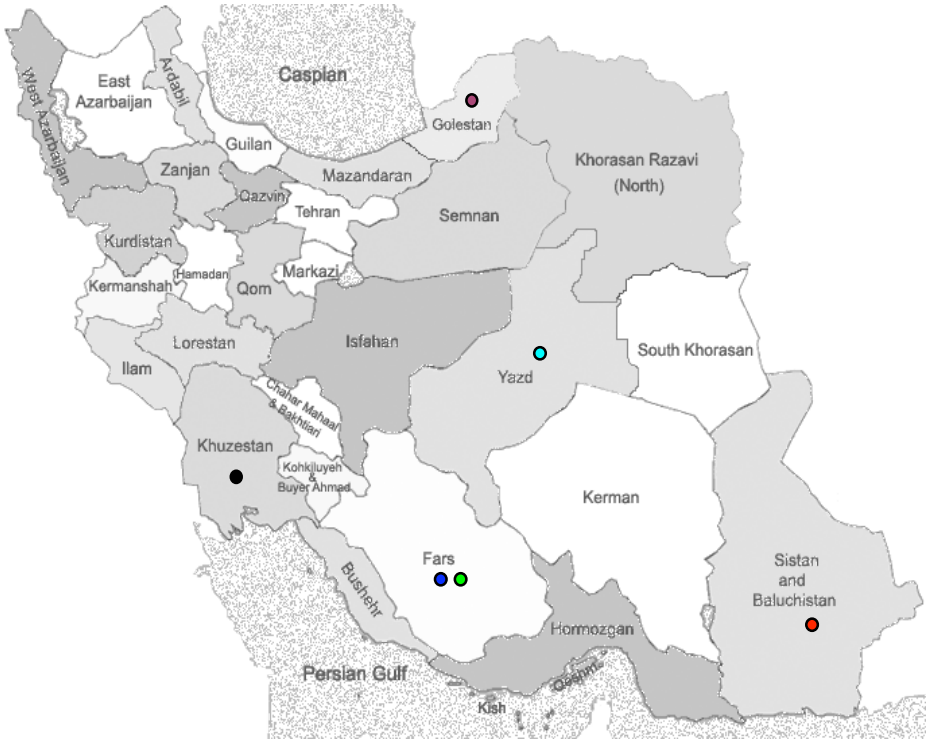

Figure 3.1.4.1 Geographical location of Iranian samples.

- = Balochs
- = Iranian Jews
- = Iranian Arabs
- = Parsees
- = Turkmens
- = Zoroastrians

3.1.5 African samples

African samples were obtained from individuals which currently live in Italy and were maintained subdivided into two different groups: Sub-Saharan Africans were represented by Ethiopians (11), Cameroonians (7), Senegalese (7), Maasai from Kenya (10), Eritreans (7) and Nigerians (15), while North Africans come from Morocco (25), Tunisia (8) and Egypt (4).

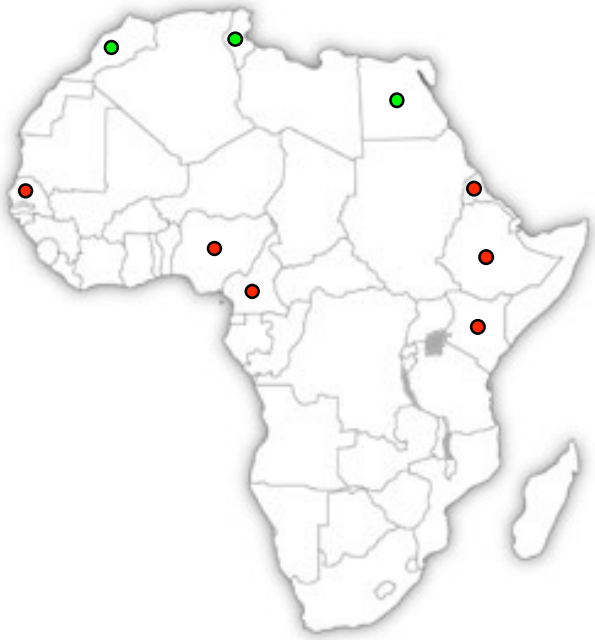

Figure 3.1.5.1 Geographical location of African samples
● = Sub-Saharan African samples
● = North African samples

3.1.6 South American samples

41 individuals from South America were selected among samples collected in 2007 during the Darwin Project expedition in Peru.
Subjects of Yanesha ethnicity (8) come from the medium-altitude Peruvian Central Jungle of the Cerro de Pasco region in the province of Oxapampa.
Individuals speaking Quechua language come from two different areas: the province of Tayacaja (9) and the high-altitude Pucachupa village in the province of Puno, close to Lake Titicaca (10).

Samples of Aymara ethnicity (10) also come from the same region, notably form the village of Chimu.

Finally, Arequipa individuals (4) were selected among samples collected by Dr. Eduardo Tarazona-Santos starting form 1998 for mtDNA and Y-chromosome genetic variability studies on South Amerindian populations.

Figure 3.1.6.1 Geographical location of Peruvian samples.
- = Quechuas from Pucachupa
- = Yanesha
- = Aymara from Chimu
- = Arequipa
- = Quechuas from Tayacaja

3.2 Laboratory Methods

3.2.1 DNA extraction

DNA was extracted from peripheral blood samples using two different methodologies. A modified Salting-Out protocol (*a*), based on the original protocol of Miller et al. (1988), was used for Middle Eastern, Central Asian, South American and Val di Scalve samples, while a QIAamp Blood Midi Kit (*b*) (QIAGEN, Hilden, Germany), which exploits ability of QIAamp Spin Columns silica-gel membrane to absorb DNA, was employed for extracting DNA from Sub-Saharan African, North African, Italian healthy, CVID and IgAD samples.

a) *Modified Salting-Out protocol:*

1. Thaw out blood samples by incubating at 37°C and then equilibrating to room temperature (15-25°C).
2. Transfer blood samples to sterile 15 ml tubes and add 12 ml *Red Cell Lising Buffer* (*RCLB*).
3. Vortex, centrifuge at 3,000 rpm for 10 min. and discard supernatant being careful to conserve the pellet. Repeat this step 3-4 times as far as the pellet loses its colour.
4. Add 3 ml *White Cell Lising Buffer* (*WCLB*), vortex and add 25 µl *Proteinase K* (20 mg/ml) and 25 µl *SDS* 20%, vortex and incubate at 55°C for 1 hour.
5. Add 1.7 ml *Sodium Acetate* (3M, pH 5.2), manually agitate and centrifuge at 3,000 rpm for 10 min.
6. Transfer the supernatant to new 15 ml tubes.
7. Add an equivalent volume of *Isopropyl alcohol* and softly agitate. At this step it should be visible the DNA "jelly-fish" (on the dimension of which depends the final amount of sterile water to add). Centrifuge at 3,000 rpm for 10 min.
8. Discard the supernatant and add 3 ml Ethanol 80%, vortex and centrifuge at 3,000 rpm for 10 min.
9. Discard the supernatant, turn upside-down the tubes and leave them drying on absorbent paper for at least 2 hours.
10. Put dry DNA in solution by using sterile water.

b) QIAamp Blood Midi Kit protocol:

1. Thaw out blood samples by incubating at 37°C and then equilibrating to room temperature (15-25°C).
2. Add 100 µl *QIAGEN Protease* in 15 ml centrifuge tubes, add 1 ml blood and mix briefly.
3. Add 1.2 ml *Buffer AL* and mix thoroughly by inverting the tubes 15 times, followed by additional vigorous shaking for at least 1 min.
4. Incubate at 70°C for 10 min.
5. Add 1 ml *Ethanol* 96-100% and mix by inverting the tubes 10 times, followed by additional vigorous shaking.
6. Transfer the solution onto *QIAamp Midi columns* placed in 15 ml centrifuge tubes. Close the cap and centrifuge at 3,000 rpm for 3 min.
7. Remove *QIAamp Midi columns*, discard the filtrate from the 15 ml centrifuge tubes and place the columns back into them.
8. Add 2 ml *Buffer AW1* to the columns. Close the cap and centrifuge at 5,000 rpm for 1 min.
9. Add 2 ml *Buffer AW2* to the columns. Close the cap and centrifuge at 5,000 rpm for 15 min.
10. Place the columns in clean 15 ml centrifuge tubes and discard the collection tubes containing the filtrate.
11. Add 200 µl *Buffer AE*, equilibrated at room temperature (15-25°C), directly onto the membrane of the columns. Close the cap and incubate at room temperature for 5 min. then centrifuge at 5,000 rpm for 2 min.
12. Add further 200 µl *Buffer AE*, equilibrated at room temperature (15-25°C), directly onto the membrane of the columns. Close the cap and incubate at room temperature for 5 min. then centrifuge at 5,000 rpm for 2 min.

Purified total extracted DNA (e.g. genomic and mitochondrial DNA) was visualized by electrophoresis on 1% agarose gel, at a voltage of 120 V for 15 min. Agarose gel was prepared with SeaKem® LE Agarose gel (CAMBREX Bio Science Rockland, Rockland, ME, USA), TAE 1X Buffer composed of 40 mM Tris, pH 8.0, 20 mM Acetic Acid, 1 mM EDTA (BIO-RAD Laboratories, Munich, Germany) and 0.1 µg/ml Ethidium Bromide (Sigma, St. Louis, MO, USA). The same visualization provided a rough DNA quantification by means of visual comparison of extracted DNA with cl857 Sam 7 λ-DNA samples at known variable concentration (Roche Diagnostic, Indianapolis, IN, USA).

3.2.2 TNFRSF13B exons amplification

Five segments, corresponding to the five *TNFRSF13B* exons and their immediately intronic flanking regions, were amplified by using Polymerase Chain Reaction (PCR) (Mullis et al. 1986), encompassing a total of 2,254 base pair (bp) for each individual. Amplification reactions were performed in a GeneAmp® PCR System 9700 thermal cycler (Applied Biosystems, USA) as previously described (Castigli et al. 2005a), except for exon 5 that required a new couple of primers and a different annealing temperature. A high efficiency FastStart Taq DNA Polymerase (Roche Diagnostic, Indianapolis, IN, USA) was used to ensure amplification even for less concentrated DNA samples. Reagents used for PCR reaction mix are reported in Table 3.2.2.1, while exon-specific couples of primers and thermal cycler setting parameters are listed in Table 3.2.2.2 and 3.2.2.3 respectively.

Table 3.2.2.1 PCR mix reagents for *TNFRSF13B* exons amplification.

Reagents	C_f	*V (µl)*
H_2O		
Buffer + $MgCl_2$	1 X	2.5
dNTPs	0.2 mM	2
Primer F	0.5 µM	1.25
Primer R	0.5 µM	1.25
Taq Polymerase	1 U	0.2
Genomic DNA	≥ 5 ng	
mix		25

C_f = final concentration; *V* = volume; *U* = units.
H_2O to reach the final volume of 25 µl.

Table 3.2.2.2 Couple of primers used for *TNFRSF13B* exons amplification.

Exon	*Amplicon (bp)*	*Primer forward (5'>3')*	T_m (°C)	*Primer reverse (5'>3')*	T_m (°C)
I	465	GCCCGGCAGGCCTTCCACT	66	GCAAGCCCCACATCCCAGAGG	70
II	336	GGCAGGAGAGGCCGTCTTGG	68	TCCTCCTGCCACCCTTTCCTCA	70
III	503	GGCTTACTCTGGAATTGCCTTCTG	72	CTTCTGGCCATTTGCTTGGACT	66
IV	497	CCAGCCTCTCCAGGAGCCAGAC	74	CCGGGTGCCACTCTCCCAGTTA	72
V	452	CCCCGGCACAGGTTCTGGTC	62	TCCTCCTTTCCCTCCCTGAC	72

bp = base pairs; T_m = melting temperature.

Table 3.2.2.3 Thermal cycler setting parameters for *TNFRSF13B* exons amplification.

Steps	*Time*	*Temp. (°C)*	*Cycles*
Initial DNA denaturation	5 min	95	
DNA denaturation	30 sec	95	
Primers annealing	30 sec	66[a]/64[b]	35
Extension	30 sec	72	
Final extension	7 min	72	

[a] for exon 1, 2, 3, 4; [b] for exon 5.

Amplicons were visualized by electrophoresis on 2% agarose gel, at a voltage of 100 V for 15 min. In order to verify if amplified DNA segments actually have the expected length, to be sure of the absence of unspecific amplifications, a GeneRuler^{TM} 100 bp DNA Ladder (Fermentas, Burlington, Ontario, Canada) was used as a reference (Figure 3.2.2.1).

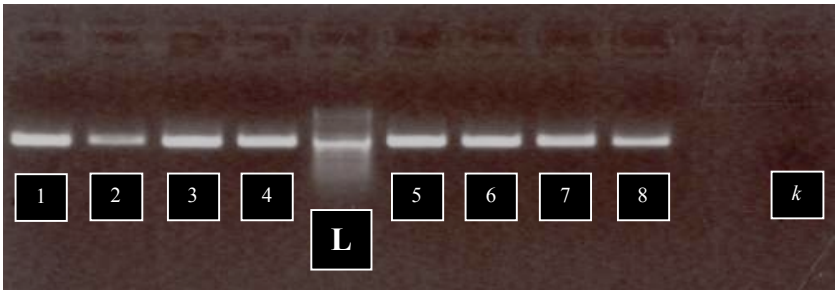

Figure 3.2.2.1 Amplification pattern for *TNFRSF13B* exon 3.
L = GeneRuler^{TM} 100 bp DNA Ladder, with the brightest band of 500 bp;
k = white control to test for possible DNA contamination of PCR reagents.

Subsequently, PCR products were added of 70 µl sterile water and transferred on a Montage PCR kit (Millipore, Bedford, MA, USA) to be purified with a vacuum pump system, at a pressure of 15 mmHg for 5 min. This typology of purification exploits the ability of a filtration membrane to eliminate residual salts, primers and not incorporated dNTPs from PCR products solution. To check if purification process has been achieved without DNA loss, purified products were visualized by electrophoresis on 1% agarose gel, at a voltage of 100 V for 15 min.

3.2.3 TNFRSF13B exons sequencing

Sequencing of *TNFRSF13B* coding region was performed by means of Chain Termination Method (Sanger et al. 1977) based on the employment of dideoxynucleotides (ddNTPs) that are devoid of the 3' OH group and labeled with four different fluorescent molecules.

In this way, ddNTPs incorporation in new DNA strands precludes incorporation of further nucleotides, producing DNA strands that differ for one base and are identifiable by a CCD (Charge-Coupled Device).

Purified PCR products were used for sequencing reaction with ABI Prism BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, USA), the protocol of which and the relative thermal cycler setting parameters are described in Table 3.2.3.1 and Table 3.2.3.2 respectively.

Table 3.2.3.1 Sequence reaction mix reagents.

Reagents	C_f	*V (µl)*
H_2O		
*Buffer**	0.5 X	2
BigDye 1.1 Kit	2.5 X	1
Primer (F o R)	0.32 µM	1
Purified PCR product		2
mix		10

C_f = final concentration; V = volume.
*BigDye® Terminator v1.1/3.1 Sequencing Buffer 5X.
H_2O to reach the final volume of 10 µl.

Table 3.2.3.2 Thermal cycler setting parameters for *TNFRSF13B* exons sequencing.

Steps	*Time*	*Temp. (°C)*	*Cycles*
Initial DNA denaturation	30 sec	96	
DNA denaturation	10 sec	96	25
Primer annealing	3 min	60	

Products of sequencing reaction were added of 10 µl sterile water and transferred on a Montage SEQ96 Sequencing Reaction Cleanup Kit (Millipore, Bedford, MA, USA) to be purified with a vacuum pump system, at a pressure of 15 mmHg for 3 min and, after re-moisturizing DNA with Injection Solution (Millipore, Bedford, MA, USA), for further 4 min.

