Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Biodiversità ed Evoluzione

Ciclo XXVI

Settore Concorsuale di afferenza: 05/B1 Zoologia e Antropologia

Settore Scientifico disciplinare: BIO/08 Antropologia

EVOLUTIONARY GENETICS OF LACTASE PERSISTENCE IN EURASIAN HUMAN POPULATIONS

Presentata da: **Sara De Fanti**

Coordinatore Dottorato

Prof. Barbara Mantovani

Relatore

Prof.ssa Donata Luiselli

Correlatore

Dott. Marco Sazzini

Esame finale anno 2014

Table of contents

1. Introduction **5**

1.1 Human Genome Variability **5**

1.1.1 Human Genome Variability 5

1.1.2 Genetic markers 11

1.1.3 Linkage Disequilibrium and Haplotypes 14

1.1.4 Processes Shaping Diversity 16

1.2 Lactase persistence **19**

1.2.1 The LCT gene 19

1.2.2 Lactase Persistence 20

1.2.3 Identification of lactase phenotypes 22

1.2.4 Hypotheses for lactase persistence 23

1.2.5 Polymorphisms involved in lactase persistence phenotypes 25

1.2.6 Lactase persistence: archaeological evidence 30

2. Aim of The Study **35**

3. Materials and Methods **39**

3.1 Population Samples **39**

3.1.1 Italian samples 39

3.1.2 Samples from Arabian Peninsula 39

3.2 SNPs Selection **40**

3.3 SNPs Genotyping and Quality Control **42**

3.3.1 PicoGreen Quantification 42

3.3.2 Multiplex PCR and SNPs Design 44

3.3.3 The Sequenom MassARRAY iPLEX Platform 45

3.3.4 Quality Control 47

3.4 Statistical Analyses **48**

3.4.1 Linkage Disequilibrium Analysis 48

3.4.2 Allele Frequency Analysis 49

3.4.3 Haplotypes Reconstruction 50

3.4.4 Summary and population differentiation statistics 50

3.4.5 LD-based SNP pruning 51

3.4.5.1 Multivariate analyses 52

3.4.6 Analysis of the Molecular Variance 53

3.4.7 Phylogenetic Methods and Networks 53

3.4.8 Correlation Analysis 56

4. Results 57

4.1 Quality Control **57**

4.1.1 Italian samples 57

4.1.2 Samples from Arabian Peninsula 57

4.2 Linkage Disequilibrium Analysis **59**

4.2.1 Italian samples 59

4.2.2 Samples from Arabian Peninsula 60

4.3 Allele Frequency Analysis **62**

4.3.1 Italian samples 62

4.3.2 Samples from Arabian Peninsula 67

4.4 Haplotype Reconstruction and Phylogenetic Analysis **74**

4.4.1 Italian samples 74

4.4.2 Samples from Arabian Peninsula 76

4.5 Summary statistics **79**

4.5.1 Italian samples 79

4.5.2 Samples from Arabian Peninsula 80

4.6 Population structure analyses **80**

4.6.1 Italian samples 80

4.6.2 Samples from Arabian Peninsula 85

4.7 Analysis of the Molecular Variance **89**

4.7.1 Italian samples 89

4.7.2 Samples from Arabian Peninsula 90

4.8 Correlation analyses **91**

5. Discussions and Concluding Remarks **93**

6. References **105**

Appendix **117**

1. Introduction

1.1 Human Genome Variability

1.1.1 Human Genome Variability

The present state of knowledge about human evolutionary history results from significant contributions of different research areas, such as Palaeontology, Archaeology, History, Biological Anthropology, Linguistics and, since more recent times, Anthropological Genetics, with each of these disciplines playing a fundamental role in the reconstruction of our past (Jobling et al. 2014).

In fact, information about current human biological and cultural variation make possible to investigate the origin of *Homo sapiens* diversity itself and the events that created this diversity during the course of our evolutionary history. Accordingly, it has to be borne in mind that genetic tools and models alone cannot completely explain this diversity, because many other factors have influenced and still influence human phenotypes.

Nevertheless, the human genome actually represents an amazing source of information about the natural history of our species. It is composed by approximately three billion of base pairs, residing in 23 pairs of chromosomes located within the nucleus of cells and containing the great majority of human hereditary information. The haploid human genome contains around 20,000 protein-coding genes, but their DNA sequences account for only a very small fraction of the genome, approximately the 2% of the whole human DNA. Therefore, more than 98% of the human genome does not encode for proteins, but it consists of introns and intergenic DNA (Jobling et al. 2014).

In a diploid set of chromosomes every gene is in double copy and this make possible the presence of two alternative forms of alleles that have a specific distribution and frequency in the different worldwide populations. Accordingly, the allele frequency is one of the most used parameter to represent the amount of genetic diversity within human groups and is described as the proportion of a particular allele among all the considered allele copies and it can be expressed as a percentage or as a value raging between 0 and 1 (Strachan & Read 2011).

Moreover, also the study of genomes of our close relatives (i.e. primates) has been proved to be informative about human evolutionary history. For instance, humans and chimps are known to have originated from a common ancestor and scientists believe that they diverged about six-seven million years ago. However, recent studies conducted comparing human and chimp genomes underlined that the values of diversity between the two species are lower than those reported after the sequencing of the whole chimpanzee genome in 2005 (Chimpanzee Sequencing and Analysis Consortium 2005). In particular, the human genome has found to contain 6.4% of all genes that do not have orthologs in the chimpanzee genome (Demuth et al. 2006). This difference is similar to the proportion of large duplicated regions that are unique to each species (2.7%) (Cheng et al. 2005), as well as to the estimates of divergence that consider both smaller insertions and deletions (5%) (Britten 2002). In 2012, Tomkins calculated that autosomal similarity between chimp and humans was on average 70.7% with a range of 66.1% to 77.9%, depending on the examined chromosome. Accordingly, at the genome-wide level, only 70% of the chimpanzee DNA was actually similar to the human one (Tomkins 2012). Moreover, a recent study conducted both on bonobo and chimpanzee genomes showed that more than 3% of the human genome is more closely related to either the two African Apes than these are to each other. In addition, about 25% of human genes contain parts that are more closely related to one of the two apes (Prüfer et al. 2012).

As expected, similarities at the nucleotide level are even more consistent within the human species, with single individuals differing approximately at only one out of every 500-1,000 nucleotide positions (Jobling et al. 2014), although few of these differences are those that actually make us unique (Tishkoff and Verrelli 2003).

Genetic diversity in human populations is the result of a relatively recent biological and cultural evolution, so that most of the human genetic background can be observed in all the regions of the world, compatibly with a recent origin of anatomically modern humans. In particular, *Homo sapiens* is thought to have originated in Africa, subsequently spreading and diversifying throughout the rest of the world, as proposed by the "Out of Africa" model (Stoneking 2008). A lot of evidences from fossils support this model, suggesting that all non-African human populations descend from an anatomically modern *Homo sapiens* ancestor that

evolved in Africa approximately 100-200 thousand years ago (Kya) and then colonized the remaining landmasses.

To further test this hypothesis, a lot of studies were conducted to explore modern human variability compared with that of archaic humans and a really recent work has reported the first high-quality Neanderthal genome sequence. This study described the first clear evidence for interbreeding between Neanderthals and modern humans and showed that several gene flow events occurred among Neanderthals, Denisovans and early modern humans, possibly including also gene flow into Denisovans from a still unknown archaic group (Prüfer et al. 2014) (Figure 1.1.1.1).

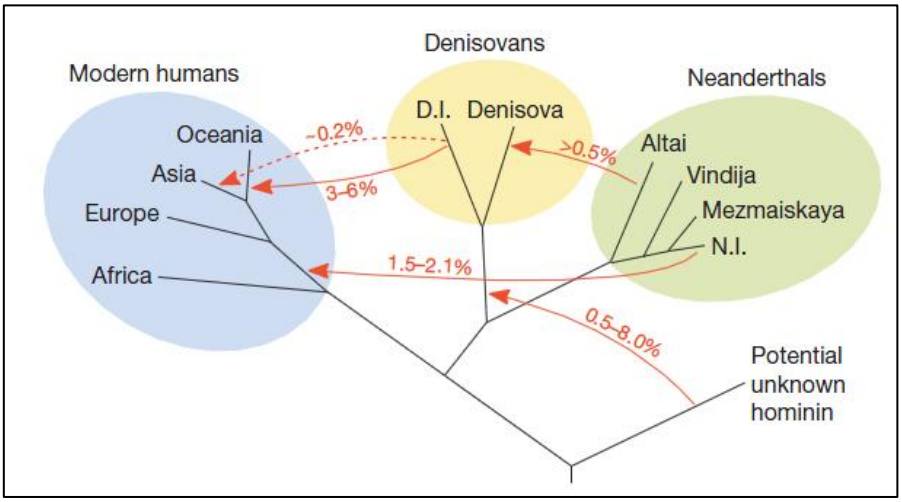


Figure 1.1.1.1: A possible model of gene flow events in the Late Pleistocene (Prüfer et al. 2014).

Surveys of mitochondrial DNA (mtDNA), Y-chromosome and various types of autosomal polymorphisms have all shown that the most human genetic diversity is found within, rather between, populations (Jorde et al. 2000; Jobling et al. 2014). On average, 83–88% of autosomal variation is found within populations and approximately 9–13% between continental groups. However, it is important to consider that such values strongly depend on the frequency of the examined polymorphisms, and would be even lower if rare variants were investigated. Data from mtDNA and the Y chromosome are somewhat different to those estimated from autosomes, with less of the variation within populations and more between

groups, plausibly according to a reduced effective population size that mainly expose these loci to the action of genetic drift (Jobling et al. 2014).

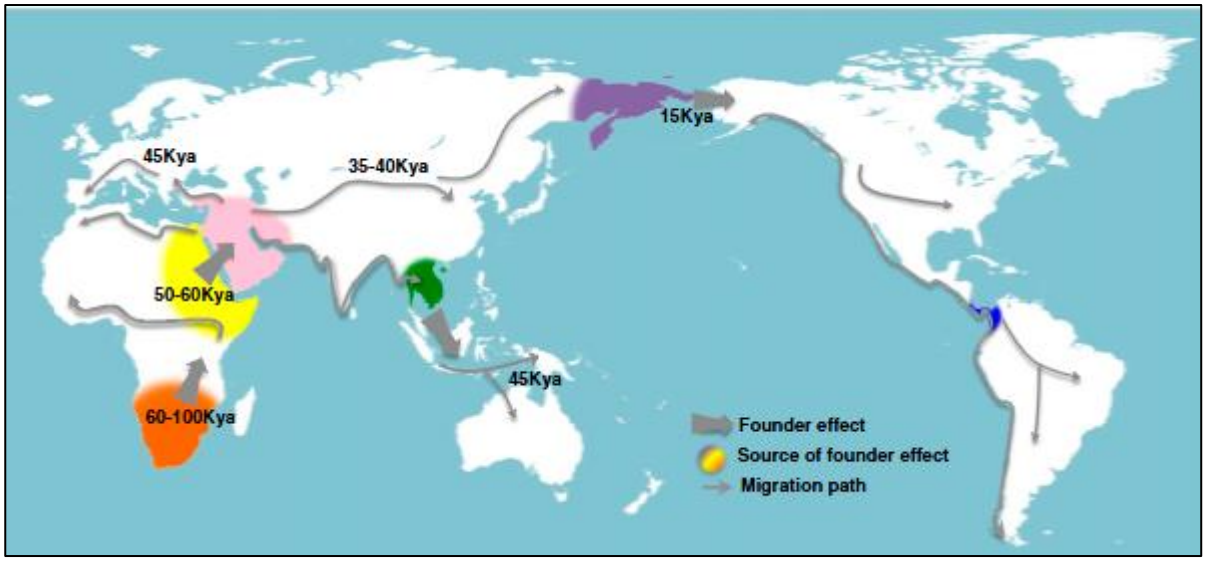


Figure 1.1.1.2: Dispersal patterns of modern humans during the past 100,000 years. (Henn et al. 2012).

For instance, the larger genetic diversity observed in present-day African populations with respect to any other continent is capable to explain phylogenetic trees in which a first deep bifurcation separates Africans from ethnic groups belonging to the other continents (Jobling et al 2014). Therefore, in non-African populations a subset of the genetic diversity present in modern Africans was observed, but the levels of diversity in non-African populations also depend on the severity of demographic bottlenecks occurred during migration out of Africa (Henn et al. 2012) (Figure 1.1.1.2 and 1.1.1.3).

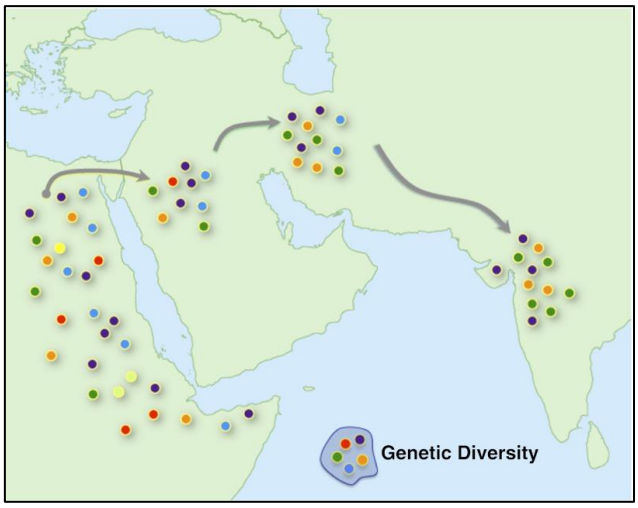


Figure 1.1.1.3: Effect of serial founder events on genetic diversity in the context of the OOA expansion. (Henn et al. 2012).

The ever-increasing new knowledge about the human genome is making the study of relationships between genotypes and phenotypes one of the central goals in present-day biology and medicine. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the human genome and only in 2003 the Consortium was able to complete and convert the draft in an accurate ad almost complete sequence that provided a foundation for the study of human genetics (International Human Genome Sequencing Consortium 2004). This actually represented a really important step in human genetics studies, but it has to be considered that a systematic investigation of human genetic and genomic diversity requires full knowledge of DNA sequence variation across the entire spectrum of allele frequencies and types of DNA differences. By 2008, the public catalogue of variant sites dbSNP contained approximately 11 million single nucleotide polymorphisms (SNPs) and 3 million short insertions and deletions (INDELs) (International HAPMAP Consortium 2007).

However, the *1,000 Genomes Project* has recently represented the most ambitious and innovative effort in the field of human genetics, being an international research aimed at creating the most detailed catalogue of human genetic variation. In this database, there is up to date about a thousand of genomes collected from anonymous participants of different ethnic groups. The *1,000 Genomes Project* was launched in January 2008 and it is the first project aimed at sequencing the whole genomes of a large number of people (1000 Genomes Project 2010).

The project unites multidisciplinary research teams from institutes around the world, including China, Italy, Japan, Kenya, Nigeria, Peru, the United Kingdom, and the United States (www.1000genomes.org) (Figure 1.1.1.4).

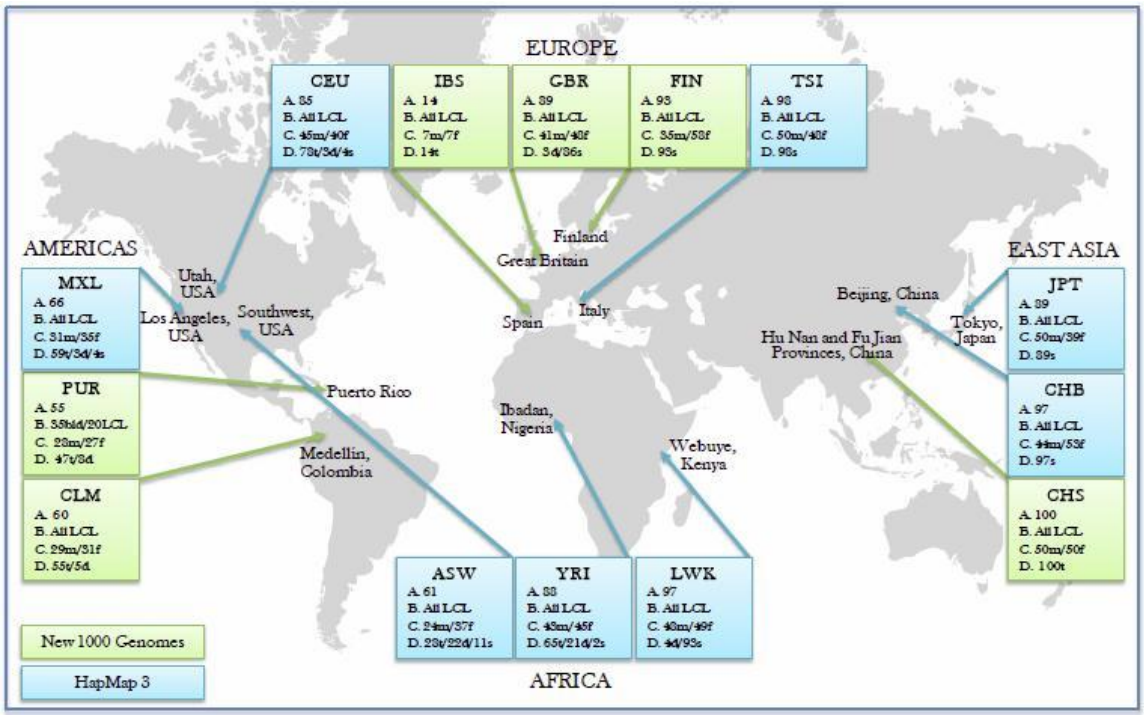


Figure 1.1.1.4: Populations included in the "1000 Genomes Project" (www.ub.edu).

The first aim of this project is to create a complete and detailed catalogue of human genetic variation, which can be used for both association studies relating genetic variation to diseases and Anthropological Genetics studies aimed at reconstructing the genetic history of human populations and the mechanisms that have enable them to adapt to very different environments. The technical goal of the *1,000 Genomes Project* is that of discovering over 95% of the variants, such as SNPs, Copy Number Variants (CNVs) and INDELs with minor allele frequencies (MAF) as low as 1% across the genome and ranging from 0.1 to 0.5% in gene regions. Furthermore, the purpose of the project is to estimate the population frequencies, haplotype backgrounds and linkage disequilibrium (LD) patterns of variant alleles, using newly developed sequencing technologies and computational approaches, which are faster and less expensive with respect to traditional ones.

The second goal of the project will include the support of better SNPs and probe selection for genotyping platforms in order to overcome the ascertainment bias

towards common European variants affecting SNP chips currently available and the improvement of the human reference sequence. Furthermore, the completed database will be a useful tool for studying regions under selection, variation in multiple populations and understanding the underlying processes of mutation and recombination (1000 Genomes Project 2012).

The development of projects such as the *1,000 Genomes* one are now possible thanks to the possibility of sequencing whole-genomes with cheaper and more routinely approaches. A so large bulk of data promises to address several questions of evolutionary interest that the analyses of uniparentally inherited markers may not completely explain and makes also possible to reconstruct ancestral information, as well as in admix populations in which mtDNA and the Y chromosome from one ancestral source have been entirely lost by genetic drift (Colonna et al. 2011) (Figure 1.1.1.5).

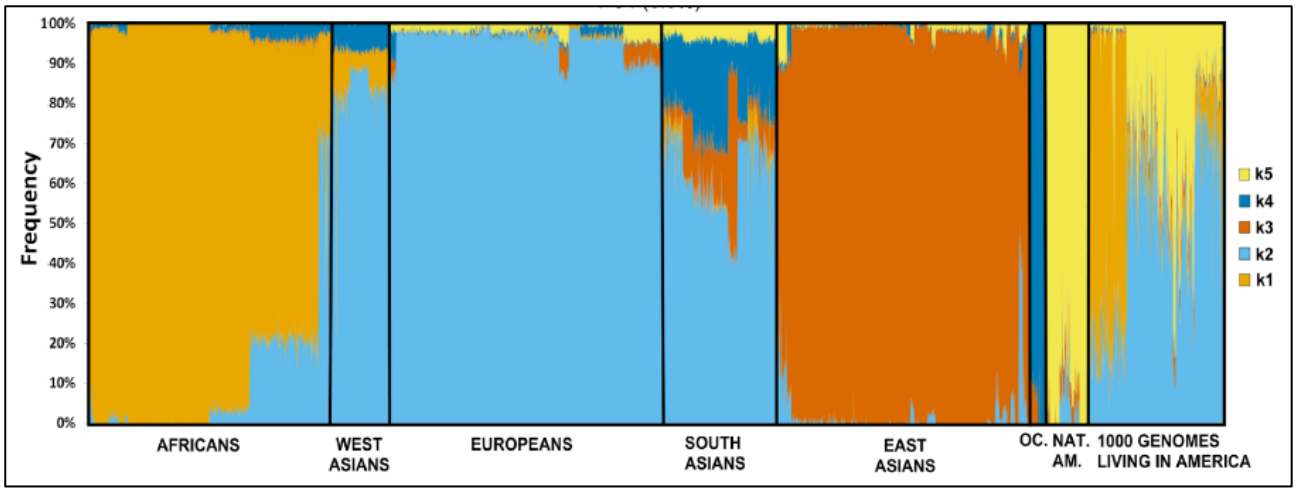


Figure 1.1.1.5: STRUCTURE-like plot, where each thin vertical line represents an individual with inferred ancestry from five clusters (Colonna et al. 2011).

1.1.2 Genetic markers

Overall, mutation has been proved to be the ultimate source of all genetic variation and it is defined as any change in the DNA sequence that may have visible effects on the phenotype (i.e. non-synonymous mutations) or that may have no effects on it (i.e. neutral mutations). Accordingly, neutral mutations do not affect the evolutionary fitness of people who carry them; hence their frequency is unaffected by natural selection.

The result of a mutation is a new gene that differs little from the one belonging to the previous generation: the two types are called alleles of a gene. When there is the presence of two or more alleles of a gene in a population, this gene is said to be polymorphic. Polymorphism occurs when a gene exists in at least two alleles and the MAF is at least 1%.

Polymorphisms can show up in multiple, not preserved forms; therefore they are mostly found in non-coding regions (e.g. gene introns). In fact, the genes are polymorphic because they are in an intermediate stage of the process, between the appearance of a mutation and its probable ultimate fate, which is its fixation or its extinction. The polymorphisms can be thus used as indicators of the variation of specific chromosomal segments; hence, they are considered genetic markers and they are the key to the understanding and the measuring of the genetic variation (Cavalli-Sforza 1994).

The possibility to accurately investigate the sequence and structural variation of the human genome has revealed the presence of several types of genetic markers, such as SNPs, short insertion/deletion (INDELs), Short Tandem Repeat (STRs, microsatellites) and minisatellites, that could be used to explore patterns of human genetic and genomic diversity.

For instance, a SNP is generated by a nucleotide substitution, which is the simplest difference between two homologous DNA sequences. The base substitution can be a transition when a pyrimidine base is exchanged for another pyrimidine, or a transversion when a pyrimidine is exchanged for a purine or vice-versa. The insertion or deletion of a single base is also included in the category of SNPs maybe unfortunately, since the mechanisms which underlie these INDELs, and the analytical treatment of them, strongly differ from those for base substitutions (Jobling et al. 2014).

Variable Number of Tandem Repeats loci (VNTRs) are changes in the number of repeated DNA sequences arranged in tandem arrays. They are classified according to the size of their repeat units. Microsatellites, also known as STRs, which are tandem arrays of repeat units 1-6 bp in length, and minisatellites consisting of repeats units from about 8 to 100 bp (Jobling et al. 2014).

Moreover, the first Biological Anthropology studies were performed contextually to research and analysis of genetic markers that are defined “classic” today, such as

blood groups. These were based exclusively on the analysis of the phenotype (i.e. the proteins encoded by the underlying genes) and thus indirectly on the analysis of mutations occurring in coding regions (Ogasawara et al. 1996a, Ogasawara et al. 1996b, Blancher et al. 1997).

The different types of mutations described above can be used as *molecular genetic markers* and they can be examined in autosomal chromosomes, Y-chromosome and mitochondrial DNA. In fact, mutation rates of such markers are very variable; SNPs and INDELs are also named "slow evolving markers". These polymorphisms are said to be biallelic or unique event polymorphisms because they have very low mutation rate (i.e. 10^{-9}). Thus, the event of mutation generally occurs only once in the course of the evolution of a chromosome. Regarding INDELs, the mutation rate is even lower (i.e. 10^{-11}), so it is possible to exclude the presence of recurrent alleles and reverse mutation (i.e. homoplasy), accordingly, every time a mutation happens, a new allele is generated. STRs and minisatellites have instead faster mutation rates (i.e. 10^{-2}-10^{-3}), and they are highly variable. Some alleles are recurring, so there may be events of homoplasy. The homoplasy creates the same patterns due to recurrent mutations and recombination events, which can have similar effects in generating variability, but do not imply the identity by descend of the examined DNA sequences (Jobling et al. 2014). Anthropological Genetics studies usually exploit differences in the mutation rates of the mentioned genetic markers to perform analyses at substantially different evolutionary scales. According to this, they are able to infer both the ancient and recent evolutionary and demographic processes occurred in human populations. As a matter of fact, SNPs turn out to be very useful for resolving deep genealogical clades, whereas STRs succeed to discriminate more recent evolutionary events, even with informative comparisons of closely related individuals (Tishkoff and Verrelli 2003).

However, among all the existing markers, SNPs represent the most striking ones for population genetics studies. In fact, they have high abundance and a relatively low mutation rate, they may be functionally relevant and are easily adapted to automated typing. In particular, SNPs represent the first dense genome-wide markers and as such their analysis has raised many challenges and insights relevant to the study of population genetics with whole-genome sequences (Novembre and Ramachandran 2011).

The mitochondrial DNA and the Y-chromosome genes are finally studied as non-recombinant loci (i.e. haploid) and they thus are lineage markers, because Y-chromosome is solely passed along the patrilineal line, from father to sons, whereas the mtDNA is passed along the matrilineal one, from mother to offspring of both sexes. Y-DNA and mtDNA are assumed to do not recombine, thus they change only by random mutations at each generation with no intermixture between parents' genetic material (Jobling et al. 2014).

Only by means of autosomal chromosomes, it is instead possible to study recombinant loci (i.e. diploid) because in every generation the event of crossing over occurs.

1.1.3 Linkage Disequilibrium and Haplotypes

In general, any class of DNA polymorphism can define a haplotype, which refers to the combination of allelic state of polymorphic markers along the same DNA molecule. In other words, they lay on the same chromosome or mtDNA (Jobling et al. 2014).

The tendency of particular alleles at separate loci to be co-inherited because of reduced recombination between them can lead to associations between alleles in a given population. This property is known as linkage disequilibrium (LD). The hope that genome-wide surveys might identify associations between anonymous markers and genetic variation contributing to common disorders has resulted in intense interest in genome-wide patterns of LD (Sabatti and Risch 2002; Conrad et al. 2006). In fact, LD describes the relationship between genotypes at a pair of polymorphic sites and depicting patterns of LD between polymorphic sites makes possible the identification of well-defined “haplotype blocks” along the human genome. These “haplotype blocks” represent genomic regions that are inherited without substantial recombination, which is generally produced by the molecular mechanisms of sexual reproduction, in the ancestors of the current examined population. One of the consequences for the presence of different dominant haplotypes between populations is the possibility that patterns of LD will differ between these populations groups (Jakobsson et al. 2008).

Several measures exist to describe LD, the two most frequently used are D’ and r^2. D’ measures the deviation of the frequency observed for a haplotype from the

expected frequency. D' ranges from –1 to 1 and if D'=1 or D'=-1 it means that no evidence of recombination between the markers was found, so that a condition of complete LD was described (Gabriel et al 2002). The r^2 ranges instead from 0 to 1 and strongly depends on the markers allele frequencies. As a matter of fact, r^2 equals 1 if, and only if, MAFs at the two loci perfectly match and the minor alleles tend to co-occur. Moreover, several empirical studies have demonstrated that the extent and conservation of different mosaic pieces depend not only on recombination rates, and thus on the potential presence of recombination hot spots, but also on mutation rates, population size and the action of natural selection on the investigated genomic regions (Gabriel et al 2002).

Haplotype analyses indicated that levels and patterns of LD are different in African respect to non-African populations. In particular, African populations typically have higher levels of genetic diversity, a complex population substructure, and low LD levels relative to non-African ones. The divergent pattern of LD in non-Africans relative to Africans is likely the result of a founding event during expansion of modern humans out of Africa within the past 100,000 years. When anatomically modern humans migrated out of Africa, they carried with them only a part of the genetic variation that existed in the ancestral population. LD established after a founding event is likely maintained in rapidly expanding non-African populations. As a result, the non-African haplotypes tend to be subsets of the African ones. Furthermore, the haplotypes in non-African populations tend to be longer than in African ones. Ancestral African populations have lower LD because more time has passed for recombination to cause LD to decay and because they have maintained a larger effective population size compared to non-Africans (Campbell and Tishkoff 2008).

Studies based on haplotypes are considered one of the most important approaches to explore the genetic bases of common complex diseases thanks to the possibility to analyse only a limited number of SNPs that is sufficient to unambiguously distinguish the haplotypes. In fact, by focusing on haplotype tagging SNPs it is possible to identify the minimum set of SNPs (i.e. htSNPs) tagging an arbitrary set of haplotypes that are actually able to describe the existing variability with no information loss (Sebastiani et al. 2003). An interesting observation using htSNPs is that the vast majority (95%) of the htSNPs in European-American samples is

actually a subset of the htSNPs of African-American groups. This result seems to provide further evidence that a severe bottleneck happened during the founding of European populations and the inferred ‘‘Out of Africa’’ event (Sebastiani et al. 2003) (Figure 1.1.3.1).

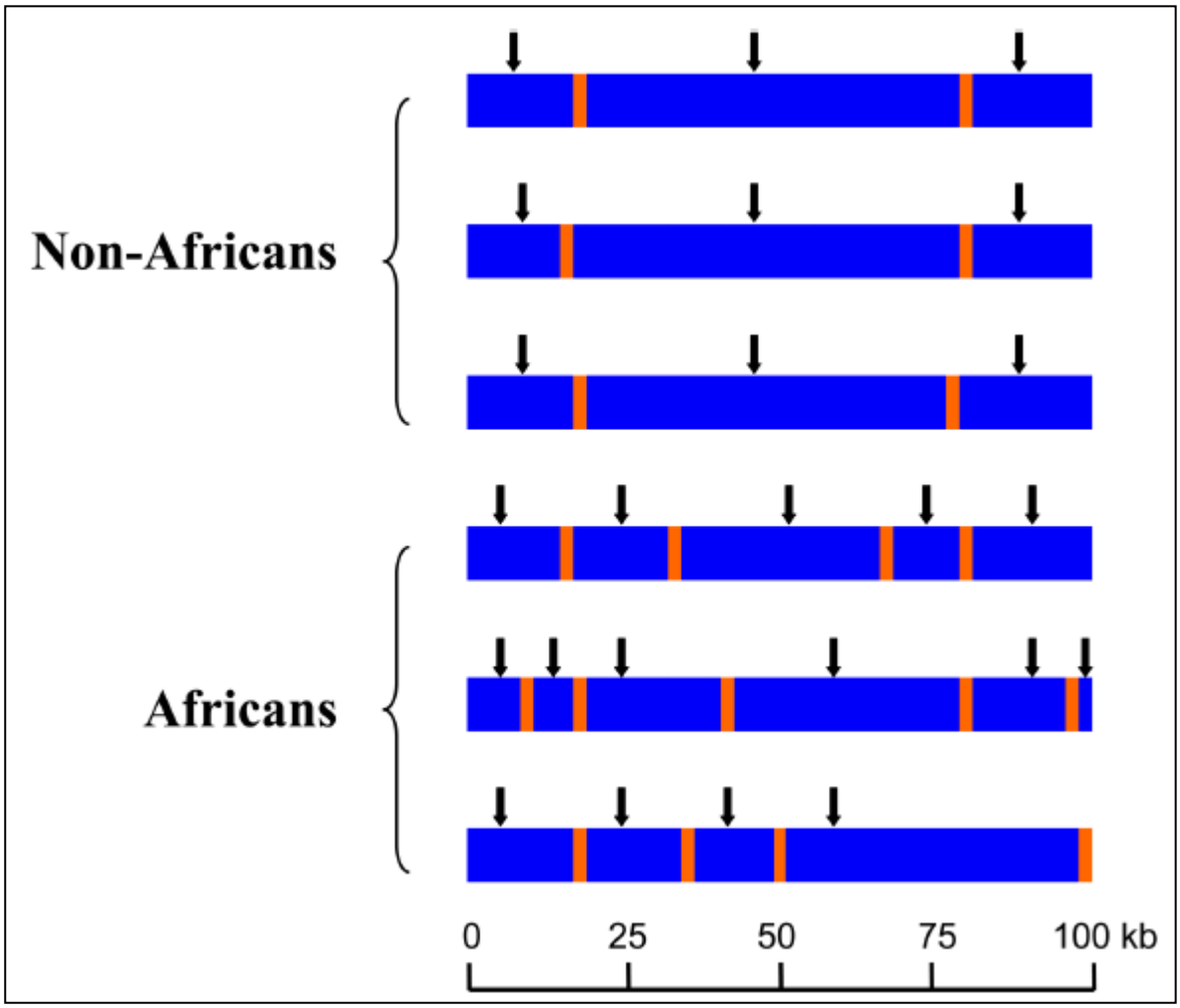


Figure 1.1.3.1: An example in LD block differences between Africans and Non-Afrincans, vertical lines indicate SNPs and vertical arrows indicate haplotype tag SNPs (htSNPs) (Campbell and Tishkoff 2008).

1.1.4 Processes Shaping Diversity

Macro-evolutionary studies focus on change that occurs at, or above, the level of species, in contrast with micro-evolutionary processes can be measured by changes in allele frequencies and diversity patterns within the populations (Reznick and Ricklefs 2009). Micro-evolutionary processes indeed create variability within species and are represented by event such as mutation, gene flow, genetic drift and natural selection.

A population that shows allele and genotype frequencies constant from one generation to another is defined to be at the Hardy-Weinberg equilibrium (HWE) and this indicate that is actually not evolving. An idealized population in HWE must have certain properties including: random mating, infinite number of individuals, isolation, absence of mutation, mortality and fertility independent of the genotype. The micro-evolutionary forces act at the opposite with respect to these conditions and are respectively: inbreeding, genetic drift, migration, mutation and natural selection. Actually, no population is infinitely large and each generation represents a finite sample from the previous one, so variation in allele frequency between generations can occur solely through a stochastic process of sampling, known as genetic drift (Jobling et al. 2014). The effects of genetic drift on variation are a function of the effective population size (N_e), whereby large effective populations can maintain higher levels of variation and small effective populations are more subject to random fluctuations in allele frequencies. Other two important processes that shape genetic diversity are founder effect and bottlenecks. The former is related to the process of colonization and the genetic separation of a subset of the diversity present within the source population, on the other hand bottlenecks are referred to the reduction in size of a single, previously large, population and the loss of a prior diversity (Jobling et al. 2014).

However, the genome of *Homo sapiens*, like those of all other species, has been shaped also by natural selection. In fact, this is the evolutionary force that drives the adaptation of one organism in the habitat where it lives. This force does not act as a random process because it increases only the frequency of alleles or genotypes that upturn the reproductive success and the survival of selected individuals. Furthermore, selection operates an adaptation that represents a cumulative process. Understanding the traits, as well as the genes underlying them, that have undergone natural selection during human evolution can provide insight into the events that have shaped our species, as well as into many of the diseases that affect present-day societies (Sabeti et al. 2006).

Several recent studies were aimed at detecting loci under natural selection without a prior knowledge, being based on genome-wide genotyping data. These methods can be used equally well to detect selection in non-coding and protein-coding

regions, but the results are usually interpreted in terms of predictions for protein-coding regions, because most functional annotation is focused on genes.

In addition to classical neutrality test based on the analysis of the site frequency spectrum, the development of relative extended haplotype homozygosity (rEHH) and integrated haplotype score (iHS) tests for incomplete selective sweeps allow to identify selection when a high-frequency haplotype with little intra-allelic variability is observed (Nielsen et al. 2010; Liu 2013). These methods have made possible to identify for the first time in particular two genes that turned out to be the most striking examples of ongoing selective sweeps in the human genome: *LCT* and *G6PD* (Voight et al .2006; Nielsen et al. 2010).

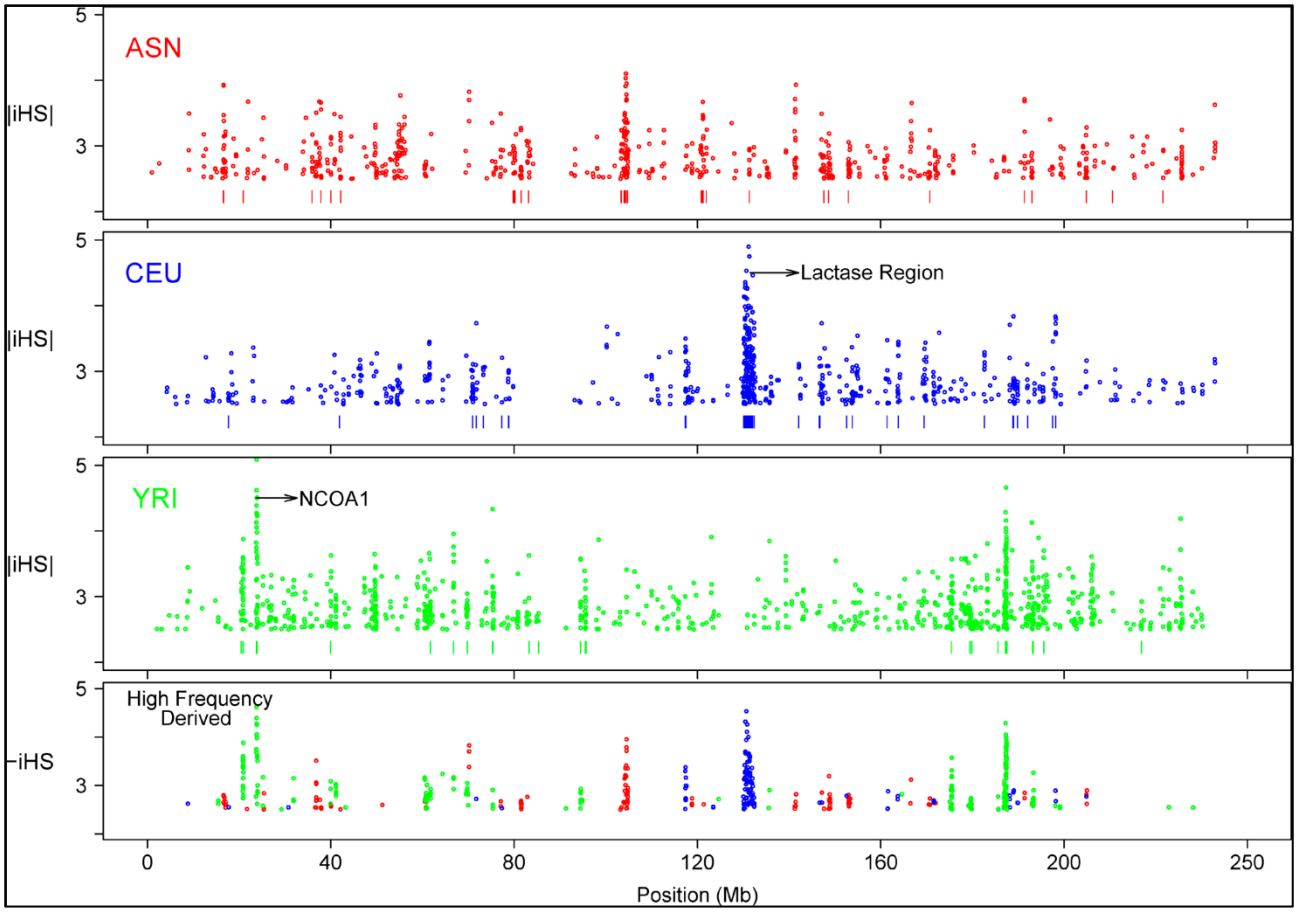


Figure 1.1.4.1: An example of a plots of chromosome 2 SNPs with extreme iHS values indicate discrete clusters of signals (Voight et al. 2006).

