Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Biodiversità ed Evoluzione

Ciclo XXVIII

Settore Concorsuale di afferenza: 05/B1 Zoologia e Antropologia

Settore Scientifico disciplinare: BIO/08 Antropologia

Detecting signatures of genetic adaptation to climate and nutrition in European population

Presentata da: **Andrea Quagliariello**

Coordinatore Dottorato

Prof.ssa **Barbara Mantovani**

Relatore

Prof.ssa **Donata Luiselli**

Correlatore

Dott. **Marco Sazzini**

Esame finale anno 2016

Table of contents

Chapter 1

Introduction

1.1 Studying Natural Selection: From first theoretical models to recent Genome-Wide studies

Natural selection is the evolutionary force able to shape patterns of human genetic variation. The detection of regions which have been targets of such force along human genome is of primary importance to investigate the human evolutionary history and so to understand the causes that have driven the pattern of genetic variability observed nowadays within different human populations.

Particularly, one of the major aims of the studies on human evolution is to identify loci that could have experienced the action of natural selection due to their influence on individuals fitness, thus on those genomic regions that are (or maybe have been) of functional importance for species survival.

In fact, the fate of every new mutation arisen within human genome depends on how this variable may affects the individual fitness when exposed to the particularly environmental conditions where the owner of the cited mutation lives. According to the neutral theory of evolution, most parts of the new molecular variations are "neutral", meaning that they did not have any appreciable effect on individual fitness (Kimura, 1968; King & Jukes, 1969; Arnheim & Taylor, 1969). Therefore, in the light of this theory, most part of genetic polymorphisms and the changes in their frequencies in a population of a finite size are

just governed by a stochastic effect of population drift. The importance of this theory is that it gives an exact quantitative prediction about what should be the expected patterns of genetic variation both between than within species. Thus, practically, it gives the null hypothesis when evidences of natural selection are tested (Charleswoth et al., 1995; Otto, 2000; Nielsen, 2001).

To date, several theoretical studies have highlighted three main kinds of selective pressures: positive, purifying and balancing selection (fig. 1.1.1). Positive selection increases the frequency of those alleles that constitute an advantage for population survival. When particular environmental conditions occurred, they can act as a strong selective pressures which can bring a specific advantageous mutation to reach fixation within the population, the so-called "selective sweep" (fig. 1.1.2).

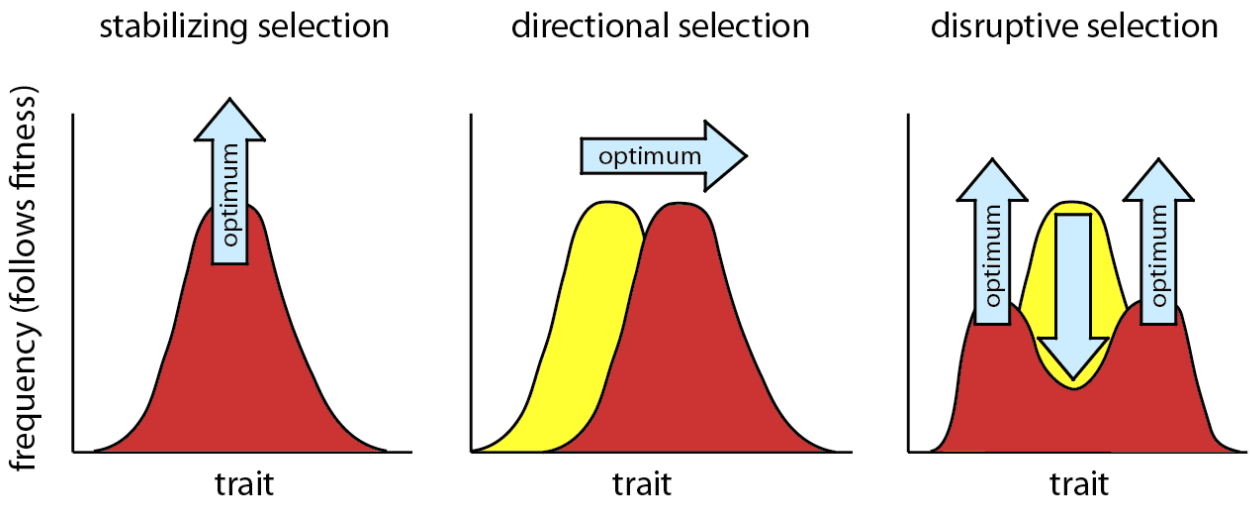


Figure 1.1.1: Schematic representation of the different kind of natural selection (Loewe, 2008)

In some cases, positive selection, may also influence the frequency of other neutral polymorphisms according to a phenomena named “hitch-hiking”, where their frequencies are influenced not from direct action of natural selection, but from their linkage level with the core SNP under positive selection (Thomson, 1977; Hedrick, 1980; Kaplan et al., 1989; Nash et al., 2005). Generally speaking, signatures of positive selection are detectable from a reduced level of variation in comparison to a neutral model and when elevated levels of linkage disequilibrium (LD) are found (Fay & Wu, 2000; Sabeti et al., 2002).

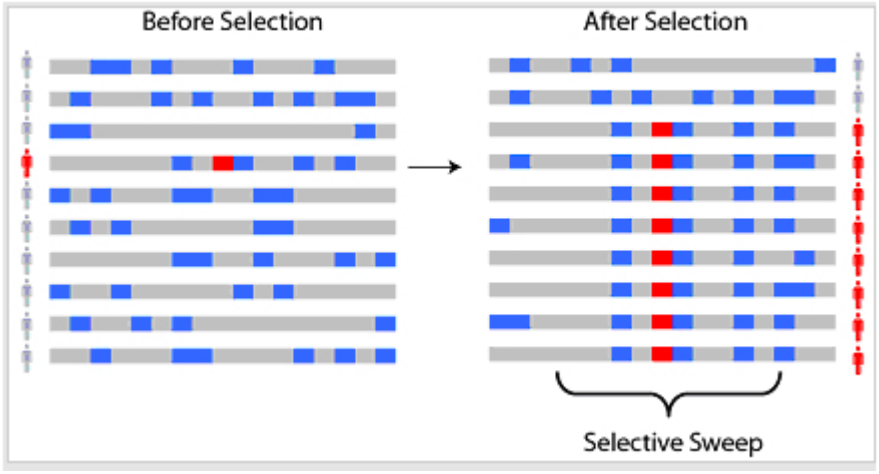


Figure 1.1.2: When a new advantageous mutation appears within a population genetic pool it will reach high frequency in a short time. Ancestral alleles are here represented in grey while derived in blue. A new derived mutation is indicated in red colour and its advantageous characteristic let it rises almost to fixation within the studied population (Schaffner et al., 2008)

On the other hand balancing selection takes place when natural selection maintain both the existing polymorphisms. Thus, in this case, signals of selection are observable from the elevated levels of variability than that expected under a neutral model, and from an elevated number of alleles with an intermediate frequency level (Kelly & Wade, 2000; Schmidt et al, 2000).

Purifying selection is at the other hand of the spectrum of positive selection. Sometimes new mutations that arise within genome could be and harmful variation, then purifying selection acts on them to maintain the general stability, thus to not let them affect individual fitness (Loewe, 2008). A simple and direct example of this assumption is provided by the so-called "living-fossil" species: when for a particular species, fossils of millions of years are more or less indistinguishable from modern living heirs of that same species, it could be a proof that purifying selection has acted on them, avoiding genetic changes (Vane-Wright, 2004). This condition is allowed when the ecological niche, where that particular species have bloomed, is maintained identical through the centuries, but when the environmental circumstances are changed some polymorphisms could undergo the effect of purifying selection, even if for a short-term period. Thus, the major consequence of negative selection is that the less-adapted variants are lead to the extinction, while if as in the fossil example, the environmental conditions are not changed, the purifying selection will just remove all new variants for that optimal trait.

Moreover as important as natural selection for its influence on the pattern of genetic variability is the action of population demography. These two forces are often confounded due to their similar result on population alleles frequencies. In fact, population demography may "mimic" the effect of natural selection, for example both positive selection and a strong increase in population size are characterized by the presence of numerous low-frequency alleles within the population (fig. 1.1.3). On the same way it is possible to confound population structure with the effect of balancing selection cause both bring to an excess of intermediate-frequency alleles (Ronald & Akey, 2005).

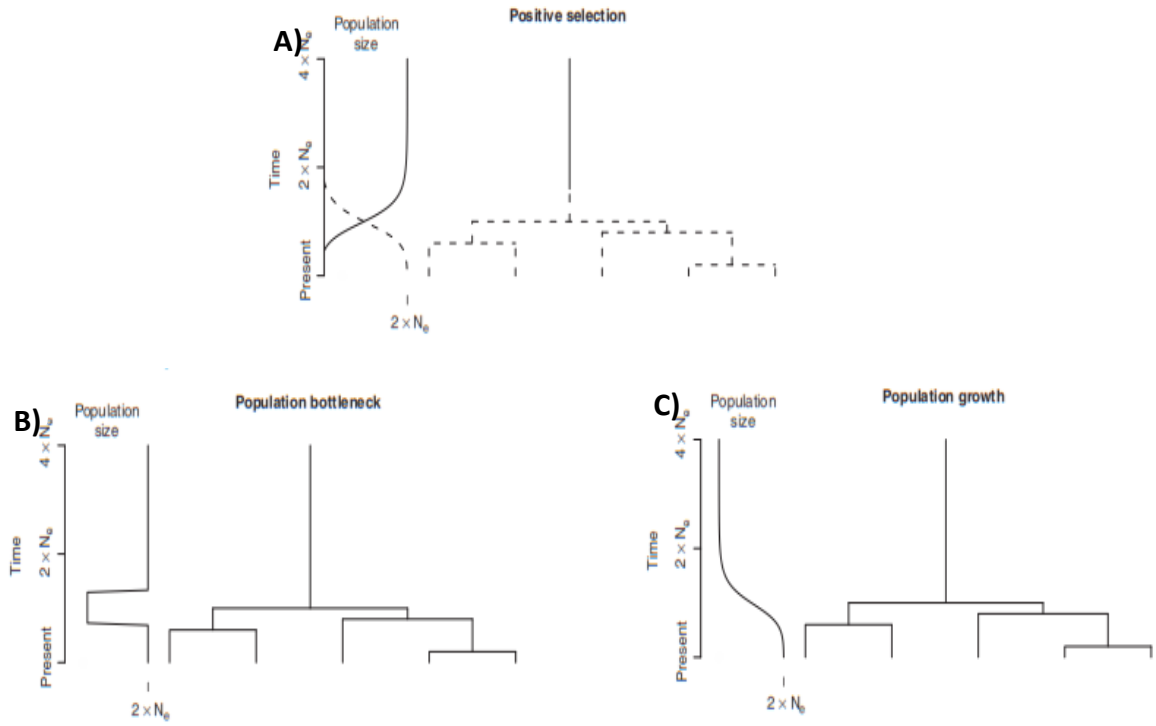


Figure 1.1.3: Effect of positive selection (A) and population demography (B,C) on the pattern of population genetic. In this case of positive selection an advantageous allele (indicated by the dashed line) appear within the population and rapidly rise to fixation. A similar scenario is even possible when the population experienced a population bottleneck or a population grow (Ronald & Akey, 2005).

To date, several methods have been produced to detect natural selection signals and further development have been carried on, thanks the contribution of genetic information due to new high-throughput methods for identification of genetic variation (Kreitman, 2000; Ronald &Akey, 2005).

These new methods have been applied on a genome-wide scale for an increasing number of populations, thus giving a complete idea of global world-wide pattern of variability. Furthermore, these studies have received recently a strong contribution from the birth of new projects that have collected complete genetic data from numerous populations, distributed world-wide: Human Genome Diversity Project (*HGDP*), *HapMap* project, *Genographic* and *1000Genomes* (The International HapMap Consortium, 2005;Cann et

al., 2002; Chitty, 2006; Buchanan et al., 2012; Altshuler et al., 2012; Zhang et al., 2015) (fig. 1.1.4).

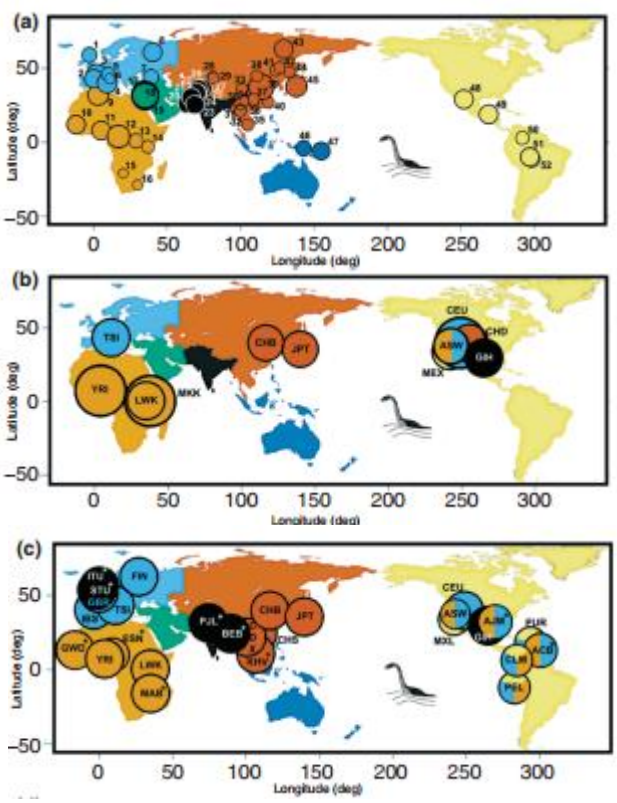


Figure 1.1.4: a) HGDP, b) *HapMap* and c) *1000Genomes*. Samples are indicated by circles of a dimension proportional to sample size. Colours on the map represent the geographical ancestry of the samples (Colonna et al., 2011).

At this point, it was possible to start analysing the genetic relationships and ancestry and to detect which are the major loci under the action of natural selection at a global level. The purpose of such international project as *1000Genomes* is mainly to create a deep catalogue of human genetic variation, that could be used in different type if genetic research studies. They made up a global estimation of population allele frequencies, haplotype inference and LD (Linkage Disequilibrium) analysis to address evolutionary

questions that were inaccessible to the common analysis of uniparentally inherited markers (Colonna et al., 2011).

As first, whole-genome data, confirmed what was already known from several studies about the global pattern of human variation, thus that African populations show a higher variability than all the other non-African ones. Moreover, the non-African populations carry an old fraction of the common African genetic background (The 1000 Genomes Project Consortium, 2010), corroborating the idea of an African origin of human species, with a progressive serial founder effect during the expansion out of Africa. Furthermore several evidences highlighted the presence of small genetic differences among populations, principally distributed with a cline of frequencies, and so changing gradually from place to place (fig. 1.1.5).

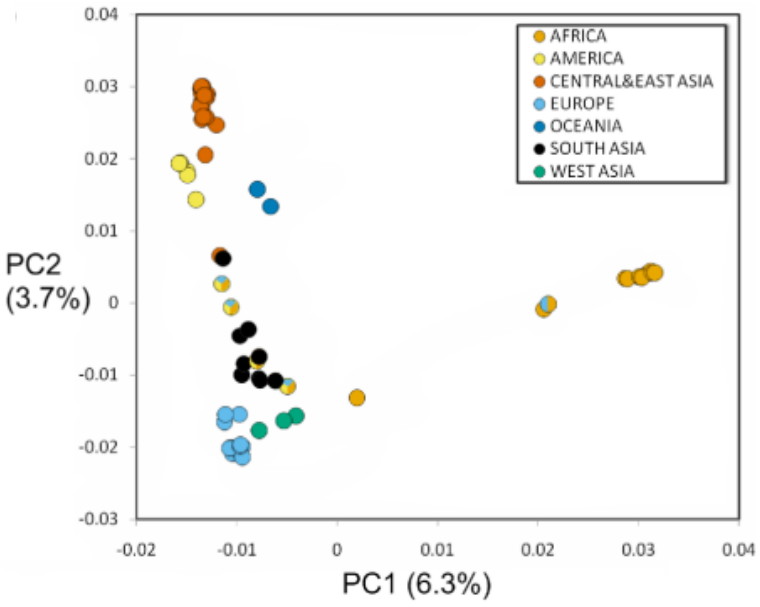


Figure 1.1.5: Principal component analysis (PCA) analysis showing the relationship among world-wide population samples

Extreme changes in frequencies between close populations are rare and generally due to geographical barriers that could have constituted a limit to the migration processes (Currat et al., 2010).

Analyses on the whole genome-wide dataset emphasize the role played by migration on human history, particularly of dispersal out of Africa process, revealing a negative correlation between genetic diversity and migration distance from Africa (Li et al., 2008; Luca et al., 2011). This decreasing in nucleotide diversity level moving away from the African continent (fig. 1.1.6), is explained by a series of founders effect: a series of subgroups division from a source population with a very high effect of genetic drift.

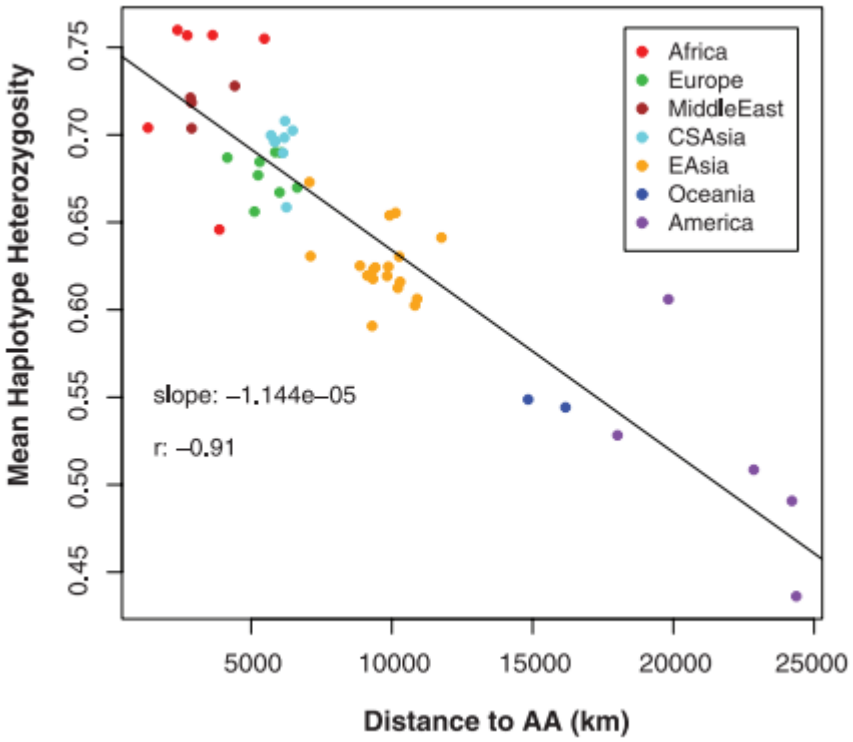


Figure 1.1.6: Nucleotide diversity decreases with distance from Africa.

1.2 The influence of natural selection on the process of local adaptation

A further interesting observation that emerged from these large-scale studies is that local adaptation processes have occurred in recent human history more than previously hypothesised. In fact, during these several migration processes over the last 100,000 years, humans subgroups have been exposed to new environmental conditions characterized by different nutritional and climatic backgrounds which have left their genetic marks (Fumagalli & Sironi, 2014). This assumption is confirmed by several studies which found more evidences for natural selection within non-African populations than within African ones (Kayser et al., 2003;Akey et al., 2004; Storz et al., 2004). A part from the skin pigmentation-related genes (i.e. *SLC24A5*, *SLC45A2*), several genes participating in various nutritional pathways have been found to be under positive selection in different human groups. For what it concerns dietary metabolism, one of the most famous is surely that of lactase persistence *LCT* gene, encoding the enzyme lactase, and with a polymorphism promoting its expression which is found at high frequency in European populations and almost absent elsewhere (Bersaglieri et al., 2004; De Fanti et al., 2015a). Another interesting signal of local adaptation process came from *PLRP2* gene which encodes for a pancreatic lipase-related protein, within a single nucleotide polymorphism that is found to result in a protein truncation which is strongly correlated to cereals consumption (fig. 1.2.1) (Hancock et al., 2010).

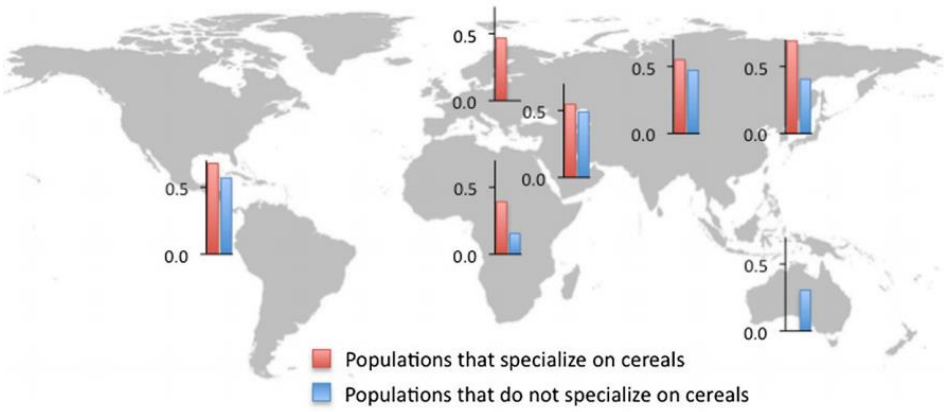


Figure 1.2.1: Average frequencies for *PRLP2* rs4751995 across populations (Hancock et al., 2010)

A comparative analysis highlighted the presence of this same protein within non-ruminant herbivore pancreas, but it is completely absent within carnivores, thus providing further elements to confirm the role of this protein in a plant-based diet. In fact, this variation is correlated even in humans to a more active enzyme to face to primarily cereals intake, therefore representing a strong example of adaptation to a specialized diet (Berton et al., 2007).

Moreover, other enzymes related to metabolism processes have been target of selective pressures always linked to dietary changes such as *NAT2*. This gene encodes for an N-acetyltransferase enzyme, which is involved in metabolism of toxic compounds that could be present in several foods resources. Luca and colleagues found that various fast acetylator alleles are more common within populations with a hunter-gatherers life-style than within farmers societies (Luca et al., 2008). Thus these gene variations are maybe the results of a reduced folate intake in diet after the Neolithic transition.

Before to be introduced and digested by the body, food is undergoing to several processing steps, during which it is seen, smelled and finally tasted. All the five senses

are employed during this process of "knowledge" of food proprieties. Thus is not surprising that even genes involved in sensory perception have been demonstrated to be as well exposed to the action of natural selection, due to their linkage with various feedings habits (Gilad et al., 2003; Ficher et al., 2005;Verrelli et al., 2008). One of the most studied gene in this field was the *TAS2R38* gene, associated with bitter-taste perception. Two principal haplotypes are associated with high-perception or no-perception phenotypes of bitter taste (also named as "taster" and "non-taster"), that showed a pattern of intermediate-level frequency among different populations. These findings bring researchers to the conclusion that there could be an heterozygote vantage maintained by the action of balancing selection in several ethnic groups (Woodings et al., 2006). This distribution could be related to the ability conferred by the heterozygote condition to detect a larger number of toxins within aliments even if, to date, no functional data are available. A similar condition is found even for *TAS2R16* gene, another bitter-taste associated gene for which signals of natural selection have been detected. Particularly the presence of the derived allele at a specific locus seems to confer a higher sensitivity to identify different glycosides. This allele seems to have been originated time before agricultural transition and out of Africa migrations, indeed there are evidences of positive selection in East Africa. Thus researchers supposed that it would have conferred the ability to detect possible toxic molecules presents in foods and so to protect ancestral hunter-gatherer human populations (Li et al., 2011; Campbell et al., 2014).

In addition other signals of selective sweeps that may have a significant effect on phenotype variation were found in relation to climate conditions. For instance, the global distribution of a large amount of SNPs observed in contemporary human populations has

been shown to reflect strong latitudinal clines, suggesting the action of sharp climatic and/or geographical-restricted selective pressures (fig. 1.2.2).

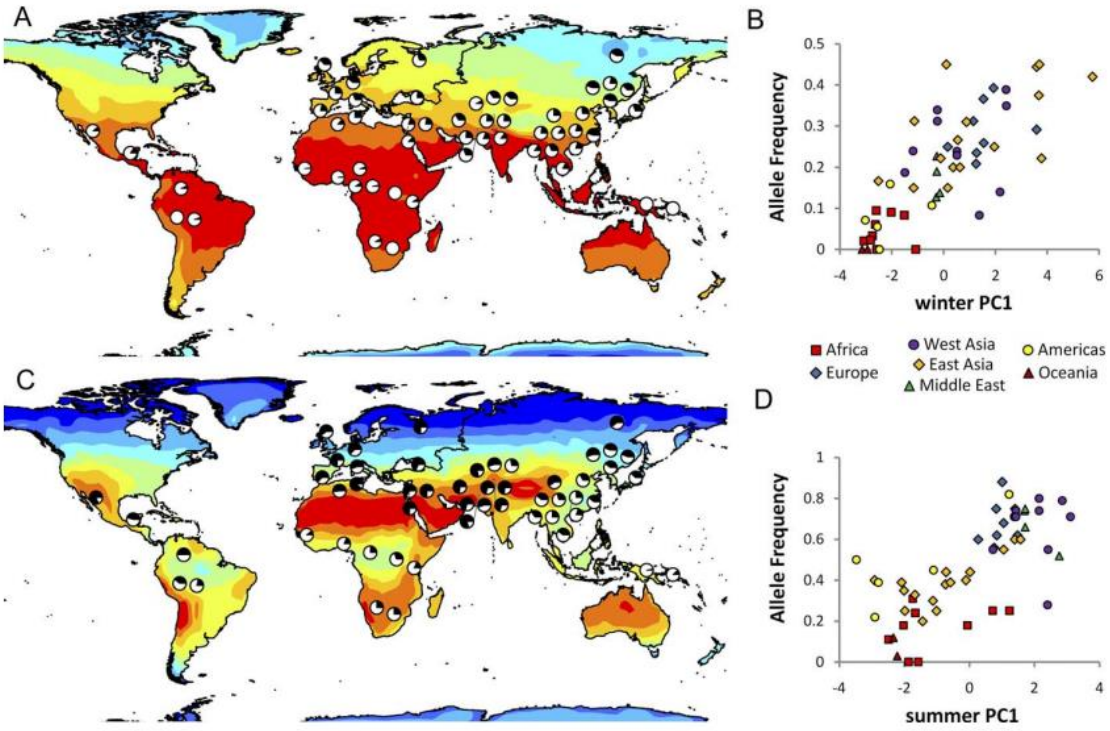


Figure 1.2.2: A) Representation of rs12946049 (*RAPTOR*) allele frequency mapped onto winter maximum temperature B) plot of the SNP against winter PC1. C) Representation rs662 (*PON1*) allele frequency mapped onto summer maximum temperature D) and plotted against summer PC1

Several genes involved in metabolic functions seems to have been influenced by environmental conditions such as *RAPTOR*, *TCF7L2* or *PPARG* (Hancock et al., 2008). Nevertheless it should be noted that climate is not only linked to latitude, but it is composed by numerous elements like temperature, humidity, solar radiation and precipitation that should be considered in the analysis of human variation. Thus when these further climatic features are taken into account, several SNPs came out to be correlated with some of the variables cited before. Particularly there are interspersed

clusters of genes which are linked to a same climatic condition. This suggests the existence of several genetic regions where are concentrated genes that fulfil a similar function (Raj et al., 2013) (fig. 1.2.3).

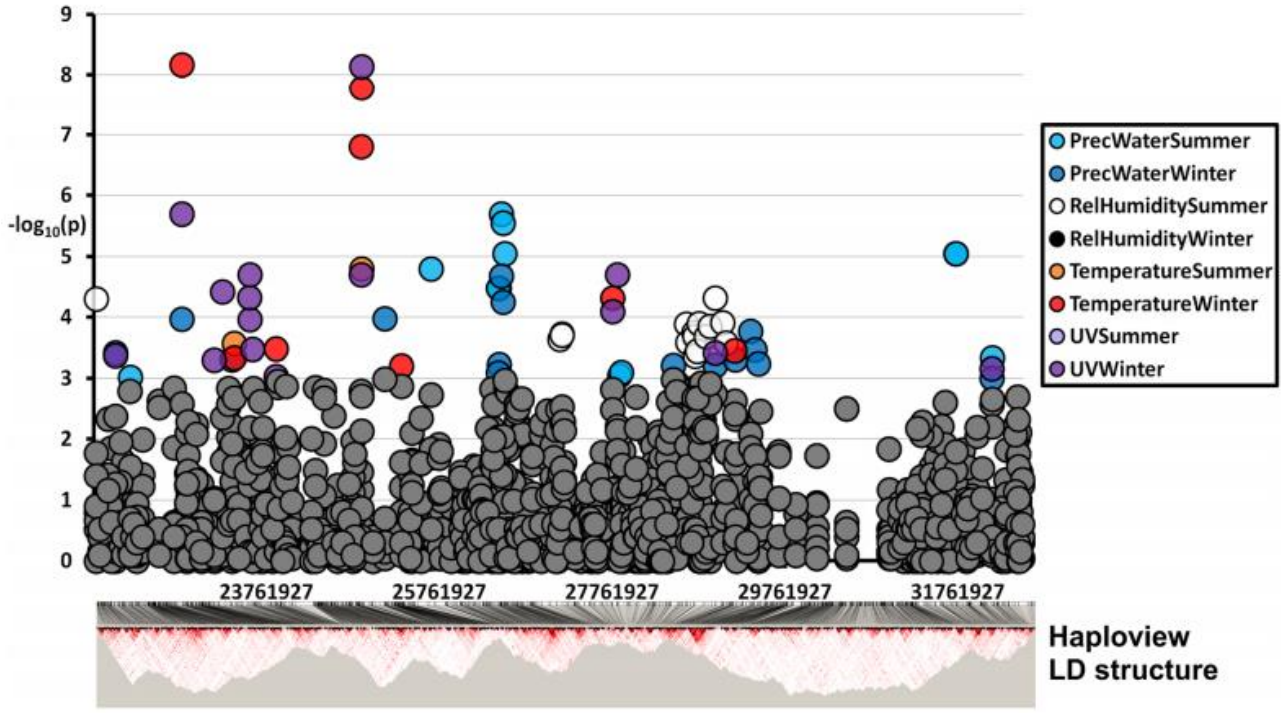


Figure 1.2.3: Manhattan plot of SNPs p-value in association with various climatic elements (Raj et al., 2013).

All these evidences, for which have been provided just a brief summary, supported the idea that when human subgroups migrated to new lands, the particular environmental condition such as temperature, climate changes, food resources and pathogens (Fumagalli & Sironi, 2014) have forced them to adapt to the local framework characteristics. Therefore, these studies highlighted how much local environmental conditions have affected human genetic background, shaping numerous processes of adaptation in the last 100,000 years (Voight et al., 2006).

1.3 Signal of local adaptation detected within European populations

As already illustrated, every population after the out of Africa migration has experienced its own local adaptation process which let to distinguish their genetic background from that of others human groups living in other regions (Roland & Akey, 2005; Lopez Herraez et al., 2009).

Numerous studies have focused on European continent both for several population demographic processes, that have strongly influenced human history, than for its bio-climatic particularities that rose from Boreal to Mediterranean environment characterized by extreme divergences in temperature and resources (fig.1.3.1) (Rivas-Martinez et al., 2004).

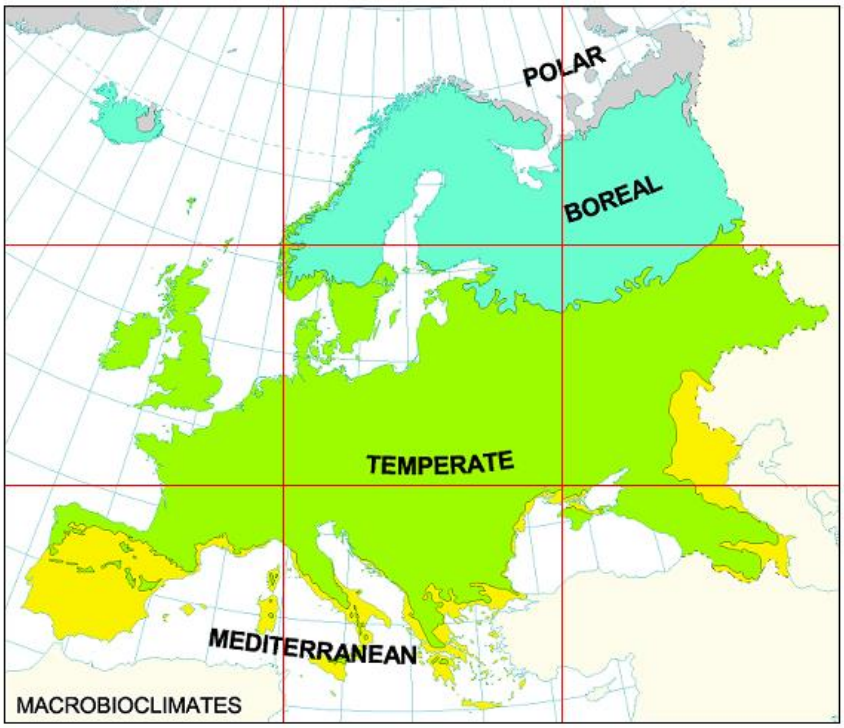


Figure 1.3.1: Bioclimatic map of European continent (www.globalbioclimatics.org)

All these peculiarities of European landscape have deeply contributed to define the modern pattern of variability found among the human groups that inhabit this continent.

During the last decade, molecular anthropologists deeply focused on the study of first European settlements to turn on the light on past events which influenced modern European variability. Despite several information obtained even from ancient remains, the processes that are at the origin of the modern genetic European landscape still remain contentious. Genome-wide analyses of both ancient and modern European DNA, highlighted the presence of three main ancestral populations: a group of western European hunter-gatherers, which seems to have contributed to all European ancestry, but not to Near Easterners; farmers Neolithic populations; and finally a steppe pastoralists (Yamnaya), that contribute to both European and Near Easterners (Lazaridis et al., 2014; Allentoft et al., 2015). First evidences showed that an ancient population, stretching from Europe to Central Asia, have contributed both to part of modern Eurasians genetic background that to the first American settlement (Seguin-Orlando et al., 2014). After that, a successive gene flow to Europe was started by Neolithic farmers around ~10,000 years ago, characterized not only by a diffusion of ideas and techniques, but really from a progressive population expansion from Anatolia into European continent via Greece (Mathieson et al., 2015; Hofmanová et al., 2015). The first farmers culture in central Europe appeared between 7500 and 6900 BP as the Linearbandkeramik (LBK) Culture (Whittle, 1996). Anatolian Neolithic farmers showed a higher genetic homogeneity with sample of European farmers from various sites: Germany, Hungary and Spain (Haak et al., 2015) (fig. 1.3.2). Thus at that moment, two deeply different cultures belonging to two genetic distinct populations were living within the same continent (Bramanti et al.,2009; Skoglund et al., 2012). European hunter-gatherers subsistence strategies seem to have been highly variable and depending on local resources, particularly it was based on

wild flora and fauna including fishes and several vegetables, with important amount of iron intake and proteins more than in present day diet (Eaton, 2006; Kuipers et al., 2010; Konner et al., 2010).

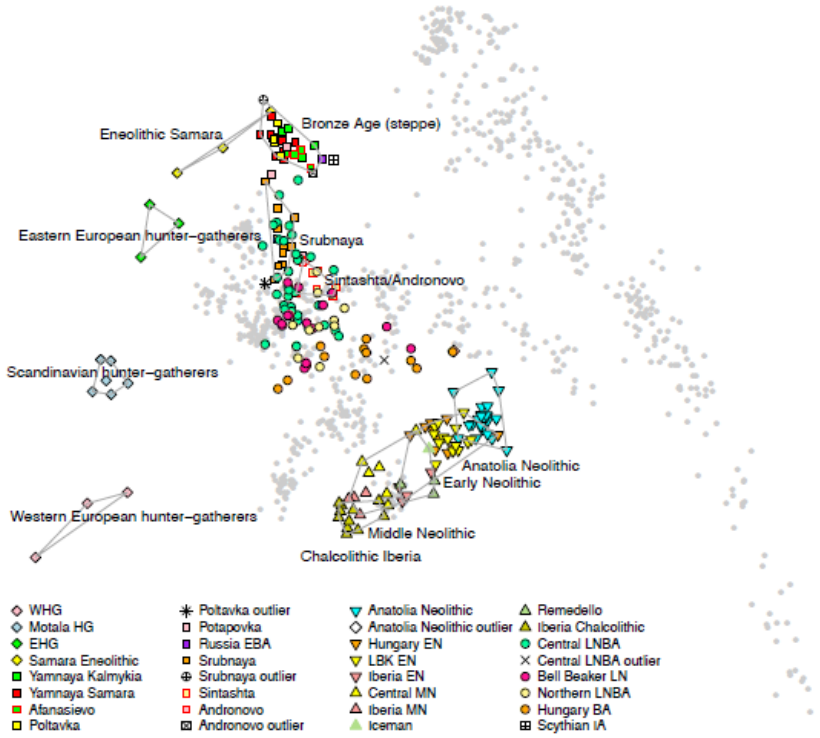


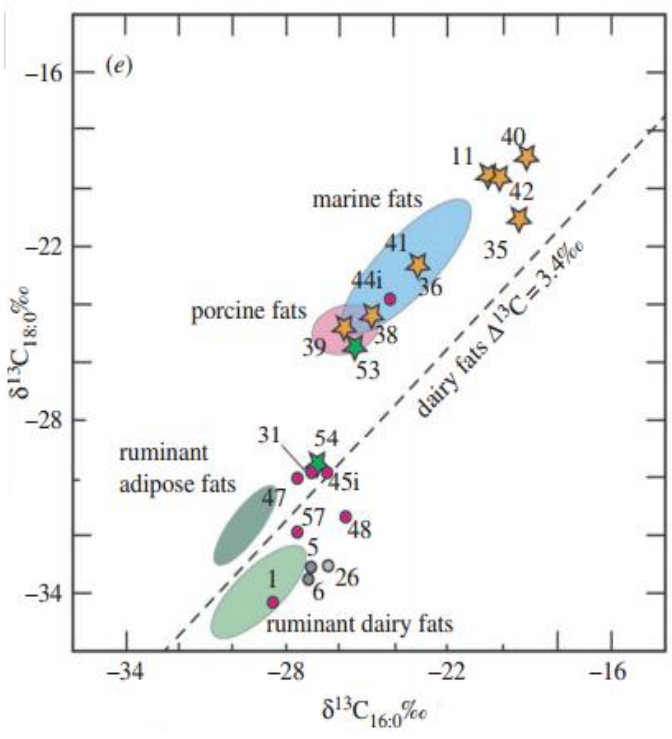
Figure 1.3.2: PCA analysis showing genetic proximity between Anatolian ancient farmers and first Neolithic people from Europe (Mathieson et al.,2015).