Finally, separation of sequencing reaction purified products, on the basis of their length and fluorescent label, was performed by capillary electrophoresis on an automatic sequencer ABI 3730 DNA Analyzer (Applied Biosystems, USA), the Data Collection software of which manage fluorescence data directly producing electropherograms (Figure 3.2.3.1).

Subsequently, the Sequencher 4.6 software (http://www.genecodes.com) was used to read electropherograms and to align obtained sequences to the *TNFRSF13B* reference sequence (NG_007281.1 GenBank) for detecting polymorphic sites.

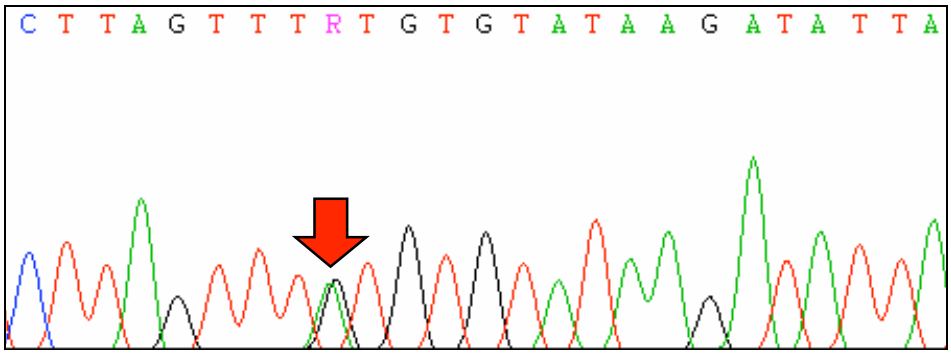

Figure 3.2.3.1 Example of ABI 3730 electropherogram.
⬇ indicates a polymorphism in heterozygous state.

3.3 Statistical Analyses

Statistical analyses were performed clustering population samples into five geographically-based groups (Sub-Saharan Africa, North Africa, Middle East, Central Asia and South America), whereas Italian healthy, CVID and IgAD individuals, as well as Italians from the isolate of Val di Scalve, were considered as four independent samples.

3.3.1 Haplotypes inference

Haplotypes were statistically inferred from unphased genotype data by means of the Bayesian algorithm implemented in the PHASE 2.1 software (Stephens et al. 2001; Stephens and Donnelly 2003). This Markov Chain-Monte Carlo (MCMC) algorithm aims to evaluate conditional distribution of unresolved haplotypes by exploiting the fact they tend to be similar to resolved haplotypes, which are certainly known since some individuals are homozygous. According to this rationale, the most likely pair of haplotypes for an ambiguous individual should be represented by two haplotypes that are similar, but not identical, to two high population frequency haplotypes.
The most plausible phase reconstructions and their confidence probabilities were estimated on the basis of such a comparison between unresolved and similar resolved haplotypes, by using default settings of the PHASE 2.1 software, with the exception of the number of iterations (1,000 rather than 100), as recommended in Graffelman et al. (2007).
Relatively low confidence probabilities are obtained if analyzed samples are made up of very few individuals, since too little information is available to allow a reliable haplotypes inference. This is the reason why data from different populations are sometimes pooled together to perform a sole haplotypes inference. Despite that, in this study haplotypes reconstruction was separately performed in each single group to avoid potential bias due to the fact that European samples are much more represented respect to other continents samples and since each group was made up of a consistent number of individuals.
All detected polymorphic sites were used for such a reconstruction, but samples with >50% missing genotypes were dropped from the analysis, as recommended in Stephens and Scheet (2005).
The chimpanzee *TNFRSF13B* reference sequence (NC_006484.2 GenBank; Pan_troglodytes 2.1 assembly) was used to define the putative ancestral allele at each SNP, except for intronic substitution C>T at position 90, where the chimpanzee sequence carries a G. In this case, the most frequent allele in analyzed samples was considered as the ancestral one.

3.3.2 Basic descriptive statistics

Estimates of several descriptive statistics, such as nucleotide diversity (π), average number of nucleotide differences (K) and haplotype diversity (H), which are useful to describe the amount of intra-population genetic variability of each sample, were computed using the DnaSp package (Rozas et al. 2003) version 4.50.2.

- ✓ *Nucleotide diversity (π)* is defined as the probability that two randomly chosen homologous nucleotides are different in the sample:

$$\pi = 1/L \left[\sum_{i=1}^{k} \sum_{j<i} p_i p_j d_{ij} \right]$$

where L is the number of loci, k is the number of haplotypes, p_i is the frequency of the i-th haplotype, d_{ij} is an estimate of the number of mutations having occurred since haplotypes i and j divergence (Tajima 1983).

- ✓ *Average number of nucleotide differences (k)* is defined as:

$$k = \left[\sum_{i<j} k_{ij} \right] / \binom{n}{2}$$

where k_{ij} is the number of nucleotide differences between the *i-th* and *j-th* nucleotides (Tajima 1983).

- ✓ *Haplotype diversity (H)* is equivalent to expected heterozygosity for diploid data so that it is expected to be higher in populations with a great number of alleles and lower in populations with few alleles or with only one common allele. It is defined as the probability that two randomly chosen haplotypes are different in the sample:

$$H = (n/n-1) \left[1- \sum_{i=1}^{k} p_i^2 \right]$$

where n is the number of gene copies in the sample, k is the number of haplotypes, and p_i is the frequency of the i-th haplotype (Nei 1987).

In addition to intra-population genetic variability, indices for total nucleotide divergence between humans and chimpanzees (Dxy) and its apportionment in non-coding (Ki), synonymous (Ks) and

non-synonymous (Ka) divergence were also computed with the same software. Finally, the *Ka/Ks* ratio was calculated as a conventional statistic for the measure of global evolutionary constraint on genes.

- ✓ A*verage number of nucleotide substitutions per site between populations (Dxy)*:

$$Dxy = \Sigma_{ij}\, x_i\, y_i\, d_{ij}$$

where x_i and y_i are the frequencies of the *i*-th haplotype in the X and Y population and d_{ij} represents nucleotide substitutions between the *i*-th haplotype from a population and the *j*-th haplotype from the other population (Nei 1987).

- ✓ *Number of non-synonymous substitutions per non-synonymous site (Ka)* and *number of synonymous substitutions per synonymous site (Ks)* for any pair of sequences:

$$Ka = N_d / N \;\; ; \;\; Ks = S_d / S$$

where N and S are the average number of non-synonymous and synonymous sites for the two sequences compared (Nei and Gojobori 1986).

- ✓ *Ka/Ks ratio* has been applied to investigate the *TNFRSF13B* evolution rate both within the human species and when compared to the chimpanzee.

3.3.3 Phylogenetic analysis and dating

Evolutionary relationships among inferred haplotypes were visualized by means of a median joining network (Bandelt et al. 1995) based on the Kruskal's minimum spanning tree and the Farris' maximum parsimony algorithms implemented in the Network 4.5.0.0 software (http://www.fluxus-engigeering.com).

First of all, the Kruskal's minimum spanning tree algorithm produces all possible trees with the shortest distance among observed haplotypes, that is with the minimal value of branch length sum. Subsequently, the Farris' maximum parsimony algorithm generates median vectors, namely the consensus haplotypes, which are necessary to joint together all produced trees. As a consequence, the median joining network represents a synthesis of all parsimony trees, so that all possible evolutionary pathways will be represented (e.g. through cycles).

One network was constructed for each group, for the total sample and for a cases/controls sample.
With the same software, a broad age estimate for analyzed *TNFRSF13B* sequences was calculated on the basis of network topology, by computing the average number of sites differing between a set of sequences and a specified common ancestor (ρ):

$$\rho = \sum_{i=1}^{m} (n_i/n)\, \rho_i$$

where ρ_i is referring to the last common ancestor of the pooled population, n_i is the sample size of each subpopulation and n is the sample size of the total population ($n = n_1 + \ldots + n_m$) (Saillard et al. 2000).
To turn ρ into a time estimate, a mutation rate per site per year (μ) of 6.22×10^{-10} was obtained by dividing total nucleotide divergence between human samples (excluding CVID and IgAD individuals) and the chimpanzee (Dxy = 0.00746) by twice the divergence time between the species (6 million years).
Finally, the most frequent haplotype in the analyzed dataset was used as the specified common ancestor, that is the root sequence of interest used to calculate ρ.

3.3.4 Analysis of population structure

Apportionment of genetic variance at different hierarchical level (F_{ct} among geographically-based groups of populations, F_{sc} within geographically-based groups of populations and F_{st} among individual populations), as well as between cases and controls, was investigated with a locus by locus Analysis of the Molecular Variance (AMOVA) (Excoffier et al. 1992), exploiting information on haplotypes allelic content and frequencies.
A F_{st} index, analogue to AMOVA Wright's F_{st} based on haplotypes pair-wise differences, was also used to calculate pair-wise genetic distances among 26 worldwide populations and CVID and IgAD samples, and to generate a Slatkin's linearized genetic distance matrix (Slatkin 1995). This matrix was finally used for a graphical representation by means of the multivariate analysis Non Metric Multidimensional Scaling (NM-MDS) (Kruskal 1964) that enables reproduction of computed distances in a bi or tri-dimensional space, reducing loss of information and giving stress values for indication about the approximation goodness.
Both AMOVA and genetic distances were computed using the Arlequin 3.01 package (Excoffier et al. 2005), while the StatSoft 6 software (http://www.statsoft.it) was employed to obtain NM-MDS.

3.3.5 Neutrality tests

Several statistical tests have been developed to infer genetic variability deviations from what is expected under neutrality, since it is well known that departures from the neutral model of evolution strongly influence gene genealogies and mutation patterns.

These tests have been classified into three different classes according to the different genetic information they are able to use: Class I tests are based on the frequency spectrum of mutations, Class II on the haplotypes distribution and Class III on the distribution of pair-wise differences (Ramos-Onsins and Rosaz 2002).

Departures form the null hypothesis of neutral evolution at the *TNFRSF13B* gene were tested by using Class I statistics, since it has been proved that they are less sensitive to misspecifying recombination in comparison to those based on haplotypes distribution or mismatch distribution (Ramírez-Soriano et al. 2008).

- ✓ *Tajima's D* is based on standardized difference between the average pair-wise difference *(π)*, which takes into account the number of differences between two sequences, and the Watterson's estimator of theta θ_W, based on the number of segregating sites:

$$D = (\theta_\pi - \theta_w) / \sqrt{\text{Var}\,(\theta_\pi - \theta_w)}$$

where $\theta_\pi = \Sigma_{ij}\, x_i\, x_j \pi_{ij}$ and $\theta_w = S / [\sum_{i=1}^{n-1} 1/i\,]$

where *n* is the number of chromosomes in the sample and *S* the number of segregating sites (Tajima 1989).

Under neutrality, both θ_π and θ_w predict the theoretical value of $\theta = 4N\mu$ so that they are equivalent. As a consequence, Tajima's *D* distribution results centered on 0 under a neutral model of evolution.

Differently, if positive selection or population expansion have acted, an excess of singletons and low frequency variants is found and the number of segregating sites *(S)* results too large compared to π, so that θ_w will be larger than θ_π, leading to more negative Tajima's D values as larger is deviation from neutrality. On the contrary, an excess of intermediate frequency variants, and so positive Tajima's D values due to a too small number of segregating sites *(S)* compared to π, will be caused by balancing selection or population substructure.

- ✓ *Fu and Li's D*, F* and D, F (with a chimpanzee sequence as outgroup)* are based on comparison between an estimator of θ and the number of derived unique mutations in the genealogy external branches. Fu and Li's *D* is computed from the normalized difference between θ_w and the expected number of derived mutations, while *F* uses π instead of θ_w:

$$D = (S - a_n\eta_e) / \sqrt{\text{Var}\,(S - a_n\eta_e)} \quad ; \quad F = (\theta_\pi - \eta_e) / \sqrt{\text{Var}\,(\theta_\pi - \eta_e)}$$

where $a_n = \sum_{i=1}^{n-1} 1/i$

where η_e is the number of derived singletons in the sample.

Under neutrality, the expected number of external mutations is $E[\eta_e] = \theta_\pi = \theta_w = 4N\mu$.

*D** and *F** statistics have been developed since it is not always possible to have an outgroup and, thus, to know whether a singleton is derived or ancestral:

$$D^* = (n/n\text{-}1)\, S - a_n\eta_s \,/\, \sqrt{\text{Var}\,[(n/n\text{-}1)\, S - a_n\eta_s]}$$

$$F^* = (n/n\text{-}1)\, \eta_s - \theta_\pi \,/\, \sqrt{\text{Var}\,[(n/n\text{-}1)\, \eta_s - \theta_\pi]}$$

where η_s is the total number of singletons in the sample (Fu and Li 1993).

- ✓ *Fay and Wu's H (with a chimpanzee sequence as outgroup)* is based on the standardized comparison between π and θ_H:

$$H = (\theta_\pi - \theta_H) / \sqrt{\text{Var}\,(\theta_\pi - \theta_H)}$$

where θ_H is an estimator of θ that gives more weight to high-frequency derived variants based on the expected number of mutations with a derived frequency *i* in the sample.

$$\theta_H = 2 / n(n\text{-}1) \sum_{i=1}^{n-1} i^2\, S_i$$

where S_i is the number of derived variants found *i* times in the sample.

Under neutrality both θ_π and θ_H predict the theoretical value of $\theta = 4N\mu$, so that they are equivalent (Fay and Wu 2000).

All these tests were performed with the DnaSp package (Rozas et al. 2003) version 4.50.2 and the same software was used to test statistical significance of such statistics by running 10,000 coalescent simulations for each one, with a constant population size and an intermediate level of recombination (R = 24.26).

Coalescent simulations produced a distribution of values for each statistic obtained under a neutral model of evolution and useful to be compared with observed values of Tajima's *D*, Fu and Li's *D**, *F**, *D*, *F* and Fay and Wu's *H* in a one-tailed test. In this way, statistical significance of each neutrality test resulted as the portion of coalescent simulations carrying more extreme values than observed ones.

In order to calculate the recombination parameter ($R = 4N_e r$) at the *TNFRSF13B* locus an effective population size for humans of N_e = 10,000 have been considered (Takahata et al. 1995). The recombination rate *r* was obtained using a weighted arithmetic mean of recombination rates between adjacent sites per generation (r_1 = 3.37 cM/Mb, r_2 = 1.83 cM/Mb, r_3 = 0.40 cM/Mb, r_4 = 0.06 cM/Mb) which are available on the on-line OXSTATS Recombination Map database (http://www.mathgen.stats.ox.ac.uk/Recombination.html) and estimated by Myers et al. (2005) for intervals spanning from position 16,764,600 to 16,817,752 (53 kb) of chromosome 17 and covering the *TNFRSF13B* region.

4. Results

4.1 Polymorphic variation at the TNFRSF13B coding region: an overview

Polymorphic variation at the *TNFRSF13B* coding region was investigated by sequencing a total of 2,254 bp for each individual. Such a survey encompassed the gene five exons and their immediately intronic flanking regions and was performed on 902 chromosomes from healthy individuals belonging to 26 worldwide populations, in addition to 154 and 76 chromosomes from Italian subjects diagnosed for CVID and IgAD respectively.

This sequencing approach led to the overall identification of 35 sequence variations, 33 of which were biallelic SNPs, while the remaining two were a single base insertion and a single base deletion (Figure 4.1.1).

