Genomic and Functional Approaches to Genetic Adaptation

Elena Carnero Montoro

TESI DOCTORAL UPF / 2013

Thesis Director

Dra. ELENA BOSCH

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Fitxer PDF de la tesi dividit en 7 parts

Part 1 de 7	pàg 0 - 17	Introduction Cap. 1 – Cap 2
Part 2 de 7	pàg 18 - 20	Introduction Cap. 3 : 3.1, 3.2
Part 3 de 7	pàg. 20 – 25	Introduction Cap. 3 : 3.2.1
Part 4 de 7	pàg. 25 – 28	Introduction Cap. 3: 3.2.2 – 3.2.3
Part 5 de 7	pàg. 29 - 64	Introduction Cap. 4 – Cap. 6
Part 6 de 7	pàg. 65 - 183	Objectives, Results, Discussion, Concluding remarks
Part 7 de 7	pàg. 184 – 233	References, Annexes

Part 5 de 7

4. More complex views on selection

4.1 The role of non-coding elements in adaptation

Protein-coding sequences are the best annotated elements on reference genomes but they only represent around 1.2% of their length. Furthermore, and surprisingly, there is a very high sequence identity of such protein-coding genes between humans and chimpanzees that do not account for all the observed phenotypic differences.

However, the relative contribution of changes in protein-coding genes versus regulatory regions in evolutionary adaptation is still a controversial issue. It has been long hypothesized that differences in gene regulation might underlie many of the phenotypic variation among populations and species (King and Wilson 1975).

Phylogenetic studies on multiple alignments of many complete mammals sequences have estimated that around 5% of the genome since the common ancestor of mouse and human is conserved, subjected to strong purifying selection, and is thus functional (Siepel et al. 2005). Because the fraction of conserved sequences is higher than the protein coding sequences, it seems obvious that a large fraction of the functional elements are non-coding sequences, and that they could also play a role in adaptive evolution.

Very few studies to date have focused on the role of non-coding elements that represent the "dark adaptive matter" for two main reasons. First, they are not as well annotated as protein-coding genes, so their identification or functional evidence beyond their constrains is still difficult to interpret; and second, it is difficult to establish a neutral sequence to use for comparison. When searching for signatures of adaptive evolution in protein-coding genes, it is an advantage to compare substitution rates of non-

synonymous and synonymous sites, because they are alternated and thus, subjected to the same background selection. For non-coding elements, the establishment of such neutral reference is clearly much more complicated.

Over the last 5 years, several works have taken a big step forward on this topic. On one hand, phylogenetic methods have been adapted to interrogate non-coding sequences using different neutrally evolving elements, such as the so-called ancestral repeats and/or pseudogenes. Under this approach Haygood et al. (2007) interrogated genome-widely the evolution of promoter sequences by comparison with close intronic sequences devoid of regulatory variants. They found a considerable amount of promoter regions with signatures of positive selection in the human and chimpanzee lineages. Interestingly, functional enrichment analyses on the positive selected set of promoters, showed a significant enrichment in promoters related to functions of the nervous-system. These had not been seen for protein-coding genes, although, they undoubtedly play a major role in the adaptation of our species.

A similar study as the one carried by Haygood, but considering ancestral repeats as neutral evolving reference, reveals a set of introns with signatures of positive selection that are enriched in neurological functions (data not published, Petit et al in prepraration).

Recently, MK-based tests have also been applied to non-coding sequences in mice and *Drosophila* and have revealed that the fraction of adaptive substitutions (α) and the adaptive substitution rate (ω_α) in these species is higher in conserved non-coding sequences than in protein-coding sequences (Kousathanas et al. 2011) (Mackay et al. 2012a).

Results in population-based studies through the detection of selective sweeps containing conserved non-coding elements also go in the same direction, and indicate that neural development and function have adapted mainly through non-coding changes

(Haygood et al. 2010) (see figure 10). Furthermore, the proposed candidate variant list for positive selection given by Grossman et al's (2013) recent work, above commented, contains numerous regions where the top-scoring variants by CMS are non-coding variants related to regulation, such as eQTL variants, variants located in transcription factor binding sites (TFBS) or predicted enhancer sequences, and variants located in lincRNAs.

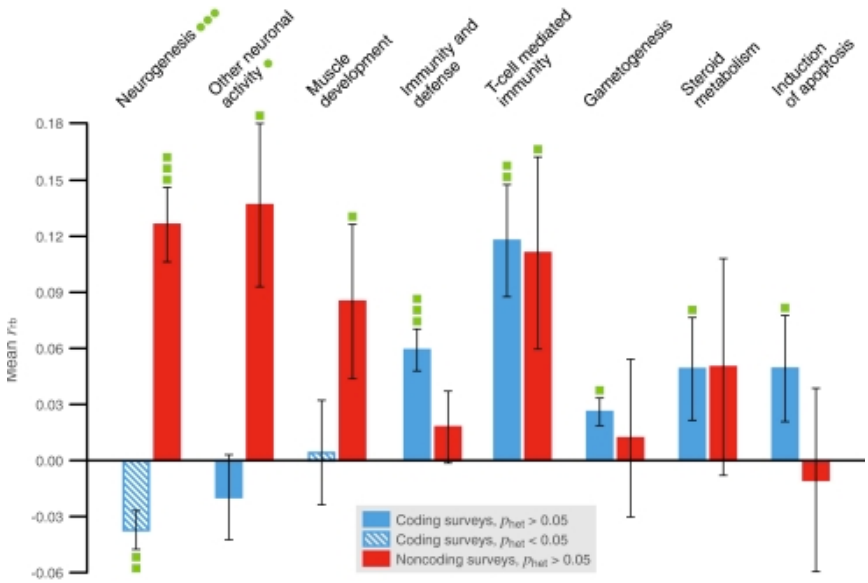

Figure 10. Functional enrichment analysis in coding and non-coding surveys of positive selection. Different biological categories are represented. Functions related to neuronal activities show different trends for coding and non-coding elements (Haygood et al. 2010).

Just a month ago, the first genome-wide study of human polygenic adaptation driven by changes in gene expression was published (Fraser 2013). The author took advantage of the extensive genotype data of diverse human populations and of the extensive data of expression profiles from different tissues and individuals. He showed that local adaptation is 10 times more likely to be driven by changes in gene expression than by changes in protein-coding genes.

In summary, there is mounting evidence for a main role of non-coding elements in adaptation and particularly in cognitive traits.

4.2 Beyond individual genes, natural selection on functional modules

Although mutations occur at a DNA level at individual loci, natural selection acts on phenotypes. Besides the rare cases of Mendelian traits, single mutations do not solely contribute to the acquisition of new functional innovations or traits. Phenotypes are rather commonly complex quantitative traits caused by the cooperation of modules of functional related and/or interacting genes, in which natural selection operated.

The current paradigm of large genome-wide analysis of adaptation is that they look for deviation of evolutionary rates of individual genes or regions from neutral expectation, making the strong assumption of independence among loci. Although further functional enrichment analysis has been able to reveal some important functions and pathways being preferably selected, the fact is that they do not formally test for selection acting on a function. Also, we have to bear in mind that usually statistical significance is difficult to achieve after multiple testing corrections.

In this regard, (Serra et al. 2011) created a new method called the Gene Set Selection Analysis (GSSA) to detect significant differences in rates of evolution over functionally related genes, and to look for their common pattern of evolution. Their method was applied genome-wide to coding regions of 5 mammals and revealed a large number of functional modules described in GO and PANTHER to have significantly higher or lower rates of adaptation in comparison to the genomic background signal in *Drosophila* and mammals.

Other studies performed on a pathway level at both comparative and population-based scales have focused on how the signatures of adaptation in coding genes are distributed in the structure of pathways, without taking into account the whole pathway as a single functional entity to be tested (Montanucci et al. 2011), (Dall'Olio et al. 2012).