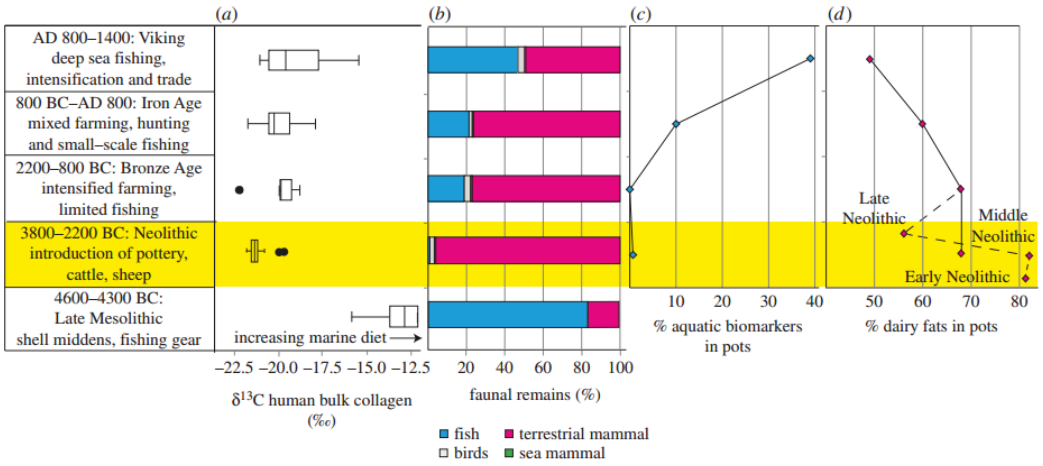
The Neolithic transition starts to take place in Europe with the arrivals of Anatolian farmers, that have profoundly changed the societies that were living there until that moment. These first farmers carried with them techniques of domestication of plants and animals that shifted eating habits to a high carbohydrate and poor iron and protein diet (Cordain et al., 2002). Modern analyses of stable isotopes on archaeobotanical and archaeozoological remains further elucidate this transition process, showing a novel subsistence strategy based on grains and dairying product (i.e. milk), but sometimes with

local variable consumption of wild resources (Fernandes et al., 2015). Despite a general nutritional shift testified by various archaeological records belonging to this historical period, some communities seem to have maintained some aspects of their hunter-gatherers dietary habits. This seems to indicate the existence of distinct routes of Neolithization, or maybe more gradual changes in dietary habits according to local resources (Cramp et al., 2014).

Indeed, northern European archaeological pottery remains, showed that within some hunter-gatherers community the nutritional shift to a Neolithic life-style was a very immediate and drastic event (i.e. data from Northeast Atlantic archipelagos and Finland), thus passing from fishing to dairying in a short period (fig 1.3.3).

A)





B)

Figure 1.3.3: A) Stable isotopes analysis ($d^{13}C_{16:0}$ and $d^{13}C_{18:0}$ values) from Comb Ware (orange), Corded Ware (pink), Kiukainen Ware (green) and Metal Age (grey) residues; It is possible to observe the shift from Comb ware (still hunter-gatherers) to the Corded ware (first farmers). B) Percentage of faunal remains founded on several archaeological sites, even here it is possible to see the highly augmentation of the percentage of terrestrial mammal remains from Late Mesolithic to Neolithic (Cramp et al., 2014).

While on the other hand stable isotopes analysis from Baltic site seems to indicate a continuation of hunter–fisher–forager dietary pattern, as if particular regions had a refuge function for last hunters-gatherer communities (Malmstrom et al., 2009; Craig et al., 2011).

This cultural and genetic change within European continent has enormously influenced modern genetic background where dietary-related signals of natural selection are detectable. One of the strongest signal of selection is the already cited *LCT* gene responsible for lactase persistence in Europe. Recent analyses on aDNA detected the appearance of the allele already in one sample from Bell Beaker culture (2300-2200 BCE) in central Europe (Mathienson et al.,2015). The positive selection of this mutation is highly linked to dairy practices and thus to Neolithization, as demonstrated by the higher presence of this allele among populations with longer dairy story (Itan et al., 2009); while a lower presence is attested where the introduction of the new dietary habit

occurred later, for example a lower frequency of lactase persistence is found in some areas of the eastern Baltic where (as said before) archaeological and genetic studies demonstrated a continuous with hunter-gatherers culture (Cramp et al., 2014). Other signals of positive selection always in relation to the agricultural revolution have been detected for FADS1 and FADS2, as well as for *DHCR7*. These first two genes are implicated in fatty acid metabolism and seem to be related to plasma lipid concentration, while the third one is associated with vitamin D levels which seem to have undergone positive selection in northern Europe, thus suggesting that changes in environment or dietary source of this vitamin may have affected the sweep of particular variants within this gene (Price et al., 2009; Teslovich et al., 2010).

The poor iron intake linked to Neolithic dietary shift, in conjunction with the exposure to wet and cold environments seem to be at the base of positive selection for the C282Y mutation (*HFE* gene) in Europe which is responsible of increased iron absorption and hemochromatosis disease (Sheftel et al., 2012). In fact, iron deficiency is associated to an inhibition of thyroid stimulating hormone (TSH) which is indispensable to regulate metabolism and body temperature (Brigham et al., 1996). Thus, first farmers which reached northern European latitudes have been exposed to higher thermal stress, and their poor iron diet was not able to face such extreme climate conditions. This is the reason why, accordingly to Health et al. (2016), this pathogenic mutation has been selected during time to assure the maximum absorption of the poor iron available in the Neolithic diet. Two other selection signals, that could be related to this period, are potentially linked to celiac disease: *SLC22A4* and *ATXN2*/*SH2B3* genes. Several variants within these genes are associated with increased rate of irritable bowel disease (IBS) and celiac disease (CD) developing risk, thus always in relation with cereals consumption and agricultural practices (Hunt et al., 2008; Mathienson et al.,2015).

1.4 Dietary shifts and its influence on human genetics and microbiota composition

Human activities have significantly affected the biosphere over the last 10,000 years, thus from when agrarian communities started to directly influence the ecosystem for their subsistence (Ruddiman & William, 2013; Lewis et al., 2015). These events occurred during human evolutionary history constitute such a great change under a biosphere point of view that it is called from geologist as "Paleo-Anthropocene" (Doughty et al., 2013) (fig. 1.4.1).

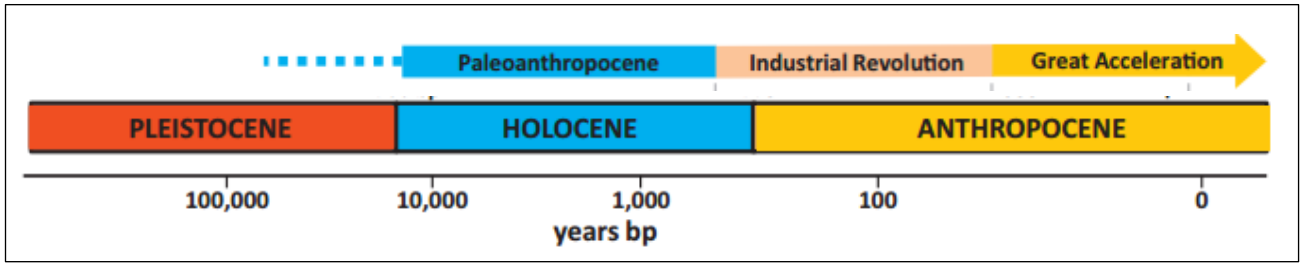


Figure 1.4.1: Distinction of three geological era: The Anthropocene period starts with the industrial revolution but numerous researchers anticipate the human influence on the biosphere from the Neolithic transition, calling this period as "Paleo-Anthropocene" (Gillings & Paulsen, 2014).

These changes have deeply influenced human genome, as highlighted from data showed in the previously paragraph, but what about its influence on human-microbiome? Surely, across this same period, humans started to influence even its own internal biosphere, exercising an unconscious selective pressure on its symbiotic microorganisms (Gillings & Paulsen, 2014). These microbial communities which co-exist within human body, are generally indicated by the term "microbiota", and the sum of the genes carried and expressed by these microorganisms is named "microbiome" (Cho & Blaser, 2012). Several studies have clearly shown how perturbations in the microbiota composition or functions may significantly affect human health, thus highlighting its central role on

human health-state, but how human gut microbiome has evolved during time still remain unclear. Particularly, despite a high variability between individuals or populations, a core functional microbiome remain relatively constant (Qin et al., 2010;Yatsuneko et al., 2012). In fact, when comparing human microbiota composition to that of African apes, there are several species which seem to characterize the *Homo* species from *Pan* and *Gorilla*. Moreover *Homo* species seems to have undergone a rapid change, which could be translated in such a "specialization" of gut microbes: in particular humans microbiota showed a decreasing level of biodiversity and a number of changes correlated to host nutrition processes (Moeller et al., 2014)(fig. 1.4.2).

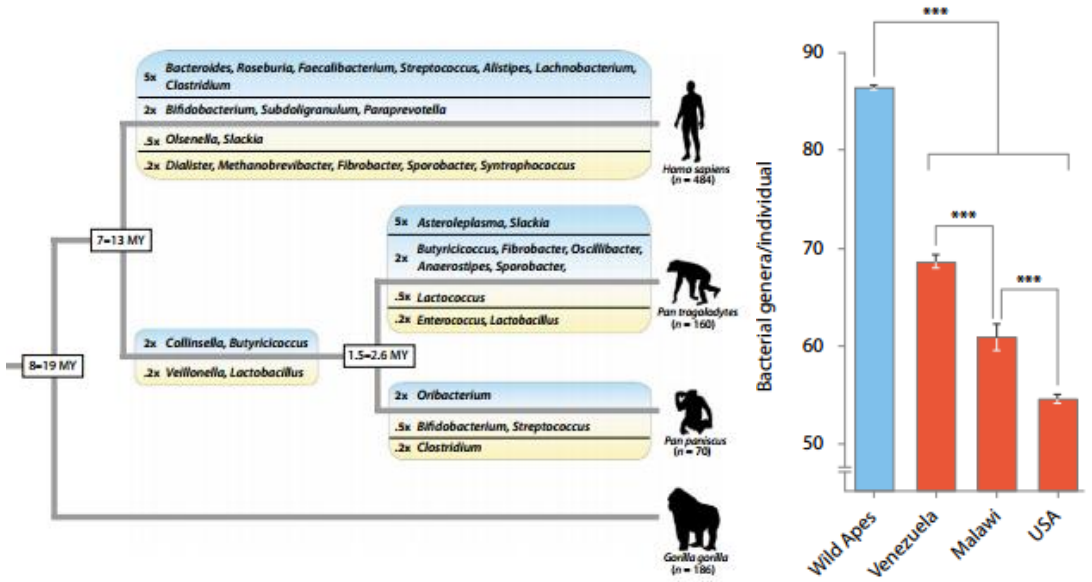


Figure 1.4.2: On the left are represented the main changes in gut microbiota composition between Homo and African apes; on the right the microbial observed biodiversity between apes and three human populations (Moeller et al., 2014).

As showed from population studies, and as visible from figure 1.4.2, it exists a high level of variability even between different human population groups. The major changes are often linked to different diet and different resources exploited, which could positive select one bacterial species while excluding others less functional to metabolism processes. For

example, various members of *Bacteroides* phylum have been positively associated with diets composed of animal fat and proteins, indeed they are mostly present within humans microbiota than within apes and, among different human groups, within people from western-societies more than among human rural populations (Wu et al., 2011). Recent studies have turned the light on the dynamics affecting host-microbiota co-evolution and environmental adaptation, starting to compare microbiota composition among different populations. For example, the study of one of the last hunter-gatherers community in the world (the Hadza) have shown the high percentage of microbial biodiversity than western societies and an enrichment in fibrolytic bacteria (as *Prevotella* and *Treponema*) which are signal of local adaptation to their plant-based diet (Schnorr et al., 2014). A total absence of *Bifidobacterium* was found within this hunter-gatherer societies while and higher enrichment of opportunists bacteria phyla like Proteobacteria. This is exactly the opposite condition to western societies where normally *Bifidobacterium* is considered to have a healthy effect and avoid auto-immune disease, while Proteobacteria are normally associated with pro-inflammation status (De Filippo et al., 2010; Ruiz Martinez et al., 2015, Angus et al., 2016). These evidences highlighted a different tolerogenic layout of Hadza immune system than modern societies, but it is interesting to note that Hadza people seem to be less subjected to metabolic and nutritional diseases (Blurton Jones et al., 1992; Bisgaard et al., 2011). Comparisons between hunter-gatherer societies, rural modern societies and western ones may help in the reconstruction of human transition from Palaeolithic life-style to Neolithic and industrial one. Generally speaking, every passage is characterized by a loss of biodiversity with the higher level among hunter-gatherers and the lowest among western populations (fig. 1.4.3).

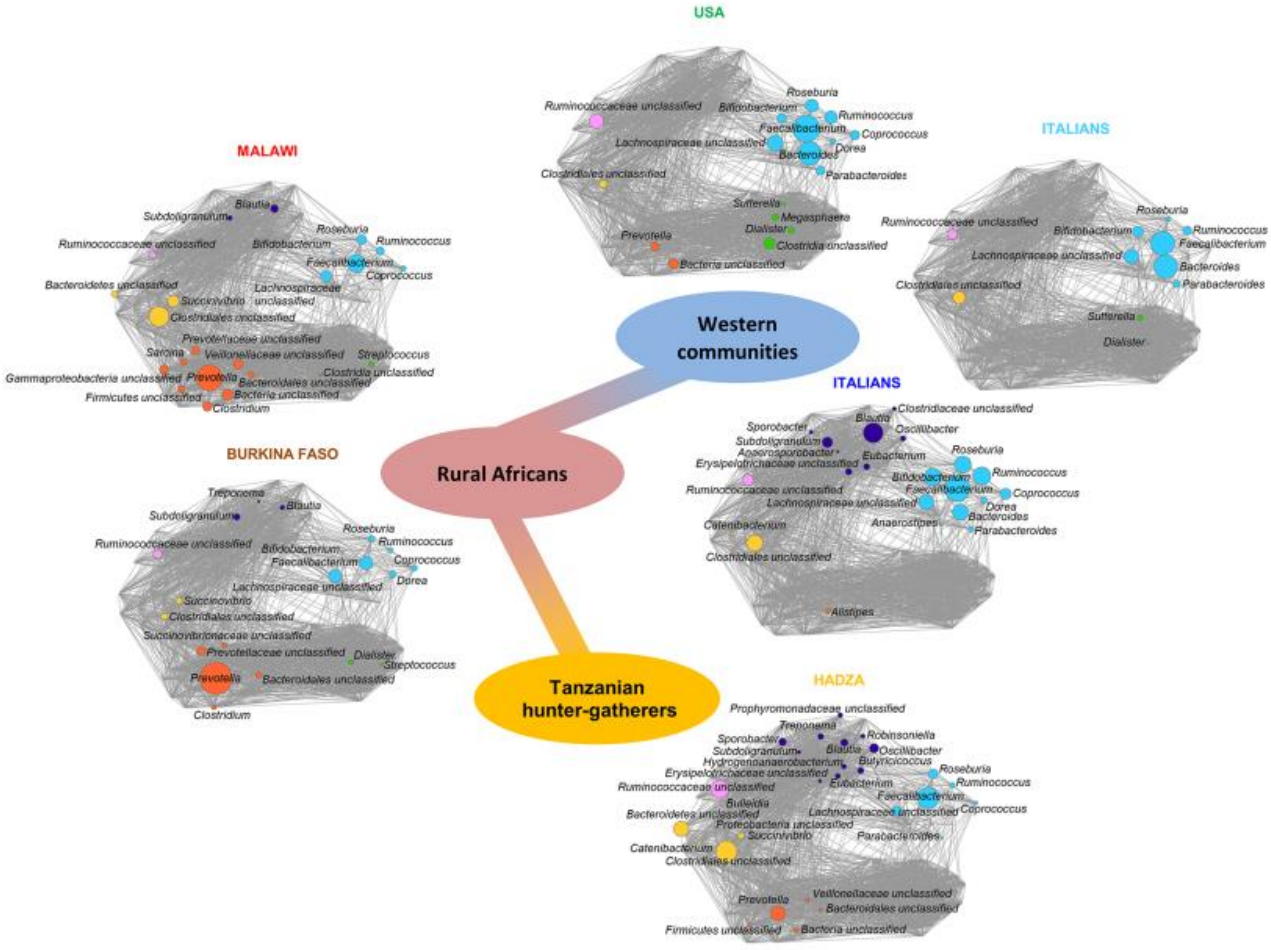


Figure 1.4.3: Wiggum plot which indicates the patterns of microbial diversity in several societies. Each node represent a bacterial genus which is represented by a dimension proportional to their relative abundance within the microbiota. The central path indicates main dietary transition happened during human history (Schnorr et al., 2014).

More in details after Neolithic transition it starts to emerge specific co-abundance groups of microbial which are highly functionally correlated such as the increase of *Ruminococcaceae* and *Clostridiales* within rural societies and of *Faecalibacterium* among western populations (Quercia et al., 2014). To date only in few cases, an evolutionary approach has been applied to microbiota analysis, thus a more precise knowledge how dietary shifts have influenced human evolution is missing, further researches are needed to elucidate the host-microbiome evolutionary dynamics.

Chapter 2

AIM OF THE THESIS

The evolution of behaviourally and anatomically modern human populations has been characterized by several dramatic changes in environment conditions and life-style. Over the last 100,000 years, our species has spread from Africa to the rest of the world, carrying on a long process of colonization that involved different ecological niches.

At the end of the last ice age ~14,000 years ago, a process of global warming has raised the temperature to the current levels. Further dramatic climate changes occurred since then, as well as several shifts in dietary habits, such as the well-known Neolithic transition around ~10,000 years ago. All these new contingences have forced the human species to adapt to a wide range of habitats, resulting in a large amount of selective pressures that have favoured specific genotypes probably more suitable to live in a new environment, and that have shaped patterns of genetic variation in modern humans.

Over the last decades, our understanding of human genetic diversity has extremely increased, also thanks to ancient DNA researches. This was made possible by the introduction of high-throughput sequencing techniques that, together with the development of sophisticated analytical tools, have allowed researchers to obtain and process a large amount of DNA data.

The present research aims at investigating the influence of such ecological and nutritional changes on the human genome, by means of three main projects:

Project 1: Detecting signals of natural selection in European populations on genes involved in nutritional and thermoregulation processes

European populations have experienced several migration and adaptation processes during their recent evolutionary history, as previously described in the first chapter. Most parts of researches on population genetics have to date strongly focalized on the detection of climate influence on SNP frequencies in the most extreme environmental conditions and by comparison between populations belonging to different continents. In this study, we explored the genetic scenario existing among populations that have adapted to different environmental conditions, but in the range of the same continent. Thus, by taking advantage from the *1000Genomes Project* database, genetic information for several genes involved in nutritional and thermoregulation processes have been analysed for the available European populations. Population structure analyses and neutrality tests have been thus performed to explore patterns of genetic variability at the cited genes, and to detect possible signals of natural selection that could be related to local adaptive events.

Project 2: Analysis of Italian genetic variability at three genes associated to metabolic functions

Previous studies carried on by our research group on several Italian samples and by means of high-throughput genome-wide approaches, highlighted the presence of high levels of genetic variability within the Italian peninsula. Particularly, human groups settled in Northern and Southern Italy seem to diverge for the frequencies of a number of

SNPs located on three different genes: *R3HDM1*, *RAB3GAP1* and *TMEM163*. To further investigate patterns of genetic variability at these genes associated to metabolic phenotypes along Italy, a targeted analysis has been conducted on 380 healthy Italians from several Italian provinces. At first, genetic data for European populations available on databases was used to made up a first screen of the genetic variability observable in Europe. Then, according to computed allele frequencies and F_{st} values, a number of SNPs were selected and genotyped with the *Sequenom* platform. In order to test the actual existence of a different pattern of variability between different macro-areas of the peninsula, various statistical tests were performed. Further descriptive analyses have been also applied to compare the Italian samples with other European populations and with worldwide genetic data to infer the recent evolutionary history of the examined genomic regions.

Project 3: Celiac disease and its influence on gut microbiota composition: an evolutionary approach to a probiotic treatment case study

Celiac disease is a common small intestinal inflammatory condition linked to gluten intake. Although the genetic and environmental factors that contribute to the development of this pathologic conditions are not fully elucidated, a high number of studies is highlighting a key role of gut microbiota composition in CD development. This metabolic autoimmune disease is probably originated due to the dietary changes that took place during Holocene transition to agriculture, a period in which human populations have been plausibly subjected to strong selective pressures, especially impacting on genes involved

in metabolic processes, as reported in chapter 1. Furthermore, under an evolutionary perspective, celiac disease represents a paradox due, differently from lactase persistence, to its higher incidence in populations with a long history of wheat consumption and agricultural habits (Sams & Hawks, 2013).

By taking advantage from data generated for a probiotic treatment case study on individual affected by celiac disease, it was possible to evaluate the influence of this pathologic condition on gut microbiota composition and thus to get evolutionary insights into the complex interplay between gut microbiota and CD susceptibility. At first, two hypervariable regions of the 16S gene (V3 & V4) were sequenced on a cohort of patients affected by celiac disease and on healthy individuals. More in details, celiac individuals were subdivided into two different groups and, during a time of three months, one group received a probiotic treatment while the other received a placebo administration. Two faecal samples have been collected at T0 (begin of the study) and T1 (after three months of probiotic or placebo administration) to evaluate the effect of probiotic on gut microbiota composition and then to compare it with that of the healthy individuals. Several bioinformatics tools were used to analyse gut microbiota changes linked to the treatment and in comparison to the healthy group. Particularly, a general evaluation of biodiversity and species relative abundance was achieved to describe how the microbial communities changed between celiac and healthy individuals and to consider how these changes could reflect those occurred in the recent evolutionary history of our species.

Chapter 3

Materials and Methods

Project 1: Detecting signals of natural selection in European populations on genes involved in nutritional and thermoregulation processes

3.1.1 Made up of the genetic dataset for genes involved in metabolic and thermogenic processes

To perform population genetic analyses, genotype data for 31 genes (showed in table 3.1.1.1) pertaining to several metabolic and thermogenic pathways were retrieved from the 1,000Genomes Project database (1000 Genomes Project Consortium et al. 2010, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp).

More in detail, the obtained dataset was composed of genotypes of 280 individuals belonging to three European populations characterized by limited genetic admixture and representative of the European genetic variability observable at different latitudes (Tuscans from Italy, **TSI**; British from England and Scotland, **GBR**; Finnish, **FIN**) (fig. 3.1.1.1). In particular, the other two European groups available in the 1,000 Genomes Project database (i.e. Utah residents with Northern or Western European ancestry, **CEU** and Iberian populations in Spain, **IBS**) were not considered due to remarkable levels of admixture and low sample size, respectively.

Table 3.1.1.1: The set of genes for which patterns of variability have been explored in European populations, with a brief explication of their function and the relative bibliography.

Gene	Chr	Role
UCP1	4	uncoupling protein 1
UCP2	11	uncoupling protein 2
UCP3	11	uncoupling protein 3
ADRB3	8	adrenoceptor beta 3
DIO2	14	deiodinase, iodothyronine, type II
PRDM16	1	PR domain containing 16
ADRA1A	8	adrenoceptor alpha 1A
FTO	16	fat mass and obesity associated
PLIN5	19	perilipin 5
INSIG2	2	insulin induced gene 2
APOA5	11	apolipoprotein A-V
MC4R	18	melanocortin 4 receptor
GNB3	12	guanine nucleotide binding protein (G protein), beta polypeptide 3

Gene	Chr	Role
VDR	12	vitamin D
MTRR	5	5-methyltetrahydrofolate-homocysteine methyltransferase reductase
PLRP2	10	pancreatic lipase-related protein 2
NAT2	8	N-acetyltransferase 2
PPARG	3	peroxisome proliferator-activated receptor gamma
TCF7L2	10	transcription factor 7-like 2 (T-cell specific, HMG-box)
MMP20	11	matrix metallopeptidase 20
ENAM	4	enamelin

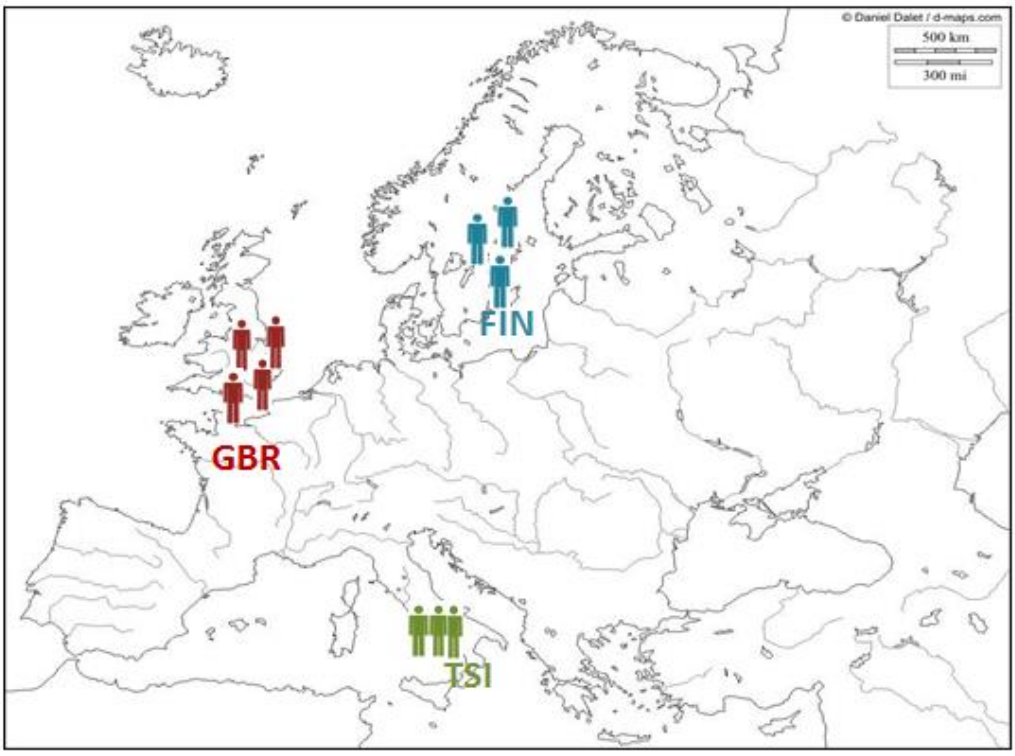


Figure 3.1.1.1: Sample distribution among the European continent

3.1.2 Population genetics analyses

To explore patterns of genetic variation among the selected populations, the PLINK package v.1.07 (Purcell et al. 2007) was used to prune SNPs in linkage disequilibrium (LD) and to prevent potential bias on the applied multivariate analyses. By proceeding with a 50 SNPs-sliding windows approach and by shifting windows of 10 SNPs every time, LD values have been calculated between each pair of SNPs and variants with a pairwise genotypic correlation (r^2) higher than 0.1 were removed.

The obtained pruned dataset was then used to perform discriminant analysis of principal components (DAPC) (Jombrat et al. 2010), which is particularly informative to explore

genetic relationships among pre-defined groups of individuals, using the R *adegenet* package (www.r-project.org) and in order to explore patterns of population structure.

3.1.3 Descriptive and site frequency spectrum-based neutrality statistics

The nucleotide diversity, estimated as the average number of pairwise difference (π), the number of polymorphic site (S), as well as Tajima's D (D), Fu and Li's D and F (D', F) statistics were calculated using the DnaSP package v.5.10 (Librando and Rozas 2009) on the whole *PRDM16* gene that turned out to be the best candidate locus among those investigated after population structure analyses. These statistics were computed for each population by applying a sliding windows approach (i.e. each window consisted of 10kb and was progressively shifted of 10kb). To test the significance of the obtained results 10,000 coalescent simulations were performed considering local recombination and mutation rates and under the assumption of a neutral model of evolution. Than, nominal *p*-values were corrected with Bonferroni procedure using the R package *multtest* in order to account for multiple testing procedure and by considering a significance threshold of $\alpha = 0.01$.

3.1.4 Haplotype and iHS Analyses

Focusing on the gene windows showing significant *p*-values for the above-mentioned neutrality tests, pairwise LD for each SNP within each window was calculated with the PLINK package v.1.07 (Purcell et al., 2007). SNPs in high LD ($r^2 > 0.90$) were then used

to infer haplotypes from unphased genotypes using the PHASE software v.2.1 (Stephens and Scheet, 2005), which implements a Bayesian algorithm for haplotype reconstruction starting from population genotype data.

The analysis output was finally used to further explore the evolutionary relationships among the inferred haplotypes by drawing a median joining network with the Network package v.4.6.1.2 (http://www.fluxus-engigeering.com).

To further investigate genetic signatures related to the action of potential selective pressures on the examined populations, statistical methods based on the evaluation of extended haplotype homozygosity (EHH) were applied. The REHH R package (Gautier et al., 2012) was used to estimate the relative extended haplotype homozygosity compared with the EHH of all other haplotypes on the gene and its progressive decay, defined as integrated EHH (iHH). The lengths of the haplotypes extending around core SNPs are deducible from the decay of EHH values until they drop below 0.2. Therefore, it is possible to calculate the haplotypes maximum dimension taking into account the distance between the core SNP and the first point to the left and to the right with an EHH > 0.2. On the basis of the decay level, the integrated haplotype score (iHS) was then calculated as the log-ratio of iHH computed at the derived and ancestral alleles (Voight et al. 2006).

The iHS statistic was finally obtained for each SNP using an homozygosity threshold of EHH > 0.05 and allele frequency bins of 0.025 to standardize iHS scores. SNPs showing |iHS| values > 2 were considered as possible candidate loci to have undergone recent positive selection (Voight et al. 2006). Bonferroni correction was then applied to account for the adopted multiple testing procedures. Several graphical tools of the REHH package were also exploited to better visualize linkage breakdown. Finally, a rough estimate of the age of selective sweeps detected by iHS analysis was obtained as described in Voight et

al. (2006) for each SNP showing significant signals of positive selection and by taking advantage from the Li and Stephens' algorithm to evaluate recombination rates (Li and Stephens, 2003).

Project 2: Analysis of Italian genetic variability at three genes associated to metabolic functions.

3.2.1 Population Samples

Replication of Sazzini et al (submitted, 2016) was carried out on 380 healthy subjects recruited from 20 provinces belonging to three different geographic areas distributed along the Peninsula. In details, 135 samples were collected from Northern Italy, 116 from Central Italy and 129 from Southern Italy (table 3.2.1.1 and fig. 3.2.1.1).

Table 3.2.1.1: Information about samples' distribution into provinces.

Macro-Area	Region	Province	ID	N°	Tot
North	Lombardia	Varese	VA	26	
	Liguria	Savona	SV	10	
		Imperia	IM	27	
		Vicenza	VI	10	
	Veneto	Feltre	BL	26	
	Friuli Venezia Giulia	Trieste	TS	36	135
Center	Toscana	Grosseto	GR	10	
	Marche	Macerata	MC	26	
		Ascoli Piceno	AP	20	
		Pesaro	PU	20	
	Umbria	Foligno	PG	10	
		Terni	TR	30	116
South	Molise	Campobasso	CB	25	
	Campania	Benevento	BN	10	
	Basilicata	Policoro	MT	15	
	Puglia	Lecce	LE	10	
	Calabria	Catanzaro	CZ	10	
		Castrovillari	CS	10	
		Reggio Calabria	RC	26	
	Sicilia	Enna	EN	23	129
					380

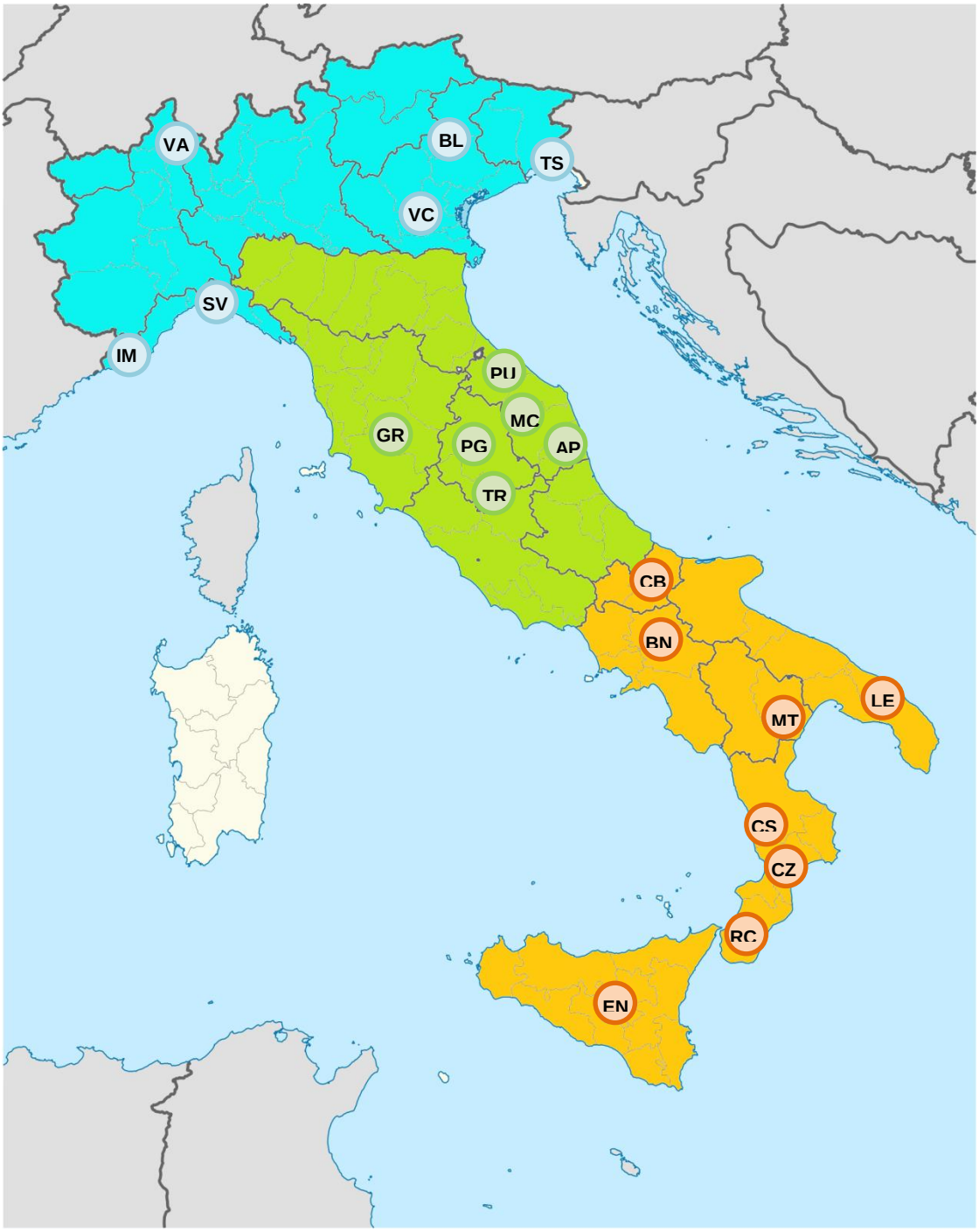


Fig. 3.2.1.1: Samples' distribution into provinces and their location along the Italian peninsula.

All individuals were selected according to two different criteria: the "grandparents criterion" and the "founder surnames' criterion" (Boattini et al., 2013). Particularly, only

individuals whose all grandparents were born in the same sampled area were included in the project, also taking into account the presence of founder surnames.

Blood samples for DNA extraction were collected thanks to collaboration with the transfusion centers of the involved provinces. Furthermore, all individuals analyzed in this project were asked to sign an informed-consent form, approved by the Ethic Committee of the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna.

Finally, data have been compared with those available for other European populations from the *1000 Genomes Project* database (1000 Genomes Project Consortium et al. 2010, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/). More in detail, the obtained dataset was composed of genotypes of 404 individuals belonging to four European populations characterized by limited genetic admixture and representative of the European genetic variability observable at different latitudes (Tuscans from Italy, TSI; British from England and Scotland, GBR; Finnish, FIN; Iberians, IBS).

3.2.2 SNPs Selection

Among the 6,648 SNPs retrieved for the available European populations sequenced by the *1000 Genomes Project* and relative to the *TMEM163*, *RAB3GAP1* and *R3HDM1* genes (+/- 10,000 bp) (fig. 3.2.2.1), a total of 23 SNPs have been selected according to their high heterozigosity values and FST computed for pairwise population comparisons. To prevent potential bias introduced in the subsequent analyses due to high LD levels, the PLINK package v.1.07 (Purcell et al. 2007) was used to prune SNPs in LD, proceeding with a 50 SNPs-sliding windows approach and by shifting windows of 10 SNPs every

time. Thanks to this approach, LD values have been calculated for each possible SNPs pair and variants with a pairwise genotypic correlation (r^2) higher than 0.1 were removed.

In addition to these SNPs, other two candidate loci (rs6723108 and rs1446585) were selected among the most informative ones pointed out by a genome-wide study performed by our research group (Sazzini et al. submitted, 2016), again according to their considerable differentiation among European populations (especially between Northern and Southern European ones).