1.2 Lactase persistence

1.2.1 The LCT gene

The main sugar in milk is the disaccharide lactose and its digestion is possible only after the hydrolysis of the disaccharide in its two monosaccharides: galactose and glucose (Swallow 2003) (Figure 1.2.1.1).

The enzyme that hydrolyzes lactose in its constituent monomers is lactase-phlorizin hydrolase (LPH), an enzyme belonging to the β-galactosidase family. It is a large glycoprotein, with two active sites, that can catalyze the hydrolysis of a variety of glucosides, including phlorizin, flavonoid glucosides, pyridoxine-5'-D glucoside and galactosides, in addition to lactose.

Monomers formed in this way easily reach the bloodstream by passing through Na-K pumps. Lactase, in fact, is expressed on the apical surface of brush border enterocytes where it is anchored into the membrane by its C-terminal end, with the bulk of the molecule projecting into the lumen of the gut on the differentiated enterocytes lining the villi of the small intestine (Swallow et al. 2003).

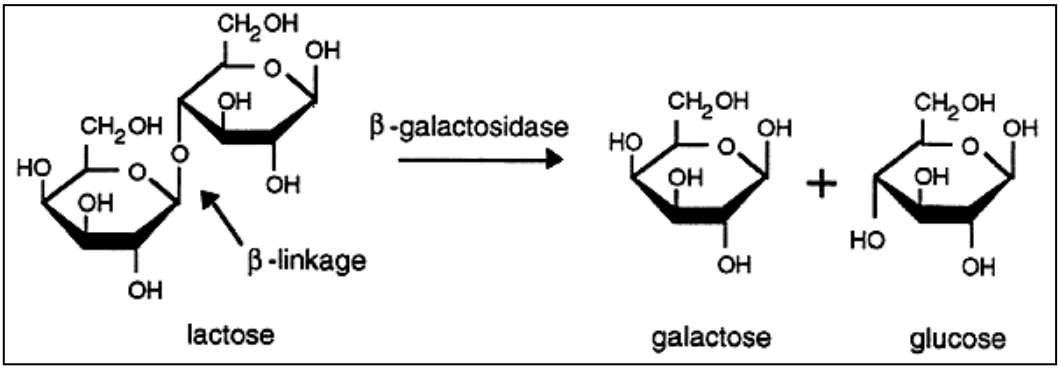


Figure 1.2.1.1: Lactase hydrolyses: digestion of the sugar lactose into two simple sugars, glucose and galactose.

In humans, lactase is encoded by a single genetic locus on the q arm of autosomal chromosome 2, the *LCT* gene, which spans approximately 50 kb. This gene has 17 exons and encodes an mRNA transcript of 6,274 nt (Genbank X07994) and a pre-protein of 1,927 aminoacid residues. This is composed of a putative signal peptide of 19 aminoacid residues, a large pro-portion of 849 amino acids and a mature protein that contains two catalytic sites, as well as, at the C-terminal end, a membrane-spanning domain and a short cytoplasmic domain. The fourfold internal

homology shown by *LCT* suggests that it arose by two duplication events. Pro-lactase is proteolytically processed to a smaller protein, and two of the four homologous regions occur in the cleaved pro-portion of the molecule, which does not have a catalytic function, but probably has a chaperone function, in that it seems to play a role in transporting the molecule to the cell surface. The expression of lactase is limited to the enterocytes or absorptive cells of the small intestine with the highest level of expression observed in the mid-jejunum. This pattern of expression closely parallels that of another digestive hydrolase, the sucrase-isomaltase (Jacob et al. 2002).

The promoter of the *LCT* gene is included within introns of the nearly located *MCM6* gene. *MCM6* contains two of the SNPs occurring on regulatory regions for *LCT*, which are located in two of the *MCM6* introns, approximately 14 kb (-13,910) and 22 kb (-22,018) upstream of *LCT*. In particular, the -13,910 locus has been shown to function *in vitro* as an enhancer element, capable of differentially activating the transcription of the *LCT* promoter (Olds & Sibley 2003) (Figure 1.2.1.2).

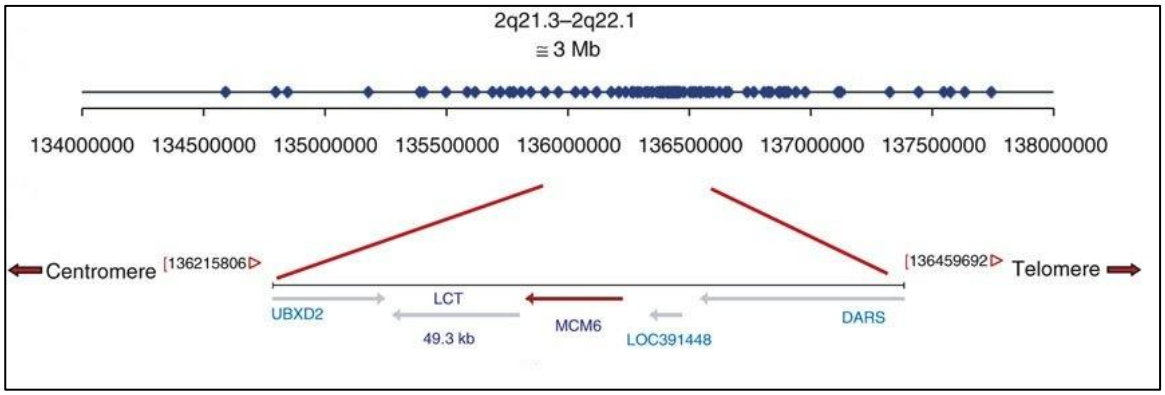


Figure 1.2.1.2: The LCT gene.

1.2.2 Lactase Persistence

During the fetal life lactase is expressed only at low levels and lactase activity is obviously elevated and of vital importance during early childhood, when milk is the main source of nutrition. In most mammals, however, lactase activity declines after the weaning period, and this is also the case in the majority of humans throughout the world who are described as lactase non-persistent subjects.

If there is no lactase enzyme activity in the small intestine, lactose is permuted by bacterial action to lactic acid, hydrogen and carbon dioxide. Being a large molecule,

when it is in the gut, lactose creates an osmotic effect and draws water, causing side effects typical of lactose intolerance: diarrhoea, bloating and flatulence (Järvelä et al. 2009).

In lactase non-persistent adults, the reduction of lactase activity and the onset of lactose intolerance usually begin between 2 and 3 years, being completed by the age of 5 to 10 years (Swallow et al. 2003).

However, in some individuals, lactase activity persists at a high level throughout adult life and this trait is known as lactase persistence (LP). People who are lactase persistent can usually hydrolyze large amounts of lactose and can thus consume large quantities of fresh milk without gastrointestinal complication. People with lactase non persistence (also referred to as adult-type hypolactasia or lactase restriction) have really lower lactose digestion capacity than those with lactase persistence and thus often, but not always, show symptoms of lactose intolerance after consumption of fresh milk.

Studies on the prevalence of the LP phenotype in worldwide human populations have shown that the frequency of this trait is highly variable in different ethnic groups and appears to be positively correlated with the importance of milk in the diet (Swallow et al. 2003) (Figure 1.2.2.1).

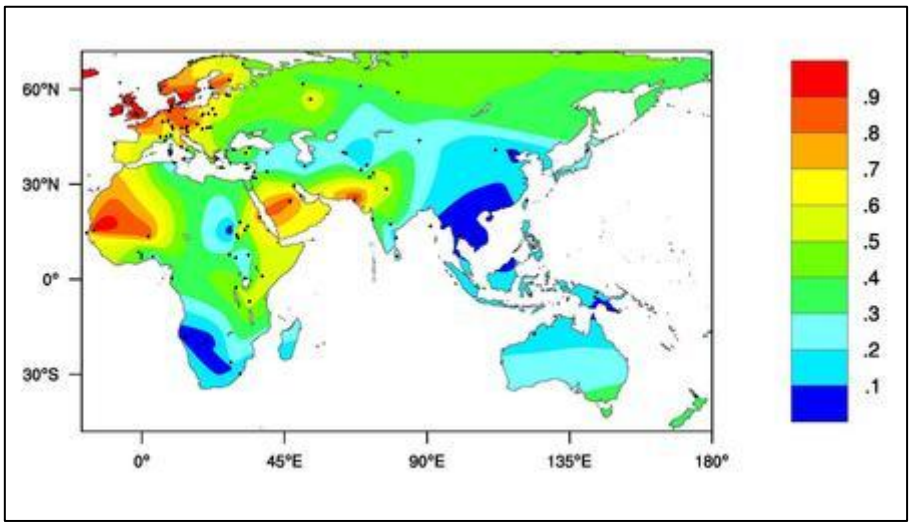


Figure 1.2.2.1: Maps of the LP phenotype distribution.

The frequency of lactase persistence is high in Northern European populations (e.g. greater than 90% in Swedes and Danes), decreases in frequency across Southern Europe and the Middle East (e.g. being around 50% in Spanish, French and

pastoralist Arab populations) and is instead low in non-pastoralist Asian and African populations, for instance reaching a frequency of barely 1% in Chinese and 5%–20% in West African agriculturalists. Notably, lactase persistence is common in pastoralist populations from Africa, which show frequency of around 90% (Tutsi) and 50% (Fulani) (Tishkoff et al. 2007).

1.2.3 Identification of lactase phenotypes

The low lactase activity in baby mammals with respect to adults was known since the end of the nineteenth century, but it was long believed that all adult humans had high levels of lactase, probably because early research was conducted in countries where lactose tolerance was the most frequent phenotype. The first examples of lactase non-persistent individuals were therefore considered as subjects carrying an abnormal trait that was described as lactase deficiency. It was soon recognized that this supposed abnormality was the most frequent trait worldwide (Swallow et al. 2003).

Currently, the adult lactase phenotypes can be determined directly by assaying lactase by means of a small-intestinal biopsy or, indirectly, by lactose-tolerance tests, although the discriminatory power of these tests is variable (Swallow et al. 2003). For instance, one indirect method to determine lactase non-persistence is the blood test. Using this approach, the milk is administered to individuals and subsequently the amount of glucose present in the blood is measured. Early population and family studies recorded an increase in blood glucose after giving a lactose load of 50 g. People with high lactase activity show a significant rise in blood glucose concentration within 15 to 45 minutes after lactose administration, whereas those with low levels do not (Swallow et al 2003). However, the most used method to assay LP in the clinical way is the breath hydrogen test. In fact, in lactase non-persistent people, undigested lactose reaches the colon, where it is fermented, leading to the production of fatty acids and gases, including hydrogen, which is excreted in the breath. This property can be exploited in a more convenient and fairly reliable lactose-tolerance test, which involves testing breath hydrogen after a lactose load (Swallow et al 2003). Urinary lactose tolerance tests is a possible supplementary test and is based on an urinary galactose measurement performed after oral lactose loading with ethanol, or strip test, which is like the former, but in

which a special test strip for urinary galactose is used.

Although, direct determination of lactase activities in intestinal samples is better than lactose-tolerance tests, it is not a practical alternative for family studies. Finally, the breath hydrogen test was proved to be less sensitive (69%) than the urinary lactose tolerance tests (81-94%), which have also high specificity (96-98%) (Arola et al. 1988).

1.2.4 Hypotheses for lactase persistence

Three main hypotheses have been formulated to describe the presence of lactose persistence in worldwide human populations, being not all each other exclusive.

The first hypothesis, which is also the most reliable one, is the *gene-culture co-evolutionary hypothesis*, which states that the geographical distribution of the lactose tolerant phenotype resembles the geographical distribution of the spread of pastoralism (Enattah et al. 2002). In fact, after the domestication of animals, milk became more available, so it became a widespread drink. Individuals carrying the gene encoding the lactase persistence were advantaged and during periods of famine they could replace the missing calories thanks to milk intake (Itan et al. 2009).

On the contrary, the reverse of the previous hypothesis states that pastoralism has spread where individuals tolerant to lactose were settled. Thus, the sheep domestication would be a result of tolerance itself. This hypothesis is currently not accredited, because no significant evidences have been found to supporting it (Aoki 2001).

In addition to these, more selection hypotheses that may explain the high frequency of LP in few specific populations have been subsequently formulated. For instance, the *calcium-absorption hypothesis* emphasizes the benefits for calcium metabolism derived from milk drinking in populations living relatively close to the poles. The calcium-absorption hypothesis is based on the fact that at higher latitudes, very little vitamin D is synthesized by the skin in the presence of sunlight (Flatz and Rotthauwe 1973; Simoons 1978). Vitamin D is involved in the gut absorption of calcium. The low intake of vitamin D at higher latitudes can cause a decline in calcium absorption that leads to low serum levels of free calcium, initiating the release of parathyroid hormone, promoting skeletal resorption and eventually bone

loss or osteomalacia (Wolff et al. 2008). Accordingly, the great advantage of consuming milk proteins and lactose is actually that they could facilitate the absorption of calcium. Hence, the ability to drink fresh milk that contains both calcium and components that stimulate its uptake may have provided an advantage to LP individuals (Ingram et al. 2009). However, Vuorisalo et al. (2012) have recently shown that the LP in the Northern European countries, especially in Sweden, is the result of a migration of tolerant peoples from Central Europe, which replaced the former local hunter-gatherers populations over time. On the other hand, in Middle East and North Africa, high lactase persistence frequencies were found in pastoralist groups living in arid regions. That being so, the underlying idea is that milk was a source of water in arid environments, important to guarantee adequate hydration and the electrolyte balance (Cook 1975).

Among all these hypotheses, the *gene-culture co-evolutionary* one was strongly supported when a high diversity in cattle milk protein genes and lactase persistence distribution was demonstrated to coincide in Europe (Beja-Pereira et al. 2003) (Figure 1.2.4.1). This study, in fact, found substantial geographic coincidence between high diversity between cattle milk genes and high frequency of present-day lactase persistence in the areas where the first European Neolithic milk-dependent groups have originated more than 5,000 years ago (Zvelebil 2000).

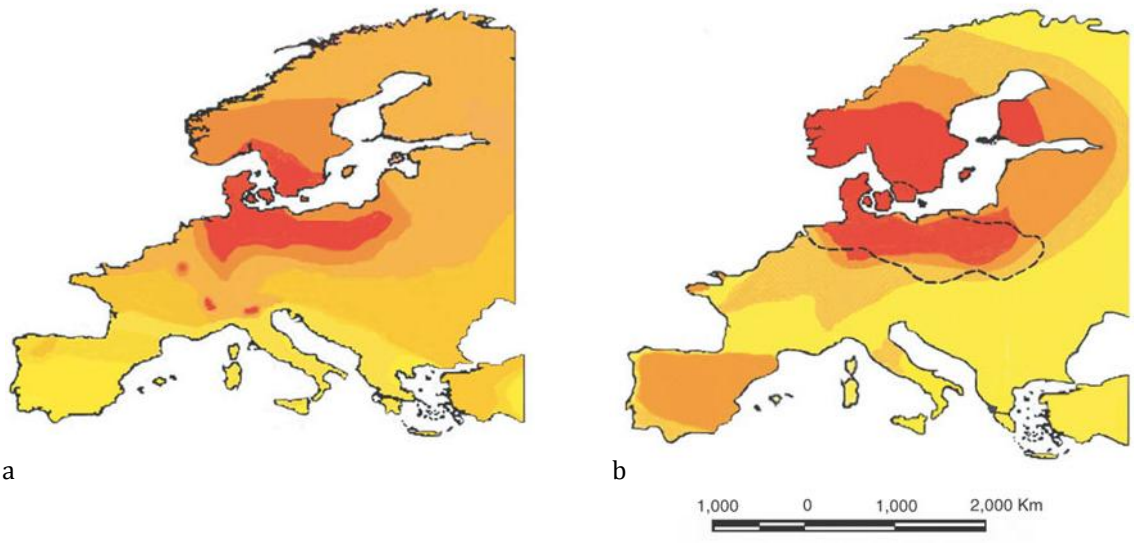


Figure 1.2.4.1: a:synthetic map showing the first principal component resulting from the allele frequencies at the cattle genes. b:geographic distribution of lactase persistence mutation in contemporary Europeans, the hotter the colour the highest the frequency (Beja-Pereira et al. 2003).

Accordingly, the *gene-culture co-evolutionary hypothesis* reliably represents a relevant example of Neolithic dietary adaptation, in fact, with the intensification of agriculture and domestication the diet of farming populations became less various in comparison with that of the hunter-gatherers, so that consuming of dairy products was important as well as to diet integration and could explain the selective advantages of LP (Losch et al. 2006).

1.2.5 Polymorphisms involved in lactase persistence phenotypes

Lactase persistence is a heritable autosomal dominant condition and it has been strongly correlated with several SNPs located 14 kb upstream of the lactase gene in different and geographically distributed populations. In fact, −13,910C/T was associated to LP in Europeans and −13,907C/G, −13,915T/G and −14,010G/C were observed in several African and Arabian pastoral populations (Enattah et al. 2002; Swallow 2003; Ingram et al. 2007; Tishkoff et al. 2007; Enattah et al. 2008;).

In 2002, Enattah et al. identified a first polymorphism located 13,910 bases upstream of the *LCT* gene that resulted fully associated with lactose tolerance in the Finnish population, being also located within an extended haplotype block of about 1 megabase. This polymorphism consists in a mutation of a C into a T nucleotide that lies at the level of the promoter of the *LCT* gene, in a binding site for the transcription factor *OCT1*. This is one example of the importance of the *cis*-acting site in determining the human phenotype and is the most important example of this event as a dietary adaptation (Jones and Swallow 2011).

In the Finnish population, tolerant individuals are homozygotes for the T allele or heterozygotes, while the intolerant ones are homozygotes for the C allele. In Finland, the most common genotypes are TT and TC, which cumulatively reach the frequency of 82.97% (Enattah et al. 2002) (Figure 1.2.5.1).

In vitro studies confirmed that the transcription factor *OCT1* binds more frequently the sequences containing the T allele (Olds & Sibley 2003). Moreover, *in vivo* studies with transgenic mice carrying a luciferase reporter gene also support a causal role for the −13,910C/T SNP (Fang et al. 2012).

Generally, in adulthood lactase continues to be produced, but not in amounts that are sufficient to digest milk, because *OCT1* binds *LCT* in a weaker way if the C allele is present.

In the Finnish population, another polymorphism located 22,018 bp upstream of the *LCT*) gene showed an appreciable association with LP, even if it was weaker than that observed for -13,910C/T (Enattah et al. 2002).

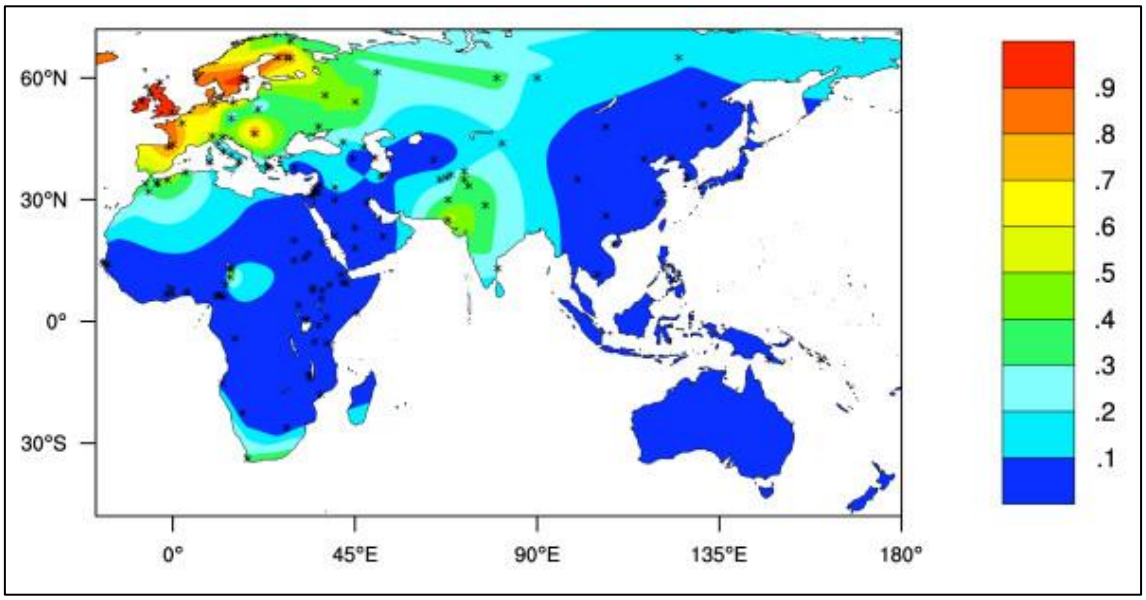


Figure 1.2.5.1: Predicted Old World LP phenotype frequencies based on -13,910 C/T allele frequency data only.

In agreement with the frequency distribution of the -13,910C/T variant, the origin of LP in Europe seems to be occurred in the North-Western part of the continent. A spatially explicit computer simulation study was indeed used to investigate the origin of lactase persistence in Europe associated with the -13,910C/T SNP by Itan and collaborators in 2009. This simulation model explored the spread of lactase persistence, dairying, other subsistence practices and unlinked genetic markers in the European and Western Asian geographic space. Accordingly, the −13,910C/T allele was demonstrated to have first underwent selection among dairying farmers between 8,683 and 6,256 years ago, in a region spanning from Central Balkans and Central Europe, possibly in association with the dissemination of the Neolithic transition (Itan et al. 2009) (Figure 1.2.5.2).

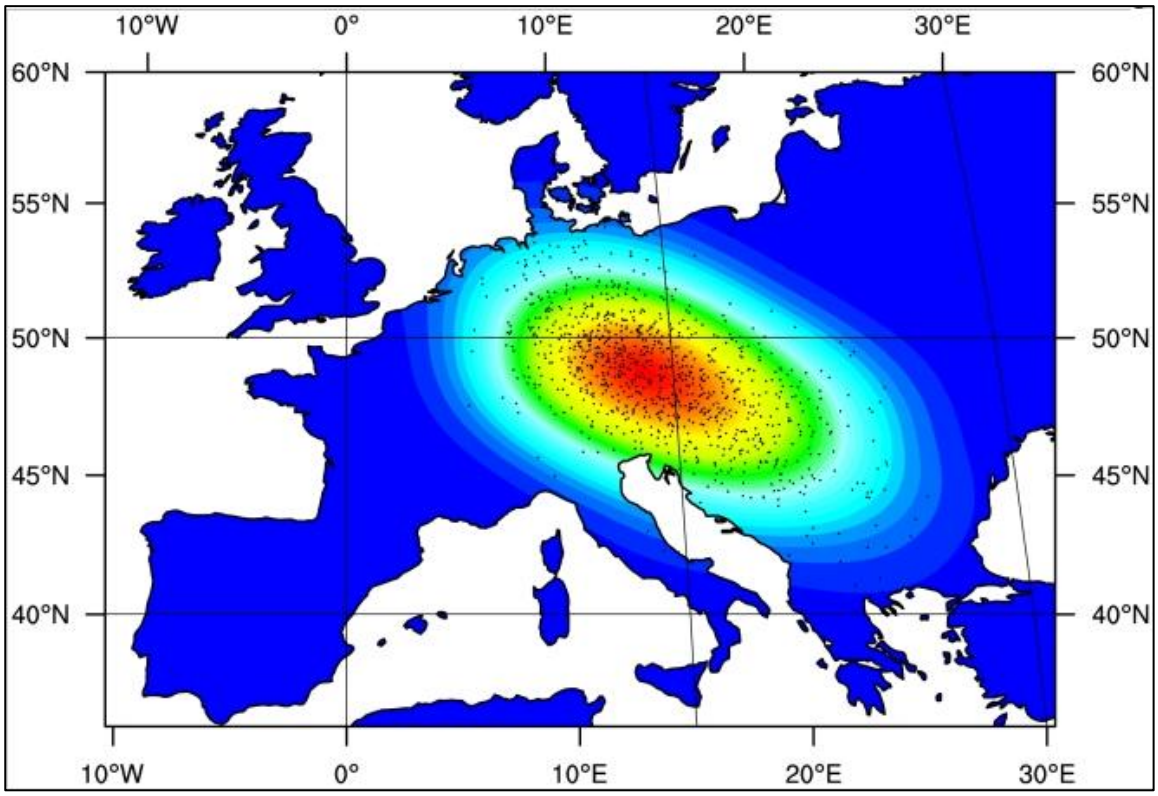


Figure 1.2.5.2: Approximate posterior density of region of origin for LP/dairying co-evolution (Itan et al. 2009).

Therefore, this study has inferred the following scenario: after the arrival of the Neolithic in South-Eastern Europe and the increasing importance of cattle herding and dairying, natural selection started to act on a few LP individuals of the early Neolithic cultures from Northern Balkans. After the initial slow increase of LP frequency in those populations and the onset of the Central European Linearbandkeramik (LBK) culture around 7,500 years ago, LP frequencies rose more rapidly in a gene-culture co-evolutionary process and on the wave front of a demographic expansion leading to the establishment of highly developed cattle (and partly also goat) based dairying economies during the Middle Neolithic of Central Europe around 6,500 years ago. A latitudinal effect on selection for LP, through an increased requirement for dietary vitamin D, is thus unnecessary to explain the high frequencies found in Northern Europe (Itan et al. 2009).

Although it has been suggested that a selective advantage based on additional nutrition from dairy explains these genetically determined population differences, formal population genetics-based evidence of selection was not provided before 2004. In fact, to assess the population genetics evidence for selection, Bersaglieri et

al. (2004) typed 101 SNPs covering 3.2 Mb around the lactase gene in American populations derived from African and European peoples. The study demonstrated that the long haplotype (1 Mb) carrying the persistence-associated alleles is much longer and more common than would be expected in absence of selection. Therefore, SNPs around the -13,910C/T variant are inherited without recombination from one generation to another.

Furthermore, the long haplotype block means that the mutation has recently occurred. On the basis of Bersaglieri's analysis of European-derived U.S. pedigrees, the best estimates of the time at which the persistence-associated haplotype began to rise rapidly in frequency are between 2,188 and 20,650 years ago, therefore consistent with the estimated origin of dairy farming in Northern Europe approximately 9,000 years ago (Bersaglieri et al. 2004).

Nevertheless, it has been proved that other polymorphisms besides than the -13,910C/T SNP and located in the mentioned haplotype block, can explain the existence of lactase tolerance in other countries in the world. In fact, Tishkoff et al. (2007) have discovered other variants in the *LCT* promoter which binds *OCT1*, such as -14,010G/C, -13,915T/G and -13,907C/G, which are associated with lactase persistence in Africa e in Arabia (Figure 1.2.5.3).

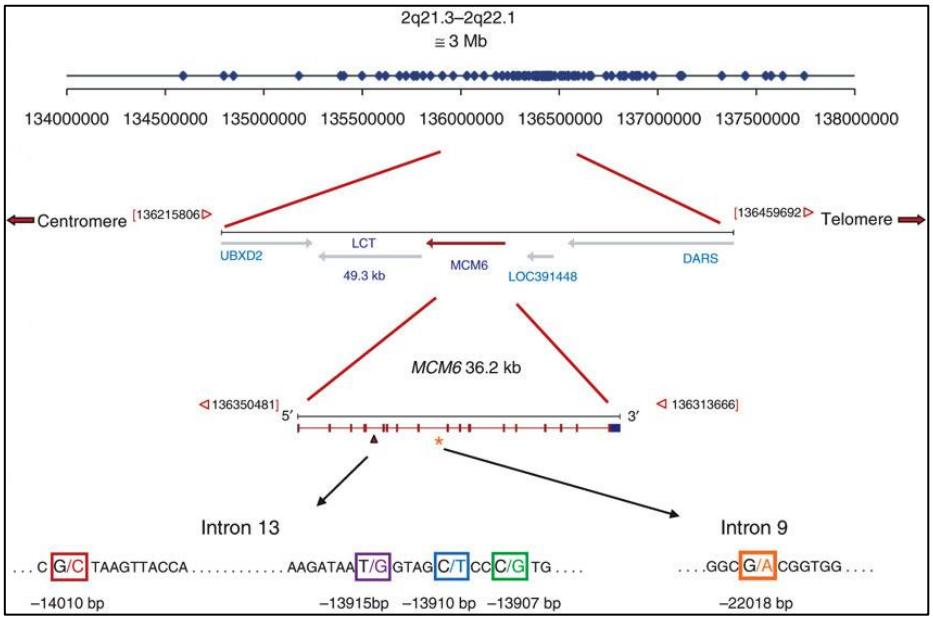


Figure 1.2.5.3: *LCT* gene and SNPs in African and European populations.

For instance, the polymorphism -14,010G/C is the principal variant particularly frequent in Nile-Saharian pastoralist populations, for example the Masaai, in Kenya and in Tanzania. Archaeological evidences suggest that cattle domestication originated in the Middle East approximately 7,000-8,000 years ago, consistent with the age estimate of around 8,000-9,000 years for the T-13,910 allele in Europeans. The more recent age estimate of the C-14,010 allele in African populations, which is 2,700-6,800 years, is instead consistent with archaeological data indicating that pastoralism did not spread south of the Sahara and into Northern Kenya until 4,500 years ago, as well as into Southern Kenya and Northern Tanzania around 3,300 years ago. Therefore, the more recent age of the C-14,010 with respect to the European -13,910C/T SNP means that African tolerant populations do not descend from North Europeans, but they became tolerant independently, in a different time. It is thus a case of evolutionary convergence, since the same phenotype is the result of different mutations occurred in different times and populations. Accordingly, lactose tolerance in desert regions has been considered as an advantage because, in times of famine, the ability to drink milk instead of water could have been advantageous, thus representing a response to a specific selective pressure (Tishkoff et al. 2007).

Two other variants found in Africa and Arabia are -13,915T/G and -13,907C/G, both with frequency generally higher than 5%. Reconstructing the natural history of the -13,915T/G polymorphism is really important to understand the evolutionary dynamics related to the LP phenotype developed by human populations from the Arabian Peninsula, since it appears to be an Arabian specific lactase persistence-associated allele (Ingram et al. 2007). The geographic distribution of the -13915*G variant suggests that this allele may have originated in the Arabian Peninsula and then spread into Northern regions of the Middle East, as well as into different regions of Africa (Ingram et al. 2007; Enattah et al. 2008). Moreover, the rise of the -13915*G allele is also thought to have occurred in the Arabian Peninsula then moving to the Middle East with decreasing frequency. In particular, it appears to be associated with the domestication of the Arabian camel and it was dated 4,000 years ago (Enattah et al. 2008). Since this variant seems to have developed independently, it could be considered another example of convergent evolution. The -13,907C/G variant has limited geographic distribution: it is present only in

populations of Sudan, Kenya and Ethiopia. It seems to be associated with the political economy of agriculture and cattle herding of the kingdom of Aksum, established around 100 AD (Tishkoff et al. 2007; Enattah et al. 2008) (Figure 1.2.5.4).

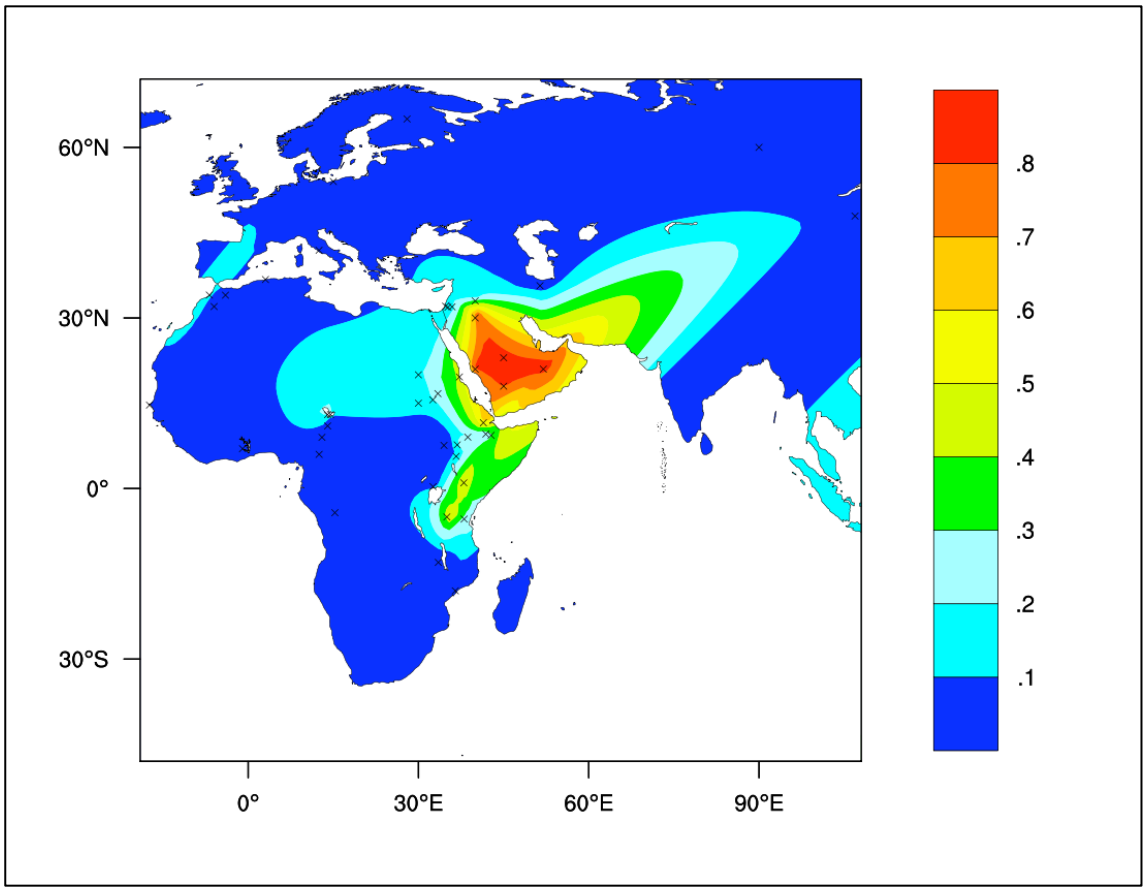


Figure 1.2.5.4: Predicted Old World LP phenotype frequencies based on -14,010 G/C, -13,915 T/G, and -13,907 C/G allele frequency data.

1.2.6 Lactase persistence: archaeological evidence

Neolithic transition was a long and complicate process that allowed the transition from a Palaeolithic-Mesolithic semi-nomadic lifestyle with an economy based on hunting and gathering, to a Neolithic sedentary culture, in which agriculture and domestic animal exploitation become the dominant subsistence strategies (Figure 1.2.6.1). Domestication of animals allowed human to using their primary products (meat, hide, bone, horn) but not only, it exists a so called "Secondary Product Revolution" that underlie the importance of the domestication also during the lifetime of the animals to use secondary products as: milk, wool, labour, dung (Sherrat 1981, 1983).

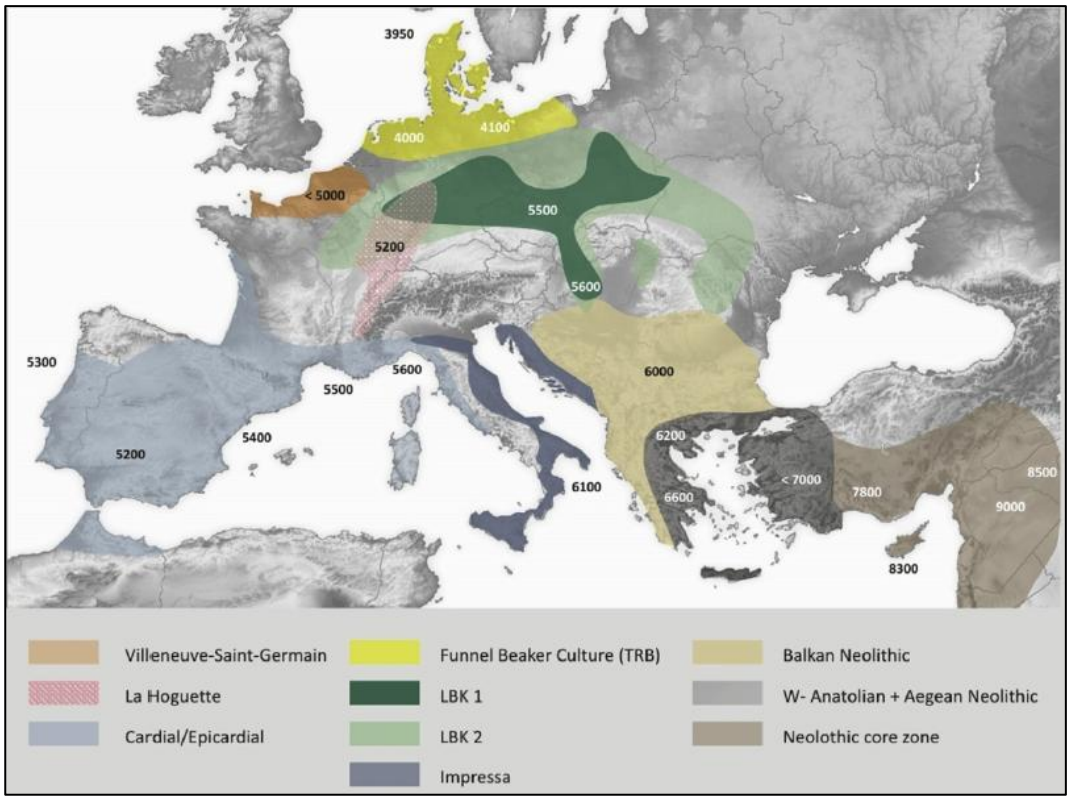


Figure 1.2.6.1: Chronological spread of the Neolithic (after Burger & Thomas, 2011).

However, this hypothesis has been widely criticized archaeozoological methods produced a lot of evidence that support this as skeletal assemblages and archaeometric analysis of organic residues in pottery. In particular analysed milk residues in pottery made possible to confirm that the exploitation of milk in the early Neolithic have occurred around 8,000 years ago in Northwestern Anatolia and Thrace (Evershed et al. 2008), around 7,000 years ago in the Carpathian Basin (Craig et al. 2005) and few hundred years later in Britain (Copley et al. 2003) (Figure 1.2.6.2).

Since the ancestral state of LP is non-persistence, and since milk exploitation is unlikely to have started before the Neolithic, could be interesting to explore the tolerance/intolerance genetics pattern by examining DNA from the skeletons of individuals living at the time. A lot of study on ancient DNA (Table 1.2.6.1) have managed to produce reliable data on the 13,910 C/T polymorphism underlined a low presence of tolerance associated allele in these individuals. For example the analyses of one Mesolithic and eight Neolithic European skeletons was conducted by Burger et al. (2007) and found not to carry the 13,910*T allele, suggesting that LP frequency was significantly lower in early Neolithic Europeans than it is now-a-

days. Analysis of 10 skeletons from a Middle Neolithic hunteregathering population in Scandinavia also indicated a large difference in LP frequency between ancient and modern populations; they found nine individuals who were homozygous for the 13,910*C allele and one heterozygous individual (Malmström et al. 2010).