Although the GSSA approach is very promising and represents a new opportunity to understand how selection shapes variability at more complex levels of the genetic architecture, it still has not been applied to data on non-coding elements, or on a population fashion, which will certainly enlighten our knowledge of how selection has affected complex phenotypes.

5. From candidate loci to advantageous phenotypes.

Divergence and population genetic studies serve as complementary tools to understand the process of natural selection shaping genetic differences among individuals, populations, and species by pinpointing potential candidate genes and regions involved in adaptation. However, bottom-up studies do not usually provide an ultimate adaptive explanation for such candidate loci.

A return to in-depth experimental follow-up studies is needed to reach actual insights of human evolutionary history, to find explanations for adaptive processes, and ultimately to fill the lack of knowledge in genotype-phenotype relationships. In order to achieve these objectives, deeper knowledge of how genetic variation affects molecular phenotypes, and how these affect fitness of individuals in challenging environments is needed. This achievement will involve several steps (see figure 11).

The first step would be identifying the causative adaptive mutation on a candidate region. Since there is no way to directly know what changes were favored by evolution, moving from candidate genomic regions to casual underlying adaptive mutations is difficult. To that end, one suggestion is to compile sequencing data along candidate regions in order to study the complete spectrum of variation of that genomic region. Current public available full-sequence data provide an excellent extended catalog of human variation to explore potential functional variants, but the small read coverage (around 6X in 1,000genomes) makes it likely that rare variants are not yet discovered in the studied populations. Although we expect selected variants to be found at high frequencies in the population where it has been selected, having access to the complete SFS may help to confirm the action of natural selection thanks to classical neutrality tests. The use of Sanger sequencing could provide an unbiased description of the variability, but depending on the length and number of candidate

loci to follow up, it could also be hard and expensive laboratory work. Alternatively, next-generation sequencing offers a great opportunity to capture the desired sequences in multiple individuals by specific capturing and parallel tagging different individuals, and by sequencing the pool at high coverage, although sequencing and mapping errors still occur at a high rate with current technology (Kircher and Kelso 2010).

Second, after the detection of potential causal variants, their possible implication in adaptation should be functionally studied. The decision on which functional approach should be undertaken ultimately depends on the nature of the variant. Transfection studies and transgenic models offer a great opportunity to explore cellular and physiological differences between putatively selected and non-selected alleles at non-synonymous positions. Another approach is to perform expression studies, based on QTLs (quantitative trait locus), on a collection of tissues if the candidate is a non-coding variant that affects the regulation of a gene. Likewise, if the variant modifies a transcription factor one could investigate its downstream effects in the regulated genes, or if the variant is sited in a CpG islands to what extent it affects the methylation pattern, etc. Although we are still far from understanding which specific role each position of the genome has, if any, the –omic revolution has allowed generating large scale high-throughput functional data, as for example, expression profiles by transcriptomics, methylation patterns by bisulfite sequencing, chromatin modifications by ChIP-seq, proteomic profiles, metabolomics description, etc., in a large tissues and cell-types. Such large functional data can help to have an initial idea on what function selection favored, and which type of functional analysis should be performed.

Third, if functional differences are finally found, there is still not direct evidence of whether they enhanced fitness of their carriers, and under which selective force. Although different population-based approaches can overcome this uncertainty.

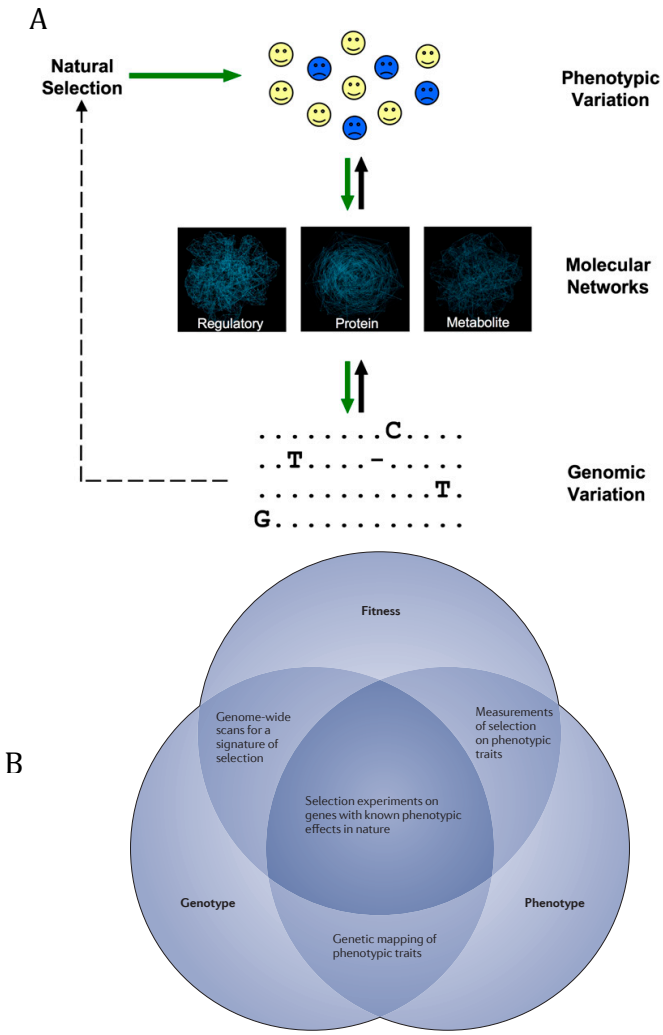

Figure 11. An integrative approach to study impact of adaptation. A) Although genome-wide studies make inferences in selection directly from patterns of genetic variation, natural selection acts merely on phenotypic variation. The link between genotype and phenotype is not direct and might involve the integration of the function of many loci in metabolic networks (Akey 2009). B) Connections between various approaches for studying the genetic of adaptation. Only when connections between genotype, phenotype and fitness are made, an allele can be considered adaptive (Barrett and Hoekstra 2011)

In summary, evolutionary genetics establish the method for the study of adaptation and it represents a starting point, but a multidisciplinary integrative approach with the involvement of many different disciplines such as epidemiology, history, ecology, anthropology, physiology, molecular and cellular biology needs to be taken into account in order to satisfactorily explain the events of adaptation and how the underlying genetic variants have generated them.

6.Cases of human genetic adaptation

Within the last 50,000 years, humans have moved out of Africa in several waves, spreading around most of the planet, colonizing a wide range of new different habitats. During this dispersal, human populations had to encounter strong alterations from their original environments such as extreme climate conditions, changes in the availability of food resources, exposition to never previously fought pathogens or to oxygen limitations at very high altitudes.

Doubtless, those changes in environmental factors triggered different processes of adaptation and can explain at least a partial fraction of the current human genetic diversity and phenotypic differences among populations.

Although many adaptive events have been theorized, only a handful of adaptive phenotypes have been described in detail, such as adaptation to malaria resistance in Africans, adaptation to lactose tolerance in African and European populations, adaptation to low UV light trough light pigmentation in Europeans and Asians or adaptation to high altitude in Tibetans (see figure 12).

The following sections will be dedicated to a more detailed explanation on two particular cases of human adaptation. The first section will describe how the immune system of different populations has adapted to new environmental challenges; whereas the second will tackle the potential role of genes related to micronutrient metabolism in the acquisition of advantageous phenotypes.

