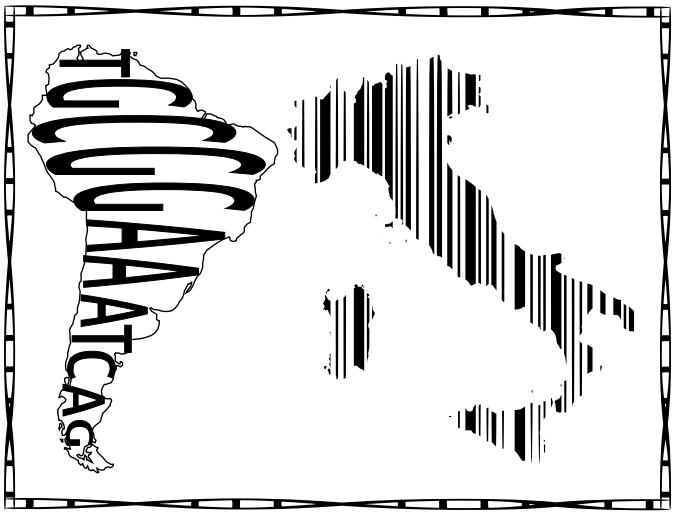
UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Human genetic history 2.0:

Y-chromosome NGS in South American populations

Genome-wide haplotype analysis in Italian populations



Alessandro Raveane

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXX – A.A. 2014-2017

On the cover:

- South America is filled by a genetic sequence since the data generated by Y-chromosome NGS is a continuous DNA sequence.

- Italy is represented as a barcode since the data used to generated genome-wide haplotypes are widespread SNPs along all the genome.

UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Human genetic history 2.0:

Y-chromosome NGS in South American populations

Genome-wide haplotype analysis in Italian populations

Raveane Alessandro

Supervised by Prof. Ornella Semino

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXX – A.A. 2014-2017

Abstract

The title of this thesis, "Genetic History 2.0", refers to novel tools in population genetics. It implies that classical molecular techniques and methodologies like PCR, RFLP, Sanger sequencing, belong to "Genetic History 1.0" in which the number "1.0" is the first version of the software named "Genetic History". Recently, high throughput sequencing methods generate a huge amount of data that could be managed and studied only with an updated version of the software, hence "Genetic History 2.0".

Two main projects with titles "Y-chromosome NGS in South American populations" and "Genome-wide haplotype analysis in Italian populations" are here presented.

Y-chromosome NGS in South American populations

South America is central to human prehistory, being the last continent colonized by modern humans and the site of multiple domestication hotspots. However, little is known about the genetic history and the demography of the first inhabitants of this area. Indeed, the drastic reduction of the native populations experienced during the colonization period, together with the post-contact newcomers' genetic contribution, have made the reconstruction of their genetic history really challenging. For the high rate of admixture between autochthonous people and the "Old World" newcomers, most of the present genetic information on Native Americans derives from studies on uniparental markers. As for Y chromosome, two founding haplogroups of Asian origin, Hg C and Hg Q, have been described so far. Hg C is virtually limited to North America. Differently, Hg Q is present all over the double continent and accounts for virtually all the Native South American Y chromosomes.

In order to shed light on the structure of haplogroup Q, in this thesis work about 1.5 Mb of the MSY have been re-sequenced in 34 unrelated males from different geographic regions and clustering within different Hg Q sub-lineages. The results analysed together with other 120 Hg Q MSY sequences available from literature.

Overall, this work provides the most updated Hg Q phylogeny and the phylogeographic distribution of all the new sub-lineages identified. Moreover, the results obtained, in agreement with recent archaeological findings in South America, are suggestive of a major phase of population growth (around 10 thousand years ago, kya), followed by a constant population size and a little sign of population growth from 3 kya, when a shift to a predominantly sedentary and agricultural subsistence occurred.

Genome-wide haplotype analysis in Italian populations

Italy territory is surrounded by the sea and bounded by the Alps, it extends over more than 1,000 km along a north-south axis and comprises the two largest islands of the Mediterranean, Sicily and Sardinia. The combination of this geographic complexity with a rich set of historical events and cultural dynamics had the potential to shape in a unique way the distribution of genetic variation within the Italian population. Recent investigations have in fact consistently reported substantial stratification in Italy when compared to other European countries, but a fine and exhaustive

characterisation of its population structure and admixture history is yet to be conducted.

In order to dissect the fine structure and the ancestry profile of Italian populations, after having genotyped 2.5 milion of SNPs in new set of Italian samples representative of the 20 administrative regions of Italy, I gathered from literature all the available genome-wide data of Italian subjects and I assembled two comprehensive genome-wide SNP datasets with different numbers of genetic markers. These two datasets, LDD (Low Density Dataset comprising 220,000 SNPs) and HDD (High Density Dataset, wiht 600,000 SNPs), included a total of almost 1,500 and 700 individuals, respectively, and were analysed together with data from additional ~ 300 world-wide reference populations.

The results of allele frequencies and haplotype-based analyses confirmed the distinctiveness of Sardinia compared to the rest of Italy and suggested a more complex distribution of genetic variation than the simple north-south differentiation previously reported. The richest biodiversity of Italy in Europe is also confirmed by F_{ST} analysis when compared with other similar data on UK and Spain populations. Recent and more ancient episodes of gene flow have contributed to shaping such diversity and I am currently working to dissect the underlying admixture history.

Acknowledgements

I am much obliged to Prof. Ornella Semino, Prof. Cristian Capelli, Dr. Francesco Montinaro, Prof. Anna Olivieri and Dr. Viola Grugni for the patient, the devotion and advices they gave me in realizing this work of thesis.

Additionally, I would like to thank Prof. Antonio Torroni together with Prof. Alessandro Achilli for the support they gave me.

Least but not last all my colleagues starting from Linda Ongaro, Serena Aneli, Marco Capodiferro, Stefania Brandini, Ryan Daniels, Miguel Gonzales Santos, Francesca Bastaroli, Nicola Rambaldi, Abir Hussain and Emily Bartolini that scientifically and not-scientifically provided me all the friendship and help that a PhD student needs. A special thank also for Simone Savino and Dr. Raefa Abou Kouzam always at my side in the last 5 years.

Finally, I also would like to acknowledge all the people that contributed in giving me published and un-published data (Dr. Barlera Simona, Prof. Bione Silvia, Dr. Boncoraglio Giorgio, Dr. Di Blasio Anna Maria, Dr. Parolo Silvia, Dr. Di Gaetano Cornelia, Dr. Dugoujon Jean-Michel, Prof. Kivisild Toomas, Prof. Lancioni Hoviragh, Prof. Metspalu Mait, Prof. Pagani Luca, Prof. Pascali Vincenzo, Dr. Brisghelli Francesca, Prof. Piazza Alberto, Dr. Ricaut François-Xavier, Prof. Paschou Peristera, Dr. Guarrera Simonetta, Dr. Cardinali Irene, Prof. Stamatoyannopoulos George, Dr. Melhaoui Mohammed, Dr. Baali Abdellatif, Dr. Cherkaoui Mohammed) , a special thanks to Clare Bycroft and Prof. Myers Simon for giving me access to the F_{ST} estimates of their unpublished work.

Finally, an aknoweledgments is for all the people that gave their DNA to be analysed.

Abbreviations

aDNA Ancient DNA

AFS Allele Frequency Spectrum

AMH Anatomically Modern Human

bp base pair

CI Confidence Interval

DHPLC Denaturing High Performance Liquid Chromatography

DNA Deoxyribonucleic acid

HDD High Density Dataset

Hg Haplogroup

HGDP Human Genome Diversity Project

HMM Hidden Markov Model

HTS High Throughput Sequencing

IBD Isolation by Distance

IBS Identical By State

kya Kilo Years Ago

LD Linkage Disequilibrium

LDD Low Density Dataset

LGM Last Glacial Maximum

MCMC Markov Chain Monte Carlo

ML Maximum Likelihood

MP Maximum Parsimony

MSY Male Specific region of the Y chromosome

mtDNA mithocondrial DNA

NGS Next Generation Sequencing

NRY Non-Recombining region of the Y chromsome

PAR Pseudo Autosomal Region

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

PSMC Pairwise Sequencing Markovian Coalescent Model

QS Phred Quality Score

RFLP Restriction Fragments Length Polymoprhism

SGDP Simon Genome Diversity Project

SNP Single Nucleotide Polymoprhism

SPR Subtree-Pruning-Regrafting

STR Short Tandem Repeat

TMRCA Terminal Most Recent Common Ancestor

YDNA Y -chromosome DNA

Contents

Abstract I
Acknowledgements III
Abbreviations IV
Contents VI
Overview 1
1. Background on population genetic-data 2
1.1 The beginning 2
1.2 The mitochondrial DNA (mtDNA) 3
1.2.3 Overview 3
1.2.3 Phylogeny 3
1.2.4 Phylogeography 5
1.2.5 Calibration of the molecular clock 7
1.3 The Y chromosome 8
1.3.1 Overview 8
1.3.2 Phylogeny 10
1.3.3 Phylogeography 13
1.3.4 Calibration of the molecular clock 15
1.3.5 Ancient Y chromosome 17
1.4 The future of uniparental markers 21
1.5 The human nuclear genome 22
1.5.1 Overview 22
1.5.2 Next Generation Sequencing (NGS) methods 23
1.5.3 SNP-typing: methods for assessing variation 23
1.5.4 Haplotype 24
1.5.5 Phylogeny 26
1.5.6 Genome wide and genome whole studies 27
2. Peopling of the world 31
2.1 The origin of the anatomically modern man 31
2.2 Dispersal routes from Africa 33
2.3 Admixture with Neanderthal 35
2.4 Peopling of Europe 35
2.5 Peopling of America 38
2.5.1 First peopling 38
2.5.2 Late peopling of Americas 39
3. Aims of the research 40
4. Materials and methods 41
4.1 NGS Y-chromosome Analysis 41
4.1.2 Variants Calling 41