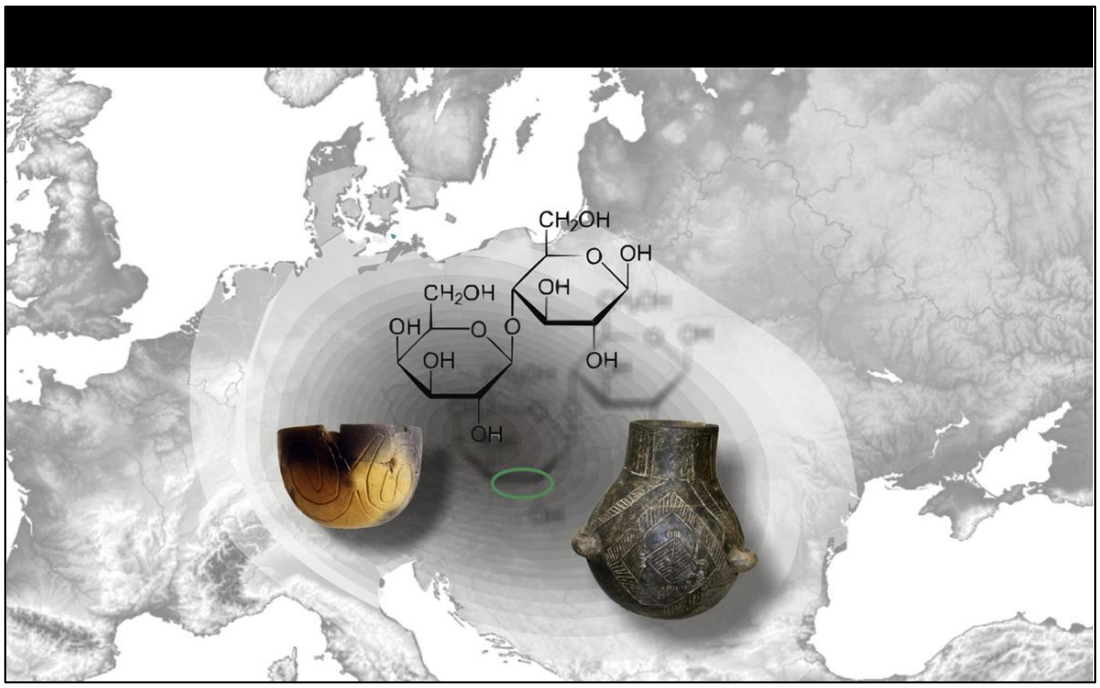


Figure 1.2.6.2: The estimates region where the lactase persistence associated allele was first subjected to selection (Itan et al. 2009) overlaps well with the region where the linear pottery culture (LBK) developed (green circle) (Leonardi et al. 2012).

Interesting the analysis of the lactase persistence status conducted on the Tyrolean Iceman showed that the mummy was lactose intolerance although was dated as a 5,300-year-old Copper age individual (Keller et al 2012). A recent paper reported a really important observation that suggest the possibility that calcium absorption was not the only driver of lactase persistence in Europe. This work examined the -13,910*T status using ancient DNA data from the skeletal remains of eight late Neolithic Iberian individuals, whom would expect to have poor vitamin D and calcium status because of relatively high incident UVB-light levels. None of the 8 samples successfully typed in the study had the derived T-allele (Sverrisdóttir et al. 2014).

Table 1.2.6.1: Summary of all the lactase persistence studies on ancient DNA.

Economy	Location	Date	N	C	T	Source
Hunter-gatherer	Loschbour, Luxembourg	6220-5990 BC	1	2	0	Lazaridis 2013
Hunter-gatherer	La Braña-Arintero, Spain	5940-5690 BC	1	2	0	Olalde 2014
Farming	Szarvas, Hungary	5840–5630 BC	3	6	0	Burger 2007
Farming	Kretuonas, Lithuania	5580-5350 BC	2	4	0	Burger 2007
Farming	Derenburg, Germany	5500–5000 BC	3	6	0	Burger 2007
Farming	Stuttgart-Mühlhausen	5100-4800 BC	1	2	0	Lazaridis 2013
Pastoralist	Ötztal Alps	3300 BC	1	2	0	Keller 2012
Farming	Sweden	3300-2500 BC	4	4-6	2-4	Malmstrom 2007
Farming	Treilles, France	3000 BC	26	52	0	Lacan 2011
Farming	San Juan Ante Portam Latinam, Spain	3000 BC	19	28	10	Plantinga 2012
Hunter-gatherer	Gotland, Sweden	2800-2200 BC	10	19	1	Malmstrom 2010
Farming	Terqa, Syria	2650–2450 BC	1	2	0	Witas 2013
Farming	Longar, Spain	2500 BC	7	12	2	Plantinga 2012
Hunter-gatherer	Drestwo, Poland	2267 BC	1	2	0	Burger 2007
Farming	Terqa, Syria	2200–1900 BC	1	2	0	Witas 2013
Farming	El Portalón de Cueva Mayor, Spain	1,735 BC	8	16	0	Sverrisdóttir 2014
Farming	Lichtenstein Cave, Germany	1000 BC	13	19	7	Schilz 2006
Farming	Tell Masaikh, Syria	200–300 AD	1	2	0	Witas 2013
Farming	Eltville, Germany	400–600 AD	1	1	1	Burger 2007
Farming	Tell Masaikh, Syria	500–700 AD	1	2	0	Witas 2013
	Csekej, Slovakia	900-1000 AD	1	2	0	Nagy 2011
Pastoralist	Besenyőtelek- Szőrhát, Hungary	900-1000 AD	1	2	0	Nagy 2011
Pastoralist	Eger- Szépasszonyvölgy, Hungary	900-1000 AD	1	2	0	Nagy 2011
Pastoralist	Harta- Freifelt, Hungary	900-1000 AD	3	6	0	Nagy 2011
Pastoralist	Orosháza- Görbics, Hungary	900-1000 AD	1	2	0	Nagy 2011
Pastoralists	Kolozsvár, Romania	900-1000 AD	1	2	0	Nagy 2011
Pastoralist	Sárrétudvar- Hízóföld, Hungary	950 AD	1	2	0	Nagy 2011
Pastoralist	Szabadkígyós- Pálliget, Hungary	950 AD	2	4	0	Nagy 2011
Medieval	Magyarhomoróg, Hungary	950 AD	4	5	3	Nagy 2011
Medieval	Dalheim, Germany	950– 1200 AD	18	18	18	Kruttli 2014
Pastoralist	Aldebrő-Mocsáros, Hungary	1000 AD	1	2	0	Nagy 2011
Pastoralist	Őrménykút, Hungary	1000 AD	2	4	0	Nagy 2011
Medieval	Fadd-Jegeshegy, Hungary	1000 AD	2	4	0	Nagy 2011
Medieval	Szegvár-Oromdűlő, Hungary	1000-1250 AD	2	2	2	Nagy 2011
Medieval	Zalavár-Kápolna, Hungary	1000-1200 AD	1	2	0	Nagy 2011
Medieval	Birger Magnusson, Sweden	d. 1266	1	1	1	Malmstrom 2011
Medieval	Erik, son of Birger, Sweden	d. 1275	1	0	2	Malmstrom 2011
Medieval	Mechtild of Holstein, Sweden	d. 1288	1	0	2	Malmstrom 2011

2. Aim of The Study

The presence of a common haplotype surrounding the *LCT* gene and laying largely undisrupted for more than 1 Mb has been demonstrated in human populations of Northern European origin (Bersaglieri et al. 2004). However, this observation was not confirmed in the other populations analysed in the same study such as African Americans, Chinese, Japanese and Southeast Asians. Moreover, this evidence was not further investigated in other human groups, especially from Southern Europe.

Accordingly, this study is expected to be the first one that surveys a large number of SNPs potentially related to the LP phenotype, in a huge sample of subjects, very well-characterized from an anthropological perspective and belonging to populations till now scarcely investigated for this adaptive trait.

In particular, assessing patterns of *LCT* variation in different Italian subpopulations promises to be highly informative according to the role that Italy has long played as a natural corridor for human migrations among Europe, Africa and Asia, according to its pivotal geographical position in the middle of the Mediterranean Sea. Therefore, dissecting gradients of genetic diversity related to the LP phenotype along the Italian peninsula has the potential to shed further light on the routes followed by LP diffusion, in conjunction with the Neolithic transition, from its area of origin to modern Europe.

Nevertheless, at the moment few studies have investigated the genetics of lactase persistence in the Italian population. In particular, Anagnostou et al. conducted the most recent study of LP in Italy in 2009, focusing only on two SNPs and the related microsatellites variation. Interestingly, they pointed out a statistically significant difference between North-Eastern and the remaining Italian populations. Moreover, the comparison of the lactose tolerance predicted by the -13,910*T allele and that assessed by studies using physiological tests showed a one-way statistically significant discrepancy that could be due to sampling differences, but also to the possible role of other genetic factors in modulating the examined phenotype (Anagnostou et al. 2009).

The present study was thus aimed at selecting the most informative SNPs located on the genomic interval showing strong signatures of positive selection in Northern Europeans and surrounding the *LCT* gene, to explore diversity patterns in a sample

of more than 400 Italian individuals. In particular, the examined subjects were recruited to be representative of the most diversified macro-areas (i.e. North-Western and Central-Western Italy, North-Eastern Italy, Central-Eastern and Southern Italy, Sardinia) pointed out by a recent analysis of the Italian population structure (Boattini et al. 2013). Accordingly, an exhaustive picture of existing gradients of nucleotide and haplotype variation potentially related to LP was described along one of the main natural corridor for westward and northward human migrations from Southern Europe and the Mediterranean area. That being so, local distribution of other variants rather than the -13,910 C/T one, and potentially involved in the modulation of the LP phenotype, has been evaluated. Moreover, together with the comparison with diversity patterns characterizing Northern European groups, drawing a detailed picture of LP-related variation in Italian subpopulations also offered a valuable opportunity to get additional insight into the routes of diffusion of this adaptive trait from its areas of origin to the regions of modern Europe in which it currently reaches the most outstanding frequency.

Finally, the described approach also promises to be useful to verify the importance of other variants, rather than the T-13,910 one, in modulating LP phenotype, as suggested by the discrepancy observed between the Italian distribution of LP obtained via physiological tests and that predicted according to -13,910 C/T genotypes.

In parallel, the present work was also focused on the exploration of LP-related diversity patterns in populations from Oman and Yemen. As already mentioned for Italy, also for this geographical area information about genetics of lactase persistence are really scarce despite the valuable and extremely complex history of human migration in these territories, that is until now not completely explained. In fact, both genetic and archaeological data show how the peopling of Arabia results from an amalgamation of internal demographic events coinciding with post-glacial climatic cycles (Al-Abri et al. 2012b; Rose et al. 2013). Recently, the two variants correlated with lactase persistence in European end Arabian populations have been studied in three groups from Oman and one with Yemeni origins. This work underlined a substantially homogeneity between all groups with the exceptions of the Omanis of Asian origin (Al-Abri et al. 2012a). To refine this picture and provide

a more detailed description of distribution of LP-related variation in this geographical area, we exploited the Sequenom chip designed for investigating Italian variability. Accordingly we explored the genetic variability of the same four human groups analysed in Al-Abri et al. (2012a) (i.e. Arabs of Northern Oman, Omanis of Asian origin, Arabs of Southern Oman (Dhofaris) and Yemeni). This extensive analysis has thus provided the opportunity to improve the knowledge about variants located in the region around the lactase persistent associated alleles and to deepen the understanding of the extent and origin of the LP selective sweep in the Arabia Peninsula.

An extended analyses of the biological evolution of adult lactose tolerance will be able to draw a picture of the different histories of the examined subpopulations and in particular, this will potentially help to disentangle some of the routes followed by LP diffusion from its area of origin to modern Europe, as well as to further explore processes of convergent evolution at LP occurred in some non-European populations.

3. Materials and Methods

3.1 Population Samples

3.1.1 Italian samples

The study of distribution of lactase persistence-related patterns of genetic variation in Italy was carried out on 453 healthy subjects recruited from four different geographic macro-areas: 106 samples from North-Western and Central-Western Italy (NCWI), 140 samples from North-Eastern Italy (NEI), 160 samples from Central-Eastern and Southern Italy (CESI) and 47 subjects from Sardinia (SARD).

All individuals were selected according to two different criteria: the "grandparent's criterion" and the "founder surnames' criterion". In more details, only individuals whose four grandparents were born in the same sampling area were included in the study, also taking into account the presence of founder surnames.

The collection of blood samples was achieved with the collaboration of transfusion centres of the involved provinces. Every subject included in the study, which has been approved by the Ethic Committee of the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna, has signed an informed-consent form.

The produced data were also compared with those publicly available for other European populations collected by the *1000 Genomes Project*. In particular, samples belonging to the Italian population (Tuscany, TSI), European Utah residents (CEU), as well as to other populations from Great Britain (GBR) and Finland (FIN), were examined. Although samples from the Iberian Peninsula (IBS) are available among those included in the 1000 Genomes dataset, they were not taken into consideration in the present study because of their high internal heterogeneity and of their extremely lower number with respect to the other investigated groups (16).

3.1.2 Samples from Arabian Peninsula

The study of distribution of lactase persistence-related patterns of genetic variation in the Arabian Peninsula was carried out on 635 samples from Oman and Yemen populations collected by the Biochemistry laboratory of the College of Medicine at the Sultan Qaboos University (Sultanate of Oman), thanks to the collaboration with Prof. Riad Bayoumi and Dr. Abdulrahim Alabri. In more detail, the Arabian sample

was composed of 215 Arabs from Northern Oman (ANO), 206 individuals from Dhofaris (DFR), 42 Omanis of Asian origins (OAO) and 172 Yemenis (YMN) (Figure 3.1.2.1).

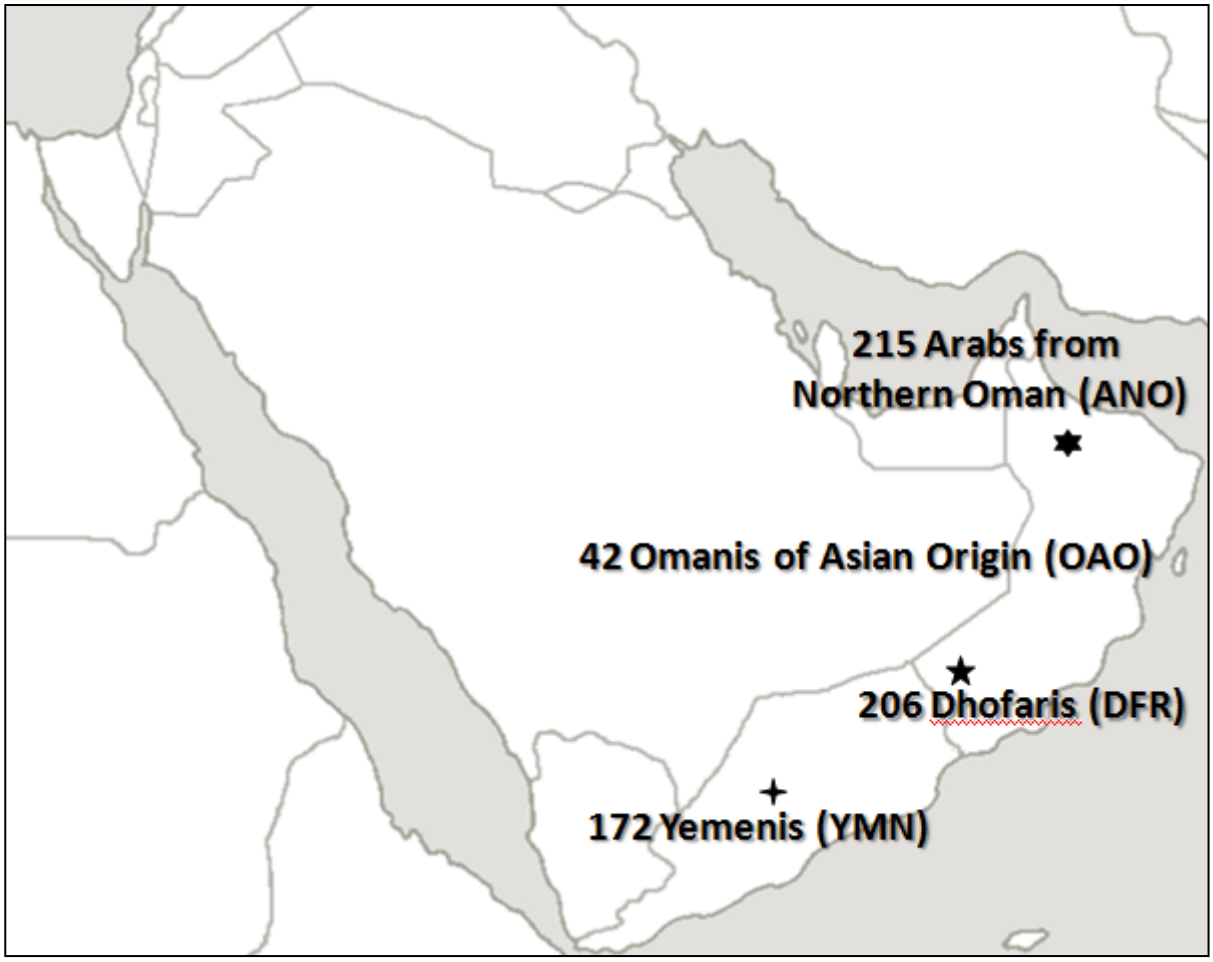


Figure 3.1.2.1: Distribution of the samples in the Arabian Peninsula.

3.2 SNPs Selection

A total of 49 single nucleotide polymorphisms (SNPs) were selected among those analysed in the Bersaglieri's study (Bersaglieri et al. 2004). In particular, these SNPs were chosen according to their highest heterozygosity values in the examined European subsample (Table 3.2.1). In addition to these SNPs, we have selected other two SNPs from those considered in the Enattah's work (Enattah et al. 2007), again according to high heterozygosity.

Table 3.2.1: Summary information on the genotyped SNPs.

ID	Position	Position from T-13,910	Polymorphism	Bersaglieri (2004) Frequencies
rs1531957	134781635	-1849339	T	21.3
rs1996589	134887524	-1743450	T	68.8
rs1257168	134986220	-1644754	A	40.4
rs1257220	135037675	-1593299	A	17.7
rs842360	135370213	-1260761	C	34.4
rs749017	135595987	-1034987	G	30.0
rs766271	135689459	-941515	C	55.4
rs2322254	135773177	-857797	C	19.8
rs1551497	135809970	-821004	C	15.0
rs2290518	135901142	-729832	G	85.4
rs2305248	135950640	-680334	A	85.4
rs935612	135963831	-667143	A	88.5
rs4954228	135998826	-632148	A	89.6
rs3754686	**136248412**	**-382562**	**C/T**	**26.0**
rs313522	136453194	-177780	T	83.0
rs1438307	136521494	-109480	T	83.0
rs3213889	136533903	-97071	G	82.6
rs1030764	136575857	-55117	T	86.5
rs1011361	136575967	-55007	A	83.3
rs2322659	136577987	-52987	C	86.4
rs892715	136598905	-32069	C	81.5
rs2164210	136602615	-28359	C	81.3
rs1470457	136604176	-26798	G	15.6
rs745500	136605520	-25454	A	81.9
rs2236783	136616486	-14488	A	81.9
rs4988235	136630974	0	T	77.2
rs4954493	**136632303**	**1329**	**C/T**	**26.0**
rs309180	136636583	5609	A	82.6
rs309181	136637141	6167	G	81.8
rs182549	136639082	8108	T	77.1
rs309176	136644544	13570	C	81.4
rs309125	136665883	34909	C	81.5
rs192822	136704602	73628	T	85.7
rs309137	136788279	157305	T	83.3
rs953388	136929457	298483	T	12.5
rs2176716	136946021	315047	T	18.8

rs1519523	136956777	325803	T	52.1
rs1519529	136996585	365611	G	19.5
rs4954411	137098753	467779	T	58.3
rs4501004	137129075	498101	T	27.1
rs1399604	137152993	522019	G	27.9
rs867563	137164828	533854	G	25.0
rs578935	137233319	602345	C	10.4
rs694510	137303189	672215	T	21.8
rs876338	137311475	680501	T	75.0
rs1427588	137514654	883680	C	43.8
rs1346731	137634915	1003941	A	40.3
rs518614	137739179	1108205	C	61.5
rs574135	137762448	1131474	G	62.5
rs1432232	137821992	1191018	C	64.6
rs882374	137935623	1304649	A	25.0

The two SNPs from the Enattah's work (Enattah et al. 2007) are reported in bold type.

As regards the Arabian samples, an additional SNP (rs4138034, -13915T/G) was previously assayed by means of a sequencing approach at the Biochemistry laboratory of the Sultan Qaboos University (Al-Abri et al. 2012a) and was thus included in the dataset used for the present study.

3.3 SNPs Genotyping and Quality Control

3.3.1 PicoGreen Quantification

The DNA quantification was obtained through the *Quant-iT dsDNA Broad-Range Assay Kit* (Invitrogen, Carlsbad, CA). This kit uses an intercalating agent of DNA, *PicoGreen*, that is an ultra-sensitive fluorescent nucleic acid stain for quantifying in solution double-stranded DNA (dsDNA), which emits a luminous signal proportional to the length of the DNA sequence. The measure of quantification is thus accurate. The adopted kit provides PicoGreen reagent, Buffer solution and eight DNA samples of bacterio-phage Lambda (λ), to be used as standards. The eight standards have a precise concentration: 0, 5, 10, 20, 40, 60, 80, 100 ng/μL. They were thus used for the calibration of the spectrophotometric instrument, being also processed with the *PicoGreen* reagent.

Each spectrophotometric reading led to the quantification of a 96-well plate

through *Microlab Star-Hamilton* technology (Figure 4.3.1.1) and was performed according to the following stages:

1) Positioning of the eight standards top-down, in order to increase concentration.
2) Placing of 16 tips 1000 CF in position nr. 1 and 208 tips 50 CF in positions nr. 2, 3, 4.
3) Placing of the PCR-Plate in position nr. 1 of "microplates" carrier.
4) Diluting PicoGreen in Buffer at a concentration of 1:200 at room temperature. The final volume of each well was 50 µL, referred 48 µL of diluted PicoGreen and 2 µL of DNA sample. Calculation of the required volumes of PicoGreen and Buffer for the reaction was performed by taking into account the overhang volume. The overhang volume must be added to compensate the lost volume during the automatic dispensation.

$$Volume =$$
$$= 2\,(replicated\ samples) * [(8\,(standards) + 96(samples)) * 50] + overhang$$
$$= 10400\mu L + overhang = 12000\mu L = 12mL$$

For *PicoGreen* Dilution:

$$Volume = \frac{12000\mu L}{200} = 60\mu L\ of\ PicoGreen$$

$$Final\ Volume = 12mL\ of\ Buffer + 60\mu L\ of\ PicoGreen$$

5) Robot dispensing of 48 µL of diluted PicoGreen, 2 µL of DNA sample and 2 µL DNA standard in apposite positions.
6) Covering of the 384-well plate with aluminium foil and activation of the shaker for 5 minutes.
7) Removing of the aluminium foil and placing of the plate in the fluorometer *Biotek Sinergy HT*. The instrument read the different wavelengths reported in a .xls file. The related software elaborated the data from the fluorometer and reported the results according to the original 96-well layout. Lastly, the final results were the concentration of DNA in each well and their average value.

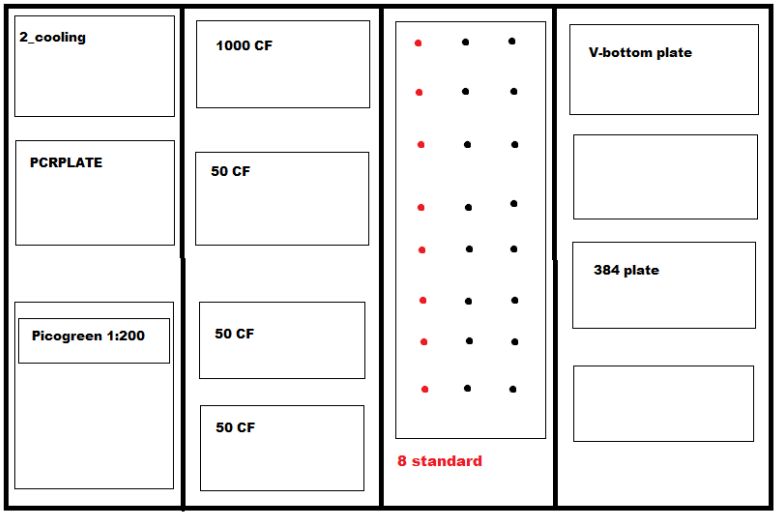


Figure 3.3.1.1: Layout of Microstar Lab Hamilton.

3.3.2 Multiplex PCR and SNPs Design

Multiplex polymerase chain reaction (Multiplex PCR) is a modification of polymerase chain reaction aimed at assaying many different genomic regions at the same time. This process amplifies genomic DNA samples using multiple primers and a temperature-mediated DNA polymerase in a thermal cycler (Hayden et al. 2008).

The *Sequenom's MassARRAY Designer software* (Sequenom, Inc., San Diego, CA) was used to design PCR and extension primers for the multiplex-PCR that investigates all SNPs at the same time, paying attention to avoid primer combinations and non-template extension products that could possibly result in non-specific extension. Moreover, multiplex levels of SNPs groups have been balanced in order to minimize the number of reactions. The obtained best multiplex design for the present project rejected three SNPs and divided the remaining 48 ones in four multiplex reactions, each one composed of a different number of SNPs:

- Multiplex nr. 1: 21 SNPs
- Multiplex nr. 2: 19 SNPs
- Multiplex nr. 3: 7 SNPs
- Multiplex nr. 4: 1 SNPs

Only two Multiplex (nr.1 and nr.2) presented an adequate number of SNPs, thus actually allowing us to genotype 40 of the 51 previously selected SNPs (Appendix Table 1).

3.3.3 The Sequenom MassARRAY iPLEX Platform

Genotyping was performed on each study participant 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 the collaboration of the Centre for Applied Biomedical Reseacrh (CRBA) of the Bologna S. Orsola University Hospital).

The adopted protocol included the following steps (Figure 4.3.3.1) (Gabriel et al. 2009):

a) DNA and oligo pool preparation

Genomic DNAs were diluted to 2.5-5 ng/µl concentration in TE buffer at a concentration of 0.25X or less, than the DNA was divided into aliquots at 2 µl/well into a 384-well PCR reaction plate from a deep-well PCR source plate (Marsh Biomedical).

b) PCR amplification of target loci

As already described, the *MassARRAY Assay Designer 3.1* software was used to design a single multiplex reaction in which all the selected SNPs were included. A multiplex-PCR reaction was used for amplifying the specific regions of genomic DNA that include all the polymorphisms chosen to be genotyped. The goal of an optimal multiplex PCR reaction is to evenly amplify many individual loci with minimal non-specific PCR products.

c) SAP reaction cleanup

Treatment with Shrimp Alkaline Phosphatase (SAP) was performed in order to remove remaining, non-incorporated, dNTPs from amplification products.

d) Primer extension

After PCR cleanup a locus-specific primer extension reaction was performed. In this reaction, an oligonucleotide primer anneals immediately upstream of the polymorphic site of interest. Primer and amplified target DNA were incubated with mass-modified dideoxynucleotide terminators. This reaction generated oligonucleotides with an allele-specific molecular mass.

e) Primer extension reaction resin cleanup

SpectroCLEAN (Sequenom, Inc., San Diego, CA) is a cationic resin pre-treated with acid reagents. This resin was added directly to primer extension reaction products to remove salts, such as Na^+, K^+, and Mg^{++} ions. This cleanup step is important to

optimize mass spectrometry analysis of the extended reaction products because if not removed, these ions can result in high background noise in the mass spectra.

f) Spotting primer extension products on SpectroCHIPs

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

g) Detection of primer extension products by Mass Spectrometry

Sample molecules were vaporized, ionized and detected on the basis of their mass-to-charge (m/z) ratios. A laser beam was used as desorption and ionization source in MALDI mass spectrometry. Detection of an ion at the end of the tube is based on its flight time, which is proportional to the square root of its m/z ratio. *Sequenom* supplies a software (*SpectroTYPER*) that automatically translates the mass of the observed primers into a genotype for each reaction (Figure 3.3.3.2).

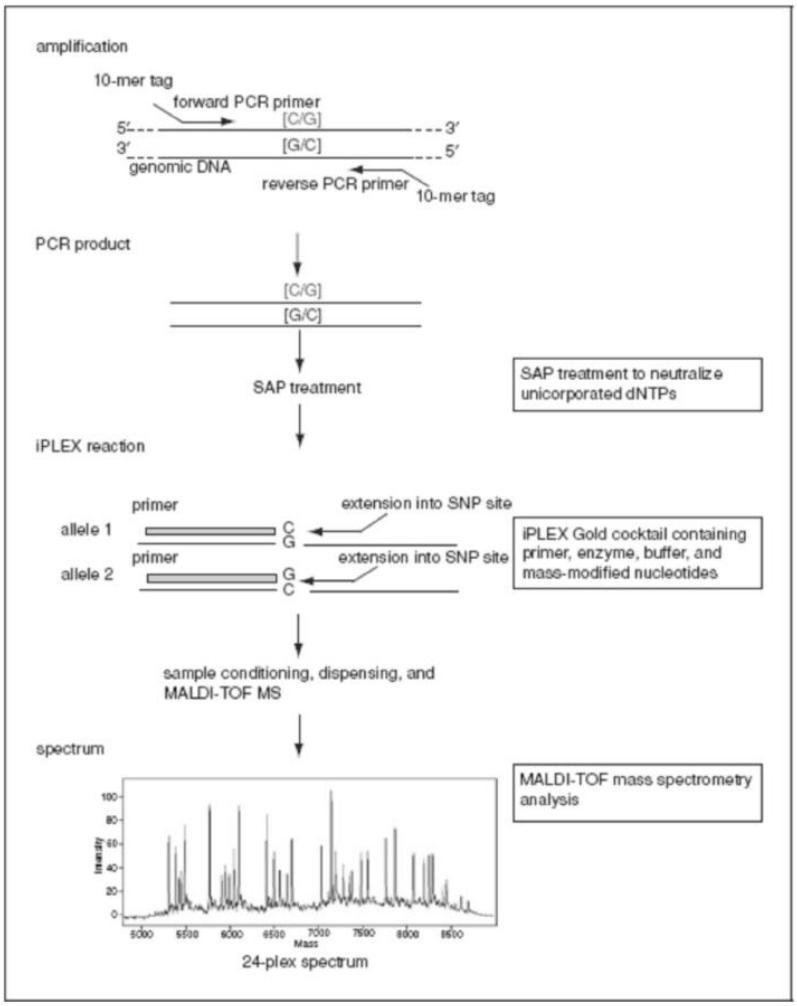


Figure 3.3.3.1: Schematic representation of genotype reaction of a C to G SNP.

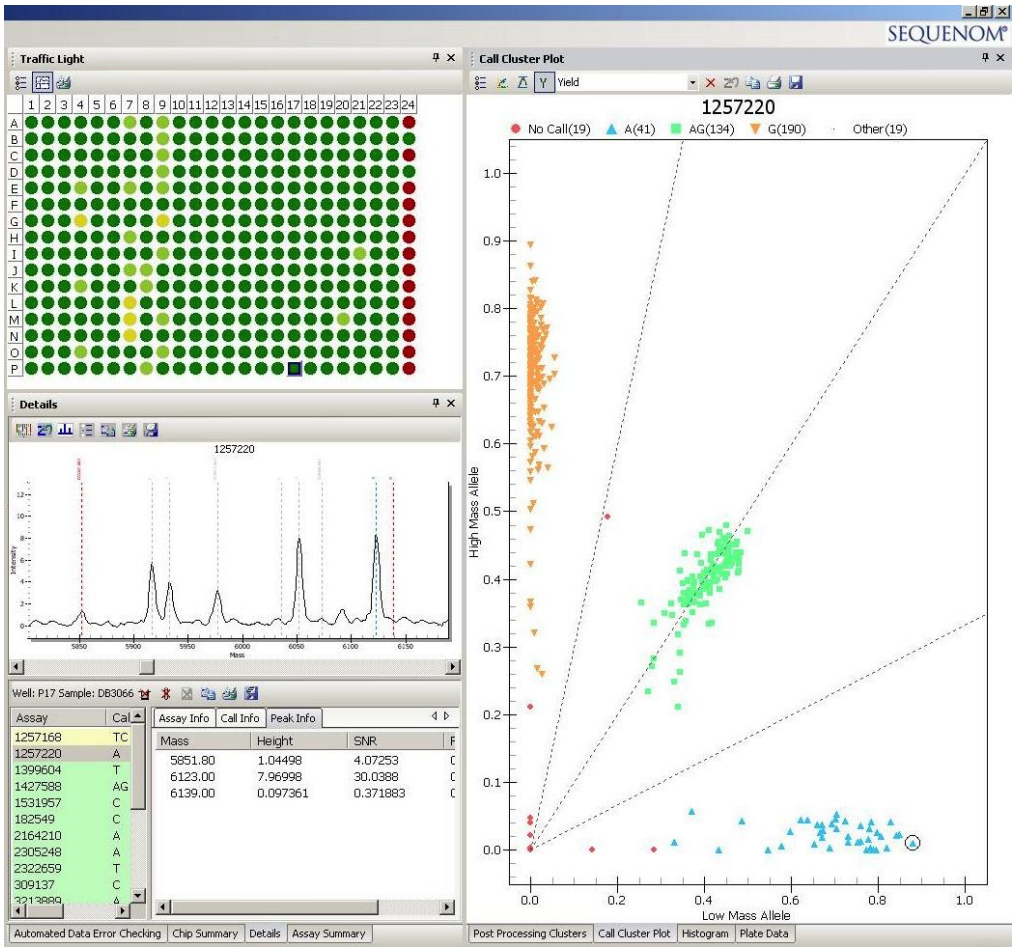


Figure 3.3.3.2: Example of Sequenom output data.

3.3.4 Quality Control

The quality control process provided several steps designed to minimize the amount of false positive results. In order to reduce these errors, the genotyped markers were filtered according to the following criteria:

- failed genotyping;
- Minor Allele Frequency (MAF) less than 5%;
- call rates less than 70%;

Significant deviations from Hardy-Weinberg equilibrium (HWE) p-value were calculated, but SNPs that does not satisfy this criteria were not removed from the analyses because of their possible implication in natural selection event.

3.4 Statistical Analyses

3.4.1 Linkage Disequilibrium Analysis

The analysis of Linkage Disequilibrium (LD) allows to measure the amount of alleles random association at one or more loci. LD reflects the differences between the observed frequency of a two-loci haplotype and its expected frequency if the alleles were segregating at random (Jobling et al. 2014).

There are two most used measures of LD: D e r^2. D is calculated as the difference between the observed frequency of a two-loci haplotype and its expected frequency, if the alleles were randomly segregating, whereas r^2 is calculated as the square of the correlation coefficient between the two loci and it is derived by dividing D^2 by the product of the four alleles at the two examined loci. D values range from -1 to 1, if D is significantly different from zero, LD exists, whether it is positive or negative it depends on the arbitrary labelling of alleles. The r^2 statistic ranges instead from 0 to 1 and strongly depends on the markers allele frequencies. In fact, $r^2 = 1$ (i.e. perfect disequilibrium) occurs if, and only if, the alleles have not been separated by recombination and thus have the same allele frequency (Jobling et al 2014).

The analysis of LD on the genotyped loci was carried out by the software *Haploview 4.2* and according to the *Solid Spine* method (Barrett et al. 2005).

Haploview is a commonly used bioinformatics software designed to analyse and visualize patterns of LD in genetic data, as well as to perform basic association studies, choose tagSNPs and estimate haplotype frequencies (*www.broadinstitute.org*).

In particular, *Haploview* allowed to reconstruct LD blocks from the analysed genotypes through the determination of r^2 values. The software also created a plot representing haplotype blocks within the examined region of the chromosome. This map represents the r^2 values of every pair of SNPs having a different colour according to their values: deep red when $r^2 = 1$ and white when $r^2 = 0$.

3.4.2 Allele Frequency Analysis

Allele frequency is calculated as the proportion of all copies of a gene constituting a particular gene variant, named allele (Strachan & Read 2011).

Statistical significance of the ratio between the frequencies can be measured by a chi-square test (χ^2). Chi-square values range from 0 to 1: $\chi^2 = 0$ is obtained when the frequencies in the considered populations are equal, whereas $\chi^2 = 1$ is obtained when they are completely different. This test can be used to verify the difference or the similarity between the allelic frequencies in two populations. In chi-squared test, there is a null hypothesis stating that the two populations present identical allelic frequencies. The aim of this test is finally the denial of null hypothesis (Iacus 2006).

Measures of allele frequencies in the examined datasets were provided by the software *PLINK 1.07*. *PLINK* is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner (*pngu.mgh.harvard.edu/~purcell/plink*). This software also performs chi-squared test and provide related p-values to test for significant differences in allele frequencies among groups.

The p-value of each SNP should be adjusted to eliminate false positives derived from multiple statistical tests. The incidence of false positives is proportional to the number of tests performed; therefore the multiple testing correction is important in order to obtain actually significant values. A kind of multiple testing correction is the Bonferroni correction (Holm 1979).

The p-value of each SNP is multiplied by the number of SNPs in the SNP list. If the corrected p-value is still below the chosen critical significance level, the SNP is considered to be actually significant.

$$Corrected\ p - value = p - value * n\ (number\ of\ SNPs\ in\ test) < 0.05$$

As a consequence, while testing 1,000 SNPs at a time, the highest accepted individual p-value is 0.00005, that makes the correction very stringent.

3.4.3 Haplotypes Reconstruction

Haplotypes can be obtained experimentally or (partially) through genotyping of additional family members, but this type of reconstruction may be an inefficient use of resources. Alternatively, a statistical method can be used to infer phase at linked loci from genotypes and thus to reconstruct haplotypes. The two most popular existing methods are a maximum likelihood method, implemented via the expectation-maximization (EM) algorithm (Excoffier and Slatkin 1995), and a parsimony method created by Clark (1990). Nowadays, a valuable method for haplotype reconstruction can be considered the one drawn up by Matthew Stephens. This is a Bayesian statistical method that allows to use a priori expectations to inform haplotype reconstruction and it has wider applicability and increased accuracy than the previous described methods. The software PHASE implements the Stephens' algorithm for estimating haplotypes from population genotype data (Stephens et al. 2001).

3.4.4 Summary and population differentiation statistics

Nucleotide diversity (π) describes the probability that two copies of the same nucleotide drawn at random from a set of sequences will be different from one another. π is a descriptive statistic that describes the genetic diversity of a population and allows the comparison of populations or loci (Jobling et al. 2014). Mean observed heterozygosity across loci (OH), number of haplotypes (k) and haplotype diversity (H) are other index to describe genetic diversity. In particular the haplotype diversity is a measure of the uniqueness of a particular haplotype in a given population. The haplotype diversity is computed as:

$$H = \frac{N}{N-1}(1 - \sum_i x_i^2)$$

where xi is the (relative) haplotype frequency of each haplotype in the sample and N is the sample size. (Jobling et al. 2014).