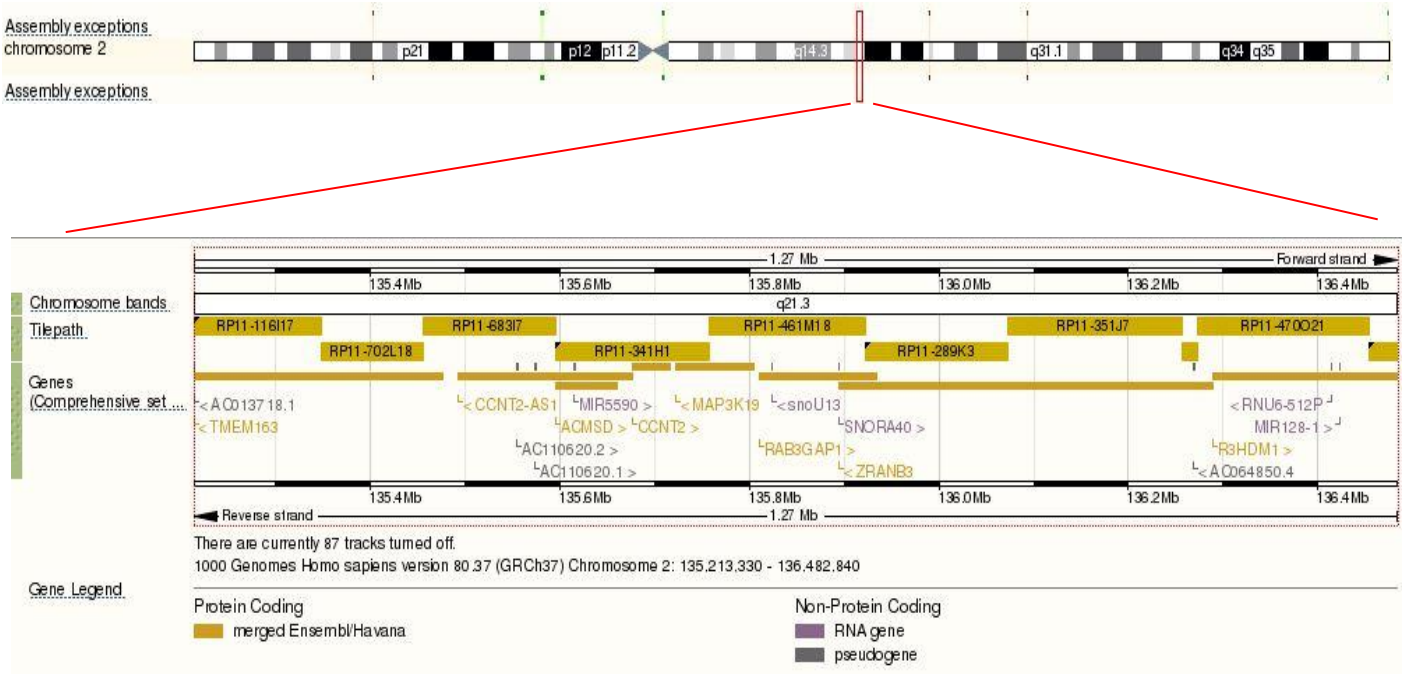


Fig. 3.2.2.1: Examined genomic region including the *TMEM163*, *RAB3GAP1* and *R3HDM1* genes.

3.2.3 DNA Quantification

DNA quantification was obtained by means of the Qubit™ 3.0 Fluorometer (Thermo Scientific) that performs a fluorometric quantification based on the detection of target-specific fluorescence emitted from an ultra-sensitive fluorescent dye which interposes in double-stranded DNA (dsDNA). The luminous signal is proportional to the length of the DNA sequence, thus the measure of quantification is very accurate and it minimizes the possibility of contamination. The Qubit dsDNA BR (Broad-Range Assay Kit - Thermo Fisher) was used to quantify the chosen samples. This kit includes concentrated assay reagents and dilution buffer and pre-diluted DNA standards, which are useful for calibration of the instrument. It allows a precise quantification for a core ranging from 0.01 μg/mL to 5 μg/mL of DNA, which is normally the suitable range for DNA extracted from blood samples.

3.2.4 Design of Multiplex PCR and Sequenom Assay

The Multiplex polymerase chain reaction (or Multiplex PCR) is a technique that uses the ability of the normal polymerase chain reaction (PCR) of amplification to amplify several different DNA sequences simultaneously. This process lets to the amplification of different genomic DNA regions thanks to the use of multiple primers and a temperature-mediated DNA polymerase in a thermal cycler (Hayden et al. 2008).

To design PCR and extension primers for the multiplex-PCR, the *Sequenom's MassARRAY Designer software* was used (Sequenom, Inc., San Diego, CA), paying close

attention to avoid primer combinations and non-template extension products. The best multiplex design obtained unified the selected SNPs in one multiple reactions.

3.2.5 Genotyping with the Sequenom MassArray iPlex Platform

Genotyping was performed on each sample using the iPLEX Gold technology (Jurinke et al., 2002) and the MassARRAY DNA analysis with Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (Sequenom, Inc., San Diego, CA), thanks to collaboration with the Centre for Applied Biomedical Research (CRBA) of the Bologna S. Orsola University Hospital.

The experimental protocol used in this project is divided in the following steps (Gabriel et al., 2009):

1) **Sample DNA preparation:**

   After Qubit quantification, the total DNA was diluted to 10 ng/µl in TE buffer concentrated 0.25X. Then, the diluted samples were divided into aliquots of 2µl/well into a 384-well PCR reaction plate for a deep-well PCR source plate (Marsh Biomedical).

2) **Amplification of target loci**

   A single multiplex-PCR reaction was performed to amplify all the target regions of genomic DNA selected as reported on paragraph 3.2.2. The aim of this first

amplification step is to amplify many individual loci with minimal non-specific PCR products.

3) PCR purification

The amplicons obtained by PCR amplification were then treated with Shrimp Alkaline Phosphatase (SAP) to remove remaining non-incorporated dNTPs from amplification products.

4) Primer extension

Cleaned amplicons were used for the primer extension reaction. In this step, the primer anneals immediately upstream of the chosen polymorphic site. Apart from the primer, the reaction mix is composed of the amplified DNA and of mass-modified dideoxynucleotide terminators, thus this reaction will generate oligonucleotides with an allele-specific molecular mass.

5) Spotting primer extension products on SpectroCHIPs

Between 15 and 20 ul were arrayed onto existing matrix spots on the silica chip to incorporate oligonucleotides with the appropriate matrix for MALDI-TOF (3-hydroxypicolinic acid). The Spectrochip was composed of 384-well microtiter plates.

6) Mass Spectrometry and genotyping

Finally, sample molecules were vaporized, ionized and detected on the basis of their mass-to-charge (m/z) ratio. The square root of m/z ratio influences the flight time of

ions that are so detected by MALDI mass spectrometry and the *Sequenom* software (*Spectro*TYPER) automatically translates the mass of the observed primers into a genotype for each reaction (fig. 3.2.5.1).

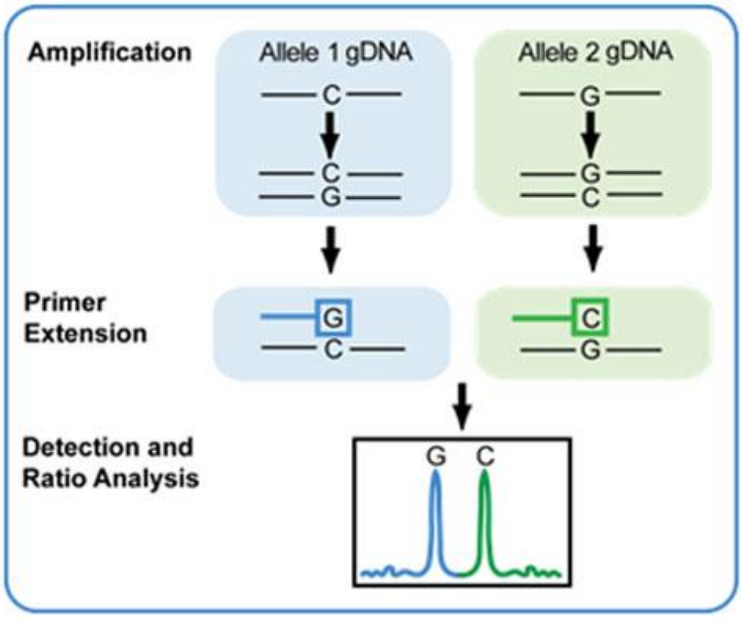


Fig. 3.2.5.1: Schematic representation of the *Sequenom* genotyping workflow.

3.2.6 Bioinformatic Analysis

3.2.6.1 Single locus frequency estimation and population genetic analyses

To estimate allele frequencies for both the genotyped Italian samples and for the European dataset from the *1000Genomes,* the PLINK package v.1.07 (Purcell et al. 2007) was used. Furthermore, this same software enables to prepare input files to be submitted to Fisher's exact test, for exploring differentiation among populations for each locus, and

to principal components analysis (PCA), which were performed using the R *stats* and *adegenet* packages, respectively.

PCA analysis was used to identify the existing patterns of population structure among the examined populations, by highlighting their similarities and differences. Since it could be hard to find a pattern of variability within genetic data of high dimension, PCA represents a powerful tool for analyzing and summarizing huge genetic and genomic datasets. Another main advantage of PCA is that once you have identified patterns of population structure within data, it can compress and reduce data variance without much loss of information. In this project, PCA was used in order to explore patterns of population structure both within the Italian sample and in comparison with the European dataset. In order to get a more clear visualization of SNPs frequencies distribution in Italy, maps were generated by applying Kriging interpolation (Relethford 2008).

3.2.6.2 Fst index calculation

Fst index was computed for all possible pairwise population comparisons for every single locus and using the Weir and Cockerham formula implemented in the R *pegas* package. Performed pairwise comparisons between population clusters enabled to explore levels of genetic differentiation and led to the identification of SNPs that drive patterns of population structure observed by PCA. Moreover, these results have been compared with those obtained by allele frequencies comparisons, as previously described.

3.2.6.3 Haplotype reconstruction

Haplotypes can be inferred from unphased genotypes through several statistical methods. Basically, the two most used methods are: a maximum likelihood approach implemented via the expectation-maximization (EM) algorithm (Excoffier and Slatkin 1995), and a parsimony method created by Clark (1990). Nowadays, the most valuable method for haplotype reconstruction is actually that developed by Stephens, which is a Bayesian statistical approach that allows to use a priori expectations to inform haplotype reconstruction as, for example, population information. In the software PHASE is implemented an improved version of the Stephens' algorithm aimed at reconstructing haplotypes from population genotype data and at estimating their frequencies among the studied groups (Stephens et al. 2001). Moreover, the *Haploview* software (Barrett JC. et al., 2005) was used for direct visualization of haplotype blocks and to select SNPs to be submitted to phasing according to calculation of LD r2 values.

3.2.6.4 Network analysis

To represent inter-population relationships on the basis of the reconstructed haplotypes, the *Network 4.6.1.0* software (www.fluxus-engineering.com) was used to reconstruct phylogenetic networks and trees. The potential of network representation is to infer the evolutionary relationships among species or populations based upon similarities in their genetic characteristics. There are several methods to construct a network and in this work the minimum spanning approach was used because it connects all given types without creating any cycles or without inferring additional nodes (Bandelt et al., 1999).

Project 3: Celiac disease and its influence on gut microbiota composition: an evolutionary approach to a probiotic treatment case study.

3.3.1 Conceived Experimental Design

The performed study was a double-blinded, placebo-controlled intervention study including 49 subjects affected by celiac disease (CD) and 18 healthy individuals as controls. All the samples were recruited at the Department of Pediatrics, University Clinical Center of Maribor (Slovenia) from October 2013 to June 2014. A written informed consent was given to patients or to parents after explanation of the study. The global protocol was approved by the Republic of Slovenia National Medical Ethics Committee. All the individuals affected by CD were randomly allocated into two groups using a computer model (http://www.randomization.com). The first group of CD patients daily received the formulation with the *Bifidobacterium breve* strain for three months, while the second group received a placebo for the same period of time. The healthy control individuals constitute the third group. Fecal material has been collected before the beginning of treatment (T0) and at the third month of treatment (T1).

Several inclusion and exclusion criteria have been applied on these samples. In particular, individuals with CD aged between 1 and 19 years were invited to participate in this study. All patients have been analyzed for their serologic markers for CD and had positive small bowel biopsy. All of them were following a gluten-free diet (GFD) from 1 to 15 years. On the other hand, individuals with acute or other chronic diseases were excluded, as well as those who used medication or follow an antibiotic treatment. All individuals of the third group, thus the healthy one, were age and gender matched with the CD patients and did

not have any acute or chronic disease or other clinically significant disorders. All inclusion/exclusion criteria are reported in further details in Klemenak et al (2015) and are resumed in fig. 3.3.1.1. DNA concentration obtained by means of Qubit quantification represented a further inclusion criteria applied in the sequencing experimental design.

Therefore, 20 individuals were finally analyzed for both the first and the second group and were sampled at T0 and T1, while 16 subjects for the control group were sampled just one time, for a total of 96 samples sequenced (fig. 3.3.1.1).

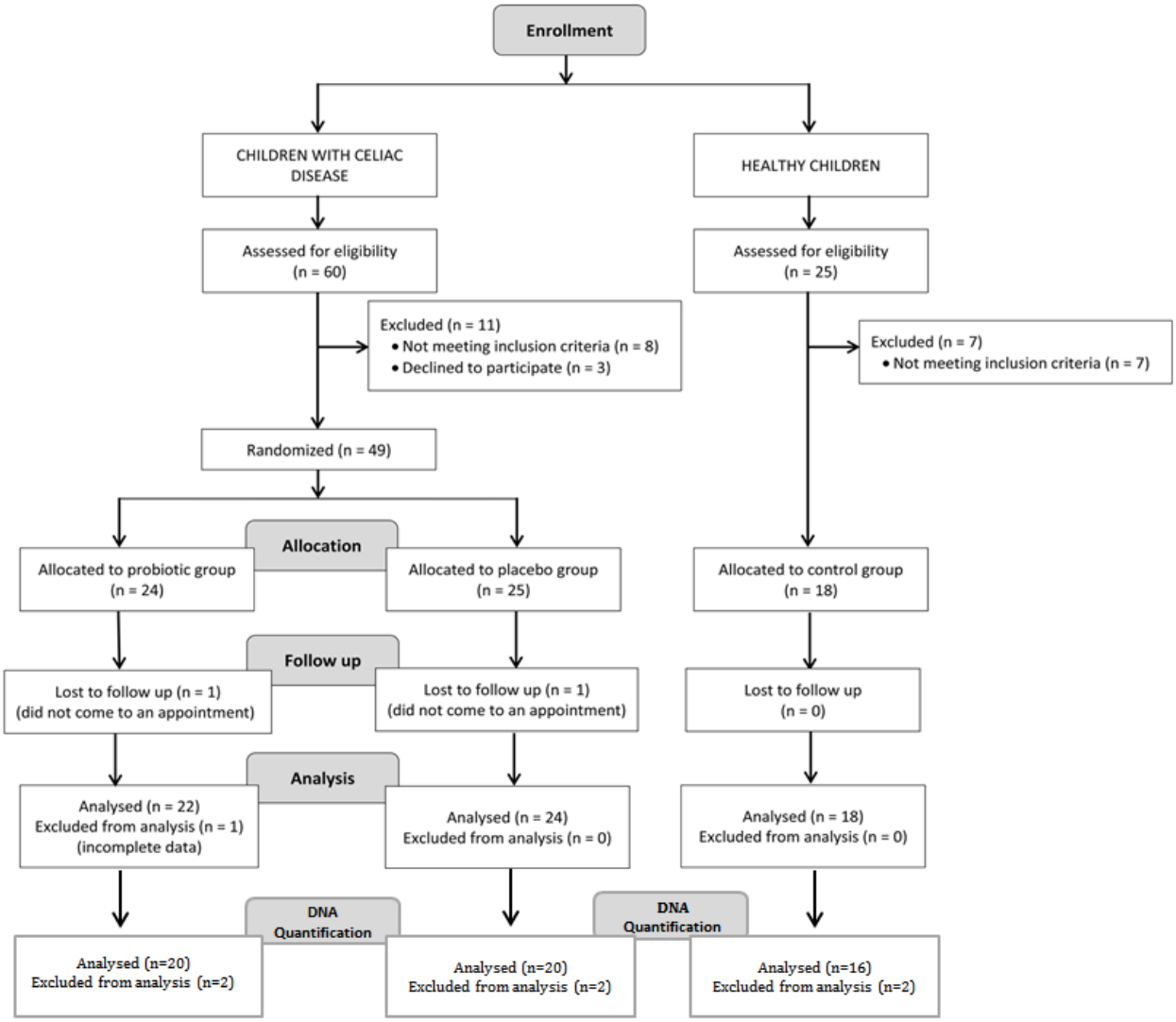


Fig. 3.3.1.1: Summary of inclusion/exclusion criteria adopted for samples selection.

3.3.2 Probiotic Treatment

Lyophilized probiotic strains composed for half part of *B.breve* BR03 and half part of *B.breve* B632, as well as placebo packages, were provided by Probiotical S.p.A. (Novara, Italy). To ensure that the study was blinded for both investigators and patients, probiotic cultures and placebo were packed in identical form and named with different codes. Each package delivered to patients contained 2g powder of probiotic or placebo, mixed with other fluids and were suggested to be ingested before breakfast during all the three months of treatment. The daily dosage of *B.breve* BR03 and *B.breve* B632 was of 10^9 CFU of each strain/g powder.

3.3.3 DNA Extraction and Quantification

For every sample, two hundred milligrams of feces (preserved at −80 °C after collection) were used for the DNA extraction using the QIAamp DNA Stool Mini Kit (QIAGEN, West Sussex, UK). A slight modification was performed during extraction: an additional incubation at 95 °C for 10 min. of the stool sample with the lysis buffer was added to the standard protocol to improve the bacterial cell rupture (Aloisio et al., 2014). Extracted DNA was stored at −80°C.

All samples were quantified by means of the Qubit™ 3.0 Fluorometer (Thermo Scientific) already described in paragraph 3.2.3 The samples had to reach a minimum of 5 ng/μl to be included in the project.

3.3.4 16S Metagenomic Sequencing with Illumina Platform

Multiple DNA sequencing methods were developed at the beginning of the 2000s and were then implemented in commercial DNA sequencers in the last ten years, reaching the ability to generate and capture millions of DNA molecules from a sample on a surface and then sequencing them simultaneously. Several sequencing platform have been thus developed, each one exploiting a different chemistry approach, but they are all named as "Next Generation Sequencing" (NGS), which are characterized by massive parallel sequencing ability.

The Illumina platforms exploit a sequencing by synthesis (SBS) technology that allows to sequence tens of millions of clusters on the flow cell surface in parallel. First of all, the sequencing template are immobilized on a flow cell surface (fig. 3.3.4.1), which is designed in a way that ensure enzyme access and high stability of surface-bound template.

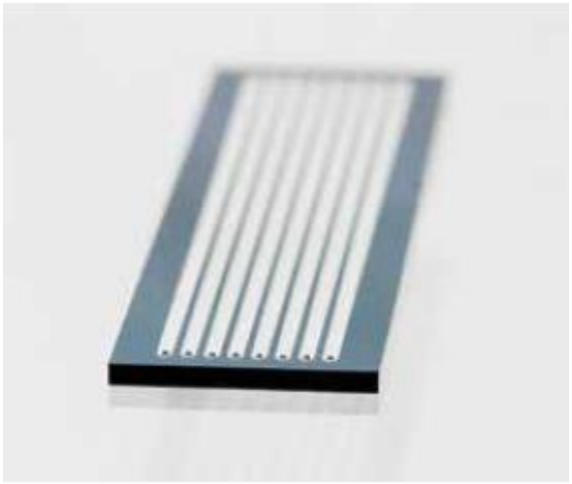


Figure 3.3.4.1: Flow cell surface used to capture sequencing template.

After library preparation, during which fragment of genomic DNA have undergone adapters ligation to both ends, the single-stranded sequencing template bind randomly to the cell surface. During every sequencing cycle, a single labeled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain. These nucleotides serve as terminators of polymerization, thus after dNTP incorporation, a fluorescent dye allows to identify the base and is then enzymatically cleaved to allow incorporation of the next nucleotide. Base calling is made directly from signal intensity measurements during each cycle. Thus, at the end of this process, the result is a highly accurate base-by-base sequencing that lets to considerable reduction of sequence-context specific errors (fig. 3.3.4.2).

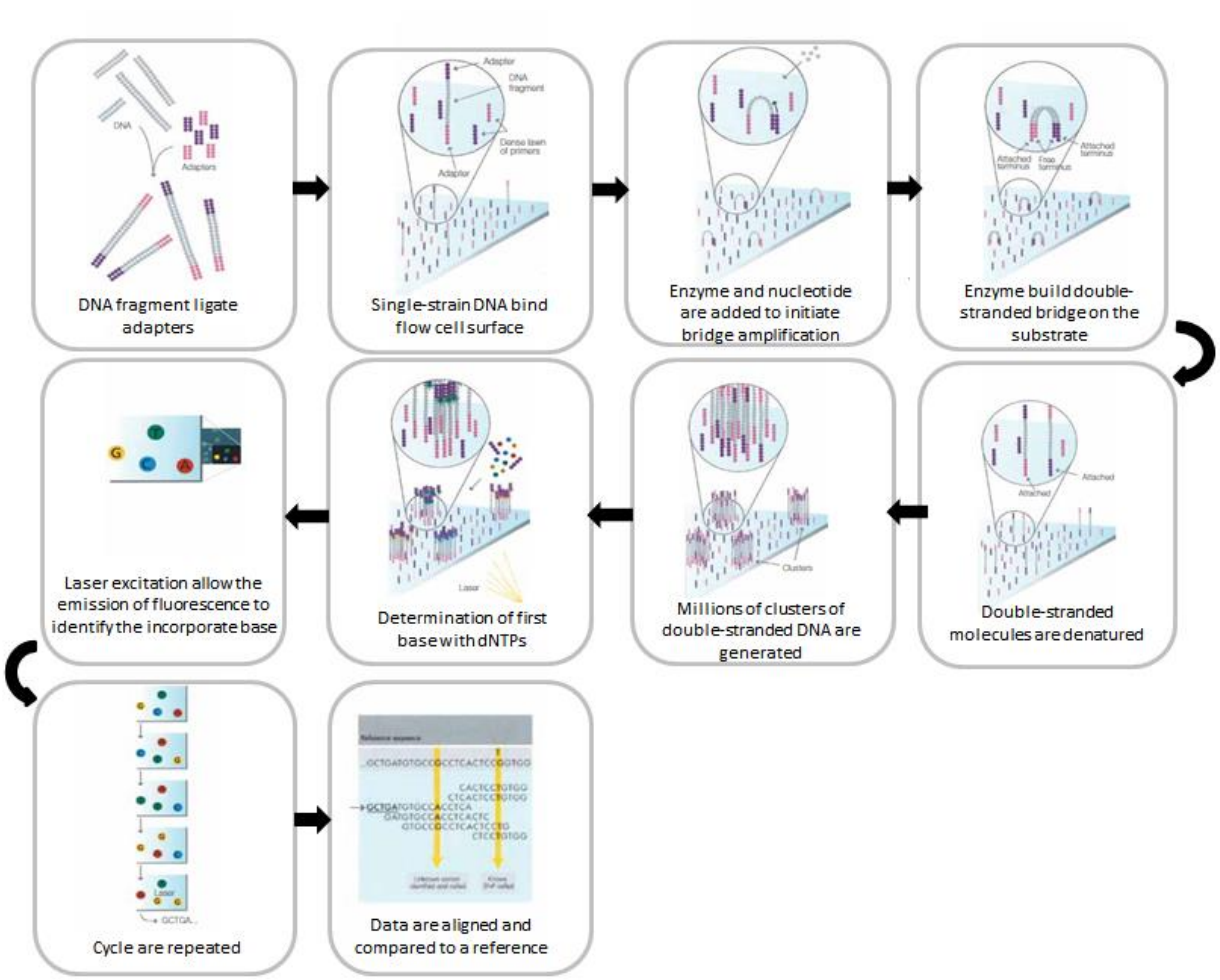


Figure 3.3.4.2: Schematic representation of the Illumina sequencing workflow.

Illumina platforms are widely used in metagenomics studies to perform the analysis of the prokaryotic 16 ribosomial RNA gene (16S rRNA), which is approximately 1,500 bp long and is composed of nine hypervariable regions interspersed between conserved regions. These variable regions are normally used in phylogenetic studies on diverse microbial populations. Which is the most informative region to be sequenced is still on debate and it depends on the scientific objectives of the study and on the typology of the available samples. Normally, for human gut microbiota analysis, V3 and V4 regions are used and Illumina developed a specific protocol to analyze these two regions on a *MiSeq* platform.

The protocol schedules the use of primer pair sequences for the V3 and V4 regions (Klindworth et al., 2013) in order to create a single amplicon of ~ 460bp, and even the overhang adapter sequences that must be appended to the primer pair sequences for the compatibility with Illumina indexing (fig. 3.3.4.3).

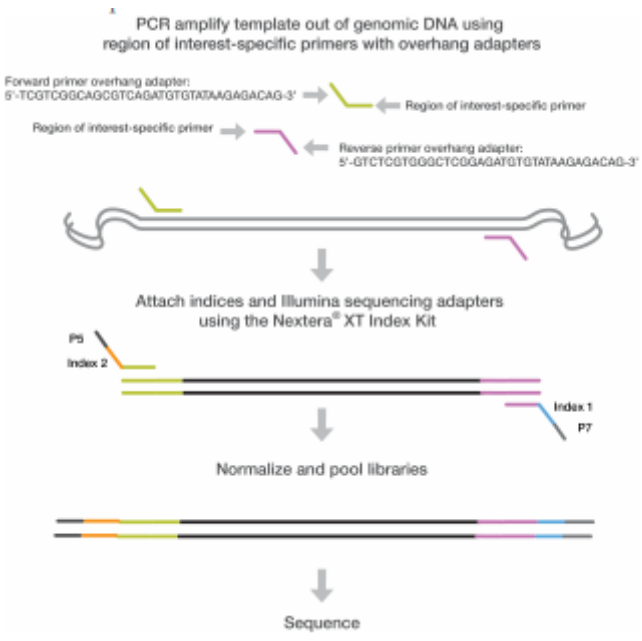


Figure 3.3.4.3: 16S V3 and V4 Amplicon Sequencing Workflow

In the first step, a preliminary PCR is performed to amplify the template of a DNA sample using 0.2 ng/ul purified gDNA and the specific primers with the overhang adapters attached. Then a cleaning process is necessary to remove free primers or primer dimers, and was achieved by using *Agentcourt® AMPure® XP* beads (Beckman Coulter Genomics, United Kingdom) on a magnetic stand.

A second PCR step is then performed to attach dual indices and Illumina sequencing adapters by using the *Nextera XT Index Kit* (Clontech, United States), followed by a cleaning step. The so prepared library are quantified with Qubit™ 3.0 Fluorometer (Thermo Scientific) and then normalized and pooled prior to sequencing.

Finally, the pooled libraries were sequenced using a 2x300 bp paired-end chemistry (MiSeq Reagent kit v3 (MS-102-3001)) on a Illumina MiSeq Sequencer according to manufacturer's instructions.

3.3.5 Bioinformatics analyses: from sequences quality controls to biodiversity analysis

Several bioinformatics pipelines have been developed to analyze the huge amount of data produced during this project. The first step of analysis is represented by quality controls of the generated raw data, which are essential to be confident of the quality of experiments' results. For this purpose, the *FastQC 0.11.4* software (Babraham Bioinformatics) and the *prinseq-lite.pl* script (Schmieder et al., 2011) were used to process fastq files produced by the adopted Illumina platform. Fastq files contain information about both read sequences and the respective quality scores for every base called in the sequence. The quality score is assigned according to the Phred Quality Score (Q score), a parameter derived from Sanger sequencing, which indicates the probability

that a given base is called correctly by the sequencer. A low Q score can lead to increased false positive variant calls and so inaccurate conclusions. The *FastQC* tool was used for a rapid visualization of sequences quality, then with the *prinseq-lite.pl* script sequences have been trimmed according to various quality criteria: first of all sequences with less than 50 bp were eliminated, then remaining reads were analyzed with a sliding-window approach of 20 bp, within this range each sequence with a mean quality lower than 20 was removed.

After that, the *fastq-join* tool from the ea-tools suite (Aronesty, 2011), was used to join forward and reverse sequences. The last quality control step was represented by the elimination of chimeric sequences using the *Usearch* tool (http://drive5.com/usearch/). Chimeras are hybrid products between multiple parent sequences that can be falsely interpreted as novel organisms, thus inflating apparent diversity. Accordingly, this quality control step is of crucial importance for the achievement of reliable results .

Once high-quality double-stranded reads were obtained, they are aligned to the 16S reference sequences database at the RDP database project (*rdp_gold.fasta*) to identify the microbial community with the *RDPclassifier* tool (Yemin Lan et al., 2012).

RDPclassifier outputs have been then processed through several R software packages, such as *vegan*, *reshape2*, *RDPutils* and *phyloseq*. These tools allow to perform the principle taxonomic analyses used in metagenomics studies and the estimation of various biodiversity indices to depict a high confident description of the examined microbiota composition.

Finally, as suggested by McMurdie et al. (2014), data have been normalized and the function *exactTest()* of the *edgeR* package was used to evaluate the effective microbial differentiation among the studied groups.

Chapter 4

Results

Project 1: Detecting signals of natural selection in European populations on genes involved in nutritional and thermoregulation processes.

4.1.1 Structure Analysis

In order to evaluate the impact of several nutritional and ecological shifts on the genome of individuals belonging to populations with European ancestry, the present research explored variability patterns at 31 genes selected from literature and public databases according to their involvement in various thermoregulation and nutritional processes (reported on table 3.1.1.1 of chapter 3).

To investigate patterns of genetic variation at the examined genes and to infer the main evolutionary forces that have shaped them in populations of European ancestry, as first DAPC was applied to the whole set of genes, as well as to data from each single gene, independently.

According to this approach and when the totality of the genes has been simultaneously analysed, signals of population structure were detected, with closer relationships being observed among Northern European populations than among Southern ones (fig. 4.1.1.1). Since it is possible that this pattern of variation could be the combined result of the same selective pressure having acted on various genes or otherwise to the substantial

contribution of few or a single gene, several independent DAPC analyses were performed gene by gene.

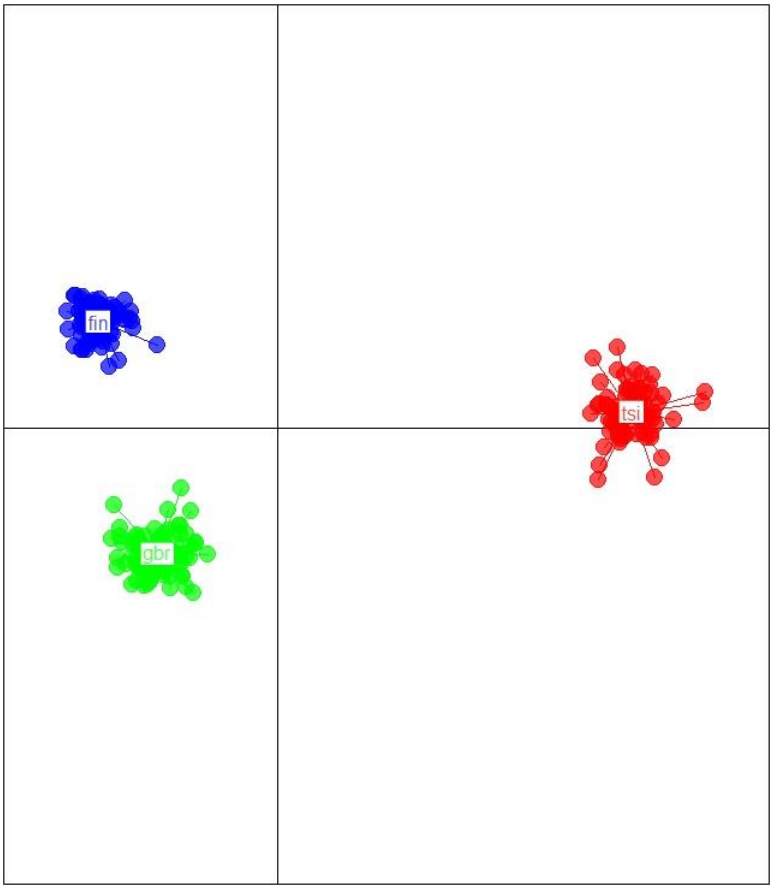
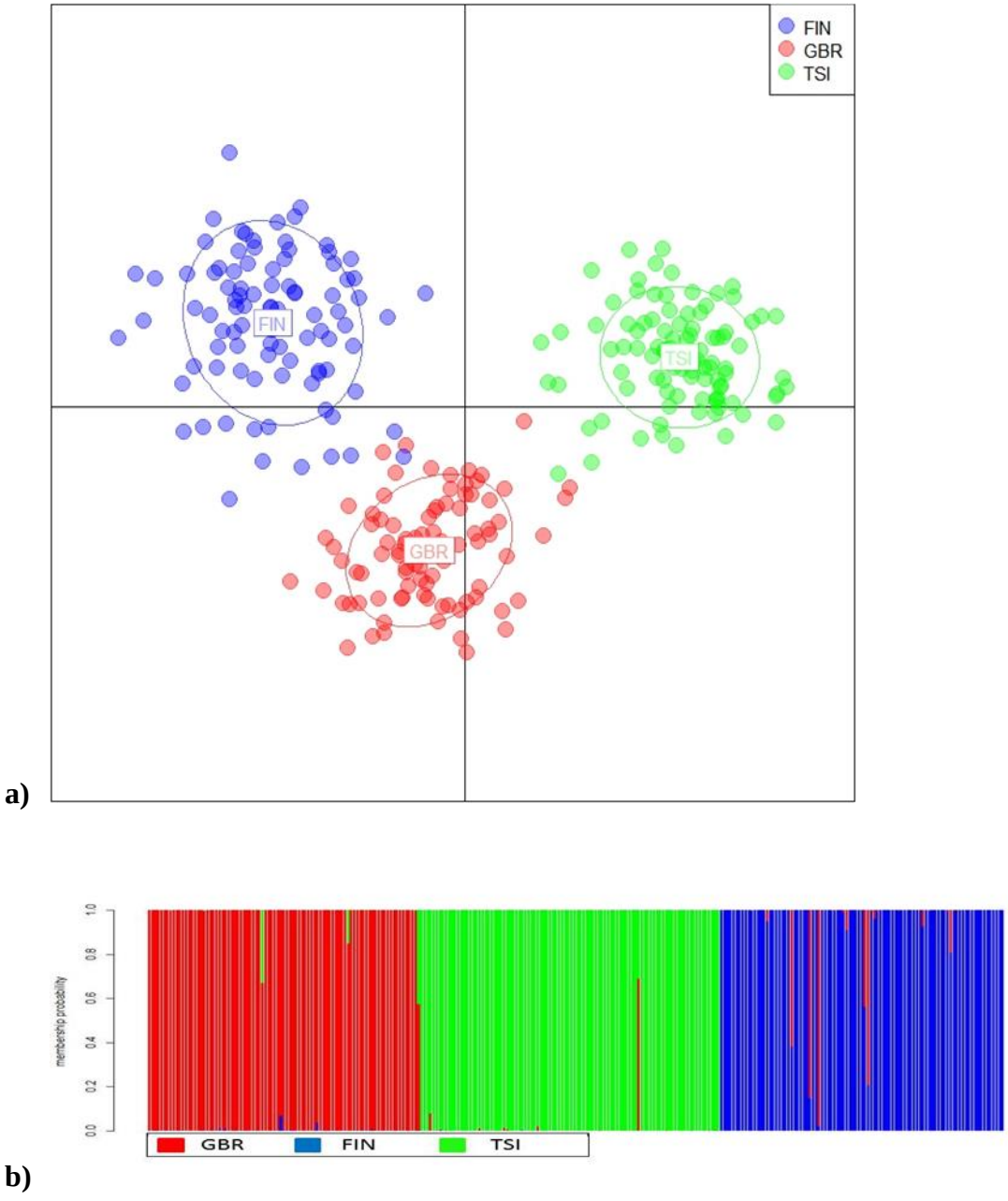


Figure 4.1.1.1: DAPC analysis on the whole pruned dataset of 31 gene showing a closer relationships in population structure between Northern European population than with TSI group. FIN are reported in blue, GBR in green and TSI in red.

Accordingly, almost the totality of the chosen genes showed highly monomorphic profiles within the selected European populations, with few or none signals of differentiation among these groups, except for the *PRDM16* gene, which turned out to be sole genomic region showing considerable heterogeneity among the examined European populations (fig. 4.1.1.2).



a)

b)

Fig 4.1.1.2: a) DAPC analysis showing clear discrimination among the three examined European populations, that are indicated by different colours and ellipses. b) Membership probabilities for each individual to belong to the identified clusters indicating high internal homogeneity within each population.

In fact, discriminant analysis of PCs computed on the pruned set of variants pointed out patterns of extremely reduced population structure for all the genes included in the dataset, except for *PRDM16* that showed a considerable level of differentiation among the examined human groups.

In particular, a closer genetic relationship was observed between the FIN and GBR groups than between TSI and the other two populations. That being so, the subsequent analyses focused on the 2,720 SNPs annotated on the 369 kb genomic interval covering the *PRDM16* locus (chr1: 2,985,732-3,355,185; GRCH37), to formally test whether the pattern of variation observed within Europe could be due or not to the action of natural selection.

4.1.2 Detecting natural selection

To unequivocally test this hypothesis, different neutrality tests were performed. First of all, site frequency spectrum-based neutrality tests have been performed by applying a sliding window approach and by considering genomic windows of 10 kb. Several windows showing significant results for the Fu & Li'D and F tests after comparison with statistic distributions, (obtained by 10,000 coalescent simulations) were identified, as reported in fig. 4.1.2.1. In particular, ten windows showed a significant value for the Fu & Li'D test and just one for the Fu & Li'F. In order to further shortlist the most plausible candidate windows, Bonferroni's correction for multiple tests was applied pointing to only four highly significant genomic intervals (adjusted $p \leq 0.01$): two for FIN and one for each of the other populations.