4.1.3 Dataset Merging 42
4.1.4 Parsimony tree and Network 43
4.1.5 Time Estimation (Rho-statistic) 45
4.1.6 Time Estimation (Bayesian methos) 45
4.2 Genotyping of South-American Y chromosomes 47
4.2.1 DNA quantification 49
4.2.2 Polymerase Chain Reaction (PCR) 50
4.2.3 Restriction Length Polymorphism (RFLP) Analysis 51
4.2.4 Electrophoretic Analysis 51
4.2.5 Sanger Sequencing 52
4.3 Analysis on the allele frequencies of Italian populations 54
4.3.1 PLINK 1.9 format and merging the data 54
4.3.2 Filtering and pruning 55
4.3.3 Principal Component Analysis (PCA) 55
4.3.4 Population cluster analysis 56
4.3.5 F3 Statistic 57
4.3.6 F_{ST} analysis 58
4.4 Haplotype-based analysis on the Italian populations 60
4.4.1 Phasing 60
4.4.2 CHROMOPAINTERv2 Analysis 61
4.4.2 FineSTRUCTURE Analysis 62
5. Results and Discussions 65
5.1 Y-chromosome NGS in South American populations 65
5.1.1 Overview 65
5.1.2 The dataset and the variants 66
5.1.3 Phylogeny of the haplogroup Q 67
5.1.3 Refinement of haplogroup Q phylogeny structure 69
5.1.4 Phylogeography of the haplogroup Q 70
5.1.5 Bayesian analysis 78
5.2 Genome-wide haplotype analysis in Italian populations 80
5.2.1 Overview 80
5.2.2 The dataset 82
5.2.3 PCA analysis of allele frequency in Italian and European context 84
5.2.4 ADMIXTURE analysis of allele frequencies in a worldwide context 86
5.2.5 Inferring signal of admixture on the Italian populations 87
5.2.6 Inferring clusters using CHROMOPAINTER and fineSTRUCTURE on the worldwide populations 89
5.2.7 Fst Analysis 98
6. Conclusion and perspective 102

6.1 Analyses of the NGS data on the South American populations..... 102
6.2 Analysis based on allele frequencies and haplotypes of Italian population..... 102
Appendix..... 105
References..... 119
List of original manuscripts..... 140

Overview

One of the main purposes of Evolutionary Genetics is to understand the relationships among species, populations and individuals, which is often reached by analyzing their genomic differences. The comparisons between distantly related groups give the opportunity to gain insights on ancient evolutionary processes; on the contrary, the analysis of the degree of relatedness between more closely related individuals can shed light on more recent and ancient events. Human Evolutionary Genetics, thus, measures the variation among human genomes and between humans and their closest primate relatives'genomes. In recent years, the ability to retrieve ancient and archaic genomic data has further extended our ability in recovering our past history, providing contribution to a variety of disciplines: from anthropology and paleoanthropology to medical and population genetics.

The study of human variation has roots to the beginning of the last century. In the early 1900 Karl Landsteiner discovered the ABO system, which allowed to classify all humans into four different blood groups (Landsteiner 1901). The variations in frequencies of these groups in the worldwide populations and their heritability led geneticists to consider these elements in the investigation of the human ancestry. Following these first attempts, in the 1940's and 50's variations in blood proteins were used as "classical markers" in the field of human evolution, paving the way to the use of genetics as a tool for understanding the historical and demographic events that shaped human populations.

In the 70's the first methods for detecting DNA variations appeared, but the real revolution occurred with the invention of the Polymerase Chain Reaction (PCR) in the mid-1980s (Mullis 1987). This methodology allows amplifying a specific fragment of DNA hundreds or thousands of times, making easy to detect similarities and differences along genomes. At the beginning, the only elements to be systematically explored were loci of haploid systems, first the mtDNA (Cann 1994), and then the non-recombining portion of the Y chromosome (Hammer and Zegura 1996). Since they are uniparentally inherited (from mother to all of her children and from father to sons, respectively) and evolve only by the accumulation of mutational events, their analysis offers a relatively straightforward view of the evolutionary history of our species.

In the last decades, the application of new techniques, which allowed the collection of information of genome-wide markers or a sequence of the whole genome (or a portion of it) (Next-Generation Sequencing: NGS), has generated an impressive amount of novel genetic information. The combination of these new techniques with the advanced biostatistical methodologies and information coming from other disciplines such as archaeology, history, linguistic among others is trying to give a clearer picture on the history and evolution of modern humans.

1. Background on population genetics

1.1 The beginning

DNA of living humans can be a really powerful source of information available nowadays. In 1984 Ammerman and Cavalli-Sforza proposed a model to explain the diffusion of the agriculture as combination of acculturation and movement of people. This model, that can be considered one of the first example of statistics applied to genetics, had some limitations. Hence, the genetic data used, belonged to modern populations that were distant thousands of years from the first appearance of agriculture, therefore on this population a genetic drift as well as multiple migration events could have easily reshaped the allele frequencies and this might bring to a misinterpretation of the results obtained. In this model, in order to prove the Neolithic diffusion, allele frequencies for classical polymorphisms had been considered. In particular 94 alleles at 34 *loci* distributed among 16 different chromosomes were analyzed on different populations. A principal component analysis (PCA) was generated to summarize all the information came from these genetic data. Moreover, Cavalli-Sforza and colleagues displayed the individual PCs values from the output of the PCA, as a synthetic map (Fig. 1). This method should allow common patterns, caused by single migrations, to be abstracted from genetic data. Geographic information became an important heuristic tool that apparently enables easy recognition of various patterns, and has led to important results in terms of migrations and archaeological records (Fig 1.)

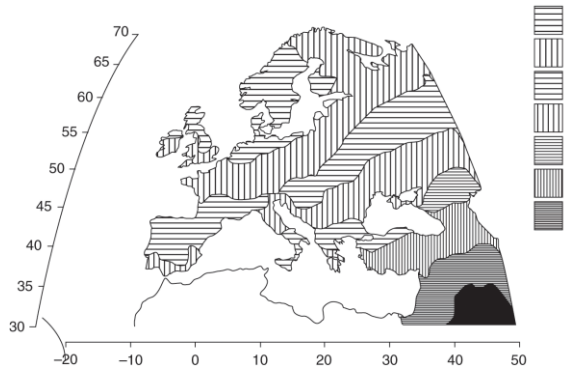


Figure 1. Synthetic map of the first principal component of variation in ninety-five classical genetic. Markers. (From Cavalli-Sforza et al., 1994)**.**

1.2 The mitochondrial DNA (mtDNA)

1.2.3 Overview

The small dimension, 16,569 bp of double-stranded DNA, together with a great abundance (on average, each cell contains between 103 and 104 copies of the mitochondrial genome), makes mtDNA an interesting tool for genetic studies. Inherited along the maternal side, the full human mtDNA sequence was published in 1981 (Anderson et al., 1981). This first sequence was used as reference with the name "Cambridge Reference Sequence (CRS)", later modified in revised CRS. The human mitochondrial genome contains only 37 genes: 13 of these genes encode for proteins and the remaining 24, consisting of 2 ribosomal RNAs (rRNAs) and 22 tRNAs, are used for the translation of the 13 structural genes. The non-coding sequence of mtDNA is the control region, also called D-loop (1,121 bp long in the human, from nucleotide position np 16024 to np 576). It is the most polymorphic region of the human mtDNA genome, with polymorphisms concentrated in three hypervariable regions: HVS-I (nps 16024-16400), HVS-II (nps 44-340) and HVS-III (nps 438-576) that is the reason of its wide use in the past population and forensic genetic studies (Brandstätter et al., 2004). With the introduction of next-generation sequencing methods, it is easier sequencing the complete mitogenome and considering it as the "whole" genetic variation inherited by the mother-side, even if it is not completely true, since the complete maternal genetic variation should consider also the autosomal chromosomes.
Summarizing: the almost complete maternal inheritance, the lack of recombination, a high copy number per cell, and a fast mutation rate (see 1.2.5 section) are the characteristics that made the mtDNA the focus of evolutionary genetics studies. In the 1980's and 1990's, when the human genomes had not been sequenced yet, and the idea of a whole nuclear genome level population genetics was only a daydream for population geneticists, this genetic locus have been a fundamental tool.

1.2.3 Phylogeny

The genetic diversity of humans is relatively low if compared with the one of other great apes with the exception of western chimpanzees and eastern gorillas (Prado-Martinez et al., 2013; Xue et al., 2015). Low genetic diversity means that for any nuclear gene, one needs to sequence from tens to thousands of individuals to have a chance of finding SNPs that are informative for population genetic purposes. In the era of PCR and Sanger sequencing, it was more cost effective to uncover DNA sequence variation at population scale from mtDNA, characterized by a high mutational rate in the non-coding region, than from any nuclear locus. Furthermore, the lack of recombination allowed the data from coding and non-coding regions of mtDNA to be easily combined into the shape of a phylogenetic tree. The branches of this ever-growing tree, as more data became available, were labelled by distinctive mutations initially identified as restriction fragment length polymorphisms (RFLPs). Alphabetic labels were assigned to the most common branches that became to be known as mtDNA haplogroups (Torroni et al., 1993; 1994).

A haplogroup (Hg) is defined as a group of uniparental systems (MtDNA or Male specific Region of Y chromosome) carrying the same mutations in the same position and order. For this reason, alleles shared by two haplogroups confirmed a common ancestry, while alleles belonging to different haplogroups showed a separation at a certain time in the past, from which a series of mutations occurred indipendetely and accumulated differentiating the two haplogroups.

In the early 1980's, RFLP based studies with limited number of polymorphic sites, placed the root of human mtDNA in Asia (Denaro et al., 1981). Subsequently, a larger and more comprehensive study, including 195 polymorphic RFLP sites in 147 mtDNAs from five geographically distinct populations, revealed that the whole genetic variation in the present-day mitochondrial DNA can be reconducted to a single female lineage that lived around 200 kya in Africa (Cann et al., 1987). Subsequently, the so-called "high level of resolution" coming out from RFLP and HVS-I sequencing based studies increased the information level.