Measures of genetic distance are instead statistics that allow to compare the relatedness of populations or molecules. Such measures enable to explore population structure and molecular diversity in greater detail, by pairwise

comparisons, rather than by averaging over all populations or molecules. As we shall see, by making certain assumptions, it becomes also possible to convert distance measures to an evolutionary time-scale.

A commonly used classical measure of genetic distance between populations is the F_{ST} index that is specifically formulated for two populations and can be defined as:

$$F_{ST} = V_P/p(1-p)$$

where p and V_p are respectively the mean and the variance of gene frequencies between the two populations. F_{ST} can be regarded as a family of distances, rather than a single distance, because there is a variety of different methods for estimating it (Jobling et al. 2014).

If we have n populations, we required n-1 dimensions to fully display their pairwise genetic distances as graphical, or Euclidean distances. However, we often have more than four populations or molecules that we want to compare; yet we are unable of conceiving of, or representing, the multidimensional spaces required to display these data. Multivariate analyses (see subsequent section) allow us to reduce this multidimensional space to the two or three dimensions we can comprehend, while reducing the inevitable loss of information (Jobling et al. 2014).

Both summary and F_{ST} statistic were calculated using *Arlequin. Arlequin 3.5* is a software package integrating methods for population genetics data analysis, like the computation of standard genetic diversity indices, the estimation of allele haplotype frequencies and the estimation of parameters from past populations expansions. This software uses two methods for the analysis of input files: a Bayesian estimation of gametic phase from multi-locus genotypes (ELB algorithm) and an estimation of the parameters of an instantaneous spatial expansion from DNA sequence polymorphism (EM sipper algorithm) (Excoffier et al. 2005).

3.4.5 LD-based SNP pruning

To avoid LD effects on multivariate analyses, a dataset containing only the subset of genotyped SNPs made up of variants in approximate linkage equilibrium with each other (i.e. tag SNPs) was created. For this purpose, SNPs that presented a LD higher than $r^2 = 0.1$ were removed using the PLINK 1.07 --indep-pairwise command that

prunes data according to a method based only on pairwise genotypic correlation. The parameters used were a window size of 50 SNPs and a number of SNPs to shift the window at each step of five. The obtained output was made up of two lists of SNPs: those that are in linkage equilibrium and those that are not. A separate command (i.e. --extract or --exclude) was then used to selected the list of interest.

3.4.5.1 Multivariate analyses

The complete and the pruned datasets of SNPs were used for applying several multivariate analyses.

Principal Components Analysis (PCA) was also performed using the R *adegenet* package and representing a commonly used example of multidimensional-scaling.

Cavalli-Sforza (1997) wrote that PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs), which sum up in a few simple figures most of all information contained in the genes analysed. Typically, the data of more than 100 gene frequencies relating to a number of populations can be synthesized with an efficiency of 20%-40%, by replacing the genetic data of each population with unique numerical value that relates to the first principal component. In this way, approximately 60%-80% of the total variation is lost, which however can be partially recovered by calculating the numerical value of a second main component. The amount of information that is summed by the second PC is less than the first one, but it is a valuable contribution to the further first. The first two PCs almost always allow us to get a good description of the genetic similarities between the populations in question (Cavalli-Sforza 1997).

Thus, using PCA it is possible to estimate the proportion of the total variance in the total dataset that has been summarized within these reduced dimensions. PC data can also be used to construct “synthetic” maps that summarize information from several alleles with similar geographic distribution.

Genetic structure at the examined loci was investigated through PCA by considering both the individual and the population levels .

Moreover, to provide further support to the identified population groups, evaluation of cluster membership probabilities for each individual was achieved by

means of *Discriminant Analysis of Principal Components* (DAPC) (Jombart et al. 2010).

In fact, this procedure is particularly well suited for depicting diversity patterns observable among pre-defined groups of observations. DAPC was repeated with different randomized groups for different numbers of retained PCs, whose optimal number was identified as that optimizing the mean α-score (i.e. the closest to one) obtained as the difference between observed and random discriminations. Retained PCs were passed to a Linear Discriminant Analysis that constructed discriminant functions as linear combinations of the original variables in order to show the largest between-group variance and the smallest within-group variance. Given the low number of clusters identified by the other population structure analyses, all discriminant functions were retained and used to compute individuals' membership probabilities.

3.4.6 Analysis of the Molecular Variance

Apportionment of genetic variance at different hierarchical level (F_{ct} among geographically-based groups of populations, F_{sc} within geographically-based groups of populations and F_{st} among individual populations) was investigated with a locus by locus Analysis of the Molecular Variance (AMOVA) (Excoffier et al. 1992), exploiting information on haplotypes allelic content and frequencies. This allowed to assess the statistical significance of variation patterns pointed out by population structure analyses.

AMOVA was computed using *Arlequin 3.5* (Excoffier et al. 2005), a software package integrating methods for population genetics data analysis, such as the computation of standard genetic diversity indices, the estimation of allele and haplotype frequencies, as well as of parameters from past populations expansions.

3.4.7 Phylogenetic Methods and Networks

Alternative methods for the graphical display of inter-population and inter-molecule relationships attempt to reconstruct the ancestral relationships between all the entities under investigation. These are known as phylogenetic methods and their outputs are phylogenies (Jobling et al. 2014).

In particular, the tree is an intuitively attractive method for displaying the relationships between many kinds of variant entities. A tree consists of branches between nodes. The ultimate aim is to relate groups of populations or molecules, known as taxa (i.e. leaves) that joined together in the tree and that are implied to have descended from a common ancestor (Figure 3.4.7.1). Trees can be rooted or unrooted.

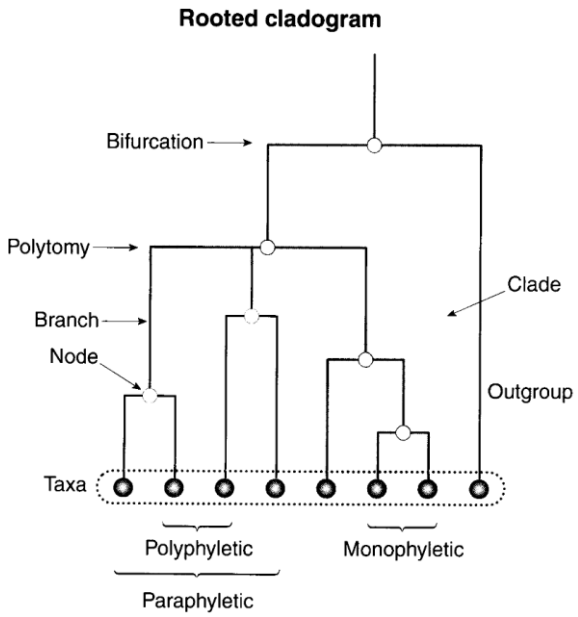


Figure 3.4.7.1: Example of a phylogenetic tree.

Rooted trees have a particular node, named root, from which all other nodes depart. That being so, this property orientates rooted trees with respect to evolutionary time, meaning that evolutionary changes have a defined direction of change, from ancestral to derived states (Jobling et al. 2014).

An important property of trees is that as evolutionary time progresses towards the present, branches diverge, but never coalesce. However, some biological processes (e.g. recombination) can cause lineages to merge, while others (e.g. parallel mutation) cause them to appear to merge. In either cases, the result can be represented as a four-sided closed structure known as a reticulation, or cycle. Trees that incorporate such structures, in the attempt to include these biological processes, are known as networks. A network is an unrooted tree and only represents the relations among taxa without connection of evolution (Figure

3.4.7.2). Network show the inferred evolutionary relationships among species or populations based upon similarities in their genetic characteristics. In particular, the network illustrates the relatedness of the leaf nodes without making any assumption about the ancestry (Jobling et al. 2014).

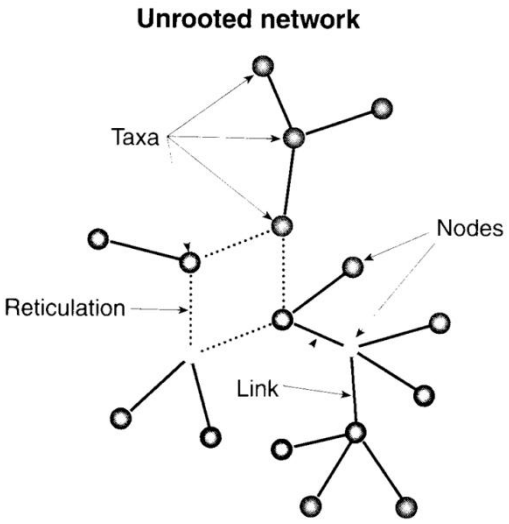


Figure 3.4.7.2: Example of network.

There are many different methods for constructing networks and they are classified using two criteria, which are the type of input data and their means. In particular, networks can be constructed from distance or character data, and there are a number of alternative methods for their construction. The easiest tree to construct is the minimum spanning because connects all given types without creating any cycles or inferring additional (ancestral) nodes (Bandelt et al. 1999).

A method for constructing networks with limited levels of reticulation is the median joining method. The algorithm used to construct these median-joining networks is based on the limited introduction of likely ancestral sequences/haplotypes into a minimum spanning network of the observed sequences. The median joining algorithm has the advantage of being applicable to multi-state markers and is useful for large datasets, but is more unreliable for phylogenies with long branches. After the creation of a minimum spanning network, the algorithm creates and adds the median vectors: consensus sequences of here mutually close sequences at a time. These median vectors can be biologically interpreted as possibly extant, unsampled,

sequences or as extinct ancestral sequences (Bandelt et al. 1999; Jobling et al. 2014). The software *Network 4.6.1.0* was used to reconstruct phylogenetic networks and trees, to infer ancestral types and potential types, as well as evolutionary branching and variants, and finally to estimate dating (www.fluxus-engineering.com).

3.4.8 Correlation Analysis

A comparison between the allelic status at lactase-related loci and at uniparentally inherited genetic markers was performed for the Italian population after the reconstruction of Y chromosome and mitochondrial DNA haplotypes taking advantage from data recently published by Boattini et al. (Boattini et al. 2013). Haplotypes correlation was tested using a Spearman's rank correlation coefficient (Spearman 1906). This coefficient is a nonparametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. The significance levels of the correlations were assessed with a p-value and the raw false discovery rate was calculated to adjust the results for multiple tests (adjusted p=0.03).

4. Results

4.1 Quality Control

4.1.1 Italian samples

Genotyping failed for two of the collected samples (one from North-Eastern Italy and one from Central-Southern Italy) and for one SNP (rs 2305248) for which experimental problems occurred during the PCR reaction. Obtained genotypes are available upon request from the Laboratory of Molecular Anthropology of the Dept. of Biological, Geological and Environmental Sciences of the University of Bologna.

The remaining 39 selected SNPs showed both MAF exceeding 5% and call rates higher than 80%. Although, three SNPs (rs749017, rs309180, rs309137) showed nominal p-values for the HWE test lower than 0.01, no variants significantly deviated from HWE after Bonferroni correction for multiple testing (adjusted p threshold = 2.56×10^{-4}) (Table 4.1.1.1)

Table 4.1.1.1 Hardy-Weinberg equilibrium p-value and SNPs call rate in our design.

Name	HWpval	Call Rate	Name	HWpval	Call Rate	Name	HWpval	Call Rate
rs1531957	0.7696	99.3	rs1011361	0.0577	100	rs1519523	1.0000	100
rs1996589	0.3308	99.6	rs2322659	0.0394	99.6	rs1519529	0.3767	99.1
rs1257168	0.4224	96.9	rs2164210	0.1982	95.1	rs4954411	0.8890	100
rs1257220	0.9633	97.5	rs1470457	0.0620	100	rs1399604	0.8380	99.8
rs842360	0.6507	96.4	rs745500	0.0434	100	rs867563	1.0000	99.8
rs749017	**0.0013**	**90.3**	rs2236783	0.0520	100	rs578935	0.7821	100
rs766271	0.0444	100	rs3754686	0.0319	100	rs876338	0.3080	100
rs2322254	0.6348	100	rs4988235	0.0616	99.8	rs1427588	0.3996	97.8
rs2290518	0.8023	98.7	**rs309180**	**0.0087**	**100**	rs1346731	0.8656	99.8
rs935612	0.3824	100	rs309181	0.0258	100	rs518614	0.8704	99.3
rs4954228	0.1740	98.9	rs182549	0.1533	100	rs574135	0.6143	99.8
rs1438307	0.0662	99.8	rs309176	0.0194	99.1	rs1432232	0.5573	100
rs3213889	0.0623	99.8	**rs309137**	**0.0055**	**96.9**	rs882374	0.2814	96.9

Level of significance: nominal p-value=0.01, p-value after Bonferroni correction=2.56×10^{-4}.

4.1.2 Samples from Arabian Peninsula

A total of 630 samples has been successfully genotyped in populations belonging to the Arabian Peninsula, since genotyping failed for five samples (one from ANO, three from DFR and one from OAO). Moreover, one SNP (rs3213889) was excluded

from the study according to the low percentage of individuals for which genotypes have been obtained, as a consequence of experimental problems occurred during the PCR reaction.

Obtained genotypes are available upon request from the Laboratory of Molecular Anthropology of the Dept. of Biological, Geological and Environmental Sciences of the University of Bologna.

The remaining 39 selected SNPs showed call rates higher than 70% and, with the sole exception of the -13,910 C/T and -22,018 G/A variants, a MAF > 5%. Nevertheless, these SNPs have been anyway retained in further analyses to explore their potential relationships with the Arabian-specific LP-associated SNPs.

Although, 15 SNPs (Table 4.1.2.1) showed nominal p-values for the HWE test lower than 0.01, a single variant (rs41380347) significantly deviated from the HW equilibrium after Bonferroni correction for multiple testing (adjusted p threshold = 2.56×10^{-4}).

Table 4.1.2.1 Hardy-Weinberg equilibrium p-value and SNPs call rate in our design.

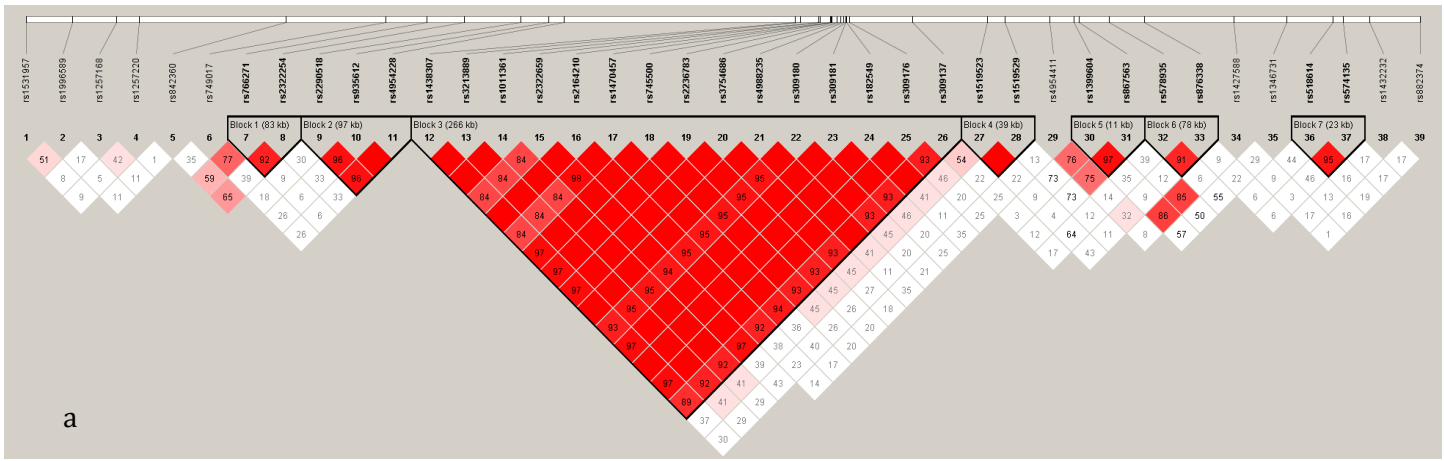
Name	HWpval	Call Rate	Name	HWpval	Call Rate	Name	HWpval	Call Rate
rs1531957	0.0636	99.5	rs2322659	0.1458	99.8	**rs1519523**	**7.00E-04**	**100**
rs1996589	0.0735	99.7	rs2164210	0.0727	99.8	**rs1519529**	**0.0045**	**100**
rs1257168	0.3747	99.8	rs1470457	0.0634	99.7	**rs4954411**	**0.0085**	**99.8**
rs1257220	0.3012	99.8	rs745500	0.1151	100	rs1399604	0.9415	99.5
rs842360	0.9042	97.3	rs2236783	0.3832	100	rs867563	0.9306	99.7
rs749017	**0.0048**	**99.8**	rs3754686	0.3311	100	**rs578935**	**3.00E-04**	**100**
rs766271	0.3726	97.3	**rs4988235**	**0.0051**	**100**	rs876338	0.0626	99.8
rs2322254	0.0235	99.5	**rs41380347**	**1.11E-52**	**87.5**	rs1427588	0.5271	99.4
rs2290518	**0.0047**	**96.5**	rs309180	0.2618	100	rs1346731	0.1628	99.7
rs2305248	**0.0034**	**99.7**	rs309181	0.2198	98.7	rs518614	0.0200	99.8
rs935612	**0.0082**	**99.7**	**rs182549**	**0.0095**	**99.8**	**rs574135**	**0.0038**	**96.6**
rs4954228	**0.0077**	**99.2**	rs309176	0.5643	70.2	rs1432232	0.1142	99.8
rs1438307	0.2215	100	rs309137	0.0624	99.8	rs882374	0.0314	99.7
rs1011361	0.3195	100						

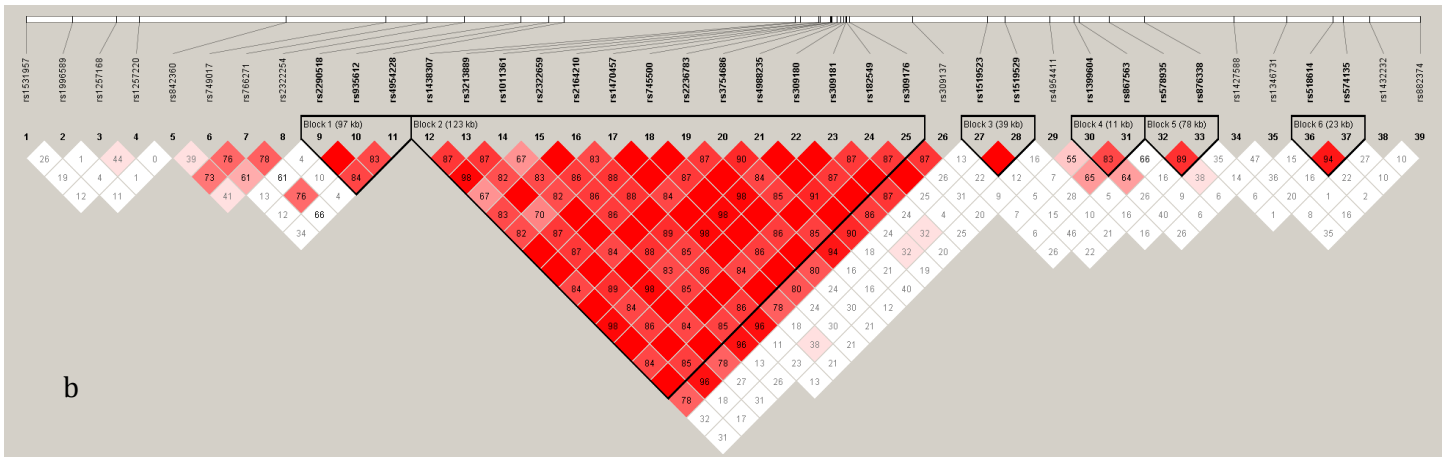
Level of significance: nominal p-value=0.01, p-value after Bonferroni correction=2.56×10^{-4}.

4.2 Linkage Disequilibrium Analysis

4.2.1 Italian samples

Analysis of LD carried out on the Italian dataset pointed out the existence of a long block of high LD (266 Kb), encompassing the functional variant -13,910C/T (rs4988235), in the North-Western and Central-Western Italian subsamples (Figure 4.2.1.1a). A slightly shorter block (123 Kb) differing for a single SNP (rs309137), was instead observed in North-Eastern Italy (Figure 4.2.1.1b). Interestingly, this region of high LD appeared to be split into two different blocks, spanning respectively 81 Kb and 184 Kb, in both the Central-Eastern/Southern Italian and Sardinian subsamples (Figure 4.2.1.1c and 4.2.1.1d).





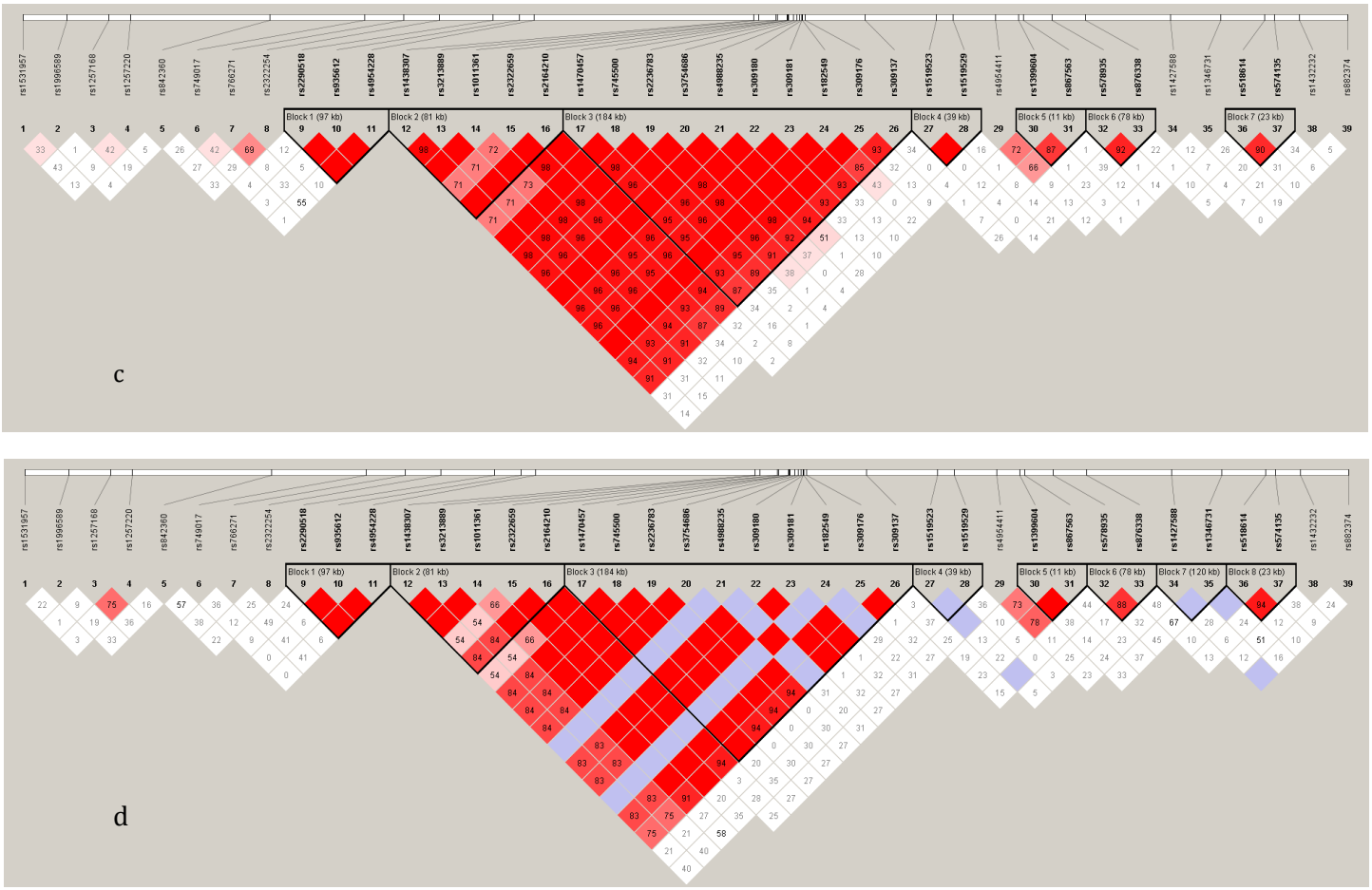


Figure 4.2.1.1: Plot of LD in NCWI, NEI, CESI and SARD.

4.2.2 Samples from Arabian Peninsula

The LD analysis was carried out also on the four populations belonging to the Arabian Peninsula, highlighting the presence of a long region of high LD surrounding the two known LP-associated loci (-13910 C/T and -13915 T/G) in ANO, OAO and YMN groups. In particular, this region appeared to be progressively reduced from OAO (266 Kb) to ANO (123 Kb) and YMN (117 Kb), whereas it was subdivided into three considerably shorter blocks of 84 Kb, 14 Kb and 2 Kb in the DFR subsample (Figure 4.2.2.1).

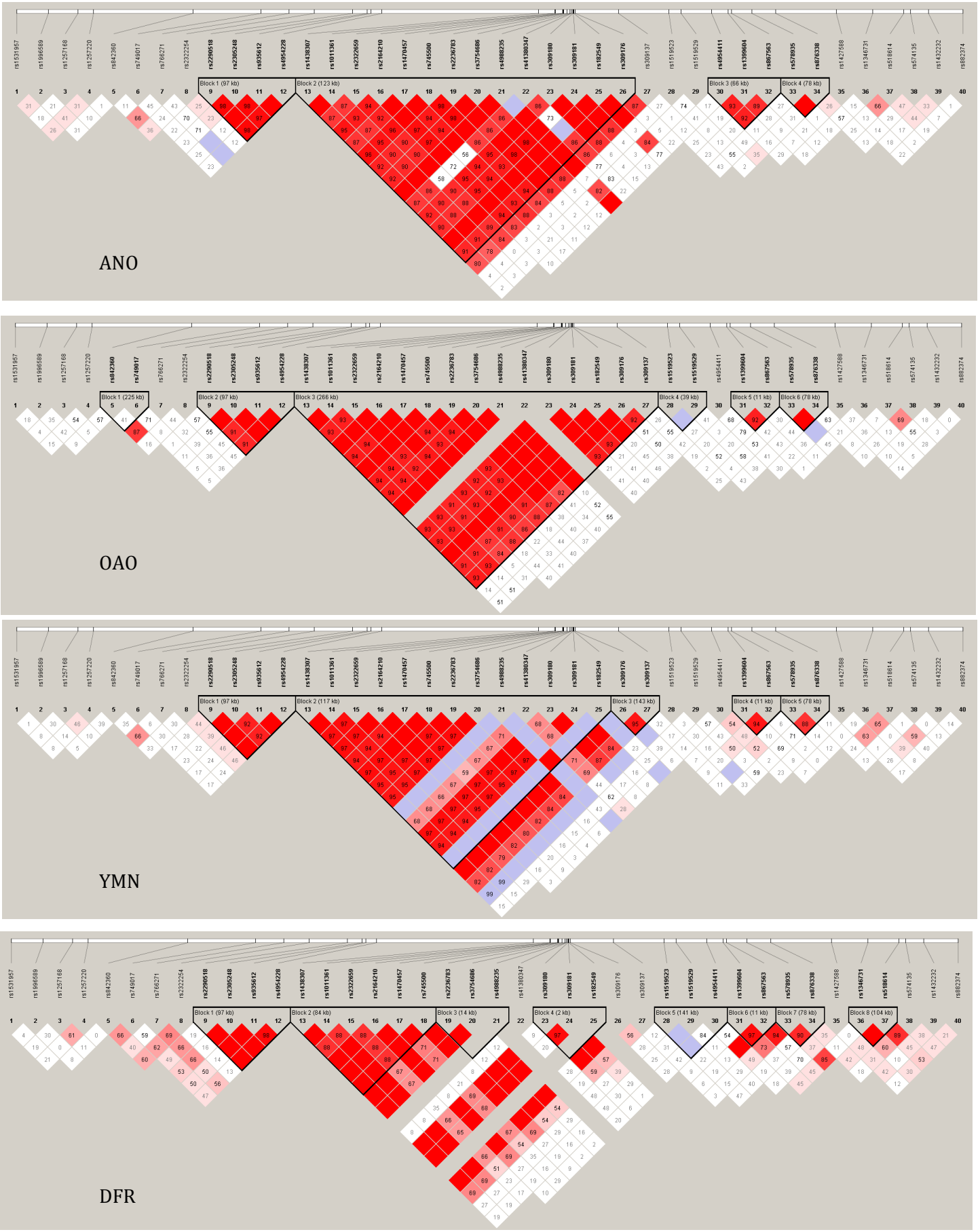


Figure 4.2.2.1: Plot of LD in ANO, OAO, YMN and DFR.

4.3 Allele Frequency Analysis

4.3.1 Italian samples

Allele frequencies were also compared among the examined Italian subgroups by applying a chi-square test. After Bonferroni correction for multiple testing, significant results were obtained only for comparisons between Northern and Southern subsamples (i.e. NCWI/NEI versus CESI/SARD), whereas no differences were observed for NCWI and NEI frequency patterns, as well as for CESI and SARD ones (Appendix table 2 and 3).

A notable difference in NCWI and CESI allele frequencies was found, with a total of five polymorphisms showing highly significant p-values (rs4988235 adjusted $p=5.91*10^{-6}$, rs182549 adjusted $p=1.91*10^{-4}$, rs745500 adjusted $p=4.36*10^{-3}$, rs309180 adjusted $p=5.03*10^{-3}$, rs2236783 adjusted $p=5.99*10^{-3}$) and being located within the previously described region of high LD in NCWI subsample (266 kb) and laying in the larger linkage block (184 kb) reconstructed in the CESI. The most significant frequency difference was especially observed for the functional variants -13,910C/T (rs4988235) and -22,018G/A (rs182549) (Table 4.3.1.1). In fact, the frequency of the -13,910*T allele ranged from 0.101 (CESI) to 0.245 (NEI) and 0.276 (NCWI) along the Italian peninsula, whereas it reached a value of 0.054 in Sardinia.

These functional variants were the sole SNPs showing significant allele frequency differences in comparisons between NEI and CESI (rs4988235 adjusted $p=1.05*10^{-4}$ and rs182549 adjusted $p=1.87*10^{-4}$) (Table 4.3.1.2), as well as NCWI and SARD (rs4988235 adjusted $p=4.91*10^{-4}$ and rs182549 adjusted $p=3.62*10^{-4}$) (Table 4.3.1.3). NEI and SARD comparison shows significant value for -13,910*C/T (rs4988235, adjusted $p=2.75*10^{-3}$) and in -22,018 *G/A (rs182549, $4.96*10^{-4}$) and an additional variant (rs1996589, adjusted $p=9.65*10^{-3}$). (Table 4.3.1.4).

Table 4.3.1.1 : Comparison of allele frequencies between CESI-NCWI.

CHR	SNP	BP	A1	CESI	NCWI	CHISQ	P	ADJ. P
2	**rs4988235**	**136608646**	**T**	**0.101**	**0.276**	**27.570**	**1.52E-07**	**5.91E-06**
2	**rs182549**	**136616754**	**T**	**0.120**	**0.276**	**20.880**	**4.90E-06**	**1.91E-04**
2	**rs745500**	**136583192**	**T**	**0.293**	**0.457**	**14.930**	**1.12E-04**	**4.36E-03**
2	**rs309180**	**136614255**	**A**	**0.277**	**0.438**	**14.660**	**1.29E-04**	**5.03E-03**
2	**rs2236783**	**136594158**	**T**	**0.296**	**0.457**	**14.330**	**1.54E-04**	**5.99E-03**
2	rs309181	136614813	G	0.277	0.429	13.050	3.04E-04	0.012
2	rs3754686	136603276	G	0.296	0.448	12.750	3.56E-04	0.014
2	rs309176	136622216	C	0.282	0.429	11.990	5.35E-04	0.021
2	rs2322659	136555659	C	0.377	0.529	11.840	5.79E-04	0.023
2	rs309137	136765951	T	0.272	0.419	11.810	5.89E-04	0.023
2	rs1470457	136581848	T	0.377	0.524	11.030	8.95E-04	0.035
2	rs1011361	136553639	T	0.355	0.491	9.554	2.00E-03	0.078
2	rs2322254	135750849	C	0.509	0.376	9.051	2.63E-03	0.102
2	rs3213889	136511575	G	0.361	0.491	8.759	3.08E-03	0.120
2	rs2164210	136580287	G	0.292	0.421	8.677	3.22E-03	0.126
2	rs1438307	136499166	A	0.361	0.486	8.139	4.33E-03	0.169
2	rs4954411	137076425	T	0.431	0.552	7.489	6.21E-03	0.242
2	rs1257168	134963892	T	0.289	0.404	7.281	6.97E-03	0.272
2	rs749017	135573659	C	0.362	0.482	6.427	0.011	0.438
2	rs766271	135667131	C	0.173	0.262	6.067	0.014	0.537
2	rs1257220	135015347	A	0.322	0.359	0.742	0.389	1.000
2	rs1346731	137612587	A	0.173	0.186	0.141	0.708	1.000
2	rs1399604	137130665	G	0.264	0.257	0.032	0.858	1.000
2	rs1427588	137492326	G	0.289	0.307	0.201	0.654	1.000
2	rs1432232	137799664	A	0.280	0.271	0.045	0.832	1.000
2	rs1519523	136934449	T	0.198	0.257	2.555	0.110	1.000
2	rs1519529	136974257	G	0.167	0.205	1.234	0.267	1.000
2	rs1531957	134759307	T	0.239	0.281	1.172	0.279	1.000
2	rs1996589	134865196	G	0.386	0.419	0.572	0.450	1.000
2	rs2290518	135878814	T	0.270	0.229	1.171	0.279	1.000
2	rs4954228	135976498	G	0.205	0.181	0.467	0.495	1.000
2	rs518614	137716851	C	0.440	0.495	1.544	0.214	1.000
2	rs574135	137740120	C	0.491	0.505	0.102	0.750	1.000
2	rs578935	137210991	C	0.248	0.271	0.350	0.554	1.000
2	rs842360	135347885	T	0.296	0.371	3.195	0.074	1.000
2	rs867563	137142500	G	0.263	0.248	0.150	0.699	1.000
2	rs876338	137289147	A	0.469	0.519	1.291	0.256	1.000
2	rs882374	137913295	A	0.253	0.232	0.286	0.593	1.000
2	rs935612	135941503	C	0.217	0.181	1.016	0.314	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in NCWI population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.1.2: Comparison of allele frequencies between CESI-NEI.

CHR	SNP	BP	MA	CESI	NEI	CHISQ	P	ADJ. P
2	**rs4988235**	**136608646**	**T**	**0.101**	**0.245**	**22.020**	**2.70E-06**	**1.05E-04**
2	**rs182549**	**136616754**	**T**	**0.120**	**0.266**	**20.910**	**4.80E-06**	**1.87E-04**
2	rs842360	135347885	T	0.296	0.423	10.350	1.29E-03	0.050
2	rs309137	136765951	T	0.272	0.392	9.324	2.26E-03	0.088
2	rs1519529	136974257	G	0.167	0.268	9.031	2.65E-03	0.104
2	rs3754686	136603276	G	0.296	0.410	8.555	3.45E-03	0.134
2	rs745500	136583192	T	0.293	0.407	8.526	3.50E-03	0.137
2	rs2164210	136580287	G	0.292	0.405	8.053	4.54E-03	0.177
2	rs2236783	136594158	T	0.296	0.407	8.044	4.57E-03	0.178
2	rs309180	136614255	A	0.277	0.381	7.387	6.57E-03	0.256
2	rs309181	136614813	G	0.277	0.378	6.906	8.59E-03	0.335
2	rs309176	136622216	C	0.282	0.380	6.429	0.011	0.438
2	rs1346731	137612587	A	0.173	0.101	6.285	0.012	0.475
2	rs1011361	136553639	T	0.355	0.450	5.498	0.019	0.742
2	rs2322254	135750849	C	0.509	0.414	5.468	0.019	0.755
2	rs2322659	136555659	C	0.377	0.471	5.391	0.020	0.789
2	rs3213889	136511575	G	0.361	0.453	5.253	0.022	0.855
2	rs1257168	134963892	T	0.289	0.377	5.101	0.024	0.932
2	rs1257220	135015347	A	0.322	0.283	1.025	0.311	1.000
2	rs1399604	137130665	G	0.264	0.326	2.737	0.098	1.000
2	rs1427588	137492326	G	0.289	0.298	0.061	0.805	1.000
2	rs1432232	137799664	A	0.280	0.234	1.642	0.200	1.000
2	rs1438307	136499166	A	0.361	0.450	4.859	0.028	1.000
2	rs1470457	136581848	T	0.377	0.450	3.201	0.074	1.000
2	rs1519523	136934449	T	0.198	0.205	0.044	0.834	1.000
2	rs1531957	134759307	T	0.239	0.290	1.973	0.160	1.000
2	rs1996589	134865196	G	0.386	0.330	2.032	0.154	1.000
2	rs2290518	135878814	T	0.270	0.282	0.096	0.757	1.000
2	rs4954228	135976498	G	0.205	0.226	0.387	0.534	1.000
2	rs4954411	137076425	T	0.431	0.489	2.037	0.154	1.000
2	rs518614	137716851	C	0.440	0.409	0.559	0.455	1.000
2	rs574135	137740120	C	0.491	0.435	1.848	0.174	1.000
2	rs578935	137210991	C	0.248	0.263	0.157	0.692	1.000
2	rs749017	135573659	C	0.362	0.423	2.143	0.143	1.000
2	rs766271	135667131	C	0.173	0.245	4.649	0.031	1.000
2	rs867563	137142500	G	0.263	0.295	0.769	0.380	1.000
2	rs876338	137289147	A	0.469	0.514	1.247	0.264	1.000
2	rs882374	137913295	A	0.253	0.276	0.390	0.532	1.000
2	rs935612	135941503	C	0.217	0.234	0.241	0.623	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in NEI population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.1.3 : Comparison of allele frequencies between SARD-NCWI.