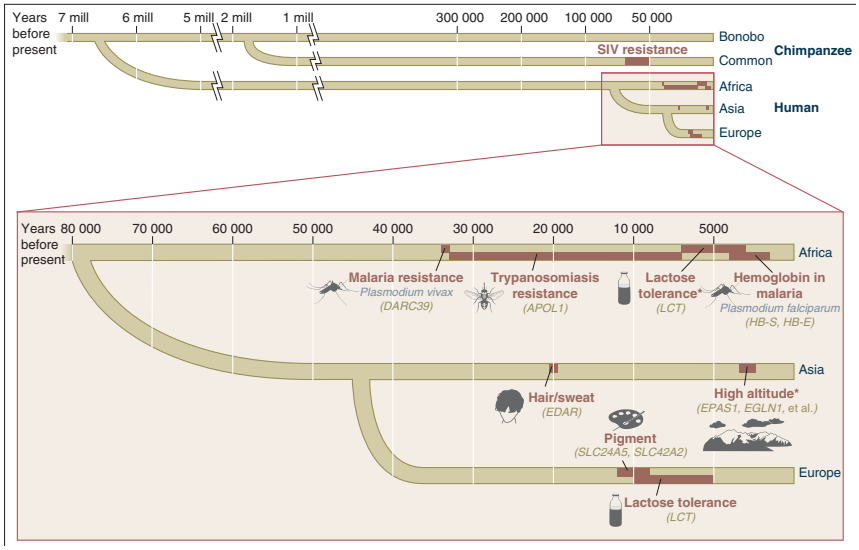

Figure 12. Some adaptive events and loci involved in human evolution. some cases, adaptation to the same enviromental challenge involving different loci has ocurred at different times in the same or different populations, representing cases of convergent adaptation (Vitti et al. 2012).

6.1 Infectious diseases as evolutionary drivers

*"Nature is the best doctor: she cures three out of four illness, and she never speaks ill on her colleges" (*Louis Pasteur)

"Now, here, you see, it takes all the running you can do, to keep in the same place" (Lewis Carroll)

Haldane first argued that infections are one of the major drivers of evolution in our species, since pathogens have been acting as a powerful selective pressure through our history (Haldane, 1932).

All animals and plants are constantly threatened by the invasion of microorganisms, and their immune system evolves accordingly to eliminate pathogens in the body. In the same way, pathogens are constantly evolving in order to acquire new strategies to avoid host

defense mechanisms. Resistance to pathogens, as opposed to other adaptations, does not increase the fitness of an individual by enhancing its reproductive chances, but allows organisms to survive ever-evolving pathogens in ever-changing environments. This process provokes a continuous pressure exerted on each other by both host and pathogen, and requires a constant process of natural selection in one species that leads to counter-adaptation in the other.

During their dispersal out of Africa to the rest of the world, humans not only faced a wide range of extreme hostile environmental conditions, including the encountering of pathogen species, but they also experienced changes in their subsistence strategies, which allowed the establishment of large, settled and interconnected populations. While in hunter/gatherer communities, factors such as the presence of animal reservoirs provoked some infections with chronic disease course due to incomplete immunity, the most important epidemic diseases appeared after the development of agriculture about 10,000 years ago. In particular, the domestication of animals and the sedentary lifestyle led to an increased exposure to zoonotic infections as well as to insect-born diseases such as malaria, yellow fever, filariasis, dengue or trypanosomiasis. Furthermore, increasing communication between urban centers since 3,000 BC allowed human settlements to be large enough to maintain infectious diseases as endemic forms (reviewed in Cagliani and Sironi 2013).

Infectious diseases represented the first worldwide cause of mortality until the very recent development of vaccines, antibiotics and hygienic measures, and remain the first cause of mortality in developing countries, where they significantly reduce life expectancy, then one can straight-forward thing that they have represented a major selective force for our species. Accordingly, a recent publication based on a genome-wide analysis of human genetic variation indicates that pathogen-driven evolution, among

other environmental factors, has the strongest influence on shaping human genetic variability through positive selection (Fumagalli et al. 2011).

Recent population genetics studies are notably contributing to the identification of defense-related genes that have allowed fighting pathogen exposures during our evolutionary history. Indeed, investigating how natural selection has affected genes related to the immune response has provided meaningful insight into the mechanisms that are crucial to survive infections.

This section will cover, first, a brief explanation of the main elements of the immune system, then, a description of the selective signatures in immune-related genes both at inter- and intraspecific levels. Later, there will be some examples of adaptation to certain pathogens, and, finally, a general picture of the pleiotropic effect of some positively selected immunity genes impacting susceptibility to high prevalent inflammatory diseases.

6.1.1 Elements of the immune system

The mammalian immune system comprises two differentiated steps in the immune response. The first is a non-clonal, non-specific, non-anticipatory response by the innate or natural system. The second is due to the action of the induced, specific, anticipatory and clonal adaptive or acquired system. Both processes function in an orchestrated way in order to react against non-self hazardous agents, with minimal inflammatory and immunopathological damage.

The innate immune system triggers a first fast host defense mechanism against the pathogen after a first interaction with it. Recognition occurs via a limited number of germline-encoded pattern-recognition receptors (PRRs) that recognize distinct

pathogen-associated molecular patterns (PAMPs) from different microorganisms. PRRs are constitutively expressed in the host and they can be located in the cell surface as membrane receptors, in the cytosol as free receptors, or they can be secreted out of the cell. PRRs can be classified according to the PAMP they recognize and to the signaling response they trigger later. The best characterized are the TLRs (Toll-like receptors) that can sense viral, bacterial and fungal components, and can be located both at the cell surface and in endosomes. CLRs (C-type lectin receptors) are another type of PRRs located at the cell surface. They recognize sugar components from different microorganisms. In the cytosol, receptors that are more specialized in sensing bacterial and viral components (such as pieces of RNA or DNA molecules) can be found. Good examples are the RLRs (RIG-I like receptor) and the NLRs (NOD-like receptors) receptors. Humoral PRRs are soluble pathogen sensors that are secreted out of the cells and circulate in tissue fluids, such as collectins, ficolins and/or elements from the complement system involved in a wide functional spectrum of processes such as opsonization, phagocytosis of microbes, and/or activation of cellular apoptosis.

Adaptor molecules (as TIR domain-containing proteins) are associated to PRRs and are ultimately responsible for initiating a downstream signaling cascade after stimulation of PRRS by PAMPs. Activation of some proteins in the signaling cascade induce the expression of effector molecules such as citokines (like interleukines or interferons (IFNs)) or antimicrobial peptides (AMPs), that are secreted out of the cells and are involved in pro-inflammatory immune responses to fight pathogens or non-self particles (see figure 13).

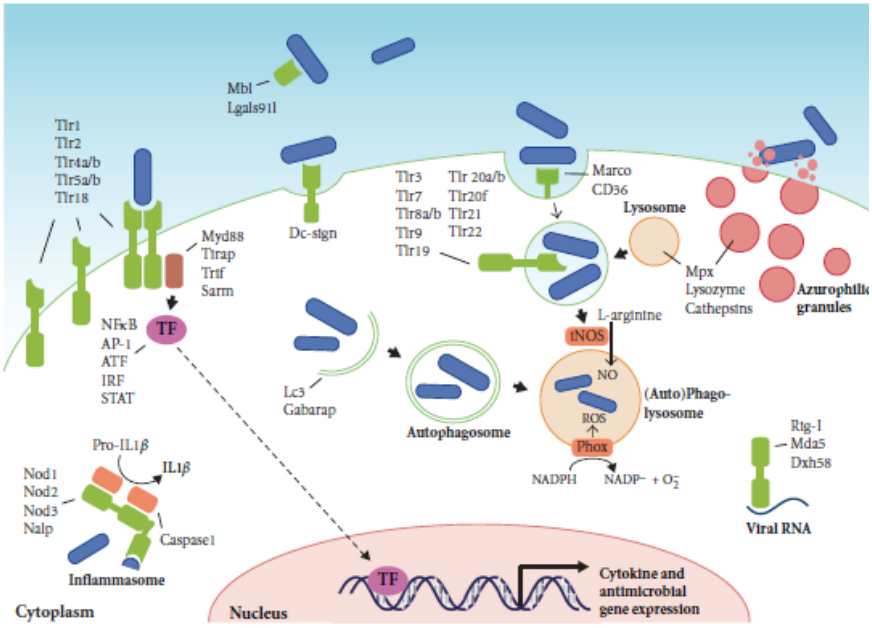

Figure 13. Pattern recognition receptors and effector mechanisms of the innate immunity Recognition of PAMPs by PRRs leads to activation of transcription factors that initiate transcription of cytokine genes. (Van der Vaart, Spaink, and Meijer 2012).

