UNIVERSITÀ DEGLI STUDI DI SASSARI

CORSO DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE

Coordinatore del Corso: Prof. Andrea Fausto Piana

CURRICULUM IN GENETICA MEDICA

Responsabile di Curriculum: Prof.ssa Margherita Maioli

XXIX CICLO

A Genomic Map of Positive Selection in Sardinia

Coordinatore:

Prof. Andrea Fausto Piana

Tutor:

Prof. Francesco Cucca

Tesi di dottorato di:

Dott. Matteo Floris

Anno Accademico 2015 – 2016

Abstract

The recent production of population-scale genomic data offers an unprecedented opportunity to understand how natural selection has shaped human phenotypic variation within populations. To identify signatures of recent positive selection in Sardinia, we used 23 million single nucleotide polymorphisms from low-coverage whole genomes of 3,514 Sardinians along with data from the 1000 Genomes project. Using single-population (iHS, nSL) and cross-population (Fst, PBS, XP-EHH) based statistics, we found many genetic regions showing evidence of positive selection.

We found that selection statistics computed for outlier variants cannot be explained by neutral forces alone. By intersecting genome-wide-association study data for hundreds of traits measured in Sardinians with publicy available functional genomic databases, we found that autoimmunity-related genes are enriched for these putatively adaptive variants.

Taken together, these results illustrate the importance of characterizing the phenotypic variation within a population, and especially the utility of whole-genome-sequence data, when proposing and interpreting genetic signatures of positive selection.

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university.

Matteo Floris

Acknowledgements

Of the many people who deserve thanks, some are particularly prominent, such as my supervisor Prof. Francesco Cucca, my colleagues Maristella Steri, Joe Marcus, John Novembre, and all the collaborators at IRGB-CNR; last but not least, I want to dedicate this work to my fantastic family: my wife Vera and my children Lorenzo, Riccardo and Nora, and my parents. Special thanks are due to Vera, a thoughtful mother and wonderful wife.

Contents

Chapter 1.

Introduction

"Without variability, nothing can be effected; slight individual differences, however, suffice for the work, and are probably the chief or sole means in the production of new species."

— Charles Darwin, *The variation of animals and plants under domestication*, 1868

1.1. What is evolution?

Evolution is the progressive accumulation of modifications that in a sufficiently large time span leads to significant changes in living organisms; evolution usually acts through two modes at different time scales [1]:

1. **microevolution**, or *adaptation*: with this term, we mean all the evolutionary processes that give rise to observable changes in gene frequencies in populations of a particular species, and which can be both seen in nature and experimentally replicated;

2. **macroevolution**, or *speciation*: this definitions comprises the arising of divisions in the taxonomic hierarchy of species, as well as the development of complex organs; unlike microevolution, macroevolution is not experimentally reproducible because of the extremely long time scale required for speciation.

Thus, evolution is therefore a phenomenon which manifests itself at different levels of complexity: from the increase in frequency of a particular DNA mutation, to the flow of processes that in the Upper Triassic leaded to the diversification of theropod dinosaurs into birds during the Jurassic (around 165-150 million years ago) [2]. In the first case an event occurs on a single population scale, while in the second case more species are involved. These two extreme cases are classic examples of micro- and macroevolution[1].

Evolution is sometimes misunderstood with natural selection, but the two concepts are not the same. Natural selection refers indeed to any process that can cause evolutionary change, but natural selection can occur without producing evolutionary change. Conversely, processes other than natural selection can lead to evolution.

The modern view of the evolution is based on the theoretical cathedral of Charles Darwin[2], who postulated the evolution of species through natural selection, combined with Gregor Mendel's theory of biological inheritance[3].

Darwin was the first to appreciate clearly that evolution depends on the existence of heritable variation within a species and acts to generate the differences between ancestral populations and descendants.

Most of the Darwin's contribution, starting from *The Origin of Species*, was intended to describe how a wide range of biological phenomena could be interpreted in terms of evolution by natural selection.

Darwin unfortunately was not aware of Mendel's work, despite its publication few years earlier. Mendel's work has permanently revolutionized our understanding of inheritance: his ability to solve the mechanism of inheritance is based on the use of a unique approach which combines rigorous experiments with quantitative probabilistic evaluations about the expected results: in other words, he used biological data to verify a quantitative hypothesis.

A detailed overview of these two fundamental theories is not the focus of this thesis.

[1] Suggested reading: http://evolution.berkeley.edu/evolibrary/article/evo_39
[2] See http://darwin-online.org.uk/converted/pdf/1861_OriginNY_F382.pdf
[3] See http://www.mendelweb.org/Mendel.html

1.2. The Hardy-Weinberg equilibrium and its assumptions

The evolution is defined with respect to a situation of equilibrium, or to a reference standard, that in genetics is summarized by the *Hardy-Weinberg principle*. The law of Hardy-Weinberg is a mathematical expression that predict the expected genotype frequencies in a new generation, given the allele frequencies of the initial population.

Suppose we have a biallelic locus: the **A** allele has a frequency of p in our population, while the **a** allele has a frequency q; independently, Hardy and Weinberg [3] [4] demonstrated that the genotype frequencies expected in the next generation can be predicted by following formulas:

$$AA = p^2 \tag{1.1}$$

$$Aa = 2pq \tag{1.2}$$

$$aa = q^2 \tag{1.3}$$

Therefore, if for instance in the starting population (the gene pool of the parents) the frequencies of alleles **A** and **a** are $p = 0.7$ and $q = 0.3$, then it follows that the expected frequencies of genotypes are equal to

$$AA = p^2 = 0.7^2 = 0.49 \tag{1.4}$$

$$Aa = 2pq = 2 * 0.7 * 0.3 = 0.42 \tag{1.5}$$

$$aa = q^2 = 0.3 * 0.3 = 0.09 \tag{1.6}$$

One useful application of this law resides in the ability to predict how many children will be born with a genetic disease caused by a recessive allele present in the population at the rate of 0.01% (ie $0.01^2 = 0.0001$, that is, 1 out of 10,000 births).

From the evolutionary point of view an important aspect, however, is that if the Hardy-Weinberg equilibrium is mantained through generations, this indicates the **absence** of evolutionary forces. Therefore, such law of equilibrium defines a starting point to explain a departure from a stationary situation: deviations from the equilibrium indicate the **presence** of evolutionary forces.

1.3. Evolution and violations of Hardy-Weinberg equilibrium

There are a number of key assumptions which ensure the Hardy-Weinberg equilibrium is not violated.[4] The first is that *in a population the mating between individuals is entirely governed by chance*. This assumption in fact means that there is no preference for a genotype with respect to another, and it can be actually violated in two ways: the first is the **inbreeding**, which is the mating between closely related individuals, and the second is the **choice of partner on the basis of the phenotype** (and consequently of the underlying genotype). In this last case we are in presence of positive assortative mating if the choice of the partner is made on the basis of phenotypic similarities (such as height, skin color, etc).

Both in the case of inbreeding than in presence of positive assortative mating, the population will experience a deviation from the Hardy-Weinberg equilibrium in terms of increase in the relative number of homozygotes. This means that there will be some change in the genotype frequencies.

Other Hardy-Weinberg equilibrium assumptions can be violated, and all of these will cause a change of allelic frequencies. These assumptions coincide exactly with the *absence of the forces that drive evolution*: indeed, despite the obvious differences between micro- and macroevolution, the deviation from equilibrium, or the evolution in itself, it is always determined by four forces, or mechanisms which, in the particular case of microevolution, are the leading causes of changes in allele frequencies over time:

1. **mutation**, the molecular event that causes any change of the DNA sequence; the mutation is always the initial event of each genetic variation, on which the other three forces act in a positive or negative direction;

2. **gene flow**, the transfer of alleles or genes from one population to another, as a result of human migration from one territory to another;

3. **natural selection**, which occurs when a given genotype or genotypes influence the fitness, by increasing or decreasing the reproductive capacity or survival of the carriers within a population;

[4] Suggested reading: John H. Relethford, *Human Population Genetics*, April 2012, Wiley-Blackwell.

4. **genetic drift**, a stochastic (random) force that can scramble the predictable effects of selection, mutation, and gene flow. Even if the contribution of a random force in the evolutionary process could be underestimated, genetic drift is an extremely important factor in evolution.

1.3.1. Mutations

A mutation is a permanent change in the DNA sequence, not corrected by the DNA quality control and repair mechanisms. Mutations can have different impacts on the original DNA sequence: can involve a single base or large chromosomal segments, and to include many genes simultaneously.

Mutations are classified into two broad categories:

1. **germline mutations**: they are inherited from one parent and are therefore present throughout the life of an individual theoretically in all body cells. These mutations are therefore present in the egg and/or sperm from which they originate an individual.

2. and **acquired somatic mutations**: occur randomly or due to environmental factors (such as solar radiation or pollutants), of chemical endogenous factors or in case of errors in the processes of replication and repair of DNA. If these mutations do affect somatic cells, they can not be transmitted to the progeny; if instead they affect germ cells, then they will be passed to the next generation.

Therefore new molecular events that cause the appearance of a change in the DNA sequence can be either inherited or somatic. If a mutation appears later during the early stages of embryonic development, it may experience a situation called *mosaicism*, where only tissues that derive from the cell in which it was presents this first mutations will display some damage, unlike the other tissues derived from the remaining embryonal cells.

The majority of pathogenic gene mutations are uncommon in the general population, while the non-pathogenic or non-deleterious mutations can reach higher frequencies. When the frequency of a mutation exceeds 1%, we define them as polymorphisms, or common variants. These are generally responsible for the phenotypic differences observed between individuals, such as height, color of skin or eyes, but they can also influence the risk of developing certain diseases later in life.

According to one school of thought in population genetics, known as the theory of neutrality [5] [6], the majority of mutations are neutral from a selective point of view, because they do not affect the fitness of the carriers. Sometimes new mutations may increase in frequency in a population due to random events, even if they do not confer any benefit on the carriers. The process of increasing allele frequencies due to random phenomena is known as genetic drift (see below). A second school of thought argues that much of the variability observed was in some way subject to Darwinian selection as capable of modifying the fitness.

From the molecular point of view, the DNA sequence of a gene can be altered in a number of ways[5]. Gene mutations have varying effects on health, depending on where they occur and whether they alter the function of essential proteins. The classes of mutations include:

- **Missense mutation** This type of mutation is a change in one DNA base pair that results in the substitution of one aminoacid for another in the protein coded by a gene.
- **Nonsense mutation** A nonsense mutation is another change in the coding part of DNA. Instead of substituting one aminoacid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.
- **Insertions** An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.
- **Deletions** A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes.
- **Duplication** A duplication consists of a piece of DNA that is abnormally copied one or more times.
- **Frameshift mutation** This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the

[5] See https://ghr.nlm.nih.gov/primer/mutationsanddisorders/possiblemutations for a primer about mutations.

grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

- **Repeat expansion** Nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated. This type of mutation can cause the resulting protein to function improperly.

1.3.2. Genetic drift

As mutations, also genetic drift is a random process: in this case from one generation to another we can observe fluctuations of allele frequencies due purely to chance. All cases of genetic drift originate from an event evolutionary equivalent to a sampling error, but there are different ways according to which, in natural populations, a sampling error can occur. First, a genetic drift can occur when *the size of the population remains constantly small for an extended period of time*: this situation is without doubt frequent, especially where populations occupy marginal habitat or when the competition limits the growth of the population.

A genetic drift can also occur in another way, through the so called *principle of the founder*: this happens when a population origins from a small number of individuals. Although a population may subsequently grow in size and later to count a large number of individuals, its gene pool derives only from genes owned by the founders. Such event can play an important role in determining which genes were present in the founding population, and this has profound effects on the gene pool of the next generation.

A third form of genetic drift that may occur in a population is the effect of *bottleneck*. This effect occurs when a population undergoes a drastic reduction in the number of individuals, during which some of the pool genes can be lost for pure effect of chance. The bottleneck effect can also be considered in terms of a kind of founder effect, given that the population is re-established starting from the few survivors to the sudden reduction.

1.3.3. Genetic flow

Gene flow is the dissemination of genes between populations, due to migration of individuals with reproductive age, and has two main effects on the populations:

- it can change the allele frequencies of the receiving population;
- it can contribute with new genes to the gene pool of the receiving population (favoring the dispersion of unique alleles and acting as a source of genetic variability alternative to the onset of new mutations, generally rare event).

Ultimately, the overall effect of the gene flow is to increase the number of polymorphisms of a population and, at the same time, to reduce the average genetic differences between the populations. This exchange of genes becomes a unifying force that tends to prevent that will populations differ from a genetic point of view.

1.3.4. Natural selection

There are many different ways to define instances of natural selection. According to a model proposed by Nielsen [7], starting from a simple model in which only biallelic loci are considered (with alleles **A** and **a**), a selective pressure is present when the fitness of the three possible genotypes is not identical:

- in particular we can see *directional selection* if the fitness of heterozygotes is intermediate compared to that of homozygotes; directional selection in fact tends to eliminate the variability within a population;
- we refer to *overdominance* when in fact the heterozygote has a fitness greater of the three possible genotypes. The *balancing selection* is a special case of overdominance, in which the variability is maintained in the population.

A classical distinction of selection types is the following:

1. *negative (or purifying) selection*: it happens when new mutations are eliminated from the population because they are deleterious for reproduction or for survival;
2. *positive selection* is the kind of selection that favors new advantageous mutations (therefore mutations with positive selection coefficient);

3. *disruptive selection* favors two extreme phenotypes simultaneously, thus tends to enhance the overall variability.

When a new mutation does not change the fitness of an individual in which it appears, it is said to be neutral. In general, the *neutrality condition* is one in which the loci under consideration are not affected by natural selection.

One of the purposes of population genetics is to distinguish the variability share under selection (in particular the positive one) from the neutral.

1.4. Signs of natural selection in human populations

A long series of studies have tried to identify the effects of natural selection in human populations, with particular emphasis to the positive selection.

1.4.1. HbS and malaria

There are at least 300 million acute cases of malaria each year globally, with more than a million deaths. Ninety percent of deaths due to malaria occur in sub-Saharan Africa, and most occur in young children. Among the four types of pathogens of malaria (Plasmodium vivax human, Plasmodium malariae, Plasmodium ovale, Plasmodium falciparum), P. falciparum is the most common in Africa, and is the major cause of mortality in sub-Saharan Africa.

P. falciparum has had profound effects on human evolution, as evidenced by the high rates of protection to malaria mutations observed in populations of historically malarial regions. Many of the protective variants are able to modify the surface of human erythrocytes, where the malaria parasite lives a crucial stage of its life cycle. Many of these mutations affect the genes which code for the hemoglobin gene. Despite the protection against malaria, these hemoglobin mutations can also cause genetic blood disorders known as hemoglobinopathies [8].

The first hints of a possible relationship between the extent and prevalence of hemoglobinopathies in some areas of the Mediterranean and malaria infection date back to the late '40s, with the work of Haldane [9]. It was later widely recognized that genetic disorders of red cells, such as thalassemia, sickle cell disease (SCD) and glucose-

6-phosphate dehydrogenase deficiency (G6PD) may confer resistance to Plasmodium infection [10].

In sickle cell anemia a single amino acid substitution in HbA results in HbS, and red blood cells are hard, sticky, and crescent-shaped [11].

While a geographical correspondence between the distribution of thalassemia and Malaria was confirmed in the Mediterranean region, as well as in other places, a similar relationship between Hb and malaria has been discovered in Africa. As for the SCD it has been suggested that, while homozygous individuals for sickle-cell allele usually die before adulthood, the gene responsible for SCD could reach high frequencies because of the resistance against malaria given by the state of heterozygous carrier, resulting in a balanced polymorphism; indeed, it has been observed that it is quite difficult for Plasmodium falciparum to develop into red blood cells HbS-contaning and is also rare to find a HbS carrier struck by cerebral malaria, a common cause of death in this disease [12].

Only recently, however, extensive geostatistical studies provided the first quantitative evidence of a geographical link between the global distribution of HbS and endemic malaria: there was a strong link between the higher frequencies allele HbS and high endemicity of malaria scale overall, but this observation is based primarily on data in Africa. The gradual increase in HBS allele frequencies from epidemic areas in endemic areas in Africa is consistent with the hypothesis that protection from malaria HbS involves the enhancement of both innate and acquired immunity to P. falciparum; interactions with hemoglobin C may explain the allele frequencies lower Hb in West Africa [12].

Many other genetic factors may be involved in the protection mechanism [13]. A number of studies have focused on different components of the red blood cell membrane and particular attention has been given to the role of the complement regulatory proteins, particularly complement receptor 1 (CR1). It has actually been shown that, in common with other genes malaria protection, the frequencies of several of CR1 polymorphisms are higher in a number of malaria endemic areas: for example, Cockburn et al. have shown that polymorphisms resulting in low expression CR1 are associated with alpha thalassemia in malaria-endemic regions of Papua New Guinea, offering protection from severe malaria by P. falciparum. Complement receptor-1 polymorphism associated with resistance to severe malaria was also demonstrated in Kenya.