The actual structure of the worldwide mito-phylogenetic tree, represented in figure 2, is based on the analysis of more than 10,000 individuals. The root of this tree is placed in Africa, from which departs the first seven bifurcations that are inclusive of the sub-Saharan African branches (L0-L6), five of which (L3 excluded) are virtually African-specific. Outside Africa, these haplogroups are restricted to those areas, as Mediterranean Europe, West Asia, and Americas that had direct influences with African continent due to ancient and more recent contact.

The nomenclature of mtDNA haplogroups was introduced in the mid-1990s with A-G labels assigned to Asian and American haplogroups and H-K to European haplogroups. Due to the small number of markers and African mtDNAs analysed in that period, only a single haplogroup, L, appeared to harbour all the African variation (Torroni et al., 1993, 1994, 1995, 1996; Chen et al., 1995). The mtDNA nomenclature that is currently used (http://www.phylotree.org/) has a robust branch structure that has been determined through the rigorous and detailed analyses of the whole mtDNA genomes (van Oven and Kayser 2009). Many researches have contributed to the topological details of this tree by sequencing whole mitogenomes from several populations across the world (Fig. 2).

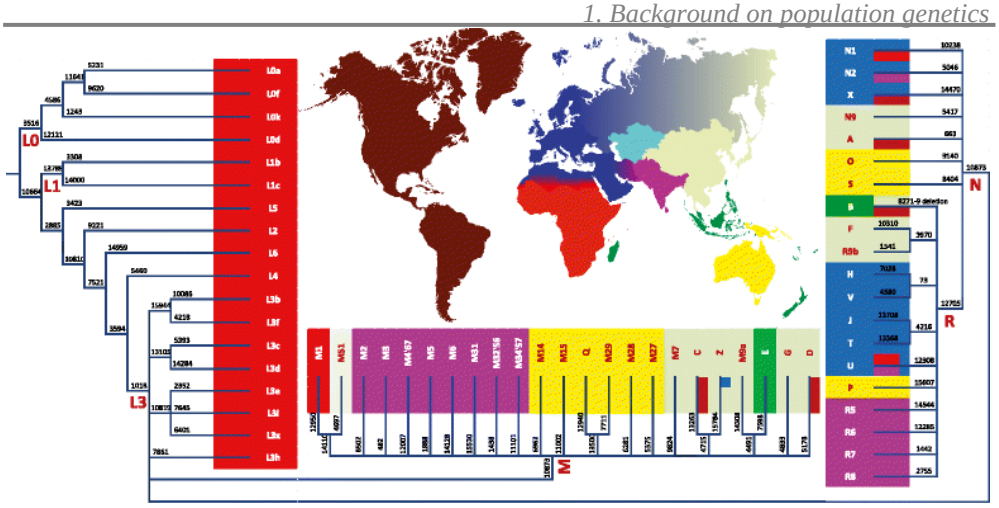


Figure 2. MtDNA Hg worldwide distribution. Labels are reported according to the van Oven and Kayser 2009 nomenclature. Only a single branch defining marker, preferably from the coding region, is shown. The main geographic features of haplogroup distribution are highlighted with colour (from Kivisild 2015).

1.2.4 Phylogeography

Phylogeographic approach represents one of the most efficient method to identify and quantify ancient and recent migrations that originated modern populations. It is an interdisciplinary approach that integrates genetic and phylogenetic data with climatological, ecological demographic, anthropological, archaeological, linguistic and historical data. Analysis of the phylogenetic position of each haplogroup, its geographical distribution and its internal variation allowed a reconstruction of the main migration routes during human history, included ancient and recent migration events.

The separation of the two sub-clades M and N from their African sister-clades L3 has been dated back to 62-95 kya (Fu et al., 2013), whereas the internal coalescent time of the M and N founders have been estimated in the range of 40 to 70 ky (Macaulay et al., 2005). This migration event has been associated with the largest known volcanic event occurred in human history: the eruption of Mount Toba occurred around 74 kya. Some archaeological artefacts (Petraglia and Dennell; 2007) and also some evidences deriving from ancient DNA (aDNA) extracted from the 45 kya Ust-Ishim (Western Siberia) skeletal remains whose mitogenome is associated with the root of haplogroup R, seem confirm this hypothesis (Fu et al., 2014). Contrary to the M and N lineages widely diffused in Asia, Australia and Oceania, their sub-clades are distributed in a more specific regional configuration (Fig. 2). The Eurasian lineages U, HV, JT, N1, N2 and X are diffused in Europe, South-West Asia and North Africa (Soares et al., 2010). Haplogroups R5-R8, M2-M6 and M4'67 are virtually confined to South Asia (Chaubey et al., 2007), while haplogroups A-G, Z and M7-M9 are widespread in East Asia (Stoneking and Delfin 2010) (Fig. 2)

Focusing on the Americas, the main Native American mtDNA haplogroups are five: A, B, C D and X (O'Rourke and Raff 2010). From results based on mitogenomes at

least 16 Native American sub-clades have been defined so far (Achilli et al., 2008, 2013; Brandini et al., 2017; O'Rourke and Raff 2010; Perego et al., 2009, 2010; Tamm et al., 2007). The diffusion of these sub-lineages is still object of study. The movement in North and South America has been associated with at least three distinct demographic events:

i) the main wave of the spread of the ancestors of both North and South American native populations occurred 15-18 kya and involving nine Pan-American founders A2*, B2*, C1b, C1c, C1d*, C1d1, D1, D4h3a, and D4e1c;

ii) approximately at the same time, Native American ancestors of C4c, X2a and X2g lineages entered America through an inland route of dispersal toward the east coast of the United States of America;

iii) the spread of Paleo-Eskimo D2a lineages around 5 kya along the Arctic through Northern Canada and Greenland, which were replaced, in the same region, by the spread of Neo-Eskimos carrying A2a, A2b, and D3 lineages. However, the association between haplogroup A2a and an ancestral Paleo-Eskimo route, inferred by the geographic distribution of modern mitogenomes (Achilli et al., 2013) has been recently questioning due to the little aDNA evidence that associated remains of Paleo-Eskimo cultures, as Saqqaq and Dorset, with haplogroup D2a (Raghavan et al., 2014). Future evidences, in particular on larger samples, will help to clarify this debate.

Several haplogroups have been associated with two major events in the peopling of Oceania: M14, M15, M27-M29, Q, P, O, and S, observed only in Australia and Melanesia, have been related to the first event of colonization and initial settlements in Sahul (Papua New Guinea and Australia) around 43-47 ka. B4a1a1 lineages, instead, is associated to a second event, more recent, often associated with the diffusion of the speaking Austronesian languages around 7 kya in Southern Australia (Kayser 2010; Tobler et al., 2017).

The results of these studies were not, however, sufficient to appease active debates on the attempts to define some issues such as, for example, the genetic source and number of Mesolithic and Neolithic gene flows in the peopling of Europe, the first lineages colonizing the Americas and the first peopling of the Oceania.

As for the peopling of Europe, initially it was proposed that the majority of the Eurasian branches were associated with a Late Glacial Maximum (LGM) colonization event (Soares et al., 2007); more recently aDNA evidence indicated (Brandt et al., 2013), that only a small fraction of European variation is attributable to a pre-Neolithic origin. This last point has been controversially discussed (Pala et al., 2012; Posth et al., 2016; Richards et al., 2016). Indeed, the European variation could be also the result of a strong introgression of a Near-Eastern component occurred around 4.5 kya as the aDNA analysis on nuclear genomes of Mesolithic and Neolithic samples revealed (Lazaridis et al., 2014, Olivieri et al., 2017).

1.2.5 Calibration of the molecular clock

Usually, patterns of uniparental-markers variations analyzed in human populations are associated with time models that create a molecular clock. The mutation rate of mtDNA genes in animals is known to be higher by at least an order of magnitude than the mutation rate of nuclear genes. This is associated with the high exposure to damage of one strand of the mtDNA molecule during the replication and/or transcription processes. Although the mutation rate can vary not only along the same species from individual to individual but also from cell to cell of the same individual, these changings do not represent a problem when calibrating a molecular clock. Indeed, these changes, which are attributed to the condition of heteroplasmy (the co-existence of wild-type and mutated mtDNAs in the same individual with ratio ranging from less than 5% up to 1:1 ratio) are usually solved after few generation, especially in somatic cells (Rebolledo-Jaramillo et al., 2014).

The divergence estimates of human mitogenomes from the chimpanzee outgroup was the first estimation of mtDNA mutation rate (1.70×10^{-8} substitution / site / year) (Horai et al., 1995; Ingman and Gyllensten 2001). Unfortunately, the use of a too distant outgroup, as in phylogenetic studies, revealed a problem: a phylogenetic approach produced mutational rate estimates which were at odds with the mutation rates estimated from pedigree data. In case of the hypervariable regions of the D-loop, several pedigree studies (Heyer et al., 2001; Howell et al., 2003; Santos, 2005) had inferred mutation rates (17×10^{-8} bp / year) that were up to an order of magnitude higher than the phylogenetic rate (5.7×10^{-8} bp / year) (Vigilant et al. 1991). Although it is encouraging to see recent aDNA based studies yielding concordant mutation rates for the whole mtDNA genome ($2.14 – 2.53 \times 10^{-8}$ bp / year) (Fu et al., 2014; Rieux et al., 2014), substantial differences are still noted among functional domains of the molecule. Now, the mutation rate in human mtDNA is estimated 2.4×10^{-8} bp / year. With this parameter it has been possible date back the TMRCA (Terminal Most Common Ancestor) for the mtDNA in the range of 100 to 200 thousand years ago (kya) (Barbieri et al., 2013; Rito et al., 2013). The estimations are close to the one of the Y-chromosome dates when considering the rare Y-chromosome haplogroup A00 lineages virtually present only in Western Africa (Mendez et al., 2013). Although the upper end of these time estimates is close to a period associated with the recent discoveries of African fossils considered the first appearance of Anatomically Modern Humans (AMH) (McDougall et al., 2011), the time back to TMRCA of a genetic *locus* depends also by the effective population size of the ancestral population. Therefore, the age of TMRCA for a single locus does not necessarily tell us the time of the emergence of the species, but instead it informs us about the origin of a small group of AMH which survived and gave origin to the actual human species (Li and Durbin 2011; Meyer et al., 2012).