The acquired or adaptive immunity is a later immune response phase involved in the complete elimination of pathogens and in the creation of memory cells that will react against them if a second infection occurs. It is mediated by T cell receptor (TCRs) and B cell receptor (BCRs) expressed in lymphocytes. The antigen receptors expressed in these cells are assembled from variable fragments encoded by the same set of genes. But their expression is determined by different recombination events mediated by the recombination activating gene (RAG) protein that produces a diverse repertory of receptors, plus other mechanisms such as gene conversion and non-template nucleotide addition, and/or somatic hyper-mutation. All these mechanisms allow a large variability of lymphocytic receptors to recognize a great amount of different pathogens.

Two types of lymphocytes express antigen receptors: conventional B and T lymphocytes, which express non-specific random antigen receptors, and innate-like lymphocytes. The major histocompatibility complex (MHC) class I and II of antigen-presenting cells (as dendritic cells) present digested antigen peptide to T cells. Once they have recognized an antigen, they trigger the cell-mediated immune response through the activation of phagocytes, antigen-specific cytotoxic T-lymphocytes and the release of various cytokines. Conventional B cells recognize antigens by binding to an epitope; they are in charge of the humoral response. They produce specific antibodies against the recognized antigen and release them as humoral circulating cells that will act as memory cells to prevent future infections.

Because the innate and the adaptive systems cooperate to fight non-self dangerous agents, it is sometimes difficult to differentiate them. This is the case of innate-like lymphocytes. They exhibit antigens in their membranes, as conventional B and T lymphocytes do. However, they do not produce a rearrangement of their receptors, and their specificities are predetermined towards particular ligands just like the case of PRRs. Moreover, unlike the conventional lymphocytes, they are often localized in the mucosa where the first contact with the pathogen is produced, and unlike T of B cells, they produce a rapid and high amount of cytokines. Furthermore, recently, new subtypes of T-cells and B-cells, such as T-reg and B-reg have been discovered. These cells act as lymphocytic suppressors once the pathogens have been eliminated and, thus, assure a balanced response of the immune system.

In summary, the immune system has evolved to defend individuals from external (and internal) pathological threats. Its ultimate function is to recognize the difference between self and non-self particles in order to maintain a balanced internal environment. Despite the high complexity that the immune system has acquired through evolution, and the high number of elements involved in

the immune response, sometimes, anomalous responses occur. What is endogenous in an organism is recognized as non-self, and vice versa, leading to misplaced immunological responses such as allergies, or autoimmune and/or cancer diseases.

6.1.2 Signatures of natural selection

Studying the impact of natural selection on immune-related elements is a powerful strategy to see the biological relevance of such elements and the molecular mechanisms behind host-defense response.

Signatures of natural selection are widespread among the immune elements with different types of selection having shaped the variability of different immune-related genes. This is mainly due to differences in the very nature of elements: the differences in the type of pathogen they recognize, in pathogenicity, and in the role they play during the host-pathogen interaction.

Although selective constraints in immune elements are widespread in primates, and some of the members show strong signals of purifying selection in humans and in other species, many cases of neutral evolution have also been observed. Many immune functions are essential to the host: some changes are deleterious, and can cause important immune disparities. But the cases of neutral evolution indicate that many others changes are redundant, and in case of function loss, they can even be replaced by alternative immune mechanisms.

Furthermore, both comparative and population studies consistently show that elements related to host defense against pathogens are enriched in signatures of positive selection. Indeed, scans of positive selection have revealed more than 300 immune-related genes as candidates having undergone recent positive selection

(see figure 14). This proves the original idea that genetic changes in immune-related genes cause the adaptation of population and species to the presence of certain pathogens (Barreiro and Quintana-Murci 2010).

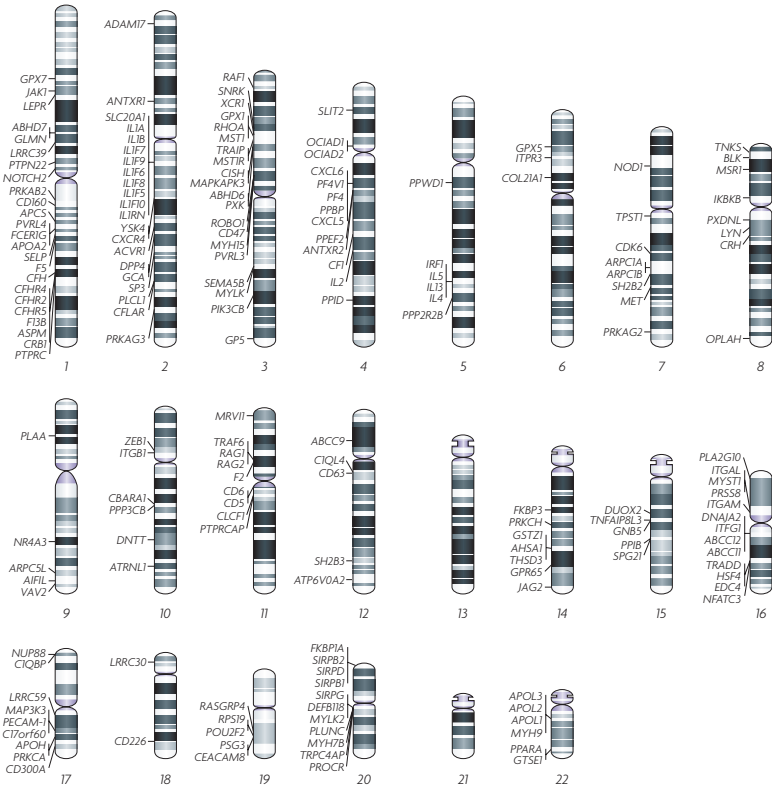

Figure 14. Genomic map of immune-related genes that are candidates for positive selection reported in at least 2 genome-wide analyses (Barreiro and Quintana-Murci 2010)

For example, comparative studies have shown high divergence rates and signatures of positive selection in many of the TLRs expressed in the cell surface (such as the *TLR1, TLR4* genes and the cluster *TLR6-TLR1-TLR10*) in humans. On the contrary, TLRs expressed in endosomes show the strongest signatures of purifying selection. These major differences observed between the two groups of receptors provide interesting insights into their different

biological roles. Molecular changes in the intracellular receptors introduced by mutation could easily make the receptor react against self- molecules and thus, lead to an autoimmune disease. To avoid such an outcome, and since long ago, purifying selection has clearly acted, eliminating new variation on the corresponding coding genes. Variation in cell surface receptors does not seem to imply such relevant phenotypic outcomes. This could be due to their redundancy and robustness in function, to the fact that in humans the selective pressure acting on this class of receptors is not present anymore (as it is in other primates), or to the fact that, for them, genetic variation is advantageous to the host (Manry and Quintana-Murci 2012) (see figure 15). Interestingly, new findings have shown that purifying selection at TLRs in more pervasive in Great Apes than in humans, where both receptors at cell surfaces and endosomes are strongly constrained. Such differences in patterns of diversity highlight differences in the importance of TLRs in sensing pathogens, which is an evidence of different selective pressures acting on different species (Quach et al. 2013).