Studies on other genetic factors have produced important results: for example, a recent multi-centre genome-wide association study (GWAS) of life-threatening P. falciparum infection (severe malaria) in over 11,000 African children with replication data in a further 14,000 individuals highlighted the role of a novel malaria resistance locus close to a cluster of genes encoding glycophorins that are receptors for erythrocyte invasion by P. falciparum [14].

Another line of research focused on the human protein CD36, which is an important receptor for P. falciparum in red blood cells (RBC) [15]. The scavenger receptor CD36 plays important roles in malaria, including the sequestration of parasite-infected erythrocytes in microvascular capillaries, control of parasitemia through phagocytic clearance by macrophages, and immunity. Several lines of evidence suggest that mutations in CD36 are protective against malaria [16]: mutations in the promoters and within introns and in exon 5 reduce the risk of severe malaria. Gene diversity studies suggest there has been positive selection on this gene presumably due to malarial selection pressure [17].

1.4.2. Lactase persistance

Lactase persistance (LP) [18] refers to a an autosomal dominant trait causing a continued expression of the lactase-phlorizin hydrolase (LPH), encoded by the LCT (lactase) gene. LP is found at high frequency in people of Northern European ancestry and populations in East Africa, Middle East, and South/Central Asia who traditionally practiced pastoralism and regularly consumed milk and other dairy products as adults.

Based on this association, LP was hypothesized to confer a selective advantage because consumption of fresh milk and other dairy products allowed for efficient caloric intake, calcium assimilation in high latitude, or increased water absorption from milk in arid environments.

Multiple genetic variants associated with LP have been found in different populations. These variants harbor signatures of recent positive selection, increase enhancer function [19], and are located in binding sites for major transcription factors in intestinal epithelia such as Oct-1, HNF1α and HNF4α. Interestingly, all of them are found within 100 bp of each other, suggesting a simple architecture for LP with a small mutational target size. Each of these variants is associated with simple haplotype patterns, consistent with single mutational events [20].

1.4.3. Inuit and tibetans

The Greenland natives, called Inuit, have long been exposed to low annual temperatures typical of the Artic, and lived on a traditional diet rich in protein mostly from marine mammals with high levels of omega-3 fatty acids. In an effort to understand how the Inuit adapted to such conditions, a group lead by Rasmus Nielsen at UC Berkeley analyzed the genomes of 191 Greenlanders with low European ancestry and compared them to the genomes of 60 Europeans and 44 Han Chinese [21]. They identified mutations occurring in a large fraction of Inuit individuals but rarely or not at all in other populations. These mutations are therefore likely to have spread throughout the Inuit because they were essential for their survival.

One cluster of these mutations was located in genes that code for enzymes that desaturate carbon-carbon bonds in fatty acids. Those genetic mutations, found in nearly 100 percent of the Inuit, are found in only 2 percent of Europeans and 15 percent of Han Chinese.

These mutations downregulate the production of omega-3 and omega-6 polyunsaturated fatty acids, possibly to account for the high intake of fatty acids from the Inuit diet. They also have other effects, as they lower *bad* LDL cholesterol, fasting insulin levels, and height, as growth is partly regulated by the fatty acid profile. Notably, the association with height was replicated in Europeans.

The same group also identified another common mutation in the Inuit located in a gene involved in the differentiation of brown and brite fat cells. As the latter is responsible for heat generation, this mutation may have helped the Inuit adapt to a cold environment.

A similar paper, previously demonstrated in residents of the Tibetan Plateau signs heritable adaptations to extreme altitude, through sequencing of 50 exomes of ethnic Tibetans, encompassing coding sequences of 92% of human genes, with an average coverage of 18x per individual [22]. Genes showing population-specific allele frequency changes, which represent strong candidates for altitude adaptation, were identified. The strongest signal of natural selection came from endothelial Per-Arnt-Sim (PAS) domain protein 1 (EPAS1), a transcription factor involved in response to hypoxia. One single-nucleotide polymorphism (SNP) at EPAS1 shows a 78% frequency difference between Tibetan and Han samples, representing the fastest allele frequency change observed at any human gene to date. This SNP's association with erythrocyte

abundance supports the role of EPAS1 in adaptation to hypoxia. Thus, a population genomic survey has revealed a functionally important locus in genetic adaptation to high altitude.

Chapter 2.

Recent positive selection in human populations

"Malaria does not ask permission before coming in"
— Malawi saying

2.1. Natural selection

The fact that in the remote past entire masses of human beings had moved to different habitats of the planet, it has meant that our ancestors have partially modified their diets, have encountered new pathogens and had been forced to adapt to new environments. In 1858 Darwin and Wallace defined in a conceptual framework the principles of natural selection, by asserting that those traits that make individuals more adapted to survival and reproduction tend to be more common in a population[1].

The first description of human adaptation has been made by Haldane [9], who observed the occurrence of hematological diseases in areas where malaria was endemic. The observational Haldane hypothesis was later confirmed by Allison, who demonstrated that mutations in the hemoglobin B gene were the molecular determinant of resistance to malaria [23].

[1] Suggested reading: https://ca1-tls.edcdn.com/documents/Special-Issue-9-Survival-of-the-Fittest.pdf?mtime=20160213060318

After this first observation, evolutionary genetics has identified several traits linked to adaptation (from lactase persistence to the coloration of human skin).

From a theoretical point of view, the approach followed in these cases is known as *forward genetics* [24], because first it is assumed that a phenotype is under selection and subsequently are identified all the loci that regulate such phenotype.

In more recent times, with the availability of genomic analysis techniques, it became possible to perform agnostic searches on the entire human genome, aimed to identify a large number of loci with evidences of selection. This transition from a scientific approach that arises from the assumptions (*hypothesis-testing*) to one that is born from the data (*data-driven*) has been made feasible by the simultaneous availability of more information on the functionality of a large part of the human genome and more sophisticated techniques for the identification of genomic signature of positive selection.

The description of loci under selection is not only important from a theoretical point of view, but also for the understanding of the mechanisms of pathogens resistance and, ultimately, also for the study of new therapies.

Natural selection is based on the principle that traits that increase fitness (ie the ability to reproduce or adapt of an organism) will be more easily transmitted to offspring, and as a consequence they will tend to be more frequent in given population. From a genetic point of view, natural selection is a not-random phenomenon which leads to the increase in frequency of an allele as a result of its effect on a favorable phenotype.

As previously mentioned, the natural selection can act in different ways:

- in a directional manner, by favoring the spread of an allele in a population: in this case we speak of **positive selection**; if instead the contrary an allele is disadvantageous, it will disappear from a population: this situation is described as **negative selection**. The negative selection acts on causal mutations, which often are deleterious and must be removed from a population especially if they modify genomic regions with important functions for the life: in this case we refer to **background selection**, which is more stringent and effective for regions very conserved along the genome.
- in a way to maintain more alleles in the same locus at appreciable frequencies, in case of **balancing selection**; this usually happens when the heterozygotes for the

alleles under consideration have a higher adaptive value than the homozygote (heterozygote advantage). In the case two alleles with opposite phenotypic effects are simultaneously favoured, the event is described as **disruptive selection**; when intermediate phenotypes are favoured, then one speaks of **stabilizing selection**.

While the negative selection usually targets functionally relevant areas of the genome and the balancing selection has more subtle effects on the genome, the positive selection has the peculiarity of leaving clearly visible signs on the genome, so that many approaches have been developed to identify the loci under selection.

When a single copy of a new beneficial allele appears in a population, it can rise rapidly to fixation: this event is known as *hard sweep*. Such classic or hard selective sweeps reduce genetic variation and increase haplotype lengths around the selected site because the time to the most recent common ancestor is shorter at that site [25] [26].

The term *soft sweep* was introduced to describe two different scenarios. In the first scenario, an allele that is already segregating in the population (standing variation) becomes selectively favored due to a new selection pressure, and sweeps up in frequency. In the second scenario, multiple independent mutations at a single locus are all favored and all increase in frequency simultaneously.

Thus, in general, selective sweeps can be *hard*, where a single adaptive allele sweeps through the population, or *soft*, when multiple adaptive alleles sweep through the population at the same time.

2.2. Statistical approaches to detect signatures of positive selection in the human genome

The methods developed to date to detect signatures of positive selection can be divided into three main categories [27]:

- **population differentiation based** tests, based on the concept that when an allele is highly impacting on a trait under selection, it will reach a very high frequency in a population, togheter with other variants in linkage disequilibrium;

- **linkage disequilibrium based** tests, aimed to identify haplotypes that achieved wider prevalence in a population; such haplotypes are preserved and persistent because not yet corroded by recombination events;
- **site frequency spectrum** based tests: site frequency spectrum (SFS) describe the distribution of allele frequencies across sites in the genome of a particular species. Near the point of fixation, the scaled SFS is characterized by an abundance of very high frequency alleles, and a near-absence of intermediate frequency alleles;
- **functional annotation** based tests, based on the assumption that only functionally relevant loci can be positively selected, while non-functional variants should be neutral and their frequencies could be affected only by genetic drift or demographic factors.

2.3. Population differentiation based tests

When a population moves to a new environment, a process of adaptation may take place, and positive selection may act on mutations that help the individuals to better adapt to their new environment. Human populations moving to different parts of the world have experienced distinct climates and natural resources. Therefore, some genetic changes have been favored in one particular population but not the others. If one or more alleles at a particular genomic locus have highly differentiated frequencies in different populations, or are even population-specific, positive selection may have acted on that particular locus in one or more populations. The fixation index (F_{ST}), first introduced by Wright, is often used to estimate population differentiation [28] [29] [30].

The value of F_{ST} ranges from 0 to 1, with a value of 0 implying complete panmixis (i.e. no differentiation), compared with a value of 1 indicating a complete separation between the two populations.

F_{ST} is often used in the detection of selective sweeps, with higher values indicating a higher probability of selection. However, this method is often criticized, as the value of F_{ST} is highly influenced by population structure and demographic history, as well as the ascertainment biases of variant discovery in different population samples. Therefore, F_{ST} values are often evaluated in the context of a genome-wide or multi-locus distribution, as demographic factors or data biases will most likely affect the whole data set equally.

Akey et al. [31] estimated locus-specific F_{ST} compared with genome-wide distribution, and identified over a hundred loci showing "signatures of positive selection" with high levels of differentiation among populations.

Briefly, the value of F_{ST} is higher when most of the variance across the total population can be attributed to the variance between groups, while low F_{ST} indicates that most of the variance is within the groups [32] [31]. If you get a low F_{ST} number then you are not getting much more information by looking at population substructure, but if you have a high F_{ST} substructure might be really informative as genetic differences break relatively cleanly along group differences.

Another test, Population Branch Statistic, or PBS, is able to identify SNPs under selection which are particularly differentiated in a population of interest [22].

Briefly, given a query population A, a closer population B and a distant population C, PBS for a given allele is calculated as follows:

$$pbs = ((-log(1 - F_{ST}\_AB)) + (-log(1 - F_{ST}\_AC)) - (-log(1 - F_{ST}\_BC)))/2 \tag{2.1}$$

2.4. Linkage disequilibrium-based tests

Linkage disequilibrium (LD) refers to the non-random associations of alleles at different loci. There are many factors that can influence the level of LD at a locus in the genome. First of all, the variation of recombination rates causes some loci to be in higher LD than others.

For example, loci within a "cold" recombination region would be more likely to be linked than those within a "hot" recombination region, even if they have similar physical distances. As linkage information is critical for many genetic studies, genetic linkage maps, often known simply as *genetic maps*, have been generated to scale the physical position of genomic variants in terms of recombination frequency. Natural selection, especially positive selection, can have a strong impact on the LD of a locus under selection, and more specifically, will cause the locus to have unusually high LD compared with neutral loci of similar frequency.

As described earlier, if a new mutation increases the fitness in the carriers, the frequency of that advantageous allele will go up rapidly in the population, and finally

Abstract

The recent production of population-scale genomic data offers an unprecedented opportunity to understand how natural selection has shaped human phenotypic variation within populations. To identify signatures of recent positive selection in Sardinia, we used 23 million single nucleotide polymorphisms from low-coverage whole genomes of 3,514 Sardinians along with data from the 1000 Genomes project. Using single-population (iHS, nSL) and cross-population (Fst, PBS, XP-EHH) based statistics, we found many genetic regions showing evidence of positive selection.

We found that selection statistics computed for outlier variants cannot be explained by neutral forces alone. By intersecting genome-wide-association study data for hundreds of traits measured in Sardinians with publicy available functional genomic databases, we found that autoimmunity-related genes are enriched for these putatively adaptive variants.

Taken together, these results illustrate the importance of characterizing the phenotypic variation within a population, and especially the utility of whole-genome-sequence data, when proposing and interpreting genetic signatures of positive selection.

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university.

Matteo Floris

Acknowledgements

Of the many people who deserve thanks, some are particularly prominent, such as my supervisor Prof. Francesco Cucca, my colleagues Maristella Steri, Joe Marcus, John Novembre, and all the collaborators at IRGB-CNR; last but not least, I want to dedicate this work to my fantastic family: my wife Vera and my children Lorenzo, Riccardo and Nora, and my parents. Special thanks are due to Vera, a thoughtful mother and wonderful wife.

Contents

Chapter 1.

Introduction

"Without variability, nothing can be effected; slight individual differences, however, suffice for the work, and are probably the chief or sole means in the production of new species."

— Charles Darwin, *The variation of animals and plants under domestication*, 1868

1.1. What is evolution?

Evolution is the progressive accumulation of modifications that in a sufficiently large time span leads to significant changes in living organisms; evolution usually acts through two modes at different time scales [1]:

1. **microevolution**, or *adaptation*: with this term, we mean all the evolutionary processes that give rise to observable changes in gene frequencies in populations of a particular species, and which can be both seen in nature and experimentally replicated;

2. **macroevolution**, or *speciation*: this definitions comprises the arising of divisions in the taxonomic hierarchy of species, as well as the development of complex organs; unlike microevolution, macroevolution is not experimentally reproducible because of the extremely long time scale required for speciation.

Thus, evolution is therefore a phenomenon which manifests itself at different levels of complexity: from the increase in frequency of a particular DNA mutation, to the flow of processes that in the Upper Triassic leaded to the diversification of theropod dinosaurs into birds during the Jurassic (around 165-150 million years ago) [2]. In the first case an event occurs on a single population scale, while in the second case more species are involved. These two extreme cases are classic examples of micro- and macroevolution[1].

Evolution is sometimes misunderstood with natural selection, but the two concepts are not the same. Natural selection refers indeed to any process that can cause evolutionary change, but natural selection can occur without producing evolutionary change. Conversely, processes other than natural selection can lead to evolution.

The modern view of the evolution is based on the theoretical cathedral of Charles Darwin[2], who postulated the evolution of species through natural selection, combined with Gregor Mendel's theory of biological inheritance[3].

Darwin was the first to appreciate clearly that evolution depends on the existence of heritable variation within a species and acts to generate the differences between ancestral populations and descendants.

Most of the Darwin's contribution, starting from *The Origin of Species*, was intended to describe how a wide range of biological phenomena could be interpreted in terms of evolution by natural selection.

Darwin unfortunately was not aware of Mendel's work, despite its publication few years earlier. Mendel's work has permanently revolutionized our understanding of inheritance: his ability to solve the mechanism of inheritance is based on the use of a unique approach which combines rigorous experiments with quantitative probabilistic evaluations about the expected results: in other words, he used biological data to verify a quantitative hypothesis.

A detailed overview of these two fundamental theories is not the focus of this thesis.

[1] Suggested reading: http://evolution.berkeley.edu/evolibrary/article/evo_39
[2] See http://darwin-online.org.uk/converted/pdf/1861_OriginNY_F382.pdf
[3] See http://www.mendelweb.org/Mendel.html

1.2. The Hardy-Weinberg equilibrium and its assumptions

The evolution is defined with respect to a situation of equilibrium, or to a reference standard, that in genetics is summarized by the *Hardy-Weinberg principle*. The law of Hardy-Weinberg is a mathematical expression that predict the expected genotype frequencies in a new generation, given the allele frequencies of the initial population.

Suppose we have a biallelic locus: the **A** allele has a frequency of p in our population, while the **a** allele has a frequency q; independently, Hardy and Weinberg [3] [4] demonstrated that the genotype frequencies expected in the next generation can be predicted by following formulas:

$$AA = p^2 \tag{1.1}$$

$$Aa = 2pq \tag{1.2}$$

$$aa = q^2 \tag{1.3}$$

Therefore, if for instance in the starting population (the gene pool of the parents) the frequencies of alleles **A** and **a** are $p = 0.7$ and $q = 0.3$, then it follows that the expected frequencies of genotypes are equal to

$$AA = p^2 = 0.7^2 = 0.49 \tag{1.4}$$

$$Aa = 2pq = 2 * 0.7 * 0.3 = 0.42 \tag{1.5}$$

$$aa = q^2 = 0.3 * 0.3 = 0.09 \tag{1.6}$$

One useful application of this law resides in the ability to predict how many children will be born with a genetic disease caused by a recessive allele present in the population at the rate of 0.01% (ie $0.01^2 = 0.0001$, that is, 1 out of 10,000 births).