CHR	SNP	BP	A1	SARD	NCWI	A2	CHISQ	P	ADJ. P
2	**rs182549**	**136616754**	**T**	**0.053**	**0.276**	**C**	**19.650**	**9.28E-06**	**3.62E-04**
2	**rs4988235**	**136608646**	**T**	**0.054**	**0.276**	**C**	**19.070**	**1.26E-05**	**4.91E-04**
2	rs1519523	136934449	T	0.096	0.257	G	10.300	1.33E-03	0.052
2	rs1470457	136581848	T	0.340	0.524	C	8.773	3.06E-03	0.119
2	rs2322659	136555659	C	0.351	0.529	T	8.207	4.17E-03	0.163
2	rs766271	135667131	C	0.117	0.262	T	8.019	4.63E-03	0.181
2	rs2236783	136594158	T	0.287	0.457	C	7.782	5.28E-03	0.206
2	rs745500	136583192	T	0.287	0.457	C	7.782	5.28E-03	0.206
2	rs309180	136614255	A	0.277	0.438	G	7.131	7.58E-03	0.295
2	rs3754686	136603276	G	0.287	0.448	A	6.971	8.28E-03	0.323
2	rs309176	136622216	C	0.277	0.429	T	6.356	0.012	0.456
2	rs309181	136614813	G	0.277	0.429	C	6.356	0.012	0.456
2	rs1011361	136553639	T	0.415	0.491	C	1.490	0.222	1.000
2	rs1257168	134963892	T	0.478	0.404	C	1.414	0.234	1.000
2	rs1257220	135015347	A	0.370	0.359	G	0.033	0.856	1.000
2	rs1346731	137612587	A	0.096	0.186	T	3.953	0.047	1.000
2	rs1399604	137130665	G	0.287	0.257	T	0.301	0.583	1.000
2	rs1427588	137492326	G	0.266	0.307	A	0.519	0.471	1.000
2	rs1432232	137799664	A	0.298	0.271	C	0.225	0.635	1.000
2	rs1438307	136499166	A	0.394	0.486	C	2.219	0.136	1.000
2	rs1519529	136974257	G	0.239	0.205	C	0.447	0.504	1.000
2	rs1531957	134759307	T	0.330	0.281	C	0.743	0.389	1.000
2	rs1996589	134865196	G	0.543	0.419	A	3.991	0.046	1.000
2	rs2164210	136580287	G	0.294	0.421	A	4.284	0.038	1.000
2	rs2290518	135878814	T	0.277	0.229	C	0.813	0.367	1.000
2	rs2322254	135750849	C	0.468	0.376	T	2.276	0.131	1.000
2	rs309137	136765951	T	0.294	0.419	C	4.216	0.040	1.000
2	rs3213889	136511575	G	0.394	0.491	A	2.452	0.117	1.000
2	rs4954228	135976498	G	0.234	0.181	A	1.155	0.282	1.000
2	rs4954411	137076425	C	0.500	0.448	T	0.716	0.397	1.000
2	rs518614	137716851	C	0.415	0.495	G	1.676	0.196	1.000
2	rs574135	137740120	C	0.447	0.505	T	0.873	0.350	1.000
2	rs578935	137210991	C	0.277	0.271	T	0.009	0.926	1.000
2	rs749017	135573659	C	0.359	0.482	G	3.706	0.054	1.000
2	rs842360	135347885	T	0.238	0.371	C	4.745	0.029	1.000
2	rs867563	137142500	G	0.245	0.248	A	0.003	0.956	1.000
2	rs876338	137289147	G	0.489	0.481	A	0.018	0.892	1.000
2	rs882374	137913295	A	0.250	0.232	G	0.108	0.742	1.000
2	rs935612	135941503	C	0.234	0.181	T	1.155	0.282	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in NCWI population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.1.4: Comparison of allele frequencies between SARD-NEI.

CHR	SNP	BP	A1	SARD	NEI	CHISQ	P	ADJ. P
2	**rs182549**	**136616754**	**T**	**0.053**	**0.266**	**19.050**	**1.27E-05**	**4.96E-04**
2	**rs4988235**	**136608646**	**T**	**0.054**	**0.245**	**15.800**	**7.04E-05**	**2.75E-03**
2	**rs1996589**	**134865196**	**G**	**0.543**	**0.330**	**13.430**	**2.47E-04**	**9.65E-03**
2	rs842360	135347885	T	0.238	0.423	9.367	2.21E-03	0.086
2	rs766271	135667131	C	0.117	0.245	6.836	8.93E-03	0.348
2	rs1519523	136934449	T	0.096	0.205	5.749	0.016	0.643
2	rs1011361	136553639	T	0.415	0.450	0.344	0.558	1.000
2	rs1257168	134963892	T	0.478	0.377	2.927	0.087	1.000
2	rs1257220	135015347	A	0.370	0.283	2.426	0.119	1.000
2	rs1346731	137612587	A	0.096	0.101	0.025	0.874	1.000
2	rs1399604	137130665	G	0.287	0.326	0.490	0.484	1.000
2	rs1427588	137492326	G	0.266	0.298	0.344	0.557	1.000
2	rs1432232	137799664	A	0.298	0.234	1.537	0.215	1.000
2	rs1438307	136499166	A	0.394	0.450	0.897	0.344	1.000
2	rs1470457	136581848	T	0.340	0.450	3.435	0.064	1.000
2	rs1519529	136974257	G	0.239	0.268	0.305	0.581	1.000
2	rs1531957	134759307	T	0.330	0.290	0.532	0.466	1.000
2	rs2164210	136580287	G	0.294	0.405	3.635	0.057	1.000
2	rs2236783	136594158	T	0.287	0.407	4.255	0.039	1.000
2	rs2290518	135878814	T	0.277	0.282	0.010	0.921	1.000
2	rs2322254	135750849	C	0.468	0.414	0.850	0.357	1.000
2	rs2322659	136555659	C	0.351	0.471	4.093	0.043	1.000
2	rs309137	136765951	T	0.294	0.392	2.851	0.091	1.000
2	rs309176	136622216	C	0.277	0.380	3.306	0.069	1.000
2	rs309180	136614255	A	0.277	0.381	3.364	0.067	1.000
2	rs309181	136614813	G	0.277	0.378	3.147	0.076	1.000
2	rs3213889	136511575	G	0.394	0.453	1.014	0.314	1.000
2	rs3754686	136603276	G	0.287	0.410	4.503	0.034	1.000
2	rs4954228	135976498	G	0.234	0.226	0.024	0.877	1.000
2	rs4954411	137076425	T	0.500	0.489	0.033	0.856	1.000
2	rs518614	137716851	C	0.415	0.409	0.009	0.926	1.000
2	rs574135	137740120	C	0.447	0.435	0.041	0.839	1.000
2	rs578935	137210991	C	0.277	0.263	0.071	0.791	1.000
2	rs745500	136583192	T	0.287	0.407	4.255	0.039	1.000
2	rs749017	135573659	C	0.359	0.423	1.168	0.280	1.000
2	rs867563	137142500	G	0.245	0.295	0.877	0.349	1.000
2	rs876338	137289147	G	0.489	0.486	0.004	0.950	1.000
2	rs882374	137913295	A	0.250	0.276	0.237	0.626	1.000
2	rs935612	135941503	C	0.234	0.234	2.07E-05	0.996	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in NEI population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

4.3.2 Samples from Arabian Peninsula

In addition to comparison of LD patterns, allele frequencies were also compared among the Arabian subgroups, by applying a chi-squared test. After Bonferroni correction for multiple testing, significant results were obtained for all the performed pairwise comparisons.

In particular, ANO showed the highest difference with DFR (adjusted $p=1.32x10^{-54}$) and YMN (adjusted $p= 2.67x10^{-33}$) for the -13,915 T/G variant (rs41380347), with the adaptive allele being more represented in DFR (allelic frequencies DFR=0.710, YMN=0.567 and ANO=0.128) and for respectively 24 and 9 additional SNPs (Table 4.3.2.1 and Table 4.3.2.2). On the contrary, significant differences in allele frequencies between ANO and OAO were found for 14 SNPs, among which the -13,910 C/T and -22,018 G/A variants showed the highest differences (adjusted p-values of $8.98x10^{-12}$ and $1.52x10^{-11}$), whereas a non-significant result was obtained for the Arabian specific mutation after Bonferroni correction for multiple testing (Table 4.3.2.3).

Interestingly, the comparison of the OAO sample with DFR and YMN revealed a particular pattern. In fact, all the three functional variants showed significantly different allele frequencies, in addition to 23 SNPs which differentiated OAO and DFR, as well as to 16 which differentiated OAO and YMN (Table 4.3.2.4 and Table 4.3.2.5).

Finally, the comparison between DFR and YMN pointed out only five significant differences in allele frequencies, including that concerning the -13,915 adaptive allele (adjust $p=6.77x10^{-3}$) (Table 4.3.2.6).

Table 4.3.2.1 : Comparison of allele frequencies between ANO-DFR.

CHR	SNP	BP	A1	DFR	ANO	CHISQ	P	ADJ. P
2	**rs41380347***	**136608651**	**G**	**0.710**	**0.128**	**249.500**	**3.31E-56**	**1.32E-54**
2	**rs749017**	**135573659**	**G**	**0.313**	**0.664**	**102.800**	**3.78E-24**	**1.51E-22**
2	**rs1996589**	**134865196**	**A**	**0.286**	**0.586**	**75.870**	**3.03E-18**	**1.21E-16**
2	**rs1519529**	**136974257**	**G**	**0.017**	**0.166**	**54.300**	**1.72E-13**	**6.87E-12**
2	**rs2322659**	**136555659**	**C**	**0.099**	**0.282**	**44.900**	**2.07E-11**	**8.28E-10**
2	**rs578935**	**137210991**	**C**	**0.552**	**0.325**	**43.660**	**3.90E-11**	**1.56E-09**
2	**rs842360**	**135347885**	**T**	**0.086**	**0.254**	**40.070**	**2.45E-10**	**9.79E-09**
2	**rs1470457**	**136581848**	**T**	**0.106**	**0.274**	**37.650**	**8.46E-10**	**3.38E-08**
2	**rs2322254**	**135750849**	**T**	**0.210**	**0.399**	**34.680**	**3.89E-09**	**1.56E-07**
2	**rs518614**	**137716851**	**G**	**0.361**	**0.561**	**33.210**	**8.26E-09**	**3.30E-07**
2	**rs1438307**	**136499166**	**A**	**0.079**	**0.209**	**28.310**	**1.03E-07**	**4.13E-06**
2	**rs1011361**	**136553639**	**T**	**0.079**	**0.208**	**28.010**	**1.21E-07**	**4.83E-06**
2	**rs867563**	**137142500**	**G**	**0.121**	**0.262**	**26.540**	**2.59E-07**	**1.03E-05**
2	**rs2164210**	**136580287**	**G**	**0.079**	**0.201**	**25.300**	**4.90E-07**	**1.96E-05**
2	**rs745500**	**136583192**	**T**	**0.079**	**0.199**	**24.790**	**6.41E-07**	**2.56E-05**
2	**rs574135**	**137740120**	**T**	**0.317**	**0.484**	**23.970**	**9.78E-07**	**3.91E-05**
2	**rs1399604**	**137130665**	**G**	**0.138**	**0.269**	**21.830**	**2.98E-06**	**1.19E-04**
2	**rs1432232**	**137799664**	**A**	**0.180**	**0.318**	**21.110**	**4.33E-06**	**1.73E-04**
2	**rs876338**	**137289147**	**A**	**0.192**	**0.322**	**18.190**	**2.00E-05**	**8.02E-04**
2	**rs2290518**	**135878814**	**T**	**0.179**	**0.303**	**17.040**	**3.67E-05**	**1.47E-03**
2	**rs2236783**	**136594158**	**T**	**0.103**	**0.203**	**15.880**	**6.75E-05**	**2.70E-03**
2	**rs3754686**	**136603276**	**G**	**0.103**	**0.203**	**15.880**	**6.75E-05**	**2.70E-03**
2	**rs2305248**	**135928312**	**G**	**0.191**	**0.308**	**15.090**	**1.02E-04**	**4.09E-03**
2	**rs935612**	**135941503**	**C**	**0.157**	**0.265**	**14.560**	**1.36E-04**	**5.44E-03**
2	**rs4954228**	**135976498**	**G**	**0.158**	**0.265**	**14.270**	**1.59E-04**	**6.34E-03**
2	rs309137	136765951	T	0.106	0.194	12.380	4.34E-04	0.017
2	rs182549*	136616754	T	0.000	0.026	10.520	1.18E-03	0.047
2	rs309180	136614255	A	0.116	0.196	10.200	1.41E-03	0.056
2	rs309181	136614813	G	0.119	0.200	9.951	1.61E-03	0.064
2	rs4954411	137076425	T	0.286	0.389	9.909	1.65E-03	0.066
2	rs4988235*	136608646	T	0.000	0.021	8.671	3.23E-03	0.129
2	rs309176	136622216	C	0.112	0.199	8.265	4.04E-03	0.162
2	rs1519523	136934449	T	0.047	0.098	8.099	4.43E-03	0.177
2	rs1257168	134963892	T	0.121	0.167	3.560	0.059	1.000
2	rs1257220	135015347	A	0.153	0.178	0.933	0.334	1.000
2	rs1346731	137612587	A	0.121	0.170	3.904	0.048	1.000
2	rs1427588	137492326	G	0.214	0.265	2.988	0.084	1.000
2	rs1531957	134759307	T	0.170	0.197	0.987	0.321	1.000
2	rs766271	135667131	C	0.054	0.090	3.942	0.047	1.000
2	rs882374	137913295	A	0.163	0.221	4.517	0.034	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in ANO population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.2.2 : Comparison of allele frequencies between ANO-YMN.

CHR	SNP	BP	A1	YMN	ANO	CHISQ	P	ADJ. P
2	**rs41380347***	**136608651**	**G**	**0.567**	**0.128**	**151.900**	**6.68E-35**	**2.67E-33**
2	**rs4954411**	**137076425**	**T**	**0.602**	**0.389**	**34.370**	**4.56E-09**	**1.83E-07**
2	**rs1996589**	**134865196**	**G**	**0.605**	**0.414**	**27.860**	**1.31E-07**	**5.22E-06**
2	**rs749017**	**135573659**	**C**	**0.512**	**0.336**	**24.220**	**8.57E-07**	**3.43E-05**
2	**rs2322254**	**135750849**	**T**	**0.234**	**0.399**	**23.560**	**1.21E-06**	**4.83E-05**
2	**rs1519529**	**136974257**	**G**	**0.055**	**0.166**	**22.670**	**1.92E-06**	**7.68E-05**
2	**rs2305248**	**135928312**	**G**	**0.181**	**0.308**	**16.060**	**6.12E-05**	**2.45E-03**
2	**rs935612**	**135941503**	**C**	**0.154**	**0.265**	**13.910**	**1.91E-04**	**7.66E-03**
2	**rs4954228**	**135976498**	**G**	**0.154**	**0.265**	**13.900**	**1.93E-04**	**7.72E-03**
2	**rs842360**	**135347885**	**T**	**0.145**	**0.254**	**13.490**	**2.40E-04**	**9.60E-03**
2	rs2290518	135878814	T	0.188	0.303	13.210	2.78E-04	0.011
2	rs867563	137142500	G	0.161	0.262	11.370	7.46E-04	0.030
2	rs578935	137210991	C	0.221	0.325	10.230	1.38E-03	0.055
2	rs1427588	137492326	G	0.177	0.265	8.409	3.73E-03	0.149
2	rs518614	137716851	C	0.541	0.439	7.858	5.06E-03	0.202
2	rs1399604	137130665	G	0.188	0.269	6.865	8.79E-03	0.352
2	rs1470457	136581848	T	0.201	0.274	5.535	1.86E-02	0.746
2	rs882374	137913295	A	0.155	0.221	5.289	0.021	0.859
2	rs1011361	136553639	T	0.154	0.208	3.688	0.055	1.000
2	rs1257168	134963892	T	0.183	0.167	0.359	0.549	1.000
2	rs1257220	135015347	A	0.181	0.178	0.018	0.894	1.000
2	rs1346731	137612587	A	0.151	0.170	0.488	0.485	1.000
2	rs1432232	137799664	A	0.252	0.318	4.072	0.044	1.000
2	rs1438307	136499166	A	0.154	0.209	3.807	0.051	1.000
2	rs1519523	136934449	T	0.061	0.098	3.500	0.061	1.000
2	rs1531957	134759307	T	0.189	0.197	0.073	0.788	1.000
2	rs182549*	136616754	T	0.009	0.026	3.088	0.079	1.000
2	rs2164210	136580287	G	0.151	0.201	3.218	0.073	1.000
2	rs2236783	136594158	T	0.154	0.203	3.110	0.078	1.000
2	rs2322659	136555659	C	0.221	0.282	3.703	0.054	1.000
2	rs309137	136765951	T	0.145	0.194	3.156	0.076	1.000
2	rs309176	136622216	C	0.149	0.199	2.083	0.149	1.000
2	rs309180	136614255	A	0.142	0.196	3.874	0.049	1.000
2	rs309181	136614813	G	0.147	0.200	3.586	0.058	1.000
2	rs3754686	136603276	G	0.151	0.203	3.508	0.061	1.000
2	rs4988235*	136608646	T	0.006	0.021	3.169	0.075	1.000
2	rs574135	137740120	T	0.412	0.484	3.659	0.056	1.000
2	rs745500	136583192	T	0.151	0.199	2.940	0.086	1.000
2	rs766271	135667131	C	0.048	0.090	4.800	0.028	1.000
2	rs876338	137289147	A	0.259	0.322	3.628	0.057	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in ANO population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.2.3: Comparison of allele frequencies between ANO-OAO.

CHR	SNP	BP	A1	OAO	ANO	CHISQ	P	ADJ. P
2	**rs4988235***	**136608646**	**T**	**0.220**	**0.021**	**53.780**	**2.25E-13**	**8.98E-12**
2	**rs182549***	**136616754**	**T**	**0.232**	**0.026**	**52.750**	**3.79E-13**	**1.52E-11**
2	**rs766271**	**135667131**	**C**	**0.293**	**0.090**	**26.130**	**3.19E-07**	**1.28E-05**
2	**rs2236783**	**136594158**	**T**	**0.439**	**0.203**	**20.900**	**4.84E-06**	**1.94E-04**
2	**rs745500**	**136583192**	**T**	**0.427**	**0.199**	**19.920**	**8.07E-06**	**3.23E-04**
2	**rs1011361**	**136553639**	**T**	**0.439**	**0.208**	**19.860**	**8.33E-06**	**3.33E-04**
2	**rs1438307**	**136499166**	**A**	**0.439**	**0.209**	**19.630**	**9.42E-06**	**3.77E-04**
2	**rs2164210**	**136580287**	**G**	**0.427**	**0.201**	**19.400**	**1.06E-05**	**4.23E-04**
2	**rs3754686**	**136603276**	**G**	**0.427**	**0.203**	**18.900**	**1.38E-05**	**5.52E-04**
2	**rs309137**	**136765951**	**T**	**0.402**	**0.194**	**17.030**	**3.69E-05**	**1.47E-03**
2	**rs2322659**	**136555659**	**C**	**0.512**	**0.282**	**16.820**	**4.11E-05**	**1.64E-03**
2	**rs309180**	**136614255**	**A**	**0.402**	**0.196**	**16.550**	**4.75E-05**	**1.90E-03**
2	**rs309181**	**136614813**	**G**	**0.402**	**0.200**	**15.880**	**6.76E-05**	**2.71E-03**
2	**rs309176**	**136622216**	**C**	**0.419**	**0.199**	**14.780**	**1.21E-04**	**4.83E-03**
2	rs842360	135347885	T	0.449	0.254	12.270	4.60E-04	0.018
2	rs1470457	136581848	T	0.463	0.274	11.690	6.27E-04	0.025
2	rs41380347*	136608651	G	0.000	0.128	11.690	6.27E-04	0.025
2	rs1257168	134963892	T	0.329	0.167	11.680	6.33E-04	0.025
2	rs749017	135573659	C	0.512	0.336	9.253	2.35E-03	0.094
2	rs1519529	136974257	G	0.305	0.166	8.700	3.18E-03	0.127
2	rs1257220	135015347	A	0.256	0.178	2.755	0.097	1.000
2	rs1346731	137612587	A	0.268	0.170	4.406	0.036	1.000
2	rs1399604	137130665	G	0.305	0.269	0.447	0.504	1.000
2	rs1427588	137492326	G	0.232	0.265	0.405	0.524	1.000
2	rs1432232	137799664	A	0.366	0.318	0.725	0.395	1.000
2	rs1519523	136934449	T	0.183	0.098	4.984	0.026	1.000
2	rs1531957	134759307	T	0.195	0.197	0.001	0.974	1.000
2	rs1996589	134865196	G	0.402	0.414	0.035	0.851	1.000
2	rs2290518	135878814	T	0.230	0.303	1.646	0.200	1.000
2	rs2305248	135928312	G	0.232	0.308	1.899	0.168	1.000
2	rs2322254	135750849	T	0.488	0.399	2.233	0.135	1.000
2	rs4954228	135976498	G	0.195	0.265	1.790	0.181	1.000
2	rs4954411	137076425	T	0.402	0.389	0.051	0.821	1.000
2	rs518614	137716851	C	0.439	0.439	0.000	0.997	1.000
2	rs574135	137740120	T	0.513	0.484	0.226	0.635	1.000
2	rs578935	137210991	C	0.293	0.325	0.326	0.568	1.000
2	rs867563	137142500	G	0.268	0.262	0.015	0.903	1.000
2	rs876338	137289147	A	0.427	0.322	3.400	0.065	1.000
2	rs882374	137913295	A	0.268	0.221	0.886	0.347	1.000
2	rs935612	135941503	C	0.195	0.265	1.785	0.182	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in ANO population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.2.4: Comparison of allele frequencies between OAO-DFR.

CHR	SNP	BP	A1	DFR	OAO	CHISQ	P	ADJ. P
2	**rs41380347***	**136608651**	**T**	**0.290**	**1.000**	**134.300**	**4.75E-31**	**1.90E-29**
2	**rs182549***	**136616754**	**T**	**0.000**	**0.232**	**97.420**	**5.61E-23**	**2.25E-21**
2	**rs4988235***	**136608646**	**T**	**0.000**	**0.220**	**92.540**	**6.61E-22**	**2.65E-20**
2	**rs1519529**	**136974257**	**G**	**0.017**	**0.305**	**92.120**	**8.18E-22**	**3.27E-20**
2	**rs2322659**	**136555659**	**C**	**0.099**	**0.512**	**83.510**	**6.34E-20**	**2.54E-18**
2	**rs1011361**	**136553639**	**T**	**0.079**	**0.439**	**73.810**	**8.61E-18**	**3.44E-16**
2	**rs1438307**	**136499166**	**A**	**0.079**	**0.439**	**73.810**	**8.61E-18**	**3.44E-16**
2	**rs745500**	**136583192**	**T**	**0.079**	**0.427**	**69.760**	**6.71E-17**	**2.68E-15**
2	**rs2164210**	**136580287**	**G**	**0.079**	**0.427**	**69.300**	**8.44E-17**	**3.38E-15**
2	**rs842360**	**135347885**	**T**	**0.086**	**0.449**	**68.980**	**9.93E-17**	**3.97E-15**
2	**rs1470457**	**136581848**	**T**	**0.106**	**0.463**	**62.990**	**2.08E-15**	**8.33E-14**
2	**rs2236783**	**136594158**	**T**	**0.103**	**0.439**	**57.210**	**3.92E-14**	**1.57E-12**
2	**rs3754686**	**136603276**	**G**	**0.103**	**0.427**	**53.690**	**2.35E-13**	**9.41E-12**
2	**rs309137**	**136765951**	**T**	**0.106**	**0.402**	**45.270**	**1.72E-11**	**6.86E-10**
2	**rs766271**	**135667131**	**C**	**0.054**	**0.293**	**44.810**	**2.17E-11**	**8.68E-10**
2	**rs309180**	**136614255**	**A**	**0.116**	**0.402**	**40.910**	**1.60E-10**	**6.39E-09**
2	**rs309181**	**136614813**	**G**	**0.119**	**0.402**	**39.250**	**3.73E-10**	**1.49E-08**
2	**rs309176**	**136622216**	**C**	**0.112**	**0.419**	**32.790**	**1.03E-08**	**4.11E-07**
2	**rs1996589**	**134865196**	**A**	**0.286**	**0.598**	**29.500**	**5.60E-08**	**2.24E-06**
2	**rs2322254**	**135750849**	**T**	**0.210**	**0.488**	**27.460**	**1.61E-07**	**6.43E-06**
2	**rs1257168**	**134963892**	**T**	**0.121**	**0.329**	**22.570**	**2.02E-06**	**8.09E-05**
2	**rs876338**	**137289147**	**A**	**0.192**	**0.427**	**21.120**	**4.31E-06**	**1.73E-04**
2	**rs1519523**	**136934449**	**T**	**0.047**	**0.183**	**19.500**	**1.00E-05**	**4.02E-04**
2	**rs578935**	**137210991**	**T**	**0.448**	**0.707**	**18.320**	**1.87E-05**	**7.49E-04**
2	**rs1432232**	**137799664**	**A**	**0.180**	**0.366**	**14.180**	**1.66E-04**	**6.64E-03**
2	**rs1399604**	**137130665**	**G**	**0.138**	**0.305**	**13.740**	**2.11E-04**	**8.42E-03**
2	rs867563	137142500	G	0.121	0.268	11.960	5.45E-04	0.022
2	rs1346731	137612587	A	0.121	0.268	11.810	5.90E-04	0.024
2	rs518614	137716851	G	0.361	0.561	11.360	7.50E-04	0.030
2	rs574135	137740120	T	0.317	0.513	11.250	7.96E-04	0.032
2	rs749017	135573659	G	0.313	0.488	9.281	2.32E-03	0.093
2	rs1257220	135015347	A	0.153	0.256	5.166	0.023	0.921
2	rs882374	137913295	A	0.163	0.268	5.160	0.023	0.925
2	rs1427588	137492326	G	0.214	0.232	0.127	0.722	1.000
2	rs1531957	134759307	T	0.170	0.195	0.301	0.584	1.000
2	rs2290518	135878814	T	0.179	0.230	1.073	0.300	1.000
2	rs2305248	135928312	G	0.191	0.232	0.727	0.394	1.000
2	rs4954228	135976498	G	0.158	0.195	0.703	0.402	1.000
2	rs4954411	137076425	T	0.286	0.402	4.382	0.036	1.000
2	rs935612	135941503	C	0.157	0.195	0.736	0.391	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in OAO population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.2.5: Comparison of allele frequencies between OAO-YMN.

CHR	SNP	BP	A1	YMN	OAO	CHISQ	P	ADJ. P
2	**rs41380347***	**136608651**	**G**	**0.567**	**0.000**	**84.440**	**3.96E-20**	**1.59E-18**
2	**rs4988235***	**136608646**	**T**	**0.006**	**0.220**	**67.580**	**2.02E-16**	**8.09E-15**
2	**rs182549***	**136616754**	**T**	**0.009**	**0.232**	**67.230**	**2.42E-16**	**9.69E-15**
2	**rs766271**	**135667131**	**C**	**0.048**	**0.293**	**44.680**	**2.32E-11**	**9.28E-10**
2	**rs1519529**	**136974257**	**G**	**0.055**	**0.305**	**44.560**	**2.47E-11**	**9.88E-10**
2	**rs842360**	**135347885**	**T**	**0.145**	**0.449**	**36.280**	**1.71E-09**	**6.82E-08**
2	**rs1011361**	**136553639**	**T**	**0.154**	**0.439**	**32.530**	**1.17E-08**	**4.69E-07**
2	**rs1438307**	**136499166**	**A**	**0.154**	**0.439**	**32.530**	**1.17E-08**	**4.69E-07**
2	**rs2236783**	**136594158**	**T**	**0.154**	**0.439**	**32.530**	**1.17E-08**	**4.69E-07**
2	**rs2164210**	**136580287**	**G**	**0.151**	**0.427**	**30.960**	**2.63E-08**	**1.05E-06**
2	**rs3754686**	**136603276**	**G**	**0.151**	**0.427**	**30.960**	**2.63E-08**	**1.05E-06**
2	**rs745500**	**136583192**	**T**	**0.151**	**0.427**	**30.960**	**2.63E-08**	**1.05E-06**
2	**rs309180**	**136614255**	**A**	**0.142**	**0.402**	**28.800**	**8.04E-08**	**3.22E-06**
2	**rs2322659**	**136555659**	**C**	**0.221**	**0.512**	**28.050**	**1.18E-07**	**4.73E-06**
2	**rs309137**	**136765951**	**T**	**0.145**	**0.402**	**27.900**	**1.28E-07**	**5.11E-06**
2	**rs309181**	**136614813**	**G**	**0.147**	**0.402**	**27.270**	**1.77E-07**	**7.07E-06**
2	**rs1470457**	**136581848**	**T**	**0.201**	**0.463**	**24.320**	**8.16E-07**	**3.26E-05**
2	**rs2322254**	**135750849**	**T**	**0.234**	**0.488**	**21.010**	**4.57E-06**	**1.83E-04**
2	**rs309176**	**136622216**	**C**	**0.149**	**0.419**	**20.130**	**7.22E-06**	**2.89E-04**
2	rs1519523	136934449	T	0.061	0.183	12.710	3.63E-04	0.015
2	rs1996589	134865196	A	0.395	0.598	11.020	9.01E-04	0.036
2	rs4954411	137076425	C	0.398	0.598	10.690	1.08E-03	0.043
2	rs876338	137289147	A	0.259	0.427	9.068	2.60E-03	0.104
2	rs1257168	134963892	T	0.183	0.329	8.485	3.58E-03	0.143
2	rs1346731	137612587	A	0.151	0.268	6.329	0.012	0.475
2	rs882374	137913295	A	0.155	0.268	5.834	0.016	0.629
2	rs1399604	137130665	G	0.188	0.305	5.401	0.020	0.805
2	rs867563	137142500	G	0.161	0.268	5.140	0.023	0.935
2	rs1257220	135015347	A	0.181	0.256	2.351	0.125	1.000
2	rs1427588	137492326	G	0.177	0.232	1.284	0.257	1.000
2	rs1432232	137799664	A	0.252	0.366	4.355	0.037	1.000
2	rs1531957	134759307	T	0.189	0.195	0.016	0.898	1.000
2	rs2290518	135878814	T	0.188	0.230	0.689	0.407	1.000
2	rs2305248	135928312	G	0.181	0.232	1.088	0.297	1.000
2	rs4954228	135976498	G	0.154	0.195	0.822	0.365	1.000
2	rs518614	137716851	G	0.459	0.561	2.743	0.098	1.000
2	rs574135	137740120	T	0.412	0.513	2.595	0.107	1.000
2	rs578935	137210991	C	0.221	0.293	1.898	0.168	1.000
2	rs749017	135573659	G	0.488	0.488	6.60E-05	0.994	1.000
2	rs935612	135941503	C	0.154	0.195	0.822	0.365	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in OAO population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Table 4.3.2.6: Comparison of allele frequencies between DFR-YMN.

CHR	SNP	BP	A1	DFR	YMN	CHISQ	P	BONF
2	**rs578935**	**137210991**	**C**	**0.552**	**0.221**	**84.900**	**3.13E-20**	**1.25E-18**
2	**rs4954411**	**137076425**	**T**	**0.286**	**0.602**	**75.850**	**3.06E-18**	**1.22E-16**
2	**rs749017**	**135573659**	**G**	**0.313**	**0.488**	**23.970**	**9.81E-07**	**3.92E-05**
2	**rs2322659**	**136555659**	**C**	**0.099**	**0.221**	**21.340**	**3.84E-06**	**1.54E-04**
2	**rs41380347***	**136608651**	**T**	**0.290**	**0.433**	**14.140**	**1.69E-04**	**6.77E-03**
2	rs1470457	136581848	T	0.106	0.201	13.140	2.89E-04	0.012
2	rs1011361	136553639	T	0.079	0.154	10.490	1.20E-03	0.048
2	rs1438307	136499166	A	0.079	0.154	10.490	1.20E-03	0.048
2	rs1996589	134865196	A	0.286	0.395	9.916	1.64E-03	0.066
2	rs745500	136583192	T	0.079	0.151	9.800	1.75E-03	0.070
2	rs2164210	136580287	G	0.079	0.151	9.650	1.89E-03	0.076
2	rs1519529	136974257	G	0.017	0.055	8.032	4.60E-03	0.184
2	rs518614	137716851	G	0.361	0.459	7.384	6.58E-03	0.263
2	rs574135	137740120	T	0.317	0.412	6.935	8.46E-03	0.338
2	rs842360	135347885	T	0.086	0.145	6.353	0.012	0.469
2	rs1257168	134963892	T	0.121	0.183	5.717	0.017	0.672
2	rs1432232	137799664	A	0.180	0.252	5.695	0.017	0.681
2	rs1257220	135015347	A	0.153	0.181	1.096	0.295	1.000
2	rs1346731	137612587	A	0.121	0.151	1.420	0.233	1.000
2	rs1399604	137130665	G	0.138	0.188	3.469	0.063	1.000
2	rs1427588	137492326	G	0.214	0.177	1.570	0.210	1.000
2	rs1519523	136934449	T	0.047	0.061	0.749	0.387	1.000
2	rs1531957	134759307	T	0.170	0.189	0.458	0.498	1.000
2	rs182549*	136616754	T	0.000	0.009	3.537	0.060	1.000
2	rs2236783	136594158	T	0.103	0.154	4.314	0.038	1.000
2	rs2290518	135878814	T	0.179	0.188	0.097	0.756	1.000
2	rs2305248	135928312	G	0.191	0.181	0.106	0.745	1.000
2	rs2322254	135750849	T	0.210	0.234	0.595	0.441	1.000
2	rs309137	136765951	T	0.106	0.145	2.584	0.108	1.000
2	rs309176	136622216	C	0.112	0.149	1.313	0.252	1.000
2	rs309180	136614255	A	0.116	0.142	1.187	0.276	1.000
2	rs309181	136614813	G	0.119	0.147	1.287	0.257	1.000
2	rs3754686	136603276	G	0.103	0.151	3.867	0.049	1.000
2	rs4988235*	136608646	T	0.000	0.006	2.367	0.124	1.000
2	rs4954228	135976498	G	0.158	0.154	0.017	0.898	1.000
2	rs766271	135667131	C	0.054	0.048	0.105	0.745	1.000
2	rs867563	137142500	G	0.121	0.161	2.497	0.114	1.000
2	rs876338	137289147	A	0.192	0.259	4.772	0.029	1.000
2	rs882374	137913295	A	0.163	0.155	0.080	0.777	1.000
2	rs935612	135941503	C	0.157	0.154	0.010	0.921	1.000

Allelic frequencies that resulted statistically different even after the Bonferroni's correction are reported in bold type. (CHR=chromosome, BP=position of the SNP, MA=minor allele in YMN population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

4.4 Haplotype Reconstruction and Phylogenetic Analysis

4.4.1 Italian samples

Haplotypes within the largest identified LD block were statistically inferred for both the Italian and European subsamples (i.e. NCWI, NEI, CESI, SARD, CEU, FIN, GBR, and TSI as well) by considering 15 SNPs in high LD that form a long haplotype considering all the samples (Table 4.4.1.1).

Accordingly, 34 different haplotypes were observed, six of them carrying the T allele at the polymorphic site -13,910C/T.

The most frequent haplotype (H_1) in the whole dataset instead carried the -13,910*C allele and was the most represented haplotype also in the Italian subpopulations. On the contrary, in the Northern European groups H_2, carrying the -13,910*T allele, was the predominant haplotype.

The haplotype carrying the reference alleles for all the 15 high LD SNPs (H_19) was present only in three Italian samples, one for every of the three regions NCWI, NEI and CESI.

Table 4.4.1.1 Haplotypes reconstruction.

	Haplotypes	CEU	FIN	GBR	TSI	NCWI	NEI	CESI	SARD	TOT
H_1	CACTACCCA**C**GCCTC	0.182	0.312	0.213	0.556	0.414	0.442	0.528	0.521	0.407
H_2	AGTCGTTTG**T**AGTCT	0.688	0.554	0.730	0.097	0.276	0.230	0.097	0.053	0.323
H_3	AGTCGTTTG**C**AGCCT	0.059	0.054	0.011	0.128	0.133	0.097	0.142	0.202	0.102
H_4	CACCATCCA**C**GCCTC	0.012	0.022	0.022	0.077	0.057	0.040	0.082	0.043	0.048
H_5	AGTTACCCA**C**GCCTC	0.012	0.005	0.006	0.041	0.033	0.040	0.063	0.128	0.038
H_6	AGTCGTTTG**C**GCCTC	0.012	0.000	0.000	0.010	0.019	0.029	0.019	0.011	0.014
H_7	CACTACCCA**C**GCCTT	0.006	0.016	0.006	0.020	0.010	0.007	0.013	0.000	0.010
H_8	CACCACCCA**C**GCCTC	0.006	0.000	0.000	0.041	0.010	0.018	0.003	0.011	0.011
H_9	AGTCGTTTG**T**AGTCC	0.012	0.032	0.006	0.005	0.000	0.000	0.003	0.000	0.007
H_10	AGTCGTTTG**C**AGCCC	0.000	0.000	0.006	0.010	0.014	0.000	0.009	0.000	0.006
H_11	AGTCGTTTG**C**AGTCT	0.000	0.000	0.000	0.000	0.000	0.018	0.013	0.000	0.006
H_12	AGTCGTTTA**C**GCCTC	0.000	0.005	0.000	0.005	0.010	0.000	0.003	0.000	0.003
H_13	CGCCGCCCG**C**GCTTT	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.003
H_14	CGCCGCCCG**C**GCCTT	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.002
H_15	AATTATTTA**C**AGCCC	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.002
H_16	CACTACCCA**C**ACCTC	0.000	0.000	0.000	0.000	0.010	0.004	0.000	0.000	0.002
H_17	AATTATTTA**T**AGCCC	0.000	0.000	0.000	0.000	0.000	0.011	0.000	0.000	0.002
H_18	AACTACCCA**C**GCCTC	0.000	0.000	0.000	0.000	0.000	0.004	0.003	0.000	0.001
H_19	CACTATCCA**C**GCCTC	0.000	0.000	0.000	0.000	0.005	0.004	0.003	0.000	0.002
H_20	AGTCGTTTG**C**AGTCC	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.001
H_21	AATCGTTTG**T**AGTCT	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
H_22	CATCGTTTG**T**AGTCT	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
H_23	AGTTACCCA**C**GCCTT	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.001
H_24	AGTCACCCA**C**GCCTC	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.001
H_25	CATTACCCA**C**GCCTC	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.001
H_26	CGTCGTTTG**C**AGCCT	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.001
H_27	AGTCGTTTA**C**AGCCT	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.001
H_28	AGTCGTTTG**T**AGCCT	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.001
H_29	CACCATCCA**C**GCCTT	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.011	0.001
H_30	CGCTACCCA**C**GCCTC	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001
H_31	CACTACCTG**C**AGCCT	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001
H_32	CACTACCCG**C**AGCCT	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001
H_33	CACTACCCA**C**GCCCC	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001
H_34	CATCGTTTG**C**AGCCT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.021	0.001

-13,910 C/T is reported in bold type.

Evolutionary relationships among the inferred haplotypes were further investigated by means of a median joining network (Figure 4.4.1.1). Accordingly, the presence of two clearly distinct groups of allelic combinations was highlighted, with haplotypes carrying LP-related alleles being circumscribed to a single cluster. Interestingly, the NEI subgroup presented a lot of private haplotypes, two of them (H_17 and H_28) carrying the functional -13,910*T allele, the same situation is observed for the CEU subsample (H_21

and H_22). Overall, none of the 221 Northern-Italian individuals presented the second most frequent functional haplotype observed in Europe (H_9).