Cytosolic receptors also show different patterns of evolutionary signatures. The RLR family, involved in the detection of viral RNA molecules, shows a variability pattern of weak constraint close to neutrality. This suggests that, although they are involved in the inflammatory response, their function is not essential. However, like in the TLRs family, the signal is not uniformly distributed among its elements. For example, some of the binding domains of the RIG-I receptor have been found to evolve under strong purifying selection, thus indicating that changes in such sequence domains could easily result in functional impairments. Likewise, the NLR cytosolic receptors also show different selective patterns. Among them, the NALP family shows strong signatures of purifying selection, while receptors such as CIITA and NAIP, show weaker constraints. Interestingly, different population genetic studies have revealed signatures of positive selection not only in less constrained genes (such as *MDA5, LGP2* and *CIITA*s), but also in some elements of the very constrained

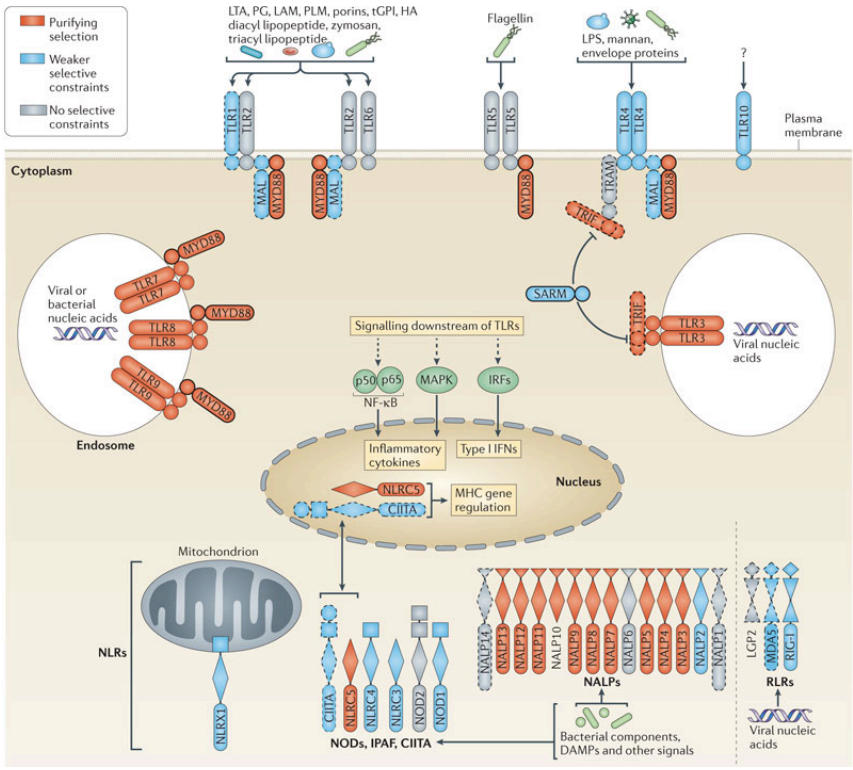

Figure 15. Different modes of selection shape variability of innate immune receptors. Differences are due not only to the family, but also to the localization and the pathogen that the receptor recognizes (Quintana-Murci and Clark 2013).

family of NALP receptors (as *NALP1* and *NALP14* genes). This study was conducted on different populations, suggesting an important role of these molecules in the immunological adaptation of populations to changing environmental pressures (Vasseur et al. 2012) (see figure 15).

There are many different signatures of selection in effector molecules. For example, some cytokines such as the IFN family play an important role in the immune response to viral infection. Most genes coding for IFNs are very conserved, and amino-acid altering changes provoke serious immunodeficiencies as a result of viral infection, and are associated with some cases of Mendelian diseases. Other members have accumulated more changes through evolution and are thought not to be indispensable for the host.

Some variants that might reflect protection against some viral pathogens are described to be under positive selection in Eurasia. The high level of variability in the signal of selection might reveal differences in the immunological relevance of the function of the different subtypes (Manry et al. 2011).

Balancing selection has been said to be pervasive among innate immunity genes (Ferrer-Admetlla et al. 2008). In fact, balancing selection is the main selective process shaping the diversity of the major histocompatibility complex (MHC). The MHC is a combination of molecules directly involved in the antigen presentation to effector immune cells, and it is the most polymorphic gene cluster in the genome (Hughes and Nei 1988), (Hedrick, Whittam, and Parham 1991). Equally, the killer-cell immunoglobin-like receptors (KIR), and the ERAP1 and ERAP2 proteins, which directly interact with the MHC, show high levels of heterozygosity. Several studies have revealed that the diversity shown in both the MHC and the KIR families, and the ERAP1 and ERAP2 proteins is the result of both balancing and directional selection (Andrés et al. 2010), and that different events of convergent adaptation have occurred at different times and in different populations.

As Haldane already noticed in the 1930s, not only the genes directly involved in the response to pathogens are targets of selective events due to interaction with these. There are other important immune elements related to determining resistance to infections that are necessary for the immune action. They show patterns of genetic diversity clearly shaped by the action of natural selection. A clear example is the case of cell surface proteins that somehow complicates the pathogen invasion of the host cell.

Besides the very well studied case of the MHC, most balancing selection cases have been detected in genes coding for proteins involved in the post-translational modification of glycan states and,

thus, in the determination of different serotypes, such as the *ABO* and the *FUT2* genes. The *ABO* gene, coding for a glycotransferase enzyme, shows very high levels of polymorphism, probably resulting from the action of balancing selection (Calafell et al. 2008). Similarly, the *FUT2* gene, a fucosyltransferase that regulates the expression of the *ABO* gene, was shown to have been subjected to long-lasting events of positive and balancing selection in different populations (Ferrer-Admetlla et al. 2009). Interestingly, both the *ABO* and the *MHC* polymorphisms are trans-specific, meaning that they are among the few cases of segregating alleles shared among different primates and preserved by selection for millions of years (Ségurel et al. 2012).

Beyond the candidate loci or the genome-wide studies, Casals et al, (2011) interrogated the action of natural selection in innate immune related elements in a network fashion. In their work they showed how different modes of selection are regularly distributed along the gene to gene interaction network. The most constrained elements tend to be located in more central positions, while the accelerated evolving loci and those under positive selection were preferentially located at the network edges. Also considering a pathway analysis, but using a different approach, Daub et al, (Daub et al. 2013) using a gene-set-enrichment test showed that many of the adaptations to pathogens are good examples of polygenic selection.

In summary, the complex evolutionary pattern that most of the different gene families involved in host defense show, reflects not only the great diversity in function and immunological relevance of the immune elements, but also the complexity of the old evolutionary history of host-pathogen interactions. The genetic variability we see today is the result of adaptation to constant environmental pressures that microorganisms have exerted in the immune system of the host. It is clear that preserving the functional integrity is essential in order to maintain pathogen

resistance, but it is also important to favor new functional changes to be able to provide new responses to the ever-changing environmental pressures exerted by pathogens.

Evolutionary genetics has allowed a deeper understanding of the role that many immune element has played in the host defense adaptation of species and populations. But one should go beyond these genetic studies in order to better understand which have been the specific pathogen agents that have provoked such particular selective events, and which are the specific advantageous properties that positively selected changes have conferred to the host.

An integrative approach comprising the nature of epidemic events as well as the physiological properties of the immune elements is needed, and studies conducted so far towards this direction are discussed in the next section.

6.1.3 Examples of genetic adaptation to certain pathogens.

Although much effort has been done to successfully identify hundreds of candidate immune-related genes having undergone positive selection, the reality is that only in a few of them their advantageous allele has been properly identified and phenotypically characterized; and in an even smaller fraction of cases, the specific selective pressure was also identified.