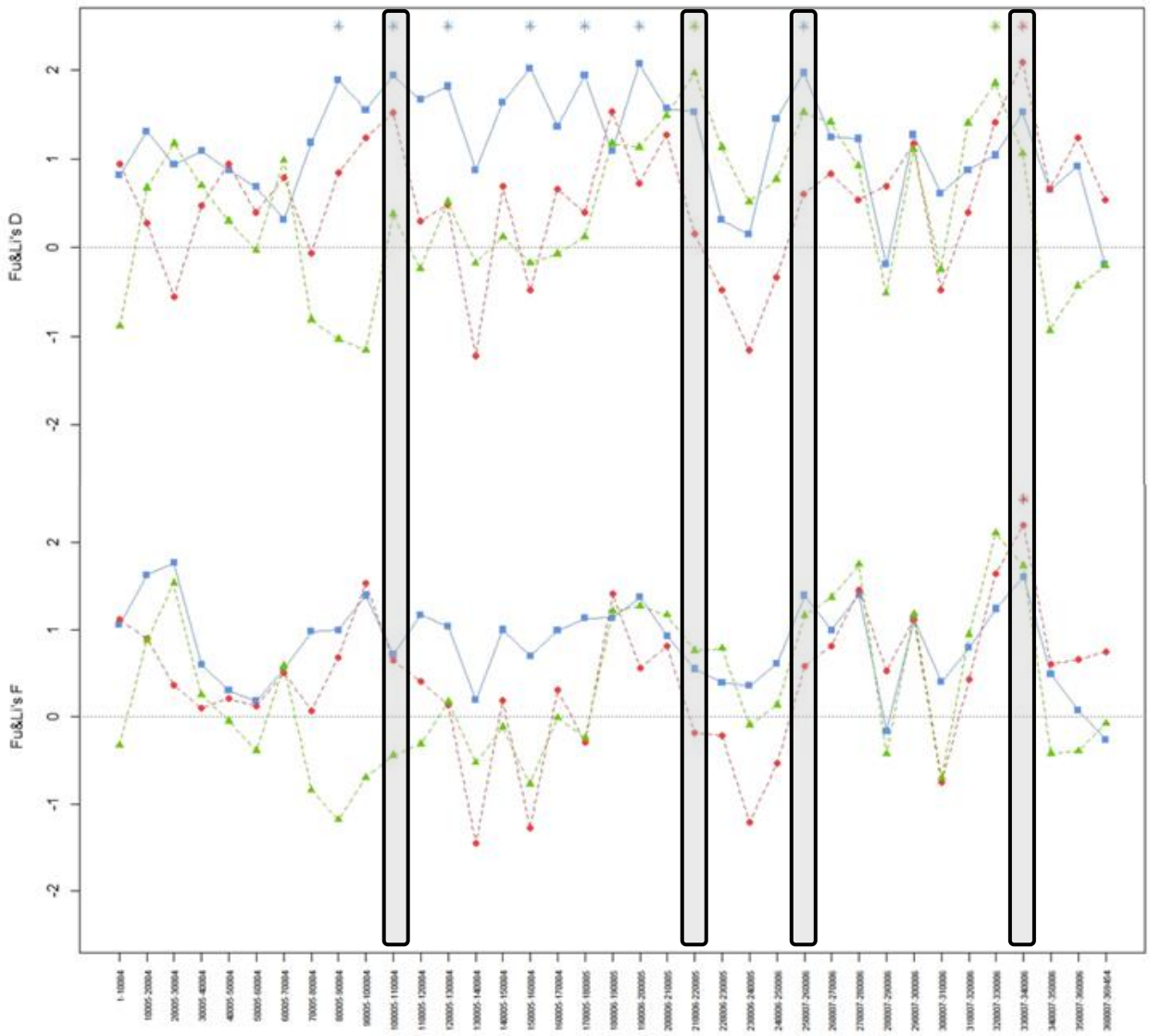


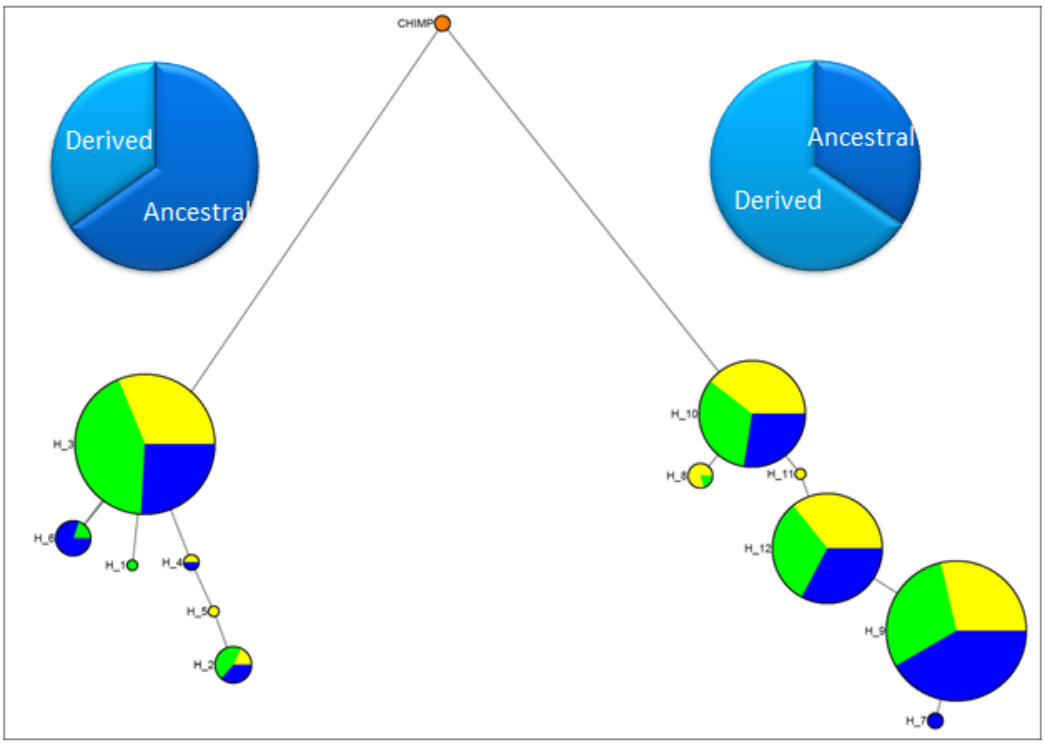
Figure 4.1.2.1: Results for the Fu & Li' D and F statistics. Each line indicates a different population: FIN (blue), GBR (red) and TSI (green). On the x-axis are reported the 37 10 kb-windows in which the *PRDM16* gene was divided. Asterisks indicate significant results after 10,000 coalescent simulations and their colour is relative to the population for which the significant result was obtained. The black columns highlight windows that remained significant after Bonferroni correction for multiple testing.

All these windows showed positive values, thus indicating possible signatures of balancing selection having acted on these genomic regions. The identified significant windows included several intronic SNPs, as well as variants related to non-coding transcripts. Furthermore, to deepen the investigation of the *PRDM16* evolutionary history

and to provide additional evidence supporting the hypothesis that balancing selection has acted on such genomic region, haplotypes analyses have been performed by focusing on the SNPs included in the identified candidate windows.

First of all, it was explored the haplotype structure at the window for which significant results were obtained for both Fu & Li tests in the GBR group (i.e. 330,007-340,006) due to its high functional relevance. A LD analysis, focused on a genomic interval including 100 kb downstream and upstream regions with respect to the candidate window, was performed to exclude hitchhiking effects due to the presence of adaptive variants on the *PRDM16* nearby genes. Accordingly, SNPs in high LD ($r^2 \geq 0.90$) with the considered variants were retrieved and used to reconstruct haplotypes at the examined genomic interval. A median joining network was finally used to explore evolutionary relationships among the inferred haplotypes, pointing to the distinction of two main haplotype clades with deep coalescence time. The former was characterized by a single common haplotype carrying mainly ancestral alleles, while the latter presented three frequent haplotypes mostly made up of derived alleles (fig.4.1.2.2a). Looking at the haplotype frequencies (fig.4.1.2.2b), GBR showed similar proportions for the above mentioned "ancestral" and "derived" haplotypes (29,8% and 27% respectively), while TSI and FIN exhibited a major presence of the "ancestral" and of the "derived" clades, respectively, suggesting that ancient balancing selection has acted on this region, leaving still appreciable genomic signatures predominantly in the GBR population, and that other evolutionary forces have instead further shaped haplotypes distribution in the other European groups.

a)



b)

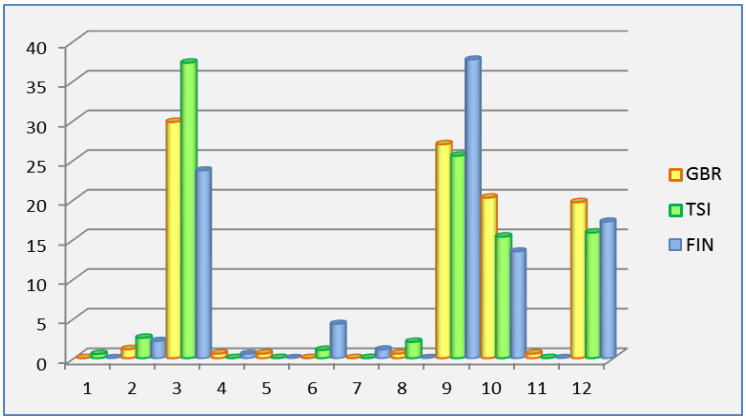


Figure 4.1.2.2: a) Median joining network of haplotypes carrying SNPs in high LD ($r^2 \geq 0.90$) with variants included in the GBR significant window spanning from nucleotide position 330,007 to 340,006. The blue pies represent the percentage of derived and ancestral alleles at the two clades. b) Haplotype distribution among the examined populations.

To better explore the potential action of more recent selective pressures, the iHS test was also computed for each *PRDM16* SNP. Accordingly, several SNPs have been identified as potential targets of positive selection in the three studied populations (fig. 4.1.2.3).

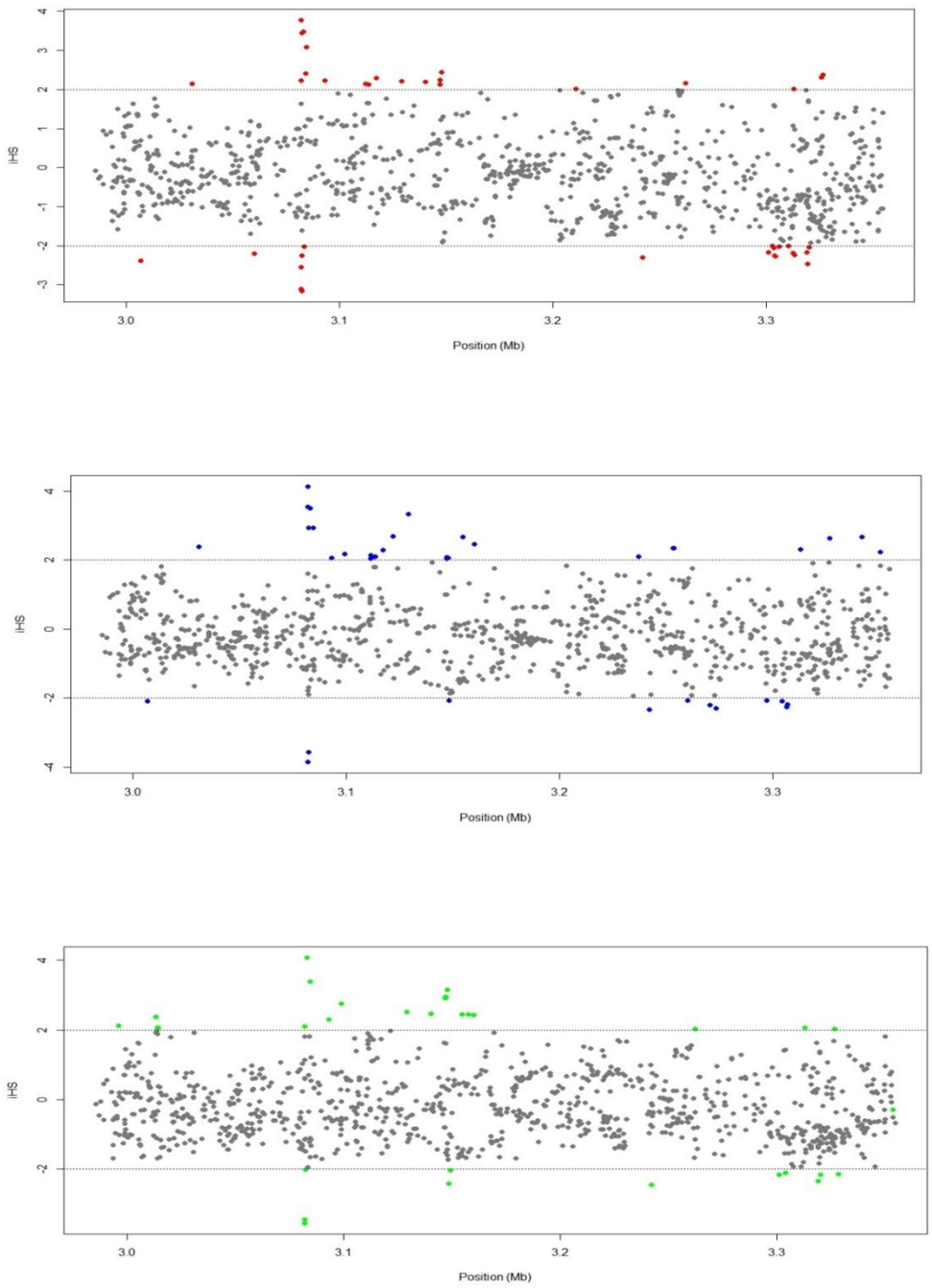
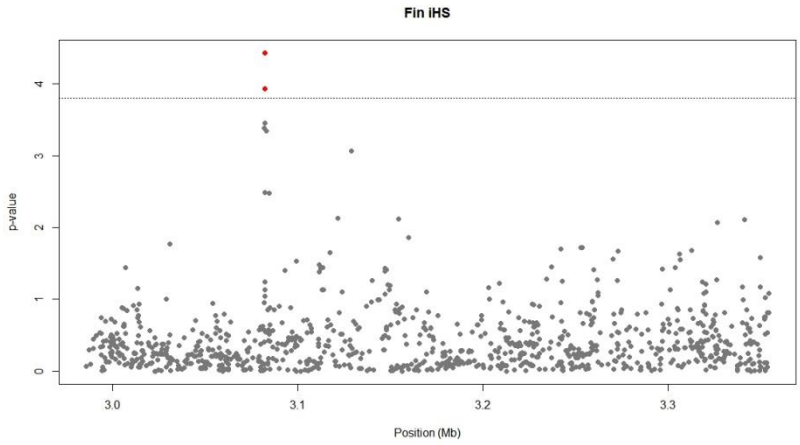


Figure 4.1.2.3: iHS plots with coloured spot indicating core SNPs with |iHS| value > 2 along the *PRDM16* gene (x axis). Blue, red and green colours indicate FIN, GBR and TSI populations, respectively.

In particular, two limited genomic intervals within the *PRDM16* gene showed a series of consecutive SNPs characterized by high iHS values. A first genomic interval spanned from nucleotide position 3,082,126 to 3,084,761, while the second covered a 3,129,182 – 3,147,970 interval. Anyway, it is interesting to note that in all populations the top 1% and 90% (for both highest positive and negative values) of iHS significant SNPs, are mainly concentrated in the first cited genomic interval (i.e. positions 3,082,093-3,084,761). When a more stringent significance threshold was considered by applying Bonferroni correction for multiple testing, a very small number of SNPs still showed statistically significant iHS values, being limited to the FIN (rs112682827 and rs144090205) and TSI (rs2817126) populations (figure 4.1.2.4, table 4.1.2.1). Both rs2817126 and rs112682827 showed highly positive iHS values (iHS = 4.07 in TSI and iHS = 4.12 in FIN), suggesting positive selection having acted on the ancestral allele, while rs144090205 was characterized by a strong negative iHS score (iHS = -3.85 for the FIN group) thus indicating that selection acted on the derived allele of this SNP (table 4.1.2.2).



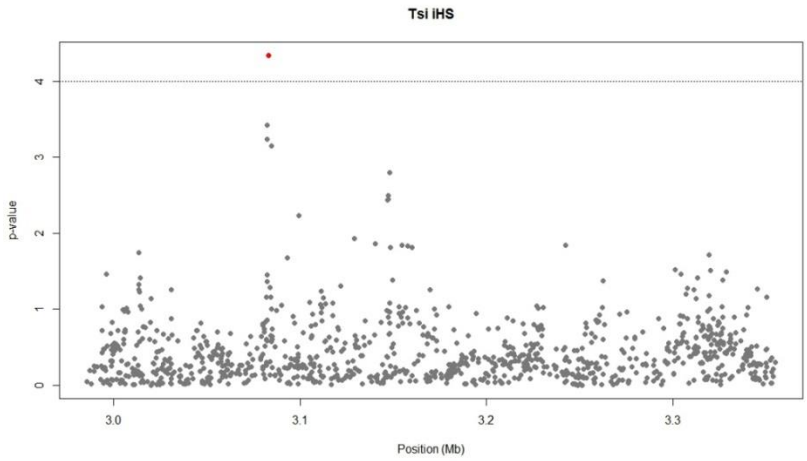


Figure 4.1.2.4: Manhattan plot of computed iHS p-values. The dotted line indicates the significance threshold identified by Bonferroni correction for multiple testing and red spots indicate SNPs with p-values lower than that threshold.

Table 4.1.2.1: Summary iHS statistics for the analysed populations.

Pop	Number of SNPs with |iHS|>2.0	Max |iHS|	Highest iHS p-value	Bonferroni threshold	Significant SNPs
GBR	42	3.75	0.00015	0.00011	0
TSI	32	4.07	0.00004	0.00009	1
FIN	40	4.12	0.00003	0.00013	2

Table 4.1.2.2: Candidate SNPs that have plausibly undergone positive selection: **POP** = Population; **Allele** = selected allele; **MA** = Minor Allele; **P-value** = p-value after Bonferroni correction for multiple testing; **EHH** = dimension of the region showing extended haplotype homozygosity and in which the core SNPs are located.

RS	Position	Pop	Allele	MA	p-value	EHH (bp)
rs2817126	3166682	TSI	G	T	0.000043	132596
rs112682827	3165673	FIN	T	G	0.000036	33508
rs144090205	3165687	FIN	T	T	0.00011	40186

These three intronic SNPs are all concentrated in the first of the above-mentioned genomic intervals and did not show appreciable LD with each other. It is interesting to note that the extension of haplotype homozygosity (EHH) around the core SNPs appreciably differed between the FIN and TSI populations (fig. 4.1.2.5), indicating that positive selection has plausibly acted on them at different times and/or with different strength.

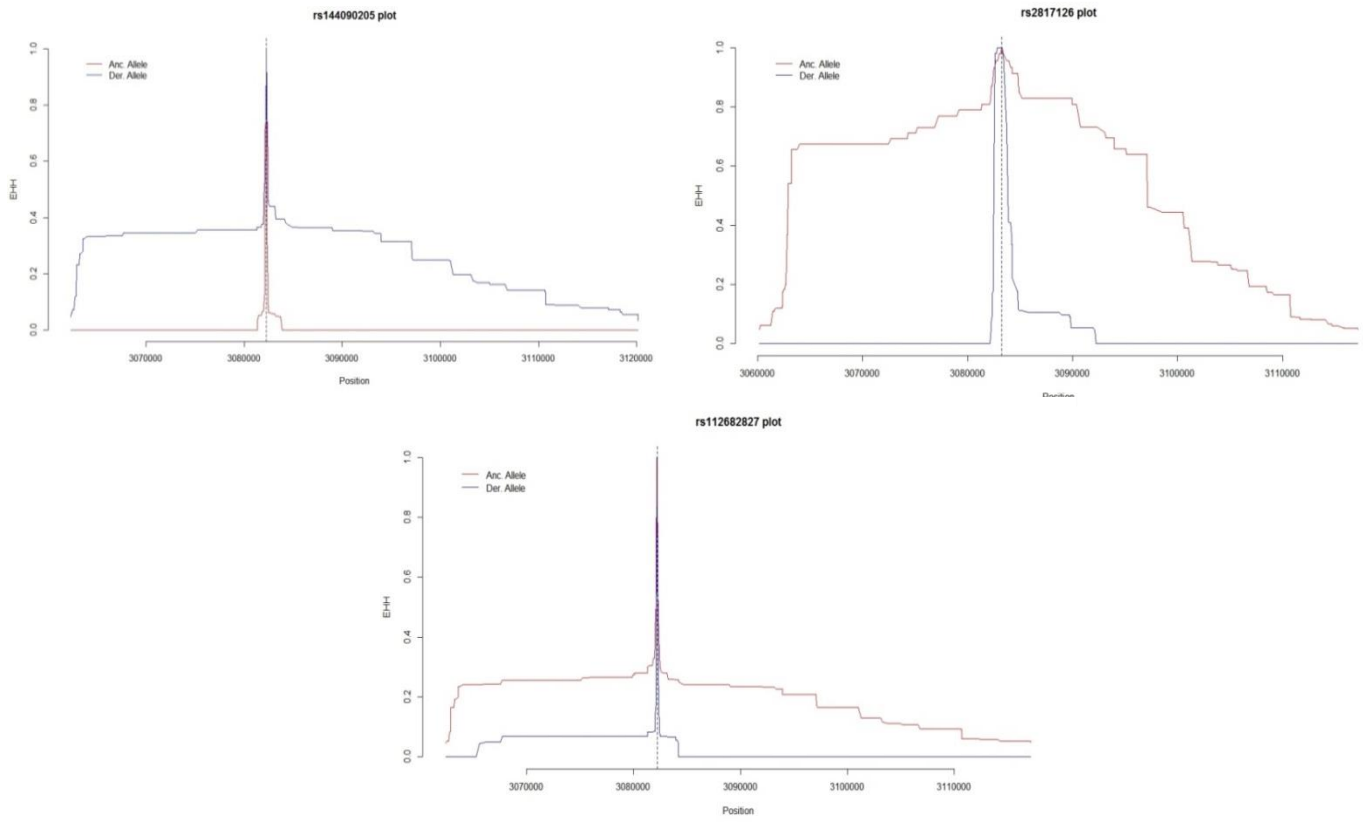


Figure 4.1.2.5: EHH decay plots for rs2817126 (TSI), which showed the longest haplotype homozigosity, rs112682827 and rs144090205 (FIN), which were characterized by reduced EHH.

Moreover, the value of extended homozygosity for the haplotype carrying the rs144090205 derived allele is shared by the majority of FIN individuals, as suggested by

branches thickness displayed in the bifurcation plot (fig. 4.1.2.6), thus providing a further evidence of positive selection having actually acted on such population.

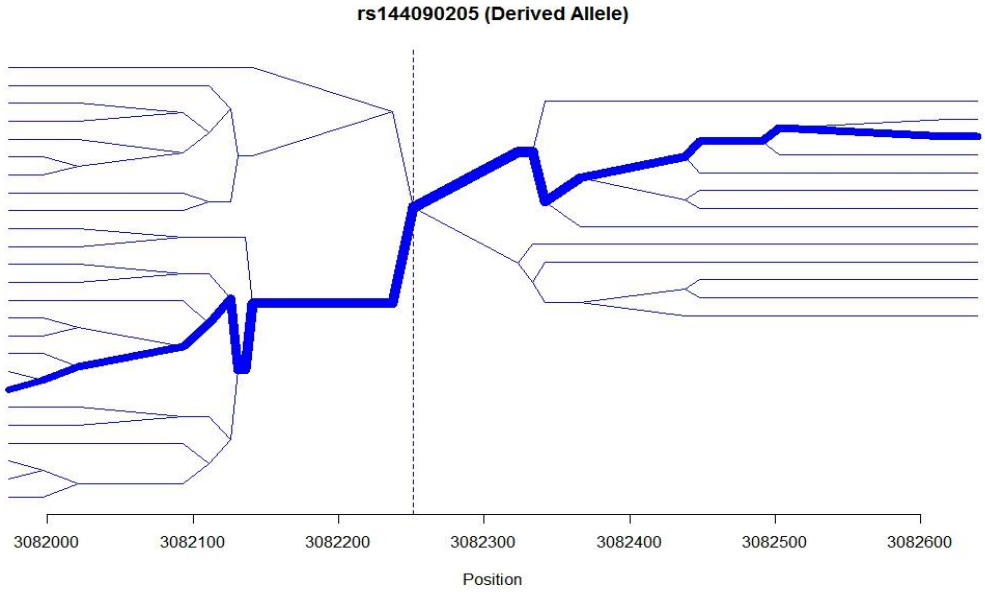


Figure 4.1.2.6: Haplotype bifurcation diagram indicating the breakdown of linkage disequilibrium in the FIN population.

To get further insights into the possible causes that could explain these multiple signatures of recent positive selection on European populations, age estimation of the detected selective sweeps was obtained as described in Voight et al. (2006). In particular, average genetic distance between each core SNP and the first point to the left and to the right dropping below EHH = 0.20 was computed. Within this range, a value of 0.39 cM and of 1.15 cM was observed for rs112682827 and rs144090205, respectively and regarding the FIN group, while a value of 1.97 cM was obtained for rs2817126 in TSI. Therefore, as deducible also from the observed haplotype structure, a more recent sweep seems to have acted on the TSI group. Although it is hard to obtain a rigorous age estimation, using the above-mentioned method and assuming a generation time of 25

years, rough ages of approximately 3,742 years, 1,429 years and 1,092 years before present were calculated for rs112682827, rs144090205 and rs2817126, respectively.

Project 2: Analysis of Italian genetic variability at three genes associated to metabolic functions.

4.2.1 Results for SNPs selection and genotyping.

Taking advantage from the *1000Genome Project*, genetic data relative to the whole *TMEM163*, *RAB3GAP1* and *R3HDM1* genes have been retrieved for all the available European populations not admixed, thus without considering the CEU group. For every gene, both the upstream and downstream regions (+/- 10.000 bp) with respect to the examined loci have been taken into account in order to consider in the study even regulatory regions of each locus. The so formed dataset, was composed by a total of 6,648 SNPs, which have been analysed for their heterozigosity level within Europe and then a further LD analysis was performed to eliminate possible bias in the results due to linkage disequilibrium. Therefore, the resulted most heterozygous SNPs have been screened for their LD value, and SNP pairs showing an r^2 value higher than 0.1 were removed, in order to retain only independent SNPs. The same approach was applied also to the F_{st} analysis, so that all the dataset has been screened locus by locus for pairwise population comparisons. Then, loci showing F_{st} values within the 1^{st} percentile of obtained F_{st} distribution were considered and the previously described LD criterion was applied. The resulted 23 independent SNPs are summarized on table 4.2.1.1.

Table 4.2.1.1: Summary information on the genotyped SNPs

Gene	Chromosome	Selection Criteria	SNP	Position
TMEM163	2	heterozigosity	rs34166010	135152287
TMEM163	2	Fst	rs1010241	135154312
TMEM163	2	heterozigosity	rs12997281	135161631
TMEM163	2	Fst	rs667430	135183066
TMEM163	2	Fst	rs535900	135202455
TMEM163	2	heterozigosity	rs681148	135206934
TMEM163	2	heterozigosity	rs554567	135216739
TMEM163	2	heterozigosity	rs501955	135267860
TMEM163	2	heterozigosity	rs701645	135287411
TMEM163	2	heterozigosity	rs35564151	135372951
TMEM163	2	heterozigosity	rs10928512	135451302
TMEM163	2	heterozigosity	rs10928513	135456759
TMEM163	2	Sazzini et al,2016	rs6723108	135479980
R3HDM1	2	heterozigosity	rs7605229	136228887
R3HDM1	2	heterozigosity	rs10187054	136388473
R3HDM1	2	heterozigosity	rs3213943	136389840
R3HDM1	2	Sazzini et al,2016	rs1446585	136407479
R3HDM1	2	Fst	rs748841	136558157
R3HDM1	2	heterozigosity	rs12475516	136560761
RAB3GAP1	2	Fst	rs1592	135722143
RAB3GAP1	2	heterozigosity	rs1530559	135755629
RAB3GAP1	2	heterozigosity	rs7575241	135768352
RAB3GAP1	2	Fst	rs10187402	135771974

These SNPs have been then successfully genotyped by means of a *Sequenom* platform on 830 healthy unrelated Italian individuals in order to explore their patterns of variation along the Italian peninsula. Only 18 failed genotyped were obtained, ten of which relative to one individual from Feltre (Belluno province, Northern Italian population group) that was eliminated from subsequent analyses; the others were randomly collocated, so without affecting the general result.

4.2.2 Structure Analysis

After quality check of the outputs generated by the genotyping platform, a series of descriptive analyses have been performed with the aim of summarizing the overall patterns of genetic structure observable along the Italian peninsula as regards the examined loci. PCA analysis (fig. 4.2.2.1) on the whole dataset highlighted a clear differentiation among the three considered geographical macro-areas. As first, it is interesting to note a deep close relation (almost complete overlap) between Southern and Central Italian samples on PC1, which explained 30.45% of the overall variance, and a greater differentiation of Northern Italians, especially along PC2 (12.83% of variance), with respect to subjects from the rest of the peninsula. Furthermore, it is possible to appreciate a higher heterogeneity of the Northern sample, with some individuals being placed in proximity to southern and central ones along the PC1.

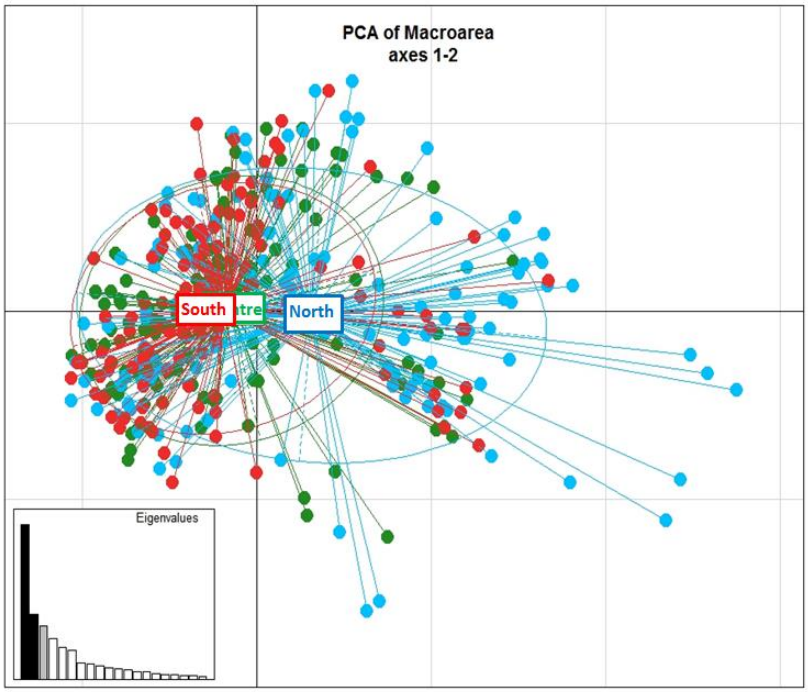


Figure 4.2.2.1: PCA plot on Northern (blue), Central (green) and Southern (red) Italian samples.

The same distribution of genetic variability came out from the DAPC analysis, which was performed to corroborate PCA results (fig. 4.2.2.2) and to further test the observed differentiation of Northern samples with respect to those from the rest of the peninsula.

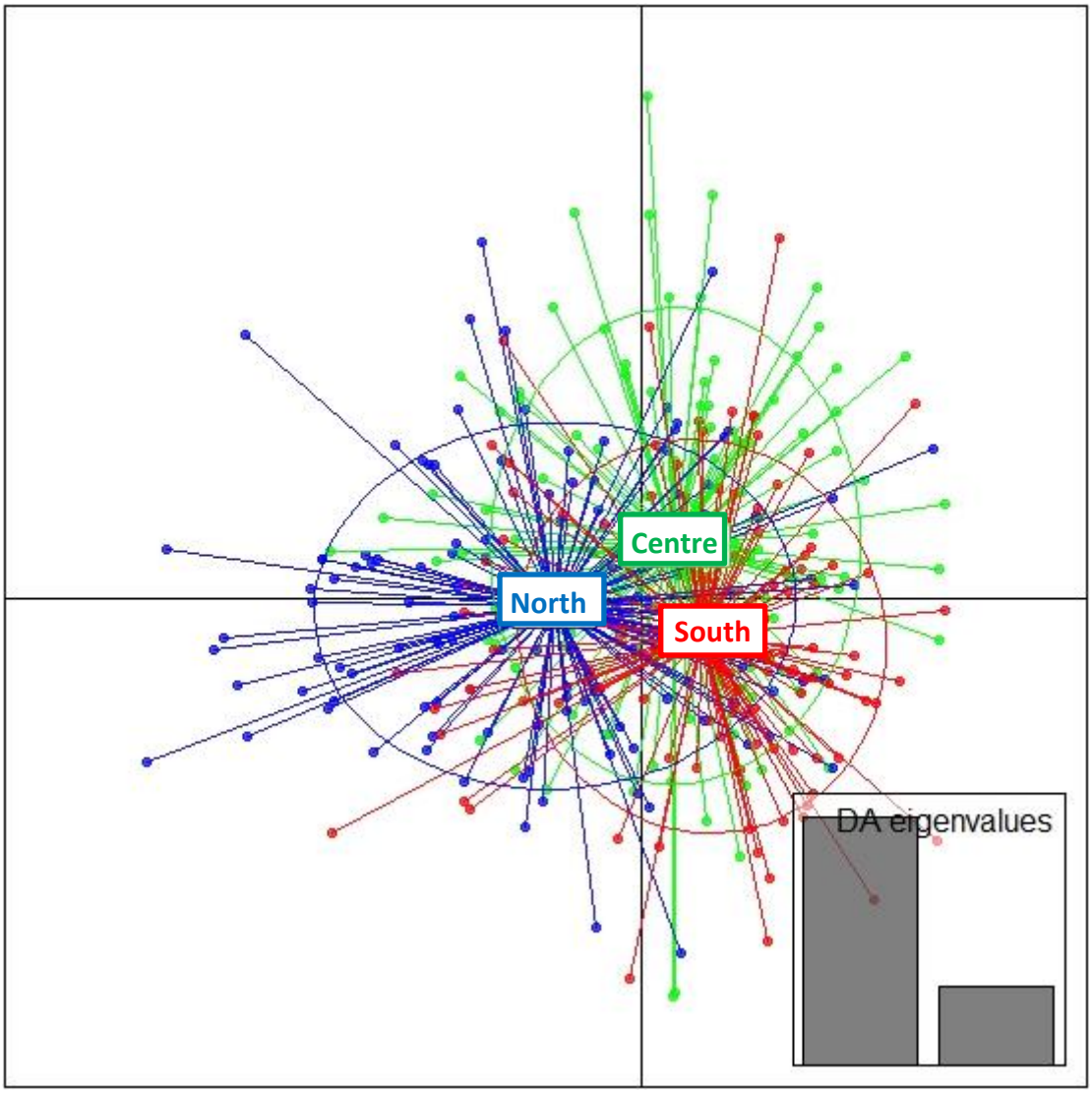


Figure 4.2.2.2 : Result from DAPC analysis performed on the Italian samples.

As visible from fig.4.2.2.3, by considering the loadings computed by PCA analysis is possible to identify those SNPs as the main responsible of this particular pattern of

variability observed within the Italian dataset. In particular, some of these SNPs seem to be associated with the characteristic differentiation of the Northern Italian samples.

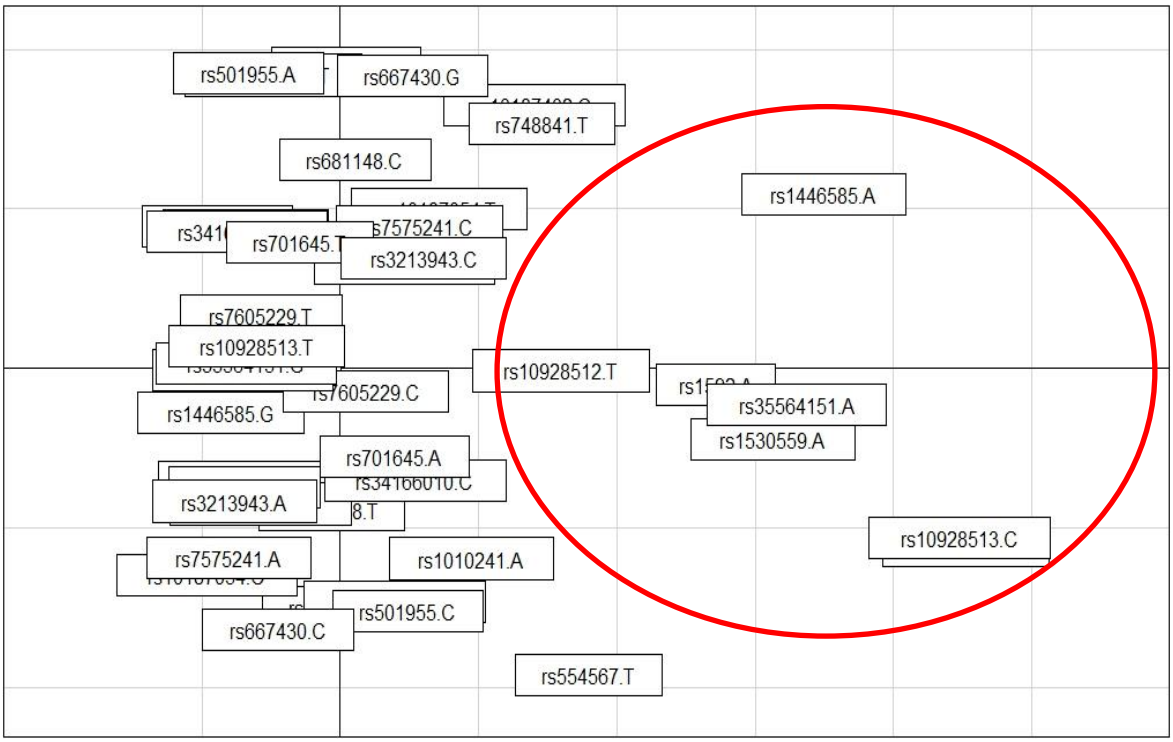


Figure 4.2.2.3 : Loadings plot obtained by PCA analysis. The circled SNPs are those that seem to distinguish Northern Italians from Central and Southern ones.

It is not surprising that these SNPs, which appear to be the main responsible of the detected North-South genetic differentiation, are also the same loci showing the highest F_{st} values in the performed pairwise comparisons between Italian population groups. Indeed, by focusing on results from F_{st} analysis, it is possible to note that most of them are following a trend characterized by high North/Centre differentiation, and by even higher North/South one. On the other hand, no appreciable genetic differences were detected between Central Italians and Southern ones, with actually most SNPs showing a

negative F_{st} value, which indicates higher differences within group than between groups (fig. 4.2.2.4).