1.3 The Y chromosome

1.3.1 Overview

The Y chromosome with its ~ 60 Mb of length is one of the smallest chromosome of the human karyotype together with chromosomes 21 and 22. It contains 63 coding genes, 109 non-coding genes and 392 pseudogenes (Hughes and Page 2015; Jobling and Tyler-Smith 2017; Verappa et al., 2013). It plays a crucial role in the determination of the sex at the level of the embryo and contains many genes influencing the male fertility. Like mtDNA, the Y chromosome avoids the recombination and is inherited, in this case, only from the paternal side of a phylogenetic tree (Jobling and Tyler-Smith 2003) (Fig. 3).

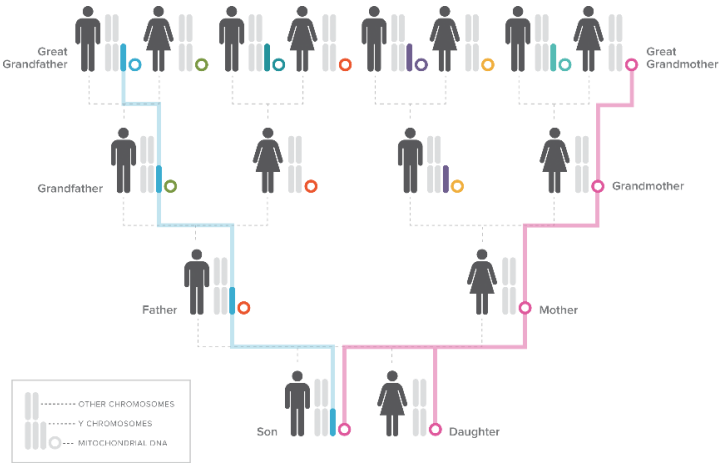


Figure 3. Uniparental transmission system. Family tree illustrating the holandric transmission (from father to son) of the Y chromosome (smallest blue bar) and the maternal transmission of mtDNA (purple circle). (From National Geographic's Genographic Project).

Because of its male specificity, Y chromosome has attracted the interest of scientists representing an important tool for various studies; for example, there are practical implications in forensic DNA analysis as well as in genetic genealogy (Jobling et al., 1997; Calafell and Larmuseau., 2017). Moreover, it has been an excellent tool for inferring human evolution and recent demographic history from a paternal perspective.

The structure of the Y chromosome consists in a major region called Male Specific region of Y chromosome (MSY), sometimes also known as the Non-Recombining region of Y chromosome (NRY), and three Pseudo Autosomal Regions (PAR) (Verrappa et al., 2013; Hughes and Page 2015). The latter are the only ones that, during meiosis, undergo to homologous recombination with their counterparts on the X chromosome; instead, the MSY is transmitted clonally in a holandric way. For these reasons, the MSY represents the counterpart of the maternally transmitted mtDNA and therefore a powerful tool in evolutionary and population genetic studies

(Fig. 3). Over the time, processes of random mutation without recombination, have generated groups of monophyletic chromosomes (haplogroups) characterized by the same combination of markers (Jobling et al., 1997; Jobling and Tyler Smith 2003), equal position and order; therefore, all modern MSY regions coalesce back to one ancestral sequence at some point in the past. Based on the number of possible different alleles, Y-chromosome variants can be subdivided into biallelic or multiallelic markers; in order to be useful in evolutionary and population genetic studies they need to be polymorphic, that means they must reach a frequency of at least 1% in a population. Alleles with lower frequencies are considered "rare variants".

Biallelic markers include any locus defined by a binary state, and between them there are Single Nucleotide Polymorphisms (SNPs), small insertions and deletions of nucleotides. MSY region accumulates SNPs through mutation at a higher average rate than other part of nuclear genome (Johnson and Lachance, 2012). This higher rate of mutation is ascribable to the larger number of cell divisions, and hence DNA replications, that occur in the male germline. In general, SNPs are considered to be unique onset markers during human evolution because they are characterized by a low mutation frequency (about 10^{-9} / base pair / generation), which makes very low the probability of a new mutation or reversion, reason why they are called biallelic.

Historically, the first Y-chromosome specific polymorphisms were reported in mid-1980s and consisted in RFLPs. The first, named 12f2/*Taq*I, was a 2kb deletion of a Y-chromosome LINE sequence (Casanova et al., 1985); the second, named 49a,f/*Taq*I, was a complex polymorphic system (Ngo et al., 1986) that identified many haplotypes. However, until 1997 only 11 binary polymorphisms that could be genotyped by PCR-based methods were known (Jobling, 1998; The Y-Chromosome Consortium YCC, 2001). Improvement of new technologies and methods (PCR, DHPLC and large-scale sequencing projects) increased the number of markers exponentially, allowing in 2000 the construction of the first worldwide phylogeny of the human Y-chromosome (Underhill et al., 2000). Nowadays, thousands of SNPs are available and specific databases for Y-chromosome have been assembled. The most complete and used are those of the International Society of Genetic Genealogy -ISOGG- (International Society of Genetic Genealogy, 2017), which includes also the information of the old YCC web site, YFULL (https://www.yfull.com/tree/) and 1000 Genomes Project (The 1000 Genome Project Consortium, 2010, 2015). In the last decades, NGS technologies have driven important progress in this field. In particular, NGS has yielded direct estimates of mutation rates, and an unbiased and calibrated molecular phylogeny that has unprecedented details. Moreover, the availability of direct to consumer NGS services is fuelling a rise of 'citizen scientists', whose interest in resequencing their own Y chromosomes is generating a wealth of new data (Jobling and Tyler-Smith 2017).

Multiallelic markers are the so-called tandem repeats consisting of a pattern of one or more nucleotides repeated adjacent to each other. They are mostly found within heterochromatin (highly condensed genome regions), thus forming non-coding DNA. They can be clustered according to the length of repeated units in: satellites, minisatellites and microsatellites. The most common multiallelic markers in Y

chromosome are microsatellite or Short Tandem Repeats (STRs). They are composed by unit of 2-6 nucleotides, which can be repeated from 10 to 20 times in tandem, forming very short fragments (less than 150 pairs of bases).
Microsatellites have been widely used in population genetics; they play also a relevant role in forensic cases and discoveries of genealogical relationships. The analysis of STRs contributes to: i) the discrimination between chromosomes belonging to the same haplogroup, thus identify potential sub-clustering; ii) the prediction of haplogroups through the use of different web tools or software; and iii) provides information on the relationships between the various evolutionary lines by estimating their internal variation. It is, indeed, possible to assign a mutation rate and time of divergence by comparing the differences between STRs of different lineages (Forster et al., 2000; Goldstein et al., 1995; Kayser et al., 2004). So far, more than 200 STRs *loci* have been identified on the MSY region of the human Y chromosome (Kayser et al., 2004) and the average Y-STR mutation rate is about 3.35×10^{-3} per generation (Ballantyne et al., 2010) but dating is still debated (Busby et al. 2012).

1.3.2 Phylogeny

Y-chromosome haplogroups consist on sets of Y chromosomes characterized by the same mutations per position and order. Modern Y-chromosome haplogroups, as the ones of mtDNA, can be organized, using the principle of maximum parsimony, in a phylogenetic tree that represents the evolution of the human Y chromosome.
In 2002, the YCC established a nomenclature for the phylogenetic tree: the main haplogroups are indicated by an uppercase letter. Their sub-haplogroups can be identified by the corresponding letter followed by the name of the terminal mutation, or by an alphanumeric code. In this case the first number after the haplogroup name indicates the sub-haplogroup, the following letter the sub-group of sub-haplogroup and so on. For example, it can be written J-M410 or J2a. In addition, internal lines to a given group, which are not characterized by any downstream markers, are defined as para-groups and they are indicated by the name of their haplogroup followed by an asterisk. A nomenclature such as J(xJ2a) means that there is a partial analysis within a clade so, in this example, it indicates all chromosomes within the J haplogroup except those belonging to sub-haplogroup J2a.
The human Y-chromosome phylogenetic tree includes 20 main haplogroups, all derived from an ancestral molecule that is considered the last common ancestral Y chromosome characterizing the most recent man to whom all living male individuals are linked by uninterrupted parental lines (Fig. 4).