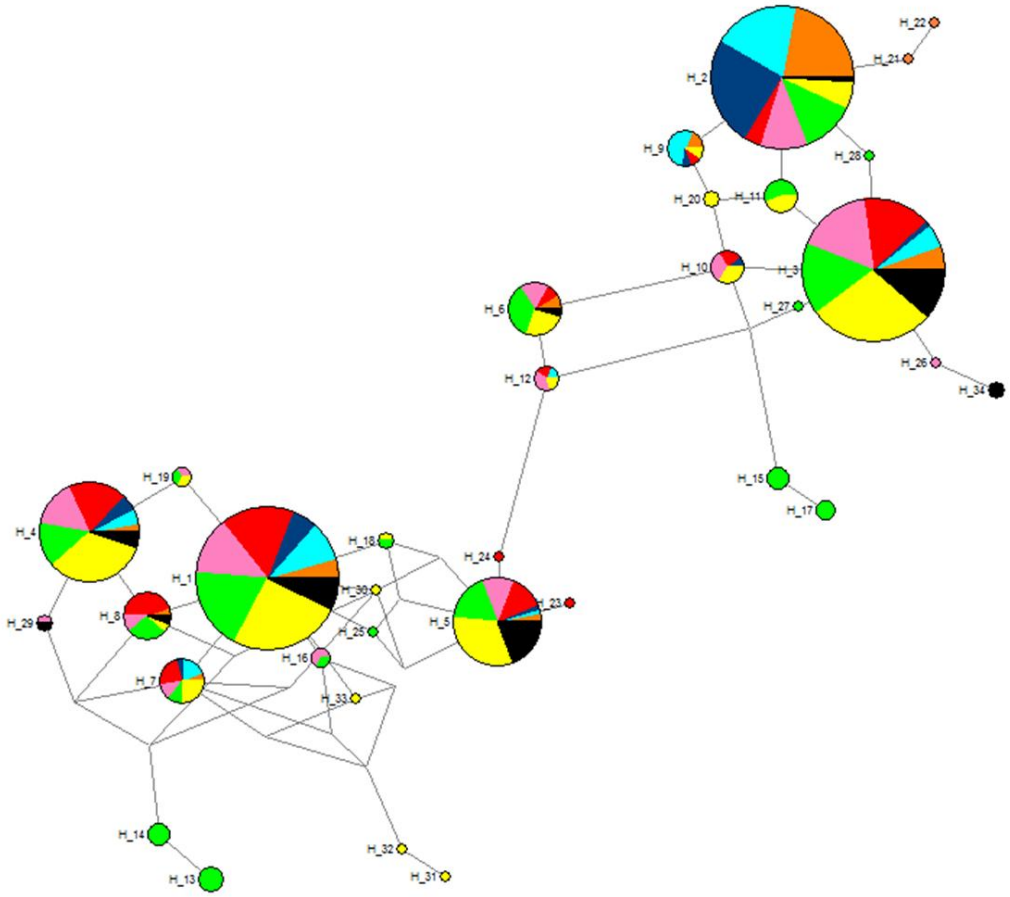


Figure 4.4.1.1 Median joining network of the inferred haplotypes.

NCWI, pink; NEI, green; CESI, yellow; SARD, black; TSI, red; CEU, orange; FIN, light blue; GBR, dark blue.

4.4.2 Samples from Arabian Peninsula

Haplotypes were statistically inferred in all the samples belonging to the Arabian Peninsula and by considering the 15 SNPs in the whole region of overall high LD that is basically shared among the examined groups.

A total of 35 different haplotypes were thus observed, 14 of which carried the functional allele at the -13,915 T/G locus and only one (H_5) showing the -13,910 T allele (Table 4.4.2.1) and being completely absent in DRF, as well as scarcely represented in the other groups, with the exception of OAO (0.220). In fact, the most frequent haplotype in the whole dataset (H_1) carried the -13,910 C allele and the -13,915 G one, being not represented only in the OAO group and reaching the highest frequencies in DFR and YMN (0.601 and 0.515, respectively). On the contrary, the second most common

haplotype (H_2), which did not carry adaptive alleles, was highly represented in ANO and OAO (0.528 and 0.415), showing considerably lower frequencies in YMN and DFR. A similar pattern was observed also for H_3, which reached an overall frequency of 0.093 in the whole dataset, whereas the remaining haplotypes appeared to be rare and, in some cases, private ones.

Table 4.4.2.1 Haplotypes reconstruction.

	Haplotypes	ANO	DFR	OAO	YMN	TOT
H_1	CCTACCCA**C**GGCCTC	0.145	0.601	0.000	0.515	0.3833
H_2	CCTACCCA**C**TGCCTC	0.528	0.212	0.415	0.227	0.3365
H_3	ATCGTTTG**C**TAGCCT	0.136	0.034	0.134	0.099	0.0929
H_4	CCCATCCA**C**TGCCTC	0.054	0.015	0.037	0.032	0.0341
H_5	ATCGTTTG**T**TAGTCT	0.023	0.000	0.220	0.006	0.0238
H_6	CCCACCCA**C**TGCCTC	0.019	0.010	0.049	0.026	0.0198
H_7	CCTACCCA**C**TGCCTT	0.007	0.022	0.037	0.017	0.0167
H_8	ATCGTTTG**C**GAGCCT	0.002	0.025	0.000	0.017	0.0135
H_9	ATCGTTTG**C**TAGCCC	0.019	0.005	0.012	0.015	0.0127
H_10	ATTACCCA**C**TGCCTC	0.016	0.000	0.012	0.006	0.0079
H_11	ATCGTTTG**C**TGCCTC	0.007	0.000	0.037	0.006	0.0063
H_12	CCTACCTG**C**TAGCCC	0.000	0.017	0.012	0.000	0.0063
H_13	CCCATCCA**C**GGCCTC	0.005	0.002	0.000	0.009	0.0048
H_14	CCTATCCA**C**TGCCTC	0.007	0.005	0.000	0.003	0.0048
H_15	CCTACCTG**C**TAGCCT	0.007	0.002	0.000	0.003	0.0040
H_16	CCCATCCA**C**TGCCTT	0.009	0.000	0.000	0.000	0.0032
H_17	CCTACCCA**C**GAGCCT	0.000	0.010	0.000	0.000	0.0032
H_18	ATCGTTTG**C**GAGCCC	0.000	0.007	0.000	0.000	0.0024
H_19	ATCGTTTG**C**TAGTCT	0.002	0.000	0.012	0.003	0.0024
H_20	ATTGTTTG**C**GAGCCC	0.000	0.007	0.000	0.000	0.0024
H_21	CCTACCCA**C**GGCCTT	0.000	0.007	0.000	0.000	0.0024
H_22	CCTACCCA**C**TGGCTC	0.000	0.005	0.000	0.003	0.0024
H_23	CCCACCCA**C**GGCCTC	0.002	0.000	0.000	0.003	0.0016
H_24	CCTACCTG**C**GAGCCC	0.000	0.005	0.000	0.000	0.0016
H_25	CCTATCCA**C**GGCCTC	0.000	0.000	0.000	0.006	0.0016
H_26	CCTATCCA**C**GGCCTT	0.000	0.005	0.000	0.000	0.0016
H_27	CCTGTTTG**C**TAGCCT	0.005	0.000	0.000	0.000	0.0016
H_28	ACCGTTTG**C**TAGCCT	0.002	0.000	0.000	0.000	0.0008
H_29	ATCGTTCA**C**TGCCTC	0.002	0.000	0.000	0.000	0.0008
H_30	ATCGTTTA**C**GGCCTC	0.000	0.000	0.000	0.003	0.0008
H_31	ATCGTTTA**C**TGCCTC	0.000	0.000	0.012	0.000	0.0008
H_32	CCCGTTTG**C**TGCCTC	0.000	0.000	0.000	0.003	0.0008
H_33	CCTACCCA**C**GACCTC	0.000	0.002	0.000	0.000	0.0008
H_34	CCTACCCA**C**TAGCCC	0.000	0.000	0.012	0.000	0.0008
H_35	CCTGTCCA**C**GGGCTC	0.002	0.000	0.000	0.000	0.0008

-13,910 C/T is reported in bold type and -13,915 T/G in red.

Evolutionary relationships among the inferred haplotypes were further investigated by means of a median joining network (Figure 4.4.2.1). In the upper part of the reconstructed topology, the ancestral haplotype (H_14), absent only in the OAO group, was observable, together with the most common H_1 and H_2 haplotypes and the great majority (71%) of allelic combinations carrying the adaptive -13,915*G allele, included five of the DFR private haplotypes (H_17, H_21, H_24, H_26, H_33). Interestingly, none of the 41 OAO individuals presented this allele.
The lower part of the reconstructed topology was instead characterized by only four haplotypes carrying the adaptive -13,915 G allele (29%), mainly distributed in DFR and YMN groups, and by the sole haplotype carrying the -13,910 T allele.

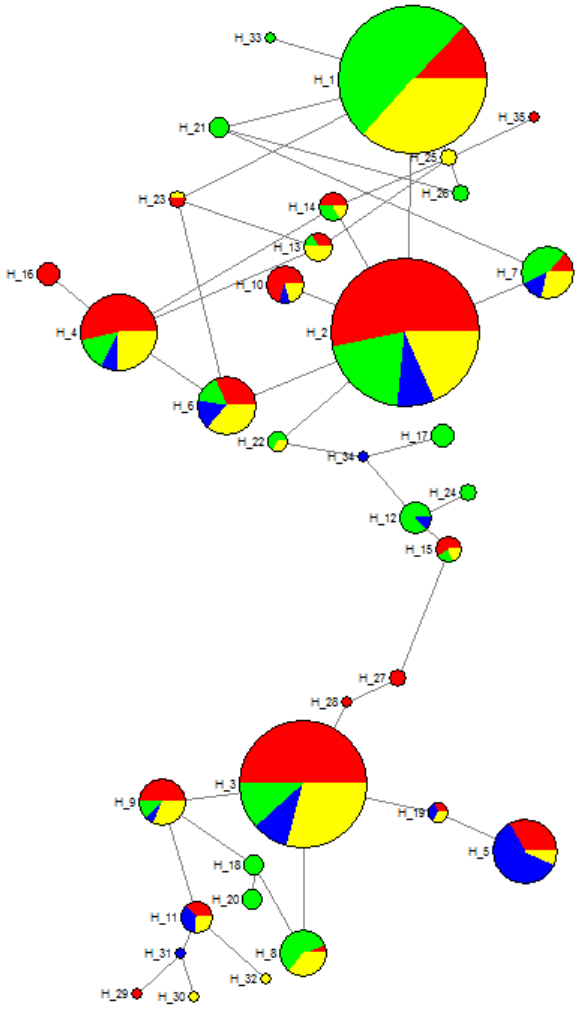


Figure 4.4.2.1 Median joining network of the inferred haplotypes.
ANO, red; DFR, green; OAO, blue; YMN, yellow.

4.5 Summary statistics

4.5.1 Italian samples

Nucleotide diversity (π) was calculated for each population sample (Table 4.5.1.1) and reached the highest value in the NCWI one, which showed a variability similar to that of NEI and FIN. The CEU and GBR samples showed the lowest level of diversity and presented similar values of π, as observed also for TSI, CESI and SARD samples. As regards the average heterozygosity across the examined loci, the highest value was found in FIN (OH = 0.443 ± 0.111), followed by NEI and CESI (OH = 0.416 ± 0.068 and OH = 0.416 ± 0.081, respectively), with NCWI showing instead an intermediate value (OH = 0.404 ± 0.056) with respect to these groups and the remaining Italian samples (TSI, OH = 0.389 ± 0.086 and SARD, OH = 0.369 ± 0.120). Again CEU and GBR presented the lowest values of the entire dataset. As pointed out for nucleotide diversity, haplotype variation appeared to be basically higher in Italian groups with respect to European ones, with the sole exception of the SARD sample, for which a remarkable reduction in the number of haplotypes and in the related variability was observed (k = 9, H = 0.390 ± 0.208). Again, the most outstanding levels of variation were found in subpopulations from Northern Italy (H = 0.739 ± 0.021, H = 0.732 ± 0.020 for NEI and NCWI, respectively), but the highest number of inferred haplotypes (k = 19) was observed in all groups from Eastern Italy (i.e. in both NEI and CESI). Among European populations, FIN showed the greatest variability (H = 0.595 ± 0.027), but, together with GBR, the lowest number of haplotypes (k = 8).

Table 4.5.1.1.Summary statistics for the examined populations.

Sample	N	π	OH	k	H
NCWI	210	0.436 ± 0.217	0.404 ± 0.056	14	0.732 ± 0.020
NEI	278	0.429 ± 0.212	0.416 ± 0.068	19	0.739 ± 0.021
CESI	318	0.394 ± 0.195	0.416 ± 0.081	19	0.682 ± 0.025
SARD	94	0.397 ± 0.198	0.369 ± 0.120	9	0.390 ± 0.208
TSI	196	0.398 ± 0.198	0.389 ± 0.086	13	0.659 ± 0.034
CEU	170	0.372 ± 0.185	0.364 ± 0.074	11	0.492 ± 0.041
FIN	186	0.421 ± 0.208	0.443 ± 0.111	8	0.595 ± 0.027
GBR	178	0.378 ± 0.188	0.352 ± 0.084	8	0.423 ± 0.038

N, n° of chromosomes; π, nucleotide diversity; OH, mean observed heterozygosity across loci; k, number of haplotypes; H, haplotype diversity

4.5.2 Samples from Arabian Peninsula

Nucleotide diversity (π) was calculated for each population sample (Table 4.5.2.1), pointing to the great variability of the OAO sample (π = 0.427 ± 0.213). DFR and YMN groups instead presented the lowest values of diversity (π = 0.236 ± 0.120 and π = 0.279 ± 0.141, respectively) and ANO an intermediate one (π = 0.346 ± 0.172).

A similar trend was observed for the average heterozygosity across the examined loci, with the highest value found in OAO (OH = 0.445± 0.115), followed by ANO (OH = 0.332 ± 0.109), DFR and YMN (OH = 0.235± 0.096 and OH = 0.270± 0.113, respectively).

Differently, haplotype variation appeared to be basically higher in the OAO sample (H=0.764±0.035), despite it showed the lowest number of inferred haplotypes (N=13), whereas ANO and YMN groups presented really similar value of diversity (H=0.679±0.021 and H=0.673±0.022, respectively). The lowest value was found in DFR (H=0.592±0.024), despite the presence in this group of a comparable number of haplotypes with respect to ANO and YMN.

Table 4.5.2.1 Summary statistics for the examined populations.

Sample	N	π	OH	k	H
ANO	428	0.346+/-0.172	0.332+/-0.109	21	0.679+/-0.021
DFR	406	0.236+/-0.120	0.235+/-0.096	20	0.592+/-0.024
OAO	82	0.427+/-0.213	0.445+/-0.115	13	0.764+/-0.035
YMN	344	0.279+/-0.141	0.270+/-0.113	20	0.673+/-0.022

N, n° of chromosomes; π, nucleotide diversity; OH, mean observed heterozygosity across loci;
k, number of haplotypes; H, haplotype diversity

4.6 Population structure analyses

4.6.1 Italian samples

PCA was firstly performed on individuals' genotypes. The first PC (PC1) accounted for 44.45% of the observed variation, whereas the second PC (PC2) accounted for 5.94% of it. The plot (Figure 4.6.1.1) illustrated a really peculiar distribution of samples along the PC1, grouping individuals in three clusters. The frequency distribution of the functional -13,910 variant in the three clusters was also characteristic, with all but one TT genotypes clustering in the left group and CC genotypes being restricted to the right one. The Italian samples clustered mainly in the central and in the right groups.

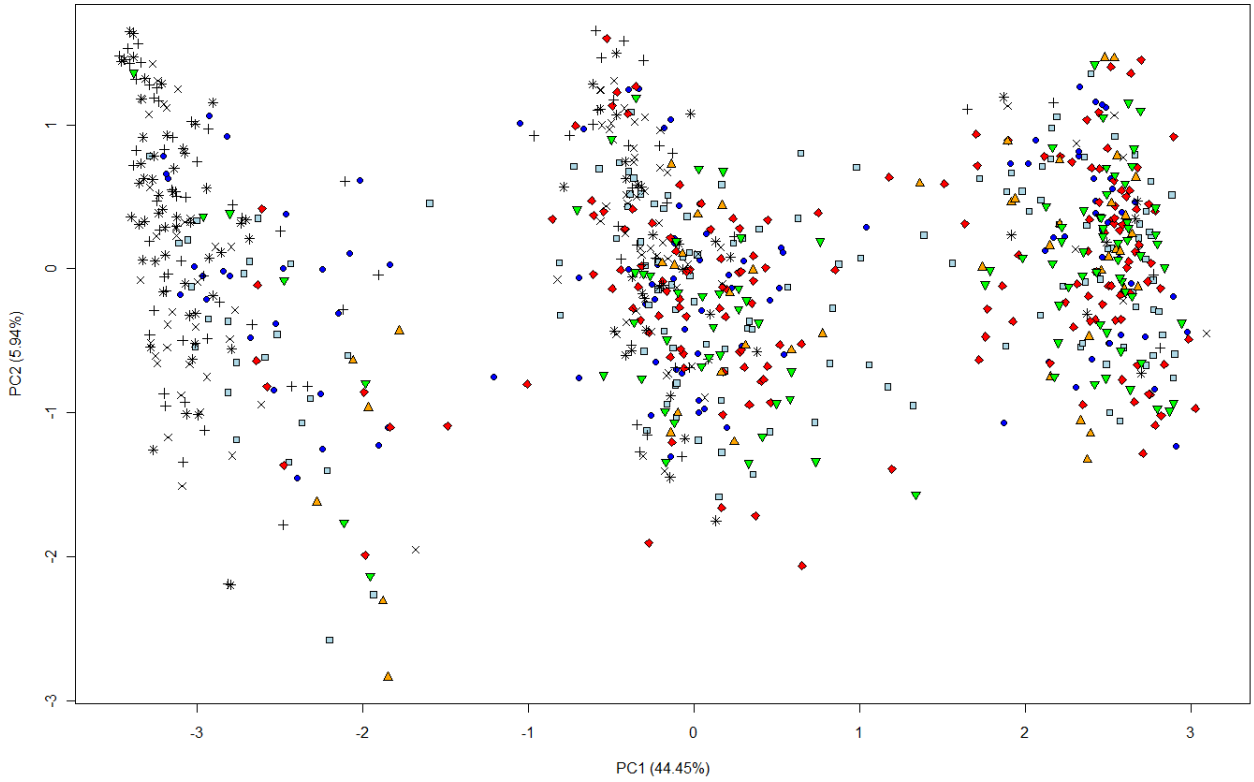


Figure 4.6.1.1: PCA based on the 39 typed SNPs.

NCWI, blue circles; NEI, light blue squares; CESI, red diamonds; SARD, orange point-up triangles; TSI, green point-down triangles; CEU, grey plus sign; FIN, grey cross; GBR, grey asterisk.

However, since LD existing among some of the examined polymorphic loci could have strongly influenced the described pattern, the obtained complete dataset was filtered to remove all the SNPs that presented r^2 values higher than 0.1. Accordingly, 13 SNPs out of the 39 genotyped and in linkage equilibrium which each other were retained to be submitted to several multivariate analyses aimed at investigating population structure.

PCA at the individual level was thus repeated showing how the percentage of variation described by PC1 decreased to 15.19%, whereas that related to PC2 increased to 10.65%. In particular, the revised plot pointed out the absence of remarkable differences among the examined populations (Figure 4.6.1.2). In fact, it was possible to observe only a weak subdivision between North European subjects (i.e. belonging to the CEU, FIN and GBR groups) and Italian ones (i.e. the samples collected in present study and the TSI).

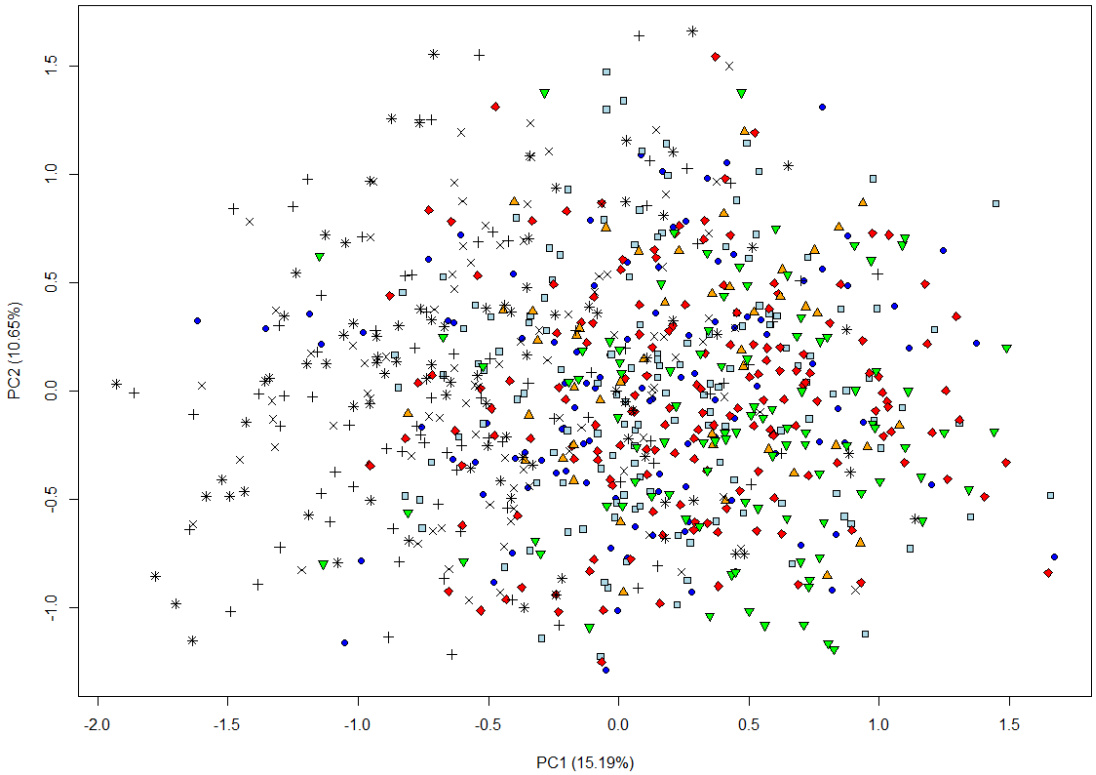


Figure 4.6.1.2: PCA based on the 13 SNPs in approximate linkage equilibrium.

NCWI, blue circles; NEI, light blue squares; CESI, red diamonds; SARD, orange point-up triangles; TSI, green point-down triangles; CEU, grey plus sign; FIN, grey cross; GBR, grey asterisk.

PCA was also conducted by considering the eight examined populations as *a priori* determined groups, showing a clear subdivision along the second PC, which accounted for 17.33% of observed variation, between North European and Italian samples. Moreover, a further separation of Italian subsamples in three clusters emerged along the first PC, accounting for 76.89% of variation, with CESI and NEI groups appearing to be more closely related with respect to NCWI and TSI, which have instead greater affinities with Northern European populations, and Sardinians standing out as potential outliers with respect to the two previously described clusters (Figure 4.6.1.3).

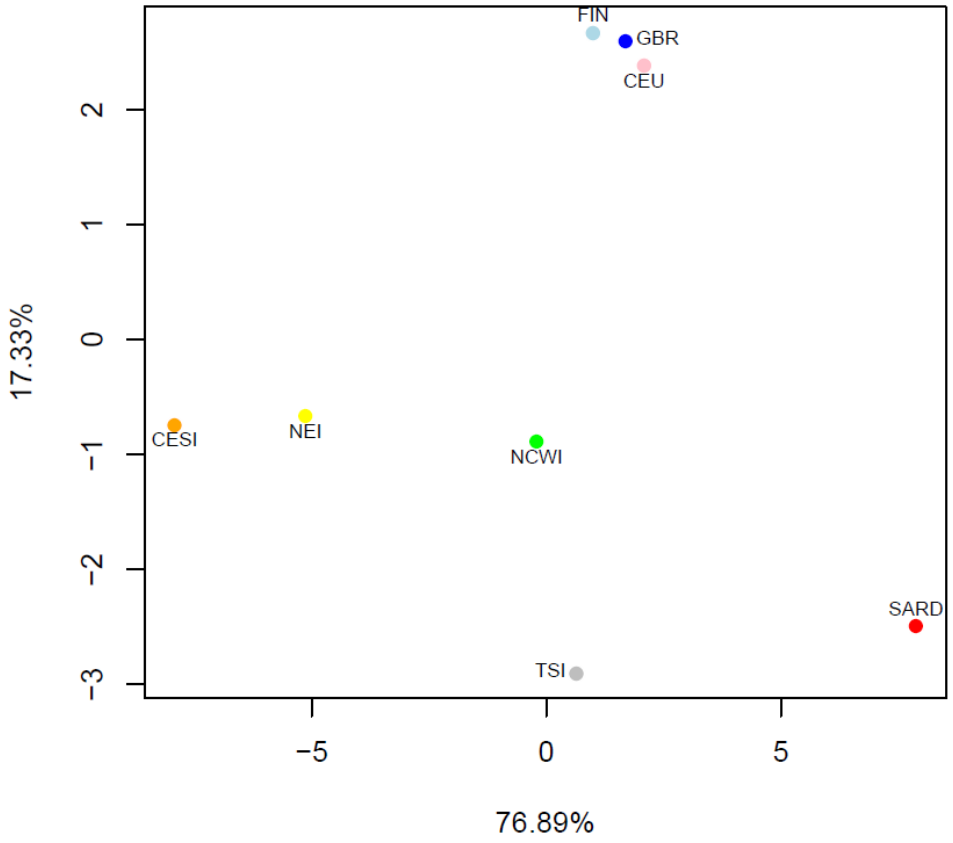


Figure 4.6.1.3: Plot of the PC1 (76.89%) against the PC2 (17.33%) considering the eight examined populations.

The exclusion of the outlier SARD sample form PCA did not substantially change the obtained results. In fact, North European populations (i.e. FIN, GBR, CEU) clustered together, whereas the Italian ones showed the same previously observed subdivision into two different groups along the first PC, which accounted for 71.05% of variation (Figure 4.6.1.4).

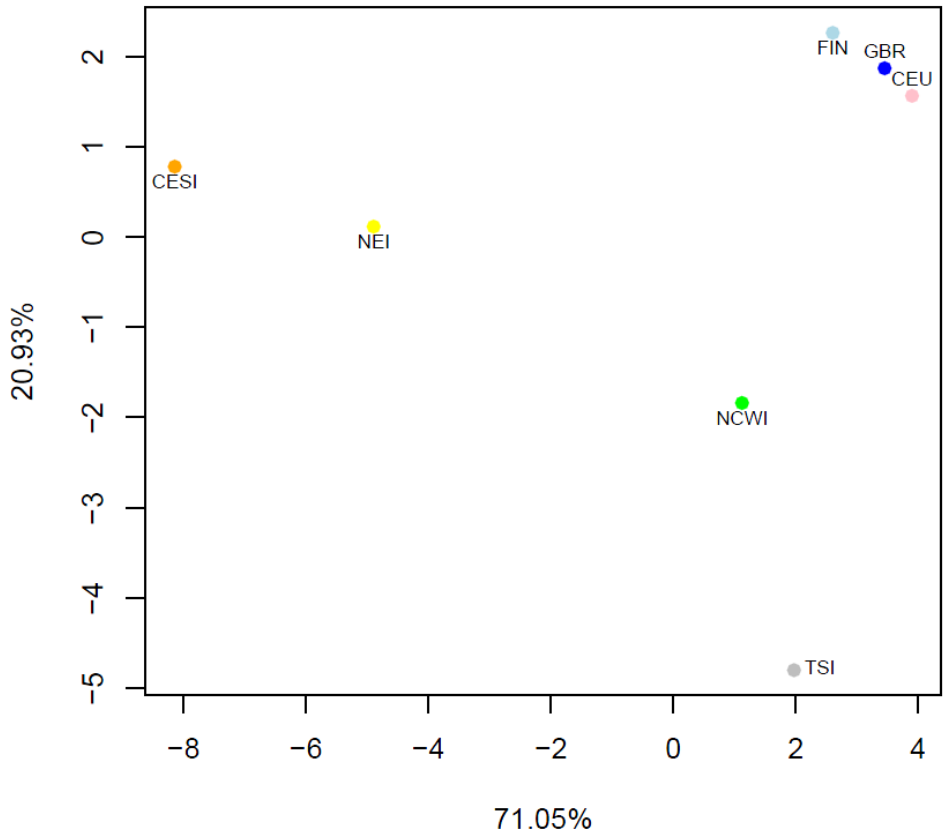


Figure 4.6.1.4: Plot of the PC1 (76.89%) against the PC2 (17.33%) in seven examined populations (SARD sample was removed).

To provide further support to the identified population structure, evaluation of cluster membership probabilities for each subject was achieved by means of DAPC conducted by specifying eight *a priori* know groups (i.e. the eight examined populations). Again, overall scarce differentiation was found between the groups, with only a weak separation observable between Northern Europeans and Italians, as already pointed out by classical PCA (Figure 4.6.1.5).

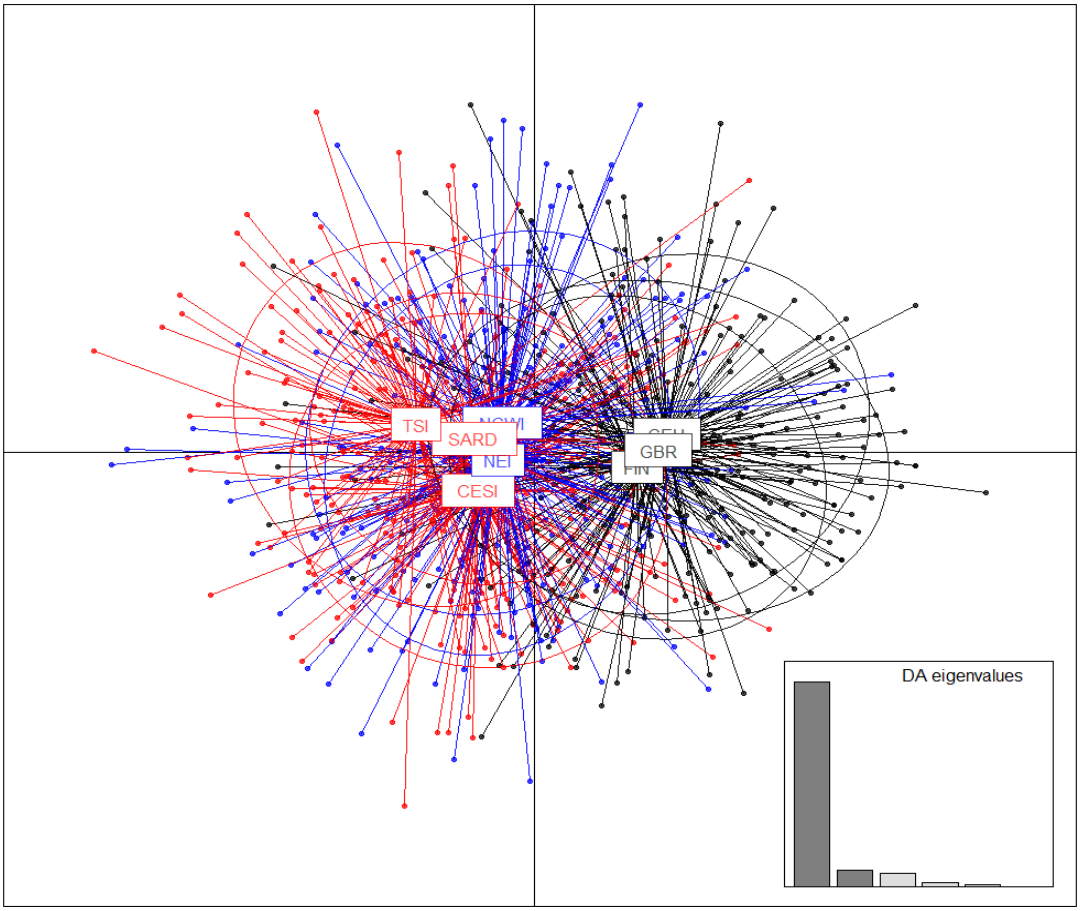


Figure 4.6.1.5: First and second principal components of DAPC.

Populations are indicated by ellipses which model 95% of the corresponding variability, as well as by different colors according to the observed structure (Central-Southern Italians in red, Northern Italians in blue, Europeans in black).

4.6.2 Samples from Arabian Peninsula

PCA was firstly performed on individuals' genotypes. The first PC (PC1) accounted for 24.93% of the observed variation and the second PC (PC2) for 13.99% of it. The plot (Figure 4.6.2.1) illustrated a really peculiar distribution along the PC1, grouping individuals in three clusters. The frequency distribution of the -13,910 C/T functional variant in the three clusters was also characteristic, with all TT genotypes clustering in the right group and the left one being composed only by CC genotypes. No clear distribution of genotypes for the -13,915 T/G variants were instead observed the three identified cluster.

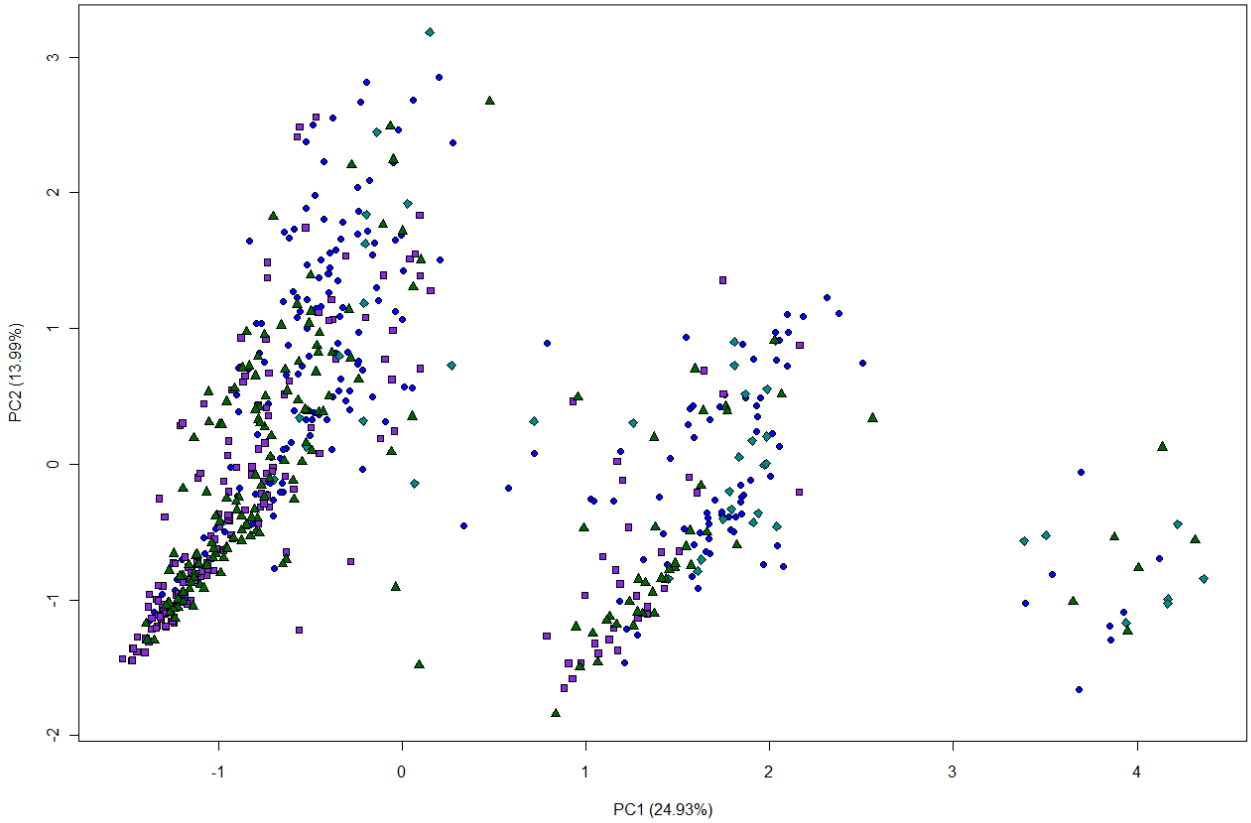


Figure 4.6.2.1: PCA based on the 40 SNPs typed in Arabian samples.
OAO, blue circles; ANO, violet squares; DFR, cyan diamonds; YMN, dark green point-up triangles.

However, since LD existing among some of the examined polymorphic loci could have strongly influenced the described pattern, the complete dataset was filtered to remove all the SNPs that presented a LD higher than $r^2 = 0.1$. Accordingly, 15 SNPs in linkage equilibrium with each other were retrieved and used to examine population structure with PCA and DAPC analyses.

PCA analysis repeated on the pruned dataset allowed to appreciably distinguish along PC1 YMN and DFR samples with respect to the bulk of those belonging to ANO and OAO. Moreover, the majority of YMN and DFR subjects differentiated to each other along PC2, while both OAO and ANO appeared to be more scattered along this PC, with the former sample showing the highest level of internal variability (Figure 4.6.2.2).

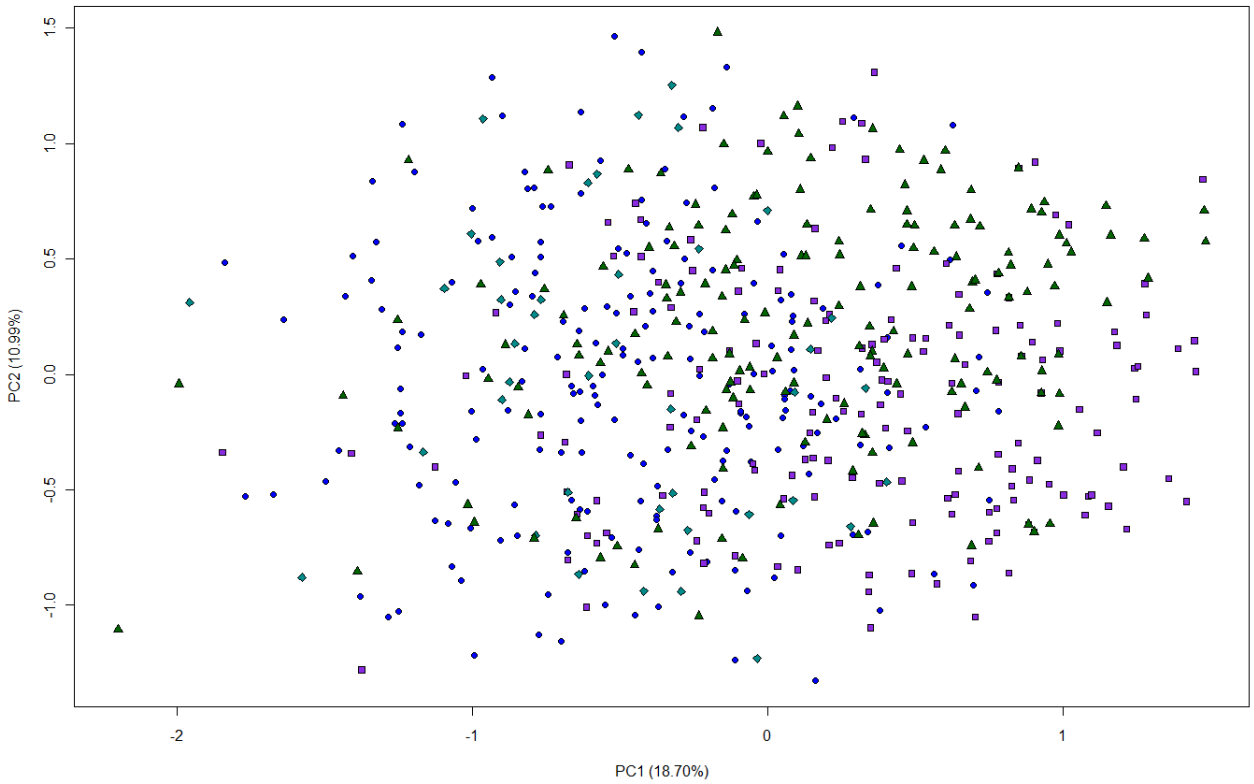


Figure 4.6.2.2: PCA based on the 15 SNPs in approximate linkage equilibrium.
OAO, blue circles; ANO, violet squares; DFR, cyan diamonds; YMN, dark green point-up triangles.