From the evolutionary point of view an important aspect, however, is that if the Hardy-Weinberg equilibrium is mantained through generations, this indicates the **absence** of evolutionary forces. Therefore, such law of equilibrium defines a starting point to explain a departure from a stationary situation: deviations from the equilibrium indicate the **presence** of evolutionary forces.

1.3. Evolution and violations of Hardy-Weinberg equilibrium

There are a number of key assumptions which ensure the Hardy-Weinberg equilibrium is not violated.[4] The first is that *in a population the mating between individuals is entirely governed by chance*. This assumption in fact means that there is no preference for a genotype with respect to another, and it can be actually violated in two ways: the first is the **inbreeding**, which is the mating between closely related individuals, and the second is the **choice of partner on the basis of the phenotype** (and consequently of the underlying genotype). In this last case we are in presence of positive assortative mating if the choice of the partner is made on the basis of phenotypic similarities (such as height, skin color, etc).

Both in the case of inbreeding than in presence of positive assortative mating, the population will experience a deviation from the Hardy-Weinberg equilibrium in terms of increase in the relative number of homozygotes. This means that there will be some change in the genotype frequencies.

Other Hardy-Weinberg equilibrium assumptions can be violated, and all of these will cause a change of allelic frequencies. These assumptions coincide exactly with the *absence of the forces that drive evolution*: indeed, despite the obvious differences between micro- and macroevolution, the deviation from equilibrium, or the evolution in itself, it is always determined by four forces, or mechanisms which, in the particular case of microevolution, are the leading causes of changes in allele frequencies over time:

1. **mutation**, the molecular event that causes any change of the DNA sequence; the mutation is always the initial event of each genetic variation, on which the other three forces act in a positive or negative direction;

2. **gene flow**, the transfer of alleles or genes from one population to another, as a result of human migration from one territory to another;

3. **natural selection**, which occurs when a given genotype or genotypes influence the fitness, by increasing or decreasing the reproductive capacity or survival of the carriers within a population;

[4] Suggested reading: John H. Relethford, *Human Population Genetics*, April 2012, Wiley-Blackwell.

4. **genetic drift**, a stochastic (random) force that can scramble the predictable effects of selection, mutation, and gene flow. Even if the contribution of a random force in the evolutionary process could be underestimated, genetic drift is an extremely important factor in evolution.

1.3.1. Mutations

A mutation is a permanent change in the DNA sequence, not corrected by the DNA quality control and repair mechanisms. Mutations can have different impacts on the original DNA sequence: can involve a single base or large chromosomal segments, and to include many genes simultaneously.

Mutations are classified into two broad categories:

1. **germline mutations**: they are inherited from one parent and are therefore present throughout the life of an individual theoretically in all body cells. These mutations are therefore present in the egg and/or sperm from which they originate an individual.

2. and **acquired somatic mutations**: occur randomly or due to environmental factors (such as solar radiation or pollutants), of chemical endogenous factors or in case of errors in the processes of replication and repair of DNA. If these mutations do affect somatic cells, they can not be transmitted to the progeny; if instead they affect germ cells, then they will be passed to the next generation.

Therefore new molecular events that cause the appearance of a change in the DNA sequence can be either inherited or somatic. If a mutation appears later during the early stages of embryonic development, it may experience a situation called *mosaicism*, where only tissues that derive from the cell in which it was presents this first mutations will display some damage, unlike the other tissues derived from the remaining embryonal cells.

The majority of pathogenic gene mutations are uncommon in the general population, while the non-pathogenic or non-deleterious mutations can reach higher frequencies. When the frequency of a mutation exceeds 1%, we define them as polymorphisms, or common variants. These are generally responsible for the phenotypic differences observed between individuals, such as height, color of skin or eyes, but they can also influence the risk of developing certain diseases later in life.

According to one school of thought in population genetics, known as the theory of neutrality [5] [6], the majority of mutations are neutral from a selective point of view, because they do not affect the fitness of the carriers. Sometimes new mutations may increase in frequency in a population due to random events, even if they do not confer any benefit on the carriers. The process of increasing allele frequencies due to random phenomena is known as genetic drift (see below). A second school of thought argues that much of the variability observed was in some way subject to Darwinian selection as capable of modifying the fitness.

From the molecular point of view, the DNA sequence of a gene can be altered in a number of ways[5]. Gene mutations have varying effects on health, depending on where they occur and whether they alter the function of essential proteins. The classes of mutations include:

- **Missense mutation** This type of mutation is a change in one DNA base pair that results in the substitution of one aminoacid for another in the protein coded by a gene.
- **Nonsense mutation** A nonsense mutation is another change in the coding part of DNA. Instead of substituting one aminoacid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.
- **Insertions** An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.
- **Deletions** A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes.
- **Duplication** A duplication consists of a piece of DNA that is abnormally copied one or more times.
- **Frameshift mutation** This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the

[5] See https://ghr.nlm.nih.gov/primer/mutationsanddisorders/possiblemutations for a primer about mutations.

grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

- **Repeat expansion** Nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated. This type of mutation can cause the resulting protein to function improperly.

1.3.2. Genetic drift

As mutations, also genetic drift is a random process: in this case from one generation to another we can observe fluctuations of allele frequencies due purely to chance. All cases of genetic drift originate from an event evolutionary equivalent to a sampling error, but there are different ways according to which, in natural populations, a sampling error can occur. First, a genetic drift can occur when *the size of the population remains constantly small for an extended period of time*: this situation is without doubt frequent, especially where populations occupy marginal habitat or when the competition limits the growth of the population.

A genetic drift can also occur in another way, through the so called *principle of the founder*: this happens when a population origins from a small number of individuals. Although a population may subsequently grow in size and later to count a large number of individuals, its gene pool derives only from genes owned by the founders. Such event can play an important role in determining which genes were present in the founding population, and this has profound effects on the gene pool of the next generation.

A third form of genetic drift that may occur in a population is the effect of *bottleneck*. This effect occurs when a population undergoes a drastic reduction in the number of individuals, during which some of the pool genes can be lost for pure effect of chance. The bottleneck effect can also be considered in terms of a kind of founder effect, given that the population is re-established starting from the few survivors to the sudden reduction.

1.3.3. Genetic flow

Gene flow is the dissemination of genes between populations, due to migration of individuals with reproductive age, and has two main effects on the populations:

- it can change the allele frequencies of the receiving population;
- it can contribute with new genes to the gene pool of the receiving population (favoring the dispersion of unique alleles and acting as a source of genetic variability alternative to the onset of new mutations, generally rare event).

Ultimately, the overall effect of the gene flow is to increase the number of polymorphisms of a population and, at the same time, to reduce the average genetic differences between the populations. This exchange of genes becomes a unifying force that tends to prevent that will populations differ from a genetic point of view.

1.3.4. Natural selection

There are many different ways to define instances of natural selection. According to a model proposed by Nielsen [7], starting from a simple model in which only biallelic loci are considered (with alleles **A** and **a**), a selective pressure is present when the fitness of the three possible genotypes is not identical:

- in particular we can see *directional selection* if the fitness of heterozygotes is intermediate compared to that of homozygotes; directional selection in fact tends to eliminate the variability within a population;
- we refer to *overdominance* when in fact the heterozygote has a fitness greater of the three possible genotypes. The *balancing selection* is a special case of overdominance, in which the variability is maintained in the population.

A classical distinction of selection types is the following:

1. *negative (or purifying) selection*: it happens when new mutations are eliminated from the population because they are deleterious for reproduction or for survival;
2. *positive selection* is the kind of selection that favors new advantageous mutations (therefore mutations with positive selection coefficient);

3. *disruptive selection* favors two extreme phenotypes simultaneously, thus tends to enhance the overall variability.

When a new mutation does not change the fitness of an individual in which it appears, it is said to be neutral. In general, the *neutrality condition* is one in which the loci under consideration are not affected by natural selection.

One of the purposes of population genetics is to distinguish the variability share under selection (in particular the positive one) from the neutral.

1.4. Signs of natural selection in human populations

A long series of studies have tried to identify the effects of natural selection in human populations, with particular emphasis to the positive selection.

1.4.1. HbS and malaria

There are at least 300 million acute cases of malaria each year globally, with more than a million deaths. Ninety percent of deaths due to malaria occur in sub-Saharan Africa, and most occur in young children. Among the four types of pathogens of malaria (Plasmodium vivax human, Plasmodium malariae, Plasmodium ovale, Plasmodium falciparum), P. falciparum is the most common in Africa, and is the major cause of mortality in sub-Saharan Africa.

P. falciparum has had profound effects on human evolution, as evidenced by the high rates of protection to malaria mutations observed in populations of historically malarial regions. Many of the protective variants are able to modify the surface of human erythrocytes, where the malaria parasite lives a crucial stage of its life cycle. Many of these mutations affect the genes which code for the hemoglobin gene. Despite the protection against malaria, these hemoglobin mutations can also cause genetic blood disorders known as hemoglobinopathies [8].

The first hints of a possible relationship between the extent and prevalence of hemoglobinopathies in some areas of the Mediterranean and malaria infection date back to the late '40s, with the work of Haldane [9]. It was later widely recognized that genetic disorders of red cells, such as thalassemia, sickle cell disease (SCD) and glucose-

6-phosphate dehydrogenase deficiency (G6PD) may confer resistance to Plasmodium infection [10].

In sickle cell anemia a single amino acid substitution in HbA results in HbS, and red blood cells are hard, sticky, and crescent-shaped [11].

While a geographical correspondence between the distribution of thalassemia and Malaria was confirmed in the Mediterranean region, as well as in other places, a similar relationship between Hb and malaria has been discovered in Africa. As for the SCD it has been suggested that, while homozygous individuals for sickle-cell allele usually die before adulthood, the gene responsible for SCD could reach high frequencies because of the resistance against malaria given by the state of heterozygous carrier, resulting in a balanced polymorphism; indeed, it has been observed that it is quite difficult for Plasmodium falciparum to develop into red blood cells HbS-contaning and is also rare to find a HbS carrier struck by cerebral malaria, a common cause of death in this disease [12].

Only recently, however, extensive geostatistical studies provided the first quantitative evidence of a geographical link between the global distribution of HbS and endemic malaria: there was a strong link between the higher frequencies allele HbS and high endemicity of malaria scale overall, but this observation is based primarily on data in Africa. The gradual increase in HBS allele frequencies from epidemic areas in endemic areas in Africa is consistent with the hypothesis that protection from malaria HbS involves the enhancement of both innate and acquired immunity to P. falciparum; interactions with hemoglobin C may explain the allele frequencies lower Hb in West Africa [12].

Many other genetic factors may be involved in the protection mechanism [13]. A number of studies have focused on different components of the red blood cell membrane and particular attention has been given to the role of the complement regulatory proteins, particularly complement receptor 1 (CR1). It has actually been shown that, in common with other genes malaria protection, the frequencies of several of CR1 polymorphisms are higher in a number of malaria endemic areas: for example, Cockburn et al. have shown that polymorphisms resulting in low expression CR1 are associated with alpha thalassemia in malaria-endemic regions of Papua New Guinea, offering protection from severe malaria by P. falciparum. Complement receptor-1 polymorphism associated with resistance to severe malaria was also demonstrated in Kenya.

Studies on other genetic factors have produced important results: for example, a recent multi-centre genome-wide association study (GWAS) of life-threatening P. falciparum infection (severe malaria) in over 11,000 African children with replication data in a further 14,000 individuals highlighted the role of a novel malaria resistance locus close to a cluster of genes encoding glycophorins that are receptors for erythrocyte invasion by P. falciparum [14].

Another line of research focused on the human protein CD36, which is an important receptor for P. falciparum in red blood cells (RBC) [15]. The scavenger receptor CD36 plays important roles in malaria, including the sequestration of parasite-infected erythrocytes in microvascular capillaries, control of parasitemia through phagocytic clearance by macrophages, and immunity. Several lines of evidence suggest that mutations in CD36 are protective against malaria [16]: mutations in the promoters and within introns and in exon 5 reduce the risk of severe malaria. Gene diversity studies suggest there has been positive selection on this gene presumably due to malarial selection pressure [17].

1.4.2. Lactase persistance

Lactase persistance (LP) [18] refers to a an autosomal dominant trait causing a continued expression of the lactase-phlorizin hydrolase (LPH), encoded by the LCT (lactase) gene. LP is found at high frequency in people of Northern European ancestry and populations in East Africa, Middle East, and South/Central Asia who traditionally practiced pastoralism and regularly consumed milk and other dairy products as adults.

Based on this association, LP was hypothesized to confer a selective advantage because consumption of fresh milk and other dairy products allowed for efficient caloric intake, calcium assimilation in high latitude, or increased water absorption from milk in arid environments.

Multiple genetic variants associated with LP have been found in different populations. These variants harbor signatures of recent positive selection, increase enhancer function [19], and are located in binding sites for major transcription factors in intestinal epithelia such as Oct-1, HNF1α and HNF4α. Interestingly, all of them are found within 100 bp of each other, suggesting a simple architecture for LP with a small mutational target size. Each of these variants is associated with simple haplotype patterns, consistent with single mutational events [20].

1.4.3. Inuit and tibetans

The Greenland natives, called Inuit, have long been exposed to low annual temperatures typical of the Artic, and lived on a traditional diet rich in protein mostly from marine mammals with high levels of omega-3 fatty acids. In an effort to understand how the Inuit adapted to such conditions, a group lead by Rasmus Nielsen at UC Berkeley analyzed the genomes of 191 Greenlanders with low European ancestry and compared them to the genomes of 60 Europeans and 44 Han Chinese [21]. They identified mutations occurring in a large fraction of Inuit individuals but rarely or not at all in other populations. These mutations are therefore likely to have spread throughout the Inuit because they were essential for their survival.

One cluster of these mutations was located in genes that code for enzymes that desaturate carbon-carbon bonds in fatty acids. Those genetic mutations, found in nearly 100 percent of the Inuit, are found in only 2 percent of Europeans and 15 percent of Han Chinese.

These mutations downregulate the production of omega-3 and omega-6 polyunsaturated fatty acids, possibly to account for the high intake of fatty acids from the Inuit diet. They also have other effects, as they lower *bad* LDL cholesterol, fasting insulin levels, and height, as growth is partly regulated by the fatty acid profile. Notably, the association with height was replicated in Europeans.

The same group also identified another common mutation in the Inuit located in a gene involved in the differentiation of brown and brite fat cells. As the latter is responsible for heat generation, this mutation may have helped the Inuit adapt to a cold environment.

A similar paper, previously demonstrated in residents of the Tibetan Plateau signs heritable adaptations to extreme altitude, through sequencing of 50 exomes of ethnic Tibetans, encompassing coding sequences of 92% of human genes, with an average coverage of 18x per individual [22]. Genes showing population-specific allele frequency changes, which represent strong candidates for altitude adaptation, were identified. The strongest signal of natural selection came from endothelial Per-Arnt-Sim (PAS) domain protein 1 (EPAS1), a transcription factor involved in response to hypoxia. One single-nucleotide polymorphism (SNP) at EPAS1 shows a 78% frequency difference between Tibetan and Han samples, representing the fastest allele frequency change observed at any human gene to date. This SNP's association with erythrocyte

abundance supports the role of EPAS1 in adaptation to hypoxia. Thus, a population genomic survey has revealed a functionally important locus in genetic adaptation to high altitude.

Chapter 2.

Recent positive selection in human populations

"Malaria does not ask permission before coming in"
— Malawi saying

2.1. Natural selection

The fact that in the remote past entire masses of human beings had moved to different habitats of the planet, it has meant that our ancestors have partially modified their diets, have encountered new pathogens and had been forced to adapt to new environments. In 1858 Darwin and Wallace defined in a conceptual framework the principles of natural selection, by asserting that those traits that make individuals more adapted to survival and reproduction tend to be more common in a population[1].

The first description of human adaptation has been made by Haldane [9], who observed the occurrence of hematological diseases in areas where malaria was endemic. The observational Haldane hypothesis was later confirmed by Allison, who demonstrated that mutations in the hemoglobin B gene were the molecular determinant of resistance to malaria [23].

[1] Suggested reading: https://ca1-tls.edcdn.com/documents/Special-Issue-9-Survival-of-the-Fittest.pdf?mtime=20160213060318

After this first observation, evolutionary genetics has identified several traits linked to adaptation (from lactase persistence to the coloration of human skin).

From a theoretical point of view, the approach followed in these cases is known as *forward genetics* [24], because first it is assumed that a phenotype is under selection and subsequently are identified all the loci that regulate such phenotype.

In more recent times, with the availability of genomic analysis techniques, it became possible to perform agnostic searches on the entire human genome, aimed to identify a large number of loci with evidences of selection. This transition from a scientific approach that arises from the assumptions (*hypothesis-testing*) to one that is born from the data (*data-driven*) has been made feasible by the simultaneous availability of more information on the functionality of a large part of the human genome and more sophisticated techniques for the identification of genomic signature of positive selection.

The description of loci under selection is not only important from a theoretical point of view, but also for the understanding of the mechanisms of pathogens resistance and, ultimately, also for the study of new therapies.