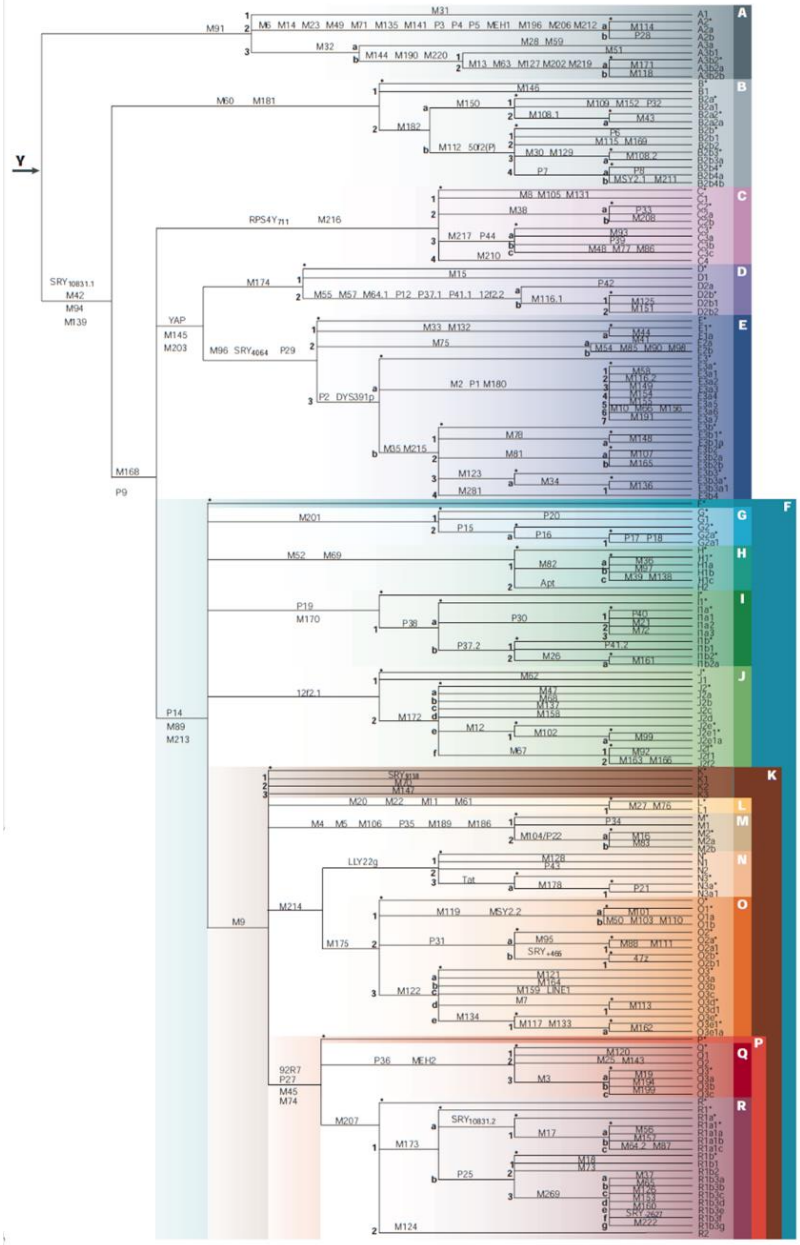


Figure 4. One of the first phylogeny of the human Y chromosome. The 20 haplogroups are indicated with different colours. At the root of the tree is displayed the Last Common Y-chromosome Ancestor (From Jobling and Tyler-Smith. 2003).

The progress of the technology in the NGS and High Throughput Sequencing (HTS) are providing a large amount of Y-chromosome SNPs (Francalacci et al., 2013; 2015; Poznik et al., 2013 and 2016; Rocca et al., 2012; Scozzari et al., 2014; Wei et al., 2013; Xue et al., 2009). This huge amount of Y SNPs is a source of complexity in defining a phylogenetic tree because most of them are redundant or extremely rare

and therefore not informative at population level. For this reason, the use Y SNPs in the field of evolution, anthropology, demography, and genealogy represents a big challenge.

Figure 5 illustrates the scheme of the updated Y-chromosome phylogenetic tree proposed by van Oven et al., (2014).

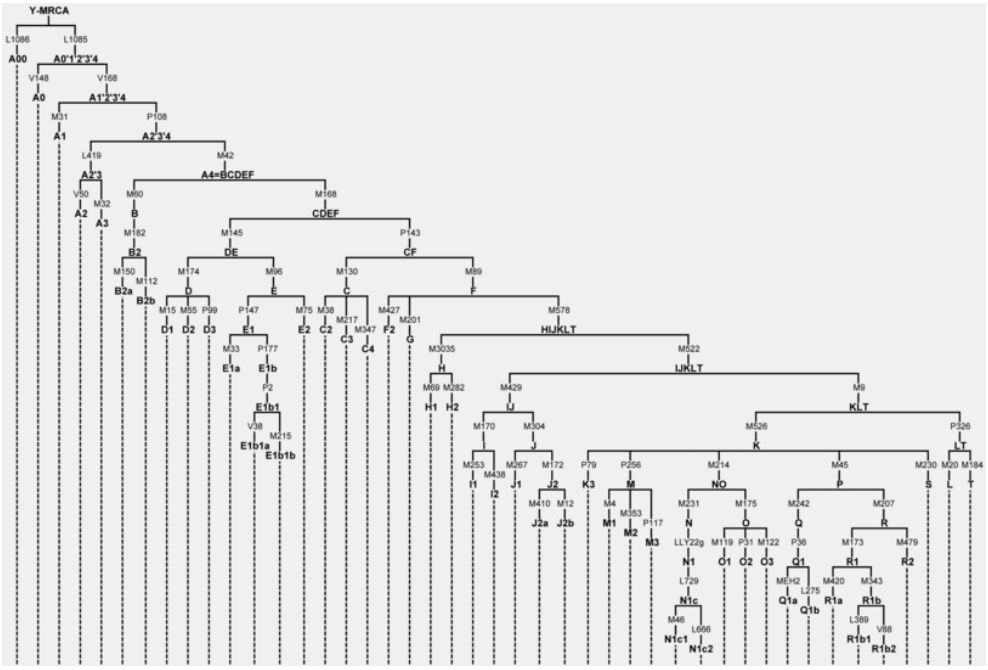


Figure 5. **Skeleton of the human Y-chromosome phylogeny**. Basal haplogroup nomenclature (in bold) and defining Y-SNP markers indicated on the branches. (From van Oven et al., 2014).

It includes the most informative Y-SNPs with the aim of providing a practical and minimal reference tree for human geneticists and citizen scientists in the human genetic fields. In the expanded version (Supplementary Fig. van Oven et al., 2014) it contains 417 primary branches deriving from thousands of already known and new Y-SNPs. The marker nomenclature is according to the research group that discovered them; therefore, they are mostly indicated with an alphanumeric code; when some redundancy occurs, the name is the most frequently used in the literature (for example, M173, P241 and Page 29 are all referring to the same mutation but in the tree it is reported as M173, the first published). The haplogroup nomenclature is according to the alphanumeric code, in which the main haplogroup is indicated with a letter followed by numbers when it is considered a sub-haplogroup.

Recently, the analysis of ~ 1,244 Y-chromosome worldwide sequences present in the 1000 genome project, allowed Ponzik et al., (2016) to refine the structure of the previous phylogenetic tree. The maximum-likelihood phylogenetic tree obtained (Fig. 6) represents all the haplogroups but M and S since none subjects belonging to these haplogroups were present in the dataset of the 1000 genome project.

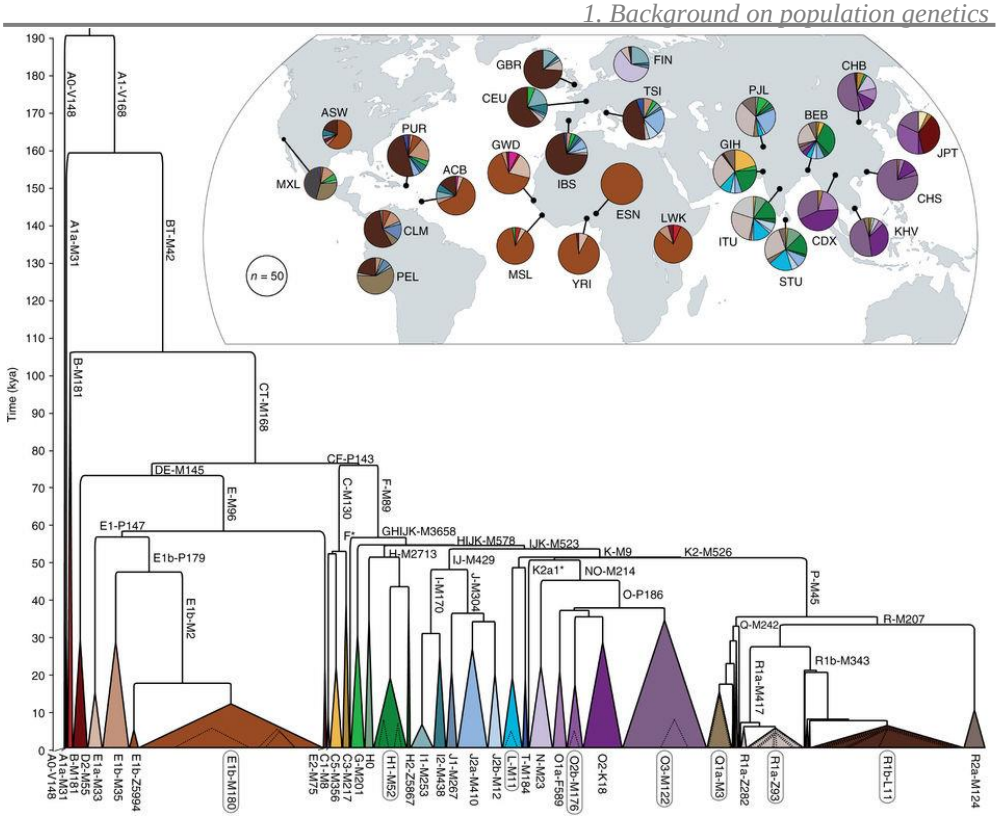


Figure 6. Phylogenetic tree obtained from the sequence of 1,244 Y chromsomes included in the 1000 Genome Project dataset. Coloured triangles represent the major clades, and the width of each base is proportional to one less than the corresponding sample size. Inset, world map indicating for each of the 26 populations: the geographic origin, sample size, and haplogroup frequencies according to the colours of the tree. (From Poznik et al., 2016).

1.3.3 Phylogeography

The worldwide distributions of the main Y-chromosome haplogroups are illustrated in figure 7.