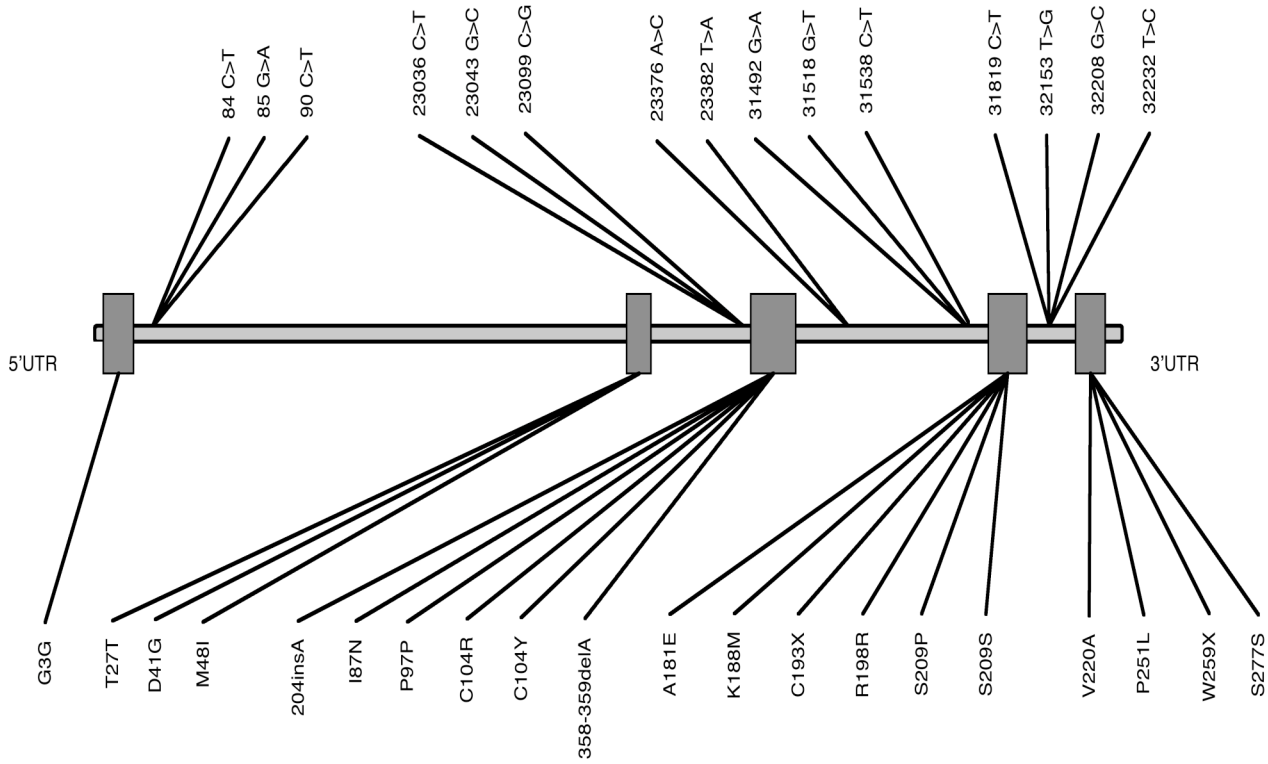

Figure 4.1.1 Genomic structure of *TNFRSF13B* and position of detected polymorphisms.
Intronic variants are reported above the gene, while coding variants are listed below.

Fifteen sequence variations were singletons, that is nucleotide changes with minor alleles which were observed only once among the total sample and in heterozygous state. Among them, six variants resulted to be exclusive of CVID and IgAD individuals. Moreover, eleven polymorphic

sites have already been reported in the on-line SNP database build 129 (dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/) (Table 4.1.1).

Table 4.1.1 Polymorphisms detected in the *TNFRSF13B* gene.

Nucleotide position [a]	*Intronic / mRNA position*	*Protein residue* [b]	*Standard nomenclature* [b]	*SNP id* [c]	*SNP alleles* [d]
22	22	Gly/Gly	G3G		C/**G**
84	*20+10*				***C/T***
85	20+11				**G**/A
90	20+14				**C**/T*
19,525	94	Thr/Thr	T27T	rs8072293	G/**A**
19,566	135	Asp/Gly	D41G		**A**/G
19,588	157	Met/Ile	M48I		**G**/C
23,036	67-70				**C**/T
23,043	*67-63*				***G/C***
23,099	67-7				**C**/G
23,110	*204*		*204insA*		*insA*
23,166	*273*	*Ile/Asn*	*I87N*		***T/A***
23,197	304	Pro/Pro	P97P	rs35062843	**T**/G
23,216	323	Cys/Arg	C104R	rs34755412	**T**/C
23,217	324	Cys/Tyr	C104Y		**G**/A
23,264	*358 359*		*358 359delA*		*delA*
23,376	149+25			rs2274892	**A**/C
23,382	149+31			rs55955502	**T**/A
31,492	150-86				**G**/A
31,518	*150-60*				***G/T***
31,538	150-49			rs56223325	**C**/T
31,674	*555*	*Ala/Glu*	*A181E*		***C/A***
31,695	576	Lys/Met	K188M		**A**/T
31,711	*592*	*Cys/stop*	*C193X*		***C/A***
31,726	606	Arg/Arg	R198R		**G**/A
31,757	638	Ser/Pro	S209P		**T**/C
31,759	640	Ser/Ser	S209S		**C**/T
31,819	210+55				**C**/T
32,153	211-139			rs11652843	**T**/G
32,208	*211-84*				***G/C***
32,232	211-60			rs11652811	**T**/C
32,319	672	Val/Ala	V220A	rs56063729	**T**/C
32,412	765	Pro/Leu	P251L	rs34562254	**C**/T
32,436	789	Trp>stop			**G**/A
32,491	844	Ser/Ser	S277S	rs11078355	T/**C**

Polymorphisms in italics are private CVID and IgAD variants, of which underlined ones are singletons.

[a] Position in the reference sequence (NG_007281.1 GenBank).
[b] Coding variants.
[c] dbSNP build 129.
[d] Inferred ancestral state in bold (see the Materials and Methods section).
* SNP with ambiguous phylogenetic information (see the Materials and Methods section).

Twenty polymorphisms lay in exons: six were synonymous changes, ten were replacement changes, two led to non-sense mutations and two led to frameshift mutations.
In particular, six coding variants (204insA, I87N, C104R, 358_359delA, A181E and C193X) have been previously observed in clinical case-control studies (Castigli et al. 2005a; Castigli et al. 2007; Salzer et al. 2005). Among them 204insA, I87N, 358_359delA, A181E and C193X resulted private diseases mutations, with I87N, 358_359delA and A181E which were exclusive of CVID individuals, whereas 204insA and C193X were found also in IgAD subjects.
Nevertheless, such coding variants extremely low frequencies resulted in little statistical power to detect significant differences of frequency between cases and controls. As a matter of fact, substitution A181E only showed a statistically higher frequency in CVID subjects respect to healthy individuals (Fisher's Exact Test, $p<0.05$).
Finally, amminoacid replacement C104R was the sole non-private diseases mutation, being found both in healthy Italians and in CVID and IgAD individuals with nearly the same frequency.

4.2 Patterns of genetic diversity

Summary statistics describing *TNFRSF13B* coding region genetic diversity were separately estimated in geographically-based groups of populations, Italian CVID and IgAD samples, Italian control group and Val di Scalve population and are reported on Table 4.2.1.
A regards nucleotide diversity (π), which was calculated as the average heterozygosity per site, a value of 0.00110 was obtained for the total sample and did not change significantly after excluding CVID and IgAD individuals from the analysis.
Compared to literature values, such a statistic resulted to be nearly twice the mean value estimated for 292 autosomal genes ($\pi = 0.00058$) (Stephens et al. 2001) and higher than the mean value calculated from the 320 genes re-sequenced by the Seattle SNPs project ($\pi = 0.00085$) (http://pga.gs.washington.edu/). Moreover, it was also higher than the value obtained as an average for eight genes involved in immune functions and selected from the Innate Immunity Program in Genomics Application database ($\pi = 0.00094$) (IIPGA, http://innateimmunity.net/), which were recently surveyed for signatures of selection (Ferrer-Admettla et al. 2008).
Interestingly, the Sub-Saharan African sample showed one of the lowest values of nucleotide diversity ($\pi = 0.00063$), as already reported in Mateu et al. (2001), but in contrast with the trend observed for several other genes (Tishkoff et al. 1996; Calafell et al. 1998; Tishkoff et al. 1998; Guthery et al. 2007) and for Seattle SNPs re-sequenced genes, for which African-Americans showed greater nucleotide diversity in comparison to European-Americans in 82.6% of cases.

Table 4.2.1 Summary statistics for *TNFRSF13B* genetic diversity.

	N	*S*	*s*	*k*	*H*	π	*K*
S.S. Africa	114	10	1	14	0.802 ± 0.022	0.00063	1.416
N. Africa	74	10	3	17	0.849 ± 0.026	0.00101	2.280
M. East	192	14	5	23	0.822 ± 0.021	0.00107	2.411
C. Asia	56	7	3	9	0.795 ± 0.039	0.00056	1.267
S. America	82	7	1	12	0.792 ± 0.028	0.00062	1.401
Italy	192	13	4	20	0.759 ± 0.025	0.00119	2.678
CVID	154	20	9	24	0.767 ± 0.032	0.00115	2.592
IgAD	76	12	3	14	0.784 ± 0.040	0.00121	2.724
Val Scalve	192	8	3	11	0.695 ± 0.026	0.00093	2.087
Total	1132	35	15	64	0.865 ± 0.006	0.00110	2.482

N, number of chromosomes; *S*, number of polymorphic sites; *s*, number of singletons; *k*, number of haplotypes; *H*, haplotype diversity; π, nucleotide diversity; *K*, average number of nucleotide differences.

Also concerning haplotype diversity (H), which is defined as the probability of two haplotypes randomly chosen in the sample to be different, a lower value was found for the Sub-Saharan African sample (H = 0.802) respect to North African and Middle Eastern ones (H = 0.849 and H = 0.822 respectively).

Finally, the same peculiar trend can be observed also for the number of polymorphic sites (S), singletons (s) and haplotypes (k), as well as for the average number of nucleotide differences (K), pointing out again an unusual scarceness of variability in Sub-Saharan Africans.

4.3 Divergence between humans and chimpanzees

Fifteen fixed nucleotide differences were observed between sequences of healthy humans and the chimpanzee reference sequence (NC_006484.2 GenBank, Pan_troglodytes 2.1 assembly), one of which, the C to G transversion at nucleotide position 23,342, was the sole leading to a replacement change (P146A).

A total nucleotide divergence (Dxy) of 0.75% was calculated, together with its apportionment in non-coding (Ki), synonymous (Ks) and non-synonymous (Ka) divergence (Table 4.3.1).

Table 4.3.1 Divergence between human and chimpanzee *TNFRSF13B* sequences.

Dxy	*Ki*	*Ks*	*Ka*	*Ka/Ks*
0.75%	1%	0.87%	0.18%	0.202
(2,254 bp)	(1,375 bp)	(667 bp)	(212 bp)	(879 bp)

Dxy, total nucleotide divergence; *Ki*, non-coding divergence; *Ks*, synonymous divergence; *Ka*, non-synonymous divergence. In brackets the number of surveyed nucleotide sites.

A Ki of 1% was estimated on the basis of the 1,375 non-coding bp analyzed and turned out to be lower than 1.27% computed as the average value for 12,997 autosomal genes, only 5% of which shows lower non-coding divergences (The Chimpanzee Sequencing and Analysis Consortium 2005).

Small values of synonymous and non-synonymous divergence (Ks = 0.87% and Ka = 0.18%) were also observed with respect to average values estimated for genomic divergence between humans and primates (Ks = 1.42% and Ka = 0.34%) (Chen and Li 2001), but they fall respectively in the 37% and 47% percentiles of such a genome-wide distribution.

Finally, as a conventional statistic for the measure of global evolutionary constraint on genes, the Ka/Ks ratio was estimated, obtaining a value of 0.202 that was quite similar to the mean value reported by The Chimpanzee Sequencing and Analysis Consortium (Ka/Ks = 0.23).

4.4 Haplotypes structure in the total sample

An ancestral haplotype, which was not found in the surveyed samples, was inferred by using the chimpanzee *TNFRSF13B* reference sequence to recover putative ancestral alleles at each polymorphic site.

Haplotypes inference was separately performed in five large geographically-based clusters of populations, in Italian groups of CVID, IgAD and healthy individuals and in the Val di Scalve sample.

Such haplotypes reconstruction led to the overall identification of 64 different haplotypes (Table 4.4.1), the frequency distribution of which showed that the five most frequent haplotypes accounted for 68% of sampled chromosomes.

Among these five most frequent haplotypes h3, h6 and h9 were found in all groups except in the Val di Scalve sample, reaching a cumulative frequency of 46%. In particular, h3 was the most represented haplotype nearly in all populations, with a frequency always ≥25%, except for the Central Asian sample (7%).

Table 4.4.1 *TNFRSF13B* haplotypes and their absolute frequencies in the surveyed groups.

Haplotype labels	11122222222222333333333333333333 99933333333333311111111112222222 55500011122233455667777781223444 28892683491691167891379125515031139 24505686390676746228845167993829261	M. East (192)	Italy (192)	Val Scalve (192)	IgAD (76)	CVID (154)	SS. Africa (114)	N. Africa (74)	C. Asia (56)	S. America (82)	Total (1132)
Chimp	CCGGAAGCGCATTTGAATGGCCACGTCCTGTTCGC * #										
h10	...T.....	1	-	-	-	-	-	-	-	-	1
h11	...T.....A.....	1	-	-	-	-	-	-	-	-	1
h2	...C.....	15	-	-	3	6	38	15	6	27	110
h63	G..C.....	-	-	-	-	-	-	-	1	-	1
h54	...C.....C....	-	-	-	-	1	-	-	-	-	1
h26	...C.G.....	-	1	-	-	1	-	-	-	-	2
h15	...C.....T.....	1	-	-	-	-	3	-	-	-	4
h60	...C.....T.....T..	-	-	-	-	-	-	1	-	-	1
h14	...C.....T..	2	4	-	1	-	8	2	-	5	22
h5	...C.....A.....T..	1	-	-	-	-	-	-	-	-	1
h41	...C.....C.....	-	-	-	2	-	-	-	-	-	2
h27	...C.....C.....T	-	2	-	-	5	-	-	-	-	7
h3	...C.....T	71	83	-	32	70	28	22	4	24	334
h53	...C.....A.....T	-	-	-	-	1	-	-	-	-	1
h47	...C.....A.....C....T	-	-	-	-	2	-	-	-	-	2
h49	...C.....A.....A.....C....T	-	-	-	-	1	-	-	-	-	1
h43	...C.....G.....T	-	-	-	-	2	-	-	-	-	2
h30	...C.....T.....T	-	1	-	-	-	-	-	-	-	1
h51	...C.....G.....T	-	-	-	-	1	-	-	-	-	1
h46	...C.....T.....T	-	-	-	-	1	-	-	-	-	1
h23	...C.....C.....T	1	-	-	-	-	-	-	-	-	1
h64	...C...T.....T	-	-	-	-	-	-	-	-	1	1
h4	...C.....G.....T	5	-	-	-	-	-	-	-	-	5
h12	...C.....C.....T	6	1	-	-	-	3	3	7	-	20
h50	...C.....A.....T	-	-	-	-	1	-	-	-	-	1
h57	.T.C.....T	-	-	-	-	-	1	-	-	-	1
h52	.T.C.....	-	-	-	-	1	1	-	-	-	2
h22	...C.....G.....	6	5	-	3	3	2	-	-	-	19
h44	...C.....G...C.....	-	-	-	-	2	-	-	-	-	2
h59	...C.....G..G.....	-	-	-	-	-	-	1	-	-	1
h32	...C.....G.....T..	-	1	-	-	-	-	-	-	-	1
h35	...CG.....G.....	-	-	1	-	-	-	-	-	-	1
h18	...CG.....	2	-	5	-	-	-	1	-	5	13
h33	...CG.....C.....T..	-	1	32	-	-	-	-	-	-	33
h38	...CG.....T..	-	-	1	-	-	-	-	-	-	1
h37	...CG.....T.T	-	-	1	-	-	-	-	-	-	1
h20	...CG.....T	3	-	93	-	-	-	-	-	-	96
h36	...CG..T.....T	-	-	1	-	-	-	-	-	-	1
h19	...CG.....T.....T	1	-	-	-	-	-	-	-	-	1
h13	...CG.....C.....T	2	3	-	-	-	1	-	-	-	6
h34	...CG.....CA.....	-	-	1	-	-	-	-	-	-	1
h31	...CG.....C.....	-	1	8	-	-	-	1	4	3	17
h1	...CG.....C.....G.C....	25	37	39	11	17	2	9	-	-	140
h21	...CG.....C.....C..G.C....	2	-	-	-	-	-	-	-	-	2
h55	...CG.....G.....C.....G.C....	-	-	-	-	1	-	-	-	-	1