Natural selection is based on the principle that traits that increase fitness (ie the ability to reproduce or adapt of an organism) will be more easily transmitted to offspring, and as a consequence they will tend to be more frequent in given population. From a genetic point of view, natural selection is a not-random phenomenon which leads to the increase in frequency of an allele as a result of its effect on a favorable phenotype.

As previously mentioned, the natural selection can act in different ways:

- in a directional manner, by favoring the spread of an allele in a population: in this case we speak of **positive selection**; if instead the contrary an allele is disadvantageous, it will disappear from a population: this situation is described as **negative selection**. The negative selection acts on causal mutations, which often are deleterious and must be removed from a population especially if they modify genomic regions with important functions for the life: in this case we refer to **background selection**, which is more stringent and effective for regions very conserved along the genome.
- in a way to maintain more alleles in the same locus at appreciable frequencies, in case of **balancing selection**; this usually happens when the heterozygotes for the

alleles under consideration have a higher adaptive value than the homozygote (heterozygote advantage). In the case two alleles with opposite phenotypic effects are simultaneously favoured, the event is described as **disruptive selection**; when intermediate phenotypes are favoured, then one speaks of **stabilizing selection**.

While the negative selection usually targets functionally relevant areas of the genome and the balancing selection has more subtle effects on the genome, the positive selection has the peculiarity of leaving clearly visible signs on the genome, so that many approaches have been developed to identify the loci under selection.

When a single copy of a new beneficial allele appears in a population, it can rise rapidly to fixation: this event is known as *hard sweep*. Such classic or hard selective sweeps reduce genetic variation and increase haplotype lengths around the selected site because the time to the most recent common ancestor is shorter at that site [25] [26].

The term *soft sweep* was introduced to describe two different scenarios. In the first scenario, an allele that is already segregating in the population (standing variation) becomes selectively favored due to a new selection pressure, and sweeps up in frequency. In the second scenario, multiple independent mutations at a single locus are all favored and all increase in frequency simultaneously.

Thus, in general, selective sweeps can be *hard*, where a single adaptive allele sweeps through the population, or *soft*, when multiple adaptive alleles sweep through the population at the same time.

2.2. Statistical approaches to detect signatures of positive selection in the human genome

The methods developed to date to detect signatures of positive selection can be divided into three main categories [27]:

- **population differentiation based** tests, based on the concept that when an allele is highly impacting on a trait under selection, it will reach a very high frequency in a population, togheter with other variants in linkage disequilibrium;

- **linkage disequilibrium based** tests, aimed to identify haplotypes that achieved wider prevalence in a population; such haplotypes are preserved and persistent because not yet corroded by recombination events;
- **site frequency spectrum** based tests: site frequency spectrum (SFS) describe the distribution of allele frequencies across sites in the genome of a particular species. Near the point of fixation, the scaled SFS is characterized by an abundance of very high frequency alleles, and a near-absence of intermediate frequency alleles;
- **functional annotation** based tests, based on the assumption that only functionally relevant loci can be positively selected, while non-functional variants should be neutral and their frequencies could be affected only by genetic drift or demographic factors.

2.3. Population differentiation based tests

When a population moves to a new environment, a process of adaptation may take place, and positive selection may act on mutations that help the individuals to better adapt to their new environment. Human populations moving to different parts of the world have experienced distinct climates and natural resources. Therefore, some genetic changes have been favored in one particular population but not the others. If one or more alleles at a particular genomic locus have highly differentiated frequencies in different populations, or are even population-specific, positive selection may have acted on that particular locus in one or more populations. The fixation index (F_{ST}), first introduced by Wright, is often used to estimate population differentiation [28] [29] [30].

The value of F_{ST} ranges from 0 to 1, with a value of 0 implying complete panmixis (i.e. no differentiation), compared with a value of 1 indicating a complete separation between the two populations.

F_{ST} is often used in the detection of selective sweeps, with higher values indicating a higher probability of selection. However, this method is often criticized, as the value of F_{ST} is highly influenced by population structure and demographic history, as well as the ascertainment biases of variant discovery in different population samples. Therefore, F_{ST} values are often evaluated in the context of a genome-wide or multi-locus distribution, as demographic factors or data biases will most likely affect the whole data set equally.

Akey et al. [31] estimated locus-specific F_{ST} compared with genome-wide distribution, and identified over a hundred loci showing "signatures of positive selection" with high levels of differentiation among populations.

Briefly, the value of F_{ST} is higher when most of the variance across the total population can be attributed to the variance between groups, while low F_{ST} indicates that most of the variance is within the groups [32] [31]. If you get a low F_{ST} number then you are not getting much more information by looking at population substructure, but if you have a high F_{ST} substructure might be really informative as genetic differences break relatively cleanly along group differences.

Another test, Population Branch Statistic, or PBS, is able to identify SNPs under selection which are particularly differentiated in a population of interest [22].

Briefly, given a query population A, a closer population B and a distant population C, PBS for a given allele is calculated as follows:

$$pbs = ((-log(1 - F_{ST}\_AB)) + (-log(1 - F_{ST}\_AC)) - (-log(1 - F_{ST}\_BC)))/2 \tag{2.1}$$

2.4. Linkage disequilibrium-based tests

Linkage disequilibrium (LD) refers to the non-random associations of alleles at different loci. There are many factors that can influence the level of LD at a locus in the genome. First of all, the variation of recombination rates causes some loci to be in higher LD than others.

For example, loci within a "cold" recombination region would be more likely to be linked than those within a "hot" recombination region, even if they have similar physical distances. As linkage information is critical for many genetic studies, genetic linkage maps, often known simply as *genetic maps*, have been generated to scale the physical position of genomic variants in terms of recombination frequency. Natural selection, especially positive selection, can have a strong impact on the LD of a locus under selection, and more specifically, will cause the locus to have unusually high LD compared with neutral loci of similar frequency.

As described earlier, if a new mutation increases the fitness in the carriers, the frequency of that advantageous allele will go up rapidly in the population, and finally

will reach fixation or near fixation. Due to the linkage disequilibrium of surrounding alleles with the selected allele, their frequencies will often go up along with the selected allele. As this process takes a much shorter time compared to random drift, it often does not allow sufficient time for recombination to break down the linkage. This will result in a long LD block at the locus, centered on the selected allele.

Therefore, by measuring the level of LD at one particular locus in a population, a selective sweep can be detected if the level of LD at this locus is high compared with other frequency-matched haplotypes in the same or different populations.

Simple measurements of LD at loci are not sufficient to detect signals of positive selection. Other factors that may influence the level of LD need to be considered and their effects need to be removed in order to isolate the long LD signal left by a selective sweep. Furthermore, the pattern of LD scores along the region of interest needs to be considered, in order to identify the most likely selection target site. Based on these principles, several statistical tests have been developed to detect signals of positive selection by measuring the decay of LD scores over long genetic distances. One of the earliest tests is the Extended Haplotype Homozogosity (EHH) test [33], which detects long range haplotypes with a high frequency in a population. Several other tests were then developed based on EHH; for example, the Cross Population EHH or XP-EHH test compares the EHH scores of a *query population* versus a *reference population* at a given locus, providing power to detect population-specific positive selection.

Another test, integrated haplotype score or iHS, calculates the integral of EHH on haplotypes carrying the ancestral allele and the integral of EHH on haplotypes carrying the derived allele, then generates a score based on the ratio of these two EHH scores. This test seems to have a higher power for detecting selective sweeps that have not yet reached the near-fixation stage.

iHS [34] and XP-EHH [35] are two complementary tests. iHS has more power to detect selective sweeps which are in transition to fixation and far less to detect those where the allele frequency is nearly fixed. In contrast XP-EHH is has more power when the sweep is near fixation, but less at lower frequencies. Neither of the techniques have much power to detect sweeps which might just be in their early stages and so exhibit a lower allele frequency.

Recently, new haplotype-based statistic were proposed for for detecting both soft and hard sweeps in population genomic data from a single population [36]. The nS_L test (number of segregating sites by length), developed by Ferrer-Admetilla, differently

from iHS does not require a genetic map. Indeed, the main difference between the nS_L and iHS statistics is in how they measure distance. The nS_L statistic uses segregating sites as a proxy for distance, while the iHS statistic uses the recombination distance.

Another statistical test has been recently proposed by Garud [37] [38], which is based on a measure of haplotype homozygosity (H12) that is capable of detecting both hard and soft sweeps with similar power. A second haplotype homozygosity statistic (H2/H1), in combination with H12, is capable of differentiating hard from soft sweeps, because the frequencies of the first and the second most common haplotype are combined into a single frequency.

Although these LD-based tests have a reasonable power for detecting signals of selective sweeps, the regions they detect are often extended from a few hundred kb to a few Mb in length, so they are generally not able to localize the selection signals into a smaller region in order to identify the exact causal variants.

A recent method is the Composite of Multiple Signals (CMS) test [39], which combines multiple EHH-based tests and measures of derived allele frequency differentiation (XP-EHH, iHS, F_{ST}, ΔDAF and ΔiHH) to generate a composite score, has been proposed to identify the exact variants selected by evolution and not only individual loci.

2.5. Frequency spectrum based tests

One of the most evident effects of positive selection is that it drives the frequency of a beneficial allele to a high frequency or even fixation. Due to the linkage of surrounding alleles with the selected allele, the frequencies of those alleles will also go up. On the other hand, the corresponding alleles on the other non-selected haplotypes will go down rapidly or even disappear from the population. Therefore, alleles in the region surrounding the advantageous allele will differentiate into either very high or very low frequencies. In contrast, frequencies of neutral alleles are only driven by genetic drift, so they fluctuate randomly and are not likely to have the highly differentiated patterns.

If we compare the allele frequency distributions of a region that has undergone a selective sweep with a neutral region, then three main differences may occur:

1. the selected region has a higher proportion of extremely low-frequency alleles than the neutral region;
2. the selected region has a higher proportion of extremely high-frequency alleles than the neutral region;
3. the selected region has a lower proportion or even absence of intermediate-frequency alleles.

Several statistical tests have been developed to detect one or more of these three features, often interpreted as evidence of selection [40] [41]. One of the earliest and still most widely used such tests is the Tajima's D statistic.

A positive Tajima's D value suggests a low level of both low and high frequency alleles in the region, indicating either balancing selection or a decrease in population size, or both. In contrast, a negative Tajima's D suggests an excess of low and high frequency alleles in the region, indicating positive selection, or population expansion.

Another widely used statistic is Fay and Wu's H, which measures an excess of high frequency derived alleles. The H statistic is similar to Tajima's D, but differs by taking into consideration of whether a particular allele is derived or not when looking at pairwise differences.

More recently developed frequency-spectrum based tests use more sophisticated algorithms to increase the robustness to demographic factors. These methods aim to capture the comprehensive spatial patterns of allele frequencies in the region, instead of focusing on just one aspect.

One example of this new generation of tests is the Composite Likelihood Ratio (CLR) test developed by Nielsen et al [7].

2.6. Functional annotation based tests

A certain allele at a genomic locus can be positively selected only if it has functional consequences that are beneficial for the carrier. Therefore, non-functional variants should be neutral and their frequencies should only be affected by genetic drift or demographic factors. By comparing patterns of functional variants versus non-functional variants in a gene or functional element, one could potentially identify signatures of

selection at this locus. The *Ka/Ks* ratio (also known as Ω, or *dN/dS*), for example, is often used for this purpose [42]. It is the ratio of the number of non-synonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) in a protein coding gene.

In the simplest analysis, a Ka/Ks ratio greater than 1 indicates a sign of positive selection, since a Ka/Ks ratio of 1 is expected for a neutral gene. However, more sophisticated statistical analysis needs to be performed to determine the significance of the Ka/Ks ratio as an indicator of positive selection, especially when the number of substitutions is low. Simulations or maximum likelihood analysis may be applied to distinguish between a neutral model and a significant Ka/Ks ratio.

The Ka/Ks ratio is a simple yet powerful tool to identify signatures of positive selection in protein-coding genes, as it uses few assumptions and has a strong functional foundation. However, it has complications and limitations.

First of all, mutation rates of different base substitutions are variable, and the codon usage is often biased, which may result in a higher probability of certain non-synonymous or synonymous changes.

Secondly, certain synonymous changes may have functional impact on the gene, and certain non-synonymous changes may result in similar amino acids and thus have no functional impact on the protein.

Thirdly, the Ka/Ks ratio can only be applied, of course, to protein coding genes, so functional non-coding genes or regulatory elements, which constitute a probably larger proportion of functional loci in the genome, are out of its radar.

Lastly, this method requires a rather strong signal of selection leading to multiple amino acid changes in the same protein, and the two lineages being compared need to be distant enough to allow for this accumulation of non-synonymous substitutions.

Chapter 3.

Gene based positive selection scan

"Sa febbre terzana non est toccu de campana"
— Sardinian saying

3.1. The cytokine BAFF

A large number of diseases occur as a result of the immune system attacking the body's own organs, tissues, and cells. Autoimmune diseases are chronic disabling disorders in which underlying defects in the immune response lead the body to attack its *self*: organs and tissues. To date, more than 80 autoimmune diseases have been classified: the most common include systemic lupus erythematosus (SLE), multiple sclerosis (MS), type 1 diabetes (T1D), autoimmune thyroid diseases (Graves' disease and Hashimoto's thyroiditis), myasthenia gravis, scleroderma, and rheumatoid arthritis (RA).

Although the causes of many autoimmune diseases remain unknown, the genetic background of an individual, in combination environmental factors (such as the personal history of infections or diet), are likely to play a significant role in the disease development. Treatments but not cures are available for the majority of autoimmune diseases. Intensive investigation of autoimmune disease genetics has the potential to fully elucidate the disease etiology and to provide new therapeutic paradigms.

A seminal paper about the process of validating therapeutic targets through human genetics was published in 2013 by Robert Plenge et al. [43] The authors show how the vast majority of the compounds that enter clinical trials fail to demonstrate sufficient

safety and efficacy, and that most of this failure is due to the ignorance regarding the consequences of perturbing specific targets over long periods of time in humans.

A solution can derive from alternative data sources to identify novel drug targets, such as from human genetics: examples are genetic variants with a subtle effect on LDL cholesterol and myocardial infarction, which could point to successful targets for cardiac prevention.

The authors postulate that dose-response curves can derive from experiments of nature, equivalent of clinical trials with an established therapeutic. An ideal scenario is the one where a gene harbours a gain-of-function allele that increases the risk of disease, the variant is also associated with an intermediate phenotype that can be used as a biomarker and the variant is within a gene that is *druggable*.

Genome-Wide Association Studies (GWAS) have provided statistical support for many associated variants, ssuch as in the case of over 110 independent signals for multiple sclerosis. Although several signals are near genes involved in immunological processes, the functional mechanisms for most associations remain unknown.

In the context of the theoretical framework of the *experiments of nature*, an useful approach to dissect disease mechanisms is to search for genetic variants that affect both autoimmune disease risk and quantitative immune variables such as circulating levels of immune cell populations, immunoglobulins, and cytokines [44].

This approach can reveal disease-related endophenotypes and is especially informative when variables are measured in healthy individuals, avoiding secondary effects of the disease and its therapy. Functional genomic annotations from these cell types and states can then be used to resolve candidate genes and causal variants.

For such studies, the genetic structure of the Sardinian population [45] (Sidore et al), with prevalence of MS and SLE among the highest world-wide [46], is likely to facilitate the detection of relevant variants missed in studies of cosmopolitan populations [47] [45] [48].

In Sardinia, in more details, about 5% of the general population is affected by one or more autoimmune diseases [49], and the prevalence values for multiple sclerosis confirm that Sardinia is among those regions with the highest prevalence of this disorder [50].

We recently reported the genetic contributions to quantitative levels of 95 cell types encompassing 272 immune traits, in a cohort of 1,629 individuals from four clustered Sardinian villages [51].

Among the others, a new association (hereafter referred to as **BAFF-var**) in the 3'untranslated region (UTR) of *TNFSF13B* (encoding tumor necrosis factor superfamily member 13b) with MS, SLE and a number of immunophenotypes (*Steri et al, NEJM, in press*). *TNFSF13B* is also known as the B-cell-activating-factor (BAFF), a cytokine essential for B-cell activation, differentiation, and survival, primarily produced by monocytes and neutrophils[1]. BAFF inhibition is also an attractive target for autoimmune disease therapy [52].

BAFF-var is common in Sardinia (MAF=26.5%) and progressively rarer proceeding from Southern to Northern Europe (and absent in Africa and Asia). Given the relative genetic isolation of the Sardinian population [45], we assessed whether the high frequency of BAFF-var on the island and the frequency differentiation we observed in Europe are expected due to genetic drift or positive selection.

3.1.1. Materials and methods

To assess whether BAFF-var was a target of positive selection, five standard frequency-based and haplotype-based statistical tests (see Appendix for details about the methods) were used: i) F_{ST}, which estimates allelic differentiation between populations; ii) Population Branch Statistics (PBS), which we used to estimate the magnitude of the Sardinian-specific allele frequency change; then iii) iHS and iv) nSL, both of which evaluate haplotype diversity among allele carriers in a single population, and v) xp-EHH, a statistic that compares the extent of haplotype diversity in different populations. Additionally, a modified version of xp-EHH was also developed applied (as-xp-iHH) to test differences in the length of the haplotype carrying BAFF-var in Sardinians compared to other populations.