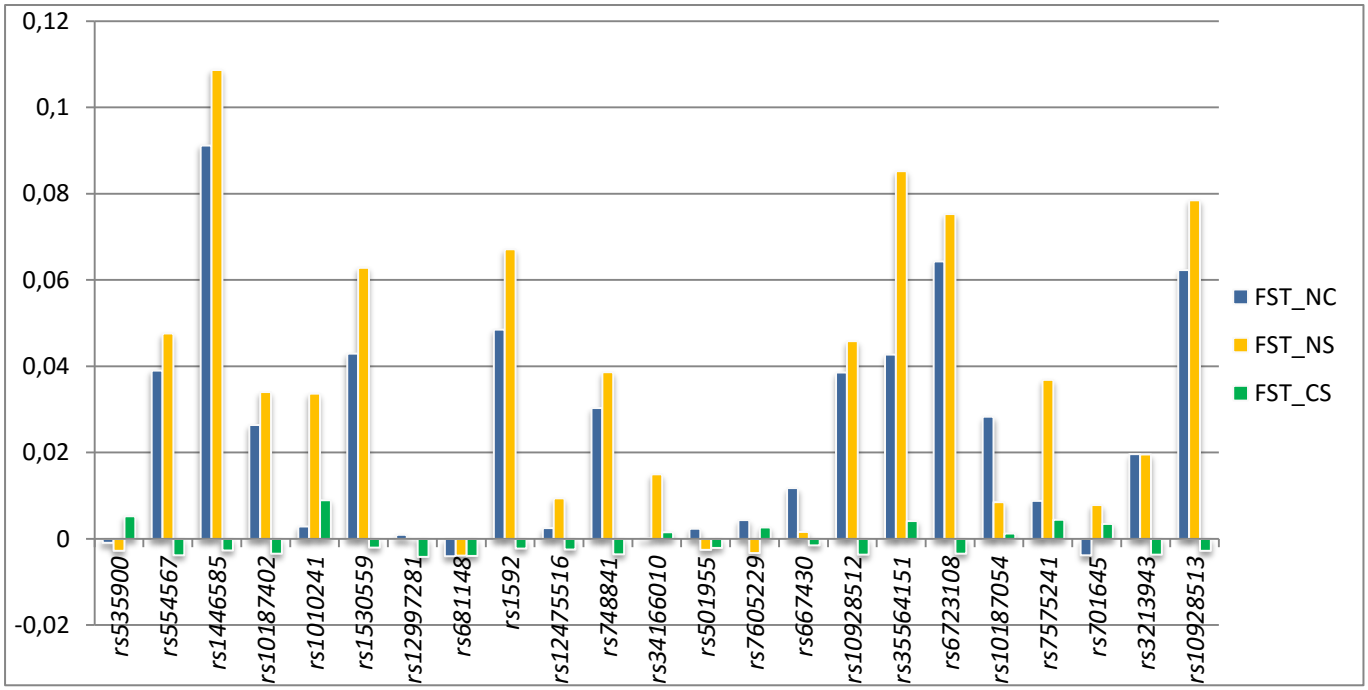


Figure 4.2.2.4: Fst values for each SNP analysed in this project.

Therefore, it is possible to appreciate a North to South gradient of variation and the robustness of this assumption was further tested by applying Fisher's exact test to contrast allele frequencies between the different groups. Accordingly, only two SNPs showed a significant p-value for the Fisher's exact test ($p < 0.01$), being the same two markers for which the highest F_{st} values were computed in the North/South comparison: rs1446585 and rs35564151 (fig. 4.2.2.5). These SNPs are not in LD with each other, so that we can rule out the fact that such finding is biased by hitchhiking effect.

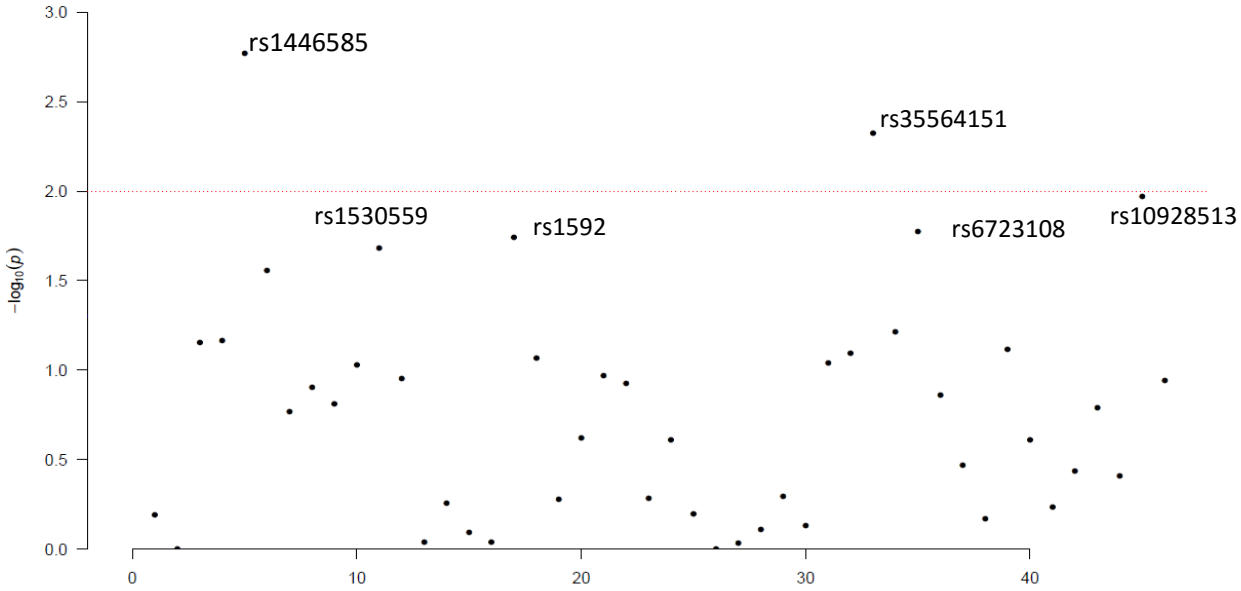


Figure 4.2.2.5: Manhattan plot of Fisher's extact test p values computed for the North/South comparison. The red line indicates the statistical threshold applied ($p < 0.01$), so that the SNPs reported above it are those showing significant p-values.

Moreover, it is interesting to note that for both the above-mentioned SNPs, the derived allele was significantly over-represented in the Northern regions of the Italian peninsula. In fact, the rs1446585 derived allele reached a mean frequency of 0.33 in Northern Italy, ranging from 0.27 at Feltre (Belluno province) to 0.38 in the province of Varese, while in Southern Italy it reached a mean frequency of 0.12, with the lowest values found at Reggio Calabria (0.02) and the highest one at Campobasso (0.20). A similar trend was observed also for the rs35564151, for which the derived allele showed a mean frequency of 0.26 in Northern Italy, again with a peak frequency in the province of Varese, being instead scarcely represented in Southern Italy (mean frequency of 0.09), and for instance totally absent at Castrovillari. In order to get a more clear visualization of the described distribution of these derived allele in Italy, two maps of SNP frequencies' gradients were generated by applying Kriging interpolation (fig. 4.2.2.6).

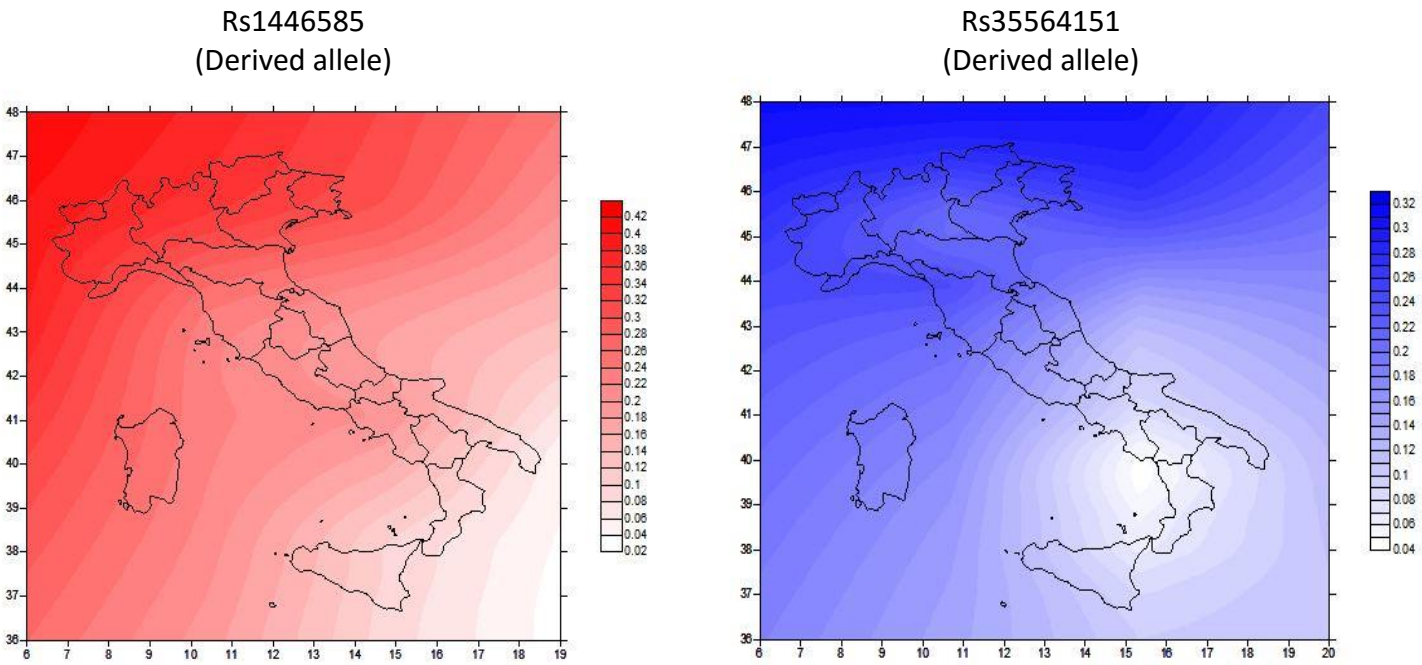


Figure 4.2.2.6: Maps of derived allele frequencies' gradients for the two significant SNPs along the Italian peninsula.

4.2.3 Comparison with genetic data from other humans populations

In order to evaluate how this pattern of variability is arranged within Europe, principal component analysis was repeated taking into account also genetic information for these loci for all available not admixed European populations sequenced by the*1000Genomes* Project (fig. 4.2.3.1). Accordingly, two main population clusters were easily discernible according to PC1, which explains 40% of the observed variability: one made up of North European samples (i.e. GBR and FIN) and another including the Italian population from Tuscany (TSI), as well as the typed Central and Southern Italians on PC2 (11%). This last group seems to distinguish itself even from the other available South European group (IBS), which instead lied in the middle of these two main clusters and with a slightly closer proximity to North Europeans.

Individuals from Northern Italy finally maintained a separated position with respect to the rest of the Italian samples, occupying a position exactly in the middle of the PCA space and thus getting closer to the rest of the European populations with respect to Central and Southern Italians.

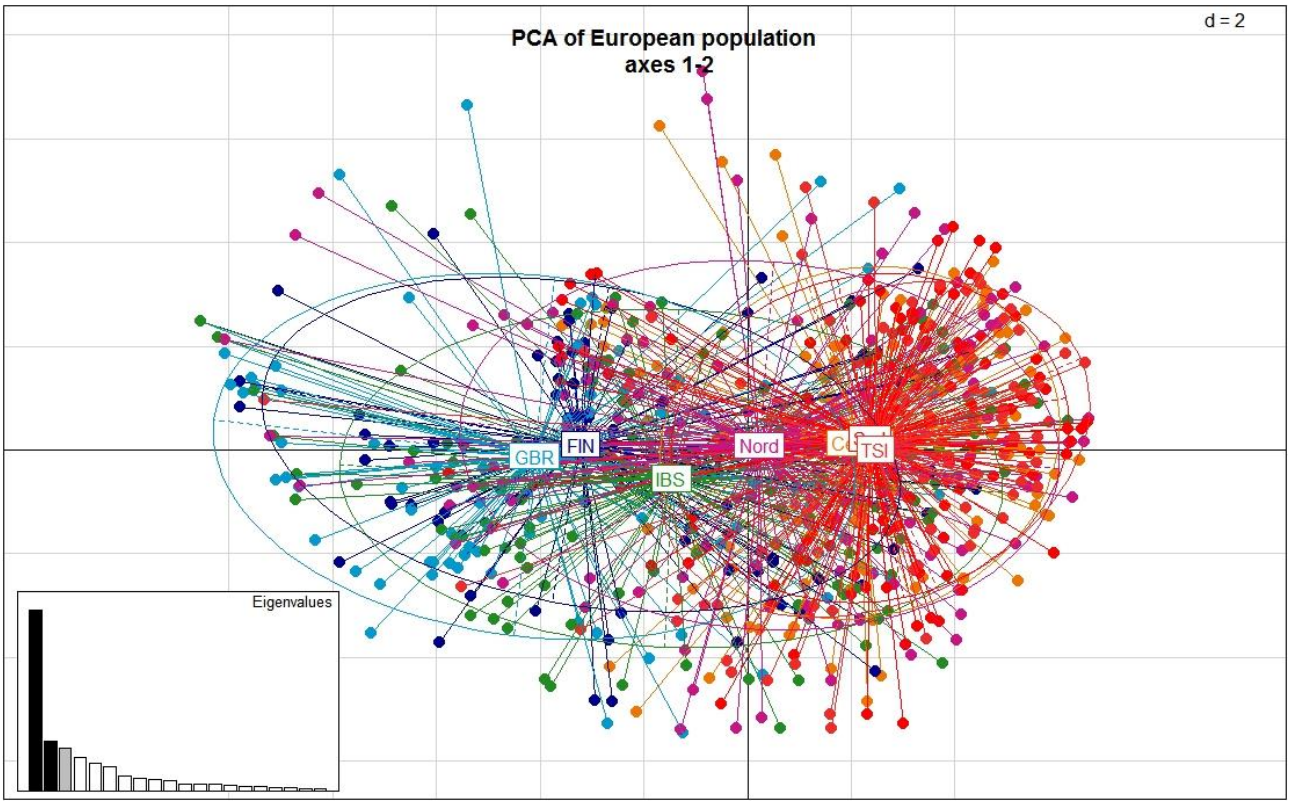
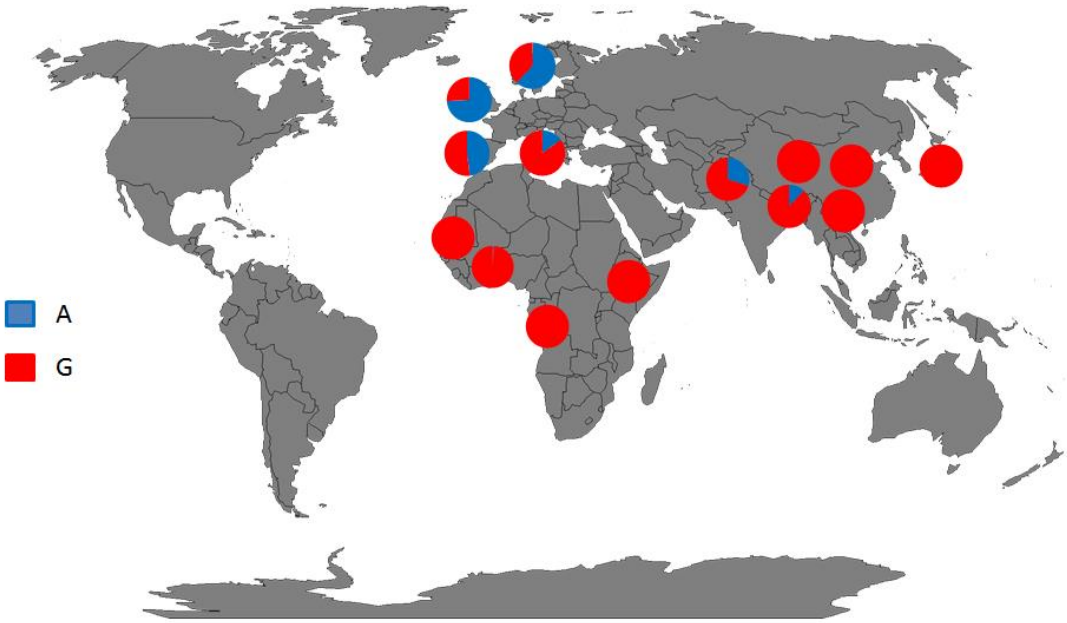


Figure 4.2.3.1: Distribution of the Italian samples with respect to the other European populations from the 1000Genomes database. FIN are displayed in dark blue, GBR in light blue, IBS in dark green, Northern Italians in magenta, TSI in light red, Central Italians in orange, Southern Italians in red.

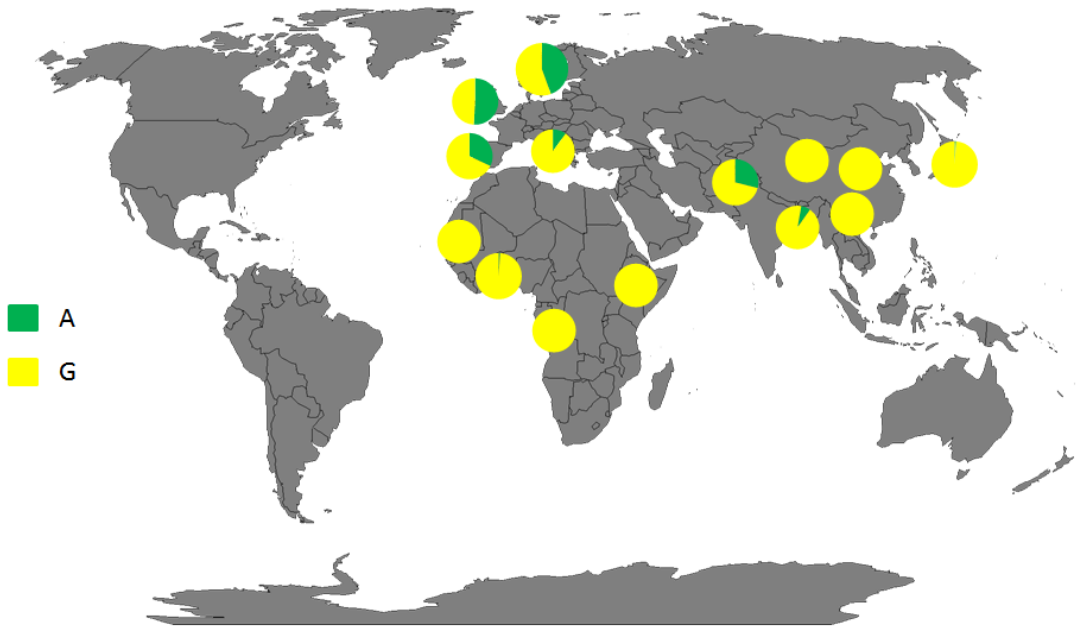
Furthermore, by focusing on the cited rs1446585 and rs35564151, the North-South gradient observed along the Italian peninsula, was still maintained at the continental level. Derived alleles for both the significant SNPs augmented their frequency proceeding towards Northern Europe, reaching a frequency of 0.74 and 0.64 (rs1446585), and of 0.5 and 0.45 (rs35564151), respectively for GBR and FIN. In fact, to be more exact, a North-

West/South-East gradient of frequency was observable across Europe for these derived alleles.

To better understand the possible adaptive events that could have shaped the particular pattern of variability for these SNPs, further genetic information has been retrieved from the *1000Genomes Project* from worldwide not admixed populations. A summary of allele frequencies at these loci are resumed in the maps below (fig. 4.2.3.2).



a)



b)

Figure 4.2.3.2: a) Allele frequency for rs1446585: derived allele (A) is coloured in blue, while the ancestral one (G) is displayed in red. b) Allele frequency for rs35564151: derived allele (A) is coloured in green, while the ancestral one (G) is displayed in yellow.

Other four African populations (i.e. LWK, GWD, YRI and ESN), two South Asian ones (i.e. BEB and PJL) and four East Asian ones (i.e. CDX, CHS, CHB and JPT) have been thus added to the previously assembled dataset.

As visible in the map (fig. 4.2.3.2), the derived allele (A) for both SNPs is almost exclusively present within the European sample, except for some South Asian populations, while all other human groups showed only the ancestral allele.

To further investigate the evolutionary histories of these two relevant SNPs, LD and haplotype analyses were performed on the whole dataset (i.e. including both the genotyped Italian samples and all the other worldwide populations) (fig. 4.2.3.3).

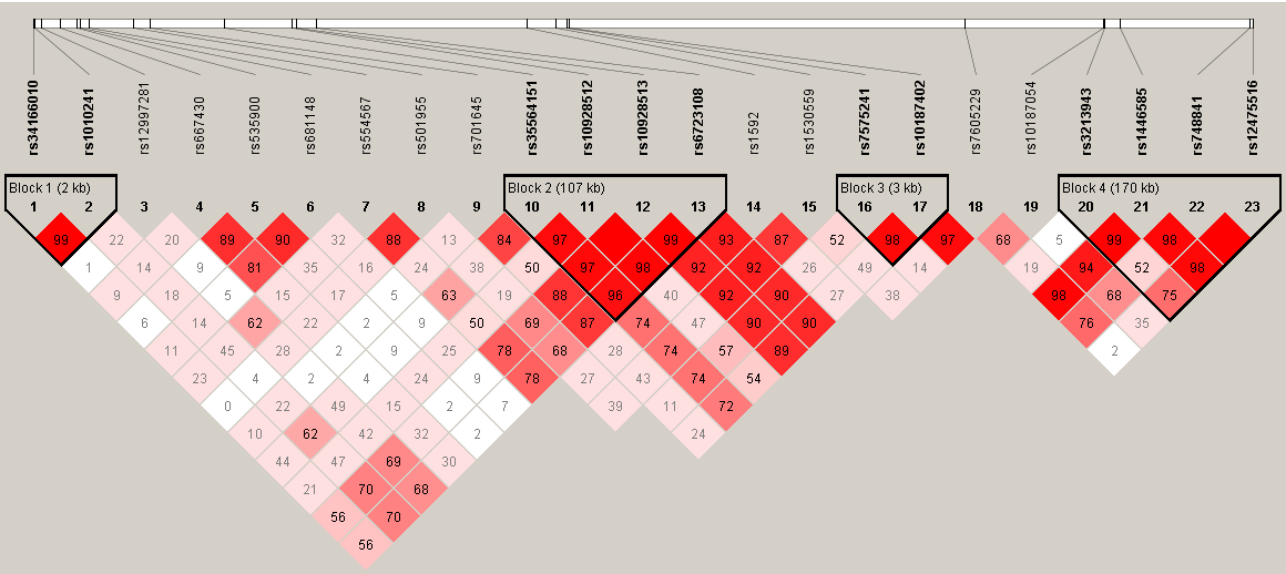


Figure 4.2.3.3: Visualization of linkage disequilibrium patterns inferred from the whole dataset.

LD analysis carried out on the whole dataset pointed out the existence of two major LD blocks, each one being composed of four SNPs, and within which are located the two investigated SNPs. One block of 107 kb included the rs35564151, which was in LD with the other SNP considered in the project (rs6723108), while the other block of 170 kb contained the rs1446585. R^2 estimation was performed to confirm that rs1446585 was in LD with all the other SNPs in the block (rs3213943, rs748841, rs12475516), as well as rs35564151 and the other SNPs of its block (rs10928512, rs10928513, rs6723108). Haplotype estimation analysis has been then carried out within these two LD blocks.

Accordingly, eight different haplotypes were observed within the block with rs35564151 and they are reported on table 4.2.3.1.

Table 4.2.3.1: Summary of the haplotypes inferred within the block where rs35564151 is located and the percentage of frequency of each haplotype within each geographic area. **AFR** = Africa; **EAS** = East Asia; **SAS** = South Asia; **NE** = Northern Europe; **SE** = Southern Europe.

Haplotype	Sequence	AFR	EAS	SAS	NE	SE	Central Italy	Northern Italy	Southern Italy
H_1	GGTT	99.8	98.0	72.8	42.4	65.4	74.6	61.1	76.0
H_2	GGTG	0	0	0.8	0.3	0.2	0	0	0
H_3	GTTT	0	1.5	7.7	9.5	13.1	11.2	13.0	14.7
H_4	GTCG	0	0	0	0.8	0.2	0.9	0.4	0
H_5	AGTT	0	0	1.1	0.5	0.2	0.4	0.4	0
H_6	ATTT	0	0.5	8.0	2.6	2.8	6.0	4.8	2.7
H_7	ATCT	0	0	0	0	0	0.4	0.4	0
H_8	ATCG	0.2	0	9.6	43.9	18.0	6.5	20.0	6.6

The most frequent haplotype (H_1) in the whole dataset carried all ancestral alleles, thus we can consider it as the "ancestral haplotype", and, in some cases (as for example in Africa or East Asia) it is almost the only haplotype present within the population. The derived allele of rs35564151 was found in haplotypes from H_5 to H_8, but only this last one showed higher frequency in several populations, especially within European groups. An high level of variability was indeed observed within this continent, with very low frequencies of this haplotype in Southern and Central Italy, while higher frequency was found within Northern Europe and Northern Italy.

The H_8 haplotype is composed of all derived alleles, Evolutionary relationships among the inferred haplotypes were further investigated by means of a median joining network (fig. 4.2.3.4), where it is possible to directly visualize the described scenario. For the second SNP, the same approach was followed to analyse worldwide haplotype distribution. In this case, haplotype reconstruction resulted in ten different haplotypes, which are resumed on table 4.2.3.2.

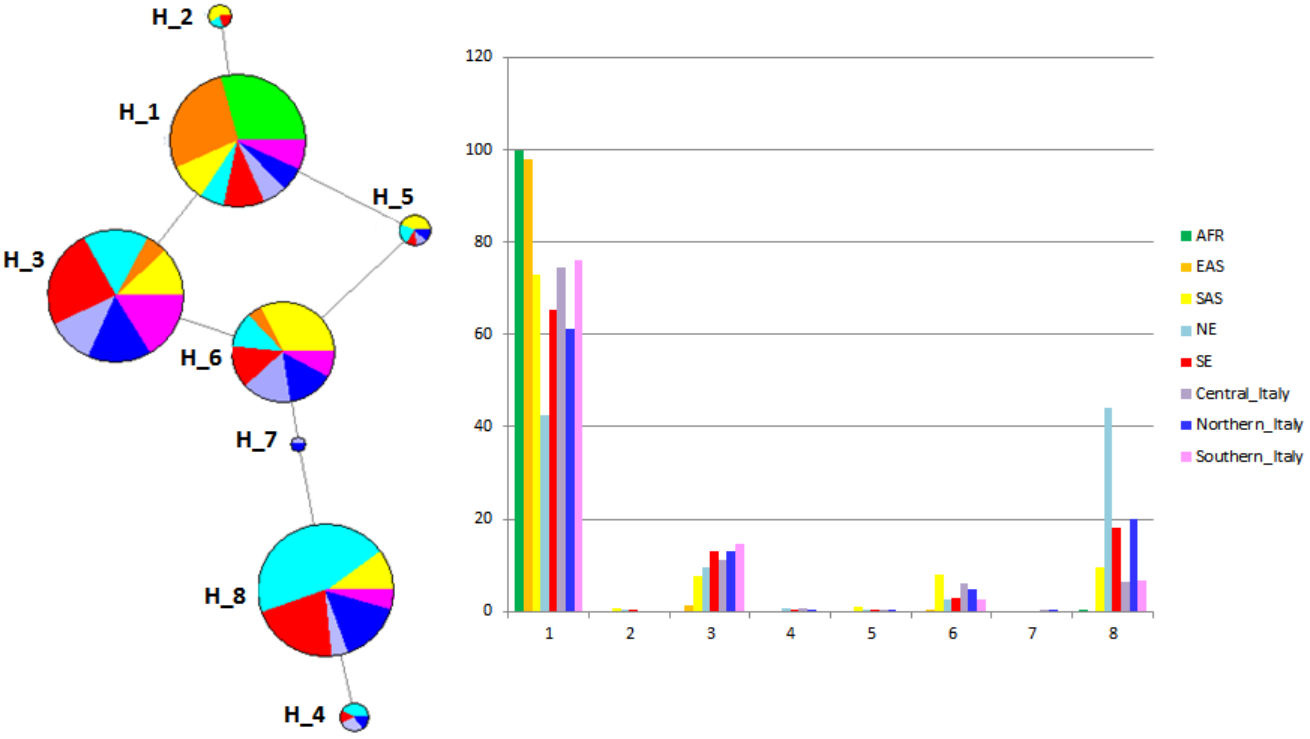


Figure 4.2.3.4: Median joining network of the inferred haplotypes. **AFR**, green; **EAS**, orange; **SAS**, yellow; **NE**, light blue; **SE**, red; **Central_Italy**, violet; **Northern_Italy**, dark blue; **Southern_Italy**, pink.

Table 4.2.3.2: Summary of the haplotypes inferred within the block where rs1446585 is located and the percentage of frequency of each haplotype within each geographic area. **AFR**=Africa; **EAS**=East Asia; **SAS**=South Asia; **NE**=Northern Europe; **SE**=Southern Europe.

Haplotype	Sequence	AFR	EAS	SAS	NE	SE	Central Italy	Northern Italy	Southern Italy
H_1	CGCC	40.4	36.3	34.1	8.9	17.1	21.6	15.2	20.5
H_2	CGCG	7.3	0.7	5.2	2.1	4.2	2.2	5.2	2.3
H_3	CGTC	10.7	12.1	7.4	0.8	9.1	12.5	8.1	14.7
H_4	CACC	0	0	0.3	0	0.2	0	0	0
H_5	CACG	0	0	0	0.8	0.2	0	0	0
H_6	CATC	0	0	21.4	66.8	30.8	13.8	32.6	12.4
H_7	AGCC	8.3	2.5	3.6	3.9	3.7	4.7	4.4	4.7
H_8	AGCG	28.2	20.6	20.9	15.0	32.9	43.1	34.1	45.3
H_9	AGTC	4.8	27.8	7.1	1.3	1.6	2.2	0.4	0
H_10	AATC	0	0	0	0.3	0	0	0	0

Among these ten haplotypes, 6 contained the ancestral allele of rs1446585. In particular, H_1 could be considered the “ancestral haplotype”. On the other hand, only four haplotypes carried the derived allele for the rs1446585 (H_4, H_5, H_6 and H_10), but only H_6 has a remarkable frequency among several populations, while all the others turned out to be private haplotypes. As visible in the second network (fig. 4.2.3.5), the ancestral haplotype H_1 (located in the middle of the plot) is that from which most part of the other haplotypes originated. It is diffused all over the world, but reaches the highest frequency especially in Africa and Asia. Interestingly, H_9 seems to be a specific East Asian haplotype that always carries the ancestral allele of rs1446585. On the other hand, H_6 is composed by more than half of Northern European population (66.8 %), followed by the South European ones and from a good portion of South Asian populations. This is the only relevant haplotype that is completely absent in both African and East Asian groups. The genotyped Italian samples carried various of the reconstructed haplotypes. Within Southern and Central Italy it is possible to find the the two major African haplotypes (H_1 and H_8), while in the North of the peninsula a good percentage (~30%) of the population carries the H_6, thus confirming its intermediate position between Northern and Southern European groups, as already seen from previous analyses.

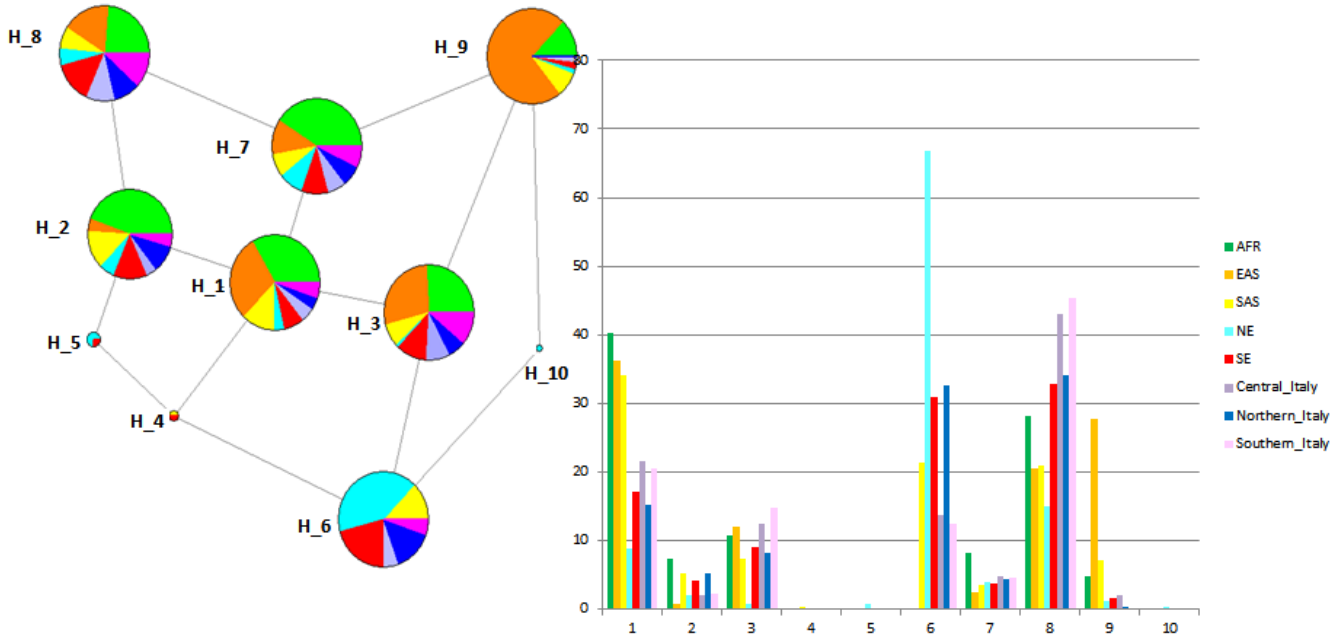


Figure 4.2.3.5: Median joining network of the inferred haplotypes for rs1446585. **AFR**, green; **EAS**, orange; **SAS**, yellow; **NE**, light blue; **SE**, red; **Central_Italy**, violet; **Northern_Italy**, dark blue; **Souther_Italy**, pink.

Project 3: Celiac disease and its influence on gut microbiota composition: an evolutionary approach to a probiotic treatment case study.

4.3.1 Analysis of Illumina experiment outputs

In this study, a cohort of 56 individuals of European ancestry has been analysed for the characterization of their faecal microbiota composition. Within these samples, 40 were subjects affected by celiac disease, further subdivided into two groups on the basis of a

probiotic/placebo treatment. Moreover, they were sampled two times during a period of three months: before the assumption of probiotic or placebo, and at the end of the treatment (i.e. at the third month). Therefore, in order to exactly identify the effect of this probiotic treatment, these two groups were further considered during the analyses as follows: Pre-Placebo; During Placebo; Pre-Treatment; During Treatment. On the other hand, 16 healthy individuals of comparable ancestry were taken into account in these analyses as a control group. According to such an experimental design, a total amount of 96 samples were sequenced in the present project.

After having performed sequencing experiment on an Illumina platform, fasta output files were processed by taking advantage from various bioinformatics tools, as described in material and methods chapter. On table 4.3.1.1, general summary of the experiment output is reported.

Table 4.3.1.1: Summary of Illumina outputs before trimming the sequences. **Number of reads**: Total number of reads for the two strands; **Means reads**: the mean number of reads per sample**; Mean Length**: mean length of the sequences; **Mean SD**: mean standard deviation of the length of sequences.

	R1	**R2**
Number of reads	14409989	14409989
Mean reads	117420	117420
Mean Length	298.9	298.7
Mean SD	6	5.20

In general, the mean quality for both forward and reverse sequences was high, as expected from Illumina platform (fig. 4.3.1.1), and with very low standard deviation.

Sequenced amplicons covered most part of the expected genomic region, thus confirming that the sequencing experiment was carried on with success.

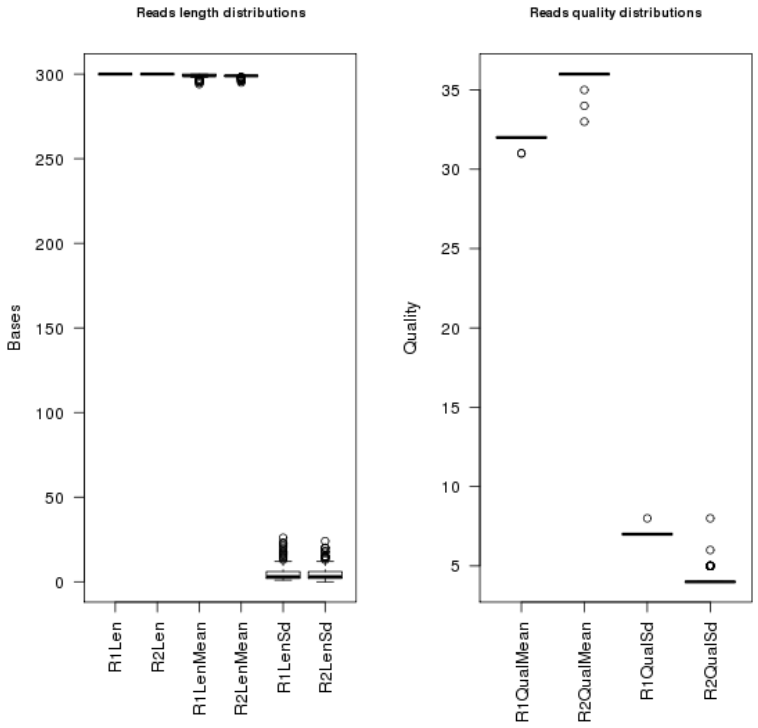


Figure 4.3.1.1 : Reads length (left) and quality (right) distributions before joining as produced directly by sequencing process. Values summarize all samples.

Only in two cases a problem during the sequencing step occurred: in sample 10A (Pre Placebo group) and sample 73 (Healthy group). In the first case, an extraordinary high number of sequences was obtained, while a very low number was observed in the second one (669 reads for both R1 and R2).

Raw data were trimmed according to several criteria: the sequences must be more than 50 bp in length and had to show at least a mean quality score higher than 20 within a window of 20 bp.

After that, trimmed sequences were joined (fig.4.3.1.2). Sample 73 was eliminated after trimming due to its very poor amount of sequences (only 69).