h40	...CG.....G.....C......A....G.C....	-	-	-	1	-	-	-	-	-	1
h45	...CG...C.......C......A....G.C....	-	-	-	-	1	-	-	-	-	1
h8	...CG..........C..........G.C...T	7	10	10	3	5	1	-	-	-	36
h25	...CG..........C..........G.CC..T	-	5	-	3	6	-	-	-	-	14
h39	...CG.....................G.C....	-	-	-	1	-	-	-	-	-	1
h62	...CG.....................G.C...T	-	-	-	-	-	-	1	-	-	1
h42	...C......................G.C...T	-	-	-	1	-	-	-	-	2	3
h58	...C...........C..........G.C...T	-	-	-	-	-	-	2	1	1	4
h61	...C...........C..........G.C..AT	-	-	-	-	-	-	1	-	-	1
h17	...C......................G.C....	3	2	-	-	-	-	1	-	1	7
h48	...C........G.............G.C....	-	-	-	-	3	-	1	-	-	4
h7	...C...........C..........G.C....	1	-	-	-	-	-	1	-	1	3
h6	...C...........C.............T..	21	22	-	10	16	9	6	9	3	96
h56	..AC...........C.............T..	-	-	-	-	-	1	-	-	-	1
h29	...C.........A.C.............T..	-	1	-	1	-	-	-	-	-	2
h24	...C..C........C.............T..	-	1	-	-	-	-	-	-	-	1
h28	...C...........C.............T.T	-	1	-	-	1	-	-	2	-	4
h16	...C...........C.........T.....T.T	1	-	-	-	-	-	-	-	-	1
h9	...C...........C..................	14	10	-	4	6	16	6	22	9	87

Polymorphisms are listed in the second column below their nucleotide position and the corresponding chimpanzee ancestral position; ancestral chimpanzee-like alleles are indicated by dots. Variants *204insA* and *358_359delA* were re-codified as biallelic SNPs: *haplotypes with the G allele at position 23,110 carry the *204insA* mutation; # haplotypes with the G allele at position 23,264 carry the *358_359delA* mutation. In brackets the number of analyzed chromosomes.

Haplotypes h1 and h2 were instead less common respect to h3, h6 and h9. Haplotype h1 was rare in Sub-Saharan Africans (1.8%) and even completely absent in Central Asians and South Americans, whereas h2 was not observed in Italian and Val di Scalve samples (Figure 4.4.1). Interestingly, this haplotype was the most closely related to the ancestral one, which was inferred from the chimpanzee sequence, showing a single divergent site at nucleotide position 90.

In these five most frequent haplotypes, intronic substitutions 23,376A>C, 32,153T>G, 32,232T>C, synonymous changes 19,525G>A (T27T), 32,491T>C (S277S) and replacement substitution 32,412C>T (P251L) were the sole represented polymorphisms.

On the contrary, three different haplotype typologies accounted for remaining 32% of analyzed chromosomes.

The first typology was actually represented by a sole haplotype (h20), which was found in only two groups and whose global frequency of 8% was mainly due to the very high frequency observed in the Val di Scalve sample (48%), in contrast to only 2% found in the Middle Eastern sample.

The second typology was represented by 18 haplotypes, whose frequencies in the total sample ranged from 0.27% to 3.18% and merged into a cumulative frequency of 20%.

Finally, the third typology was that of rare haplotypes (40), which were found only once or twice

in the total sample and which showed a cumulative frequency of 4%.

Interestingly, the higher percentage of rare haplotypes (18%) was found in North Africans and was higher also in comparison to CVID and IgAD percentages (11% and 9% respectively), followed by Sub-Saharan African and Middle Eastern samples (8%), Italians, Central Asians and South Americans (7%) and, at the end, Val di Scalve sample (3%).

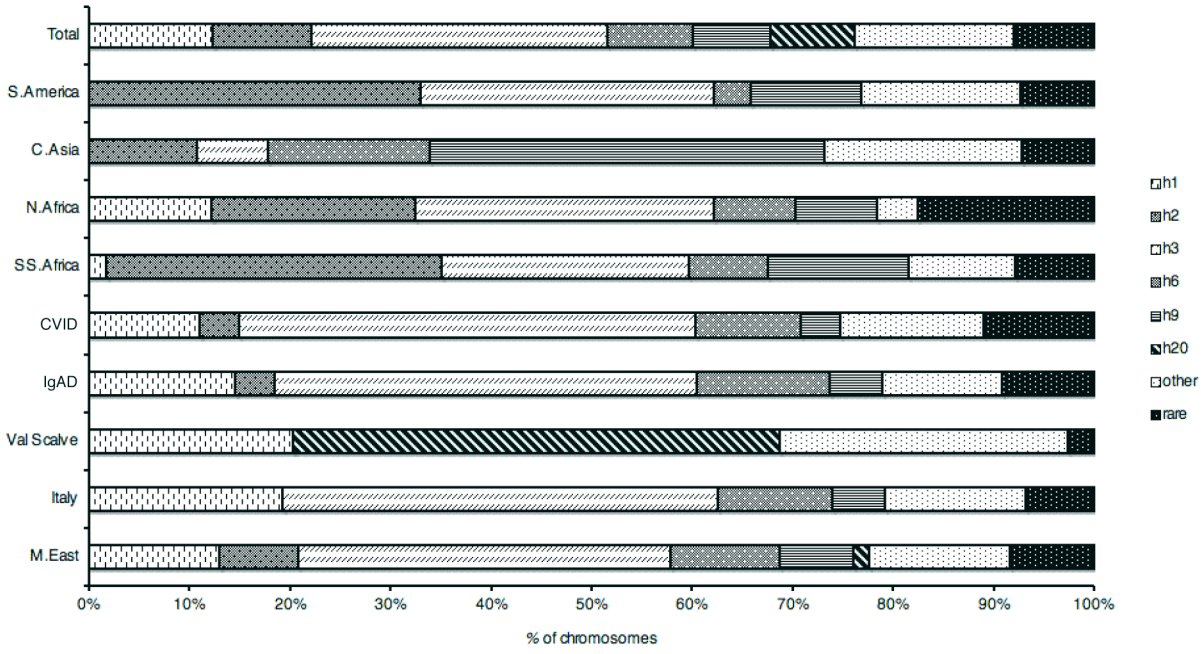

Figure 4.4.1 Relative frequencies of *TNFRSF13B* haplotypes in the surveyed groups. The five most frequent haplotypes were individually represented, haplotypes with frequencies ranging from 0.27% to 3.18% are pooled as *other* and haplotypes found only once or twice are pooled as *rare*.

4.5 Haplotype structure in Italian CVID, IgAD and healthy samples

A very similar haplotypes frequency distribution was shown by Italian CVID, IgAD and healthy samples, especially as regards the five most frequent worldwide haplotypes (Figure 4.4.1) and Italian and Middle Eastern distinctive haplotypes h8, h22, h25 (Table 4.4.1).

In particular, these haplotypes carried synonymous substitution 23,197T>G (P97P) and replacement change 32,319T>C (V220A), in addition to nearly the same polymorphisms of the five haplotypes mentioned above.

On the whole, such a eight-haplotypes cluster accounted for 89% of Italian healthy chromosomes and for 91% and 84% of IgAD and CVID chromosomes respectively.

The group of individuals affected by CVID presented another non-rare haplotype (h27), which

carried supposed pathogenic substitution 23,216T>C (C104R) (Garibyan et al. 2007) and showed a higher, but not significantly different, frequency in the CVID sample respect to the control one (3.25% vs. 1.04%; Fisher's Exact Test, $p = 0.25$).

As discussed in the previous section, rare haplotypes accounted for 7% of Italian healthy chromosomes, while their frequency rose to 9% in IgAD and 11% in CVID groups. In more detail, three rare haplotypes only turned out to be exclusive of controls and of IgAD individuals (on a total of 11 and six rare haplotypes respectively), while they reached the number of 11 in the CVID sample (on a total of 14 rare haplotypes).

In that case, haplotypes h53, h47 and h49 carried replacement change 31,674C>A (A181E), which was the sole supposed pathogenic variant with a significantly higher derived allele frequency in individuals diagnosed for CVID respect to healthy subjects (Fisher's Exact Test, $p < 0.05$).

4.6 Phylogenetic analysis of Italian CVID, IgAD and healthy haplotypes

Evolutionary relationships among inferred CVID, IgAD and healthy haplotypes were visualized by means of a median joining network constructed for a cases/controls pooled sample (Figure 4.6.1).

This graphical representation disclosed that haplotypes h50 and h51, which carried substitution 23,217G>A (C104Y) and 358_359delA deletion respectively, as well as already described h53, h47 and h49, were one-step neighbors of the most represented haplotype h3.

On the contrary, haplotype h45 and h55, which carried substitution 31,711C>A (C193X) and 204insA insertion respectively, derived from the second most frequent haplotype (h1), as well as IgAD private haplotype h40, which carried both mutations at the same time.

Moreover, haplotype h41, with substitution 23,216T>C (C104R), also resulted exclusive of IgAD individuals, while the remaining four private IgAD haplotypes carried only intronic variants 32,153T>G, 32,232 T>C, 31,518 G>T, 23,376 A>C and the synonymous change 23197T>G (P97P).

Finally, haplotype h29, which was derived from the cosmopolitan haplotype h6 and shared between controls and IgAD subjects, carried another potential pathogenic mutation, substitution 23,217G>A (C104Y).

Identification of circles originated from some branches of the median joining network, in accordance to observation of 64 different haplotypes with respect to 35 variants only, also represented a reliable clue about the level of recombination occurred within the *TNFRSF13B* gene.

A minimum number of six recombination events (Rm) was indeed inferred for the total sample and each single group also presented a certain degree of recombination, with Sub-Saharan Africans

showing the lowest one (Rm = 1), contrary to all expectations.

Although changes in nucleotide patterns along sequences could be used to infer which ones may be the outcome of recombination (Parida et al. 2008), the pattern of analyzed sequences resulted too complex to allow a reliable characterization of recombinant haplotypes.

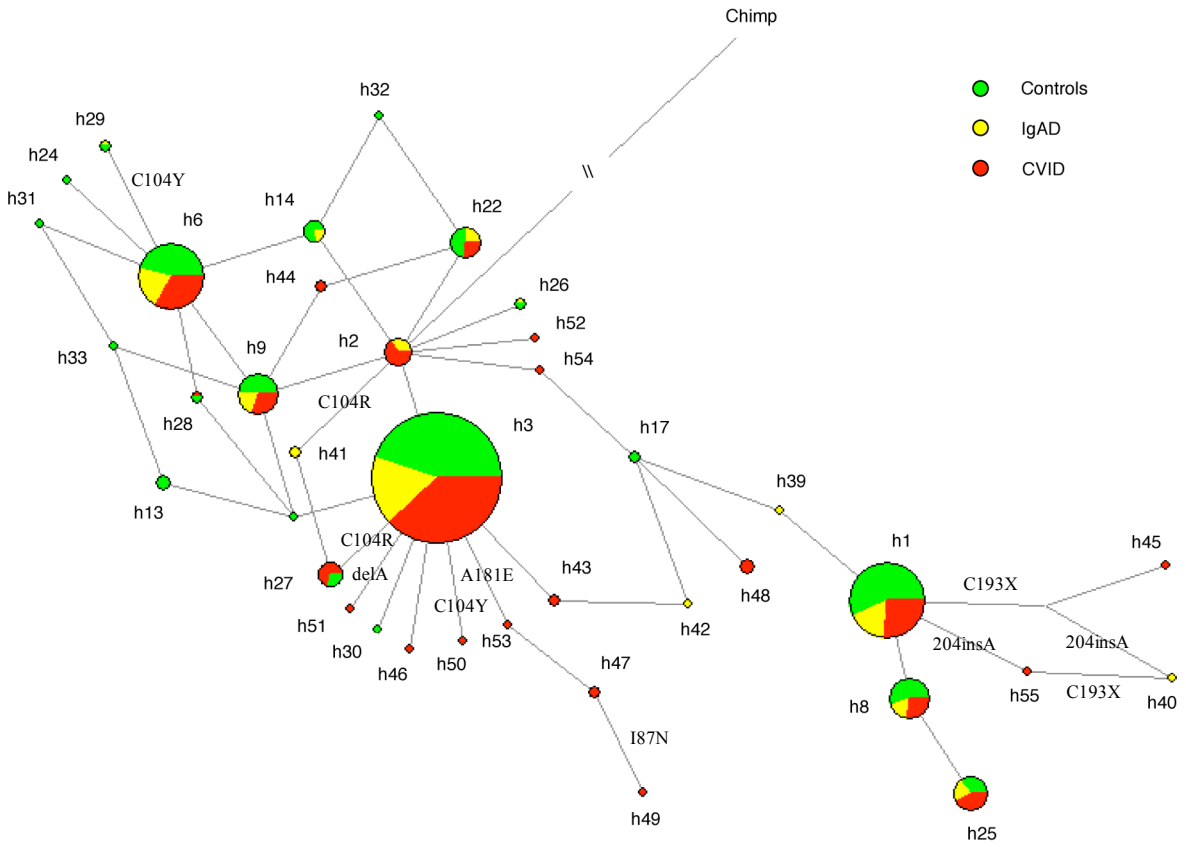

Figure 4.6.1 Median joining network of *TNFRSF13B* haplotypes in Italian cases and controls. The nodes are proportional to haplotype frequencies, while the branch lengths are proportional to the number of variants occurred in the sequences, except for the branch tracing back to the chimpanzee, which encompasses 15 divergent nucleotide positions between humans and chimpanzee. Supposed pathogenic variants only are reported on branches.

4.7 Dating of healthy and diseases haplotypes

As discussed in the previous section, recombination has undoubtedly played a role in shaping the genealogy of *TNFRSF13B* sequences, so that unambiguous time estimates for inferred haplotypes are hardly achievable.

Nevertheless, it was at least possible to get glimpses about the rough history of events that describes the gene evolution using total divergence between humans and chimpanzee (Dxy = 0.00746), in order to estimate the time to the most common recent ancestor.

Taking into account a divergence time of six million years for *H. sapiens* and *P. troglodytes*, a mutation rate per site per year (μ) of 0.622×10^{-9} was obtained and corresponded to an average number of base substitutions, from inferred ancestral haplotype to each analyzed sequence, of 1.853 ± 0.623. Such amount of variation required about 1.323 ± 0.444 million years to accumulate.