The F_{ST} statistic, a measure of population differentiation, when close to 1 means that the maximal possible differentiation between two populations is observed at a segregating site. To calculate the F_{ST} statistic, we used the Weir-Cockerham formula implemented in vcftools v.0.1.12b (http://vcftools.sourceforge.net/53) to compare Sardinians with populations from the 1000 Genomes Project [53].

[1] See https://omim.org/entry/603969

F_{ST} is a measure of differentiation between two populations at a given locus, but is not informative about the direction of the differentiation and could be influenced by demographic events. To further investigate the strength and the direction of this difference in allele frequency, we used another well known test, the population-branch statistic (PBS), which estimates the allele frequency difference on a specific population branch of a 3-taxa population tree. PBS was calculated comparing Sardinians with Tuscans from Italy (TSI) and with British from England and Scotland (GBR), in order to estimate the magnitude of allele frequency change since the divergence of Sardinian and Tuscan populations. PBS values were calculated combining the F_{ST} among the three tested populations.

Among the haplotype-based tests, iHS is based on the ratio of integrated haplotype homozygosity for the haplotypes carrying the derived allele (iHH_D) and the ancestral allele iHH_A at a candidate site. An alternative approach, nSL, has been proposed as a haplotype-based statistic for detecting, in a single population, both hard and soft sweeps. A hard sweep denotes an instance when a new advantageous mutation arises and spreads quickly to fixation due to natural selection. A soft sweep indicates an instance when a neutral allele becomes favored due to a driving force of positive selection and increases in frequency, or when multiple independent mutations at a single locus are all favored and all increase in frequency simultaneously until the sum of the frequencies is 1 but no single favored allele will reach fixation during the selective event. The main difference between iHS and nSL statistics is the method used to calculate the length of a segment of haplotype homozygosity: iHS is based on the genetic distance, while nSL relies on the number of mutations in the region. Therefore, no genetic map is required to calculate the nSL statistic.

Another haplotype-based score, the xp-EHH (cross population-Extended Haplotype Homozygosity), compares the integrated EHH profiles between two populations at the same SNP; the xp-EHH only gains power for nearly fixed alleles, because it does not consider the specific allele under selection: this means that sampling error from haplotypes of the unselected allele is minimized when the selected allele is near fixation. Indeed, the xp-EHH test has the most elevated power for selective sweeps in which the selected allele has risen to high frequency or fixation in one population, but remains polymorphic in the human population as a whole. The as-xp-iHH (allele specific - cross population - iHH), is defined as:

$$as - xp - iHH = log(iHH_{\mathrm{d1}}/iHH_{\mathrm{d2}}) \tag{3.1}$$

where iHH_{d1} and iHH_{d2} refer to the same derived allele in populations 1 and 2. Whereas the iHS statistic compares the integrated EHH profiles between two alleles at a given SNP in the same population, the as-xp-iHH statistic compares the integrated EHH profiles of the specific allele putatively under selection between two populations. With as-xp-iHH, we assessed evidence for the specific hypothesis that BAFF-var resides on a longer haplotype in Sardinians with respect to other populations.

The unstandardized iHS, nSL and xp-EHH values were calculated using selscan (release 1.1.0 - 07MAY2015) [54].

In order to assess a significance for the values of BAFF-var in each test, an empirical percentile calculation was performed. In more details, a genomic background distribution was constructed by calculating the statistics on a set of 3,042 randomly selected variants matching BAFF-var in three genetic features in Sardinians: minor allele count, measure of background selection, and recombination rate in a 50kb region around the variant.

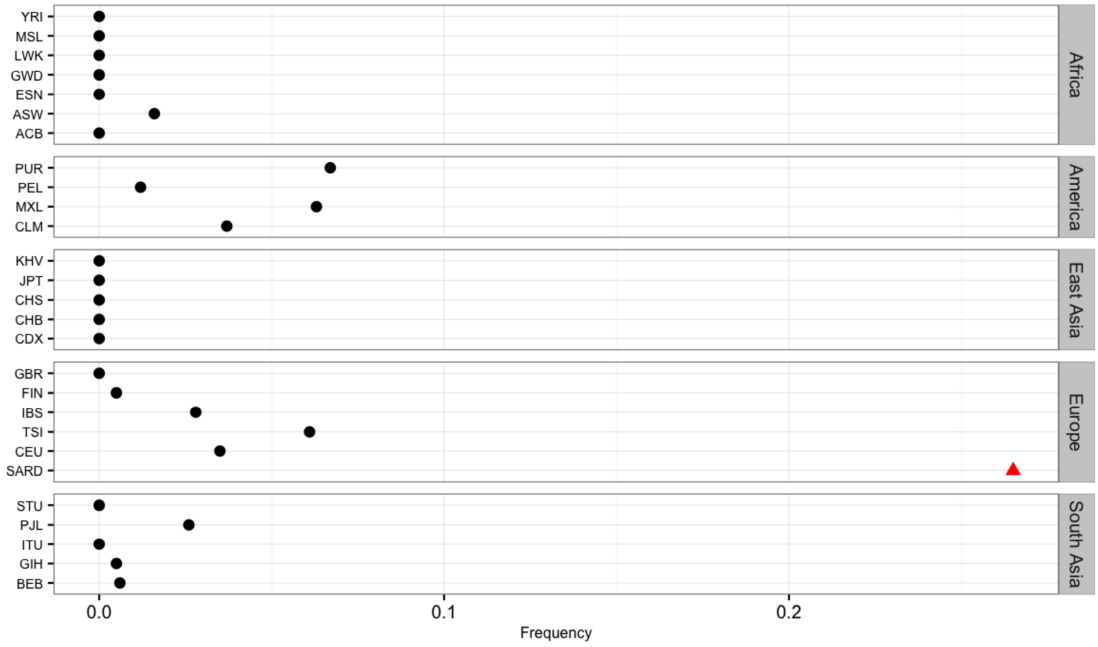
To exclude variants in LD, variant pairs with r2 $\geq$ 0.1 were filtered using a *pruning* procedure implemented in Plink (https://www.cog-genomics.org/plink2/). A high-resolution genetic map based on LD patterns was then estimated by linearly interpolating the genetic map files in IMPUTE format (physical positions in NCBI b37/hg19 coordinates) from https://mathgen.stats.ox.ac.uk/impute/1000GP. The estimated genetic distances were then used when calculating iHS and xp-EHH statistics.

For both the frequency-based and the haplotype-based analyses, variants with MAF<0.01 and Hardy-Weinberg proportion test P value<10^{-6} in Sardinian samples were first removed; this conservative filter allowed a better reconstruction of haplotypes by reducing errors in genotype calls.

3.1.2. Results

Although BAFF-var is relatively common in Europe (Figure 3.1), we estimated that BAFF-var shows substantially greater differentiation between Sardinians and other Europeans (F_{ST}=0.21, 99.97th genomic percentile, Table 3.1 and Population Branch Statistic - PBS, Figure 3.2).

Figure 3.1.: BAFF-var frequency in Sardinians and 1000G populations.



Moreover, the variation at BAFF-var is consistent with a selective sweep reducing haplotype diversity in Sardinia (iHS=3.38, 99.91th genomic percentile, Figure 3.3; nSL=1.96, >99.99th genomic percentile, Figure 3.4). These findings suggest that the high frequency of BAFF-var in Sardinia arose as adaptation to selective pressure particularly prevalent there.

We examined the extent of haplotype homozygosity to see if it reinforced the strong evidence for BAFF-var differentiation between Sardinians and the 1000G populations, inferred from two frequency-based tests - F_{ST} (Table 3.1) and Population Branch Statistic (PBS) (Figure 3.2). With this aim, we applied three haplotype-based statistics: iHS, nSL, xp-EHH, and a modified version of xp-EHH, as-xp-EHH.

Table 3.1.: BAFF-var differentiation between Sardinians and the 1000G populations

Population	# Chromosomes	BAFF-var frequency	F_{ST}	Genomic percentile
Europeans (EUR)	1006	0.027	0.207	99.97
Tuscans in Italy (TSI)	214	0.061	0.163	99.97
Europeans w/o TSI (EUR not TSI)	792	0.018	0.196	99.97
Utah Residents (CEU)	198	0.035	0.162	99.90
Iberian Population in Spain (IBS)	214	0.028	0.162	99.93
Americans of African Ancestry (ASW)	122	0.016	0.156	85.98
Admixed Americans (AMR)	694	0.045	0.190	99.13
Colombians from Medellin (CML)	188	0.037	0.161	99.53
Mexican Ancestry from USA (MXL)	128	0.063	0.156	94.90

Compared to haplotypes possessing similar genetic features in Sardinians, haplotypes carrying BAFF-var are significantly longer than those carrying the ancestral allele (Table 3.2 and Figures 3.4 and 3.5a). Less significant results were found for the other European populations, although the estimate of the extent of haplotype homozygosity could be affected by smaller sample size of those population samples.

Listed, from left to right, are: (1st column) the population code; (2nd) the integrated Haplotype Score (iHS) values; (3rd) the number of matching variants with BAFF-var; (4th) the number of matching variants with BAFF-var with an absolute iHS value better then BAFF-var iHS; (5th) the genomic percentile.

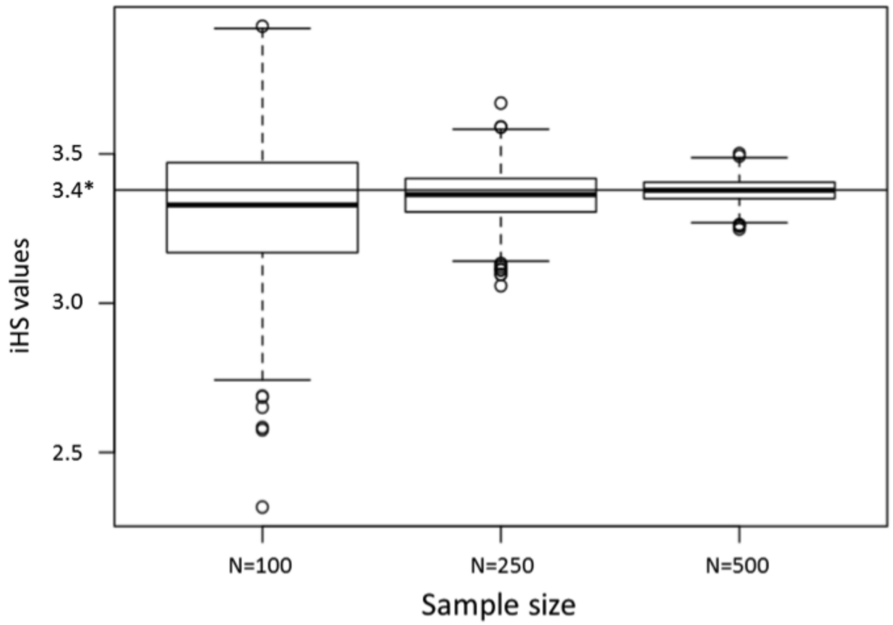
Table 3.3.: iHS results in Sardinians and in 1000 Genomes populations.

Population code	iHS	# matching variants	# matching variants with abs(iHS) > BAFF-var	iHS Genomic percentile
Sardinians	3.3790	3042	5	99.91
ASW	1.8039	27713	23718	19.50
CEU	3.2868	11896	1220	94.27
CLM	3.4215	9487	855	91.38
IBS	3.4305	11112	821	95.80
MXL	3.1348	4108	284	97.03
PEL	1.8941	10150	3888	60.05
PJL	4.0280	14597	430	85.70
PUR	4.0654	6859	62	99.19
TSI	2.9306	7985	531	96.48
EUR	3.2013	76602	11463	99.93

To assess the impact on EHH (and consequently on iHS) of the slight differences in sample size between Sardinians and 1000 Genomes Project populations, 1,000 simulations were performed at the *TNFSF13B* locus by randomly sampling 100, 250 and 500 Sardinian individuals (equally distributed among the SardiNIA and the MS

case-control studies) to reproduce the average 1,000 Genomes Project population size (Figure 3.2). This simulation showed that the higher the sample size, the more

Figure 3.2.: Simulations at the TNFSF13B locus.



robust the estimate of the haplotype homozygosity. Thus, we can state that the observed selection signals in the Sardinian population are robust; however, we cannot completely exclude that a stronger signal might be observed in the 1000 Genomes population if the sample size for European populations were larger. Moreover, xp-EHH and as-xp-EHH results showed no significant differences in the length and structure of the core haplotype around BAFF-var in Sardinians compared to other European populations in which the variant is detected at appreciable frequencies (Figure 3.5b and Table 3.3).

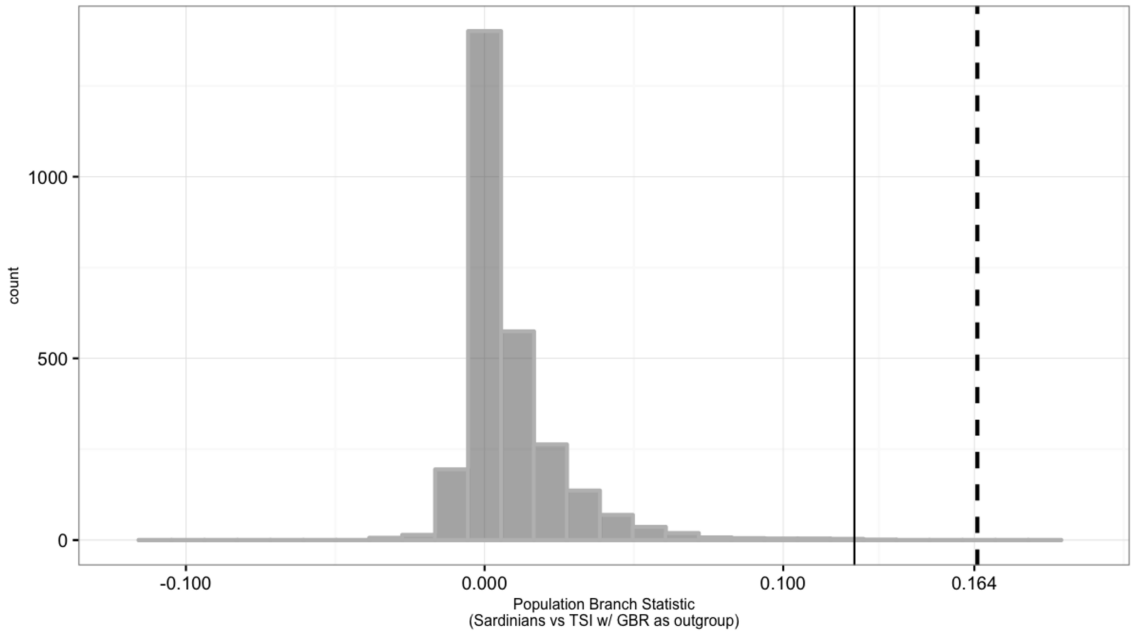
Table 3.5.: Cross-population results when comparing Sardinian vs 1000 Genomes populations.

Pop. code	BAFF-var freq	iHHd Sard.	iHHd 1000G pop.	xp-EHH	xp-EHH gen. perc.	as-xp-iHH	as-xp-iHH - anc. gen. perc.	as-xp-iHH - der. gen. perc.
EUR	0.026	0.008	0.007	0.109	20.3	0.047	36.0	24.1
CEU	0.035	0.010	0.008	0.177	65.9	0.077	76.5	80.2
TSI	0.060	0.010	0.008	0.190	93.9	0.082	49.0	88.5

Listed, from left to right, are: (1st column) the population code; (2nd) the BAFF-var frequency, (3rd) the integrated Haplotype Homozygosity of the derived allele (BAFF-var, iHHd) in Sardinians; (4th) the iHHd in the 1000 Genomes populations; (5th) the cross population-Extended Haplotype Homozygosity (xp-EHH) value, and (6th) its genomic percentile; (7th) the allele specific cross population iHH (as-xp-IHH) value and its genomic percentile with respect to (8th) the ancestral and (9th) the derived allele.

Overall, these findings point to high frequencies of BAFF-var arising substantially as an adaptation to a selective pressure that has been relatively common in Southern Europe and extremely prevalent in Sardinia.

Figure 3.3.: Genomic distribution for Sardinian PBS relative to Tuscans-TSI and British-GBR from 1000 Genomes Project, where the BAFF-var score (the vertical dotted lines) is compared with about 3,000 variants matched with BAFF-var by allele frequency in Sardinians, local recombination rate and B score.



3.1.3. Discussion

BAFF-var may have been positively selected in Sardinia by providing resistance to infections by Plasmodium Falciparum and/or Vivax malaria, both strikingly prevalent in Sardinia until their eradication in the 1950s [55].

Malaria selective pressure on BAFF levels is supported by mouse malaria models, in which BAFF-overexpression protects from lethal malaria infection [56]. In addition, as shown here, BAFF-var increases antibody production, and classic findings showed that antibody transfer from immune adults to acutely infected children reduced blood-stage parasitemia and disease severity. Overall, the evolutionary scenario we propose is that BAFF-var was selected as an adaptive response to malaria infection, resulting in increased present-day risk for autoimmunity.