The most extended study so far is the study of the adaptive events that malaria has imposed upon human populations. Malaria is caused by different *Plasmodium* species. Historically, it has been one of the main causes of mortality of our species, and it is still so for many developing countries where the vector of the parasites, a tropical mosquito, is frequent. Unquestionably, the presence of

such an important selective pressure has shaped the variability of the genome of the exposed populations. Through history, different loci that have been selected to fight malaria represent one of the clearer examples of convergent adaptation. The selected loci which participated in resistance to malaria can be divided in different groups:

i) genes directly involved in the immune activation response after the infection, such as *HLA*, *IFNG* (interferon γ), *TNF* (tumor necrosis factor) and/or *CD40LG* (CD40 ligand);

ii) genes involved in erythrocyte metabolism that somehow obstruct or prevent the establishment of the parasite in red blood cells, where the parasite spends part of its life cycle feeding on hemoglobin, including variants in genes such as the *DARC* (Duffy bloody group chemokine receptor), *G6PD* (glucose-6-phosphate deshydrogenase), glycophorins A and C (GYPC and GYPA) and the globin genes *HBB;*

iii) genes that mediate cellular adherence of the parasite such as the *CR1* (complement component receptor) or the *ICAM1* (intracellular adhesion molecule 1)

Very interestingly, the appearance of some of these variants has been estimated to be correlated with the dramatic population expansions of the African mosquito vector due to the establishment of the first Neolithic societies in African human populations (reviewed in Cagliani and Sironi 2013).

Other links have been established between selected variants and other specific pathogen presence but to a much lower extent, and without as much compiled evidence as in the case of malaria.

One of the approaches used to search evidences of pathogen-host interaction is looking for relations among their genetic variability.

For example, *HLA* diversity is correlated to viral richness but is not correlated to bacterial or protozoa richness, which provides evidence for viruses having exerted a stronger selective pressure. The same is observed for some TLR genes. Curiously, although we do not know much about selective events caused by worm infections, it was shown that *IL* genes diversity is better correlated to helminthes diversity than to intracellular parasites (Fumagalli et al. 2009).

Another approach taken to link the effect of putative adaptive variants with their phenotypic outcome is studying the distribution of some linked pleiotropic effects in the population. For example, a deletion in the cytokine receptor gene *CCR5*, expressed in T lymphocytes, while absent in African and Asian populations, has shown to have reached high frequencies in Europeans by a process of positive selection that took place around 1,000 years ago (Sabeti et al. 2005). In that case, several lines of evidence support the notion that the deleted allele was selected to confer protection against the smallpox *Variola major* virus, which caused a high mortality rate. Interestingly, the CCR5 deletion is associated with protection against HIV infection, which is currently affecting human populations. In the wild type form, the CCR5 co-receptor is exploited by the HIV virus to enter lymphocytes. Although representing a major threat in current populations, HIV infection is not considered to have been the actual selective pressure increasing the frequency of the deleted variant in Europeans, as this virus only appeared recently. On the contrary, it is thought to be an immune side consequence of a past adaptive event, that represents a new advantage in case of HIV infection today (Galvani and Slatkin 2003).

As discussed previously, resequencing is a good approach to study the complete allele frequency spectrum, confirm signatures of selection and detect the candidate variants that could result in the advantageous phenotype. Resequencing approaches of cytosolic

microbial sensors (members from the NOD-like receptor family) have identified signatures of adaptive evolution in the *NLRP1*, *NLRP14* and *CIITA* genes from the NALP family. They have also further identified a set of candidate variants which have been the target of positive selection in these genes in African and European populations based on neutrality tests (Vasseur et al. 2012). It is interesting to see that the *NLRP1* gene has been associated with different susceptibilities to *Toxoplasmosis* congenital infection, revealing the important role played by the NLRP1 protein in detecting protozoan microbes. However, as in the case of the *CCR5* gene, it is believed that the *Toxoplasma gondii* intracellular parasite could not have exerted enough selective pressure to be the real target of positive selection, and thus, that another more ancient cause might have contributed to the *NLRP1* allelic repertoire.

Functional genomic screens, together with genome-wide association studies related to the immune response, have also yielded important knowledge about the genetic elements associated with specific pathogen resistances. For example, IFITM is a family of interferon-inducible transmembrane proteins which participate as effectors in the early stages of the innate immunity restricting the replication of multiple viruses. Specifically, IFITM3 has been proved to confer resistance to Influenza 1 H1N1, West Nile and Dengue viruses, suggesting that its genetic diversity and evolutionary history has been, respectively, influenced and driven by viral infections. Moreover, both *in vitro* and *in vivo* studies with knockout mice have demonstrated its essential role in fighting influenza. Resequencing the whole sequence of *IFITM3* in patients has shown that a variant influencing the alternative splicing of the protein is significantly more frequent in the case group than in controls. Finally, it was also recently demonstrated that IFITM3 was targeted by recent positive selection in African populations; an evolutionary pattern that could easily result from a scenario where IFITM3 has a protective role against one of these infections (Everitt et al. 2012).

Genome-wide scans have revealed that some genes, like *LARGE* and *IL21*, related with Lassa hemorrhagic fever (LS), show signatures of positive selection in some West African populations where the viral infection is endemic. The *LARGE* gene codes for a glycosilase that specifically post-translationally modifies α-distroglycan (α-DG), the cellular receptor of the Lassa virus. *IL21* is involved in the systemic clearance of viral components after an infection has occurred (when the virus has reached the inside of the cell). Although it is biologically linked to the Lassa virus, its function is not restricted to this specific virus, but applies more generally to all viral infections. Thanks to the application of the CMS method, the signatures of selection, initially spanning a large region of around 300kb of the genome where more genes are located, have now been reduced to a shorter region where only a few candidate variants outside the ORF region of both genes are localized. This suggests that the adaptive advantage behind such a signal might be conferred through regulatory changes. Functional follow-up studies on candidate variants within these genes are needed to understand how they have conferred resistance to the viruses.

One of the most complete studies on positive selection was performed by Genovese et al (Science 2010), where they incorporated association, genetic, epidemiological and functional experiments. In their work, they show how the increased prevalence of kidney diseases in African Americans is due to variability shaped by positive selection at the *APOL1* gene in African populations. Genome-wide studies had previously shown strong evidences of positive selection in a region containing both *MYH9* and *APOL1* genes. Furthermore, association studies of kidney diseases had found susceptibility associated to variants at the *MYH9* locus. By using resequencing data from the 1,000 genome project, the investigators could identify candidate derived variants, not at the *Myh9* locus but at the nearby-locus *Apol1* that showed high allele differentiation among Africans and the rest of

HapMap populations (Europeans, Han Chinese and Japanese). Association studies confirmed that such variants were significantly more present in African Americans cases with kidney diseases, than in African American controls with no family history of related diseases. The authors functionally proved that the derived state of *Apol1* present in Africans shows higher lytic activity for some species of *Trypanosoma* sps. Thus, researchers proved that the selected variant had resistance ability for *Trypanosoma* sp. Whether the selected variant became more frequent in African populations to confer resistance to this pathogen is, of course, a speculation, but all the evidence in this study points to a mode of selection through a heterozygous advantage in which homozygotes for the ancestral allele are more vulnerable against the infection, homozygotes for the derived allele suffer from severe kidney diseases, and heterozygotes show higher fitness (Genovese et al. 2010).

The examples detailed in this section show how, thanks to population genetics tools, many elements involved in pathogen-host interaction have been successfully identified, and how the impact and pressures exerted by specific pathogens can now be studied within a very wide scope. However, many of the studies already performed are still incomplete stories that are based on much speculation. It is still difficult to reach enough knowledge to close the gap between past selective signatures, pathogen agents that drove the selective event, as well as the corresponding link between causal variants and the mechanistic way by which they conferred a functional advantage to the host. An integrated approach, as the one used by Genovese et al, is needed to efficiently understand how past adaptations have shaped our genomes and which consequences they have had.