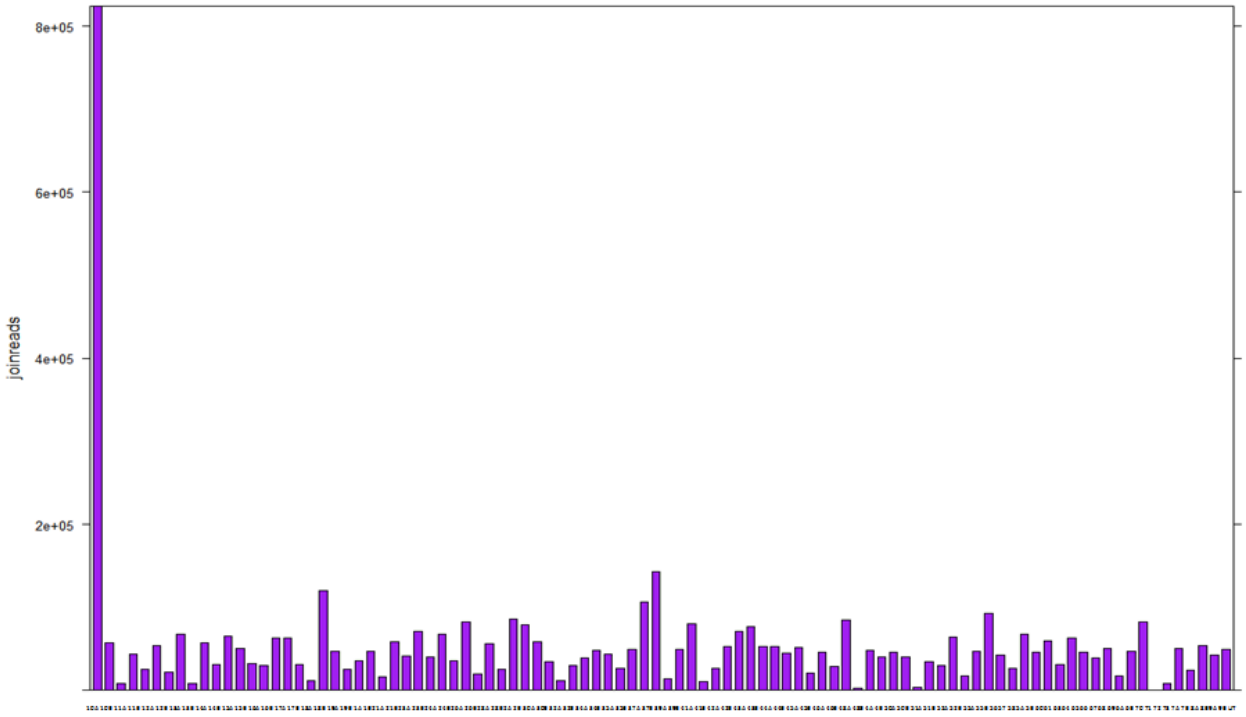


Figure 4.3.1.2: Plot showing the number of reads obtained for each sample after quality and read length control.

Then, the so joined reads, were screened for the presence of chimeric sequences using the *usearch* tool. After this step, the number of sequences for sample 10A were strongly reduced (reaching a very low number as in the case of sample 73), so that even this sample was removed from the general dataset to do not affect subsequent analyses.

A total dataset of 4,348,432 filtered high-quality joined reads (excluding the undetermined sequences) was thus generated, with an average of ~46,259 sequences per sample.

4.3.2 Gut microbiota analysis

By considering all microbial phyla present at least at more than 1% in one sample, it is possible to estimate that more than 98% of the sequences in all the samples studied were found to belong to the most populated bacterial phyla, namely Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria and Verrucomicrobia. Interestingly, also one Archaea phylum (i.e. Euryarchaeota) was present in more than 1% in some samples (fig. 4.3.2.1).

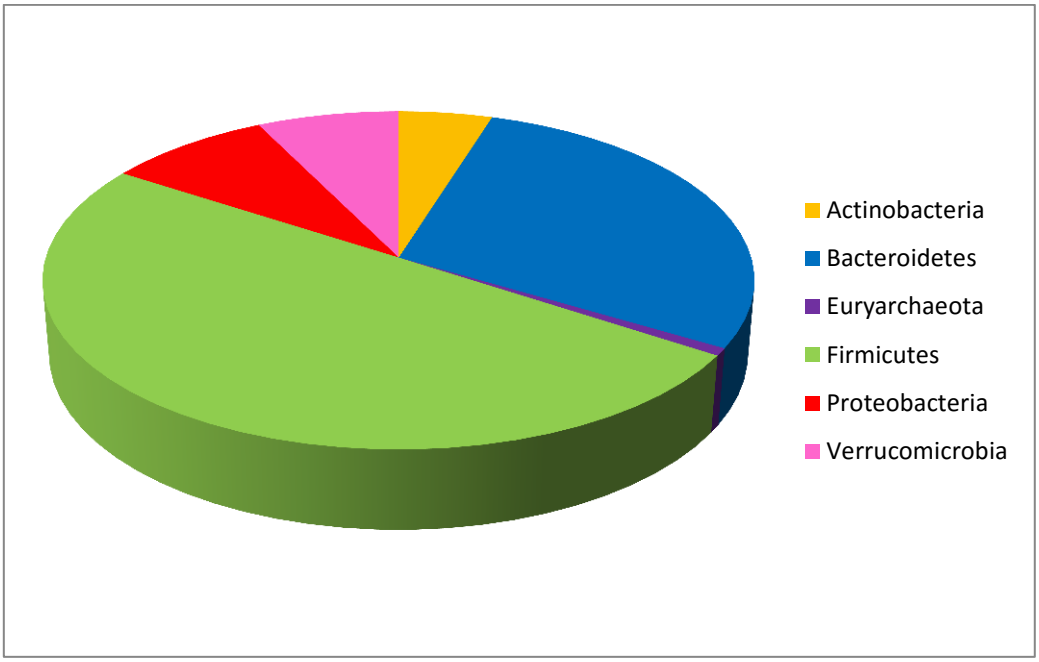


Figure 4.3.2.1: Phyla distribution among all the samples sequenced in this project. Firmicutes, green; Bacteroidetes, blue; Euryachaeota, purple; Actinobacteria, yellow; Proteobacteria, red; Verrucomicrobia, pink.

Anyway, the distribution of relative abundances of these phyla among the studied groups was not exactly the same, as clearly showed in figure 4.3.2.2. In fact, the heatmap picture highlighted that several phyla showed different relative abundance within each group. For instance, the Firmicutes phylum was characterized by really higher percentage in the

healthy group (accounting for ~60-70% of the total microbial community). On the other hand, Bacteroidetes seem to have a different trend, with higher frequency in all celiac groups and a lower percentage within the healthy one. Moreover, it is important to note that cluster analysis pointed out that the During Treatment group occupied an intermediate position between the healthy group and the rest of celiac individuals, being thus considered as an outlier with respect to the other disease clusters.

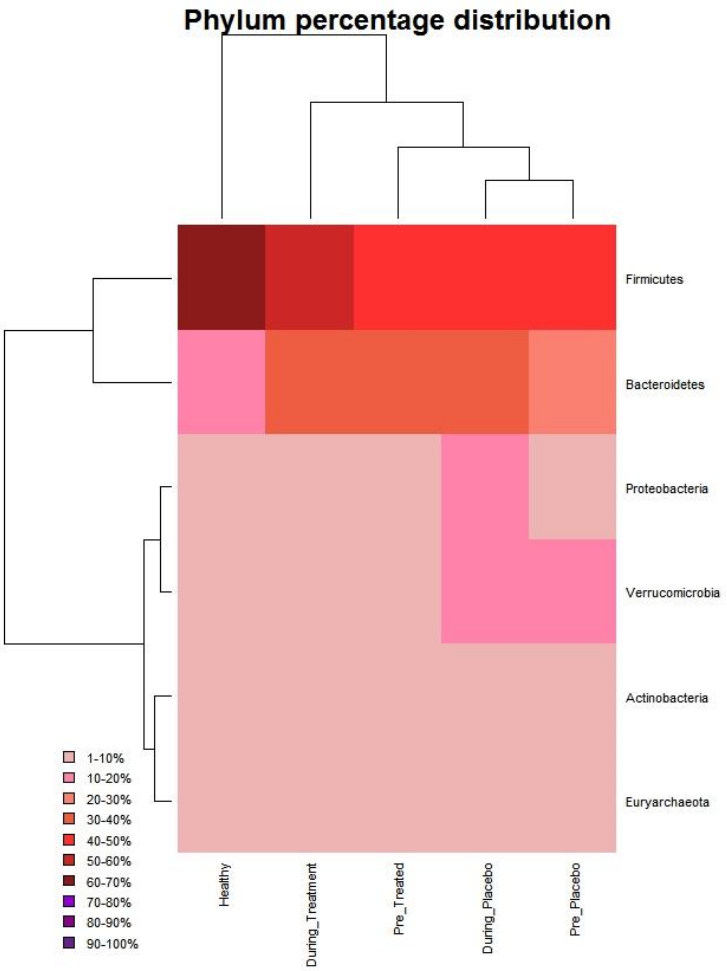


Figure 4.3.2.2: Heatmap analysis indicating the percentage of relative abundance of each phylum within the different groups.

Following the data normalization procedure and assignation of statistical significance described in the material and method chapter, several comparisons between pair of groups were performed in order to identify which phyla distinguish the microbiota of healthy

people from that of people affected with celiac disease. Furthermore, these analyses would be useful for the detection of possible signals of treatment effect on the general microbiota composition.

A first comparison was made between pre-treated, during treatment and healthy groups (fig. 4.3.2.3).

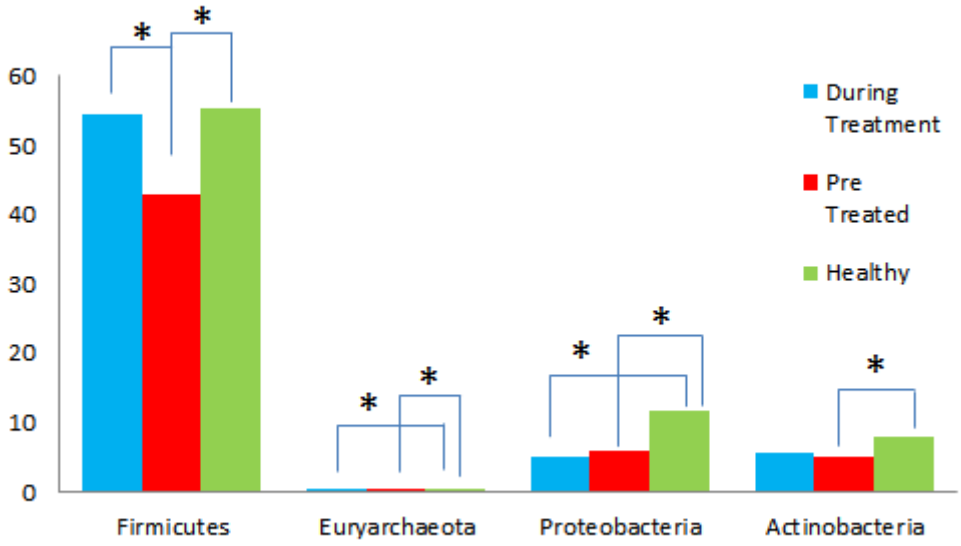


Figure 4.3.2.3: Barplots of several significant phyla standing out from comparison between pre-treated, during treatment and healthy groups. The asterisk indicates for which comparison a significant p-value was obtained ($p < 0.01$).

At the phyla level, the Firmicutes showed a significant difference in the relative abundance in both comparisons between pre-treated/during treatment and between pre-treated/healthy groups. Particularly, this phylum showed a lower abundance in the pre-treated group with respect to the other two, since no statistical significance was obtained from the exact test performed between during treatment and healthy groups. For the Proteobacteria and the only Archaea phylum, a higher frequency was observed within the

healthy group. Furthermore, the Actinobacteria phylum showed a significant p-value only when the percentage of pre-treated and healthy samples were compared. Accordingly, it seems to have increased its percentage of abundance during the treatment, even if it did not change so far from the pre-treated value.

These results are in accordance with those obtained by considering also the other two placebo groups (fig. 4.3.2.4).

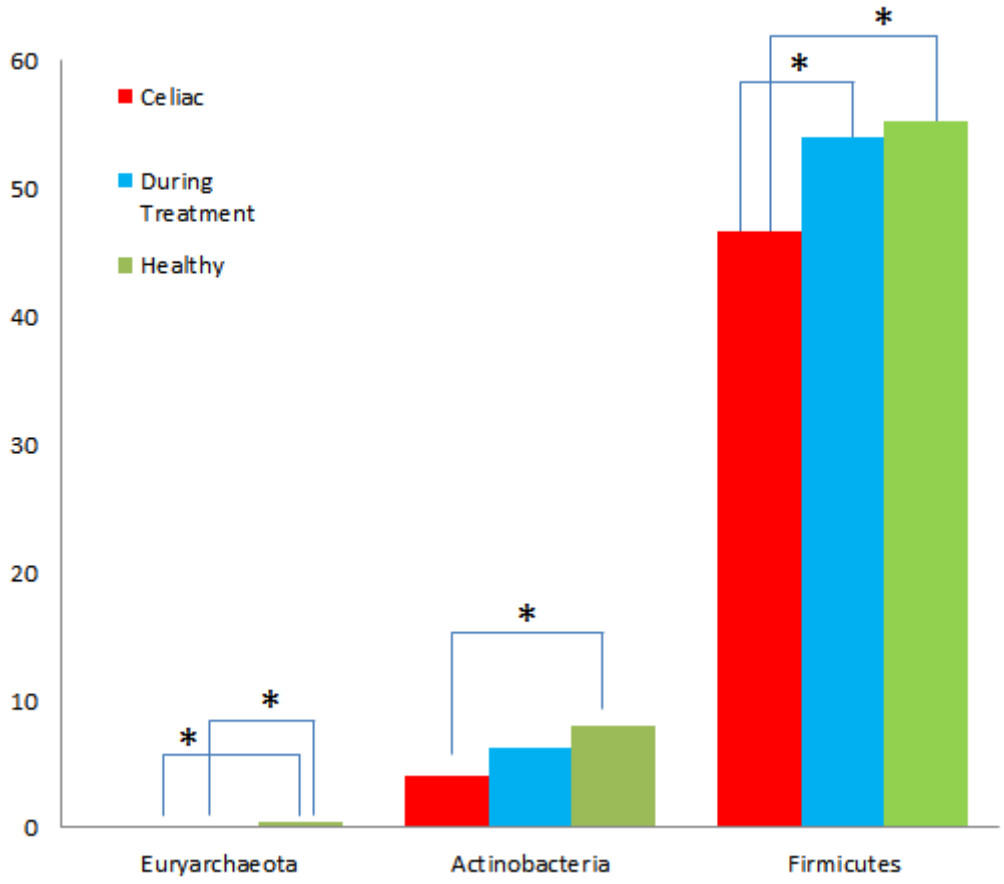


Figure 4.3.2.4: Replication of the exact test using all the groups of people affected by celiac disease in comparison with the celiac treated group and the healthy group.

In fact, when all the placebo and the pre-treated groups (that we can mention as “celiac group”) are compared against the celiac treated group and the healthy one, most part of the same phyla already observed in the previous analyses were detected (fig.4.3.2.5). More in details, the celiac group showed a lower percentage of Firmicutes, this same bacterial phylum is also the only one that showed a statistical significance difference between all celiac individuals and the treated group. Moreover, as seen before, the healthy group had a major presence of both Actinobacteria and Euryarchaeota than the rest of the analysed samples. It seems to be confirmed that Actinobacteria augmented a bit their percentage during the treatment, but without reaching any statistical significance. Interestingly, with respect to the first comparison, here we did not find anymore the group of Proteobacteria, plausibly due to the high variability that this phylum showed in each group studied.

Repeating this approach at the bacterial family level, it is possible to get deeper inside into the differentiation between groups and to get more information about the effectiveness of the probiotic treatment.

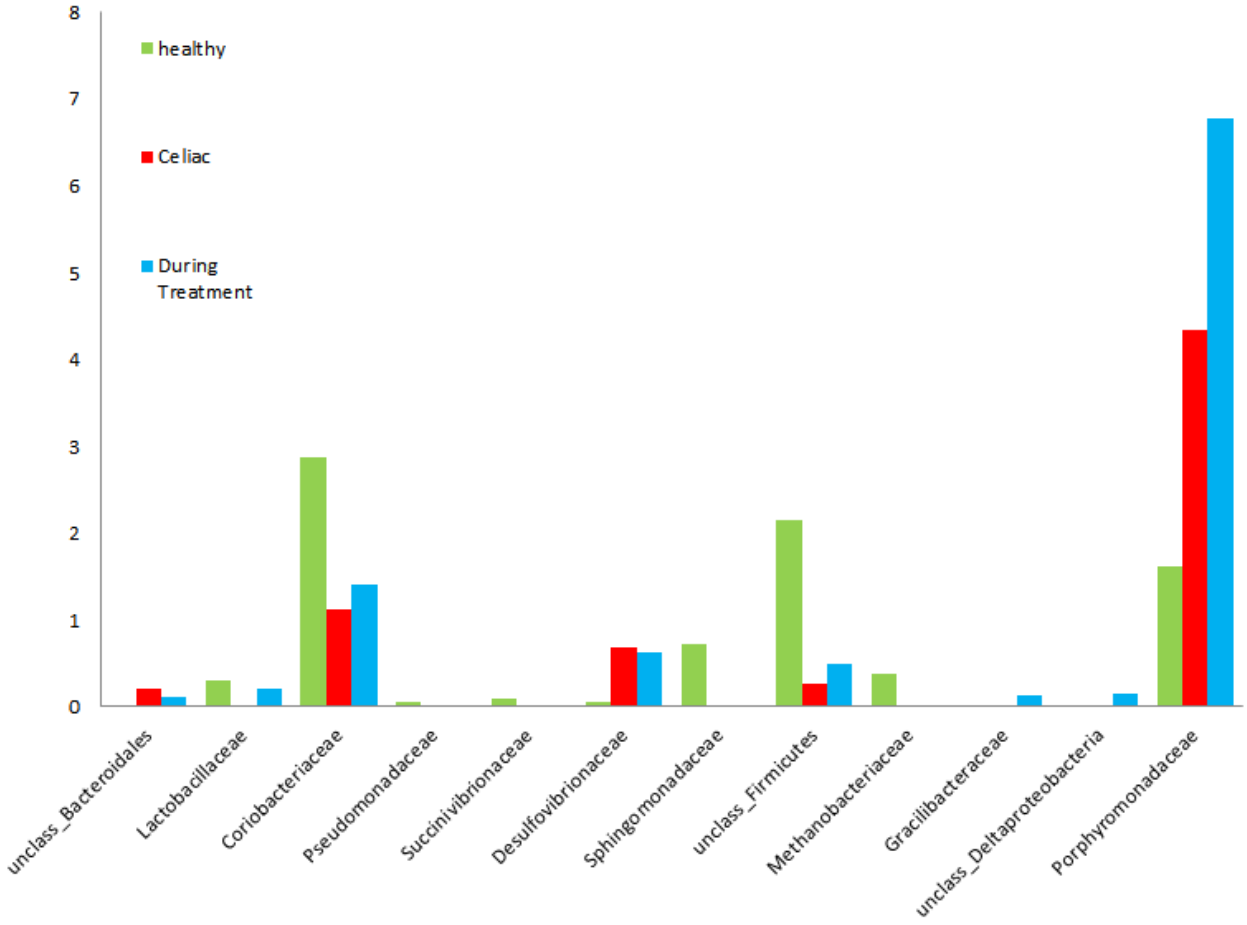


Figure 4.3.2.5: Barplots of the bacterial families that showed a statistical significant difference when celiac, during treatment and healthy groups were compared.

By comparing bacterial family abundance among groups, a clearest description of the pattern observed at the phylum level was obtained. Particularly, within the Firmicutes phylum, three families have been detected to be really poorly represented in the celiac groups, showing instead a higher level in both the treated and the healthy groups. These families are: *Lactobacillaceae*, *Gracilibacteraceae* and general unclassified Firmicutes. Following the same trend, other bacterial families were found to be more represented in the healthy microbiota, such as *Coriobacteriaceae* (belonging to Actinobacteria phylum) and three Proteobacteria families (*Pseudomonadaceae*, *Succinivibrionaceae*, and *Sphingomonadaceae*). On the other hand, two additional families, such as

Desulfovibrionaceae and *Deltaproteobacteria*, seem to be present within both the celiac group and the celiac treated one.

Interestingly this family analysis enabled to the point out the archaea *Methanobacteriaceae* as a family almost exclusive of the healthy subjects. Finally, although at the phyla level the Bacteroidetes did not get any significant value, at the family level it occurred for the *Phorphyromonadaceae*, which reached really high percentage within the treated group with respect to all the remaining ones, and for a general unclassified Bacteroidetes that, as expected from results of the heatmap analysis, showed a major percentage within the celiac sample and a bit lower value in the treated microbiota, while being almost absent in healthy individuals.

A more accurate comparison also enabled to detect exactly which bacterial families have acted as the probiotic administrated to the treated group. In fact, within the Firmicutes phyla, the *Lactobacillaceae* were found to reach almost the same percentage of abundance that they show in the healthy group also within the treated cohort. At the same time, also the *Gracilibacteriaceae*, as well as, the unclassified Deltaproteobacteria, jointly reached a higher percentage in the microbiota of the treated group (fig. 4.3.2.6).

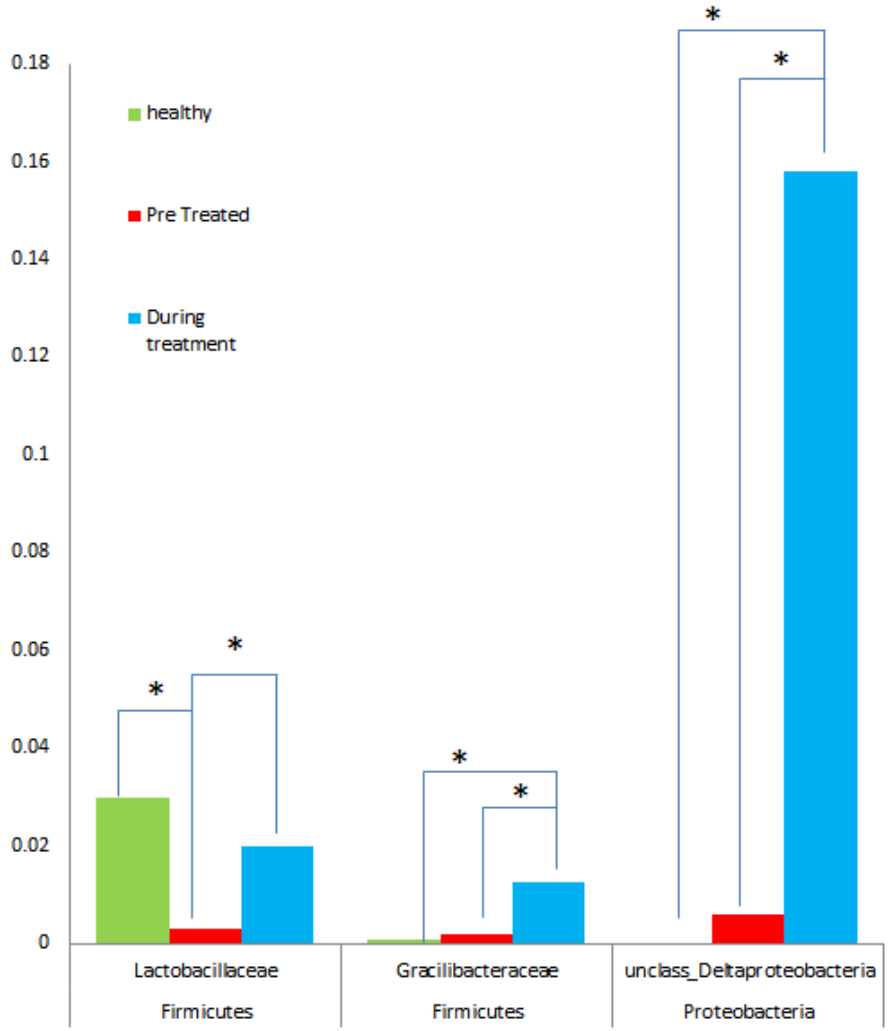


Figure 4.3.2.6: Comparison between pre-treated, treated and healthy groups to detect the probiotic effect on the examined samples.

Then, the overall picture that emerged from the analysis of bacterial families on the basis of the exact test was that the microbiota composition of individuals that received the probiotic treatment assumed some aspects of the typical microbial composition observable in the healthy group. To confirm this finding, several cluster analyses have been performed. As first, a hierarchical method was used to assess the relationships between different groups at the family level (fig. 4.3.2.7).

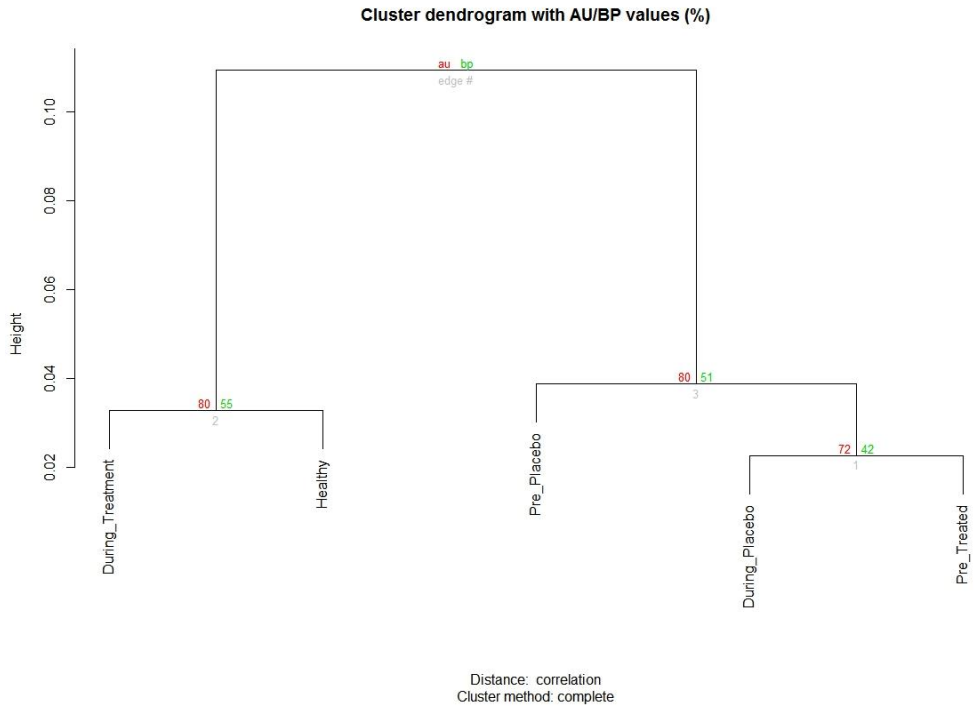


Figure 4.3.2.7: Cluster analysis performed with a hierarchical method on microbial composition at the family level. AU value is indicated in red, while BP value is reported in green.

The AU indicates the approximately unbiased p-value, which is computed by multiscale bootstrap resampling, as a better approximation to unbiased p-value than BP value computed by normal bootstrap resampling. By considering the p-value indicated in the plot, as suspected, the group of treated individuals clustered on the same branch of the healthy group, with a AU p-value of 0.80. On the other cluster were instead unified all the other celiac individuals reported in this study.

PCA analysis was performed as well to further corroborate these results (fig. 4.3.2.8). Also in this case, the treated group was found to be in proximity of the healthy one along PC1, and to be more precise, it lied in the middle between the pre-treated and the healthy groups. This cluster analysis, a part from giving confirmation of the first dendrogram, showed high microbial biodiversity within each group, highlighting how each microbiota has its own characteristics.

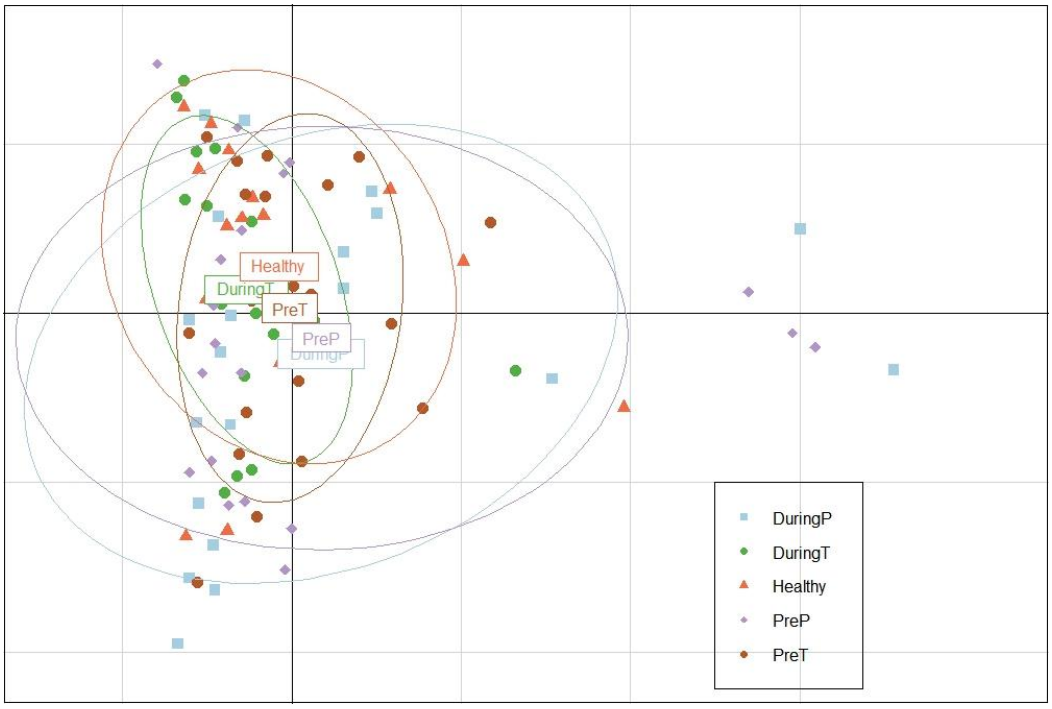


Figure 4.3.2.8: PCA analysis performed on bacterial families. **DuringP** = During placebo; **DuringT** = During treatment; **PreP** = pre-placebo; **PreT** = pre-treated.

On the light of these results, a further analysis was performed to estimate the alpha diversity within each group (fig. 4.3.2.9). The observed raw biodiversity within the healthy sample seems to be little higher than those of all the other groups considered in the project. Looking at the Chao1 index, the values appeared to be more or less similar, except for the one of the placebo group for which the higher value was founded only in one sample and, anyway, it was characterized by high standard deviation with respect to all the other groups. Even the Shannon index indicated similar trends among all groups, with a mean value of ~2. This similarity among groups is further confirmed by the application of Wilcoxon test on these indices, which indicated the totally absence of significant differences (data not showed).

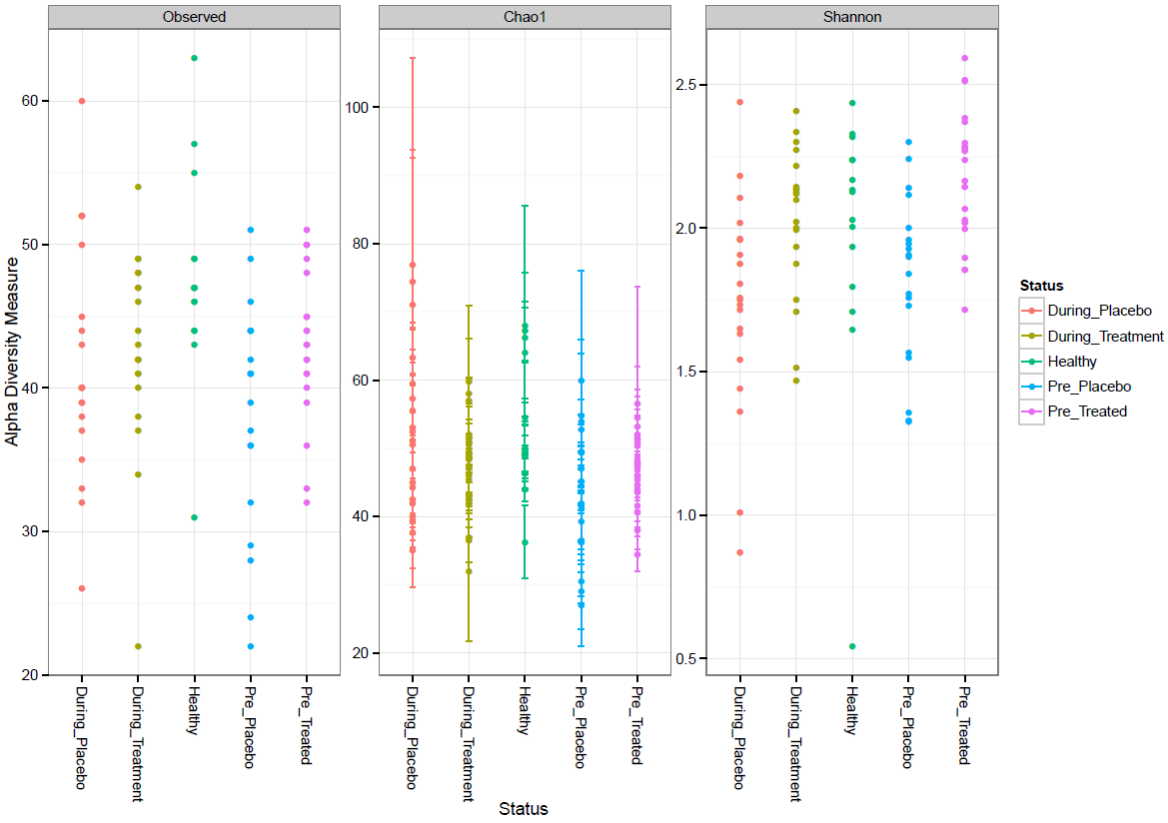


Figure 4.3.2.9: Computation of alpha diversity most common indices for each group studied.

Chapter 5

Discussion and Concluding remarks

Project 1: Detecting signals of natural selection in European populations on genes involved in nutritional and thermoregulation processes.

5.1 Discussion of Project 1 results.

After the exit from Africa of their ancestors, population of anatomically modern humans have experienced strong evolutionary shifts in terms of occupied habitats, exploited food resources, experienced climate conditions, as well as demography and pathogen exposure. These events have certainly left sizeable marks on the human genome, as suggested by an increasing number of studies (Voight et al., 2006; Barreiro et al., 2008; Hernandez et al., 2011; Stewart & Stringer, 2012). Nonetheless, the impact of natural selection on the processes especially related to human nutrition and energy expenditure has been to date poorly understood (Heyer & Quintana-Murci, 2009; Hancock et al., 2011; Ye & Gu, 2011; Breslin, 2012).

This research project is thus aimed at detecting possible genetic signatures in the gene pool of European populations, that could be associated to the above-mentioned major evolutionary events, in the attempt to understand how relatively recent ecological and cultural shifts related to nutritional habits and environmental changes could have shaped current patterns of genetic variation and human biodiversity within the European continent.

Accordingly, starting with an extensive literature survey, a total of 21 genes differently implicated in several nutritional and thermoregulation processes have been selected and the corresponding genetic data have been retrieved from the *1000Genomes Project* database for three European non-admixed populations from different latitudes (i.e. TSI, GBR, FIN).

In order to explore geographical patterns of genetic variation at the selected genes, population structure was investigated by means of DAPC, first by considering the whole dataset, and then via a gene-by-gene approach. Accordingly, appreciable intra-continental variability associated to this subset of genes, was observed, with Northern European populations showing closer genetic affinities with each other than with the Southern European group. Even when the analysis was repeated without giving any indication on population identity, FIN and GBR tended to cluster together, while most TSI individuals were arranged on a different cluster. To dig up if the observed structure was due mainly to one or more genes, the same analysis was repeated gene-by-gene. While almost all the explored genes showed a substantial monomorphic profile, *PRDM16* turned out to be the unique locus characterized by considerable heterogeneity among the considered European populations. Accordingly, genetic structure pointed out by global DAPC analysis seems to be principally related to variation at this gene. That being so, further analyses have focused on this particular genomic region.

To further investigate the underlying causes of this particular genetic pattern, it would be useful to recapitulate what is known about this specific locus. The *PRDM16* gene encodes for a 140 kDa protein and is characterized by a positive regulatory (PR) domain similar to a SET domain, in addition to 10 zinc finger domains, which modulate the transcriptional activity of the brown fat (BAT) thermogenic genes programme. *PRDM16* has also a major role in transcription and regulation of adipocyte differentiation (Frühbeck et al.,

2009). In fact, it is definitely a master regulator of the processes of BAT adipogenesis, differentiation and function, as well as of the so-called "browing" process of white adipose tissue (WAT). In the last years, a large number of studies have indeed confuted the old binomial view concerning the existence of only two adipose tissues, distinct from both histological and functional characteristics, by the discovery of a new kind of tissue named "beige" or "brite" (i.e. the contraction from "brown" into "white") (Wu et al., 2012). This last type of adipocytes constitutes a link between the two well-known adipose tissues, although its role in the physiological metabolism is still far to be elucidated.

The pattern of population differentiation observed for *PRDM16* within Europe could basically result from demographic processes and/or the action of natural selection, even if the absence of population structure at the other 20 examined loci suggests that the latter evolutionary force might have substantially contributed to shape variation at *PRDM16*. To corroborate this hypothesis, several neutrality tests have been applied on data from the whole *PRDM16* genomic interval. After Bonferroni correction for multiple tests, four significant genomic windows were obtained as concerns the Fu & Li' D test (two for FIN, one for GBR and one for TSI) and just one (the same previously observed for GBR) as concerns the Fu & Li' F test. For all these windows, a positive statistics was calculated, which indicates the possibility that balancing selection has acted on these regions. Network analysis conducted by considering the window significant in GBR seems to corroborate this hypothesis. Moreover, within such window, a Zinc finger 1 domain of the *PRDM16* gene is present, corresponding to amino acids 224-454 (Kajimura et al., 2008). This is a DNA binding domain that is mainly implicated in the activation of *PRDM16* thermogenic function by specifically binding the *PGC-1α* and the *PGC-1β* genes and thus increasing their transcriptional activity (Puigserver et al., 1998). These two proteins regulate both the development of brown fat cells identity (Seale et al., 2007;Seale et al.,

2008) and the thermogenic gene programme of BAT adipocytes by promoting BAT-specific gene induction (Hondares et al., 2011). In fact, through the activation of *PGC-1α*, which then co-activates *PPARα* and *PPARγ*, *PRDM16* actually regulates both mitochondrial biogenesis and thermogenesis processes (Barbera et al., 2001). Thus, according to what already described about the *PRDM16* function, it is highly plausible that selective pressures responsible for such putative events of balancing selection were related to a climate-driven adaptive scenario.