Figure 3.4.: Genomic distribution of the absolute unstandardized iHS in Sardinians.

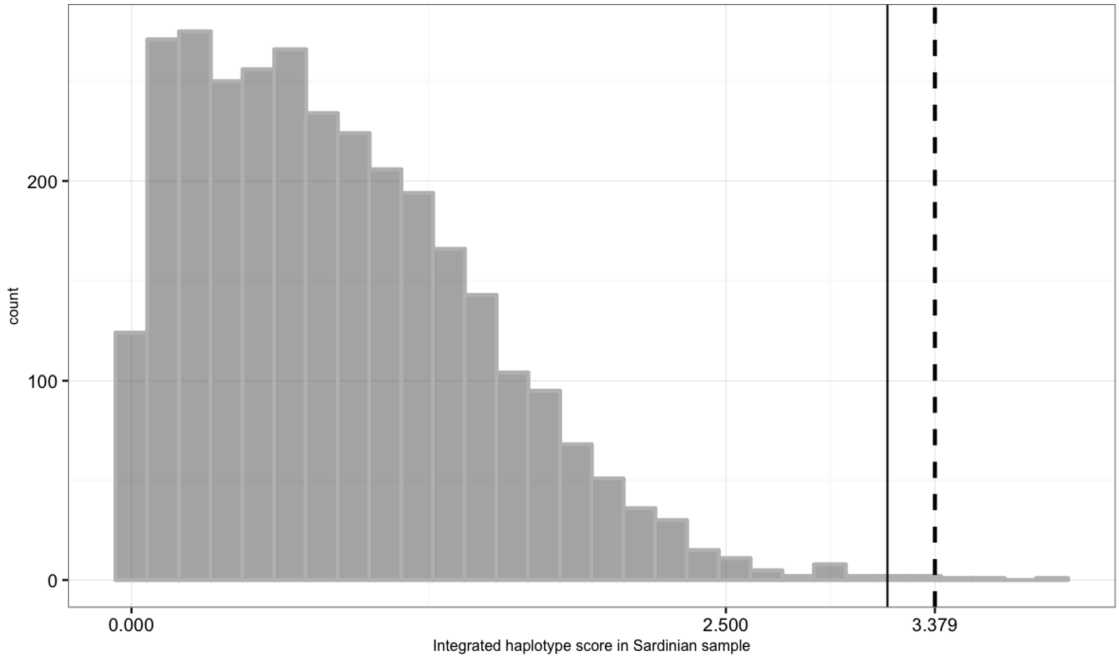
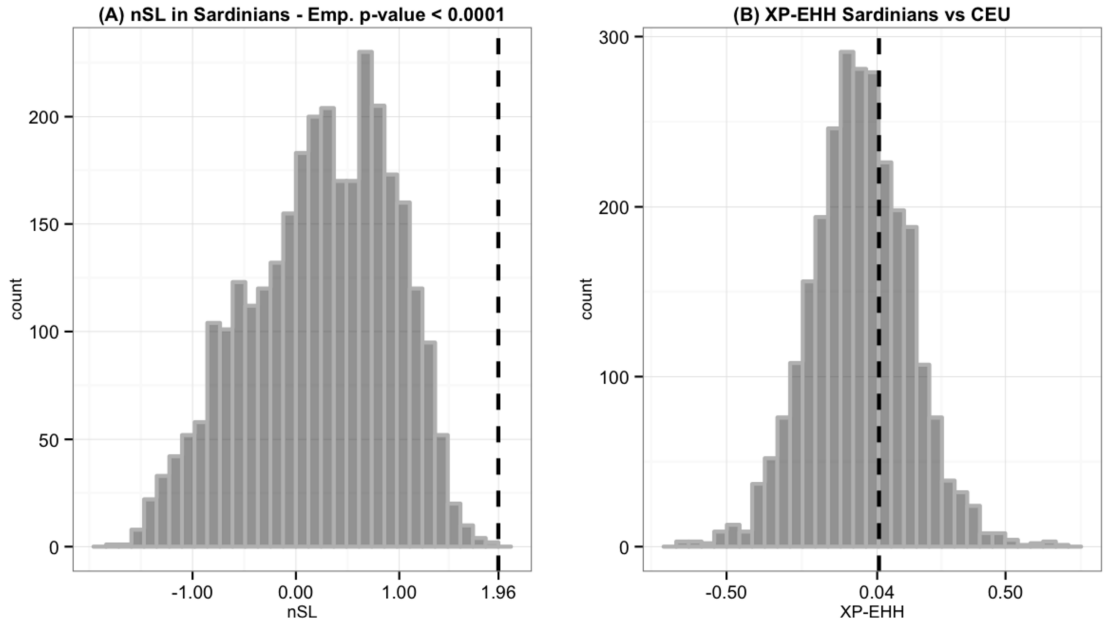


Figure 3.5.: Haplotype-based tests for BAFF-var selection.



3.2. Polygenic patterns of selection: the example of short stature in Sardinia

Recently, many papers focused on human height reported its variation across populations and dissected the genetic basis of this quantitative trait. Briefly, it emerges that height varies across Europe in a North-South gradient, with Sardinia being an exception. Furthermore, some of this variation in height is genetic and some of the genetic variation in height is driven by selection.

In a set of recent sequencing-based whole-genome association analyses to evaluate the impact of rare and founder variants in 6,307 individuals on the island of Sardinia, the group of Prof. Cucca at IRGB-CNR identified two variants with large effects on stature [48]. The first of the two mutation is a stop codon in the *GHR* gene, relatively frequent in Sardinia (0.87% vs <0.01% elsewhere), which in homozygosity causes the short stature Laron syndrome. This mutation reduces height in heterozygotes by an average of 4.2 cm (-0.64 s.d).

The other variant, in the imprinted *KCNQ1* gene (MAF = 7.7% vs <1% elsewhere) reduces height by an average of 1.83 cm (-0.31 s.d.) when maternally inherited.

To evaluate the overall impact of known variants on the average short stature observed in Sardinia relative to other populations and to test the possibility that short stature might be selected for in this island population, we used polygenic height scores. These scores measure the total frequency of height-changing alleles in a population, weighting each allele by its effect size.

In the population genetic analyses, we focused on a subset of 1,081 unrelated sequenced individuals. To investigate whether height-decreasing loci have been under selection in Sardinia, we calculated a polygenic height score for each population m as

$$Z_m = 2 \sum_{l=1}^{L} \beta_i p_{ml} \tag{3.2}$$

where β_{i} is the effect size of the height-increasing allele l and $\mathrm{p_{ml}}$ is the frequency of allele l in population m.

To avoid biases and to ensure uniformity of the source of effect size estimates, we used the effect size estimates from the Sardinian dataset regardless of whether the variant is significantly associated with height in this dataset.

We first calculated the polygenic height score Z_m based on the 691 height loci identified by the GIANT consortium [57] with effect sizes estimated in the Sardinian dataset and then added the two top variants reported, totaling 693 height alleles. To test if there were a signature of polygenic adaptation on height in Sardinia, we adopted a framework developed by Berg and Coop [58], which builds a multivariate normal model based on matched, presumably neutral variants, to account for relationships among populations.

Populations with extreme polygenic scores relative to the expectation (pvalue = 0.01) are likely to have undergone selection. To construct a null distribution of frequencies needed for the multivariate normal framework, we obtained for each of the height loci all variants in the 1000 Genomes phase 3 European data with minor allele count $\pm$ 10 counts ($\sim$1% in frequency), B score (35) $\pm$ 50 units, and local recombination rates $\pm$ 0.5 cM/Mb. A random subset of 509,386 SNPs, representing 10% of the union of the matched SNPs, were then used as a set of matched SNPs for the analysis. Of note, we also repeated the calculation using effect sizes estimated by the GIANT consortium as well as using only a subset of 162 SNPs that are not subject to population stratification.

We observed a significantly lower polygenic height score in Sardinia compared to other European populations examined in the 1000 Genomes project, including the Southern European Tuscans and Spanish. Adding the *KCNQ1* and *GHR* variants to the previously described 691 alleles, the polygenic score of Sardinians decreased by 3.8%. Overall, Sardinian scores are lower than would be expected compared to other European populations ($p=1.62 \times 10^{-6}$, -5.9 cm relative to CEU, 1.6% average increase in frequency for height decreasing alleles), even when calibrating for genome-wide patterns of differentiation due to genetic drift, suggesting that selection has played a role in decreasing height in Sardinia. The differences in height explained by the polygenic score are in accord with the observed $\sim$10 cm of phenotypic differences between Sardinians and the other European populations.

We have also considered the possibility that Sardinians might have an additional contribution of reduced height due to the expression of recessively acting height-decreasing alleles exposed due to founder effects. However, the impact of elevated

homozygosity among Sardinians on height appears to be small (0.129 s.d.) relative to the effects predicted by the polygenic score (0.910 s.d.).

Intriguingly, the increased frequencies of height-decreasing alleles at *GHR* and *KCNQ1*, and especially the polygenic height scores in this population, are also consistent with the long- standing observation of an *island effect* in which many large animals become adaptively smaller on islands relative to their mainland counterparts.

Chapter 4.

Genome-wide positive selection scan

4.1. Introduction

4.1.1. Detecting positive selection in geographic isolates

Isolated populations, such as Finnish, Old Order Amish (a North American ethno-religious group), Hutterites (an ethno-religious group that is a communal branch of Anabaptists, most of which live in Western Canada and the upper Great Plains of the United States), Sardinian and Jewish communities, present an inspiring opportunity to comprehend the genetic basis of adaptation, and many recent genomic scans for positive selection in such populations have led to candidate genes directly linked to adaptive phenotypes [59].

In geographic isolates, longer stretches of haplotypes are expected than those observed in cosmopolitan populations [60] [61]. On the one hand, this phenomenon may increase the chances of identifying genuine sweeps, by emphasizing the positive selection signals around them. On the other hand it could also complicate the attribution of an increase length of core haplotypes around a given site to positive selection vs demographic forces. However, geographic isolates may also offer a good opportunity to discriminate between selection and demography by assessing a large number of unlinked loci throughout the genome all sharing the same demographic background. Indeed, demographic forces affect genetic patterns at all loci to a similar extent, whereas natural selection acts only upon specific loci [62].

Probably, it could be assumed that population demographic history affects patterns of variation at all loci in a genome in a similar extent, whereas natural selection acts

upon specific loci. Thus, by sampling a large number of unlinked loci throughout the genome, it is in principle possible to discriminate between selection and demography [63].

To date, many recent studies of positive selection in isolated populations have detected very clear signals of adaptation. For instance, very recently whole-genome high-coverage sequence data from native Siberians have highlighted one of the strongest selective sweeps reported in humans, in *CPT1A*, a key regulator of mitochondrial long-chain fatty-acid oxidation [64].

Fumagalli et al performed a positive selection scan on a population of Greenland natives, the Inuit [21]; the results clearly show that Inuit have mutations in genes that control how the body uses fat which provides the clearest evidence to date that human populations are physically adapt to survive dramatic conditions and live healthly with a traditional diets rich in omega-3 polyunsaturated fatty acids from marine mammal fat. The SNP with the highest PBS (Population Branch Statistic) value falls within *FADS2*, which encodes a delta-6 desaturases, responsible of the conversion of linoleic acid (omega-6) and a-linolenic acid (omega-3) to biologically active fatty acids. This study was based on data from 191 previously genotyped Greenlandic individuals by using the Illumina MetaboChip.

Another study in the same population found evidence that adaptation to the traditional hypoglycemic diet may have favored a mutation in *TBC1D4* that affects glucose uptake and occurs at high frequency only among the Inuit [65].

A similar study focused on the exomes of 50 individuals from the Tibetan Plateau. Notably, the strongest signal of natural selection came from *EPAS1*, a transcription factor involved in response to hypoxia. *EPAS1* was the most significant gene in a PBS-based scan [22].

4.1.2. Positive selection in the Sardinian population

A recent work aimed to study positive selection in Sardinia by analyzing about 300 individuals (from the Ogliastra region and from Southern Sardinia) genetically characterized with the Affymetrix Genome-Wide Human SNP 6.0 Array [66]. In Appendix more classical natural selection studies are reported.

The study presented here is different for many reasons. First of all, it includes a much larger sample set, deeply covering the world-wide genetic variability, specifically focused on the Sardinia population. Secondly but not secondarily, our analyses are based on sequencing data: this means higher genetic definition and sensibility. Indeed, the Affymetrix array (likewise all commercial Chip Arrays) is not specifically designed for Sardinians, allowing us to observe in many cases only "surrogated" signals; while using sequencing, we can focus on Sardinian specific variation in an exhaustive manner.

Furthermore, Piras et al. compared two sardinian sub-populations with the aim to explore their genetic structure and to identify signatures of natural selection; instead, in our study, we compare Sardinians with mainland populations, and our analysis consider both Sardinians as unique group, and also Sardinians divided into sub-clusters (a smaller unadmixed subset of people from the Arzana village and a sample representing the overall sardinian population); in this way, we are able now to check if alleles differentiated among Sardinians and other populations, for example Tuscans, are also differentiated among sardinian sub-clusters (unadmixed and admixed subsets), indicating selective/demographic events peculiar for Sardinian population.

In general, a weakness of genome wide positive selection scans is a weak follow-up; now, by coupling them with GWAS, we are in conditions to explain the meaning of a sweep, because more phenotypic and functional roles of the SNPs putatively under selection are made available through GWAS; good examples of this approach are Fumagalli et al, but also Raj et al papers.

In this context, therefore, we have an unprecedented opportunity to couple a genome wide positive selection scan on a population with many association studies [51] [45] [47] [48]. We can say our study is a *second generation selection scan* (or *post-GWAS scan*), where with this definition we mean a direct identification of traits under selection (more than SNPs under selection) through a proper intersection of genomic

regions with evidence of selection and genomics regions associated with many trait or diseases.

During the years, several GWAS studies have been performed in Sardinia [51] [45] [47] [48]. Starting from the work by Pilia et al 2006 [67], in the context of the ProgeNIA project[1], aimed to identify genes and environmental factors responsible for human aging, clinical data and DNA genotyping allowed the identification of genetic determinants associated with obesity, regulation of uric acid, cholesterol, triglycerides, TSH (the thyroid stimulating hormone), fetal hemoglobin, and those involved in the determination of the height of an individual.

Therefore, the concomitant availability of association studies, clinical data and sequencing data creates a unique opportunity to identify variants detected by GWAS affected by positive selection and to facilitate the finding of true causal loci for complex traits of an unique population.

4.2. Materials and methods

4.2.1. Datasets

1. Sardinian population: 1,577 unrelated samples extracted from 3,514 sequenced individuals belonging to the SardiNIA and autoimmunity case-control Sardinian cohorts were used for our analysis. We estimated relatedness by computing the genome-wide proportion of pairwise IBD (π) on a random set of 1 million SNPs with an MAF>0.05 in 1,000 Genomes populations (Phase 3 v5). For each pair of individuals with $\pi > 0.05$, we preferentially removed the offspring if it occurred in a trio; otherwise, we removed the individual with the larger summed value of π across all other relationships with $\pi > 0.05$, leading a total of 1,577 samples for the analyses.
2. 1000 Genomes populations: we used the Phase 3 VCF files, version 5, from which we extracted variation data at single population levels and at superpopulation levels (continental blocks).

[1] See https://sardinia.nia.nih.gov/

4.2.2. Software

Variant processing tools

1. VCFTools: development version downloaded May 2015 [68].
2. other *in-house* tools were developed.

Selection scan tools

1. Selscan: release 1.1.0 (07MAY2015), downloaded May 2015 [54].
2. Scripts released by Matteo Fumagalli at https://github.com/mfumagalli/EvoGen_course

4.2.3. Data preparation

For both Sardinian and 1000 Genomes data analysis, we retained only biallelic SNPs with minor allele frequency MAF $\geq$ 1% and with genotypes in Hardy-Weinberg Equilibrium (i.e. sites with a p-value below 10^{-6} are taken to be out of Hardy-Weinberg Equilibrium, and therefore excluded). Both MAF and Hardy-Weinberg p-value filters were applied at single population level. In the Sardinian dataset, we retained a total of 8,001,711 biallelic SNPs.

Specific filters for Population Branch Statistic

To discard potential genotyping allele swapping, which can result in false positives in frequency based tests, at each site we computed the absolute difference in allele frequencies between the 91 individuals from England (GBR) included in the 1000 Genomes Project and 3,781 samples from the UK10K Study [69]. We first discarded 252 SNPs from the top 80,000 PBS results (top 1% outliers) because they did not pass the Hardy-Weinberg filter; from the remaining SNPs, we discarded further 1,295 SNPs because the allele frequency difference among GBR and UK10K dataset was significantly high on the basis of a chi-square test (p-value < 0.05). We then applied the same filter by comparing allele frequencies in 503 1000 Genomes European (EUR) individuals and 449 GoNL samples [70]; with this filter, further 3,513 SNPs were discarded. The filtering resulted in a total of 74,940 SNPs. We considered only the 58,498 for which minor allele frequency in both 1000 Genomes Tuscans (TSI) and

Yorubans (YRI) was ≥ 0.01. We considered only the top 8,000 of the resulting SNPs for the PBS analysis.