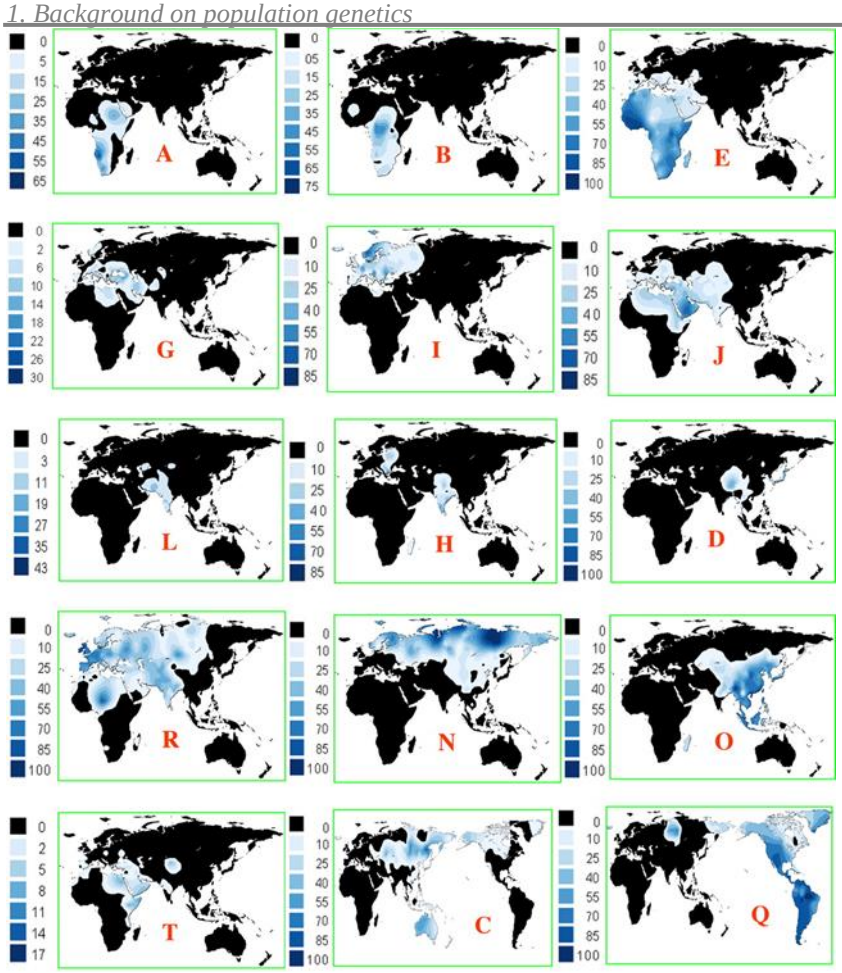


Figure 7. Geographic distributions of the main Y-chromosome haplogroups (From Chiaroni et al., 2009).

By considering the phylogenetic position and the geographic distribution of the main 20 Y-chromosome haplogroups it was possible to reconstruct their phylogeography. The most divergent haplogroups are A and B. They show a wide distribution at low frequencies and have been mainly found in many African regions. Out of Africa, only haplogroup A was observed in Sardinia where, however, it is ascribable to a recent event of migration. All the other haplogroups are observed mainly outside Africa. This indicates that current populations are the descendants of people who migrated out of Africa. Taking into account that uniparental systems evolve only by accumulation of mutational events, all the contemporary haplogroups are the results of the molecular differentiation occurred during and after the different migratory events in the last 50 kya. Therefore, haplogroups and sub-haplogroups tend to be restricted to specific geographic areas and population groups.

The Y-chromosome haplogroups observed in South-Eastern Asia are clearly different from the great part of haplogroups located in Central-Western Asia and

Europe. Haplogroups C and O have been found at high frequencies in the totality of South-Eastern Asia, haplogroups D is present in Japan and China but it was not observed in India while haplogroup M is present in New Guinea. In Europe and North-Western Asia, the most common haplogroups are the R, I, J and N. Something different come up when referring to the haplogroup Q. This haplogroup, which probably originated in Central Asia (Balanosky et al., 2017), is observed at low frequency in Europe and is the only Pan-American haplogroup characterizing more than 85% of Native American Y chromosomes.
By considering the distribution of haplogroups C, D, M and O and I, J, N and R in Eurasia two distinct routes of migration were proposed after the out of Africa along a southern costal route (Jobling and Tyler-Smith, 2003): one, more ancient, toward south-east, which have reached Melanesia and Australia, the another toward north-west reaching Western Eurasia (Hallast et al., 2014; Karmin et al., 2015). Then, a last event of migration from Central Asia open the road for the colonization of the Americas (Dulick et al., 2012). Analysis of the mtDNA show a similar scenario with a clear distinction between haplogroup N widely distributed in Europe and haplogroup M observed mainly in Eastern Asia (Chaubery et al., 2007; Stoneking and Delfin 2010). Based on the Y-chromosome tree illustrated in figure 6 and on the worldwide haplogroup variation (Fig. 7), the pattern of the world colonization by modern human was the following: i) Africa, ii) South Asia, iii) Southeastern Asia and Australia, iv) Central and Western Europe, iv) Pacific and Americas.
The new phylogenetic tree (Fig. 6) obtained by Poznik et al., 2016 confirm the structure of the Y-chromosome phylogeny and using a mutational rate of 0.76×10^{-9} mutations/bp/year (Fu et al., 2014), dated the TMRCA of the tree 190 ky old, and the TMRCA of all non-African branches (haplogroups DE and CF) ~76 ky old. In addition, they identified a clear expansion of different lineages around 50-55 kya, probably reflecting the expansion and the differentiation of Eurasian population. In this scenario, haplogroup E, the most frequent haplogroup in Africa, would be arisen out of the continent as previously proposed by Hammer et al., 1998. The new analysis identified a new out-group associated with a Vietnamese rare lineage of the haplogroup F. A new megagroup harbouring the non-African haplogroups G, H, I, J and K as well a new haplogroup H0, separated from the rest of haplogroup H have been identified.

1.3.4 Calibration of the molecular clock

A molecular clock is a tool that allows to measure the chronological distance between genetic regions on the basis of molecular distance. It is based on fossil and molecular studies and it is relevant for estimating the TMRCA. For a long time, the most used strategy for calibrating the Y-chromosome molecular clock was based on the microsatellite variation. The mutation rate of STR *loci* is several orders of magnitude higher than that of SNPs, with a high heterogeneity among *loci*. The STR estimated rate of mutations was obtained with two different methods: the analysis of deep rooted pedigrees (Bianchi et al., 1998; Forster et al., 2000; Heyer et al., 1997; Weber and Wong, 1993) and the comparison between father's and son's haplotypes

(Burgarella and Navascuè, 2011; Dupuy et al., 2001; Gusmão et al., 2005; Kayser et al., 2000). Both methods have estimated a mutation rate around 2×10^{-3} per *locus* per generation, a value very similar to that previously obtained for the autosomal microsatellites (Weber and Wong, 1993). However, these methods, which determine the mutation rate at meiosis, turned out to be an order of magnitude greater than that (about 10^{-4} per *locus* per a generation of 20 years) estimated on an evolutionary base (Forster et al., 2000). Thus, evolutionary methods were also set up. In particular, Zhivotovsky et al., (2004) evaluated an average mutation rate from a population rather than a family, using the known founding effect as a starting point for the production of the present diversity. The obtained evolutionary mutational rate was three to four times lower than that estimated by direct methods, and this result was explained as due to genetic drift caused by multiple bottlenecks during random fluctuations (Zhivotovsky et al., 2006) or due to the use of different set of STRs in the estimation (Busby et al., 2012). More recently, the possibility of sequencing large portions of the human genome has paved the way to new experimental approaches. These allowed to calibrate the molecular clock directly by counting the accumulated SNPs in the Y chromosome since the MRCA by using as out-group great apes' sequence (Cruciani et al., 2011; Wei et al., 2012) or by considering demographic events known from the archaeological records (Francalacci et al., 2013; Poznik et al., 2013).

Ponzik et al., 2016 used, instead, a slightly different approach; they considered Y-SNP phylogeny sufficiently detailed and precise to estimate a Y-STR's mutational dynamics. Therefore, they build a single phylogeny based only on the SNPs extracted from the Y chromosome sequencing. Then, they generated an error model for each microsatellite, considering the artefact and the alignment errors that can be generated by PCR, in particular in 1000 Genome Project's low-coverage data. Thanks to this model they generated, for each read along all the dataset, an expectation-maximization algorithm. The combination of this algorithm with uniform prior and a learned stutter model allows to generate the posterior genotype for each sample that, at the end, is representative of the probability of each leaf of the phylogeny. To calibrate the tree they used a mutation rate of 0.76×10^{-9} mutation per base pair per year based on their sequence analysis of a 45 ky old specimen (Fu et al., 2014). To validate their STR estimations they compared the mutation rates obtained with those extracted from a large scale father-son study (Willuweit and Roewer, 2007) and from an orthogonal high coverage dataset. The estimation obtained with the different methods resulted well correlated. The mutation rate of Ponzik et al. (2016) was used by Mendez et al. (2016) in order to estimate the Y-chromosome divergence of 120 kb orthologous region between a Neanderthal individual from Spain and modern human. From this comparison, they estimated TMRCA for Neanderthal and modern human Y chromosomes of about 588 kya while that of all the modern human Y chromosomes including A00, was evaluated by Catellano et al., (2014) as 275 kya.

1.3.5 Ancient Y chromosome

Improvements in the estimation of the Y-chromosome mutation rate derive from the analysis of Y-chromosome data from ancient remains. Indeed, using these data it is possible to test and define the number of mutations accumulated during the time that separate the remains from the recent samples. The difficulty in the analyses of ancient Y chromosome consists in common issues present in general ancient DNA studies. Especially, the DNA might be highly damaged or contain low level of endogenous DNA that, in addition, can be confused with X chromosome due to the high homology and the high level of repetition in some fragments. Most of the ancient Y chromosome are not suitable to build a tree with informative branch length, due to the low quality of the DNA that make not possible distinguish between a true mutation from an error caused by a damage.

As reviewed by Pickrell and Reich in the 2014 new capture-based methods, which allow only certain sequences surrounding specific SNPs to be sequenced, became available. This technology has the advantages of being certain of traits of DNA even if they are of poor quality but, on the other hand, its limitation is that it can identify only variants present on the capture chip. It was successfully applied for the first time on the nuclear DNA of a 40 ky old human specimen from Tianyuan Cave outside Beijing (Fu et al., 2013), thus suggesting a possible application also for regions of the Y chromosome.

So far, the most common method used to obtain Y-chromosome specific variation is to sequence long regions of this chromosome and then filter out SNPs from regions highly homologue to X chromosome (Karmin et al., 2015; Ponzik et al., 2013; Scozzari et al., 2014; Wei et al., 2013).