The same rationale has been applied in order to date expansion times of haplotypes carrying variants of CVID and IgAD individuals. In this way, it has been estimated that 61,283 ± 34,577 years were required for C104R, 358_359delA and A181E changes to accumulate from the most frequent haplotype h3, whereas 52,483 ± 25,711 years were required for 204insA and C193X to appear on the second most represented haplotype h1.

However, this kind of inference necessarily represents an overestimate of actual haplotypes expansion times, since it was based on a "biased" sample (i.e. individuals affected by CVID and IgAD) rather than on a random population sample.

4.8 Analysis of population structure

Excluding Arequipa and Egyptians because of their very small sample sizes, genetic distances among 24 worldwide populations, as well as between CVID and IgAD samples, were measured using F_{st} as an estimate of their allele frequency differentiation.

A graphical representation of such genetic distances was obtained from the computed genetic distance matrix by means of a Non Metric Multidimensional Scaling (NM-MDS) (Figure 4.8.1), revealing that a clear pattern of geographical structure for *TNFRSF13B* genetic diversity was hardly recognizable.

As a matter of fact, North African, Middle Eastern, Italian CVID, IgAD and healthy samples were represented in the NM-MDS as gathered into an indiscernible single cluster, South Americans instead occupied an intermediate position between the previous cluster and Sub-Saharan Africans, while Central Asian populations remained nearly isolated probably as a consequence of their lowest sequence diversity, due to smaller sample sizes.

Moreover, in a first plotting the Val di Scalve sample also stood out as a clear outlier, relegating other populations in a cloud very hard to disentangle, so that it was subsequently excluded from the NM-MDS computation.

Analysis of Molecular Variance (AMOVA) (box in Figure 4.8.1) also provided statistical support for the absence of a sharp geographical structure for *TNFRSF13B* genetic diversity, showing a very low and barely significant level of differentiation among geographically-based groups of populations (Sub-Saharan Africa, North Africa, Middle East, Europe, Central Asia and South

America) (F_{ct} = 0.0658, p = 0.026), if compared to the average value of 0.10 estimated for 109 DNA loci (Barbujani et al. 1997).

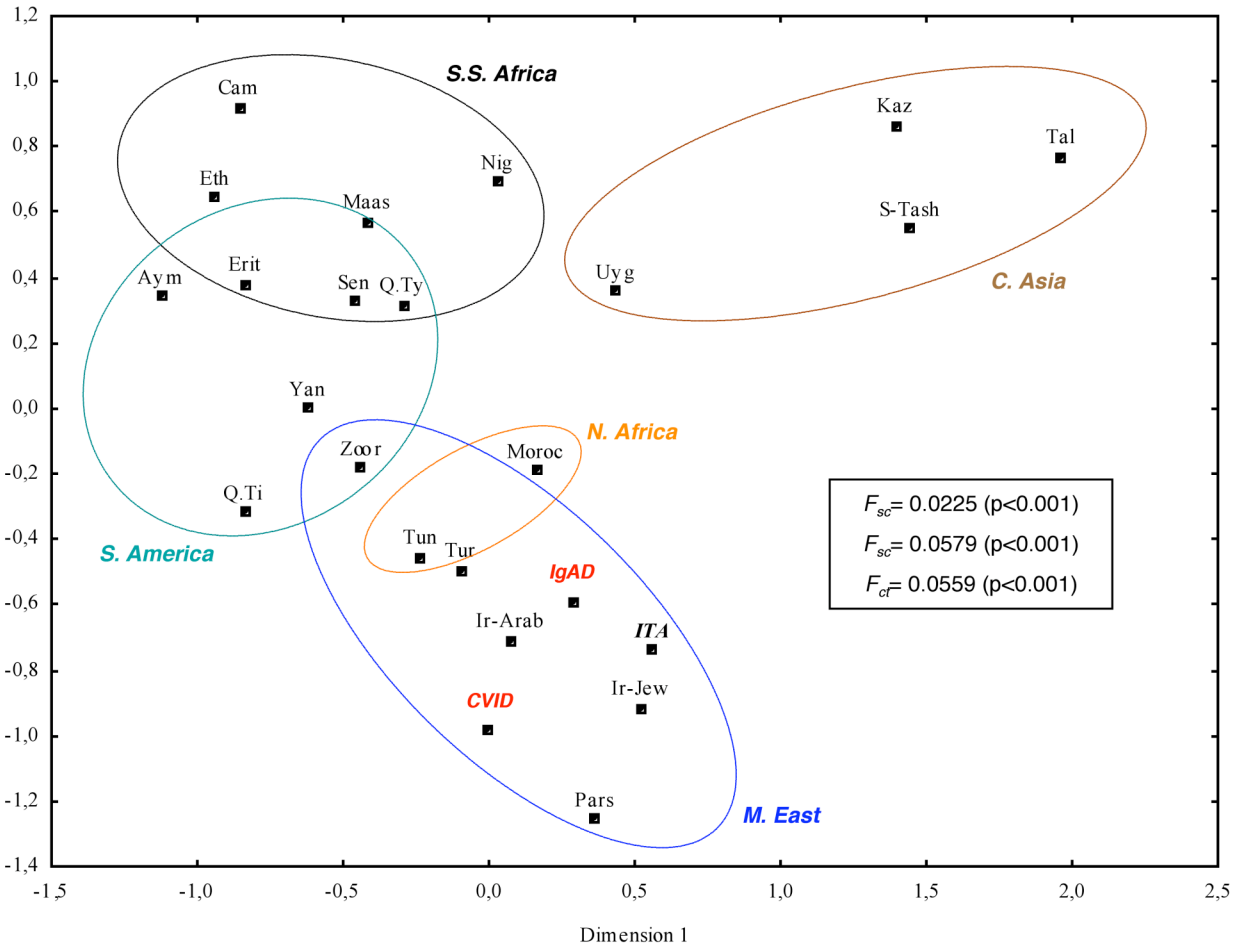

Figure 4.8.1 Multidimensional Scaling of genetic distances among analyzed groups. Val di Scalve sample was excluded from the MDS. CVID and IgAD samples in red bold, controls in black bold. The apportionment of genetic variance at the *TNFRSF13B* coding region resulting from AMOVA is reported in the box.

However, such a minimal differentiation increased its statistical significance after removing Val di Scalve sample from the analysis (F_{ct} = 0.0559, p<0.001), as well as the level of differentiation among populations within the same geographically-based group, which shifted from F_{sc} = 0.0761 to F_{sc} = 0.0225, but maintaining a significance of p<0.001.

Most importantly, after removing Val di Scalve from AMOVA computation the level of differentiation among populations, represented by the F_{st} value, drastically dropped from 0.14, an extent well-comparable to the average computed for several human populations (F_{st} = 0.15) (Romualdi et al. 2002) and to the mean value for the Alfred database (F_{st} = 0.13) (http://alfred.med.yale.edu/alfred/), to only 0.0579 (p<0.001).

Nevertheless, even if very low, such a value perfectly lay within known distribution of F_{st} and

resulted very similar to those observed for other genes recently surveyed for signatures of selection, such as *OR511* (F_{st} = 0.06) (Moreno-Estrada et al. 2008), *CD14* and *TLR9* (F_{st} = 0.07 and F_{st} = 0.05 respectively) (Ferrer-Admetlla et al. 2008) and *ABO* (F_{st} = 0.06 for Seattle SNPs data and F_{st} = 0.07 for re-sequencing data) (Calafell et al. 2008).

The same F_{st} index was also used to measure genetic differences between Italian diseases and healthy groups of individuals. Comparing IgAD and control samples an F_{st} value of -0.00629 (p = 0.823) was found, whereas a thin difference was observed between CVID sample and the control one, but with a borderline statistical significance (F_{st} = 0.01046, p = 0.043).

4.9 Neutrality tests

Genetic footprints of selection at the *TNFRSF13B* gene were sought for by computing several neutrality tests, such as Tajima's *D*, Fu and Li's *D*, *F* and Fay and Wu's *H*, which are based on the frequency spectrum of mutations, since it has been proved that they are less sensitive to misspecifying recombination in comparison to those based on haplotypes distribution or mismatch distribution (Ramírez-Soriano et al. 2008). In this way, a proper test for departures from the null hypothesis of neutral evolution within the pattern of observed sequence variation was ensured.

After coalescent simulations performed to evaluate their statistical significance, Tajima's *D* estimates resulted not significantly different from zero, showing very small negative and positive values, except for Val di Scalve and total samples (*D* = 1.156, p = 0.914 and *D* = -1.163, p<0.05 respectively) (Table 4.9.1). Although not significantly different from zero, the Val di Scalve large positive Tajima's *D* value suggested a condition close to a scarceness of segregating sites and an excess of intermediate frequency alleles; however, the other sole noteworthy statistic for this sample (Fu and Li's *D* = -1.451, p<0.05) did not result significantly different from zero after correction for multiple testing.

An excess of singletons and low frequency variants generally causes a negative Tajima's D statistic, as observed for the total sample. In this case, such an excess of rare variants was mainly contributed by polymorphisms of CVID and IgAD samples, since total Tajima's *D* value notably decreased, and became non significant, after removing diseases individuals from the analysis (*D* = -0.832, p = 0.107).

As regards Fu and Li's *D* and *F* statistics, the sole significant values were associated with CVID and total samples (*F* = -2.354, p<0.01; Fu and Li's *D* = -2.743, p<0.01 and *F* = -3.917, p<0.001; Fu and Li's *D* = -4.926, p<0.001 respectively). Both showed significant large negative values even after

correction for multiple testing. Nevertheless, excess of singletons and low frequency variants in the CVID sample is presumably due to an actual relationship between *TNFRSF13B* sequence variations and pathological conditions of surveyed individuals (unlike for IgAD subjects), while total sample negative values, which remained nearly the same excluding CVID and IgAD individuals from the analysis, could be more plausibly linked to demographic expansion of anatomically modern humans, rather than to an effective selective sweep, as seen in most of the human genome (Stephens et al. 2001).

Table 4.9.1 Neutrality tests for *TNFRSF13B* in the surveyed groups.

	N	*D (p)*	*F (p)*	*D* (p)*	*H (p)*
SS.Africa	114	-0.623 (0.229)	0.236 (0.575)	0.652 (0.585)	1.009 (0.816)
N.Africa	74	0.298 (0.668)	-0.389 (0.330)	-0.659 (0.339)	1.381 (0.938)
M.East	192	0.011 (0.533)	-1.319 (0.087)	-1.875 (0.066)	1.241 (0.859)
C.Asia	56	-0.437 (0.313)	-1.146 (0.139)	-1.227 (0.188)	-0.266 (0.312)
S.America	82	-0.009 (0.515)	0.276 (0.583)	0.355 (0.797)	1.024 (0.934)
Italy	192	0.501 (0.756)	-0.668 (0.237)	-1.197 (0.186)	1.042 (0.777)
CVID	154	-0.751 (0.141)	-2.354 (0.007)	-2.743 (0.008)	1.273 (0.052)
IgAD	76	0.311 (0.683)	-0.132 (0.425)	-0.344 (0.459)	1.342 (0.862)
Val Scalve	192	1.156 (0.914)	-0.627 (0.253)	-1.451 (0.038[a])	0.719 (0.761)
Total	1132	-1.163 (0.025)	-3.917 (0.000)	-4.926 (0.000)	1.249 (0.698)

N, number of chromosomes; *D*, Tajima's D; *F*, *D**, Fu and Li's F and D, with chimpanzee as outgroup; *H*, Fay and Wu's H; *(p)*, p-value obtained by coalescent simulations.
[a] non significant after correction for multiple testing.

5. Discussion and Concluding Remarks

Human Primary Immunodeficiencies Diseases (PIDs) represent a heterogeneous group of disorders in which inherited genetic defects compromise the ability to produce effective immune responses.
Until the last decade PIDs were thought to be few rare, familial and monogenic, recessive traits impairing the development of one or several leukocyte subsets and resulting in multiple, recurrent, fatal infections in infancy. In the recent years this conventional view has instead undergone a substantial change (Casanova and Abel 2007), opening the way for broader and deeper genetic analyses on such a class of diseases. These studies quickly turned out to be very important both for elucidating crucial functional pathways in immune responses and for prompting the design of new diagnostic tools and therapeutic researches.
Humoral PIDs, which are antibody-related defects characterized by B cells differentiation and immunoglobulins production defects, account for 65% of all primary immunodeficiencies (Yin et al. 2001), with Common Variable Immunodeficiency (CVID) that stands out as the most common clinically relevant primary immunodeficiency, representing about 30% of all PIDs affected individuals in Europe (Eades-Perner et al. 2007).
As regards humoral immunity, it has been far-back proved that differentiation of mature B cells into effectors capable of specific immune responses is strictly regulated and that Tumor Necrosis Factor Receptor Superfamily (TNFRSF) members undoubtedly play important and diversified roles in the regulation of activation and apoptosis of several immune cell types.
In particular, the *TNFRSF13B* protein product TACI plays a pivotal role in a very complex ligands/receptors network by binding the TNF ligands B cell-activating factor (BAFF) and a proliferating inducing ligand (APRIL) (Mackay et al. 2003), so that its function in regulation of isotype switching, survival and differentiation of B lymphocytes is nowadays accepted, even if only partially understood (Mackay and Schneider 2008).
Conversely, what still remains enigmatic is the degree of association of such gene variants with the heterogeneous spectrum of CVID. Although *TNFRSF13B* defects represent the most common DNA sequence variations in individuals affected by CVID, being found in about 10-20% of disease subjects (Park et al. 2008), after the initial claim of their strict association with CVID and Selective IgA Deficit (IgAD) (Castigli et al. 2005a; Salzer et al. 2005), further analyses have indeed shown the existence of some of these variants also in healthy individuals (Castigli et al. 2007; Pan-Hammarstrom et al. 2007; Lee et al. 2008; Salzer et al. 2008) and, besides, their functional effects in relation to the development of these diseases have not been yet established.

Moreover, up to the present, clinical genetic studies which investigate on *TNFRSF13B* involvement in CVID, IgAD, autoimmune disorders and some other diseases are principally based on European and North American of European ancestry cases and controls, with very few exceptions (Inoue et al. 2006; Lee et al. 2007), so that this work actually represents the first global survey of *TNFRSF13B* nucleotide diversity and haplotype structure.

The rationale behind this population-based approach is that analysis of samples with different ancestry ensures the reconstruction of a plausible evolutionary history for the examined genomic region and this could turn out to be truly significant for shedding light on its actual role in immune functions and for facilitating distinction between variants falling into the standard degree of intra-specific variation and changes which are potentially related to a so complex and multifaceted disease such as CVID.

For this purpose *TNFRSF13B* exons and their immediately intronic flanking regions were sequenced in a worldwide panel of 26 human populations from Sub-Saharan Africa, North Africa, Middle East, Central Asia and South America, for a total of 451 healthy individuals, in addition to 96 healthy, 77 CVID and 38 IgAD individuals from Italy.

This led to identification of 13 unpublished and 22 known sequence variations, almost all in heterozygous state, in accordance to clinical studies from which it has been estimated that less than 2% of CVID subjects carry biallelic mutations (Salzer et al. 2008).

In particular, already described 204insA, I87N, 358_359delA, A181E and C193X coding variants (Salzer et al. 2005; Castigli et al. 2005a; Castigli et al. 2007) were observed in diseases-individuals only, even if nearly all their extremely low frequencies resulted in little statistical power to detect significant differences between cases and controls.

Supposed pathogenic substitution A181E only showed a statistically higher frequency in CVID subjects, but not in IgAD ones, respect to healthy individuals, in accordance to several studies which report a significant association of this mutation with CVID (Castigli et al 2007; Pan-Hammarstrom et al. 2007; Zhang et al. 2007; Park et al. 2008). Such an observation supports the hypothesis that replacement of a neutral to a negatively charged amminoacid in the protein transmembrane domain may prevent downstream signaling by interfering with ligand-induced conformational changes of the whole receptor also in a mutant-wild-type complex.