Of note, many SNPs with extreme PBS were removed by the above filters: of the SNPs with the top 1% PBS values, 7,436 are not in the UK10K dataset, 7,547 are not in the GoNL dataset (overlap = 3,343 SNPs).

4.2.4. Tests for differentiation and positive selection

For detecting positive selection in our Sardinian genomes, we applied the two following test cathegories:

1. Frequency-based tests

    a) F_{ST} [32]: the pairwise per-site Fst values, following the diploid method in Weir [71] implementated in VCFtools, were computed for 1,577 unrelated Sardinians versus each of the 1000G population and superpopulation, as described earlier for every biallelic SNP with with minor allele frequency > 1% and not excluded by the Hardy-Weinberg Equilibrium test. We next computed the F_{ST} values by using a method-of-moments estimator described in Reynolds et al. (1983) [72]. The correlation of the F_{ST} values calculated with the Weir and the Reynolds estimators was close to 1 (data not shown).

    b) Population-Branch-Statistic: we detected recent differentiation by comparing 1,577 unrelated Sardinians with 107 Tuscans (here considered as closer population) and 108 Yorubans (outgroup, distant population); for each SNP, we calculated the corresponding PBS value by considering the per-site F_{ST} values of each pair of populations [22].

2. Haplotype-based tests

    a) nSL: we used the nSL as implemented in Selscan for detecting both soft and hard sweeps; the nSL method has beed show to have much power as other methods under a number of different selection scenarios, most notably in the cases of sweeps from standing variation. nSL statistics was calculated for 1,577 unrelated Sardinians [36].

    b) iHS: we calculated iHS for 1,577 unrelated Sardinians and for all the 1000G populations and superpopulations [34].

c) xp-EHH: we calculated xp-EHH by using 1,577 unrelated Sardinians as query population and 107 Tuscans as reference population [35].

4.2.5. Strategies for detection of outlier loci

The aim of positive selection studies is to identify outlier loci. Classical approaches are based on coalescent simulations to mimic genome-wide analysis by simulating both neutral and positively selected genes, and use the resulting simulations to establish the threshold for a given test statistics.

A recently prevailing approach is to produce genome-wide data and assume that selection acts on one or a few loci while demographic processes act across the genome. Usually, only strong outliers of the genome-wide distributions are considered as true candidates. This is the strategy we adopted in our genomic scan for positive selection.

Finally, for single loci analysis it is usually generated a null distribution, based on genomic features (such as minor allele frequency or local recombination rate); than, the variant of interest is ranked against this null distribution. We used this approach in the case of BAFF var (see Chap. 3).

Table 4.1.: Summary of the outliers for each test statistic.

Test	**# outliers SNPs**	**# clusters (size>2)**
PBS	8000	641
iHS	8000	262
nSL	8000	271
XP-EHH	8000	146

4.3. Results

4.3.1. Signals of differentiation in Sardinians

To detect signals of positive selection, and in particular to estimate the magnitude of the Sardinian-specific allele frequency changes, we used the Population Branch Statistic (PBS), which identifies alleles that have experienced a frequency change relative to two reference populations (in our case, Tuscans-TSI and Yorubans-YRI); we only considered SNPs whose MAF in both TSI and YRI is higher than 1% .

The risk of considering isolated signals is to generate hypotheses based on many false positives. For this reason, the analysis of clustered signals should reveal more genuine results. Our cluster analysis identified several SNP windows with high PBS values, indicative of directional selection. We built clusters by considering groups of SNPs where the maximum allowed pairwise distance is 250 Kb.

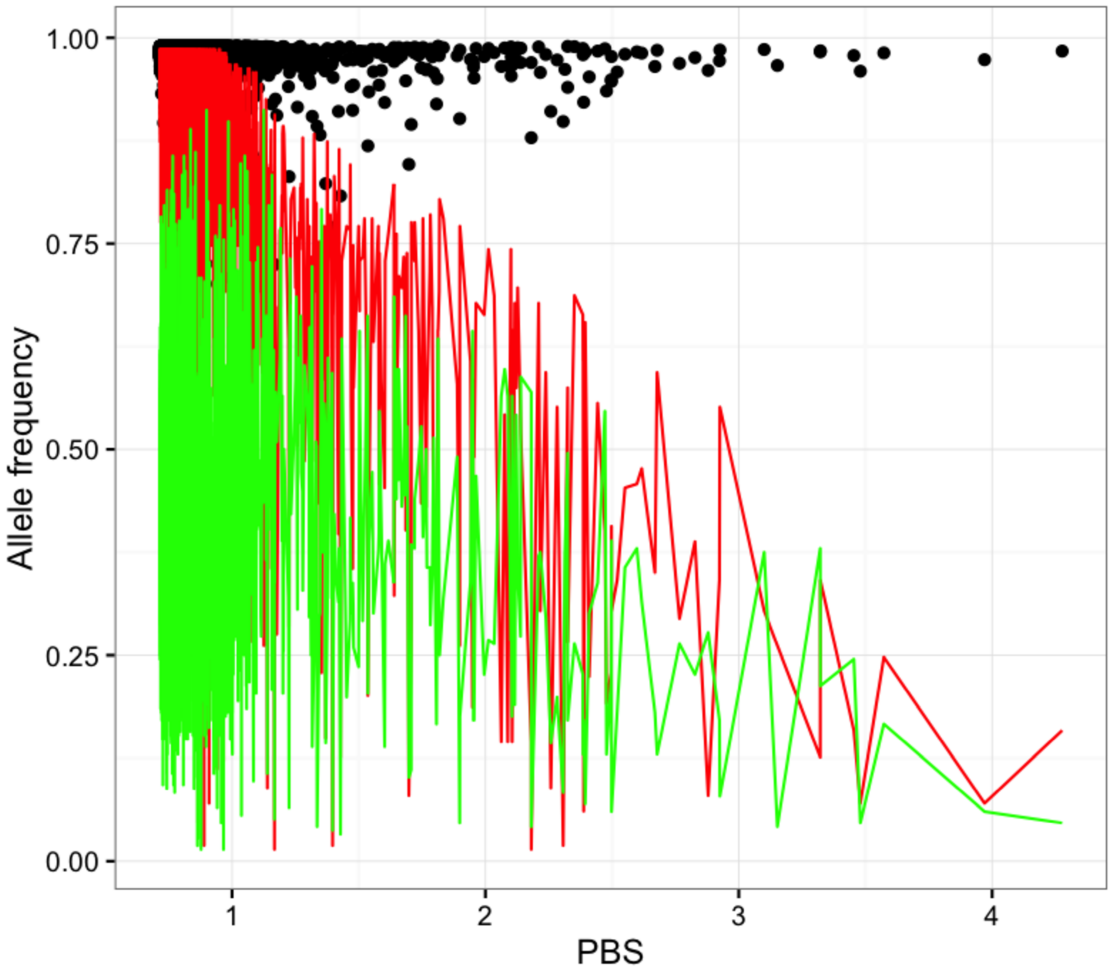
Regions with less than 2 SNPs were discarded and not considered for further analyses. A total of 641 clusters were identified. The allele frequencies in Sardinians (black points), TSI (red line) and YRI (green line) of the top SNPs in each cluster can be seen in Fig. 4.1. We applied a sorting strategy of the PBS clusters. Indeed, we first sorted by PBS score of the top signal in clusters, then considered only clusters with at least 5 total PBS outliers, and finally sorted again by the median PBS value of the SNPs in the cluster (this in order to emphasize those clusters with many strong signals). Among the strongest signals of selection (see Table 4.2), a relevant is the one represented by the SNP 12:56726518; such cluster, composed by 22 PBS outliers with a top PBS value of 1.06 and a median of 1.04, is located within a region on chromosome 12 and encompasses 7 genes: ANKRD52, PAN2, STAT2, IL23A, CS, NABP2 and SLC39A5. Notably, this is an important autoimmunity related locus[2].

[2] See https://www.immunobase.org/

Table 4.2.: Genomic regions representing the top PBS results

top SNP	cluster size	median PBS	top PBS	Genes
19:20407749	41	0.91	1.82	ZNF826P, ZNF626
9:6720202	5	1.14	1.27	KDM4C
9:108550613	5	0.99	1.09	proximal to TMEM38B
12:56726518	22	1.04	1.06	ANKRD52, PAN2, STAT2, IL23A, CS, NABP2, SLC39A5
22:23714788	5	0.99	1.04	C22orf43, CES5AP1
11:80870304	29	0.95	1.03	intergenic
6:5741280	5	0.99	1.03	FARS2
6:103908744	6	0.96	1.01	intergenic
9:17547852	15	0.97	0.99	CNTLN, SH3GL2
11:61144652	13	0.93	0.98	DDB1, CYB561A3, VWCE, DAK
10:77411624	23	0.92	0.98	intergenic
4:72839147	17	0.95	0.98	NPFFR2
2:18661982	6	0.97	0.98	intergenic
6:3192118	6	0.93	0.98	BPHL, LOC100507194
4:53088516	14	0.94	0.97	intergenic
15:69130997	9	0.90	0.97	NOX5, MIR548H4
1:207338691	11	0.93	0.96	FAIM3, C4BPA
1:20256006	6	0.93	0.96	intergenic
2:138225792	6	0.91	0.95	THSD7B
2:188136191	8	0.91	0.94	intergenic
20:47150428	6	0.91	0.93	intergenic
6:129548839	17	0.90	0.92	LAMA2
1:54070831	29	0.90	0.92	GLIS1
3:102797474	6	0.90	0.90	intergenic

Figure 4.1.: Frequencies of representative SNPs of top PBS clusters.



4.3.2. Signatures of positive selection from extended haplotypes

Many signals of very recent positive selection have been highlighted by EHH-based tests (iHS and XP-EHH). This is also true with Sardinian genomes. As in the case of PBS analysis, we adopted a similar cluster-based approach for the identification of footprint of positive selection with iHS and XP-EHH. Clusters were built by considering only the top 0.1% scores for each test.

Among the top iHS clusters (a short list with top 10 is available in Table 4.4), the most dense signals are the following:

1. a cluster of 5,000 base pairs on chromosome 10, represented by the SNP 10:56485762, is located within the intron of the PCDH15 gene; the cluster is composed by 65

strong outliers, which are about 80% of the SNPs in the region; interestingly, as reported by Akey in 2009 [73], in a rare example of multiple analyses converging on a single gene, PCDH15 was identified in six out of the nine genome-wide scans. Mutations in this gene, important for retinal and cochlear function, can result in Usher syndrome type IF and Autosomal Recessive Deafness 23 [74].

2. a cluster in chromosome 4, about 26,5 Kb long, centered on the top SNP 4:78243065 and composed by 224 SNPs (84% of outliers), is located within the promoter region of the chemokine CXCL13.

Other clusters point to genes involved in olfactory functions.

Table 4.4.: Top iHS clusters in coding genes

top SNP	cluster size	median std. iHS	top std. iHS	Genes
3:101119754	48	5.091585	10.978	SENP7
10:56485762	65	9.53881	10.5657	PCDH15
3:120708493	98	5.07435	10.3158	STXBP5L
4:78243065	224	7.00579	10.1279	promoter CXCL13
1:222987443	30	6.21526	9.37278	DISP1
6:32751349	1282	4.796125	8.66354	HLA region
2:201885727	19	4.37023	7.71575	ORC2, FAM126B
9:108552253	58	6.553925	7.64748	proximal to TMEM38B
2:44092618	93	6.11154	7.33934	ABCG8
4:170428901	22	4.44103	7.04864	NEK1

Probably the most relevant results - in the context of the two historical plagues of Sardinia island, malaria in the past and autoimmunity in recent years - are revealed by the cross-population EHH (xp-EHH) test, which discover alleles that have swept to near-fixation within Sardinians when compared to Tuscans. Among the top 10 clusters the following are particularly worthy of mention:

1. the top signal is a relatively small cluster located within one intron of CNTNAP2, a key neurally expressed gene, which has been reported in a positive selection survey in European populations [75].

2. a very interesting cluster, centered on the SNP 4:144679704 and spanning on about 400Kb, is located within a Malaria protective locus recently discovered, in particular on the GYPA gene. The protective role of this locus was recently discovered by an international team with a GWAS on about 14,000 malaria cases and controls from The Gambia, Ghana, Malawi, Tanzania, Burkina Faso, and Cameroon [76]. Glycophorin genes encode proteins that are receptors for erythrocyte invasion by P. falciparum. Interestingly, this locus, that provides 33% protection against severe malaria, has been previously linked to ancient balancing selection, on the basis of haplotype sharing between humans and chimpanzees. A recent preprint [77] reported that this association with severe malaria is explained by a complex structural variant that involves the loss of GYPB and gain of two hybrid genes, each with a GYPB extracellular domain and GYPA intracellular domain.

   Analyses performed in 2002 by Baum et al [78] on GYPA sequence variation among six higher primates and within a human population, have shown a large excess of nonsynonymous substitutions along each primate lineage and a significant excess of polymorphisms in exon 2 of this gene within humans. These two signatures suggested a strong positive selection on this receptor driving GYPA divergence during primate evolution and balancing selection maintaining allelic variation within human populations.

   In an attempt to identify between human and other primates the most fast-evolving genes among 280 genes, in 2003 Wang et al [79] identified that the three glycophorins, GPA, GPB, and GPE, have the highest rate of nonsynonymous substitutions among the surveyed (by performing a Ka/Ks test), this indicating positive selection. They first hypothesized that GPA has been evolving rapidly to evade malaria parasites.

   Ko et al. [80] in 2011 identified an excess of genetic variation in the coding region of GYPA, and this signature of selection was found to be strongest in African populations with the highest levels of malaria exposure.

   These findings are confirmed by a global overview of the number of outlier signals for the iHS test (see Table 4.6): the stronger presence of extended haplotypes is evident for african populations, with Sardinians being an exception among the European populations.

Table 4.6.: Number of iHS outliers in the region of glycophorin genes (chr4:144M-145,5M)

POP	BLOCK	iHS outliers
African Caribbeans in Barbados	Africa	260
Americans of African Ancestry in SW USA	Africa	138
Esan in Nigeria	Africa	340
Gambian in Western Divisions in the Gambia	Africa	230
Luhya in Webuye, Kenya	Africa	284
Mende in Sierra Leone	Africa	203
Yoruba in Ibadan, Nigeria	Africa	170
Colombians from Medellin, Colombia	America	77
Mexican Ancestry from Los Angeles USA	America	10
Peruvians from Lima, Peru	America	36
Puerto Ricans from Puerto Rico	America	28
Chinese Dai in Xishuangbanna, China	East_Asia	26
Han Chinese in Bejing, China	East_Asia	40
Japanese in Tokyo, Japan	East_Asia	33
Kinh in Ho Chi Minh City, Vietnam	East_Asia	70
Southern Han Chinese	East_Asia	27
British in England and Scotland	Europe	39
Finnish in Finland	Europe	38
Iberian Population in Spain	Europe	20
Sardinians	Europe	73
Toscani in Italia	Europe	25
Utah Residents	Europe	28
Bengali from Bangladesh	South_Asia	35
Gujarati Indian from Houston, Texas	South_Asia	18
Indian Telugu from the UK	South_Asia	14
Punjabi from Lahore, Pakistan	South_Asia	60
Sri Lankan Tamil from the UK	South_Asia	10

3. other two clusters are worthy of mention because they are autoimmunity related loci. The first one encompass the promoter region of IL23R, a common susceptibility genetic factor in autoimmunity[3]; the second one encompasses the whole region of the TNFRSF13B gene, which encodes the protein TACI, one of the three receptors of soluble BAFF (see Chapter 3) [81].

The last test we used, nSL [36], has been developed for detecting both soft and hard sweeps in population genomic data. Of the top cluster from this last analysis (see Table 4.9), probably the only one of interest is the signal downstream to the HGF gene. Indeed, wounding of hepatocytes by sporozoite migration induces the secretion of hepatocyte growth factor (HGF), which renders hepatocytes susceptible to infection by Plasmodium. HGF/MET signaling induces rearrangements of the host-cell actin cytoskeleton that are required for the early development of the parasites within hepatocytes [82].