6.1.4 The hygiene hypothesis. Pleiotropic effects of adaptive alleles.

"You cannot have your cake and eat it too"
(General wisdom)

It has been already widely discussed that pathogens have represented a major historical threat for human evolution, and that, as a consequence, hundreds of variants conferring resistance to infections have been targeted by natural selection in recent human history.

However, over the last few decades, due to the development of vaccines and the implantation and improvement of good healthcare systems, infections no longer represent a survival challenge in developed industrialized countries, and are not among the top mortality causes anymore. On the contrary, other new diseases have rapidly appeared in these populations at high prevalence, as it is the case of autoimmune disorders.

An explanation for this phenomenon was provided long ago by the hygiene hypothesis. It postulates that today's industrialized societies experience imbalanced immune responses that predisposed them to a higher risk of autoimmune diseases, as a consequence of living under new pathogen-free non-challenging environmental conditions to which the population has not had enough time to adapt (Sironi and Clerici 2010).

Resistance variants in immune-related genes that have been selected to fight infection, generally act by enhancing pro-inflammatory responses of the immune response to be more efficient, as it has been demonstrated by many functional studies. Such immune enhancing today is not favorable for the new pathogen-free conditions and provokes an unbalanced hyper-

reactivity of the immune elements that can lead to autoimmune/inflammatory disorders.

Supporting evidence for this hypothesis originally came from both the epidemiological and the immunological fields. For, example, notable differences were noticed in prevalence of autoimmune diseases between the developed countries and the developing countries where infection still is a main cause of morbidity.

Evolutionary evidence for the hygiene hypothesis can be obtained as well by adding data from scans of recent human positive selection to data from genome-wide association studies about the many complex diseases related to the immune system. Indeed, many of the regions that have been positively selected harbor genetic variants involved in inflammatory diseases, suggesting that some variants that played a role in past adaptations to infections confer susceptibility to inflammatory disease today.

Barreiro, back in 2010, and using functional enrichment analysis, noted that SNPs with signatures of recent positive selection based on iHS values are more present among the susceptibility variants found in GWAs related to autoimmune diseases than what is expected by chance, or in other GWAs of complex traits (see figure 16). Since then, many studies have observed this relationship between signatures of positive selection and association with inflammatory diseases.

For instance, the interleukin receptor genes that show signatures of positive selection, such as *IL2, IL21, IL18-RAP,* have been associated with inflammatory bowel and celiac diseases (Cagliani et al. 2013). Variants at the selected *CIITA* loci have been linked to rheumatitis (Eike et al. 2012), and variants at the positively targeted *FUT2* gene for conferring protection from norovirus infections, are risk variants for Crohn's disease (Franke et al. 2010).

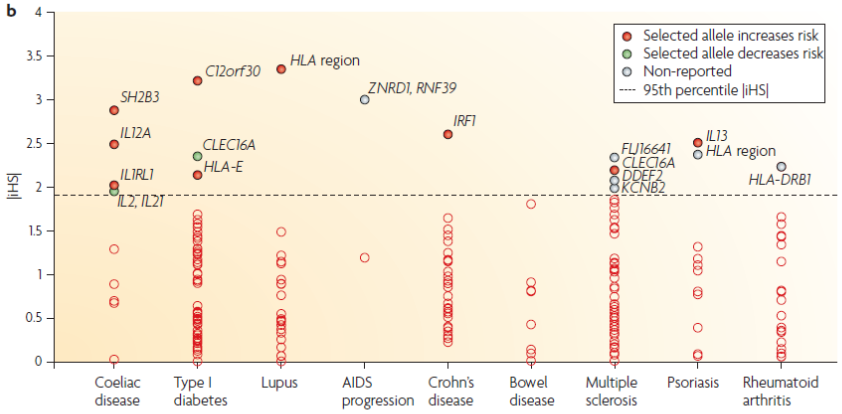

Figure 16. SNPs associated with diseases are enriched in signatures of positive selection. From (Barreiro and Quintana-Murci 2010)

Although a systematic analysis is already available and confirms that generally adaptive loci play a key role in influencing susceptibility to inflammatory diseases, a more comprehensive work is needed to first, identify causal alleles and not just the variation associated with the disease which is probably linked to the actual causal variant, as it is the case for most of the known associated variants, and second, to understand the biological relevance of the impact of the adaptive allele on influencing the pathogenesis of the disease (Raj et al. 2013). Furthermore, some aspects of such process still need to be investigated, such as: 1) which is the percentage of risk variants targeted by pathogens; and 2) which type of infectious agents has exerted the pressure at each disease. For example, in a recently published paper, Cagliani et al. (2013) have demonstrated that Crohn's disease loci are common targets of protozoa-driven selection, and not of other pathogen agents such as virus or bacteria.

Some precaution should still be taken in the study of pathogen-driven evolution. It is generally considered that immune-related genes showing signatures of positive selection are due to pathogen-driven evolution. Although pathogen-driven evolution is a main selective force shaping genome diversity, it could be that positive

selected changes were not due, indeed, to selective pressures driven by pathogens. This is due to the fact that most of the immune elements not only play one role in the physiology of organisms, but many, and some are not necessarily related to the defense against pathogens.

6.2 The role of genes related to zinc metabolism in genetic adaptation.

The maintenance of micronutrient homeostasis is fundamental in living beings to ensure the correct molecular and cellular functions that depend on metal presence. Micronutrients are essential metals that are not endogenously produced by organisms. They are incorporated from the trophic chain, primarily, by the consumption of plants that have incorporated them from the soil, and secondarily, by the consumption of animals that have, themselves, consumed plants.

The concentration of micronutrients within the cell has been proved to be tightly regulated by numerous membrane proteins that participate in the micronutrient metabolism. This, indeed, proves the importance of the homeostasis maintenance. In fact, developmental impairments have been associated with deleterious mutations within those proteins (Kambe, Weaver, and Andrews 2008).

The colonization by human populations of numerous challenging environments also includes inhabiting regions with different soil metal concentrations. In order to ensure correct amounts of metals within the cells, several adjustments of the elements in charge of the metal homeostasis maintenance must have occurred in response to the heterogeneous distribution of metal concentration in soil around the globe. Despite their recognized importance, genetic adaptations towards this environmental pressure have not been described yet.

The importance of diseases related to the metabolism of micronutrients such as iron and magnesium have been long known and widely studied. However, despite its essential function, the importance of zinc for the body balance was only truly revealed a few decades ago, when Dr. Prasad noticed that people from South and North Asia (Iran, Pakistan, India, Bangladesh), where zinc in soil is known to be present at low concentrations, suffered from zinc deficiency as an endemic condition which provoked growth and mental retardation, gonadal dysfunction, cognitive impairments and immune disorders.