The other sole supposed pathogenic substitution C104R was instead present both in healthy Italians and in CVID and IgAD individuals with nearly the same frequency and was also the sole homozygous variant observed in the CVID sample. It disrupts a disulphide bond required for ligand biding CRD-2 in the receptor extracellular domain. However, it has been proved that only homozygous C104R individuals are unable to bind APRIL (Salzer et al. 2005), whereas in

heterozygotes, a mutant-wild-type complex binds the ligand, but is not able to induce intracellular signaling (Garibyan et al. 2007). A similar impact can be also supposed for the never described before C104Y substitution, which was found in both healthy and CVID and IgAD individuals, only in a heterozygous state.

Despite these considerations, effects of heterozygous C104R and A181E substitutions on B cells function in individuals with CVID are still under debate, as well as whether their presence in healthy people means that they have incomplete penetrance or simply that a delayed CVID onset in controls used for association studies is possible. As regards this issue, blood donors without any evident immunological manifestations and with an age of over than 55 years, were collected to be used as control group in this study, since a mean age of about 25 years at the onset of CVID symptoms has been estimated (Cunningham-Rundles and Bodian 1999). In this way, possibility that healthy subjects carrying C104R/C104Y replacement might subsequently develop the disease has been strongly reduced, suggesting that incomplete penetrance or presence of additional genetic–environmental factors are responsible for the disease manifestation.

Although this advises that CVID does not fit a monogenic disease model, extremely low frequencies of *TNFRSF13B* and other supposed CVID-causing variants reveal at the same time that it does not fit neither the Common Disease/Common Variant paradigm, for which alleles contributing risk for a common complex disease are supposed to be frequent in the general population (Reich and Lander 2001).

However, there exist examples of rare variants influencing common diseases (Romeo et al. 2007), reinforcing the idea that both mutation typologies may play a role in such pathological manifestations, although to date it is not known which of them is more important. Even for disorders in which common nucleotide changes have been found, most genetic variation is indeed still uncovered and it is not possible to rule out the possibility that much genetic variation is due to rare variants. Therefore it is not yet known whether observed common disease-associated mutations represent only the tip of an undiscovered iceberg (Iles 2008).

Consequently, if rare variants were the primary cause of common complex disease, so that for a disease to be common there would be many different causative alleles, traditional genetic association studies would have little power to detect them and an evolutionary approach can offer a new and useful perspective to face the matter.

In accordance to this view, comparison of *TNFRSF13B* genetic diversity among samples with different ancestry turned out to be significant to understand its potential role in different CVID susceptibilities of different populations, as it is known that such a disease is essentially more common in individuals of European ancestry (Eades-Perner at al. 2007).

Low values of *TNFRSF13B* genetic variability were found for Sub-Saharan African, Central Asian, South American and Val di Scalve populations. Excluding Sub-Saharan Africans results, observed values may be due to small sample sizes, especially for Central Asians, and, most importantly, to the strong action that genetic drift has historically exerted on small isolated South American and Val di Scalve populations.

A completely different explanation can be instead invoked for values obtained for the Sub-Saharan African sample. Interestingly, an unusual remarkable scarceness of variability was observed for this group of populations, which displayed one of the lowest values of nucleotide diversity, as well as a lower value of haplotype diversity, respect to North African and Middle Eastern samples, in contrast with the trend observed for several other genes (Tishkoff et al. 1996; Calafell et al. 1998; Tishkoff et al. 1998; Guthery et al. 2007). Usually, a lower level of nucleotide diversity is found in non-African populations, as exemplified by Seattle SNPs data, and, at the same time, haplotype diversity is found to decrease as distance from Africa increases (The International HapMap Consortium 2005). It has also been observed that African groups exhibit lower levels of LD, concomitant with higher population recombination rates (ρ), compared to Europeans and Asians (Sawyer et al. 2005), suggesting that patterns of genetic variation observed in present-day populations are not the mere result of the action of mutation, recombination and natural selection, but are also strongly affected by demographic histories of populations. Shorter African LD blocks are indeed a consequence of larger effective ancestral population size (N_e) and of the fact that there has been more time for recombination to disrupt LD (Tishkoff and Verrelli 2003b). On the contrary, the greater LD in non-Africans is the result of founding events experienced by the groups of modern humans that migrated out of Africa starting from 100,000-50,000 years ago (Stoneking 2008). Therefore, *H. sapiens* migration from Africa to Eurasia and the rest of the world is thought to be accompanied by a strong population bottleneck that produced an inevitable loss of genetic diversity (Liu et al. 2006).

That being so, the higher *TNFRSF13B* variability observed in North-African, Middle Eastern and European samples respect to Sub-Saharan African one may be principally due to the presence of recent variants, which have arisen in ancestors of such populations during the early dispersal of anatomically modern humans out of Africa. In ancestral Sub-Saharan African populations an excessive number of *TNFRSF13B* changes should be not the optimum for survival, since the huge amount of bacterial pathogens encountered during childhood. As a matter of fact, human immune system remains quite immature during the first two years of life, being mainly dependent on its innate component. Thus, if *TNFRSF13B* was also involved in innate immunity, as recently proposed (Mackay and Schneider 2008), its defects would doubly affect the health of carrier

individuals, altering both innate responses and humoral immunity maturation, being rapidly removed from the population. On the contrary, the presence of endemic tuberculosis since in ancestral northern African populations might has allowed the spread of *TNFRSF13B* defects, as it seems that a weak wastefulness in B cell function can boost a stronger inflammatory response, which results just essential against mycobacteria.

This may be one of the reasons why new *TNFRSF13B* changes started to be maintained in first modern human groups that left Sub-Saharan Africa and were subsequently spread in Eurasian populations. In this way, *TNFRSF13B* changes in modern human populations, especially those of European ancestry, in which health care and hygiene conditions have been strongly improved, are tolerated in early childhood, perhaps with IgAD manifestations, and might subsequently lead to an increased susceptibility to CVID in adulthood.

The hypothesis of the presence of recent non-Sub-Saharan African private mutations is also confirmed by time estimates for haplotypes carrying coding variants found in CVID and IgAD individuals. Although unambiguous estimates were hardly achievable, since recombination undoubtedly played a role in shaping the genealogy of *TNFRSF13B* sequences, some glimpses about a rough dating can be indeed obtained. Moreover, 61,283 ± 34,577 years required for C104R, 358_359delA and A181E to accumulate from haplotype h3 and 52,483 ± 25,711 years required for 204insA and C193X to accumulate from haplotype h1, are overestimates, since they are based on a biased sample, made up of CVID and IgAD affected individuals, rather than on a random population sample. Therefore, actual haplotype expansion times are bound to be even more recent, perfectly falling into the 60-40 thousands years ago range that saw the first colonization of Eurasia by *H. sapiens*.

Results of human/chimpanzee nucleotide divergence estimate show that, both within our species and when compared to our closest ancestor, *TNFRSF13B* seems to be evolving at a slightly slower, though not unusual, rate. This suggests that a striking differentiation among worldwide human populations can not be expected for this genomic region, unless recent population specific selective pressures have affected it.

Observed haplotype structure at the *TNFRSF13B* locus indeed emphasizes that its five most frequent haplotypes accounted for 68% of sampled chromosomes, being found in almost all groups and carrying the more ancient and globally widespread variants, while remaining haplotypes are less widespread, and in some cases population specific, but rare.

This nearly homogeneous genetic background of worldwide populations, as regards *TNFRSF13B*, was also confirmed by computation of F_{st} genetic distances among them and by Analysis of Molecular Variance (AMOVA) results. They revealed the absence of a sharp geographical structure

for *TNFRSF13B* genetic diversity, with very low levels of differentiation among large geographically-based groups of populations, among populations within such groups and simply among individual populations. Moreover, as already inferred from nucleotide and haplotype diversity values, both F_{st} and AMOVA analyses pointed out a peculiar pattern of variability for the Val di Scalve population, suggesting again its outstanding genetic isolation.

The absence of recent population specific selective pressures was further and more reliably verified by searching for genetic footprints of selection at the *TNFRSF13B* locus by means of specific neutrality tests, such as Tajima's *D*, Fu and Li's *D*, *F* and Fay and Wu's *H*, since this matter could represent a potential key to understand the actual role of this gene in immune functions.

Looking back to the evolutionary history of such a young and cosmopolitan species as *H. sapiens*, it is coherent to believe that our pathogens, which live in the extremely diversified environments colonized by modern humans, may have represented one of the major selective pressures on our genome. Thereby, genes whose function is strictly related to the immune system are supposed to be more likely subjected to the action of natural selection respect to other typology of genes (Sabeti et al. 2006). In particular, genes involved in adaptive immunity, such as *TNFRSF13B*, are likely to be subjected to geographically localized selective pressures, as they may interact with local pathogen landscapes, resulting in increased inter-population genetic differentiation.

However, as already suggested by the absence of such a remarkable inter-population differentiation, also neutrality tests showed not significant results for examined groups, leading to the acceptance of a neutral model of evolution for the analyzed genomic region. Excluding CVID and IgAD individuals from the analysis, significant values for total sample Fu and Li's *D* and *F* statistics only were obtained, so that they seem to be more plausibly due to demographic expansion of anatomically modern humans rather than to the action of selection, as seen in most of human genes (Stephens et al. 2001).

That being so, genetic drift and gene flow only might have driven *TNFRSF13B* evolution, unless it might be very anciently shaped by selective pressures which have predated the exit of anatomically modern humans from Sub-Saharan Africa, resulting homogeneous on early *H. sapiens* populations. This latter hypothesis might be consistent with several recent remarks which are more and more emphasizing relationships between innate and adaptive components of the immune system (Pancer and Cooper 2006; Groom et al. 2007; Katsenelson et al. 2007) and with what already discussed about the potential involvement of *TNFRSF13B* in innate immunity. Such a typology of immune responses represents a generic and not pathogen specific defense, so that related genes may be subjected to similar selective forces also in different populations (Ferrer-Admetlla et al. 2008). Therefore, it can not be unlikely that equal, but too ancient to be recognized, selective pressures

may have been responsible for low *TNFRSF13B* genetic diversity of ancestral Sub-Saharan Africans and, since its slow rate of evolution, also for present-day Sub-Saharan Africans low diversity.

Moreover, neutrality tests results for the Val di Scalve sample were again the most peculiar, with a large, even if not significant, positive Tajima's *D* value that suggests a condition close to a scarceness of segregating sites and an excess of intermediate frequency alleles. This pattern is generally consistent with population substructure, bottlenecks or even balancing selection, although, in this case, it could be simply interpreted as a consequence of a founder effect, followed by strong genetic drift.

Together with described population-based evolutionary analyses, a direct comparison of Italian CVID, IgAD and control samples was performed applying the same methods.

Haplotype structure in such groups was found to be characterized by a very similar haplotype frequency distribution with respect to that observed in worldwide populations and, in particular, to that of Italian and Middle Eastern samples. More in details, taking into account percentages of cosmopolitan and rare haplotypes, an extremely subtle difference is noticed between IgAD individuals and healthy controls, whereas just a bit greater difference is found when CVID subjects are examined. A statistical support for these cases-controls differences is additionally given by relative AMOVA results, which showed a barely significant F_{st} value for CVID-control comparison only.

Neutrality tests also turned out to be very useful for further characterizing and distinguishing CVID, IgAD and control subjects, again emphasizing a difference between groups of individuals affected by the two studied diseases.

As already observed for worldwide populations, the absence of significant departures from the null hypothesis of neutral evolution was also verified for IgAD and healthy Italian groups. Significant values only for CVID sample Fu and Li's *D* and *F* large negative statistics were obtained, reflecting a substantial excess of singletons and low frequency variants in such a group. These features, which are not confirmed by Tajima's *D* and Fay and Wu's *H* statistics, are unlikely due to positive selection acting on CVID subjects, resulting more plausibly consistent with an actual involvement of some of these changes in the disease. On the contrary, this seems not true for variants of IgAD individuals, for which smaller negative and not significant values were observed.

In conclusion, evolutionary analyses performed on *TNFRSF13B* coding region have demonstrated that populations in which CVID is rare are characterized by a low variability of the examined genomic region, and that, at the same time, individuals affected by CVID carry a little, but

significant, excess of rare derived alleles, respect to healthy individuals belonging to the same population. This leads to the conclusion that some of these *TNFRSF13B* changes may actually contribute to the development of the disease.

However, the extent of such disease/healthy samples difference and the fact that geographical distribution of this gene diversity is more plausibly related to its potential involvement in innate immunity rather than to its involvement in adaptive immunity, suggest that CVID might be more likely related to still unknown environmental and genetic factors, rather than to the nature of *TNFRSF13B* variants only.

That being so, for populations in which health care and hygiene conditions have been strongly improved, and especially for those of European ancestry, it seems that changes in *TNFRSF13B* coding region can be tolerated in early childhood, perhaps with IgAD manifestations, but might subsequently lead to an increased susceptibility to CVID in adulthood, acting as genetic risk factors rather than causative mutations.

6. References

Abecasis, G. R., D. Ghosh, and T. E. Nichols. 2005. Linkage disequilibrium: ancient history drives the new genetics. *Hum. Hered.* 59: 118–124.

Aghamohammadi, A., J. Mohammadi, N. Parvaneh, N. Rezaei, M. Moin, T. Espanol, and L. Hammarstrom. 2008. Progression of selective IgA deficiency to common variable immunodeficiency. *Int. Arch. Allergy Immunol*. 147(2): 87–92.

Aiuti, A., S. Vai, A. Mortellaro, *et al*. 2002. Immune reconstitution in ADA–SCID after PBL gene therapy and discontinuation of enzyme replacement. *Nat. Med.* 8(5): 423–425.

Akey, J. M., K. Zhang, M. Xiong, and L. Jin. 2003. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* 20: 232–242.

Aldea, A., F. Calafell, J. L. Aróstegui, O. Lao, J. Rius, S. Plaza, *et al*. 2004. The west side story: MEFV haplotype in Spanish FMF patients and controls, and evidence of high LD and a recombination "hot–spot" at the MEFV locus. *Hum. Mutat.* 23(4): 399.

Altshuler, D., and M. Daly. 2007. Guilt beyond a reasonable doubt. *Nat. Genet.* 39: 813–814.

Antoine, C., S. Muller, A. Cant, *et al*. 2003. Long–term survival and transplantation of haemopoietic stem cells for immunodeficiencies: report of the European experience 1968–99. *Lancet.* 361(9357): 553–560.

Antonarakis, S. E., and J. S. Beckmann. 2006. Mendelian disorders deserve more attention. *Nat. Rev. Genet*. 7: 277–282.

Bacanu, S. A., B. Devlin, and K. Roeder. 2002. Association studies for quantitative traits in structured populations. *Genet. Epidemiol.* 22: 78–93.

Bailey, J. A. and E. E. Eichler. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev Genet*. 7: 552–564.

Bamshad, M., and S. P. Wooding. 2003. Signatures of natural selection in the human genome. *Nat. Genet.* 4(2): 99–111.

Bandelt, H. J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141: 743–753.

Barbujani, G., A. Magagni, E. Minch, and L. L. Cavalli–Sforza. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* 94(9): 4516–4519.

Bayry, J., O. Hermine, D. A. Webster, Y. Levy, and S. V. Kaveri. 2005. Common variable immunodeficiency: the immune system in chaos. *Trends Mol. Med.* 11: 370–376.

Beck, G., and G. S. Habicht. 1996. Immunity and the Invertebrates. *Scientific American* 60–66.

Bentley, D. R. 2006. Whole–genome resequencing. *Curr. Opin. Genet. Dev.* 16: 545–552.