[3] See https://www.immunobase.org/gene/ENSG00000162594/

Table 4.7.: Top xp-EHH clusters in coding genes

top SNP	cluster size	median std. xp-EHH	top std. xp-EHH	Genes
7:146523947	34	3.33021	6.26188	CNTNAP2
17:44173215	819	3.76572	5.66862	KANSL1, STH, MAPT, CRHR1
5:266041	117	4.68397	4.92245	SDHA, PDCD6
16:76540194	5	4.74565	4.83897	CNTNAP4
4:144679704	410	3.500535	4.63192	GYPA, BC029578
1:67588901	18	3.37521	4.56117	IL23R, C1orf141
16:18886875	151	3.85348	4.54974	SMG1
15:45070913	103	3.61738	4.54316	TRIM69
19:20756966	100	4.156955	4.53314	ZNF737
17:16855222	50	3.554355	4.52902	TNFRSF13B (TACI)

Table 4.9.: Top nSL clusters in coding genes

top SNP	cluster size	median std. nSL	top std. nSL	Genes
1:13176234	317	5.41738	7.75336	PRAMEF family
21:10868983	306	4.538455	7.71672	TPTE, AK311573
14:19031620	83	6.03623	7.61021	olfactory receptors
14:20337774	101	5.96731	7.58169	olfactory receptors
7:81107705	87	4.99852	6.70674	downstream to HGF
2:44092497	83	5.19077	6.58788	ABCG8
7:149926945	219	4.91783	6.45638	ACTR3C
2:132228140	12	4.236125	6.30277	TUBA3D
10:47107016	45	4.84926	5.97318	ANXA8, NPY4R, LINC00842
3:129851796	81	4.61229	5.86599	FAM86HP
6:32624586	189	4.11617	5.80937	HLA-DQA1 region

4.3.3. Quantitative trait loci with evidences of positive selection in the Sardinian population

With the aim to uncover traits whose associated variants show signatures of positive selection, we extracted SNPs with significant association (p-value <= 1e-8) with any trait from 4 GWAS resources (ProgeNIA [67] [51], GWAS Catalog and ImmunoBase) and checked how many of them were strong outliers (here we mean among the top 0,1% genome wide outliers for each test of positive selection, or in LD with any of these outliers - considering an r2 $\geq$ 0.8).

Results are listed in Table 4.10, where for each trait are listed the loci putatively under selection (for each locus, the number of associated SNPs coincident with signals of selection is reported).

Among the 67 traits, about 35% of them were associated with SNPs from the ATXN2/SH2B3 locus; the pleiotropic effect of the protein products of this locus is extensively studied, and here we show that these variants should be sitting on an extended haplotype under positive selection. Another redundant locus is revealed by SNPs in the SLC45A2 gene, all of them associated with traits such as skin pigmentation

and hair color; this observation is in line with the results of the paper that introduced the xp-EHH method [35].

A deeper investigation would be needed for the xp-EHH signal located in NFE2, which encodes erythroid nuclear factor 2; in this locus, here under selection in Sardinians and associated with the trait "Mean platelet volume", suggestive signals of association with regulation of hemoglobin levels were found by Danjou et al [47].

MLPH, here associated with Prostate Cancer, is known to influence skin pigmentation and to have a strong xp-EHH signature in non African population [83].

An interesting pleiotropic effect is observed for the SNPs 12:56737973 (STAT2), a strong outlier for PBS associated with both Psoriasis and Height.

Notably, about 25% of the traits are autoimmunity related.

Table 4.11.: Traits with positive selection signatures

TRAIT	iHS	XP-EHH	nSL	PBS
Alopecia Areata	SH2B3/ATXN2 (1)	-	-	-
Autoimmune hepatitis type-1	SH2B3/ATXN2 (1)	-	-	-
Beta-2 microglubulin plasma levels	SH2B3/ATXN2 (1)	-	-	-
Bitter taste perception	PRH1-PRR4 (3)	-	-	-
Hair color	-	SLC45A2 (1)	-	-
Blood metabolite levels	SH2B3/ATXN2 (1)	-	-	-
Blood pressure	SH2B3/ATXN2 (1)	-	-	-
Blood pressure measurement (low sodium intervention)	-	-	-	PIBF1 (1)
Bone mineral density	-	LRP5 (1)	-	-
Celiac dis.	SH2B3/ATXN2 (2)	downstr. TNFSF18 (1)	-	-
Chronic kidney dis.	SH2B3/ATXN2 (1)	-	-	-
Colorectal or endometrial cancer	SH2B3/ATXN2 (1)	-	-	-

-				
Coronary artery dis.	SH2B3/ATXN2 (1)	-	-	-
Coronary heart dis.	-	-	STK32B (1)	-
Crohn's dis.	SH2B3/ATXN2 (1), KIAA1109 (1), IPMK (4)	downstr. TNFSF18 (8)	-	-
Diastolic blood pressure	SH2B3/ATXN2 (2)	-	-	-
Dupuytren's dis.	-	downstr. c22orf26 (2)	-	-
Eosinophil counts	SH2B3/ATXN2 (1)	-	-	-
Eye color	-	SLC45A2 (2)	-	-
Fibrinogen levels	SH2B3/ATXN2 (1)	-	-	-
Hair color	-	SLC45A2 (2)	-	-
HDL cholesterol	-	promoter ADH5 (1)	-	-
Height	-	-	-	STAT2 (1)
Helicobacter pylori serologic status	-	promoter TLR10 (4)	-	-
Hematological parameters	SH2B3/ATXN2 (1)	-	-	-
Hypertension	intergenic (3)	-	-	-
Hypothyroidism	SH2B3/ATXN2 (1)	-	-	-

Inflammatory bowel dis.	SH2B3/ATXN2 (2), KIAA1109 (1), IPMK (4)	downstream TNFSF18 (1)	-	-
Inflammatory skin dis.	-	-	-	STAT2 (2)
Juvenile idiopathic arthritis	SH2B3/ATXN2 (1)	-	-	-
Lipoprotein (a) levels	-	-	-	downstr. ARID1B (1)
Mean platelet volume	-	NFE2 (4)	-	-
Motion sickness	-	HOXB3 (1)	-	-
Myocardial infarction	SH2B3/ATXN2 (1)	-	-	-
Non-albumin protein levels	-	TNFRSF13B (1)	-	-
Obesity	-	HOXB5 (1)	-	-
Primary Biliary Cirrhosis	-	MAPT (9)	-	-
Perceived skin darkness	-	SLC45A2 (1)	-	-
Platelet count	SH2B3/ATXN2 (2)	COPZ1 (1)	-	-
Post-traumatic stress disorder	-	-	-	TLL1 (1)
Prostate cancer	MLPH (2)	-	MLPH (2)	LOC727677 (1)

Primary Sclerosing Cholangitis	SH2B3/ATXN2 (1)	-	-	-
Psoriasis	-	-	-	STAT2 (1)
Red blood cell traits	SH2B3/ATXN2 (1)	-	-	-
Retinal vascular caliber	SH2B3/ATXN2 (1)	-	-	-
Rheumatoid arthritis	SH2B3/ATXN2 (2)	-	-	-
Rheumatoid arthritis (ACPA-negative)	-	-	-	intergenic (2)
Skin colour saturation	-	SLC45A2 (1)	-	-
Skin pigmentation	-	SLC45A2 (2)	-	-
Systemic lupus erythematosus	SH2B3/ATXN2 (2)	-	-	-
Systolic blood pressure	SH2B3/ATXN2 (1)	-	-	-
Tanning	-	SLC45A2 (2)	-	-
TB_trait5 [51]	SH2B3/ATXN2 (1)	-	-	-
TB_trait50 [51]	SH2B3/ATXN2 (1)	-	-	-
TB_trait8 [51]	SH2B3/ATXN2 (1)	-	-	-

Thyroid peroxidase antibody levels	SH2B3/ATXN2 (1)	-	-	-
Thyroid peroxidase antibody positivity	SH2B3/ATXN2 (1)	-	-	-
TR_trait115 [51]	HLA region	-	-	-
TR_trait117 [51]	HLA region	-	-	-
TR_trait3 [51]	HLA region	-	-	-
Type 1 diabetes	SH2B3/ATXN2 (1)	-	-	-
Type 1 diabetes autoantibodies	SH2B3/ATXN2 (1)	-	-	-
Ulcerative colitis	KIAA1109 (1)	-	-	-
Urate levels	SH2B3/ATXN2 (1)	-	-	-
Vitiligo	SH2B3/ATXN2 (2)	-	-	-

4.4. Conclusions

The simultaneous availability of a detailed description of the common variability in the Sardinian population and the dissection of genetic basis of many quantitative/qualitative variables, jointly with genome wide scan for positive selection, has made possible an unprecedented analysis in such a large population. The results of such analysis can be summarized by two messages: firstly, the majority of the traits with signatures of positive selection are immunity and/or autoimmunity related (this finding being in line with the observations published by Barreiro in 2010 [84]); secondly, new evidences of recent selection driven by Malaria pathogens are highlighted by the extended haplotypes in the glicophorins locus. Apart of this general analysis, the resource here presented will be surely a valuable tool for many other genome wide associations studies in the Sardinian population.

Appendix A.

Sardinia, the *unhealthy island*

The Sardinian island population has been the focus of many classical natural selection studies [85]. This is largely due to the fact that this isolated ancient population has been severely exposed to malaria both from Plasmodium Vivax and Plasmodium Falciparum. Indeed, signals of natural, mainly balanced selection have been attributed to variants involved in haematological diseases such as beta and alpha Thalassemia and G6PD deficiency*G6PD* deficiency, thalassemia or Wilson disease [85] [86] [87]. More recentlym genome wide array-based data have suggested evidence of positive slection at the CR1 locus. In addition to the variants involved in haematological diseases there is also recent evidence of variants that have been positively selected because they conferred resistance to pathogens, again most likely to malaria, and that are now a risk factor for autoimmunity (see discussion about citokyne sBAFF, Chap. 3).

Indeed, a peculiarity of Sardinia is the unhappy coincidence of two different but probably related facts: a past long history of malaria [55] and a recent high incidence of autoimmune disease [88] [46].

Malaria is believed to have been introduced to Sardinia by infected workers imported from North Africa after the Carthaginian conquest of Sardinia in 502 bc. According to Tognotti E. [55], such plague became endemic to this region during the medieval age, but since the classical ages the islands and their people had a status as an *unhealthy island*. Sardinia kept the unfortunate primacy of being the most malaria-ridden region in Italy because of the high prevalence of P. falciparum and its associated high mortality rates. Rates were particularly high for children <5 years of age in highly malaria-endemic areas.

An evolutionary scenario that we could hyphotize is that several mutations, now predisposing to autoimmunity, were selected as part of a selective response to improve the genetic fitness to combat overwhelming malaria infection or other pathogens.

A link between the disappearance of malaria with the increase of multiple sclerosis (MS) in Sardinia has been proposed [89]. As stated by the hygiene hypothesis, the establishment of new hygienic conditions could have played a role in the appearance of autoimmunity in *westernalised* countries. In recent years, multiple genome scans for signatures of selection on common variation have identified many immune-related loci [84] [90] [91] [92].

Bibliography

[1] D. N. Reznick and R. E. Ricklefs, Nature **4**57, 837 (2009).

[2] S. Brusatte, J. O'Connor, and J. E.D., Curr Biol. (2015).

[3] H. G.H., Science (1908).

[4] W. W., Jahreshefte des Vereins fÃijr vaterlÃd'ndische Naturkunde in WÃijrttemberg (1908).

[5] L. Duret, Nature Education (2008).

[6] M. Kumura, Nature (1968).

[7] R. Nielsen, Annu Rev Genet. (2005).

[8] D. Kwiatkowski, Am J Hum Genet. (2005).

[9] J. Haldane, Hereditas (1949).

[10] V. Mangano and D. Modiano, Curr Opin Immunol. (2014).

[11] M. Siniscalco and L. Bernini, Bull World Health Organ. (1966).

[12] F. Piel, Nature Communications (2010).

[13] S. Eridani, Hematol Rep. (2011).

[14] M. G. E. Network, Nature (2015).

[15] F. Hsieh, Nat Commun. (2016).

[16] M. Nagaraj, Nat Commun. (2013).

[17] K. Omi, Am J Hum Genet. (2003).

[18] Y. Itan, PLoS Comput Biol. (2009).

[19] T. Bersaglieri, Am J Hum Genet. (2004).

[20] C. Jeong and A. Di Rienzo, Curr Opin Genet Dev. (2014).

[21] M. Fumagalli *et al.*, Science (2014).

[22] X. Yi *et al.*, Science (2010).

[23] A. Allison, Br Med J. (1954).

[24] E. Moresco, Am J Pathol. (2013).

[25] P. Messer and D. Petrov, Trends in Ecology and Evolution (2013).

[26] J. Pritchard, J. Pickrell, and G. Coop, Current Biology (2010).

[27] R. Nielsen, Nature Reviews Genetics (2007).

[28] G. Bhatia, N. Patterson, S. Sankararaman, and A. Price, Genome Res. (2013).

[29] N. Bierne, D. Roze, and J. Welch, Mol Ecol. (2013).

[30] L. PortoNeto, S. Lee, H. Lee, and C. Gondro, Methods Mol Biol. (2013).

[31] M. Akey, Genome Res. (2002).

[32] K. Holsinger and B. Weir, Nature Rev. Genet. (2009).

[33] P. Sabeti, Nature (2002).

[34] B. Voight, S. Kudaravalli, X. Wen, and J. Pritchard, PLOS Biology (2006).

[35] P. Sabeti, Nature (2007).

[36] A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen, Mol. Biol. Evol. (2014).

[37] N. Garud, PLOS Genetics (2015).

[38] N. Garud, Theoretical Population Biology (2015).

[39] S. Grossman, Science (2010).

[40] R. Ronen, Genetics (2013).

[41] E. Han, Bioinformatics (2014).

[42] A. Nekrutenko, Genome Res. (2002).

[43] R. Plenge, Nature Reviews Drug Discovery (2013).

[44] M. Gutierrez-Arcelus, Nat Rev Genet. (2016).

[45] C. Sidore, F. Busonero, A. Maschio, E. Porcu, and S. Naitza, Nat Genet. (2015).

[46] M. Pugliatti *et al.*, European Journal of Neurology (2006).

[47] F. Danjou, M. Zoledziewska, C. Sidore, M. Steri, and F. Busonero, Nat Genet. (2015).

[48] M. Zoledziewska, C. Sidore, C. W. Chiang, S. Sanna, and A. Mulas, Nat Genet. (2015).

[49] C. Sardu, Plos One (2012).

[50] M. Marrosu, Lancet (2002).

[51] V. Orru, M. Steri, G. Sole, C. Sidore, and V. F., Cell (2013).

[52] I. Moisini, Clin Exp Immunol. (2009).

[53] T. . G. P. Consortium, Nature (2015).

[54] A. Szpiech and R. Hernandez, Mol Biol Evol. (2014).

[55] E. Tognotti, Emerg Infect Dis. (2009).

[56] X. Liu, Eur J Immunol. (2012).

[57] A. Wood, Nat. Genet. (2014).

[58] J. Berg and G. Coop, PLoS Genet. (2014).

[59] J. L. Crisci, M. D. Dean, and P. Ralph, Molecular Ecology (2016).

[60] M. Arcos-Burgos and M. Muenke, Clin Genet. (2002).

[61] L. Peltonen, A. Palotie, and K. Lange, Nat Rev Genet. (2000).

[62] C. Pedersen *et al.*, Biorxiv (2016).

[63] J. M. Akey *et al.*, PLoS Biol. (2004).

[64] F. Clemente *et al.*, Am J Hum Genet. (2014).

[65] I. Moltke *et al.*, Nature (2014).

[66] I. Piras *et al.*, European Journal of Human Genetics (2012).

[67] G. Pilia, W. Chen, A. Scuteri, M. Orru, and G. Albai, PLOS Genetics (2006).

[68] P. Danecek, Bioinformatics (2011).

[69] T. U. Consortium, Nature (2015).

[70] L. Francioli and A. Menelaou, Nature Genetics (2014).

[71] B. Weir and C. Cockerham, Evolution (1984).

[72] J. Reynolds, B. Weir, and C. Cockerham, Genetics (1983).

[73] J. Akey, Genome Res. (2009).

[74] Z. Ahmed *et al.*, Hum. Mol. Genet. (2003).

[75] Q. Ayub, B. Yngvadottir, and Y. Chen, The American Journal of Human Genetics (2013).

[76] M. G. E. Network, Nature (2015).

[77] M. Leffler, Biorxiv (2016).

[78] J. Baum, R. Ward, and D. Conway, Mol. Biol. Evol. (2002).

[79] H. Wang, H. Tang, C. Shen, and C. Wu, Mol. Biol. Evol. (2003).

[80] W. Ko, K. Kaercher, and E. Giombini, Am J Hum Genet. (2011).

[81] H. Jacobs, C. D. Thouvenel, S. Leach, and A., The Journal of Immunology (2016).

[82] M. Carrolo *et al.*, Nat Med (2003).

[83] J. Pickrell, Genome Res. (2009).

[84] L. Barreiro and L. Quintana-Murci, Nat Rev Genet (2010).

[85] M. Siniscalco, L. Bernini, B. Latte, and A. G. Motulsky, Nature (1961).

[86] M. Rosatelli *et al.*, Hum Genet. (1992).

[87] G. Loudianos *et al.*, Hum Mutation (1999).

[88] M. Karvonen, J. Tuomilehto, I. Libman, and R. LaPorte, Diabetologia (1993).

[89] S. Sotgiu, A. Sannella, B. Conti, and G. Arru, Journal of Neuroimmunology (2007).

[90] S. Ramos, S. Shaftman, R. Ward, and C. Langefeld, Autoimmune Dis. (2014).

[91] L. Jostins *et al.,* Nature (2012).

[92] T. Raj *et al.,* The American Journal of Human Genetics (2013).

List of figures

List of tables