By the combination of ancient and modern Y-chromosome data, and knowing the dates of the aDNA it is possible to build a phylogenetic tree illustrating the positions of the ancient samples in the ancestral variation inferred from modern data (Fig. 8).

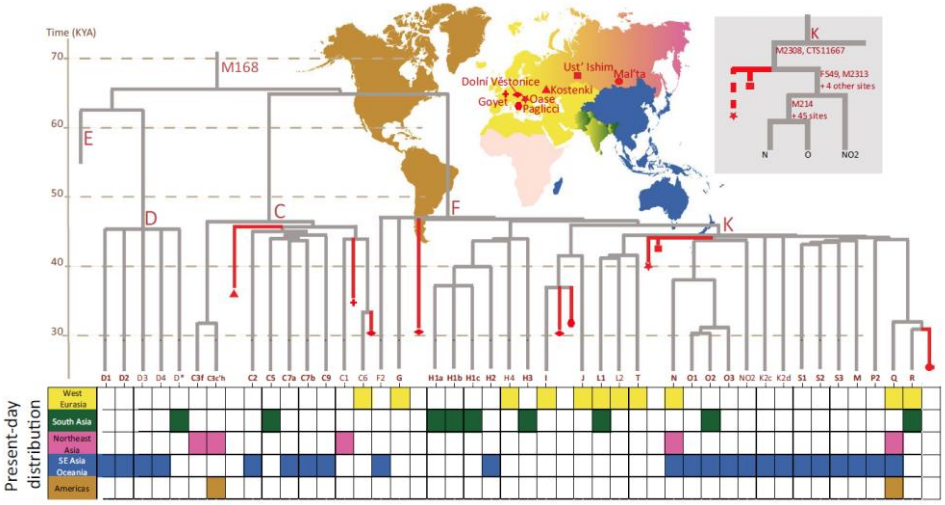


Figure 8. Human Y chromosome diversity outside Africa 20-50 thousand years ago. The branching structure of 41 extant Y-chromosome clades inferred to be older than 30,000 years is shown according to the tree based on high coverage Y chromosome sequence (from Kivisild, 2017).

The nine oldest European ancient Y chromosomes (Fu et al., 2014, 2015, 2016; Raghavan et al., 2014b; Seguin-Orlando et al., 2014) are represented by symbols in figure 8 (Kivisild, 2017) and they have been integrated in a tree with modern samples. All the ancient DNAs fell into one of the three major founding lineages inferred by the analysis of modern DNAs. The overlap between modern and ancient variation confirms an intense replacement about 20-50 kya of the previous male populations present in Europe right after the first "Out-of-Africa" (Lippold et al., 2014). The two oldest aY chromosomes, Ust'Ishim Man (Fu et al., 2014) and the Oase Man (Fu et al., 2015) belong to haplogroups K, the most frequent Y-chromosome lineage alive today. They both contain the M2308 (Poznik et al., 2016), which is representative of the basal root of two common haplogroups N and O. The haplogroups deriving from the M2308 are widespread in Eurasia today (Ilumae et al., 2016; Poznik et al., 2016). Although the Y-chromosome analysis of these ancient samples testifies a continuum between the Y-chromosome lineages of the past and present day populations of Eurasia, something different came out from autosomal analysis. Indeed, it was not observed a clear contribution of the genome of ancient samples in the modern populations living in the regions in which these ancient specimens have been discovered (Fu et al., 2015; Fu et al., 2016). Regarding the Y chromosome, the Kostenki-14 Man, a sample found in Southwestern Russia and dated 37 kya (Seguin-Orlando et al., 2014), belong to haplogroup C, which is virtually absent in most of the European modern populations (Rosser et al., 2000; Semino et al., 2000; Wells et al., 2001). In contrast with genome-wide results, which associate this sample to modern European populations, the phylogenesis of the Y chromosome revealed a link between this sample and present day Siberia, South East Asia and Oceania populations where haplogroup C is frequent (Bergstrom et al., 2016; Karafet et al., 2001, 2002; Kayser 2010). The haplogroup C affiliation of La

Brana (Olalde et al., 2014) and of three Neolithic farmer samples from Anatolia and Central Europe (Mathieson et al., 2015) shows that a diverse set of haplogroup C lineages may have been common and widely spread throughout Eurasia before Middle Holocene. Beside haplogroup C, hunter-gatherer and farmer populations of Europe and the Middle East were characterised by a diverse set of haplogroups, such as G, H, I, J and R, which are restricted in their present-day distribution by and large to Europe, the Middle East, North Africa, South and Central Asia. Ancient DNA evidence from Anatolia and Iran confirms haplgroups G and H as the most common Y-chromosome haplogroups of the early farmers in these areas. These haplogroups have been also found in European Early Neolithic populations who show low autosomal genetic distances with Anatolian farmers (Broushaki et al., 2016; Hofmanova et al., 2016; Lazaridis et al., 2016; Mathieson et al., 2015). The Y chromosome of Tyrolean Iceman, nicknamed Ötzi, associated with present-day populations of Sardinia and Corsica by the analysis of autosomal markers (Keller et al., 2012), is belonging to the haplogroup G2a-L166. This lineage descends from the G2a-L91, the most common haplogroup among Anatolian Farmers living 8 kya (Lazaridis et al., 2016). Today, 1 % of the living Sardinian Y chromosomes belong to haplogroup G and among this the 33 % of the sub-lineage G2a-L91 belong to G2a-L166 (Francalacci et al., 2013). For these reasons, Sardinians are considered the closest population to Early Farmers (Sikora et al., 2014; Skoglund et al., 2012). Haplogroup H is at present almost entirely restricted to South Asia, while one of its sub-clades, H2-M282 is an extremely rare lineage of some European populations. This lineage has also been found in the ancient Y-chromosome sequences of the Anatolian and Levantine farmers as well as in Iberian Chalcolithic samples (Gunther et al., 2015; Lazaridis et al., 2016). High frequencies of these sub-haplogroup G and H in modern populations can be found only in some geographically isolated areas such as the Caucasus, Sardinia, Corsica, whereas their frequency in main parts of Europe dropped later due to income of other Y-chromosome lineages.

Ancient DNA evidence suggested that also haplogroup I was common in Palaeolithic hunter-gatherers of Europe (Fu et al., 2016; Lazaridis et al., 2014). A complex scenario emerges from these studies in which both hunter-gatherers and farmers are carrying haplogroup I as well as haplogroup J lineages. Some minor sub-clades of haplogroup I have been found spread across a wide geographic area in Early and Middle Holocene samples, being found in Anatolian Farmers (Lazaridis et al., 2016) as well as in Scandinavian hunter gatherers from Motala (Mathieson et al., 2015). Similarly, to its present day peak frequency area, ancient Y-chromosome sequences falling into this clade characterize three Nordic Late Neolithic and Bronze Age samples (Allentoft et al., 2015). Haplogroup J, which, based on its present-day distribution follows a southeast-northwest decreasing cline, has been associated with the early spread of farming toward Europe (Rosser et al., 2000; Semino et al., 2000). It was found only in hunter-gatherers from geographically distant areas (Caucasus and Karelia), as well as in two early farmers from Iran and one from Anatolia (Jones et al., 2015; Lazaridis et al., 2016; Mathieson et al., 2015). J sub-lineages emerged in Central and Western Europe during the Bronze Age, likely being part of the demographic processes and population movements initiated from the North

Caucasus area during that period. It seems that, together with haplogroup J, some E sub-lineages have been introduced in the same period (Cinnioğlu et al., 2004; Cruciani et al., 2007; Trombetta et al., 2015). They have been associated with the Natufian culture by recent ancient DNA data (Gallego Llorente et al., 2015; Lazaridis et al., 2016). Haplogroup R is the most common haplogroup in Western Europe. Its oldest R1b-M343 lineage characterizes a 14 kya Villabruna Man from Italy (Fu et al., 2016), three European hunter-gatherers and three early farmer samples that did not belong to the R1b-M269 sub-clades. Hence, the R1b-M269, previously associated to the European hunters and gatherers (Cinnioğlu et al., 2004) has been found only after the Late Neolithic/Bronze Age (Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2015). Several sub-clades of haplogroup R have discovered in several ancient samples within a period that goes from the Late Neolithic until the Iron Age in Central Europe, Northern Caucasus and the Steppe belt of Russia (Allentoft et al., 2015; Broushaki et al., 2016; Cassidy et al., 2016; Haak et al., 2015; Mathieson et al., 2015; Schiffels et al., 2016).

American ancient samples helped to unravel the history of the two Y-chromosome Native American lineages, the haplogroups C and Q. These haplogroups, which are frequent in Central and North-East Asia (Zhong et al., 2011), have been involved in multiple migrations in the "New Continent" from Siberia, starting around 15 kya (Battaglia et al., 2013; Dulik et al., 2012a; Dulik et al., 2012b; Grugni et al., 2015; Lell et al., 2002; Zegura et al., 2004). Analysis on ancient DNA confirmed that a loss of rare Native American lineages in post-European contact is common and possibly it is part of the extensive lineage extinction process that has been observed in mitochondrial lineages (Llamas et al., 2016). Strangely enough, the analysis of a 24 ky old genome from Central-Southern Siberia, also known as Mal'ta boy, revealed a 'dual ancestry': one in Europe rather than East Asia and one in America. The most likely scenario able to explain this finding is that 24 kya a population like the one living in Siberia, mixed with the ancestors of East Asians at some point after the boy died. Although the result seems astonishing the European ancestry might be interpreted like a different source that now disappeared (Raghavan et al., 2014b). Also from Y-chromosome analysis, Mal'ta boy is affiliated to West Eurasian R haplogroup rather than to East Asian D, C or O haplogroups (Fig. 8). Probably it might represent a lineage at the basis of an extinct haplogroup R present right after the split of haplogroups Q and R.