Bossen, C., and P. Schneider. 2006. BAFF, APRIL and their receptors: structure, function and signaling. *Semin. Immunol.* 18: 263–275.

Bossen, C., T. G. Cachero, A. Tardivel, *et al*. 2008. TACI, unlike BAFF–R, is solely activated by oligomeric BAFF and APRIL to support survival of activated B cells and plasmablasts. *Blood*. 111: 1004–1012.

Byng, M. C., J. C. Whittaker, A. P. Cuthbert, *et al*. 2003. SNP subset selection for genetic association studies. *Ann. Hum. Genet.* 67: 543–556.

Calafell, F., A. Shuster, W. C. Speed, J. R. Kidd, and K. K. Kidd. 1998. Short tandem repeat polymorphism evolution in humans. *Eur. J. Hum. Genet.* 6(1): 38–49.

Calafell, F., F. Roubinet, A. Ramírez–Soriano, N. Saitou, J. Bertranpetit, and A. Blancher. 2008. Evolutionary dynamics of the human ABO gene. *Hum. Genet.* 124(2): 123–135.

Cambien, F., and L. Tiret. 2007. Genetics of cardiovascular diseases: From single mutations to the whole genome. *Circulation.* 116: 1714–1724.

Carlson, C. S., M. A. Eberle, M. J. Rieder, *et al*. 2004. Selecting a maximally informative set of single–nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* 74: 106–120.

Carvalho Neves Forte, W., F. Ferreira De Carvalho Junior, N. Damaceno, F. Vidal Perez, C. Gonzales Lopes, and R. A. Mastroti. 2000. Evolution of IgA deficiency to IgG subclass deficiency and common variable immunodeficiency. *Allergol. Immunopathol.* 28: 18–20.

Casanova, J. L., and L. Abel. 2005. Inborn errors of immunity to infection: the rule rather than the exception. *J. Exp. Med.* 202: 197–201.

Casanova, J. L., and L. Abel. 2007. Primary immunodeficiencies: a field in its infancy. *Science.* 317: 617–619.

Casrouge, A., S.Y. Zhang, C. Eidenschenk, E. Jouanguy, A. Puel, K. Yang, *et al*. 2006. Herpes simplex virus encephalitis in human UNC–93B defi–ciency. *Science*. 314: 308–312.

Castigli, E., S. A. Wilson, L. Garibyan, R. Rachid, F. Bonilla, L. Schneider, and R. S. Geha. 2005a. TACI is mutant in common variable immunodeficiency and IgA deficiency. *Nat. Genet.* 37: 829–834.

Castigli, E., S. A. Wilson, S. Scott, *et al*. 2005b. TACI and BAFF–R mediate isotype switching in B cells. *J. Exp. Med.* 201: 35–39.

Castigli, E., and R. S. Geha. 2006. Molecular basis of common variable immunodeficiency. *J. Allergy Clin. Immunol.* 117: 740–746.

Castigli, E., S. A. Wilson, L. Garibyan, R. Rachid, F. Bonilla, L. Schneider, *et al.* 2007. Reexamining the role of TACI coding variants in common variable immunodeficiency and selective IgA deficiency. *Nat. Genet.* 39(4): 430–431.

Chen, F. C., and W. H. Li. 2001. Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. *Am. J. Hum. Genet.* 68: 444–456.

Clark, A. G., R. Nielsen, J. Signorovitch, T. C. Matise, S. Glanowski, *et al*. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single–nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.* 73: 285–300.

Conrad, D. E., M. Jakobsson, G. Coop, *et al*. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* 38: 1251–1260.

Cooper, G. M., T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson. 2008. Systematic assessment of copy number variant detection via genome–wide SNP genotyping. *Nat Genet*. 40(10): 1199–1203.

Crawford, D. C., D. T. Akey, and D. A. Nickerson. 2005. The patterns of natural variation in human genes. *Ann. Rev. Genomics Hum. Genet*. 6: 287–312.

Cunningham–Rundles, C., and C. Bodian. 1999. Common variable immunodeficiency: clinical and immunological features of 248 patients. *Clin. Immunol.* 92: 34–48.

Dahlman, I., I. A. Eaves, R. Kosoy, *et al*. 2002. Parameters for reliable results in genetic association studies in common disease. *Nat. Genet.* 30: 149–150.

Dean, M., M. Carrington, S. J. O'Brien. 2002. Balanced polymorphism selected by genetic versus infectious human disease. *Annu. Rev. Genomics Hum. Genet*. 3: 263–292.

deBakker, P. I., N. P. Burtt, R. R. Graham, *et al*. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* 38: 1298–1303.

Dobzhansky, T. 1973. Nothing in Biology Makes Sense Except in the Light of Evolution. *The American Biology Teacher*. 35: 125–129.

Eades–Perner, A. M., B. Gathmann, V. Knerr, *et al*. 2007. The European internet–based patient and research database for primary immunodeficiencies: results 2004–06. *Clin. Exp. Immunol.* 147: 306–312.

Easton, D. F., K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson, *et al*. 2007. Genome–wide association study identifies novel breast cancer susceptibility loci. *Nature*. 447: 1087–1093.

Ehl, S., K. Schwarz, A. Enders, U. Duffner, U. Pannicke, J. Kuhr, *et al*. 2005. A variant of SCID with specific immune responses and predominance of gamma delta T cells. *J. Clin. Invest.* 115: 3140–3148.

Enard, W., M. Przeworski, S. E. Fisher, C. S. L. Lai, V. Wiebe, *et al*. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature.* 418: 869– 872.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479–491.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1: 47–50.

Fay, J. C., and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.

Ferrer–Admetlla, A., E. Bosch, M. Sikora, T. Marquès–Bonet, A. Ramírez–Soriano, A. Muntasell, *et al*. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J. Immunol.* 181(2): 1315–1322.

Feuk, L., A. R. Carson, and S. W. Schreier. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* 7: 85–97.

Fischer, A. 2007. Human primary immunodeficiency diseases. *Immunity*. 27: 835–845.

Fleischmann, R.D, M. D. Adam, O. White, R. A. Clayton, E. F. Kirkness, *et al.* 1995. Whole–genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* 269: 496–512.

Fu, Y. X., and W. H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, *et al.* 2002. The structure of haplotype blocks in the human genome. *Science*. 296(5576): 2225–2229.

Garibyan, L., A. A. Lobito, R. M. Siegel, M. E. Call, K. W. Wucherpfennig, and R. S. Geha. 2007. Dominant–negative effect of the heterozygous C104R TACI mutation in common variable immunodeficiency (CVID). *J. Clin. Invest*. 117: 1550– 1557.

Garrigan, D., and M. F. Hammer. 2006. Reconstructing human origins in the genome era. *Nat. Rev. Genet*. 7: 669–680.

Garrigan, D., S. B. Kingan, M. M. Pilkington, J. A. Wilder, M. P. Cox, *et al*. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics.* 177: 2195–2207.

Geha, R. S., L. D. Notarangelo, and J. L. Casanova. 2007. Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *J. Allergy Clin. Immunol*. 120: 776–794.

Goldstein, D. B., and G. L. Cavalleri. 2005. Understanding human diversity. *Nature*. 437: 1241–1242.

Graffelman, J., D. J. Balding, A. Gonzalez–Neira, and J. Bertranpetit. 2007. Variation in estimated recombination rates across human populations. *Hum. Genet.* 122(3–4): 301–310.

Green, R. E., J. Krause, S. E. Ptak, *et al.* 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature*. 444: 330–336.

Greenwald, R. J., G. J. Freeman, and A. H. Sharpe. 2005. The B7 family revisited. *Annu. Rev. Immunol.* 23: 515–548.

Grimbacher, B., A. Hutloff, M. Schlesier *et al*. 2003. Homozygous loss of ICOS is associated with adult–onset common variable immunodeficiency. *Nat. Immunol.* 4: 261–268.

Groisman, E. A., and S. D. Ehrlich. 2003. Genomics. A global view of gene gain, loss, regulation and function. *Curr. Opin. Microbiol.* 6: 479–481.

Groom J. R., C. A. Fletcher, S. N. Walters, S. T. Grey, S. V. Watt, M. J. Sweet, *et al*. 2007. BAFF and MyD88 signals promote a lupuslike disease independent of T cells. *J. Exp. Med.* 204(8): 1959–1971.

Guthery, S. L., B. A. Salisbury, M. S. Pungliya, J. C. Stephens, and M. Bamshad. 2007. The structure of common genetic variation in United States populations. *Am. J. Hum. Genet.* 81: 1221–1231.

Hamblin, M. T., E. E. Thompson, A. DiRienzo. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* 70: 369–383.

Hammarstrom, L., I. Vorechovsky, and D. Webster. 2000. Selective IgA deficiency (SIgAD) and common variable immunodeficiency (CVID). *Clin. Exp. Immunol.* 120: 225–231.

Harding, R. M., E. Healy, A. J. Ray, N. S. Ellis, N. Flanagan, *et al*. 2000. Evidente for variable selective pressures at MC1R. *Am. J. Hum. Genet.* 66: 1351–1361.

Harris, R. A., J. Rogers, and A. Milosavljevic. 2007. Human specific changes of genome structure detected by genomic triangulation. *Science*. 316: 235–237.

Hinds, D. A., L. L. Stuve, G. B. Nilsen *et al*. 2005. Wholegenome patterns of common DNA variation in three human populations. *Science* 307: 1072–1079.

Hirschhorn, R. 2003. In vivo reversion to normal of inherited mutations in humans. *J Med Genet*. 40(10): 721–728.

Hoggart, C., E. Parra, M. Shriver, *et al*. 2003. Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* 72: 1492–1504.

Hunter, D. J., and P. Kraft. 2007. Drinking from the fire hose–statistical issues in genomewide association studies. *N. Engl. J. Med.* 357: 436–39.

Hurst, L. D., C. Pa´ l, and M. J. Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet*. 5: 299–310.

Hymowitz, S. G., D. R. Patel, H. J. A. Wallweber, S. Runyon, M. Yan, J. Yin, *et al*. 2005. Structures of APRIL–receptor complexes: like BCMA, TACI employs only a single cysteine–rich domain for high affinity ligand binding. *J. Biol. Chem.* 280: 7218–7227.

Iles, M. M. 2008. What can genome–wide association studies tell us about the genetics of common disease?. *PLoS Genet*. 4(2): e33.

Inoue, K., Y. Mineharu, S. Inoue, S. Yamada, F. Matsuda, K. Nozaki, *et al*. 2006. Search on chromosome 17 centromere reveals TNFRSF13B as a susceptibility gene for intracranial aneurysm: a preliminary study. *Circulation* 113(16): 2002–2010.

International HapMap Consortium. 2003. The International HapMap Project. *Nature.* 426: 789–794.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 449: 851–617.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–941.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945.

Jakkula, E., K. Rehnström, T. Varilo, O. P. Pietiläinen, T. Paunio, N. L. Pedersen *et al*. 2008. The genome–wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.* 83(6): 787–794.

Janeway, C. A., L. Apt, and D. Gitlin. 1953. Agammaglobulinemia. *Trans. Assoc. Am. Physicians.* 66: 200–202.

Katsenelson, N., S. Kanswal, M. Puig, H. Mostowski, D. Verthelyi, and M. Akkoyunlu. 2007. Synthetic CpG oligodeoxynucleotides augment BAFF– and APRIL–mediated immunoglobulin secretion. *Eur. J. Immunol.* 37(7): 1785–1795.

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat. Genet.* 41(1): 66–70.

Khaja, R., J. Zhang, and J. R. MacDonald. 2006. Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet*. 38: 1413–1418.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, London, UK.

Komura, D., F. Shen, S. Ishikawa, *et al*. 2006. Genome–wide detection of human copy number variations using highdensity DNA oligonucleotide arrays. *Genome Res*. 16: 1575–1584.

Krings, M., A. Stone, R. W. Schmitz, *et al*. 1997. Neanderthal DNA sequences and the origin of modern humans. *Cell*. 90: 19–30.

Kruskal, J. 1964. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*. 29: 28–42.

Ku, C. L., H. von Bernuth, C. Picard, S. Y. Zhang, H. H. Chang, K. Yang, *et al*. 2007. Selective predisposition to bacterial infections in IRAK–4–deficient children: IRAK–4–dependent TLRs are otherwise redundant in protective immunity. *J. Exp. Med.* 204: 2407–2422.

Lander, E. S. 1996. The new genomics: Global views of biology. *Science*. 274: 536–539.

Lao, O., J. M. de Gruijter, K. van Duijn, A. Navarro, and M. Kaiser. 2007. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann. Hum. Genet.*71: 354–369.

Lee, W. I., J. L. Huang, M. L. Kuo, S. J. Lin, L. C. Chen, M. T. Chen, and T. H. Jaing. 2007. Analysis of genetic defects in patients with the common variable immunodeficiency phenotype in a single Taiwanese tertiary care hospital. *Ann. Allergy. Asthma. Immunol.* 99(5): 433–442.

Lee, J. J., E. Ozcan, I. Rauter, and R. S. Geha. 2008. Transmembrane activator and calcium–modulator and cyclophilin ligand interactor mutations in common variable immunodeficiency. *Curr. Opin. Allergy Clin. Immunol*. 8(6): 520–526.

Little, P. 1992. Human Genome Project: Mapping the way ahead. *Nature.* 359: 367–368.

Liu, H., F. Prugnolle, A. Manica, and F. Balloux. 2006. A geographically explicit genetic model of worldwide human–settlement history. *Am. J. Hum. Genet.* 79: 230–237.

Lohmueller, K. E., C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschorn. 2003. Meta analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33: 177–182.

Macaulay, V., C. Hill, A. Achilli, C. Rengo, D. Clarke, *et al*. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science.* 308: 1034–1036.

Mackay, F., and C. Ambrose. 2003. The TNF family members BAFF and APRIL: the growing complexity. *Cytokine Growth Factor Rev.* 14: 311–324.

Mackay, F., P. Schneider, P. Rennert, and J. Browning. 2003. BAFF AND APRIL: a tutorial on B cell survival. *Annu. Rev. Immunol,* 21: 231–264.

Mackay, F., and P. Schneider. 2008. TACI, an enigmatic BAFF/APRIL receptor, with new unappreciated biochemical and biological properties. *Cytokine Growth Factor Rev*. 19 (3–4): 263–276.

Manolio, T. A., J. D. Brooks, and F. S. Collins. 2008. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* 118: 1590–1605.

Mardis, E. R. 2008. Next–Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet.* 9: 387–402.
Mateu, E., F. Calafell, O. Lao, B. Bonné–Tamir, J. R. Kidd, A. Pakstis, *et al*. 2001. Worldwide genetic analysis of the CFTR region. *Am. J. Hum. Genet.* 68(1): 103–117.

Matzinger, P. 2002. The danger model: a renewed sense of self. *Science.* 296(5566): 301–305.

McPherson, R., A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, *et al*. 2007. A common allele on Chromosome 9 associated with coronary heart disease. *Science*. 316: 1488–1491.

Medzhitov, R. 2007. Recognition of microorganisms and activation of the immune response. *Nature.* 449(7164): 819–826.

Miller, S. A., D. D. Dykes, and H. F. Polesky. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16(3): 1215.

Moreno–Estrada, A., F. Casals, A. Ramírez–Soriano, B. Oliva, F. Calafell, J. Bertranpetit, and E. Bosch. 2008. Signatures of selection in the human olfactory receptor OR5I1 gene. *Mol. Biol. Evol.* 25(1): 144–154.

Mullis, K., F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* 51: 263–273.

Myers, S., L. Bottolo, C. Freeman, G. Mc Vean, and P. Donnely. 2005. A fine–scale map of recombination rates and hotspots across the human genome. *Science* 310: 321–324.

Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*. 40(9): 1124–1129.