Further information about the evolutionary processes that could have shaped the observed genetic landscape came out from the application of the iHS test on the examined dataset. This analysis pointed out the presence of two SNPs for FIN (rs112682827 and rs144090205) and one for TSI (rs2817126) showing significant p-values, even after Bonferroni correction. This could be explained by positive selection having acting on these populations in relatively recent times. Both rs2817126 and rs112682827 were characterized by positive iHS values, suggesting that the ancestral allele was subjected to positive selection, while rs144090205 showed a strong negative iHS value, indicating that the derived allele could have exerted an adaptive role. Interestingly, these SNPs are distributed in a restricted genomic range (from position 3,082,126 to 3,084,761, GRCh37), in which a lot of other SNPs showed high absolute, even if not significant after Bonferroni correction, iHS values. This could indicate that probably the entire considered genomic interval has been subjected to recent positive selection in the European populations. Moreover, these SNPs did not show any appreciable LD with each other, and they are all intronic substitution related to a non-coding transcript variant (ensemble code: ENST00000607632). In particular, the two SNPs significant in FIN are in a *PRDM16* regulatory region associated with the gene promoter (ensemble code for regulatory feature: ENSR00001516847).

To get deeper insights into *PRDM16* recent evolutionary history and to better understand which selective pressures could have mainly triggered such adaptive events, age estimation for the putative adaptive haplotypes was obtained. Analysis of the extension of haplotype homozigosity (EHH) around the core SNPs first pointed out interesting differences between the examined populations. Particularly, the two SNPs positively selected in FIN showed significantly smaller EHH extension than the SNP identified in TSI, indicating that positive selection have plausibly acted in different time periods in the two populations. More specifically, the presence of larger haplotype homozigosity around the TSI significant SNP probably means that such group have experienced a very recent event of selection, while FIN were subjected to older selective pressures thus enabling recombination to erode LD around their adaptive SNPs. The algorithm developed by Voight et al. (2006) was then applied to perform a formal age estimation and confirmed the EHH results by pointing to very recent selective sweeps occurred in the populations considered in this study. In the line with the specificity of the iHS test to detect recent signs of positive selection, it is plausible that selective pressures have started to act long time before the obtained dating, but that their intensity strongly increased in more recent time, mainly determining the signatures detected by the applied test. In fact, it is interesting to note that the obtained age estimations reflect the timing of different major climate changes occurred in Europe from the Bronze Age until more recent historical periods. These shifts were linked to natural circumstances that seem to have strongly influenced European populations' demography finally leading to the end of advanced civilizations or to intense migrations (Li et al., 2011; Büntgen et al., 2011; Miller et al., 2012; Kaniewski et al., 2013). Particularly, recent studies have recorded that exactly the Bronze Age was a period of great changes in climate conditions, which were probably causes of mass migration from the Steppe to Central/Northern Europe and Asia, as

suggested by both genetic and archaeological records (Anthony, 2009; Itan et al., 2009; Cassidy et al., 20016). These migrations seem to have deeply influenced the Neolithic European genetic background, leading to a distinction of northern European populations from southern ones as regards several genomic traits (Allentoft et al., 2015). Therefore, it is plausible to hypothesize that these major climate changes have exerted strong selective pressures on the genetic background of European human populations, especially as concerns loci involved in thermoregulation processes.

In conclusion, in accordance with other studies that demonstrated how European populations have been exposed to strong selective pressures even within the last 10,000 years (Sabeti et a., 2002; Voight et al., 2006), different footprints of natural selection were detected on the *PRDM16* locus. These signals probably reflect local adaptation occurred with different intensities among populations and also in different times. Results from both Fu & Li' D, F and iHS test suggest that this gene has been subjected to multiple selective events during the history of European populations. Moreover, populations-specific footprints of selection were also found to affect different regions of the gene.

As commented before, *PRDM16* is mainly expressed in both BAT and beige tissue and controls the cell fate switch between skeletal myoblast and brown fat cells, also enhancing the thermoregulation process. Brown and beige adipose tissues share many biochemical characteristics and both are activated in response to cold exposure and in energy balance activity (Ouellet et al., 2012). Several studies highlighted how much climate changes could affect BAT activity, as for example how reduced seasonal cold exposure could influence energy balance and obesity risk (Johnson et al., 2011), as well as BAT amount around arteries to assure survival at extreme cold temperature (Huttunen et al., 1981).

Furthermore, *PRDM16* serves as positive regulator of the whole pathway expressed within the BAT and beige tissues, by mediating enviroment-incuced beige adipocyte development and thus testifying its major role in climate adaptation processes (Ohno et al., 2012; Ohno et al., 2013; Cohen et al., 2014). That being so and on the light of the obtained results, an evolutionary balance acting on BAT activity could be helpful both to face cold temperatures and in the process of energy storage necessary to survive during eventual periods of famine. On the other hand, the identified candidate SNPs that plausibly undergone recent positive selection could be barely associated to unique events, but more probably to multiple shifts in climate conditions and nutritional habits occurred over the last 5,000 years, which could have contributed to shape the current picture of European population structure at *PRDM16*.

Project 2: Analysis of Italian genetic variability at three genes associated to metabolic functions.

5.2 Discussion of Project 2 Results

The Italian peninsula represents an interesting case study under an anthropological and evolutionary point of view due to its particular geographical collocation exactly in the middle of the Mediterranean sea. This position assigned a central importance to the peninsula, placing it at the heart of migration movements during both prehistoric and historic periods.

Since the Upper Palaeolithic, the south of the country was reached by maritime migration movements of people belonging to the Aurignacian culture from the Balkans (Benazzi, 2011). During the Last Glacial Maximum most part of modern northern Europe was

completely covered by ice (Strauss, 2015) and characterized by rigid climate conditions, while Central and Southern Italy were probably under warmest conditions making these areas a possible refuge for several human groups (Pala et al, 2009; Tallavaara et al., 2015, De Fanti et al, 2015a). After that, first Neolithic agricultural populations reached the peninsula in two distinct time, first from the Aegean Sea (Tresset and Vigne, 2011) and by occupying the southern coastlines, then with people belonging to the Impressa and Cadial culture. These populations have brought agricultural techniques in Italy, probably long time before the rest of the continent (Pessina and Tinè, 2008). Subsequently, other migration flows of people belonging to the Linearbandkeramik culture reached Northern Italy from a Central Europe corridor (De Fanti, 2015b). For instance, recent studies showed how the lactase persistence phenotype has been brought to Italy probably through these population movements(De Fanti, 2015b).

The archaeological record testifies also major cultural changes in Europe and Asia after the Neolithic period (Harrison and Heyd, 2007), which are considered to be at the origin of the most renowned proto-historic Italic populations (Boattini, 2013). Finally, further migration processes have interested the Italian peninsula, starting from the establishment of Phoenician and Greek colonies, through the Roman Empire and the more recent Medieval migrations of northern European and Arab populations (Sarno et al, 2014). All these events are supposed to have determined a consistent gene flow through time, affecting enormously the genetic landscape of the Italian population, which is consequently highly heterogeneous (Capocasa et al., 2014). In addition to these complex demographic events, the Italian peninsula is also characterized by a complex environmental landscape that entails different climate conditions (KleinTank et al., 2002; Rivas-Martinez et al., 2004), which have potentially influenced the processes of local

adaptation to specific environments, thus contributing to define the high genetic variability observed within the Italian population.

In order to provide further information about how different climatic and environmental conditions, as well as different demographic histories, have shaped the genetics of Italian people, the present study focused its attention on three genes (*TMEM163*, *RAB3GAP1*, *R3HDM1*) that have been selected according to their high variability within the Italian population, as emerged from a previous study conducted by our research group (Sazzini et al., under review). Therefore, in order to validate results obtained by previous genome-wide analyses, and to provide new data useful for a deepen knowledge of patterns of genetic variation associated to these genes at a micro-geographic level, several SNPs belonging to these loci have been genotyped on 380 healthy Italian samples. Considering that this study was focused on Italy, a first screening of the variability level associated with these genes was performed on genetic data of European populations available from the *1000Genomes Project*. Thus, after filtering for their heterozigosity levels, F_{st} and LD values, 23 SNPs were selected as the most informative for a population genetic study. These cited genes exert their role in different metabolic processes. *RAB3GAP1* is involved in the presynaptic neurotransmitter release and contributes to the maintenance of neurotransmission; *R3HDM1* seems to be involved in the interactions with single-strained nucleic acids, which is probably linked to mRNA stabilization and whose several variants are found to be associated with triglycerides levels (Teslovich et al., 2010, Shin et al., 2014); finally, *TMEM163* is probably linked to modulation of zinc cellular level (Cuajungco et al., 2014).

Globally, the genotyping experiment was carried out with success, only one sample showed a large number of not typed SNPs, maybe due to its DNA deterioration, and thus it was excluded from further analyses.

In order to depict patterns of variability of the selected SNPs within the Italian samples, population structure analyses have been performed. Both DAPC and PCA pointed out a similar trend of distribution for these loci, with groups from Central and Southern Italy being highly homogenous between themselves and also internally, while Northern ones showed a discrepancy from the rest of the peninsula and also higher internal heterogeneity. The fact that this distribution is confirmed by both structure analyses corroborate the results obtained, furthermore high percentages of variation were explained by the most informative PCs (PC1 30.45% and PC2 12.83%). Moreover, thought the visualization of the PCA's loadings it was also possible to observe that differentiation of Northern Italians with respect to the other groups was principally driven by few SNPs.

These same variants were also pointed out by F_{st} analysis that described a trend of differentiation between Northern and Southern Italian subpopulations similar to that previously described by other analyses. Almost the totality of the analysed SNPs reached higher F_{st} values when Northern and Southern groups were compared than when Northern and Central ones were. On the other hand, between Central and Southern populations almost exclusively negative values were obtained, indicating that the variability within each group is higher than that between the two groups analysed.

When the Italian dataset was compared to the rest of Europe, people from the North of the peninsula maintained their genetic distance from groups from the Centre and the South of Italy, which still exhibited close genetic relation. Particularly, the northern samples were collocated in the middle of the PCA space thus close to both Northern and Southern European populations, as well as to IBS. The TSI group instead clustered perfectly with samples from South and Centre of Italy, confirming that it could not be considered as representative of the whole Italian population.

By considering results from the Fishers' exact test applied on allele frequencies, only two SNPs already indicated by Fst analysis showed statistical significance values: rs1446585 and rs35564151, which are located on *R3HDM1* and *TMEM163*, respectively. Although these two SNPs follow a similar pattern of distribution among the typed samples, very low LD level was proved to exist between them, thus suggesting that they could represent two independent signals of differentiation between Northern and Southern Italians. In fact, in both cases the derived allele turned out to be more represented in the North of the country (0.33 and 0.26, respectively), being at very low frequency or almost absent in some southern provinces. Moreover, it is possible to observe a decreasing North-South gradient for these derived alleles along the peninsula. Comparison with the other European populations testifies that this gradient of distribution is evident also in the rest of Europe, where the derived alleles are highly represented in GBR (0.74 and 0.64, respectively) and FIN (0.5 and 0.45, respectively). Interestingly, when the focus was moved at a worldwide level to further investigate patterns of variability for these two SNPs, it is possible to discover that the derived alleles seem to be really Europeans-specific. All other African and Asian populations available from the *1000Genomes Project*, showed almost exclusively the ancestral allele for both SNPs. The only exception was represented by South Asian groups, and in particular by Punjab population from Pakistan. This is probably due to the several human migrations occurred across Eurasia during the Bronze ages and recently confirmed by many studies (Allentoft et al., 2015). For long time, anthropologists and archaeologists have debated if the major cultural changes that occurred during this period were caused by human migrations or by simple ideas circulation across countries. Recent findings highlighted how the most important events happened during the Bronze Ages are actually related to multiple human migration events (Mallory et al., 1987; Lalueza-Fox et al., 2004; Itan et al., 2009; Cassidy et al.,

20016). Particularly, a first migration was performed by people belonging to the Yamnaya/Afanasievo culture that was directed towards both Central/Northern Europe and Central Asia (fig. Ra). Thus, the Bronze Age Northern European population was composed of a mixture of the first hunter-gatherers, Neolithic farmers and of this "Caucasian" recently arrived population group. The result of this genetic admixture was the birth of the Corded Ware culture. It also seems that the Southern European populations were less affected by this Yamnaya genetic pool, as demonstrated from the totally absent "Caucasian" genetic landscape in Remedello sample (fig.5.2.1).

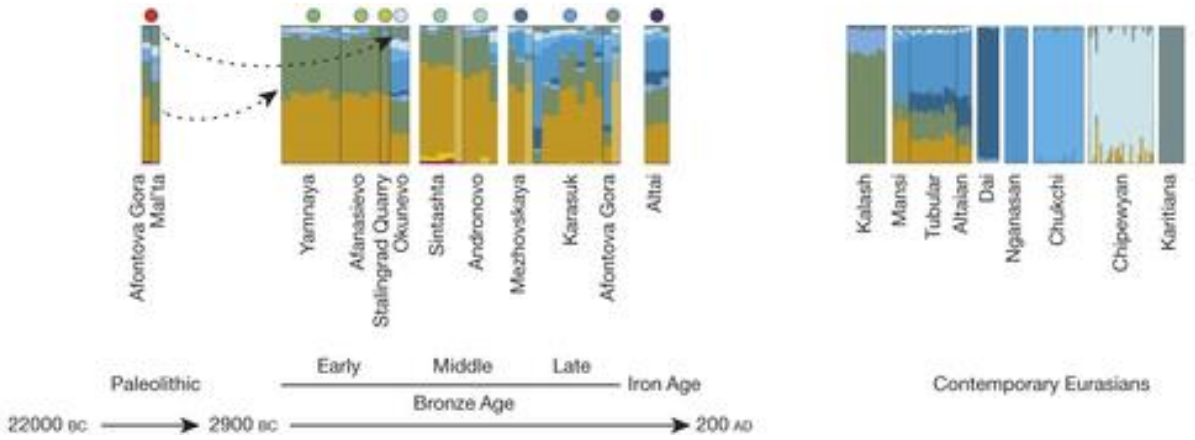


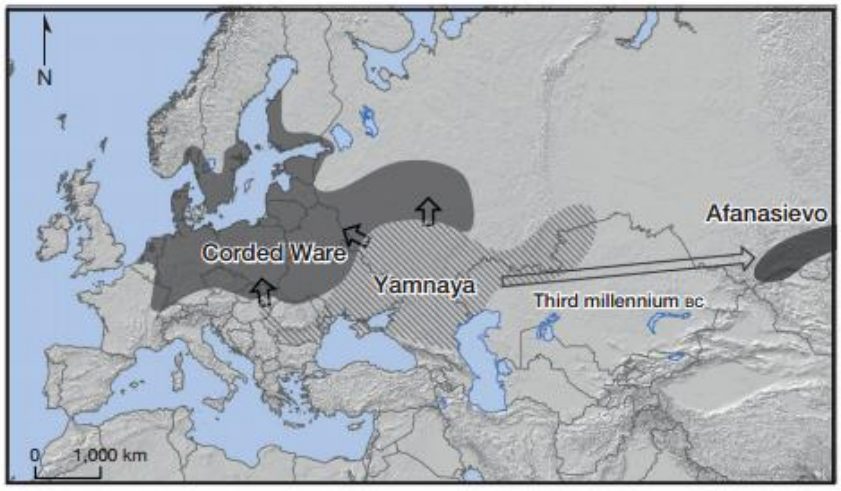
Figure 5.2.1: Admixture plot demonstrating the influence of Bronze Age migration on European genetic pool (Allentoft et al., 2015).

Thus this Italian sample clustered with the Neolithic farmer group, testifying the existence of a closer genetic relationship with this population substrate. Furthermore, while modern Northern European populations showed a lower genetic differentiation with local ancient Bronze Age populations, nowadays southern European populations (particularly Sicilians) showed higher Fst values with them and the lowest values as regards comparison with Neolithic farmers (Allentoft et al., 2015). These data could

probably be at the base of why the ancestral haplotypes investigated in this study are founded in the Southern and Central Italian samples, but not in the Northern European ones, as highlighted from the network analysis.

It is also known that the Sintashta culture was derived from the Corded Ware culture, as confirmed by their genetic homogeneity, having then migrated eastward into the Asian continent (fig.5.2.2). Accordingly, it has evolved in Central Asia into the Andronovo culture, which gradually expanded in South and East Asia admixing with local populations. These last migrations are maybe linked with major climate changes (Anthony, 2009). Therefore, it is highly probable that the presence of this European haplotype carrying the derived allele for both the SNPs under consideration in South Asian populations is due to the migration processes that characterize all the Bronze Age period.

a)



b)

Figure 5.2.2: a) Yamnaya migration from Northern Europe to East Asia. B) Sintashta evolution and diffusion in South Asia (Allentoft et al., 2015)

No information relative to rs35564151 are instead available from scientific literature. Nevertheless, haplotype analysis suggest that it is included in the same LD block of rs6723108, a SNP located near the *TMEM163* gene and for which recent researches proposed the involvement in the development of type 2 diabetes (T2D). Particularly, the ancestral allele of this SNP is found to reduce insulin secretion (Tabassum et al., 2014). The haplotype analysis pointed out that both ancestral and derived allele for these SNPs are associated exactly with the ancestral and derived alleles of rs35564151, thus showing the same North-South distribution.

The rs1446585 is a missense variant and is in moderate LD with other two SNPs of the *R3HDM1* gene ($r^2 > 0.7$), for which the ancestral alleles are found to be associated with high levels of LDL (rs12465802 and rs4954280) (Ma et al., 2010). These two SNPs share the same worldwide genetic distribution with rs1446585, with the derived allele being almost exclusively present in Europe, particularly in northern countries, and in South Asia.

The South-North decreasing frequency of the ancestral disease-associated alleles, which are highly represented in South and Central Italy, could be probably related to the higher incidence of T2D in Southern Italy than in Central and Northern one (fig. 5.2.3) (http://www.instat.it/). This same situation is found also within other ethnic groups. In fact, globally speaking the prevalence of T2D stood around 2% within Europe, while reaching a frequency of 9% and 13% among Chinese in Singapore and African Americans, respectively, populations in which the ancestral haplotype is almost exclusively observed (Diamond, 2003; Das & Elbein, 2006).

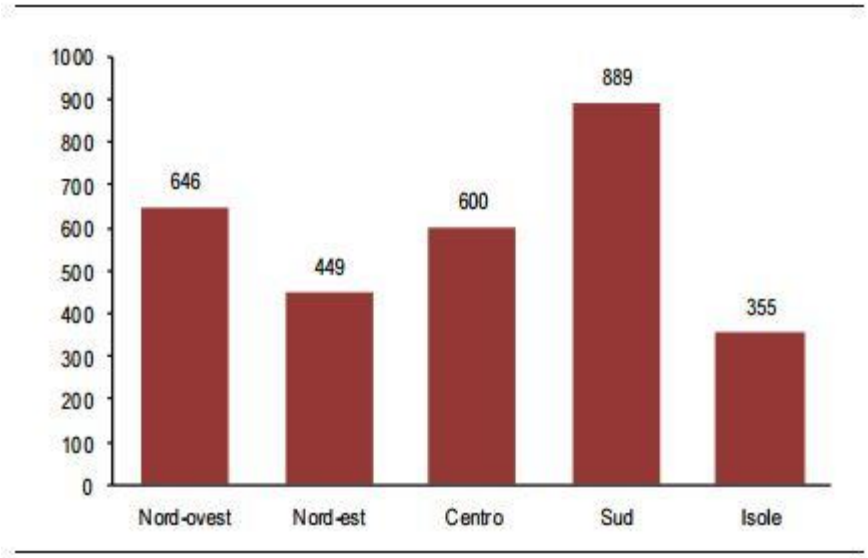


Figure 5.2.3: T2D incidence rate in Italy in 2011. (http://www.instat.it/)

A probable relationship could be found also between distribution of the rs1446585 ancestral allele and incidence of high LDL and triglycerides levels within the Italian population, as represented in figure 5.2.4.

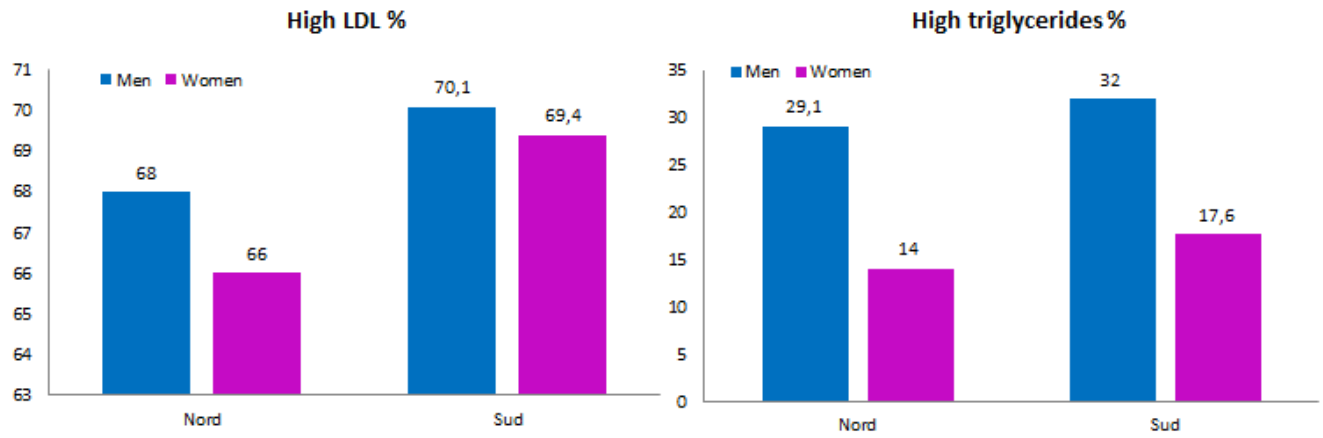


Figure 5.2.4: Data retrieved from "Progetto Cuore" for high LDL and triglycerides incidences in Italy (http://www.cuore.iss.it/fattori/)

Both high LDL levels and T2D are complex multifactorial conditions for which a single allele is not sufficient to be causative, so that the concurrence of genetic traits and environmental factors could represent the actual scenario that determines increased susceptibility to such disease conditions (Dandona & Roberts, 2014; Wu et al., 2014; Hong et al., 2015; Ma et al., 2015; Nikoyeeh et al., 2015; Parhofer et al., 2015). Therefore, it would be reductive to consider these SNPs as the primary causes of the disease development, but at the same time it would not be a farfetched idea to hypothesise their contribute in these pathologic conditions, as also highlighted by recent studies (Ma et al., 2010; Tabassum et al., 2014). Moreover, it is highly plausible that the ancestral disease-associated alleles constitute a risky factor only since recent times, due to the general mismatch between how we have evolved in the past and the very recent dietary shifts to which we have been subjected (Konner & Eaton, 2010). This theory is confirmed by numerous case studies in which the introduction of high-fat diet and a general Western life-style are associated with the gradual spread of several metabolic diseases (particularly T2D) among developing populations (O'Dea et al., 1984; Tabassum et al., 2014).

These two derived alleles were previously not associated with the cited diseases, and are not found in the Neanderthal and Denisova genomes (https://bioinf.eva.mpg.de/jbrowse/), so that appear to be new variants specific of our species.

Furthermore, iHS analysis showed high negative values for both rs1446585 and rs6723108 (in LD with the rs35564151) in CEU (Northern Europeans from Utah), thus indicating that recent positive selection may have acted on the derived alleles in Northern Europe (http://haplotter.uchicago.edu/). According to the association of these alleles with high LDL levels and insulin resistance, it is plausible to hypothesize that natural selection has determined their higher frequencies in North European populations in response to specific dietary-related selective pressures. In fact, populations involved in Bronze Age migrations have been forced to interact with different nutritional resources and climate conditions. Moreover, several climate changes affected this period, resulting in years of extremely cold climate. Particularly, the Eurasian steppe region reached the coldest and driest peak around 2,200-2,000 BCE, as confirmed by pollen analysis from Russia to Samara valley (Kremenetski et al., 1999; Klimento et al., 2000), and to Caucasus, East Africa and India (Weiss, 2000). In conclusion, these populations were likely subjected to important changes in their environmental conditions and available foods, being thus subjected to several dietary shifts. These changes might have influenced the evolutionary forces that maintained the ancestral alleles at high frequency across the world until that moment, and perhaps allowed the diffusion of derived alleles that could have played a crucial adaptive role in cold climates.

Project 3: Celiac disease and its influence on gut microbiota composition: an evolutionary approach to a probiotic treatment case study.

5.3 Discussion of Project 3 results

Celiac disease (CD) is a chronic autoimmune enteropathy caused by the consumption of several type of cereals. In particular, CD immune system is activated in response to specific cereals proteins, such as gliadin, hordein and secalin, which are present respectively in wheat, barley and rye. These proteins share a similar structure that is the reason why they are all named as "gluten" (Fernandez et al., 2010; Roma et al., 2010).

At the moment, the only therapy knew for such disease condition is to complete eliminate grains from dietary sources, as well any other possibility of gluten intake.

CD is a multifactorial disease and, as in other cases, there are evidences of genetic predisposition in developing this pathology. Its development is associated mainly to the human leukocyte antigen system (HLA) for which two risk genotypes are known: HLA-DQ2 and HLA-DQ8 (Spurkland et al., 1990; Fernandez et al.,2013). These two genotypes seem to explain an high percentage of CD incidence, but recent genome-wide association studies (GWAS) suggested an additional role for further 39 non-HLA loci as coeliac disease risk factors (Trynka et al., 2011). However, these new non-HLA genotypes have been estimated to contribute to only 14% of the genetic variance associated with CD, thus researchers are still far from fully understand CD genetic background and from defining a precise disease genetic profile.

The effect of these genetic variants consists in the production of receptors that bind gliadin peptides more than other HLA profiles, leading to the activation of CD4+ T immune cell (Cenit et al., 2015) and to secretion of several inflammation markers, such as gamma interferon (IFN-γ) and tumour necrosis factor alpha (TNF-α). All these elements contribute to developing injury signal on intestinal mucosa, thus promoting matrix degradation and villous atrophy (Lionetti et al., 2008).

Interestingly, a great number of autoimmune diseases have been linked to several changes in the gut microbiota composition, as well as CD (Bisgaard et al., 2011; Collado et al., 2009; Wacklin et al., 2013). To prevent CD exacerbate condition, several potential treatment have been identified and one of the most promising therapy in this field is the administration of probiotics (Bakshi et al., 2012). A probiotic is defined by the FAO as a "*live microorganism, which when administrated in adequate amounts confers a health benefit on the host*" (ftp://ftp.fao.org/es/esn/food/wgreport2.pdf).

In this study, 40 CD patients accepted to test the effect of a specific probiotic treatment composed of *Bifidobacterium breve* strain, through the analysis of their microbiota composition. Among them, 20 received probiotic administration, while the other 20 a placebo. Furthermore, the microbiota of 16 healthy individuals was also analysed as control. The fundamental aim of this study was thus to analyse how the supply of *Bifodobacterium* could affect CD microbiota composition, and to compare these data to those generated for the healthy samples. Particularly, it is well-known that CD patients show high inflammation molecules pattern (as described before) and a general decrease in anti-inflammatory bacteria (Collado et al., 2008), therefore in principle the administration of such strain could positively affect the general microbial composition.

To better evaluate probiotic treatment effect, faecal samples from the 40 CD individuals have been collected at two distinct time: one before the treatment and one at its end. According to this approach, a total of 96 faecal samples have been sequenced for two hypervariable regions of the 16S rRNA gene with an Illumina platform. High-throughput sequencing has been considered the best approach to be applied for such kind of study according to the peculiar advantages that characterize this technique. In fact, as already highlighted by other studies, in the last decade the research on the microbiota has moved into the "*metagenomic era*", being characterized by an increasing number of generated microbial sequences thanks to the recent introduction of massive parallel DNA sequencing techniques (Frank et al., 2008). This approach is particularly useful when applied on 16S rRNA sequencing, because it can provide very detailed information about identity and phylogenetic relationships between microbes. Furthermore, it turned out to be able to describe complex bacterial profile environments, such as the gastrointestinal microbiota composition (Fouhy et al., 2012). Other methods could be indeed indirect or less accurate (e.g. DGGE or TGGE), or maybe very precise, but focused on a single species, thus losing the global bacterial composition (i.e. qPCR). Therefore, this technique has been considered the most performing to characterize CD microbiota and treatment effect on the whole microbial environment composition. Furthermore, numerous researches have already used this kind of approach to study gut microbiota association with different diseases, such as diabetes, Chron disease, IBS, cancer and obesity (Gophna et al., 2006; Roesch et al., 2009; Kassinen et al., 2007; Ley et al., 2010).

As expected, the performed Illumina experiments generated a huge number of sequence data with a general high-quality profile. Two samples have been removed from further analyses due to too few sequences or too high chimeric sequences. A part from these two extreme cases, the general experiment was carried out successfully.

From a first screening about phyla presence within the whole dataset, it is possible to observe the presence of the well-known five dominant bacterial phyla, which have been reported in all the metagenomic studies conducted on human gut: Actinobacteria, Proteobacteria, Verrucomicrobia, Bacteroidetes and Firmicutes. Despite changes in dietary habits, health conditions or through the lifetime, these main phyla are the most common in the gut of modern humans and together they constitute around 90% of the gut ecosystem (Costello et al., 2009; Rodriguez et al., 2015 Wu et al., 2011). These data are in accordance with the results presented here, showing that these main phyla reach the 98% of global biodiversity distribution. Interestingly, it must be noted the presence of Euryarchaeota phylum within the analysed samples. This phylum have been reported in other researches and it is well-established that it could be part of the human gut environment. To date, great attention was given to archaea within human gut microbiota, because they are poorly represented within this particular environment (Horz & Conrads, 2010) and due to several researches that have supposed a previously underestimate role of these species in different pathologic conditions (Cavicchioli et al., 2003; Eckburg et al., 2003).

The cited phyla are obviously not represented with the same relative abundance in all the groups investigated. In fact, they showed a high rate of variability among groups, that was already clearly observable at the phyla level distribution. More in detail, the healthy group showed a very high presence of Firmicutes, which constitute ~60-70% of their gut microbial composition, differently from the CD individuals whose did not received probiotic treatment (named generally as “celiac”). In addition, the healthy microbiota seems to be characterized by a more elevated percentage of Actinobacteria and Euryarchaeota. The association between celiac disease status and a lower presence of Actinobacteria was already described by several studies, so this result is not surprising

(Palma et al., 2012). In the same way, Euryarchaeota are present within the healthy group, but almost completely absent within the celiac groups. This evidence could be conceivably linked to differences in the dietary habits of the two groups. In fact, recent researches focused on the Euryarchaeota phylum, highlighted its ability in promoting polysaccharide degradation and absorption of fatty acids, thus they seem to play a role in energy extraction from degradation of organic compounds (Samuel & Gordon, 2006). All the CD individuals here analysed were following a gluten-free diet from years, therefore the absence of this phylum is highly probable due to the absence of polysaccharide intake in their daily diet. On the other hand, the celiac group seems to be characterized by higher presence of the Bacteroidetes phylum. This difference did not reach a statistical significance after data normalization, but anyway they are more present in celiac microbiota than in the healthy group and this trend is corroborated by results from another study on CD patients with gluten-free diet (Collado et al., 2009).

These divergences came out also when the celiac microbiota is compared to that of healthy people and to that of individuals who received probiotic treatment. The microbiota of this last group seems to share characteristic of both celiac and healthy microbiota. In particular, there was no presence of Euryarchaeota (always linked to the gluten free diet followed), but on the other hand there was a high presence of Firmicutes, as for the healthy group. The Actinobacteria phylum showed no significant difference if compared with the other two groups because it reached values which were exactly in the middle between the two extremes. To detect with more accuracy what may have been the action of probiotic on the treated group, the phyla relative abundances were compared between both pre-treated and treated groups and between treated and healthy groups. In this case, a clear correspondence with the analysis carried on the whole CD dataset compared to the treated group was observed and it is plausible to suppose that the

probiotic administration has acted principally on the Firmicutes phylum, that indeed augmented it percentage within the treated group, and a bit also on the Actinobacteria, (even if not at a statistical significant level). It should not be a surprise due to the fact that the probiotic was composed of an Actinobacteria strain. These evidences about the changes experienced by the microbiota of treated individuals, that get some specific characteristics of the healthy group and maintained others of the celiac one, were highlighted by the cluster analysis, which collocated them exactly in the middle between the other two groups of subjects.

By reaching the microbial family level of analysis, it was possible to get more detailed information about the differences among the examined microbial communities. As previously supposed, the archaea family that characterizes the healthy group is the *Methanobacteriaceae*, a methanogenic archaea that seems to play a role in polysaccharides degradation and that, from what is known until now about our ancestral microbiome, seems to co-exist with humans since several centuries, decreasing its abundance only since the 14th century (Samuel & Gordon, 2006; Huynh et al., 2015). As mentioned before, poor information are available at the moment about ancient human microbiome composition, and on the basis of what is known about *Methanobacteriaceae* function it is plausible to suppose that it was part of our microbiota from long time before the 14th century. This assumption is confirmed from the presence of this same archaea family in the extant species of African apes (Moeller et al., 2014). Particularly, it seems that the *Methanobacteriaceae* have undergone a fivefold reduction within human gut when compared to that of our closest ancestors. This could be interpreted as an evident signal of nutritional shift that human have experienced during their evolutionary history, whereas taxa associated with digestion of complex plant polysaccharides have become less prominent. Grain is the most common source of polysaccharides in modern human

populations, thus the totally absence of this archaea family within celiac group with a gluten-free diet is linked to this further nutritional shift that almost eliminates this species from the gut microbial community.

In contrast with expectations, within the Actinobacteria phylum the family that differs between healthy and celiac groups is the *Coriobacteriaceae*. Therefore, even if a *Bifidobacterium* strain was administrated as probiotic treatment, no difference were founded within treated microbiota composition respect to the pre-treated group or any other group analysed in the present study. It seems that this treatment has thus influenced other bacterial species, without changing at all *Bifidobacterium* presence within the whole microbiota composition, despite what is known from literature (i.e. a reduction in *Bifidobacterium* spp.) (Cenit et al., 2015). Accordingly, further analyses would be necessary to understand what happened during the process of digestion of the probiotic.

By comparing how individuals' microbiota has changed during the administration of the treatment, it came out that two Firmicutes families (*Lactobacillaceae* and *Gracilibacteraceae*) changed their relative abundances, reaching almost the values that characterized the healthy group. Indeed, these two families showed a significant different abundance between pre-treated and treated groups and no differences between treated and healthy individuals. This means that the probiotic has restored the normal amount of microbes belonging to these families within the treated individuals. The augmentation of *Lactobacillaceae* is of relevant interest especially if related to a strong decreasing level of tumour necrosis factor alpha (TNF-α) that has been reported on the same samples (Klemenak et al., 2015). Several evidences demonstrate that *Lactobacillaceae,* a bacterial group characterized by the formation of lactic acid at the end of carbohydrates metabolism, seem to have health-promoting proprieties. In nature, it is possible to find them in plants and fermented food, while within human body they are present in both the

oral cavities and the gastrointestinal tract (Berg et al., 1996; Reuter et al., 2001; Frank et al., 2008). Recent studies indicate that *Lactobacillaceae* can antagonize pro-inflammatory signals from both the human host and the other bacterial species, as for example antagonizing the induction of pro-inflammatory genes in *Shigella flexneri*, also if the mechanism is still not fully elucidated (Servin et al., 2004; Tien et al., 2006). Therefore, it is highly probable that the decrease of TNF-α observed within treated individuals, is closely linked to the increase of *Lactobacilli*, with their anti-inflammatory function, promoted by the administration of *Bifidobacterium*.

It remains to be explained why the administration of such a *Bifidobacterium* strain have affected *Lactobacillaceae* species. This could be related to a high ability of *Bifidobacterium* to deep influence gut microflora composition, by enhancing the blooming of some species and antagonizing others (Li et al., 2004). In particular, there are evidences that *Bifidobacterium* support *Lactobacillaceae* development, although is not so clear the relation existing between these species (Ahmed et al., 2007; Ohutsuka et al., 2012). Probably, an evolutionary link exists between them and with gut microbial composition, which could explain their tight relation. They are both able to convert fermentable sugars to lactic acid, thus allowing food fermentation and preservation (Steinkraus, 2002). Therefore, although *Lactobacillaceae* are present in almost all mammals, it is possible to suppose that they started their co-evolutionary processes with human host during the Neolithic transition (~10,000 years ago). In fact, with the spread of agricultural practices, even fermentation techniques started to be used and diffused all over the world, following different root and in different times, as confirmed by several archaeological records (McGovern et al., 2004). This hypothesis is further corroborated by the very low presence of *Lactobacillaceae* and of *Bifidobacterium* in all the others African apes and even among modern human hunter-gatherer societies (Moeller et al.,

2014; Schnorr et al., 2014). On the other hand, multiple studies noted that these two bacterial families are highly present within modern rural societies, thus in all humans groups in which a strictly Neolithic life-style is maintained (Benno et al., 1986; De Filippo et al., 2010; Turpin et al., 2011). Moreover, other studies observed the lower presence of *Lactobacillaceae* in CD patients (Di Cagno et al., 2011), thus a close relationship between this pathological condition and the bacterial family detected also in this research probably exists, but it is still not fully understand which is the relation between bacterial strains and CD. Multiple factors coexist with this pathological condition, so it is hard to assign a major role to the absence/presence of specific bacterial strain as the one detected in the present study, to the host genetic influence on microbiota composition (Olivares et al., 2015), or to the introduction of novel foods with different process of fermentation (Gobetti et al., 2007). In fact, currently consumed cereals are manufactured by highly accelerated processes in which the normal fermentation bacteria (such as Lactobacilli) are replaced by new chemical yeast, thus deeply changing nutrients profile of foods which did not experience the normal degradation processes traditionally exerted by natural bacteria, resulting in a less digestible food (Gobetti, 1998).

Finally, it is interesting to note that after probiotic treatment, individuals exhibited a close relationship at the family level with the healthy group, as further highlighted by structure analysis. PCA highlighted also the existence of higher variability in species within the healthy group, which is confirmed by the α-diversity observed index. From the other two indices, it is also possible to note that although the healthy group seems to have a slightly higher mean value for Chao1 (thus a higher number of species), no statistical differences were observed among groups. This result highlighted the absence of microbial disbiosis within celiac individuals, probably related to the fact that they are following a gluten-free

diet from years, thus their microbiota had found a new equilibrium which is characterized by a different species distribution than the healthy one.

5.4 Concluding remarks

The present project showed that different technical approach are able to analyse into deep natural selection events, demonstrating how different study perspectives may highlight different variables that may have contributed to local adaptation phenomena.