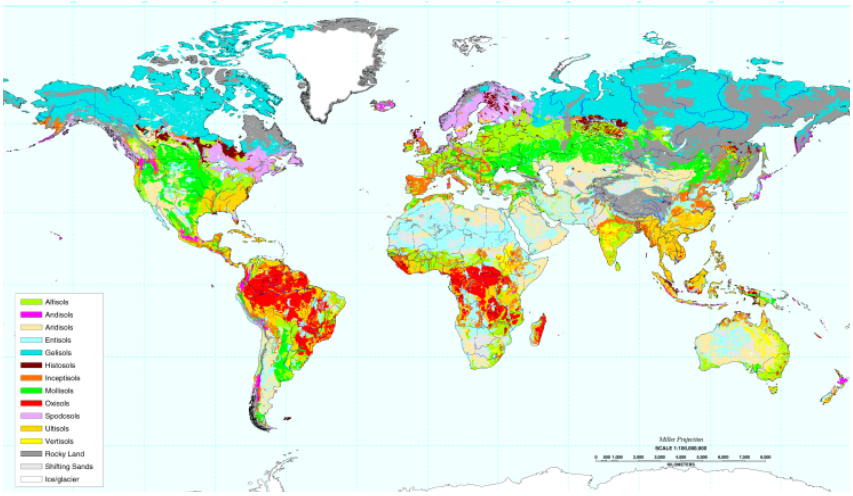

Figure 17. Global distribution of soil type. Taxonomy from the USDA (United States Department of Agriculture). Green color refers to the best indicator of high soil fertility, while red color refers to the worst. (From Wikipedia)

Since then, it has been recognized that, at a molecular level, zinc is required for the function of more than 300 metalloproteins. We also know of the existence of more than 2,000 zinc-depending transcription factors, and we know that zinc homeostasis is critically important to human health, since it influences processes such as aging and it is an essential element for immunity and diseases such as diabetes or cancer (Rink and Haase 2007).

Two families of genes are involved in the regulation of zinc homeostasis. The *SLC30A* gene family, which contains 10 zinc transporters (ZnTs) and it is in charge of decreasing intracellular zinc levels by transporting zinc from the cytoplasm to the organellar lumen or out of the cell. And, on the contrary, the gene family *SLC39A,* which is formed by 14 zinc influx transporters (ZIPs, Zrt-, Irt-like proteins) and participates in increasing intracellular zinc levels by either transporting the metal from the extracellular space, or from the organellar lumen into the cytoplasm (see figure 18) .

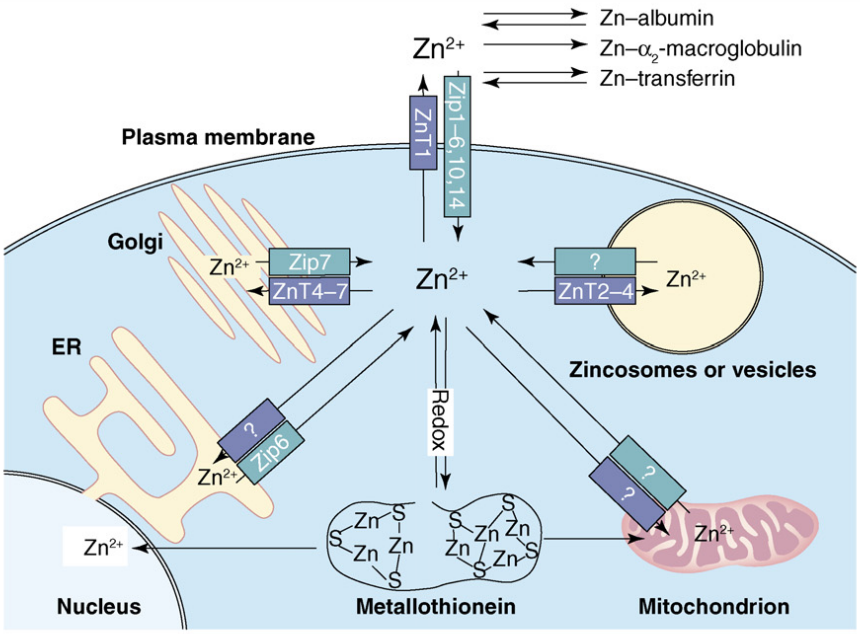

Figure 18. Zinc homeostasis in mammalian cells. Availability of cellular zinc is under tight control mediated by interaction of expression, localization and affinity of zinc transporters and binding proteins. While Zip proteins mediate the transport of zinc into the cytosol, ZnTs mediate zinc efflux from the cytosol. In the cytoplasm, zinc is available under its free metal form (Zn^{2+}) or bound to metalloproteins, that exchange Zn molecules by redox reactions. From Rink and Haase (2007)

Different transporters are expressed in different tissues; their expression is regulated by different factors, such as hormones, cytokines or the metal presence itself, and the disruption of their function has been linked to different diseases, which reveal their non-redundant roles.

Although adaptive phenotypes involving genes related to zinc metabolism have not been described so far, genome-wide scans have shown significant signatures of positive selection for some of the genomic regions coding for them (Grossman et al. 2013), revealing their possible role in population adaptation to changes in zinc availability. However, little attention has been given to them.

Again, further follow-up studies are necessary, first, to investigate the action of natural selection on these genes; second, to identify selected variants, and then, to understand how individuals have adapted to the micronutrient changing availabilities in their diet.

6.2.1 Nutritional immunity: reduced zinc availability as a target of pathogen-driven selection.

It has been described that pathogens, when infecting a host, express different factors that steal micronutrients from the host, such as iron, to benefit their own growth. In the opposite direction, the host has developed defense mechanisms to sequester iron out of the cells to prevent microbial growth, and thus, infections. This observation led to the concept of "nutritional immunity" (Hood and Skaar 2012) (see figure 19).

Interestingly, a very recent paper has shown that polymorphisms in a gene related to iron intake confer susceptibility to tuberculosis (Baker et al. 2012). The same stealing-sequestering dynamics have been seen for other micronutrients such as zinc, and it has been

proved that several members of the ZIP and ZnT families are involved in limiting the availability of zinc to prevent infections (Kehl-Fie and Skaar 2010).

For example, after stimulation of dendritic and T-cells (simulating a response after a pathogen recognition), it was observed that the ZIP proteins had a lower expression in the cells, while the ZnT proteins were highly expressed. As a consequence of the regulation of both processes, the zinc is not available in the cell, and such outcome is a defense mechanism (see figure 19).

The identification of genes related to micronutrient transport under signatures of positive selection might suggest that, likely, they also play an important role in the continuous adaptation to pathogen threats. The investigation of their adaptive function is an intriguing new field in the study of human immune genetic adaptation.

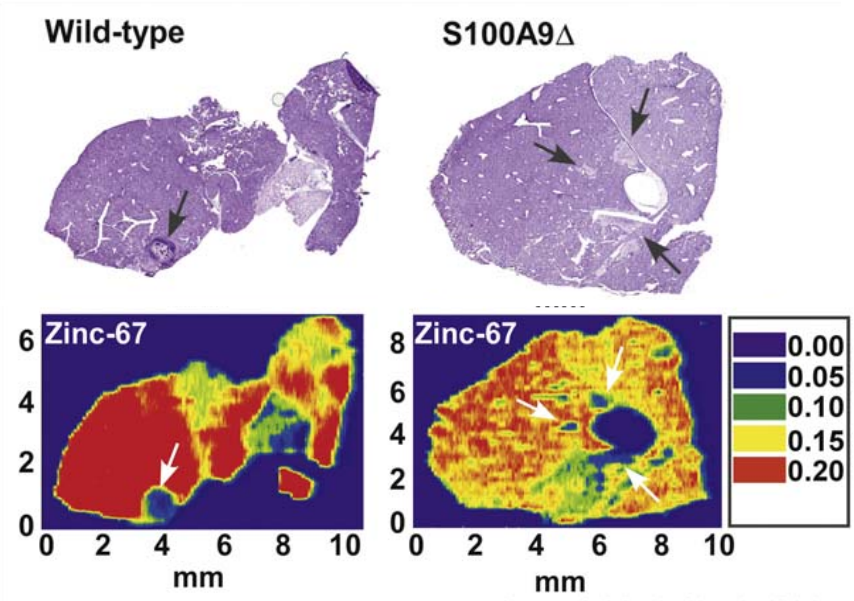

Figure 19. Zinc is found at reduced concentrations in infected liver samples compared to wild-type non-infected ones. From Kehl-Fie and Skaar (2010).