When conducted by considering the four examined populations as *a priori* determined groups, PCA showed a clear subdivision of populations along both the first PC, which accounted for 79.55% of the observed variation, and the second PC, accounting for 17.67% of variation, with a particularly outstanding position occupied by OAO (Figure 4.6.2.3).

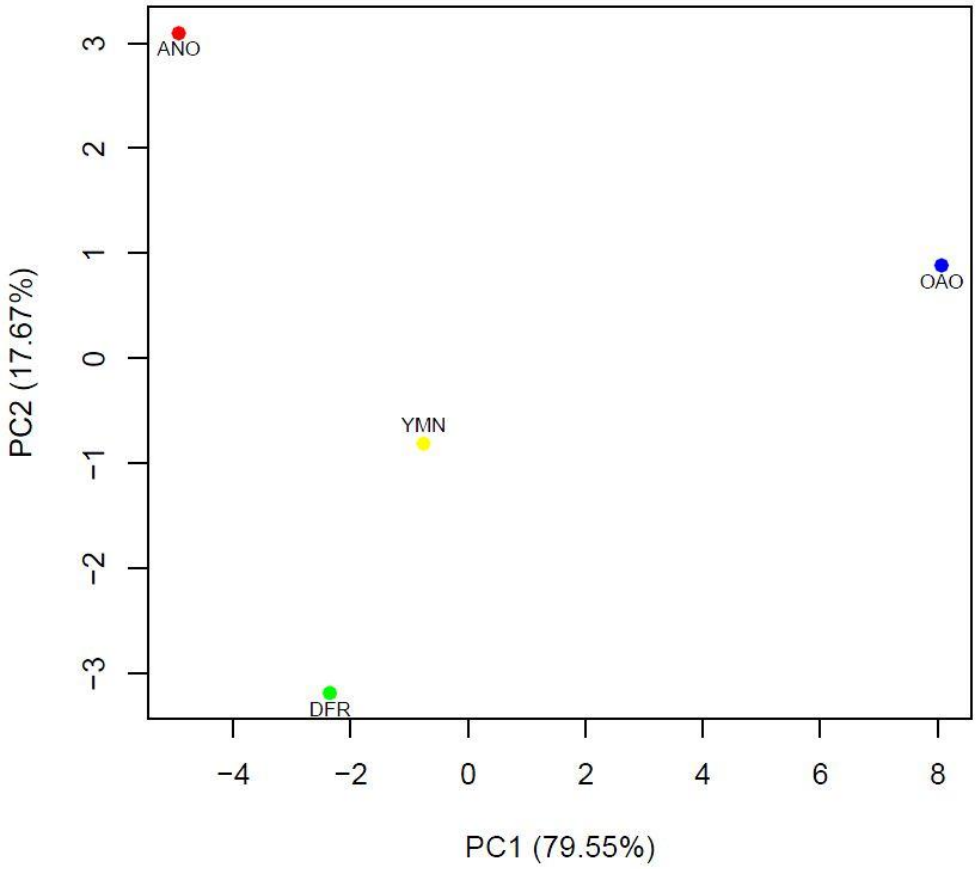


Figure 4.6.2.3: Plot of the PC1 (79.55%) against the PC2 (17.67%) considering the examined populations.

To more accurately describe patterns of population structure, evaluation of cluster membership probabilities for each subject was achieved by means of DAPC conducted by specifying four *a priori* know populations.

Accordingly, the DFR group appeared to be the most differentiated from the remaining ones, showing also greater affinity to YMN, whereas individuals belonging to OAO and ANO turned out to be considerably overlapped (Figure 4.6.2.4).

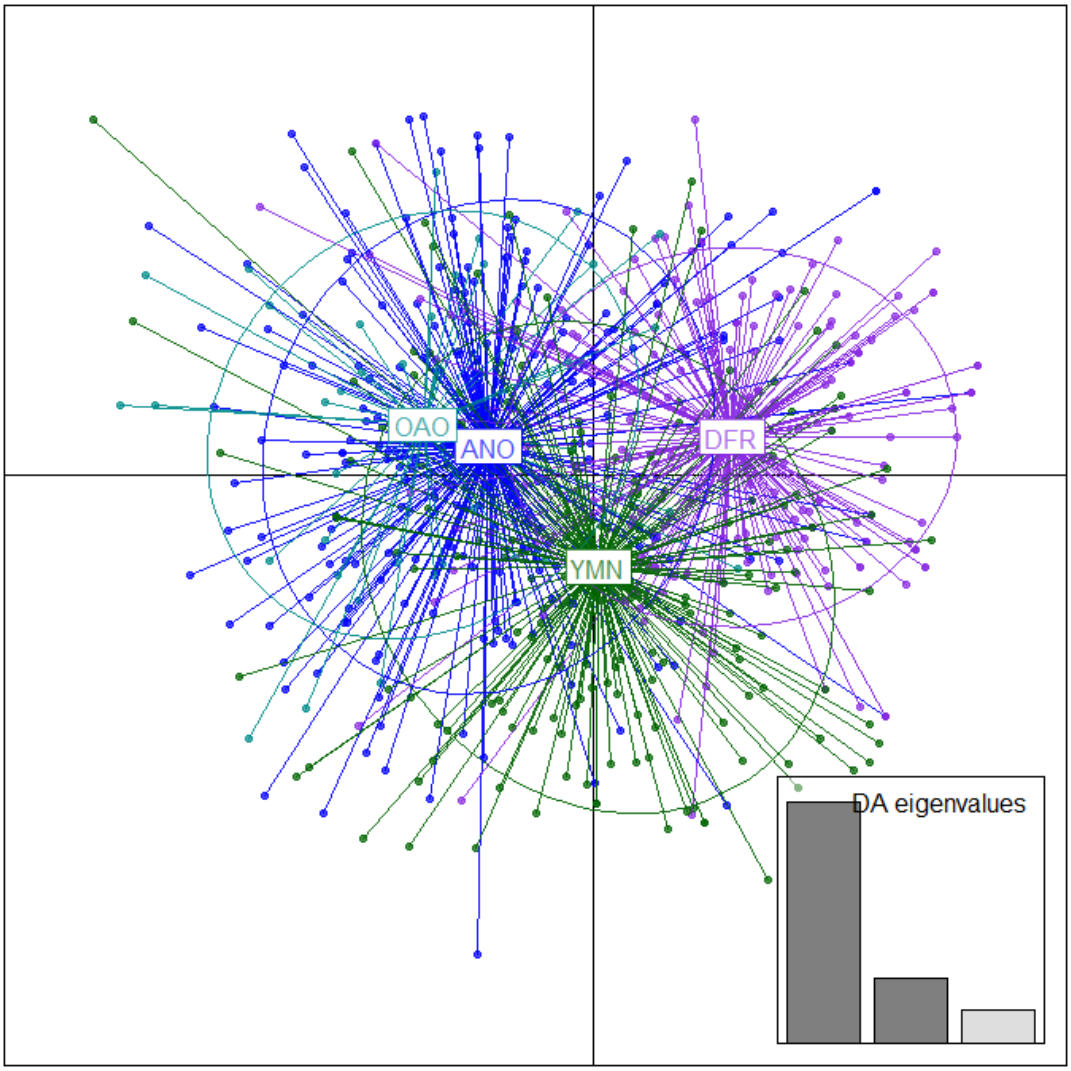


Figure 4.6.2.4: First and second principal components of DAPC.

Populations are indicated by ellipses which model 95% of the corresponding variability, as well as by different colors according to the observed structure.

4.7 Analysis of the Molecular Variance

4.7.1 Italian samples

Patterns pointed out by population structure analyses were statistically tested by AMOVA results (Table 4.7.1.1), which showed overall significant, but relatively low, differentiation among the examined populations (F_{ST} = 0.094, p < 0.001). Interestingly, when samples were subdivided according to macro-geographical criteria, a low but significant variance was observed among populations belonging to the same cluster (F_{SC} = 0.012, p < 0.001), but not among different clusters. On the contrary, when subpopulations form Northern and Central-Southern Italy were considered as independent samples, a significant among-groups component of

variance appeared (F_{CT} = 0.114, p < 0.01), pointing to the presence of an appreciable structure along the Italian peninsula.

Table 4.7.1.1 Analysis of molecular variance (AMOVA) according to different levels of population clustering.

Classifications	*Percentages of genetic variance*			*Fixation indices*		
	Among pop/groups	*Among pop within groups*	*Within pop*	F_{CT}	F_{SC}	F_{ST}
No grouping	9.41		90.59			0.094***
Macro-geographical groups [1]	15.03	1.04	83.93	0.150	0.012***	0.161***
Micro-geographical groups [2]	11.39	0.39	88.21	0.114**	0.004	0.118***

[1] Europe, Italy
[2] Europe, Northern Italy, Central-Southern Italy
** p<0.01, *** p<0.001.

4.7.2 Samples from Arabian Peninsula

Patterns pointed out by population structure analyses were statistically tested by AMOVA computations. Accordingly, an overall low but significant differentiation was observed among the examined populations (F_{ST} = 0.052, p < 0.001). When samples were subdivided according to geographical criteria, a low but significant variance was observed among populations belonging to the same cluster (i.e. among Oman and Yemen or Northern and Southern populations) (F_{SC} = 0.074, F_{SC}=0.037, p < 0.001), but not among the supposed clusters. Moreover, when structure observed by previous analyses was considered (i.e. ANO, OAO and DFR, YMN) a significant variance was observed among populations belonging to the same cluster (F_{SC} = 0.062, p < 0.001).

Table 4.7.2.1 Analysis of molecular variance (AMOVA) according to different levels of population clustering.

Classifications	*Percentages of genetic variance*			*Fixation indices*		
	Among pop/groups	*Among pop within groups*	*Within pop*	F_{CT}	F_{SC}	F_{ST}
No grouping	5.16		94.84			0.052***
DAPC results [1]	-1.14	6.24	94.90	-0.011	0.062***	0.051***
Oman/Yemen [2]	-4.38	7.74	96.64	-0.044	0.074***	0.034***
North/South [2]	2.22	3.60	94.18	0.022	0.037***	0.058***

[1] ANO and OAO, DFR and YMN
[2] Geographical groups
*** p<0.001.

4.8 Correlation analyses

Potential concordance between the distribution of haplotypes carrying the tolerance or the intolerance allelic status (LP/NLP haplotypes) at the -13,910 C/T locus and the frequency of Y chromosome and mitochondrial DNA haplogroups was investigated by means of a correlation analysis.

Data for the mentioned uniparentally inherited genetic markers were retrieved from a recent study carried out by our research group (Boattini et al. 2013), which analyzed a panel of Italian samples belonging to the same geographic macro-areas examined in the present study, and of which individuals genotyped for LP-related SNPs represented a selected subsample.

Interestingly, Boattini et al. have also performed an estimate of haplogroups' ages, so that potential correlation between the considered LP-related and neutral datasets allowed to infer an age estimate also for the entrance of the -13,910 C/T mutation in the Italian peninsula.

The analysis carried out by considering LP and NLP haplotypes including the 15 SNPs in high LD with each other showed a positive correlation between the most frequent LP haplotype (H_2) in the complete dataset carrying the functional -13,910 C/T site and some Y-chromosome haplogroups. In particular, occurrence of this haplotype was positively correlated with that of haplogroup R1b-L2, showing an estimated time of entrance in the Italian peninsula of ~ 3,200 years ago (ya), as well as with that of G2a-L497 (time of entrance ~ 3,600 ya), I1-M253* (no dating reported) and I2-P215* (no dating reported). The same LP haplotype showed a negative correlation with Y-chromosome G2a-M406 and M527. Moreover, an important observation that make sense at our speculations, was that the most frequent NLP haplotype (carrying the -13,910*C allele) showed a negative correlation with the above mentioned haplogroups (Appendix Table 4 and 5).

LP haplotypes do not show correlation with mtDNA haplogroups, all the haplogroups reported in Boattini et al. (Boattini et al. 2013) show an age estimation higher than 10,000 yrs, confirming a most recent entrance of the mutation in Italy (Appendix Table 6 and 7).

5. Discussion and Concluding Remarks

Increasing genomic evidences suggest that cultural practices have long and heavily shaped the human genome variation, so that the evolution of our species has been actually driven by genome-culture co-evolutionary processes, which are generally faster and operate over a broader range of conditions with respect to strictly biological ones (Laland et al. 2010).

Accordingly, taking into account a gene-culture co-evolutionary perspective in the study of recent human evolutionary history can help researchers to understand the processes by which culture has been shaped by biological imperatives, as well as those in which biological properties were altered by genetic evolution in response to cultural history. For instance, hundreds of genes involved in particular human traits, such as diet and resistance to infectious diseases or language and learning, have been supposed to have undergone natural selection in response to modifications in cultural practices. Cultural determinants have been thus invoked to explain many signatures of selection identified in nutrition-related genes. In particular, dietary shifts due to the introduction of agricultural practices have been associated to adaptive events on carbohydrates metabolism genes (e.g. selection on the *AMY1* gene) (Groot et al. 1989; Perry et al. 2007), while response to nutritional and thermal stresses encountered during out-of-Africa migrations were proposed to be responsible for selection on genetic variants enabling an increased energetic efficiency in nomadic lifestyles (e.g. thrifty phenotype hypothesis, Neel 1999).

In particular, the interdependence of the biological evolution of adult lactose tolerance and the cultural evolution of dairy farming is probably the most extensively studied example of gene-culture coevolution (Enattah et al. 2002; Beja-Pereira et al. 2003; Swallow et al. 2003; Itan et al. 2009).

In fact, the main carbohydrate in milk is lactose, which must be hydrolysed to glucose and galactose before it can be digested. While 65%, or more, of the global human population is lactose intolerant, in some human groups lactase activity commonly persists into adulthood. For instance, lactose tolerance is exceptionally widespread in Northern European countries, such as Sweden and Finland, with tolerance levels of 74% and 82%, respectively (Vuorisalo et al. 2012). Moreover, in the last decade, scientists discovered a polymorphism completely associated with

LP in individuals of Finnish origin (i.e. the -13,910C/T SNP) and the presence of a common large haplotype block, spanning around 1 Mb in individuals of Northern European origins, thus confirming the strong positive selection that have acted on this genomic region in the last few thousand years of evolution of certain human groups (Enattah et al. 2002; Bersaglieri et al. 2004).

Moving from these evidences, the present study aims at representing the first exploration of genetic variation patterns potentially related to the LP phenotype in populations from two geographical areas (i.e. the Italian and the Arabian peninsulas) till now scarcely investigated for this adaptive trait. In fact, the study of these populations promises to be highly informative about regions which have long played key roles as natural corridors for human migrations across the Mediterranean basin (i.e. Italy) and from Africa towards the Middle East (i.e. the Arabian peninsula). This will potentially help to disentangle some of the routes followed by LP diffusion from its area of origin to modern Europe, as well as to further explore processes of convergent evolution at LP occurred in some non-European populations.

For this purpose, 51 SNPs were selected among those analysed in Bersaglieri et al. (2004) and Enattah et al. (2007) and according to their high heterozigosity values, to be typed on a wide and very well-characterized sample of Italian and Arabian subjects.

After the design of PCR and extension primers for the multiplex PCR assay, only 40 out of 51 SNPs were genotyped in the two collected samples made up of 453 Italian healthy subjects and 635 healthy individuals from Oman and Yemen.

As regards the Italian sample, it was recruited in order to obtain a geographic distribution of individuals that could be informative of the demographic processes occurred across different Italian subpopulations and recently hypothesized by archaeological and genetic data. In fact, evidences from both research fields have suggested that the diffusion of Neolithic along the Italian peninsula was the result of two independent and parallel events involving the Adriatic and Tyrrhenian coastlines (Pessina 2008; Boattini et al. 2013).

In parallel, Arabian samples were also selected according to the complex and not completely explained population history of this region, which has represented one of the main corridor for the exit of modern humans from Africa, but also a plausible

crossroads for human migration in more recent times. For instance, different and conflicting scenarios were proposed for the recent peopling of the Arabian Peninsula during the Early Holocene, between 10,000 and 8,000 years ago. In fact, a first hypothesis postulates a "Levantine Expansion" as the main non historical demographic event that has shaped the current Arabian population picture, with substantial migration of Pre-Pottery pastoralists from the Levant (Uerpmann et al. 2009). On the contrary, a second thesis proposes a "refugia" model that argues for the spread of indigenous groups within Arabia itself (Fedele 2009; Rose 2010).

To draw a first picture of the potentially LP-related different histories to which the examined subpopulations undergone within both Italian and Arabian peninsulas, the analysis of LD patterns at the investigated genomic region was carried out for each of them.

As regards the four Italian subsamples, a long undisrupted haplotype block, including the functional variant -13,910C/T, was observed only in NCWI and NEI subjects, with overall higher LD values in the former. On the contrary, this region of strong LD turned out to be subdivided into two different blocks in both the CESI and SARD samples.

A similar pattern was observed also within the Arabian Peninsula, with a progressive reduction of a long LD block from OAO to ANO and YMN and its breaking down into three small LD blocks in the DFR population.

The presence of a long haplotype block into Northern Italians and in OAO and ANO Omani groups can be interpreted as a recent origin of the underlying combination of allelic states or, more plausibly, as a footprint of the action of a strong selective pressure on the genomic region of interest and the consequent unchanged conservation of the linkage block through generations.

Interestingly, comparison of allele frequencies of typed SNPs pointed out statistically significant differences (represented in bold in Table 4.3.1.2) between CESI and NEI, as concerns the two functional variants -13,910 and -22,018, as well as between CESI and NCWI, with three additional SNPs (rs745500, rs309180 and rs2236783) in high LD with them functional variant presenting different frequencies after Bonferroni correction (Table 4.3.1.1). Nevertheless, in both cases the highest differences were observed for the functional variants. These differences of allelic frequencies distribution for the persistence-associated -13,910*T allele are

in accordance with the North-South decreasing pattern observed by Anagnostou et al. (2009). This observation is also in accordance with results from LD analyses and suggests that the high occurrence of this variant is strongly preserved by positive selection in Northern Italy only, but not in Central-Southern Italy or Sardinia. In fact, selection on the -13,910*T allele has led to the concomitant rise in frequency of other polymorphisms near to the functional SNP and the consequent creation of a conserved haplotype block, as observed in Bersaglieri et al. (2004), although at a lower scale. The comparison of allelic frequencies of the same SNPs between the two Northern Italian and the Sardinian samples also showed similar results ,but highlighting the almost complete absence of the lactose tolerance-associated alleles in Sardinia (Table 4.3.1.3 and 4.3.1.4).

A more complicated scenario was instead observed when allele frequencies were compared among the four groups belonging to the Arabian Peninsula, with a larger variability observed between groups as compared with what described for the Italian Peninsula. In particular, the lower number of differences was observed between DFR and YMN, with a significant value obtained for the -13,915 variant, the functional allele of which showed an higher frequency in DFR (frequency of G allele 0.710). Moreover, significant p-values were observed for all the three comparisons involving this group, suggesting that a genetic background more similar to the Arabian one could be invoked for DFR, and in general for the southern part of the Peninsula, with respect to ANO and OAO populations. In the OAO sample we can indeed observe the complete absence of the -13,915*G allele and an outstanding frequency of the -13,910 SNP, which turned out to be higher than in the other three Arabian groups and similar to that observed in Northern Italy. Accordingly, an appreciable genetic affinity between OAO and European populations could be depicted for the lactase persistence trait.

To further confirm these findings and in the attempt to reconstruct the evolutionary history of the best candidate allelic combinations responsible for adaptation to lactose consumption in the examined populations, patterns of haplotype variation were also investigated.

The haplotypes reconstruction carried out on the 15 SNPs included in the larger observed LD block allowed us to identify 34 different haplotypes in our Italian samples, as well as in the European *1,000 Genomes Project* ones used as a reference

dataset. The most frequent haplotype considering the whole sample (H_1) carried the -13,910*C allele and represented the most frequent haplotype in all the Italian subgroups (i.e. TSI, NCWI, NEI, CESI and SARD). The second most frequent haplotype observed in the complete sample (H_2), carrying the -13,910*T allele, instead showed the highest frequencies in Northern European populations (i.e. CEU, FIN and GBR). These results are clearly illustrated by the Network analysis, by which the higher frequencies for haplotypes carrying the functional T allele at the -13,910 locus are highlighted in groups from Northern Europe. Interestingly, the NEI subgroup presented a lot of private haplotypes carrying the T or the C allele at the functional variant site. Although their frequency is very low, this points to an considerable variability within the NEI sample, with respect to the NCWI one.

A similar picture characterized by a lot of private haplotypes is observable also for CESI, but no one of these haplotypes presented the functional T allele at the -13,910 site. The second most frequent haplotype carrying the persistence-associated allele (H_9) showed a really low frequency, but being present in all the samples excluded NEI, NCWI and SARD. Finally, even if with a low presence in the considered samples, the haplotype carrying the reference allele at all the 15 SNPs (H_19) is found only in our Italian samples (NCWI, NEI and CESI).

The high variability observed in the Italian samples and the presence of the ancestral haplotype only in Italy suggesting that once arrived in the Italian peninsula LP have evolved differently and independently with respect to what occurred in northern European populations. It is possible to hypothesize an ancient entrance of the mutation in Italy than in North Europe and due to the geographical isolation for the presence of the Alps an independent evolution in the peninsula.

As regards samples from the Arabian Peninsula, we reconstructed 35 haplotypes considering again the 15 SNPs spanning the longer LD block, but including also the -13,915T/G polymorphism and excluding rs3213889, for which genotyping failed. The most common haplotype in the complete dataset (H_1) presented the intolerance-associated -13,910 allele and the tolerance-associated -13,915 one. Its distribution appeared to be really peculiar, with elevated frequencies in DFR and YMN (i.e. in the Southern part of the Peninsula), lower presence in ANO (i.e. in Northern Oman) and complete absence in OAO. These observation confirms inferences made on the basis of allele frequencies comparison, suggesting that YMN

and DFR are the most “Arabs” groups among those explored, while ANO and OAO showed LP-related genetic background more similar to those observed in Northern Italian and European populations. The second most frequent haplotype (H_2) presented a similar frequency in the total sample, but a really different distribution with respect to H_1. This haplotype does not carry lactase persistence-associated alleles and is mostly found in the ANO sample, as well as in the OAO one, in which it accounts for nearly half of the collected chromosomes. The H_2 frequencies observed in DFR and YMN are, on the other hand, considerably lower than the H_1 ones. In the 35 observed haplotypes, only one carries the LP-related T allele at the -13,910 site, being absent or showing very low frequencies in all the samples with the exception of OAO, in which it reaches a frequency of 22%. Again, this observation confirms results from the allelic frequencies analyses, highlighting the notable OAO affinity with European populations. Network analyses finally pointed out an high variability in the DFR sample, due to a lot of private haplotypes that in the majority of the cases presented the lactose-associated G allele at the -13,915 site, suggesting that this region could be one of the candidate place of origin for this Arabian-specific adaptive variant.

Summary statistics of genetic diversity calculated for each the examined populations were overall in line with the above described picture of population differentiation. In fact, nucleotide diversity values highlighted an appreciable similarity between the Finnish *1,000 Genomes Project* population and Northern Italians. The highest value of nucleotide diversity was observed for the NCWI sample and presented actually a little difference with respect to those calculated for NEI and Finnish. As expected, diversity levels of samples from Central-Southern Italy, Sardinia and Tuscany are almost the same and this is true also for GBR and CEU values. A similar pattern was observed for the mean observed heterozygosity across loci (OH), although in this case the highest value was observable in the FIN population, with the same values obtained in CESI and NEI. In line with that haplotype variation is unquestionably larger in the Italian samples, with the highest number of haplotypes inferred in NEI and CESI and the highest haplotype diversity found in Northern Italy. These observations confirm what already remarked by Network analyses, with NCWI, NEI and CESI showing an consistent number of private haplotypes and the most substantial diversity.

As regards the Arabian Peninsula samples, OAO showed the highest values for nucleotide diversity, mean observed heterozygosity across loci and number of haplotypes, followed by ANO that, as already observed for previous analyses, showed intermediate values between OAO and populations from the Southern part of the Peninsula (i.e. YMN and DFR) which are characterized by the lower values of diversity. On the contrary, OAO presented the lowest number of reconstructed haplotypes (13), while ANO, DFR and YMN show the highest and almost identical numbers.

To explore overall pattern of genetic structure at the examined genomic region, multivariate analyses were applied to the generated datasets.

The PCA plot obtained from the analysis of individuals' genotypes showed, along the first dimension, a peculiar distribution of samples in three main clusters, that perfectly reflected the distribution of genotypes of the 13,910 C/T variant in the European samples. However, this distribution is probably affected by the high LD that characterizes a large fraction of the analysed SNPs. In fact, the left cluster included all but one the TT genotypes and is composed by a high percentage of Northern European individuals. The right cluster, on the other hand, was almost exclusively composed by Italian samples, as expected by its main composition of CC genotypes. The distribution of the NEI and NCWI subjects is the most scattered along the plot dimensions, according to the highest value of nucleotide diversity observed in such subsamples. This observation was also evident when we considered a pruned dataset containing only the SNPs in linkage equilibrium which each other. This dataset presented only 13 SNPs and, as expected, the presence of the three groups observed in the previous PCA were not confirmed. It was anyway possible to observe a separation between Northern European and Italian samples.

PCA plot obtained by considering the eight populations, although underlining along the second dimension a sharp detachment between Italian and Northern European samples, further validated these observations. Along the first dimension, CESI and NEI appeared to be considerably distinct from NCWI and TSI, which showed instead great affinities to the three Northern European populations. The Sardinian sample appeared to be an outlier in the obtained picture of genetic diversity, but relative positions of the examined samples in the PCA plot did not change when this outlier was removed from the analyses. Finally, an additional issue that underlined the

peculiarity of the Italian Peninsula with respect to Northern Europe was represented by results from DAPC. In fact, all the Italian samples clustered together according to this analysis, being appreciably separated from Northern Europeans with NCWI and NEI being in an intermediate position between the two clusters.
A considerable genetic structure was observed also within the Arabian Peninsula. When we considered the individuals' genotypes and all the 40 typed SNPs, probably due to the presence of two functional variants, it was not possible to define a clear pattern of persistence/not persistence genotypes along the computed PCs. As already done for the European dataset, we tried to explore the variability of the investigated genomic region after pruning SNPs in high LD. Again, it was not possible to define a clear clustering between the examined populations, although the higher internal variability of the OAO sample was appreciable. As expected from results of the other analyses, the distribution of the four population in the PCA plot pointed out the proximity between YMN and DFR on both the first and the second components. In particular, along the first component, the OAO population appeared to be an outlier and the ANO group was close to, but not clustering with, Southern Arabian populations. All these results are confirmed by DAPC analysis, which highlighted how OAO and ANO showed higher affinity with respect to YMN and DFR.
Analyses of molecular variance (AMOVA) were conducted considering different levels of population clustering in order to provide statistical support to the observed population structure. In particular, for the European samples we considered three patterns: a single population group, macro and micro geographical groups. In all the three clustering, Fst values were significant, underling a high variance within populations. On the other hand, when we considered macro geographical groups (i.e. Italy and Northern Europe as two separated groups), we obtained also a significant Fsc value, suggesting that considerable differences could be observed also among populations belonging to the same cluster and thus plausibly between Northern and Central-Southern Italians. In fact, this hypothesis was confirmed by the significant value of the Fct index found by considering Europe, Northern Italy and Central-Southern Italy as three different groups. In the same way the Arabian Peninsula samples were analysed considering four subdivisions: a single population group, groups derived by DAPC analyses and two geographical clustering (i.e. Oman and Yemen or Northern and Southern

populations). In all the four clustering p-values associated to Fst presented significant results, thus demonstrating the presence of high variability within the four explored populations. All the three considered subdivisions showed significant values of Fsc indices, but not of Fct ones. These results underlined a higher variation among the populations within groups than among the population groups.

Potential concordance between the distribution of haplotypes carrying the tolerance- or the intolerance-associated allelic status (LP/NLP haplotypes) at the -13,910C/T locus and that of Y chromosome and mitochondrial DNA haplogroups was investigated in the Italian samples by means of a Spearman's rank correlation analysis. For this purpose, data regarding uniparentally inherited genetic markers were retrieved from a recent study carried out by our research group (Boattini et al. 2013), which analyzed a panel of Italian samples belonging to the same geographic macro-areas examined in the present study, and of which individuals genotyped for LP-related SNPs represented a subsample. Boattini et al. have also calculated an estimate of haplogroups' ages, so that it was possible also to infer an age estimate for the entrance of the -13,910 C/T mutation in the Italian peninsula.

A positive correlation was thus observed between the most frequent LP haplotype (H_2) and some Y-chromosome haplogroups. In particular, H_2 showed a positive correlation with haplogroup R1b-L2, which reached higher frequencies in the Northern part of the peninsula. The age estimated for R1b-L2 was about 3,200 years (Boattini et al. 2013) and this allows to hypothesize an entrance of the functional mutation in Italy from Central Europe together with the R1b-L2 lineage, after the early arrival of the Neolithic, thus suggesting that not only the mutation but also the adaptive haplotype was not originated in Italy and was already shaped by natural selection before its arrival in the Italian peninsula. These hypothesis could be confirmed also by the negative but significant correlation observed between one NLP haplotypes (H_1) with this Y-chromosome linage.

A comparable scenario was observed also for haplotypes that Boattini et al. (2013) identified as belonging to a specific cluster of sequences of haplogroup G2a. Again these Y-chromosome haplotypes presented higher frequencies in North-Western Italy and a similar estimated age of about 3,600 years, together with a comparable Central European origin (Boattini et al. 2013). Accordingly the G2a-L497 showed a positive correlation with the most frequent LP haplotype (H_2). An opposite picture

was observed for the other clusters of the G2a (M406 and M527) haplogroup, with a significant negative correlation between LP haplotype (H_2). All these considerations allowed to paint a possible route for the entrance and diffusion of LP in Italy, confirming the occurrence of two independent events of Neolithisation in the Peninsula (Pessina et al. 2008). In fact, it is possible to correlate the NLP haplotype with the first event involving the Adriatic and the Tyrrhenian coasts and dating as early as 8,100 YBP (Apulia, South-Eastern Italy) and the LP haplotype with the second event from the North-Western Italy. A positive association was observed also between the LP haplotype H_2 and haplogroups I1-M253* and I2-P215*, but for these two linages was not reported an estimate date and their origin and distribution are less clear and defined (Boattini et al, 2013).

Observations and hypotheses formulated considering the distribution of Y-chromosome haplogropus were confirmed also comparing that of mitochondrial DNA haplogroups with those of LP/NLP haplotypes. In fact, frequency patterns of LP haplotypes did not show any correlation with that of mtDNA haplogroups probably because all of them showed age estimations higher than 10,000 yrs (Boattini et al. 2013), thus suggesting a most recent entrance of LP in Italy. Interestingly, the distribution of a lot of mtDNA haplogroups showed instead a significant association with that of NLP haplotypes.

Lactase persistence is unquestionably one of the most studied and explored traits implicated in the complex and still on-going nutritional shifts experienced by human populations after the Neolithic revolution. Although the high number of studies that inspected the variability of the related genomic region, a great part of the underlying variation was, until now, not completely investigated and understood. Furthermore, it represents a clear example of the importance of genome-culture co-evolutionary processes and of how their investigation could depict a new and more exhaustive history of human evolution and migrations.

That being so, the conducted extensive analyses of a large region of the human genome implicated in the lactase persistence status made possible to confirm evidences already observed in other human populations, as well as to lay the foundations for new hypothesis and speculations.

The presence of a longer block of LD in the two Northern Italian samples compared with the Central-Eastern/Southern and Sardinian ones suggests a possible stronger

impact of natural selection in these Italian subpopulations, preventing recombination to break down the adaptive haplotype. The presence of the highest haplotype variability, nucleotide diversity, as well as of a great percentage of unique haplotypes, among which many carrying the functional variant, also suggests that these subpopulations may represent an independent source point for the evolution of LP-involved haplotypes with respect to other northern European populations.

Actually similar consideration may be done also for the Omanis of Asian origin (OAO) subsample considered in our Arabian Peninsula dataset, and, although to a less extent, for the Arabs belonging to Northern Oman (ANO). In fact, these two samples showed higher values of diversity with a pattern often comparable with the European groups. As confirmation of these results, a recent study used computer simulations in order to address important questions regarding the mode and direction of spread of the T-13,910 allele and the precise nature of the selective advantage conferred by LP (Itan et al. 2009). This study highlighted the importance of the combined effects of the selective pressure and of the demographic expansion of farmers during the Neolithic to explain the present distribution of the -13,910 variant. Moreover, Itan et al. stressed the importance of positive selection in all groups that adopted dairying cultures. For this reason, scientists inferred the most probable place of origin of the co-evolution between LP and dairying, proposing that it lies in a region between the central Balkans and Central Europe. This co-evolution, which occurred along the wave of advance of the Neolithic demographic expansion, is the key fact explaining the actual distribution of lactase persistence in Europe and its high frequencies in the North-western part of the continent (Itan et al. 2009).

Therefore, the results of this work support the hypothesis of possible differential selective pressures on LP-related genetic variants in different Italian subpopulations, also suggesting that once arrived in the Italian peninsula LP have evolved differently and independently with respect to what occurred in northern European populations.

The recent and actually detailed survey achieved by our research group about the Italian variability related to uniparentally inherited genetic markers (Y-chromosome and mitochondrial DNA, Boattini et al. 2013) allowed to further explore this scenario in the Italian samples correlating the distribution of LP or NLP

haplotypes with those of the more frequent Y-chromosome and mitochondrial DNA haplogroups in the Italian Peninsula. These analyses made possible to speculate a possible entrance and an inference of the age of the adaptive haplotype in Italy. We observed a correlation between two Y-chromosome haplogroups (i.e. R1b-L2 and G2a-L497) that presented a similar distribution, origin and estimated age. In particular the obtained results suggested that the functional mutation was entered in Italy about 3,000 years ago from Central Europe. This means that the adaptive haplotype was not arisen in Italy, but the high variability present in Northern Italy allows to speculate an independent evolution of LP-related variation in the Peninsula. Furthermore, the distribution of the functional variant in Italy confirms the two already described different migrations that characterized the Italian Neolithic period.

In conclusion, as expected considering their geographical, historical and archaeological records, in both our samples we observed a high and complex variability of the explored genomic trait that allowed to suppose the origin of the -13,915*G in Southern Arabia and an independent evolution of adaptive haplotypes carrying the -13,910*T allele in the Italian Peninsula.

6. References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. *An integrated map of genetic variation from 1,092 human genomes.* Nature.;491(7422):56-65. doi: 10.1038/nature11632.

Al-Abri AR, Al-Rawas O, Al-Yahyaee S, Al-Habori M, Al-Zubairi AS, Bayoumi R. 2012a. *Distribution of the lactase persistence-associated variant alleles -13910* T and -13915* G among the people of Oman and Yemen.* Hum Biol.;84(3):271-86. doi: 10.3378/027.084.0310.

Al-Abri A, Podgorná E, Rose JI, Pereira L, Mulligan CJ, Silva NM, Bayoumi R, Soares P, Cerný V. 2012b. *Pleistocene-Holocene boundary in Southern Arabia from the perspective of human mtDNA variation.* Am J Phys Anthropol.;149(2):291-8. doi: 10.1002/ajpa.22131. Epub 2012 Aug 24.

Anagnostou P, Battaglia C, Coia V, Capelli C, Fabbri C, Pettener D, Destro-Bisol G, Luiselli D 2009. *Tracing the distribution and evolution of lactase persistence in Southern Europe through the study of the T-13910 variant*. Am J Hum Biol. 21: 217-219.

Aoki K 2001. *Theoretical and empirical aspects of gene-culture coevolution*. Theor Popul Biol. 59: 253-261.

Arola H, Koivula T, Jokela H, Jauhiainen M, Keyriläinen O, Ahola T, Uusitalo A, Isokoski M 1988. *Comparison of indirect diagnostic methods for hypolactasia*. Scand J Gastrenterol. 23: 351-357.

Bandelt HJ, Forster O, Röhl A 1999. *Median- joining networks for inferring intraspecific phylogenies.* Mol Biol Evol. 16(1): 37-48.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. *Haploview: analysis and visualization of LD*

and haplotype maps. Bioinformatics. 21: 263-265.

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, Erhardt G. 2003. *Gene-culture coevolution between cattle milk protein genes and human lactase genes.* Nat Genet. 35(4):311-3.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN 2004. *Genetic Signatures of strong recent positive selection at the lactase gene.* Am J Hum Genet. 74: 1111-1120.

Blancher A, Socha WW 1997. *The ABO, Hh and Lewis blood groups in man and nonhuman primates. In: Blancher A, Jan Klein J, Socha WW (eds) Molecular biology and evolution of blood group and mhc antigens in primates.* Springer, Heidelberg, pp 30–92

Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, Sanz P, Yang-Yao D, Manry J, Ciani G, Luiselli D, Quintana-Murci L, Comas D, Pettener D; Genographic Consortium. 2013. *Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata.* PLoS One 8(5):e65441.

Britten RJ 2002. *Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels.* Proceedings of the National Academy of Sciences of the United States of America 99: 13633–13635.

Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG. 2007. *Absence of the lactase-persistence-associated allele in early Neolithic Europeans.* Proc Natl Acad Sci U S A;104(10):3736-41.

Burger J, & Thomas MG 2011. *The palaeopopulation genetics of humans, cattle and dairying in Neolithic Europe.* In R. Pinhasi, & J. Stock (Eds.), The bioarchaeology of the transition to agriculture (pp. 371e384). Chichester, UK: Wiley Blackwell.

Campbell MC, Tishkoff SA. 2008. *African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.* Annu Rev Genomics Hum Genet.;9:403-33. doi: 10.1146/annurev.genom.9.081307.164258.

Cavalli- Sforza LL, Menozzi P, Piazza A 1997. *Storia e geografia dei geni umani.* Adelphi. Milan. Italy.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, et al. 2005. *A genomewide comparison of recent chimpanzee and human segmental duplications.* Nature 437: 88–93.

Clark AG. 1990. *Inference of haplotypes from PCR-amplified samples of diploid populations.* Mol Biol Evol.;7(2):111-22.

Colonna V, Pagani L, Xue Y, Tyler-Smith C. 2011. *A world in a grain of sand: human history from genetic data.* Genome Biol. 12(11):234. doi: 10.1186/gb-2011-12-11-234.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. 2006. *A worldwide survey of haplotype variation and linkage disequilibrium in the human genome.* Nat Genet. Nov;38(11):1251-60. Epub 2006 Oct 22.

Cook GC, Al-Torki MT 1975. *High intestinal lactase concentrations in adult Arabs in Saudi Arabia*. Br Med J. 3: 135-143.

Copley MS, Berstan R, Dudd SN, Docherty G, Mukherjee AJ, Straker V, Payne S, Evershed RP. 2003. *Direct chemical evidence for widespread dairying in prehistoric Britain.* Proc Natl Acad Sci U S A;100(4):1524-9.

Craig OE, Chapman J, Heron C, Willis LH, Bartosiewicz L, Taylor G, Whittle A and Collins M. 2005. *Did the first farmers of central and eastern Europe produce dairy foods?* Antiquity, 79 (306). pp. 882-894. ISSN 0003-598 Xentral and Eastern Europe

produce dairy foods? Antiquity, 79, 882e894.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW 2006. *The evolution of mammalian gene families.* PLoS One. 1:e85.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I 2002. *Identification of a variant associated with adult-type hypolactasia*. Nat Genet 30: 233-237.