Navarro, A., and N. H. Barton. 2002. The effects of multilocus balancing selection on neutral variability. *Genetics.* 161: 849–863.

NCI–NHGRI working group on replication in association studies. 2007. Replicating genotype–phenotype associations. *Nature*. 447: 655–660.

Neel, J. V. 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?. *Am. J. Hum. Genet.* 14: 353–362.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol*. 3(5): 418–426.

Nei, M. 1987. Molecular Evolutionary Genetics. Eds. Columbia University Press, New York, NY, USA. 180.

Ng, L. G., A. P. Sutherland, R. Newton, F. Qian, T. G. Cachero, M. L. Scott, *et al*. 2004. B cell–activating factor belonging to the TNF family (BAFF)–R is the principal BAFF receptor facilitating BAFF costimulation of circulating T and B cells. *J. Immunol*. 173(2): 807–817.

Noonan, J. P., G. Coop, S. Kudaravelli, *et al*. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science*. 314: 1113–1118.

North, M. E., A. D. Webster, and J. Farrant. 1998. Primary defect in CD8+ lymphocytes in the antibody deficiency disease (common variable immunodeficiency): abnormalities in intracellular production of interferon–gamma (IFN–gamma) in CD28+ (‘cytotoxic’) and CD28– (‘suppressor’) CD8+ subsets. *Clin. Exp. Immunol.* 111: 70–75.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics.* 154: 931–942.

Ochs, H., C. I. E. Smith, and J. Puck. 2007. Primary Immunodeficiencies: A Molecular and Genetic Approach. Oxford University Press, New York, NY, USA.

Orlando, L., P. Darlu, M. Toussain, *et al*. 2006. Revisiting Neanderthal diversity with a 100,000 year old mtDNA sequence. *Curr. Biol*. 16: R400–R402.

Paabo, S. 2003. The mosaic that is our genome. *Nature*. 421: 409–412.

Pan–Hammarström, Q., U. Salzer, L. Du, J. Björkander, C. Cunningham–Rundles, D. L. Nelson, *et al*. 2007. Reexamining the role of TACI coding variants in common variable immunodeficiency and selective IgA deficiency. *Nat. Genet.* 39(4): 429–430.

Pancer, Z., and M. Cooper. 2006. The evolution of adaptive immunity. *Annu. Rev. Immunol.* 24: 497–518.

Parida, L., M. Melé, F. Calafell, J. Bertranpetit and The Genographic Consortium. 2008. Estimating the Ancestral Recombinations Graph (ARG) as Compatible Networks of SNP Patterns. *J. Comput. Biol.* 15(9): 1133–1154.

Park, M. A., J. T. Li, J. B. Hagan, D. E. Maddox, and R. S. Abraham. 2008. Common variable immunodeficiency: a new look at an old disease. *Lancet*. 372(9637): 489–502.

Pastinen, T., B. Ge, and T. J. Hudson. 2006. Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet*. 15: R9–R16.

Payseur, B. A., A. D. Cutter, and M. W. Nachman. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* 19: 1143–1153.

Ramírez–Soriano, A., S. E. Ramos–Onsins, J. Rozas, F. Calafell, and A. Navarro. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* 179(1): 555–567.

Ramos–Onsis, S. E., and J. Rosaz. 2002. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* 19(12): 2092–2100.

Ramoz, N., L. A. Rueda, B. Bouadjar, L. S. Montoya, G. Orth, and M. Favre. 2002. Mutations in two adjacent novel genes are associated with epidermodysplasia verruciformis. *Nat. Genet*. 32: 579–581.

Redon, R., S. Ishikawa, K. R. Fitch, *et al*. 2006. Global variation in copy number in the human genome. *Nature*. 444: 444–454.

Reed, F. A., and S. A. Tishkoff. 2006. African human diversity, origins and migrations. *Curr. Opin. Genet. Dev.* 16: 597–605.

Reich, D. E. and E. S. Lander. 2001. On the allelic spectrum of human disease. *Trends Genet.* 17: 502–510.

Reich, D. E., M.Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, *et al*. 2001. Linkage disequilibrium in the human genome. *Nature*. 411: 199–204.

Reich, D. E., S. F. Schaffner, M. J. Daly, *et al*. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet*. 32: 135–142.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science.* 316: 222–234.

Rigaud, S., M. C. Fondaneche, N. Lambert, B. Pasquier, V. Mateo, P. Soulas, *et al*. 2006. XIAP deficiency in humans causes an X–linked lymphoproliferative syndrome. *Nature*. 444: 110–114.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science*. 273: 1516–1517.

Romeo, S., L. A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg–Hansen, *et al*. 2007. Population–based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* 39: 513–516.

Romualdi, C., D. Balding, I. S. Nasidze, G. Risch, M. Robichaux, S. T. Sherry, M. Stoneking, M. A. Batzer, and G. Barbujani. 2002. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res.* 12(4): 602–612.

Rozas, J., J. C. Sancez–DelBarrio, X. Messeguer, and R. Rosaz. 2003. DnaSp, DNA polymorphism analyses by the coalescent and others methods. *Bioinformatics* 19: 2496–2497.

Sabater–Lleal, M., J. M. Soria, J. Bertranpetit, L. Almasy, J. Blangero, J. Fontcuberta, and F. Calafell. 2006. Human F7 sequence is split into three deep clades that are related to FVII plasma levels. *Hum. Genet.* 118(6): 741–751.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, *et al*. 2006. Positive natural selection in the human lineage. *Science.* 312(5780): 1614–1620.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, *et al*. 2007. Genome–wide detection and characterization of positive selection in human populations. *Nature*. 449(7164): 913–918.

Saillard, J., P. Forster, N. Lynnerup, H. J. Bandelt, and S. Nørby. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet*. 67: 718–726.

Salonen, J. T., P. Uimari, J. M. Aalto, M. Pirskanen, J. Kaikkonen, *et al*. 2007. Type 2 diabetes whole–genome association study in four populations: The DiaGen consortium. *Am. J. Hum. Genet.* 81: 338–345.

Salzer, U., A. Maul–Pavicic, C. Cunningham–Rundles, *et al*. 2004. ICOS deficiency in patients with common variable immunodeficiency. *Clin. Immunol.* 113: 234–240.

Salzer, U., H. M. Chapel, A. D. Webster, Q. Hammarstrom, A. Schmitt–Graeff, M. Schlesier, *et al.* 2005. Mutations in TNFRSF13B encoding TACI are associated with common variable immunodeficiency in humans. *Nat. Genet.* 37: 820–828.

Salzer, U., and B. Grimbacher. 2006. Common variable immunodeficiency: The power of co–stimulation. *Semin. Immunol.* 18(6): 337–346.

Salzer, U., C. Bacchelli, S. Buckridge, Q. Pan–Hammarstrom, S. Jennings, V. Lougaris, *et al*. 2008. Relevance of biallelic versus monoallelic TNFRSF13B mutations in distinguishing disease–causing from risk–increasing TNFRSF13B variants in antibody deficiency syndromes. *Blood*. Epub ahead of print.

Samani, N. J., J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, *et al*. 2007. Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* 357: 443–453.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain–terminating inhibitors. *Proc. Natl. Acad. Sci. USA*. 74(12): 5463–5467.

Sawyer, S. L., N. Mukherjee, A. J. Pakstis, I. Feuk, J. R. Kidd, *et al*. 2005. Linkage disequilibrium patterns vary substantially among populations. *Eur. J. Hum. Genet.* 13: 677–686.

Saxena, R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, *et al*. 2007. Genome–wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science.* 316: 1331–1336.

Schroeder, H. W. Jr, H. W. 3rd Schroeder, and S. M. Sheikh. 2004. The complex genetics of common variable immunodeficiency. *J. Investig. Med.* 52: 90–103.

Serre, D., A. Langaney, M. Chech, *et al*. 2004. No evidente of Neanderthal mtDNA contribution to modern humans. *PLos Biol.* 2: 313–317.

Sham, P. C., S. S. Cherny, S. Purcell, *et al*. 2000. Power of linkage versus association analysis of quantitative traits, by use of variance–components models, for sibship data. *Am. J. Hum. Genet.* 66: 1616–1630.

Shapiro, J. A. 2005. A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering. *Gene*. 345: 91–100.

Sharp, A. J., Z. Cheng, and E. E. Eichler. 2006. Structural variation in the human genome. *Annu. Rev. Genomics Hum. Genet*. 7: 407–442.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462.

Smith, D. J., and A. J. Lusis. 2002. The allelic structure of common disease. *Hum. Mol. Genet*. 11(20): 2455–2461.

Soldevila, M., A. M. Andrés, A. Ramírez–Soriano, T. Marquès–Bonet, F. Calafell, A. Navarro, and J. Bertranpetit. 2006. The prion protein gene in humans revisited: lessons from a worldwide resequencing study. *Genome Res.* 16(2): 231–239.

Stephens, J. C., D. E. Reich, D. B. Goldstein, H. D. Shin, M. W. Smith, *et al*. 1998. Dating the origin of the CCR5– Delta 32 AIDS–resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* 62: 1507–1515.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, *et al*. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*. 293: 489–493.

Stephens, M., N. J. Smith, and P. Donnely. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68: 978–989.

Stephens, M., and P. Donnelly. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet*. 73: 1162–1169.

Stephens, J. C., and P. Scheet. 2005. Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing–Data Imputation. *Am. J. Hum. Genet.* 76: 449–462.

Stoneking, M. Human origins. 2008. The molecular perspective. *EMBO Rep*. 9(1): S46–50.

Stranger, B. E., M. S. Forrest, A. G. Clark, *et al*. 2005. Genome–wide associations of gene expression variation in humans. *PLoS Genet*. 1: e78.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 105: 437–460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

Takahata, N., Y. Satta, and J. Klein. 1995. Divergence and population size in the lineage leading to modern humans. *Theoretical Population Biology* 48: 198–221.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437: 69–87.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature*. 437: 1299–1320.

Tian, C., R. Kosoy, A. Lee, M. Ransom, J. W. Belmont, P. K. Gregersen, and M. F. Seldin. 2008. Analysis of East Asia genetic substructure using genome–wide SNP arrays. *PLoS ONE*. 3(12): e3862.

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, *et al*. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271(5254): 1380–1387.

Tishkoff, S. A., A. Goldman, F. Calafell, W. C. Speed, A. S. Deinard, B. Bonne–Tamir, *et al*. 1998. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am. J. Hum. Genet.* 62(6): 1389–1402.

Tishkoff, S. A., and B. C. Verrelli. 2003a. G6PD deficiency and malarial resistance in humans: insights from evolutionary genetic analyses. In *Infectious Disease: Host–Pathogen Evolution*, ed. K. Dronamraju. Cambridge University Press, New York, NY, USA.

Tishkoff, S. A., and B. C. Verrelli. 2003b. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Curr. Opin. Genet. Dev.* 13: 569–575.

Trevathan, W. R. 2007. Evolutionary Medicine. *Annual Review of Anthropology*. 36: 139–154.

Tsuiji, M., S. Yurasov, K. Velinzon, S. Thomas, M. C. Nussenzweig, and H. Wardemann. 2006. A checkpoint for autoreactivity in human IgM+ memory B cell development. *J. Exp. Med.* 203: 393–400.

Tuzun, E., A.J. Sharp, J. A. Bailey, *et al*. 2005. Fine–scale structural variation in the human genome. *Nat. Genet*. 37: 727–732.

van Zelm, M. C., I. Reisli, M. van der Burg, *et al*. 2006. An antibodydeficiency syndrome due to mutations in the CD19 gene. *N. Engl. J. Med.* 354: 1901–1912.

Vogt, G., B. Vogt, N. Chuzhanova, K. Julenius, D.N Cooper, and J. L. Casanova. 2007. Gain–of–glycosylation mutations. *Curr. Opin. Genet. Dev*. 17: 245–251.

Wacholder, S., N. Rothman, and N. Caporaso. 2002. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev.* 11: 513–520.

Wakeley, J., R. Nielsen, S. N. Liu–Cordero, and K. Ardlie. 2001. The discovery of single–nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.* 69: 1332–1347.

Warnatz, K., U. Salzer, and S. Gutenberger. 2005. Finally found: human BAFF–R deficiency causes hypogammaglobulinemia. *Clin. Immunol.* 115(suppl 1)**:** 820.

Wellcome Trust Case Control Consortium and Australo–Anglo–American Spondylitis Consortium. 2007. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet*. 39(11): 1329–1337.

Weller, S., M. C. Braun, B. K. Tan, *et al*. 2004. Human blood IgM "memory" B cells are circulating splenic marginal zone B cells harboring a prediversified immunoglobulin repertoire. *Blood.* 104: 3647–3654.

Xia, X. Z., J. Treanor, G. Senaldi, *et al*. 2000. TACI is a TRAF–interacting receptor for TALL–1, a tumor necrosis factor family member involved in B cell regulation. *J. Exp. Med.* 92: 137–143.

Yan, M., H. Wang, B. Chan, *et al*. 2001. Activation and accumulation of B cells in TACI–deficient mice. *Nat. Immunol*. 2: 638–643.

Yin, E. Z., D. P. Frush, L. F. Donnelly, and R. H. Buckley. 2001. Primary Immunodeficiency Disorders in paediatric patients: Clinical features and imaging findings. *A. J. R.* 176: 1541–1551.

Zhang, L., L. Radigan, U. Salzer, *et al*. 2007. Transmembrane activator and calcium–modulating cyclophilin ligand interactor mutations in common variable immunodeficiency: clinical and immunologic outcomes in heterozygotes. *J. Allergy Clin. Immunol.* 120: 1178–1185.

Zhang, S. Y., E. Jouanguy, S. Ugolini, A. Smahi, G. Elain, P. Romero, *et al*. 2007. TLR3 deficiency in patients with herpes simplex encephalitis. *Science*. 317: 1522–1527.

Zhao, Z., N. Yu, Y. X. Fu, and W. H. Li. 2006. Nucleotide variation and haplotype diversity in a 10–kb noncoding region in three continental human populations. *Genetics.* 174: 399–409.

Zondervan, K. T., and L. R. Cardon. 2004. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet*. 5: 89–100.

Acknowledgments

Thanks to Prof. Donata Luiselli and Prof. Davide Pettener who gave me the chance to carry out this PhD program and disclosed to me the challenging field of modern Anthropology.
A grateful thanks also to Angela Ventrella, Antonella Useli, Alessio Boattini, Cristina Fabbri, Chiara Consiglio, Graziella Ciani, Daniele Yang Yao, Lisa Argnani, Loredana Castrì, Marco Milella, Paolo Garagnani and Stefania Zampetti who have played along with me during these three intense, but very pleasant, years in Anthropology laboratories.

Thanks to Prof. Giovanni Romeo and especially to Dr. Simona Ferrari who allowed me to design and carry out the experimental phase of this study at the Laboratory of Medical Genetics of the S. Orsola Malpighi Hospital.
A sincere thanks also to Roberta Zuntini, Michela Bonaguro and all the Laboratory of Medical Genetics members, who have welcomed me in the lab for several months.

Thanks to Prof. Jaume Bertranpetit and notably to Prof. Francesc Calafell, who allowed me to perform data processing at the Barcelona Biomedical Research Park and who certainly contributed to shape the final outcome of this thesis.
A great thanks also to all the members of the Unitat de Biologia Evolutiva of the Pompeu Fabra University of Barcelona and in particular to the "equipo" of BioEvo, who have made me feel at home during four unforgettable months.

Finally, a special thanks to my parents, my friends and, above all, to Francesca, who bore with love and patience both my physical absence during the Spanish stay and my being chronically engaged in a multitude of things to do (at the end the writing of this thesis), since they have always unconditionally supported and encouraged me in my personal and professional choices.