More in details the analysis of human genetic variability (both thought an *in silico* approach than thought *ex novo* genotyping experiment of highly informative variants) and of microbial biodiversity, led to the identification of multiple signal of local recent adaptation events to climate changes (as highlighted in the first project) than to the evolution of several dietary habits (as observed both within Italy than in the microbiota composition).

References

The International HapMap Consortium increases access to data(2005).Pharmacogenomics 6, 6-6.

Agus, A., Denizot, J., Thevenot, J., Martinez-Medina, M., Massier, S., Sauvanet, P., Bernalier-Donadille, A., Denis, S., Hofman, P., Bonnet, R., et al. (2016). Western diet induces a shift in microbiota composition enhancing susceptibility to Adherent-Invasive E. coli infection and intestinal inflammation. Scientific reports 6, 19032-19032.

Ahmed, M., Prasad, J., Gill, H., Stevenson, L., and Gopal, P. (2007). Impact of consumption of different levels of Bifidobacterium lactis HN019 on the intestinal microflora of elderly human subjects. Journal of Nutrition Health & Aging 11, 26-31.

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., and Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. Plos Biology 2, 1591-1599.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

Arnheim, N., and Taylor, C.E. (1969). Non-Darwinian evolution: consequences for neutral allelic variation. Nature 223, 900-903.

Barbera, M.J., Schluter, A., Pedraza, N., Iglesias, R., Villarroya, F., and Giralt, M. (2001). Peroxisome proliferator-activated receptor alpha activates transcription of the brown fat uncoupling protein-1 gene. A link between regulation of the thermogenic and lipid oxidation pathways in the brown fat cell. J Biol Chem 276, 1486-1493.

Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat Genet 40, 340-345.

Benazzi, S., Douka, K., Fornai, C., Bauer, C.C., Kullmer, O., Svoboda, J., Pap, I., Mallegni, F., Bayle, P., Coquerelle, M., et al. (2011). Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature 479, 525-U249.

Benno, Y., Suzuki, K., Narisawa, K., Bruce, W.R., and Mitsuoka, T. (1986). COMPARISON OF THE FECAL MICROFLORA IN RURAL JAPANESE AND URBAN CANADIANS. Microbiology and Immunology 30, 521-532.

Berg, R.D. (1996). The indigenous gastrointestinal microflora. Trends in Microbiology 4, 430-435.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 74, 1111-1120.

Berton, A., Sebban-Kreuzer, C., and Crenon, I. (2007). Role of the structural domains in the functional properties of pancreatic lipase-related protein 2. Febs Journal 274, 6011-6023.

Bisgaard, H., Li, N., Bonnelykke, K., Chawes, B.L., Skov, T., Paludan-Müller, G., Stokholm, J., Smith, B., and Krogfelt, K.A. (2011a). Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. J Allergy Clin Immunol 128, 646-652.e641-645.

Bisgaard, H., Li, N., Bonnelykke, K., Chawes, B.L.K., Skov, T., Paludan-Mueller, G., Stokholm, J., Smith, B., and Krogfelt, K.A. (2011b). Reduced diversity of the intestinal

microbiota during infancy is associated with increased risk of allergic disease at school age. Journal of Allergy and Clinical Immunology 128, 646-U318.

Blurton Jones, N.G., Smith, L.C., O'Connell, J.F., Hawkes, K., and Kamuzora, C.L. (1992). Demography of the Hadza, an increasing and high density population of Savanna foragers. American journal of physical anthropology 89, 159-181.

Boattini, A., Martinez-Cruz, B., Sarno, S., Harmant, C., Useli, A., Sanz, P., Yang-Yao, D., Manry, J., Ciani, G., Luiselli, D., et al. (2013). Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. PLoS One 8, e65441.

Bramanti, B., Thomas, M.G., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M.N., Jankauskas, R., Kind, C.J., et al. (2009). Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. Science 326, 137-140.

Breslin, P.A. (2013). An evolutionary perspective on food and human taste. Curr Biol 23, R409-418.

Brigham, D., and Beard, J. (1996). Iron and thermoregulation: A review. Critical Reviews in Food Science and Nutrition 36, 747-763.

Brown, E.A. (2012). Genetic explorations of recent human metabolic adaptations: hypotheses and evidence. Biol Rev Camb Philos Soc 87, 838-855.

Buchanan, C.C., Torstenson, E.S., Bush, W.S., and Ritchie, M.D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. Journal of the American Medical Informatics Association 19, 289-294.

Büntgen, U., Tegel, W., Nicolussi, K., McCormick, M., Frank, D., Trouet, V., Kaplan, J.O., Herzig, F., Heussner, K.U., Wanner, H., et al. (2011). 2500 years of European climate variability and human susceptibility. Science 331, 578-582.

Campbell, M.C., Ranciaro, A., Zinshteyn, D., Rawlings-Goss, R., Hirbo, J., Thompson, S., Woldemeskel, D., Froment, A., Rucker, J.B., Omar, S.A., et al. (2014). Origin and Differential Selection of Allelic Variation at TAS2R16 Associated with Salicin Bitter Taste Sensitivity in Africa. Molecular Biology and Evolution 31, 288-302.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science 296, 261-262.

Capocasa, M., Anagnostou, P., Bachis, V., Battaggia, C., Bertoncini, S., Biondi, G., Boattini, A., Boschi, I., Brisighelli, F., Caló, C.M., et al. (2014). Linguistic, geographic and genetic isolation: a collaborative study of Italian populations. J Anthropol Sci 92, 201-231.

Cassidy, L.M., Martiniano, R., Murphy, E.M., Teasdale, M.D., Mallory, J., Hartwell, B., and Bradley, D.G. (2016). Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. Proc Natl Acad Sci U S A 113, 368-373.

Cavicchioli, R., Curmi, P.M.G., Saunders, N., and Thomas, T. (2003). Pathogenic archaea: do they exist? Bioessays 25, 1119-1128.

Cenit, M.C., Olivares, M., Codoñer-Franch, P., and Sanz, Y. (2015). Intestinal Microbiota and Celiac Disease: Cause, Consequence or Co-Evolution? Nutrients 7, 6900-6923.

Chakravarthy, M.V., and Booth, F.W. (2004). Eating, exercise, and "thrifty" genotypes: connecting the dots toward an evolutionary understanding of modern chronic diseases. J Appl Physiol (1985) 96, 3-10.

Charlesworth, D., Charlesworth, B., and Morgan, M.T. (1995). THE PATTERN OF NEUTRAL MOLECULAR VARIATION UNDER THE BACKGROUND SELECTION MODEL. Genetics 141, 1619-1632.

Chitty, M. (2006). Deep ancestry: Inside the genographic project. Library Journal 131, 161-161.

Cho, I., and Blaser, M.J. (2012). APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease. Nature Reviews Genetics 13, 260-270.

Cohen, P., Levy, J.D., Zhang, Y., Frontini, A., Kolodin, D.P., Svensson, K.J., Lo, J.C., Zeng, X., Ye, L., Khandekar, M.J., et al. (2014). Ablation of PRDM16 and beige adipose causes metabolic dysfunction and a subcutaneous to visceral fat switch. Cell 156, 304-316.

Collado, M.C., Donat, E., Ribes-Koninckx, C., Calabuig, M., and Sanz, Y. (2008). Imbalances in faecal and duodenal Bifidobacterium species composition in active and non-active coeliac disease. BMC Microbiol 8, 232.

Collado, M.C., Donat, E., Ribes-Koninckx, C., Calabuig, M., and Sanz, Y. (2009). Specific duodenal and faecal bacterial groups associated with paediatric coeliac disease. J Clin Pathol 62, 264-269.

Colonna, V., Pagani, L., Xue, Y., and Tyler-Smith, C. (2011). A world in a grain of sand: human history from genetic data. Genome Biology 12.

Cordain, L., Eaton, S.B., Miller, J.B., Mann, N., and Hill, K. (2002). The paradoxical nature of hunter-gatherer diets: meat-based, yet non-atherogenic. European Journal of Clinical Nutrition 56, S42-S52.

Cordain, L., Eaton, S.B., Sebastian, A., Mann, N., Lindeberg, S., Watkins, B.A., O'Keefe, J.H., and Brand-Miller, J. (2005). Origins and evolution of the Western diet: health implications for the 21st century. Am J Clin Nutr 81, 341-354.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. Science 326, 1694-1697.

Craig, O.E., Steele, V.J., Fischer, A., Hartz, S., Andersen, S.H., Donohoe, P., Glykou, A., Saul, H., Jones, D.M., Koch, E., et al. (2011). Ancient lipids reveal continuity in culinary practices across the transition to agriculture in Northern Europe. Proceedings of the National Academy of Sciences of the United States of America 108, 17910-17915.

Cramp, L.J.E., Evershed, R.P., Lavento, M., Halinen, P., Mannermaa, K., Oinonen, M., Kettunen, J., Perola, M., Onkamo, P., and Heyd, V. (2014a). Neolithic dairy farming at the extreme of agriculture in northern Europe. Proceedings of the Royal Society B-Biological Sciences 281.

Cramp, L.J.E., Jones, J., Sheridan, A., Smyth, J., Whelton, H., Mulville, J., Sharples, N., and Evershed, R.P. (2014b). Immediate replacement of fishing with dairying by the earliest farmers of the northeast Atlantic archipelagos. Proceedings of the Royal Society B-Biological Sciences 281.

Currat, M., Poloni, E.S., and Sanchez-Mazas, A. (2010). Human genetic differentiation across the Strait of Gibraltar. Bmc Evolutionary Biology 10.

D'Arienzo, R., Maurano, F., Luongo, D., Mazzarella, G., Stefanile, R., Troncone, R., Auricchio, S., Ricca, E., David, C., and Rossi, M. (2008). Adjuvant effect of Lactobacillus casei in a mouse model of gluten sensitivity. Immunology Letters 119, 78-83.

D'Arienzo, R., Stefanile, R., Maurano, F., Mazzarella, G., Ricca, E., Troncone, R., Auricchio, S., and Rossi, M. (2011). Immunomodulatory Effects of Lactobacillus casei Administration in a Mouse Model of Gliadin-Sensitive Enteropathy. Scandinavian Journal of Immunology 74, 335-341.

Daikoku, T., Shinohara, Y., Shima, A., Yamazaki, N., and Terada, H. (2000). Specific elevation of transcript levels of particular protein subtypes induced in brown adipose tissue by cold exposure. Biochim Biophys Acta 1457, 263-272.

Dandona, S., and Roberts, R. (2014). The role of genetic risk factors in coronary artery disease. Curr Cardiol Rep 16, 479.

Das, S.K., and Elbein, S.C. (2006). The Genetic Basis of Type 2 Diabetes. Cellscience 2, 100-131.

De Fanti, S., Barbieri, C., Sarno, S., Sevini, F., Vianello, D., Tamm, E., Metspalu, E., van Oven, M., Hubner, A., Sazzini, M., et al. (2015a). Fine Dissection of Human Mitochondrial DNA Haplogroup HV Lineages Reveals Paleolithic Signatures from European Glacial Refugia. PloS one 10, e0144391-e0144391.

De Fanti, S., Sazzini, M., Giuliani, C., Frazzoni, F., Sarno, S., Boattini, A., Marasco, E., Mantovani, V., Franceschi, C., Moral, P., et al. (2015b). Inferring the genetic history of lactase persistence along the Italian peninsula from a large genomic interval surrounding the LCT gene. Am J Phys Anthropol 158, 708-718.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proceedings of the National Academy of Sciences of the United States of America 107, 14691-14696.

De Palma, G., Capilla, A., Nova, E., Castillejo, G., Varea, V., Pozo, T., Antonio Garrote, J., Polanco, I., Lopez, A., Ribes-Koninckx, C., et al. (2012). Influence of Milk-Feeding Type and Genetic Risk of Developing Coeliac Disease on Intestinal Microbiota of Infants: The PROFICEL Study. Plos One 7.

Devkota, S., and Chang, E.B. (2015). Interactions between Diet, Bile Acid Metabolism, Gut Microbiota, and Inflammatory Bowel Diseases. Digestive Diseases 33, 351-356.

Di Cagno, R., De Angelis, M., De Pasquale, I., Ndagijimana, M., Vernocchi, P., Ricciuti, P., Gagliardi, F., Laghi, L., Crecchio, C., Guerzoni, M.E., et al. (2011). Duodenal and faecal microbiota of celiac children: molecular, phenotype and metabolome characterization. Bmc Microbiology 11.

Diamond, J. (2003). The double puzzle of diabetes. Nature 423, 599-602.

Doughty, C.E. (2013). Preindustrial Human Impacts on Global and Regional Environment. Annual Review of Environment and Resources, Vol 38 38, 503-527.

Dronfield, J. (1996). Europe in the Neolithic: The creation of new worlds - Whittle,A. Antiquity 70, 985-987.

Duar, R.M., Clark, K.J., Patil, P.B., Hernandez, C., Bruening, S., Burkey, T.E., Madayiputhiya, N., Taylor, S.L., and Walter, J. (2015). Identification and characterization

of intestinal lactobacilli strains capable of degrading immunotoxic peptides present in gluten. Journal of Applied Microbiology 118, 515-527.

Eaton, S.B. (2006). The ancestral human diet: what was it and should it be a paradigm for contemporary nutrition ? Proceedings of the Nutrition Society 65, 1-6.

Eckburg, P.B., Lepp, P.W., and Relman, D.A. (2003). Archaea and their potential role in human disease. Infection and Immunity 71, 591-596.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics 155, 1405-1413.

Fernandes, R., Grootes, P., Nadeau, M.-J., and Nehlich, O. (2015). Quantitative diet reconstruction of a Neolithic population using a Bayesian mixing model (FRUITS): The case study of Ostorf (Germany). American Journal of Physical Anthropology 158, 325-340.

Fischer, A., Gilad, Y., Man, O., and Paabo, S. (2005). Evolution of bitter taste receptors in humans and apes. Molecular Biology and Evolution 22, 432-436.

Fouhy, F., Ross, R.P., Fitzgerald, G.F., Stanton, C., and Cotter, P.D. (2012). Composition of the early intestinal microbiota: knowledge, knowledge gaps and the use of high-throughput sequencing to address these gaps. Gut Microbes 3, 203-220.

Frank, D.N., and Pace, N.R. (2008a). Gastrointestinal microbiology enters the metagenomics era. Curr Opin Gastroenterol 24, 4-10.

Frank, D.N., and Pace, N.R. (2008b). Gastrointestinal microbiology enters the metagenomics era. Current Opinion in Gastroenterology 24, 4-10.

Frühbeck, G., Sesma, P., and Burrell, M.A. (2009). PRDM16: the interconvertible adipo-myocyte switch. Trends Cell Biol 19, 141-146.

Fumagalli, M., and Sironi, M. (2014). Human genome variability, natural selection and infectious diseases. Current Opinion in Immunology 30, 9-16.

Gautier, M., and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28, 1176-1177.

Gilad, Y., Bustamante, C.D., Lancet, D., and Paabo, S. (2003). Natural selection on the olfactory receptor gene family in humans and chimpanzees. American Journal of Human Genetics 73, 489-501.

Gillings MR(2014). Microbiology of the Anthropocene. Anthropocene 5, 1 - 8.

Giralt, M., and Villarroya, F. (2013). White, brown, beige/brite: different adipose cells for different functions? Endocrinology 154, 2992-3000.

Gobbetti, M. (1998). The sourdough microflora: Interactions of lactic acid bacteria and yeasts. Trends in Food Science & Technology 9, 267-274.

Gobbetti, M., Rizzello, C.G., Di Cagno, R., and De Angelis, M. (2007). Sourdough lactobacilli and celiac disease. Food Microbiology 24, 187-196.

Hancock, A.M., Alkorta-Aranburu, G., Witonsky, D.B., and Di Rienzo, A. (2010a). Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond B Biol Sci 365, 2459-2468.

Hancock, A.M., Clark, V.J., Qian, Y., and Di Rienzo, A. (2011a). Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. Mol Biol Evol 28, 601-614.

Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., Beall, C.M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. (2011b). Adaptations to climate-mediated selective pressures in humans. PLoS Genet 7, e1001375.

Hancock, A.M., Witonsky, D.B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J., Coop, G., et al. (2010b). Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proceedings of the National Academy of Sciences of the United States of America 107, 8924-8930.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. (2008). Adaptations to climate in candidate genes for common metabolic disorders. Plos Genetics 4.

Heath, K.M., Axton, J.H., McCullough, J.M., and Harris, N. (2016). The evolutionary adaptation of the C282Y mutation to culture and climate during the European Neolithic. Am J Phys Anthropol.

Hedrick, P.W. (1980). Hitchhiking: a comparison of linkage and partial selfing. Genetics 94, 791-808.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G., Przeworski, M., and Project, G. (2011). Classic selective sweeps were rare in recent human evolution. Science 331, 920-924.

Heyer, E., and Quintana-Murci, L. (2009). Evolutionary genetics as a tool to target genes involved in phenotypes of medical relevance. Evol Appl 2, 71-80.

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez del Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., et al. (2015). Early farmers from across Europe directly descended from Neolithic Aegeans. bioRxiv.

Hondares, E., Rosell, M., Díaz-Delfín, J., Olmos, Y., Monsalve, M., Iglesias, R., Villarroya, F., and Giralt, M. (2011). Peroxisome proliferator-activated receptor ? (PPAR?) induces PPAR? coactivator 1? (PGC-1?) gene expression and contributes to thermogenic activation of brown fat: involvement of PRDM16. J Biol Chem 286, 43112-43122.

Hong, Y.J., Kang, E.S., Ji, M.J., Choi, H.J., Oh, T., Koong, S.S., and Jeon, H.J. (2015). Association between Bsm1 Polymorphism in Vitamin D Receptor Gene and Diabetic Retinopathy of Type 2 Diabetes in Korean Population. Endocrinol Metab (Seoul) 30, 469-474.

Horz, H.P., and Conrads, G. (2010). The discussion goes on: What is the role of Euryarchaeota in humans? Archaea 2010, 967271.

Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A.R., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D., et al. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics 40, 395-402.

Huttunen, P., Hirvonen, J., and Kinnula, V. (1981). The occurrence of brown adipose tissue in outdoor workers. Eur J Appl Physiol Occup Physiol 46, 339-345.

Itan, Y., Powell, A., Beaumont, M.A., Burger, J., and Thomas, M.G. (2009). The origins of lactase persistence in Europe. PLoS Comput Biol 5, e1000491.

Jablonski, N.G., and Chaplin, G. (2010). Colloquium paper: human skin pigmentation as an adaptation to UV radiation. Proc Natl Acad Sci U S A 107 Suppl 2, 8962-8968.

Johnson, F., Mavrogianni, A., Ucci, M., Vidal-Puig, A., and Wardle, J. (2011). Could increased time spent in a thermal comfort zone contribute to population increases in obesity? Obes Rev 12, 543-551.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11, 94.

Kajimura, S., Seale, P., Tomaru, T., Erdjument-Bromage, H., Cooper, M.P., Ruas, J.L., Chin, S., Tempst, P., Lazar, M.A., and Spiegelman, B.M. (2008). Regulation of the brown and white fat gene programs through a PRDM16/CtBP transcriptional complex. Genes Dev 22, 1397-1409.

Kaniewski, D., Van Campo, E., Guiot, J., Le Burel, S., Otto, T., and Baeteman, C. (2013). Environmental roots of the late bronze age crisis. PLoS One 8, e71004.

Kaplan, N.L., Hudson, R.R., and Langley, C.H. (1989). THE HITCHHIKING EFFECT REVISITED. Genetics 123, 887-899.

Kassinen, A., Krogius-Kurikka, L., Mäkivuokko, H., Rinttilä, T., Paulin, L., Corander, J., Malinen, E., Apajalahti, J., and Palva, A. (2007). The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects. Gastroenterology 133, 24-33.

Kayser, M., Brauer, S., and Stoneking, M. (2003). A genome scan to detect candidate regions influenced by local natural selection in human populations. Molecular Biology and Evolution 20, 893-900.

Kelly, J.K., and Wade, M.J. (2000). Molecular evolution near a two-locus balanced polymorphism. Journal of Theoretical Biology 204, 83-101.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature 217, 624-626.

King, J.L., and Jukes, T.H. (1969). Non-Darwinian evolution. Science (New York, NY) 164, 788-798.

Konner, M., and Eaton, S.B. (2010a). Paleolithic Nutrition Twenty-Five Years Later. Nutrition in Clinical Practice 25, 594-602.

Konner, M., and Eaton, S.B. (2010b). Paleolithic nutrition: twenty-five years later. Nutr Clin Pract 25, 594-602.

Kuipers, R.S., Luxwolda, M.F., Dijck-Brouwer, D.A.J., Eaton, S.B., Crawford, M.A., Cordain, L., and Muskiet, F.A.J. (2010). Estimated macronutrient and fatty acid intakes from an East African Paleolithic diet. British Journal of Nutrition 104, 1666-1687.

Lalueza-Fox, C., Sampietro, M.L., Gilbert, M.T., Castri, L., Facchini, F., Pettener, D., and Bertranpetit, J. (2004). Unravelling migrations in the steppe: mitochondrial DNA sequences from ancient central Asians. Proc Biol Sci 271, 941-947.

Lan, Y., Wang, Q., Cole, J.R., and Rosen, G.L. (2012). Using the RDP Classifier to Predict Taxonomic Novelty and Reduce the Search Space for Finding Novel Organisms. Plos One 7.

Lappalainen, T., Salmela, E., Andersen, P.M., Dahlman-Wright, K., Sistonen, P., Savontaus, M.L., Schreiber, S., Lahermo, P., and Kere, J. (2010). Genomic landscape of positive natural selection in Northern European populations. Eur J Hum Genet 18, 471-478.

Lewis, S.L., and Maslin, M.A. (2015). Defining the Anthropocene. Nature 519, 171-180.

Ley, R.E. (2010). Obesity and the human microbiome. Curr Opin Gastroenterol 26, 5-11.

Li, H., Pakstis, A.J., Kidd, J.R., and Kidd, K.K. (2011a). Selection on the Human Bitter Taste Gene, TAS2R16, in Eurasian Populations. Human Biology 83, 363-377.

Li, H., Zhao, X., Zhao, Y., Li, C., Si, D., Zhou, H., and Cui, Y. (2011b). Genetic characteristics and migration history of a bronze culture population in the West Liao-River valley revealed by ancient DNA. J Hum Genet 56, 815-822.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.

Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213-2233.

Li, Y.D., Shimizu, T., Hosaka, A., Kaneko, N., Ohtsuka, Y., and Yamashiro, Y. (2004). Effects of bifidobacterium breve supplementation on intestinal flora of low birth weight infants. Pediatrics International 46, 509-515.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451-1452.

Lopez Herraez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M.R., Gross, A., Scholz, M., and Stoneking, M. (2009). Genetic variation and recent

positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PloS one 4, e7888-e7888.

Lorenzo Pisarello, M.J., Vintini, E.O., Gonzalez, S.N., Pagani, F., and Medina, M.S. (2015). Decrease in lactobacilli in the intestinal microbiota of celiac children with a gluten-free diet, and selection of potentially probiotic strains. Canadian journal of microbiology 61, 32-37.

Luca, F., Bubba, G., Basile, M., Brdicka, R., Michalodimitrakis, E., Rickards, O., Vershubsky, G., Quintana-Murci, L., Kozlov, A.I., and Novelletto, A. (2008). Multiple Advantageous Amino Acid Variants in the NAT2 Gene in Human Populations. Plos One 3.

Luca, F., Hudson, R.R., Witonsky, D.B., and Di Rienzo, A. (2011). A reduced representation approach to population genetic analyses and applications to human evolution. Genome Research 21, 1087-1098.

Luca, F., Perry, G.H., and Di Rienzo, A. (2010). Evolutionary adaptations to dietary changes. Annu Rev Nutr 30, 291-314.

Ma, E., Wang, H., Guo, J., Tian, R., and Wei, L. (2015). The association between the rs11196218A/G polymorphism of the TCF7L2 gene and type 2 diabetes in the Chinese Han population: a meta-analysis. Clinics (Sao Paulo) 70, 593-599.

Malmstrom, H., Gilbert, M.T.P., Thomas, M.G., Brandstrom, M., Stora, J., Molnar, P., Andersen, P.K., Bendixen, C., Holmlund, G., Gotherstrom, A., et al. (2009). Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians. Current Biology 19, 1758-1762.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528, 499-+.

McGovern, P.E., Zhang, J.H., Tang, J.G., Zhang, Z.Q., Hall, G.R., Moreau, R.A., Nunez, A., Butrym, E.D., Richards, M.P., Wang, C.S., et al. (2004). Fermented beverages of pre- and proto-historic China. Proceedings of the National Academy of Sciences of the United States of America 101, 17593-17598.

McMurdie, P.J., and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. Plos Computational Biology 10.

Miller, G.H., Geirsdóttir, Áslaug,Zhong, Yafang,Larsen, Darren J.,Otto-Bliesner, Bette L.,Holland, Marika M.,Bailey, David A.,Refsnider, Kurt A.,Lehman, Scott J.,Southon, John R.,Anderson, Chance,Björnsson, Helgi,Thordarson, Thorvaldur (2012). Abrupt onset of the Little Ice Age triggered by volcanism and sustained by sea-ice/ocean feedbacks (Geophysical Research Letters).

Mina, S.S., Azcurra, A.I., Dorronsoro, S., and Brunotto, M.N. (2008). Alterations of the oral ecosystem in children with celiac disease. Acta odontologica latinoamericana : AOL 21, 121-126.

Moeller, A.H., Li, Y., Ngole, E.M., Ahuka-Mundeke, S., Lonsdorf, E.V., Pusey, A.E., Peeters, M., Hahn, B.H., and Ochman, H. (2014). Rapid changes in the gut microbiome during human evolution. Proceedings of the National Academy of Sciences of the United States of America 111, 16431-16435.

Nash, D., Nair, S., Mayxay, M., Newton, P.N., Guthmann, J.P., Nosten, F., and Anderson, T.J.C. (2005). Selection strength and hitchhiking around two anti-malarial

resistance genes. Proceedings of the Royal Society B-Biological Sciences 272, 1153-1161.

NEEL, J.V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14, 353-362.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. Heredity 86, 641-647.

Nikooyeh, B., and Neyestani, T.R. (2015). Oxidative stress, type 2 diabetes and vitamin D: past, present and future. Diabetes Metab Res Rev.

O'Dea, K. (1984). Marked improvement in carbohydrate and lipid metabolism in diabetic Australian aborigines after temporary reversion to traditional lifestyle. Diabetes 33, 596-603.

Ohno, H., Shinoda, K., Ohyama, K., Sharp, L.Z., and Kajimura, S. (2013). EHMT1 controls brown adipose cell fate and thermogenesis through the PRDM16 complex. Nature 504, 163-167.

Ohno, H., Shinoda, K., Spiegelman, B.M., and Kajimura, S. (2012). PPAR? agonists induce a white-to-brown fat conversion through stabilization of PRDM16 protein. Cell Metab 15, 395-404.

Olivares, M., Neef, A., Castillejo, G., De Palma, G., Varea, V., Capilla, A., Palau, F., Nova, E., Marcos, A., Polanco, I., et al. (2015). The HLA-DQ2 genotype selects for early intestinal microbiota composition in infants at high risk of developing coeliac disease. Gut 64, 406-417.

Orlando, A., Linsalata, M., Notarnicola, M., Tutino, V., and Russo, F. (2014). Lactobacillus GG restoration of the gliadin induced epithelial barrier disruption: the role of cellular polyamines. Bmc Microbiology 14.

Otto, S.P. (2000). Detecting the form of selection from DNA sequence data. Trends in Genetics 16, 526-529.

Ouellet, V., Labbé, S.M., Blondin, D.P., Phoenix, S., Guérin, B., Haman, F., Turcotte, E.E., Richard, D., and Carpentier, A.C. (2012). Brown adipose tissue oxidative metabolism contributes to energy expenditure during acute cold exposure in humans. J Clin Invest 122, 545-552.

Pala, M., Achilli, A., Olivieri, A., Kashani, B.H., Perego, U.A., Sanna, D., Metspalu, E., Tambets, K., Tamm, E., Accetturo, M., et al. (2009). Mitochondrial Haplogroup U5b3: A Distant Echo of the Epipaleolithic in Italy and the Legacy of the Early Sardinians. American Journal of Human Genetics 84, 814-821.

Parhofer, K.G. (2015). Interaction between Glucose and Lipid Metabolism: More than Diabetic Dyslipidemia. Diabetes Metab J 39, 353-362.

Price, A.L., Helgason, A., Palsson, S., Stefansson, H., Clair, D.S., Andreassen, O.A., Reich, D., Kong, A., and Stefansson, K. (2009). The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland. Plos Genetics 5.

Puigserver, P., Wu, Z., Park, C.W., Graves, R., Wright, M., and Spiegelman, B.M. (1998). A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. Cell 92, 829-839.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-

genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-U70.

Raj, S.M., Pagani, L., Gallego Romero, I., Kivisild, T., and Amos, W. (2013). A general linear model-based approach for inferring selection to climate. BMC Genet 14, 87.

Reuter, G. (2001). The Lactobacillus and Bifidobacterium microflora of the human intestine: composition and succession. Current issues in intestinal microbiology 2, 43-53.

Rizzello, C.G., De Angelis, M., Di Cagno, R., Camarca, A., Silano, M., Losito, A., De Vincenzi, M., De Bari, M.D., Palmisano, F., Maurano, F., et al. (2007). Highly efficient gluten degradation by lactobacilli and fungal proteases during food processing: New perspectives for celiac disease. Applied and Environmental Microbiology 73, 4499-4507.

Rodriguez, J.M., Murphy, K., Stanton, C., Ross, R.P., Kober, O.I., Juge, N., Avershina, E., Rudi, K., Narbad, A., Jenmalm, M.C., et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. Microbial ecology in health and disease 26, 26050-26050.

Roesch, L.F., Lorca, G.L., Casella, G., Giongo, A., Naranjo, A., Pionzio, A.M., Li, N., Mai, V., Wasserfall, C.H., Schatz, D., et al. (2009). Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. ISME J 3, 536-548.

Ronald, J., and Akey, J.M. (2005). Genome-wide scans for loci under selection in humans. Human genomics 2, 113-125.

Ruddiman, W.F. (2013). The Anthropocene. Annual Review of Earth and Planetary Sciences, Vol 41 41, 45-68.

Ruiz Martinez, R.C., Bedani, R., and Isay Saad, S.M. (2015). Scientific evidence for health effects attributed to the consumption of probiotics and prebiotics: an update for current perspectives and future challenges. British Journal of Nutrition 114, 1993-2015.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002a). Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-837.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002b). Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-837.

Sams, A., and Hawks, J. (2013). Patterns of Population Differentiation and Natural Selection on the Celiac Disease Background Risk Network. Plos One 8.

Samuel, B.S., and Gordon, J.I. (2006). A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. Proceedings of the National Academy of Sciences of the United States of America 103, 10011-10016.

Sanz, Y., Sanchez, E., Marzotto, M., Calabuig, M., Torriani, S., and Dellaglio, F. (2007). Differences in faecal bacterial communities in coeliac and healthy children as detected by PCR and denaturing gradient gel electrophoresis. Fems Immunology and Medical Microbiology 51, 562-568.

Sarno, S., Boattini, A., Carta, M., Ferri, G., Alu, M., Yao, D.Y., Ciani, G., Pettener, D., and Luiselli, D. (2014). An Ancient Mediterranean Melting Pot: Investigating the Uniparental Genetic Structure and Population History of Sicily and Southern Italy. Plos One 9.

Sazzini, M., Schiavo, G., De Fanti, S., Martelli, P.L., Casadio, R., and Luiselli, D. (2014). Searching for signatures of cold adaptations in modern and archaic humans: hints from the brown adipose tissue genes. Heredity (Edinb) 113, 259-267.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78, 629-644.

Schmidt, P.S., Bertness, M.D., and Rand, D.M. (2000). Environmental heterogeneity and balancing selection in the acorn barnacle Semibalanus balanoides. Proceedings of the Royal Society B-Biological Sciences 267, 379-384.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-864.

Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. Nature Communications 5.

Seale, P., Bjork, B., Yang, W., Kajimura, S., Chin, S., Kuang, S., Scimè, A., Devarakonda, S., Conroe, H.M., Erdjument-Bromage, H., et al. (2008). PRDM16 controls a brown fat/skeletal muscle switch. Nature 454, 961-967.

Seale, P., Kajimura, S., Yang, W., Chin, S., Rohas, L.M., Uldry, M., Tavernier, G., Langin, D., and Spiegelman, B.M. (2007). Transcriptional control of brown fat determination by PRDM16. Cell Metab 6, 38-54.

Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., et al. (2014). Genomic structure in Europeans dating back at least 36,200 years. Science 346, 1113-1118.

Servin, A.L. (2004). Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. Fems Microbiology Reviews 28, 405-440.

Sheftel, A.D., Mason, A.B., and Ponka, P. (2012). The long history of iron in the Universe and in health and disease. Biochimica Et Biophysica Acta-General Subjects 1820, 161-187.

Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. Nature Genetics 46, 543-550.

Skoglund, P., Malmstrom, H., Raghavan, M., Stora, J., Hall, P., Willerslev, E., Gilbert, M.T.P., Gotherstrom, A., and Jakobsson, M. (2012). Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336, 466-469.

Steinkraus, K.H. (1994). NUTRITIONAL SIGNIFICANCE OF FERMENTED FOODS. Food Research International 27, 259-267.

Stewart, J.R., and Stringer, C.B. (2012). Human evolution out of Africa: the role of refugia and climate change. Science 335, 1317-1321.

Storz, J.F., Payseur, B.A., and Nachman, M.W. (2004). Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Molecular Biology and Evolution 21, 1800-1811.

Straus, L.G. (2015). THE HUMAN OCCUPATION OF SOUTHWESTERN EUROPE DURING THE LAST GLACIAL MAXIMUM Solutrean Cultural Adaptations in France and Iberia. Journal of Anthropological Research 71, 465-492.

Sturm, R.A. (2009). Molecular genetics of human pigmentation diversity. Hum Mol Genet 18, R9-17.

Tallavaara, M., Luoto, M., Korhonen, N., Jarvinen, H., and Seppa, H. (2015). Human population dynamics in Europe over the Last Glacial Maximum. Proceedings of the National Academy of Sciences of the United States of America 112, 8232-8237.

Tank, A., Wijngaard, J.B., Konnen, G.P., Bohm, R., Demaree, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. International Journal of Climatology 22, 1441-1453.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-713.

Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. (2004). CYP3A variation and the evolution of salt-sensitivity variants. Am J Hum Genet 75, 1059-1069.

Thomson, G. (1977). The effect of a selected locus on linked neutral loci. Genetics 85, 753-788.

Tien, M.T., Girardin, S.E., Regnault, B., Le Bourhis, L., Dillies, M.A., Coppee, J.Y., Bourdet-Sicard, R., Sansonetti, P.J., and Pedron, T. (2006). Anti-inflammatory effect of Lactobacillus casei on Shigella-infected human intestinal epithelial cells. Journal of Immunology 176, 1228-1237.

Turchin, M.C., Chiang, C.W., Palmer, C.D., Sankararaman, S., Reich, D., Hirschhorn, J.N., and Consortium, G.I.o.A.T.G. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44, 1015-1019.

Turpin, W., Humblot, C., and Guyot, J.-P. (2011). Genetic Screening of Functional Properties of Lactic Acid Bacteria in a Fermented Pearl Millet Slurry and in the Metagenome of Fermented Starchy Foods. Applied and Environmental Microbiology 77, 8722-8734.

U, G. (2006). Differences between Tissue-Associated Intestinal Microfloras of Patients with Crohn's Disease and Ulcerative Colitis G.S. Sommerfeld K, Doolittle WF, Veldhuyzen van Zanten SJO, ed. (Journal of Clinical Microbiology), pp. 4136-4141

Vane-Wright, D. (2004). Entomology - Butterflies at that awkward age. Nature 428, 477-+.

Verdu, E.F., Galipeau, H.J., and Jabri, B. (2015). Novel players in coeliac disease pathogenesis: role of the gut microbiota. Nature Reviews Gastroenterology & Hepatology 12, 497-506.

Verrelli, B.C., Lewis, C.M., Jr., Stone, A.C., and Perry, G.H. (2008). Different Selective Pressures Shape the Molecular Evolution of Color Vision in Chimpanzee and Human Populations. Molecular Biology and Evolution 25, 2735-2743.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol 4, e72.

Wacklin, P., Kaukinen, K., Tuovinen, E., Collin, P., Lindfors, K., Partanen, J., Mäki, M., and Mättö, J. (2013). The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease. Inflamm Bowel Dis 19, 934-941.

Walker, W.A. (2013). Initial Intestinal Colonization in the Human Infant and Immune Homeostasis. Annals of Nutrition and Metabolism 63, 8-15.

Watanabe, R., Iino, R., Shimabukuro, K., Yoshida, M., and Noji, H. (2008). Temperature-sensitive reaction intermediate of F1-ATPase. EMBO Rep 9, 84-90.

Watve, M.G., and Yajnik, C.S. (2007). Evolutionary origins of insulin resistance: a behavioral switch hypothesis. BMC Evol Biol 7, 61.

West, C.E., Renz, H., Jenmalm, M.C., Kozyrskyj, A.L., Allen, K.J., Vuillermin, P., Prescott, S.L., MacKay, C., Salminen, S., Wong, G., et al. (2015). The gut microbiota and inflammatory noncommunicable diseases: Associations and potentials for gut microbiota therapies. Journal of Allergy and Clinical Immunology 135, 3-14.

Wooding, S., Bufe, B., Grassi, C., Howard, M.T., Stone, A.C., Vazquez, M., Dunn, D.M., Meyerhof, W., Weiss, R.B., and Bamshad, M.J. (2006). Independent evolution of bitter-taste sensitivity in humans and chimpanzees. Nature 440, 930-934.

Wright, J.D. Deglaciation triggered by the resumption of North Atlantic Deep Water.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011a). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science 334, 105-108.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011b). Linking long-term dietary patterns with gut microbial enterotypes. Science 334, 105-108.

Wu, J., Boström, P., Sparks, L.M., Ye, L., Choi, J.H., Giang, A.H., Khandekar, M., Virtanen, K.A., Nuutila, P., Schaart, G., et al. (2012). Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. Cell 150, 366-376.

Wu, N.Q., and Li, J.J. (2014). PCSK9 gene mutations and low-density lipoprotein cholesterol. Clin Chim Acta 431, 148-153.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature 486, 222-+.

Ye, K., and Gu, Z. (2011). Recent advances in understanding the role of nutrition in human genome evolution. Adv Nutr 2, 486-496.