Enattah NS, Trudeau A, Pimenoff V, Maiuri L, Auricchio S, Greco L, Rossi M, Lentze M, Seo JK, Rahgozar S, Khalil, Alifrangis IM, Natah S, Groop L, Shaat N, Kozlov A, Verschubskaya G, Comas D, Bulayeva K, Mehdi SQ, Terwilliger JD, Sahi T, Savilahti E, Perola M, Sajantila A, Järvelä I, Peltonen I 2007. *Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans*. Am J Hum Genet. 81: 615-625.

Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, Natah A, Ali A, Natah S, Comas D, Mehdi SQ, Groop L, Vestergaard EM, Imtiaz F, Rashed MS, Meyer B, Troelsen J, Peltonen L. 2008. *Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture.* Am J Hum Genet.;82(1):57-72. doi: 10.1016/j.ajhg.2007.09.012.

Evershed RP, Payne S, Sherratt AG, Copley MS, Coolidge J, Urem-Kotsu D, Kotsakis K, Ozdoğan M, Ozdoğan AE, Nieuwenhuyse O, Akkermans PM, Bailey D, Andeescu RR, Campbell S, Farid S, Hodder I, Yalman N, Ozbaşaran M, Biçakci E, Garfinkel Y, Levy T, Burton MM. 2008. *Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding.* Nature;455(7212):528-31. doi: 10.1038/nature07180.

Excoffier L, Laval G, Schneider S 2005. *Arlequin (version 3.0): An integrated software package for population genetics data analysis*. Evol Bioinfom Online. 1: 47-50.

Excoffier L, Slatkin M 1995. *Maximum-likelihood estimation of molecular haplotype*

frequencies in a diploid population. Mol Biol Evol 12: 921-927.

Fang L, Ahn JK, Wodziak D, Eric Sibley E 2012. *The human lactase persistence-associated SNP –13910*T enables in vivo functional persistence of lactase promoter–reporter transgene expression.* Human Genet. 131:1153-1159.

Fedele, F. 2009. *Early Holocene in the highlands: data on the peopling of the eastern Yemen plateau, with a note on the Pleistocene evidence.* Pages 215–236 in Petraglia & Rose 2009.

Flatz G, Rotthauwe HW 1973. *Lactose nutrition and natural selection*. Lancet. 302: 76-77.

Gabriel S, Ziaugra L, Tabbaa D 2009. *SNP Genotyping using the sequenom MassARRAY iPLEX platform*. Curr Protoc Hum Genet. 60: 2.12.1-2.12.18.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. *The structure of haplotype blocks in the human genome.* Science. Jun 21;296(5576):2225-9. Epub 2002 May 23.

Groot PC, Bleeker MJ, Pronk JC, Arwert F, Mager WH, Planta RJ, Eriksson AW, Frants RR. 1989. The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics;5(1):29-42.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ 2008. *Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping*. BMC Genomics. 18: 9-80.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD, Comas D. 2012. *Genomic ancestry of North Africans supports back-to-Africa migrations.* PLoS Genet. 8(1):e1002397. doi: 10.1371/journal.pgen.1002397.

Holm, S 1979. *A simple sequentially rejective Bonferroni test procedure.* Scandinavian

Journal of Statistics. 6: 65-70.

Iacus SM 2006. *Statistica*. McGraw-Hill. Milano. Italy

Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N, Swallow DM. 2007. *A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence?* Hum Genet. 120(6):779-88. Epub 2006 Nov 21.

Ingram CJ, Mulcare CA, Itan Y, Thomas MG, Swallow MD 2009. *Lactose digestion and the evolutionary genetics of lactase persistence.* Hum Genet. 124: 579-591.

International HapMap Consortium. 2007. *A second generation human haplotype map of over 3.1 million SNPs*. Nature. 449(7164):851-61.

International Human Genome Sequencing Consortium. 2004. *Finishing the euchromatic sequence of the human genome*. Nature. 431,931-945 doi:10.1038/nature03001.

Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG 2009. *The origins of lactase persistence in Europe*. PLoS Comput Biol. 5: e1000491.

Jacob R, Peters K, Naim HY. 2002. *The prosequence of human lactase-phlorizin hydrolase modulates the folding of the mature enzyme*. J. Biol. Chem. 277:8217–25

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. 2008. *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature. Feb 21;451(7181):998-1003. doi: 10.1038/nature06742.

Järvelä I, Torniainen S, Kolho KL 2009. *Molecular genetics of human lactase deficiencies*. Ann Med. 41: 568-575.

Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. 2014. *Human evolutionary genetics 2Ed.* Taylor & Francis. New York. USA.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet.;11:94. doi: 10.1186/1471-2156-11-94.

Jones BL, Swallow DM. 2011. *The impact of cis-acting polymorphisms on the human phenotype.* Hugo J.;5(1-4):13-23. doi: 10.1007/s11568-011-9155-4. Epub 2011 Jul 20.

Jorde LB, Watkins WS, Bamshad MJ, Dixon MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA 2000. *The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosome data.* Am J Hum Genet. 66(3): 979-988.

Jurinke C, van den Boom D, Cantor CR, Koster H 2002. *Automated genotyping using the DNA MassArray technology*. Methods Mol Biol. 187: 179-192.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A. 2012. *New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing.* Nat Commun.;3:698. doi: 10.1038/ncomms1701.

Laland KN, Odling-Smee J, Myles S. 2010. *How culture shaped the human genome: bringing genetics and the human sciences together.* Nat Rev Genet.;11(2):137-48. doi: 10.1038/nrg2734.

Leonardi M, Gerbault P, Thomas MG, and Burger J. 2012. *The evolution of lactase persistence in Europe. A synthesis of archaeological and genetic evidence.* International Dairy Journal 22(2):88-97.

Liu X, Ong RT, Pillai EN, Elzein AM, Small KS, Clark TG, Kwiatkowski DP, Teo YY. 2013. *Detecting and characterizing genomic signatures of positive selection in global populations.* Am J Hum Genet.;92(6):866-81. doi: 10.1016/j.ajhg.2013.04.021.

Lösch S, Grupe G, Peters J. 2006. *Stable isotopes and dietary adaptations in humans and animals at pre-pottery Neolithic Nevalli Cori, southeast Anatolia.* Am J Phys Anthropol.;131(2):181-93

McCorriston, J. and Martin, L. 2009. *Southern Arabia's early Pastoral Population History: Some Recent Evidence. The Evolution of Human Populations in Arabia, Vertebrate Paleobiology and Paleoanthropology*. Springer Science +Business Media B.V. Chapter17, pp 237-250.

Malmström H, Linderholm A, Lidén K, Storå J, Molnar P, Holmlund G, Jakobsson M, Götherström A. 2010. *High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe.* BMC Evol Biol.;10:89. doi: 10.1186/1471-2148-10-89.

Neel JV. 1999. The "thrifty genotype" in 1998. Nutr Rev.;57(5 Pt 2):S2-9.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. *Recent and ongoing selection in the human genome.* Nat Rev Genet. Nov;8(11):857-68.

Novembre J, Ramachandran S. 2011. *Perspectives on human population structure at the cusp of the sequencing era.* Annu Rev Genomics Hum Genet.;12:245-74. doi: 10.1146/annurev-genom-090810-183123.

Ogasawara K, Bannai M, Saitou N, Yabe R, Nakata K, Takenaka M, Fujisawa K, Uchikawa M, Ishikawa Y, Juji T, Tokunaga K 1996a. *Extensive polymorphism of ABO blood group gene: three major lineages of the alleles for the common ABO phenotypes.* Hum Genet 97:777–783

Ogasawara K, Yabe R, Uchikawa M, Saitou N, Bannai M, Nakata K, Takenaka M, Fujisawa K, Ishikawa Y, Juji T, Tokunaga K 1996b. *Molecular genetic analysis of variant phenotypes of the ABO blood group system.* Blood 88:2732–2737

Olds LC, Sibley E 2003. *Lactase persistance DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element.* Hum Mol Genet. 12: 2333-2340.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet.;39(10):1256-60.

Pessina A, Tine` V. 2008. *Archeologia del Neolitico. L9Italia tra il Vi e il IV millennio a.C.* Roma: Carrocci editore. 375

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight JR, Mullikin JC, Meader SJ, Ponting CP, Lunter G, Higashino S, Hobolth A, Dutheil J, Karakoç E, Alkan C, Sajjadian S, Catacchio CR, Ventura M, Marques-Bonet T, Eichler EE, André C, Atencia R, Mugisha L, Junhold J, Patterson N, Siebauer M, Good JM, Fischer A, Ptak SE, Lachmann M, Symer DE, Mailund T, Schierup MH, Andrés AM, Kelso J, Pääbo S. 2012. *The bonobo genome compared with the chimpanzee and human genomes.* Nature. 486(7404):527-31. doi: 10.1038/nature11128.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL,

Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S 2014 *The complete genome sequence of a Neanderthal from the Altai Mountains.*Nature. 505(7481):43-9. doi: 10.1038/nature12886.

Reznick DN, Ricklefs RE. 2009. Darwin's bridge between microevolution and macroevolution. Nature. Feb 12;457(7231):837-42. doi: 10.1038/nature07894.

Rose, J.I. 2010. *New light on human prehistory in the Arabo-Persian Gulf Oasis.* Current Anthropology 51: 849–883

Rose JI, Černý V, Bayoumi R. 2013. *Tabula rasa or refugia? Using genetic data to assess the peopling of Arabia.* Arabian Archaeology and Epigraphy. Special Issue: The Neolithic of Arabia – New Paradigms and Future Perspectives Volume 24, Issue 1, pages 95–101.

Sabatti, C & Risch, N. 2002. *Homozygosity and linkage disequilibrium.* Genetics 160, 1707–1719.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. *Positive natural selection in the human lineage.* Science. Jun 16;312(5780):1614-20.

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. 2003. *Minimal haplotype tagging.* Proc Natl Acad Sci U S A. Aug 19;100(17):9900-5. Epub 2003 Aug 4.

Spearman C. 1906 *"Footrule" for measuring correlation*. Brit. J Psychol. 2, 89-108.

Sherratt, A. 1981*. Plough and pastoralism: aspects of the secondary products revolution.* In I. Hodder, G. Isaac, & M. Hammond (Eds.), Pattern of the past: Studies in honour of David Clarke (pp. 261e302). Cambridge, UK: Cambridge University Press.

Sherratt, A. 1983. *The secondary exploitation of animals in the Old World.* World Archaeology, 15, 90e104.

Simoons FJ 1978. *The geographic hypothesis and lactose malabsorption: A weighing of the evidence.* Am J Dig Dis. 23: 963-980.

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

Strachan T & Read A 2011. *Human Molecular Genetics*. Garland Science. New York. USA

Stoneking M. 2008. Human origins. The molecular perspective. EMBO Rep.;9 Suppl 1:S46-50. doi: 10.1038/embor.2008.64.

Sverrisdóttir OO, Timpson A, Toombs J, Lecoeur C, Froguel P, Carretero JM, Arsuaga Ferreras JL, Götherström A, Thomas MG. 2014. *Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe.* Mol Biol Evol.

Swallow DM 2003. *Genetics of lactase persistence and lactose intolerance.* Annu Rev Genet. 37: 197-219.

Tishkoff SA, Verrelli BC 2003. *Patterns of human genetic diversity: implications for human evolutionary history and disease.* Annu Rev Genomics Hum Genet. 4: 293-340.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P 2007. *Convergent adaptation of human lactase persistence in Africa and Europe.* Nat Genet. 39: 31-40.

Tomkins JP 2013. *Comprehensive Analysis of Chimpanzee and Human Chromosomes Reveals Average DNA Similarity of 70%.* Answers Research Journal 6:63–69.

Uerpmann, H.P., Potts, D.T. & Uerpmann, M. 2009. *Holocene (re-)occupation of Eastern Arabia.* Pages 205–210 in Petraglia & Rose 2009.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. *A map of recent positive selection in the human genome.* PLoS Biol. Mar;4(3):e72. Epub 2006 Mar 7.

Vuorisalo T, Arjamaa O, Vasemägi A, Taavitsainen JP 2012. *High Lactose Tolerance in North Europeans: A Result of Migration, Not In Situ Milk Consumption*. Perspsect Bio Med. 55: 163-174.

Wolff AE, Jones AN, Hansen KE 2008. *Vitamin D and musculoskeletal health*. Nat Clin Pract. 4: 580-588.

Zvelebil M 2000. *Archaeogenetics: DNA and the population history of Europe.* 57-70. Ed Boyle K., MacDonald Institute Cambridge, Cambridge.

Websites visited

www.ub.edu
www.r-project.org
www.1000genomes.org
www.broadinstitute.org
pngu.mgh.harvard.edu/~purcell/plink/
www.fluxus-engineering.com
stephenslab.uchicago.edu

Appendix

Appendix Table 1. Primers Design.

WELL	SNP ID	2nd-PCRP	1st-PCRP	Single Base Exstension
W1	*rs1519529*	ACGTTGGATGTCAGTAGAGGAACCATTTGC	ACGTTGGATGCTCTCCTTTAGGGTTTTCTC	GCTCTCTTGCATTTCCT
W1	*rs1519523*	ACGTTGGATGGGGACTTGGGAGAAACGACT	ACGTTGGATGGCAGTCAAGCAAATGGACAA	GAGAAACGACTTCCCTG
W1	*rs1346731*	ACGTTGGATGAATTAGTCTCAATGTGGTC	ACGTTGGATGACCACAAATCCATCAGAGGC	CAATGTGGTCTGCTGTG
W1	*rs2322254*	ACGTTGGATGCTGAAGTTTCACACGGAGAG	ACGTTGGATGGTTTTTCACCTGGGATTGGC	ACACGGAGAGAAAAAGAC
W1	*rs1470457*	ACGTTGGATGAGTAGCACAGGCTCTGGATG	ACGTTGGATGTACAGGGCACCTCTTAGAGC	GTGGCCTTGTTCCTGATCC
W1	*rs745500*	ACGTTGGATGTTTCTGATAAGTGGATGGG	ACGTTGGATGAATGAAGAGCTGCGGCACAC	cACCTCCTCGTGTGCCTCCT
W1	*rs1011361*	ACGTTGGATGTTACTTGTGAGCCTGGGAAC	ACGTTGGATGAGGCTTTTCACTCAGAGTGC	tAACTTTGGCTTCTTAACCT
W1	*rs2236783*	ACGTTGGATGTAGGCTTTTCTGTGGGCATC	ACGTTGGATGAAGTGATCAAGGAGCTTCCC	gTTCTGTGGGCATCACTGAG
W1	*rs1432232*	ACGTTGGATGAATTGCTCTGCTGATACCCC	ACGTTGGATGATGTGAAGGTAGAACGCAGG	CTGCTGATACCCCTGAAATAT
W1	*rs1438307*	ACGTTGGATGAGTCCAGGAAGTGGAGGAA	ACGTTGGATGATCATGCCTCACTGCAGGCG	GCGGTAAGAGGGATGACTAAC
W1	*rs4954411*	ACGTTGGATGATCATGGCCCACCTTTTCAC	ACGTTGGATGTGCAGCAAGACGTTTTCAGG	tCTTTTCACCCCAGCCGATGTC
W1	*rs867563*	ACGTTGGATGATCCCTCTACATCTTGGGAC	ACGTTGGATGTCTCTGCCTCTGAGCCAAAC	AAACAGTTCTTGTAGATGTTCC
W1	*rs842360*	ACGTTGGATGCAAACCAGCTTGTCTGTCTC	ACGTTGGATGTGGGAAGAAGGGTTAGGTAG	cAGCCCTGTGTTCCTCCTGGCCT
W1	*rs935612*	ACGTTGGATGGTGGGTTAAAGAGCAGTGTA	ACGTTGGATGCCTCTGAATCTGCTAGATGG	tAGCAGTGTAAATAAAACATGGC
W1	*rs309181*	ACGTTGGATGTGGCAATCTTTTACCAGCAC	ACGTTGGATGGCTGTGGTATACACAATTGA	ACCAGCACATATTAGATCAACTTT
W1	*rs4988235*	ACGTTGGATGATGTACTAGTAGGCCTCTGC	ACGTTGGATGAGGAGGAGAGTTCCTTTGAG	GGCAATACAGATAAGATAATGTAG
W1	*rs1996589*	ACGTTGGATGCCTGCTCTCTTGAACGTTTC	ACGTTGGATGTCTCTTGGGACCACAGGGC	aTTGAACGTTTCCATTGTTATAAGC
W1	*rs766271*	ACGTTGGATGCACACAGTAACATCTCTAC	ACGTTGGATGTAGGAGGAAGACCTGTTTAC	CACAGTAACATCTCTACTAGCATTCA
W1	*rs2290518*	ACGTTGGATGGCGAATGATTAACATGTCTG	ACGTTGGATGCTCAGTTTATGAGGGTGAGC	gAACATGTCTGTTTTTAGACTATTAT
W1	*rs309176*	ACGTTGGATGGCCACTGAACTTATACACTG	ACGTTGGATGTAAAATTGTGATAAAAATAC	ATACACTGAAAAATGGTTAAAATGAT
W1	*rs309180*	ACGTTGGATGGTGTATCTTCCAGCCTTGTC	ACGTTGGATGAATGAGGTAACCCCATGAGC	cGTGTATCTTCCAGCCTTGTCTCACTC
W2	*rs2322659*	ACGTTGGATGACGGATCCGCGTCTTCATC	ACGTTGGATGAGTTCATGGGAGGCTGGTTT	GCGTCTTCATCACCTCA
W2	*rs309137*	ACGTTGGATGACAGTGGCCGTTGGTGGTG	ACGTTGGATGTCACTTCTGCTGTTGCCACC	GTGACAAAGGGAAACGC
W2	*rs749017*	ACGTTGGATGAGGCATGCCTGCATAACCAC	ACGTTGGATGCACAGCCCCTCTTGAAACTG	CTGCATAACCACCTGGAA
W2	*rs1257168*	ACGTTGGATGGTCAAGTATGCTAATGGGAG	ACGTTGGATGCCCCTACTATTCACATCAGC	aTAATGGGAGAAGGAGCA
W2	*rs182549*	ACGTTGGATGGTCCTTAAAAACAGCATTCTC	ACGTTGGATGCCAAAGTACTGGGACAAAGG	cCAGCATTCTCAGCTGGGC
W2	*rs1257220*	ACGTTGGATGTGCTGGTGCAGAGACTGTAG	ACGTTGGATGCTAATGGCCTATGCATCCAC	AGAGACTGTAGGATTCTCA
W2	*rs2164210*	ACGTTGGATGTTGTAGGGACAAACCACATC	ACGTTGGATGCCCTGTGGAGCCAACCCA	caACAAACCACATCCAAACC
W2	*rs4954228*	ACGTTGGATGAGAGTCAAATCACTGGTTGC	ACGTTGGATGCATATATTCTTAGCACTGGG	TGCATTTTCATCATTCTAGG
W2	*rs578935*	ACGTTGGATGTTGTGGTGGTAGGAAGTTGG	ACGTTGGATGCAAGGTCAGACAGAAAATGG	AGGAAGTTGGTACTCTGAAT
W2	*rs3213889*	ACGTTGGATGTAATTTCAAGTACTGCATC	ACGTTGGATGCTTCACTAGAATATACCTC	CAAGTACTGCATCATTATACC
W2	*rs3754686*	ACGTTGGATGGCTTCCTACCCTTAACTGC	ACGTTGGATGCATTTTTCCTTCTGCTTTG	cTTCAAAGTTTTATCACCCTAT
W2	*rs876338*	ACGTTGGATGTGCTCTCTTTATGGACTTGG	ACGTTGGATGGACCCTGTAGGTCTCTTTGA	GACTTGGAACATATAAAGACTT
W2	*rs1531957*	ACGTTGGATGGGGATATTTCACTGCAAAAC	ACGTTGGATGATGGGGTGCCATTCACATTG	ggCACTGCAAAACTCTAATGGCA
W2	*rs518614*	ACGTTGGATGTCTGTTTCCATCTCCACTCC	ACGTTGGATGTGTCTGATAGTGCAGCATTG	ttTTTCCATCTCCACTCCTTTAAG
W2	*rs2305248*	ACGTTGGATGTCACACCCTCTGGGGTAATT	ACGTTGGATGTGCAGCAATGTCATAAAGTG	aCTCTGGGGTAATTTTCCTGATCC
W2	*rs882374*	ACGTTGGATGGTCTCACTGCCTCCTTGATG	ACGTTGGATGTTCCCCTCCTTCATTGCTAC	GGGGACCATTTGGGAAAAGACTGG
W2	*rs1427588*	ACGTTGGATGAATTGCTGATCACAGCTGAC	ACGTTGGATGGGCCTATCAGGTATTTTGAG	gTCACAGCTGACATCTTTTATCTCG
W2	*rs1399604*	ACGTTGGATGGGAACACTGAGATGAAGACC	ACGTTGGATGTTTTGTGAAGAATTCTTGG	gCCAATGACAACTGTGAAAGTTCTT
W2	*rs574135*	ACGTTGGATGCTCACGAGGCTTTAAAAAGT	ACGTTGGATGAAGGGATGTGGGTAATGATG	TTTGTCTTATGGGTAAATGGTGGCTA

Appendix Table 2. : Comparison of allele frequencies between NEI-NCWI

CHR	SNP	BP	MA	NEI	NCWI	CHISQ	P	ADJ. P
2	rs1346731	137612587	A	0.101	0.186	7.125	7.60E-03	0.297
2	rs1011361	136553639	T	0.450	0.491	0.801	0.371	1.000
2	rs1257168	134963892	T	0.377	0.404	0.354	0.552	1.000
2	rs1257220	135015347	A	0.283	0.359	3.027	0.082	1.000
2	rs1399604	137130665	G	0.326	0.257	2.719	0.099	1.000
2	rs1427588	137492326	G	0.298	0.307	0.046	0.830	1.000
2	rs1432232	137799664	A	0.234	0.271	0.903	0.342	1.000
2	rs1438307	136499166	A	0.450	0.486	0.626	0.429	1.000
2	rs1470457	136581848	T	0.450	0.524	2.636	0.105	1.000
2	rs1519523	136934449	T	0.205	0.257	1.848	0.174	1.000
2	rs1519529	136974257	G	0.268	0.205	2.625	0.105	1.000
2	rs1531957	134759307	T	0.290	0.281	0.046	0.830	1.000
2	rs182549	136616754	T	0.266	0.276	0.061	0.805	1.000
2	rs1996589	134865196	G	0.330	0.419	4.091	0.043	1.000
2	rs2164210	136580287	G	0.405	0.421	0.113	0.737	1.000
2	rs2236783	136594158	T	0.407	0.457	1.254	0.263	1.000
2	rs2290518	135878814	T	0.282	0.229	1.745	0.187	1.000
2	rs2322254	135750849	C	0.414	0.376	0.702	0.402	1.000
2	rs2322659	136555659	C	0.471	0.529	1.580	0.209	1.000
2	rs309137	136765951	T	0.392	0.419	0.355	0.551	1.000
2	rs309176	136622216	C	0.380	0.429	1.150	0.284	1.000
2	rs309180	136614255	A	0.381	0.438	1.601	0.206	1.000
2	rs309181	136614813	G	0.378	0.429	1.291	0.256	1.000
2	rs3213889	136511575	G	0.453	0.491	0.666	0.414	1.000
2	rs3754686	136603276	G	0.410	0.448	0.690	0.406	1.000
2	rs4954228	135976498	G	0.226	0.181	1.490	0.222	1.000
2	rs4954411	137076425	C	0.511	0.448	1.912	0.167	1.000
2	rs4988235	136608646	T	0.245	0.276	0.623	0.430	1.000
2	rs518614	137716851	C	0.409	0.495	3.531	0.060	1.000
2	rs574135	137740120	C	0.435	0.505	2.348	0.126	1.000
2	rs578935	137210991	C	0.263	0.271	0.048	0.827	1.000
2	rs745500	136583192	T	0.407	0.457	1.254	0.263	1.000
2	rs749017	135573659	C	0.423	0.482	1.461	0.227	1.000
2	rs766271	135667131	C	0.245	0.262	0.190	0.663	1.000
2	rs842360	135347885	T	0.423	0.371	1.312	0.252	1.000
2	rs867563	137142500	G	0.295	0.248	1.346	0.246	1.000
2	rs876338	137289147	G	0.486	0.481	0.010	0.919	1.000
2	rs882374	137913295	A	0.276	0.232	1.142	0.285	1.000
2	rs935612	135941503	C	0.234	0.181	2.007	0.157	1.000

(CHR=chromosome, BP=position of the SNP, MA=minor allele in NCWI population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Appendix Table 3. Comparison of allele frequencies between CESI-SARD

CHR	SNP	BP	A1	CESI	SARD	CHISQ	P	ADJ. P
2	rs1257168	134963892	T	0.289	0.478	11.550	6.79E-04	0.026
2	rs1996589	134865196	G	0.386	0.543	7.273	7.00E-03	0.273
2	rs1519523	136934449	T	0.198	0.096	5.272	0.022	0.845
2	rs1011361	136553639	T	0.355	0.415	1.105	0.293	1.000
2	rs1257220	135015347	A	0.322	0.370	0.736	0.391	1.000
2	rs1346731	137612587	A	0.173	0.096	3.297	0.069	1.000
2	rs1399604	137130665	G	0.264	0.287	0.196	0.658	1.000
2	rs1427588	137492326	G	0.289	0.266	0.180	0.671	1.000
2	rs1432232	137799664	A	0.280	0.298	0.116	0.734	1.000
2	rs1438307	136499166	A	0.361	0.394	0.336	0.562	1.000
2	rs1470457	136581848	T	0.377	0.340	0.425	0.514	1.000
2	rs1519529	136974257	G	0.167	0.239	2.507	0.113	1.000
2	rs1531957	134759307	T	0.239	0.330	3.111	0.078	1.000
2	rs182549	136616754	T	0.120	0.053	3.412	0.065	1.000
2	rs2164210	136580287	G	0.292	0.294	0.001	0.981	1.000
2	rs2236783	136594158	T	0.296	0.287	0.024	0.876	1.000
2	rs2290518	135878814	T	0.270	0.277	0.014	0.906	1.000
2	rs2322254	135750849	C	0.509	0.468	0.496	0.481	1.000
2	rs2322659	136555659	C	0.377	0.351	0.202	0.653	1.000
2	rs309137	136765951	T	0.272	0.294	0.157	0.692	1.000
2	rs309176	136622216	C	0.282	0.277	0.011	0.918	1.000
2	rs309180	136614255	A	0.277	0.277	6.49E-06	0.998	1.000
2	rs309181	136614813	G	0.277	0.277	6.49E-06	0.998	1.000
2	rs3213889	136511575	G	0.361	0.394	0.336	0.562	1.000
2	rs3754686	136603276	G	0.296	0.287	0.024	0.876	1.000
2	rs4954228	135976498	G	0.205	0.234	0.362	0.548	1.000
2	rs4954411	137076425	T	0.431	0.500	1.405	0.236	1.000
2	rs4988235	136608646	T	0.101	0.054	1.862	0.172	1.000
2	rs518614	137716851	C	0.440	0.415	0.184	0.668	1.000
2	rs574135	137740120	C	0.491	0.447	0.557	0.456	1.000
2	rs578935	137210991	C	0.248	0.277	0.303	0.582	1.000
2	rs745500	136583192	T	0.293	0.287	0.010	0.922	1.000
2	rs749017	135573659	C	0.362	0.359	0.003	0.953	1.000
2	rs766271	135667131	C	0.173	0.117	1.687	0.194	1.000
2	rs842360	135347885	T	0.296	0.238	1.070	0.301	1.000
2	rs867563	137142500	G	0.263	0.245	0.122	0.727	1.000
2	rs876338	137289147	A	0.469	0.511	0.515	0.473	1.000
2	rs882374	137913295	A	0.253	0.250	0.004	0.950	1.000
2	rs935612	135941503	C	0.217	0.234	0.123	0.726	1.000

(CHR=chromosome, BP=position of the SNP, MA=minor allele in SARD population, CHISQ=chi square value, P=p value, ADJ.P= p value after Bonferroni correction)

Appendix Table 4: Spearman's rank correlation coefficient between LP/NLP haplotype and Y-chromosome haplogroups (only those haplotypes and haplogroups with at least N=10 are reported).

			R1b1b2a1b4-U152	G2a-(xL497)	E1b1b1a2-V13	J2a-M410	R1b1b2-M269	R1b1b2a1b4c-L2	R1b1b2a1b-P312
	Haplotypes	**N**	97	68	67	67	57	51	44
H_1	CACTACCCACGCCTC	371	-0.02	0.23	0.02	0.24	0.03	**-0.52**	-0.25
H_2	**AGTCGTTTGTAGTCT**	118	0.05	**-0.55**	0.08	-0.30	-0.20	**0.60**	0.39
H_3	AGTCGTTTGCAGCCT	98	0.20	0.09	-0.36	-0.13	0.00	-0.08	0.14
H_4	CACCATCCACGCCTC	47	-0.10	0.47	0.03	-0.15	-0.05	-0.32	-0.29
H_5	AGTTACCCACGCCTC	40	-0.26	0.36	-0.28	0.32	0.41	-0.50	-0.21
H_6	AGTCGTTTGCGCCTC	14	0.15	0.01	0.00	0.14	0.12	0.14	-0.22
H_7	CACTACCCACGCCTT	10	0.11	-0.12	-0.30	0.35	0.23	0.01	0.20

	I2a1-M26	R1a1a-M17	G2a-L497	E1b1b1c-M123	I1-M253	J2a2-M67	R1b1b2a1a-U106	J1-M267	J1e-P58	J2a2a-M92
	36	27	25	23	21	19	16	15	15	15
H_1	-0.20	0.21	-0.31	0.20	-0.49	-0.23	-0.27	0.30	0.28	-0.02
H_2	-0.15	-0.27	**0.64**	-0.22	**0.68**	-0.03	0.22	-0.17	-0.31	-0.19
H_3	0.29	**0.56**	-0.06	0.17	-0.31	0.25	0.00	-0.05	-0.29	-0.19
H_4	0.09	0.33	-0.26	-0.06	-0.27	-0.06	0.37	-0.34	0.15	0.20
H_5	0.20	0.06	-0.28	**0.51**	-0.41	0.49	-0.26	0.31	-0.23	0.07
H_6	0.14	-0.26	-0.25	0.38	0.05	0.31	-0.27	0.00	-0.24	0.15
H_7	0.25	-0.03	0.05	-0.17	-0.29	0.18	0.17	0.19	0.21	0.18

	R1b1b2a1a4-L48	L-M20	T-M70	J2b2-M241	I2-P215	R1b1b2a1b5-L21
	15	13	13	12	10	10
H_1	-0.07	0.22	0.16	0.07	**-0.69**	-0.20
H_2	0.24	-0.01	-0.09	-0.25	**0.59**	0.14
H_3	-0.29	-0.19	0.03	0.19	-0.29	0.12
H_4	-0.22	0.03	0.17	0.13	-0.14	0.04
H_5	-0.07	-0.24	-0.11	-0.03	-0.34	-0.18
H_6	-0.36	0.15	-0.13	-0.22	-0.18	-0.51
H_7	0.09	-0.32	-0.14	**0.52**	0.26	-0.05

Correlations that resulted statistically significant are reported in bold type.

Appendix Table 5: p-values obtained from the correlation index between LP/NLP haplotype and Y-chromosome haplogroups (only those haplotypes and haplogroups with at least N=10 are reported).

			R1b1b2a1b4-U152	G2a-(xL497)	E1b1b1a2-V13	J2a-M410	R1b1b2-M269	R1b1b2a1b4c-L2	R1b1b2a1b-P312
	Haplotypes	**N**	97	68	67	67	57	51	44
H_1	CACTACCCACGCCTC	371	0.945	0.348	0.951	0.331	0.916	**0.028**	0.312
H_2	**AGTCGTTTGTAGTCT**	118	0.842	**0.019**	0.751	0.226	0.426	**0.008**	0.107
H_3	AGTCGTTTGCAGCCT	98	0.416	0.735	0.141	0.612	0.989	0.738	0.571
H_4	CACCATCCACGCCTC	47	0.700	0.050	0.900	0.544	0.845	0.189	0.248
H_5	AGTTACCCACGCCTC	40	0.293	0.137	0.252	0.194	0.095	0.034	0.404
H_6	AGTCGTTTGCGCCTC	14	0.562	0.983	0.986	0.590	0.625	0.584	0.381
H_7	CACTACCCACGCCTT	10	0.664	0.649	0.219	0.159	0.358	0.955	0.435

	I2a1-M26	R1a1a-M17	G2a-L497	E1b1b1c-M123	I1-M253	J2a2-M67	R1b1b2a1a-U106	J1-M267	J1e-P58	J2a2a-M92
	36	27	25	23	21	19	16	15	15	15
H_1	0.420	0.396	0.214	0.427	0.039	0.355	0.276	0.233	0.259	0.931
H_2	0.540	0.278	**0.004**	0.385	**0.002**	0.913	0.389	0.502	0.213	0.456
H_3	0.250	**0.015**	0.811	0.506	0.208	0.320	0.998	0.838	0.251	0.459
H_4	0.718	0.180	0.293	0.827	0.275	0.828	0.131	0.163	0.560	0.427
H_5	0.419	0.803	0.260	**0.030**	0.093	0.040	0.292	0.216	0.353	0.784
H_6	0.567	0.294	0.315	0.119	0.834	0.215	0.273	0.993	0.329	0.559
H_7	0.308	0.916	0.847	0.506	0.241	0.466	0.502	0.440	0.413	0.468

	R1b1b2a1a4-L48	L-M20	T-M70	J2b2-M241	I2-P215	R1b1b2a1b5-L21
	15	13	13	12	10	10
H_1	0.791	0.381	0.532	0.785	**0.001**	0.423
H_2	0.337	0.982	0.729	0.307	**0.009**	0.579
H_3	0.249	0.445	0.897	0.455	0.243	0.634
H_4	0.376	0.897	0.489	0.616	0.583	0.863
H_5	0.783	0.343	0.657	0.897	0.161	0.472
H_6	0.147	0.558	0.616	0.389	0.464	0.031
H_7	0.730	0.192	0.590	**0.028**	0.303	0.852

Correlations that resulted statistically significant are reported in bold type.

Appendix Table 6: Spearman's rank correlation coefficient between LP/NLP haplotype and mitochondrial DNA haplogroups (only those haplotypes and haplogroups with at least N=10 are reported).

			H1	H	H5	K1a	HV	H3	T1a	J1c	T2b	U4
	Haplotypes	**N**	85	83	33	33	31	31	26	26	25	25
H_1	CACTACCCACGCCTC	371	-0.11	0.33	-0.21	-0.09	**-0.57**	0.07	-0.11	-0.17	-0.29	0.42
H_2	**AGTCGTTTGTAGTCT**	118	-0.29	-0.19	0.38	0.44	0.32	-0.13	0.01	0.04	0.06	-0.13
H_3	AGTCGTTTGCAGCCT	98	**0.73**	-0.29	-0.27	0.00	**0.51**	-0.01	-0.05	0.11	**0.53**	-0.23
H_4	CACCATCCACGCCTC	47	0.28	-0.02	0.21	-0.30	0.28	-0.30	-0.12	0.18	0.08	**-0.61**
H_5	AGTTACCCACGCCTC	40	0.49	-0.25	**-0.53**	0.14	-0.16	0.15	0.34	-0.27	-0.10	0.11
H_6	AGTCGTTTGCGCCTC	14	0.11	-0.06	-0.13	0.34	-0.25	0.14	-0.02	-0.12	0.18	0.13
H_7	CACTACCCACGCCTT	10	-0.06	0.49	0.01	-0.24	-0.10	0.01	-0.11	-0.25	-0.24	0.15

	H8	J2b	N1	H13	V	T2	U4a	I	W	T	J2a	J1b
	21	18	16	16	14	14	14	13	12	11	11	10
H_1	-0.01	0.36	-0.15	0.16	-0.11	0.00	-0.12	-0.26	0.19	-0.22	-0.19	0.03
H_2	-0.08	-0.32	0.23	0.09	0.40	-0.06	-0.18	-0.03	-0.01	-0.01	-0.38	-0.07
H_3	-0.13	-0.03	0.22	-0.49	-0.06	0.03	0.27	0.18	-0.42	-0.02	0.30	0.00
H_4	-0.48	0.21	-0.12	-0.31	-0.45	0.21	0.36	0.41	-0.03	0.42	0.49	0.03
H_5	0.17	0.14	-0.03	0.13	0.07	-0.04	0.27	-0.04	-0.14	-0.06	**0.55**	0.22
H_6	0.13	-0.21	-0.15	0.08	0.14	-0.33	0.18	0.13	-0.30	-0.15	0.09	0.19
H_7	0.31	0.11	-0.15	0.12	-0.23	-0.08	-0.12	-0.14	-0.16	-0.03	0.33	0.28

Correlations that resulted statistically significant are reported in bold type.

Appendix Table 7: p-values obtained from the correlation index between LP/NLP haplotype and mitochondrial DNA haplogroups (only those haplotypes and haplogroups with at least N=10 are reported).

			H1	H	H5	K1a	HV	H3	T1a	J1c	T2b	U4
	Haplotypes	**N**	85	83	33	33	31	31	26	26	25	25
H_1	CACTACCCACGCCTC	371	0.670	0.182	0.397	0.715	**0.014**	0.782	0.669	0.510	0.239	0.086
H_2	**AGTCGTTTGTAGTCT**	118	0.247	0.448	0.124	0.071	0.195	0.599	0.960	0.883	0.805	0.602
H_3	AGTCGTTTGCAGCCT	98	**0.001**	0.248	0.282	0.993	**0.030**	0.984	0.857	0.667	**0.023**	0.365
H_4	CACCATCCACGCCTC	47	0.261	0.935	0.408	0.232	0.258	0.222	0.625	0.464	0.748	**0.007**
H_5	AGTTACCCACGCCTC	40	0.039	0.315	**0.024**	0.591	0.529	0.556	0.168	0.287	0.707	0.664
H_6	AGTCGTTTGCGCCTC	14	0.670	0.801	0.602	0.170	0.317	0.571	0.938	0.631	0.466	0.610
H_7	CACTACCCACGCCTT	10	0.808	0.039	0.979	0.336	0.698	0.963	0.670	0.314	0.342	0.540

	H8	**J2b**	**N1**	**H13**	**V**	**T2**	**U4a**	**I**	**W**	**T**	**J2a**	**J1b**
	21	18	16	16	14	14	14	13	12	11	11	10
H_1	0.983	0.143	0.559	0.531	0.662	0.988	0.626	0.290	0.440	0.382	0.450	0.897
H_2	0.763	0.203	0.361	0.710	0.103	0.820	0.476	0.898	0.958	0.954	0.115	0.780
H_3	0.598	0.902	0.375	0.039	0.799	0.900	0.280	0.468	0.087	0.950	0.225	0.993
H_4	0.045	0.398	0.621	0.207	0.063	0.413	0.141	0.095	0.913	0.079	0.041	0.893
H_5	0.511	0.573	0.900	0.609	0.768	0.871	0.271	0.863	0.571	0.825	**0.017**	0.385
H_6	0.600	0.395	0.542	0.743	0.580	0.181	0.464	0.597	0.219	0.549	0.711	0.444
H_7	0.210	0.654	0.544	0.623	0.349	0.742	0.626	0.569	0.533	0.919	0.177	0.259

Correlations that resulted statistically significant are reported in bold type.