Haplogroup Q has two ancient sub-clades, Q-M3 and Q-L54(xM3) (now identified as Q-Z780), which were likely born somewhere in Siberia before the first dispersal into Americas, and which together capture the overwhelming majority of extant Native American Y chromosomes today (Battaglia et al., 2013; Grugni et al., 2015, Jota et al., 2016; Zegura et al., 2004). The analysis of the 10.3 ky-old sample named On Your Knees Cave Man (OYKCM) confirmed the affiliation of its Y chromosome to the sublineage Q-M3 (Kemp et al., 2007). On the other hand, from the shotgun sequences of two other ancient genomes from the Americas, the Anzick Boy (Rasmussen et al., 2014) associated with Clovis Culture and dated 11 ky, and the Kennewick Man dated 9 ky (Rasmussen et al., 2015), turned out that their Y chromosmes belong to Q-M3 and Q-Z780, respectively. Thus, these results

confirmed that the two sub-lineages Q-M3 and Q-Z780 are in the Americas for at least 10.3 kya. A fourth ancient American sample, technically from Greenland, has been analysed; it is a 4 ky-old specimen referred as Saqqaq from the name of the site where it was discovered. The Y-chromosome lineage of Saqqaq is separated from the other sub-groups of haplogroup Q diverging from the others more than 25 kya. Therefore, it has been associated with a founding lineage, that from Asia moved through the Bering Strait toward Greenland, where it has been isolated, maybe in a population that lived closely related with the Paleo-Eskimos. So far, this lineage was not found in any samples from South America (Jota et al., 2016).
Further breakthroughs of ancient DNA success in regions like Africa, South-East Asia and Oceania will be most desirable to tackle broader range of questions about the continuity and nature of male-specific dispersals and admixture in human evolutionary history.

1.4 The future of uniparental markers

Genome-wide and whole genome sequences are answering most of the demographic history of population questions, thus, now the question is: is there still be enough space for more discoveries using uniparental markers? The answer is yes, uniparental markers are still playing a major role in evolutionary genetics studies.
Almost tens of thousands of mitogenomes and Y chromosomes covering the entire world have been sequenced and they are supposed to be public available for further analysis. MtDNA is widely used for the estimation level of contamination when aDNA is extracted. Indeed, due to the high number of copies per cell, it is very easily extracted also starting from ancient tissues and therefore it allows to obtain information also when the archeological specimens are too old and/or the biological material too damaged. Sex specific pattern of human migrations can be analyzed, only in part, by genome-wide data, therefore whole mtDNA and Y-chromosome data can go step by step with nuclear genetic *loci* and historical records. Their real utility comes from their uniparental modes of inheritance, which can provide insights into past social structure and the potentially different behaviours of men and women (Lippold et al., 2012; Tumongor et al., 2013); these areas are of considerable interest to historians, archaeologists and anthropologists. Due to prevailing patrilocality (almost 70% of the human societies), which means that after the marriage the family is going to be established near the man's birthplace rather than the woman one, the genetic differences among population are typically higher for Y chromosome than for mtDNA, although this effect has been mostly noticed at local rather than global scale (Wilder et al., 2004; Mark et al., 2012). The use of the full power of whole mtDNA together with Y-chromosome sequences is the right way to unravelling this kind of problem (Gunnarsdottir et al., 2011).
"Citizen science" in the Oxford English Dictionary is defined as: "scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions" (Oxford dictionaries 2016). Usually, these studies use genealogical internet forums to discuss about population genetics and results of academic paper. Often, many citizen scientists

analyse their genome using public company such as 23andme (www.23andme.com), ancestry (www.ancestry.co.uk) and Familytree (www.findmypast.com) and with them they make public their genealogical pedigree and the results of their analyses. An example of application of the citizen science to the study of uniparental markers, was provided by Balanovsky et al., (2017). This paper, in which I am a co-author, Y-chromosome sequences obtained by academic researchers and citizen scientists were combined providing a highly defined phylogeny of the Y-chromosome sub-haplogroup Q-L275.

1.5 The human nuclear genome

1.5.1 Overview

The length of a haploid nuclear genome is 3 billion of base pairs organized in 23 separate molecules, the chromosomes. In a human being, therefore there will be a total of 46 chromosomes organized in 23 pairs; each member of the pair is inherited from a parent. Twenty-two of them are called autosomes, while the remaining are known as the sex chromosomes due to their different presence in male and female.
Differently from the uniparental systems, autosomes undergo recombination during meiosis (Jeffreys et al., 2001). Recombination process can be detected through pedigree analysis, by investigating what fragments of the parent DNA have been passed to the offspring (Kong et al., 2002). The detection of the fragments can be directly shown by homologous sequences (alleles) or indirectly, through the manifestation of a phenotype.
Variants at *loci* of different chromosomes are inherited from parents in an independent way; in this case, they are defined unlinked. On the contrary, variants at *loci* close together on the same chromosome can be co-inherited from the same parent and that are known as linked. Recombination can interrupt this co-inheritance, and after the meiosis the child might have inherited a different association of variants at the two *loci*. The count of these recombination events allows to determine the genetic distance between the variants after having understood their order; this is genetic mapping (Strachan and Read 2010). One unit of recombination, also known as centimorgan (cM) is the genetic distance representative of 1% of recombination frequency between two variant *loci*. Technology and advanced sequencing produced a physical map of the human nuclear genomes, measured in megabases (Mb); comparison of the physical map and genetic map shows a broad correspondence of 1cM to 1Mb (Strachan and Read 2010). It is important to remember that recombination does not occur along all the genome in the same position and at the same frequency: there are regions, called hotspot, in which the frequency of recombination is higher, these segments are small (1-2 kb) and are separated by larger regions with a low recombination activity (Neumann and Jreffreys 2006). However, even in the so-called hotspot the recombination rates can vary a lot also in different population. This variation has to be taken into consideration when population genetic studies are carried on.

1.5.2 Next Generation Sequencing (NGS) methods

The advance in human genetic diversity studies has been improved by the determination of the human reference genome sequence, which was made possible by the automation of Sanger sequencing (Metzker 2010). However, this methodology, now referred as the "first generation DNA sequencing", presented some limitation being slow and very expensive. These disadvantages, led in the first 2000's to the development of a new set of different technologies characterized by reduced costs and increased sequencing throughput. These new methods, named Next Generation Sequencing (NGS) can be distinguished in second-generation and third-generation according to the technique used (Internation human genome sequencing consortium 2001).

The second-generation sequencing methods are nowadays at the basis of different genetic fields, allowing to shotgun sequencing million or billions of nucleotides in parallel with lower cost (Ajay et al., 2011). These methods usually need a library-preparation step in which DNA is fragmented (in 200 to 500 bp fragments), end-repaired, linked to adapters, bound to a solid surface (micro bead or solid body) and clonally amplified. Then these fragments can be sequenced in parallel through a number of NGS platforms using different sequencing technologies.

Then, bioinformatics analyses are applied to assemble these many 100 bp fragments with the reference genome.

Second generation sequencing methods sequence multiple times each genome base, providing high depth and, therefore, delivering accurate data. However, the library preparation can introduce biases in the recovery of sequence and cause particular problems for ancient DNA analysis (Hebsgaard et al., 2005).

Now, new methods have been developed which are able to produce sequence data from single original template DNA molecules and this are called third-generation sequencing methods (Schadt et al., 2010). Here there is no need of any library-preparation and amplification. The HeliScopeTM sequencer has a throughput of 37 Gb per run and mean read length of 32 bp, and uses much less DNA than established NGS platforms. The recently released Pacific Biosciences sequencer has lower throughput but the potential to achieve very long (mean of 3-kb) read lengths (albeit with >10% error rates) that could allow the direct determination of haplotypes. The company Oxford Nanopore promises hand-held devices for single-molecule sequencing. The area of sequencing technology continues to advance remarkably rapidly.

1.5.3 SNP-typing: methods for assessing variation

Before the advance of next-generation sequencing, a few million SNPs had been discovered, and a wide variety of methods were developed to type subsets of them in genomic DNA samples. Among these methods, whole-genome SNP chips can provide, at a reduced costs, the remarkable achievement of assaying million SNPs simultaneously in a DNA sample with >99% accuracy and reproducibility (Nielsen et al., 2011).

Some SNPs are preliminary selected by a company on the basis of previous studies, based on genome-whole or NGS sequencing (i.e.: HapMap, 1000 Genome). For examples, thanks to the HapMap project (International HapMap Consortium 2005) (see haplotype chapter for details), it had been possible to select some variants, named tag-snps, according to both their Minor Allele Frequencies (MAF) and their recombination association or Linkage Disequilibrium (LD). Considering these two features, almost 1 million SNPs, mirroring all the variation at all common SNPs in four populations (Japanese, Chinese, Yoruba and North Western European from USA), have been extracted (International SNP map Working group 2001). Now, the resolution of these chips increased a lot; there are microarrays based on the 1000 Genome project (The 1000 Genome Project Consortium 2010; 2015) which includes from 1 million to 5 million of SNPs with MAF > 1% in all the populations present in the study. Not only autosome markers are present on these chips, but also Y-chromosome and mtDNA markers can be found. The great advantage of these arrays is the certainty into calling the markers and the variants. Whereas not all these chips have been designed for population genetic studies, they have some drawbacks that have to be considered (Alkan et al., 2011). One of these disadvantages is that the set of SNPs would be bias towards the population or populations from which the markers were selected; therefore, when referring to a particular population, part of the genetic variation might be lost. For this reason, Affymetrix Axiom Human Origin Array was designed (Ha et al., 2014); it takes into consideration 630,000 SNPs extracted from 13 population-specific panels, analyzed in the Human Genome Diversity Project (HGDP) as well as in the Neanderthal Genome Project (Green et al., 2010; 2015). The array also contains 87,000 SNPs that include sets of markers from mtDNA and the Y chromosome, and variants from standard chips for data comparison purposes. Genotypes of individuals obtained from array analysis come out as a .txt file format; usually they are converted in other formats in which allele frequencies of the individuals are displayed (see the 4.3.3 section for details).

1.5.4 Haplotype

The combination of allelic states (including SNPs and CNVs) found on the same chromosome or on mtDNA is known as haplotype. The list of variants present in uniparental systems or on the X chromosome of a male, represent a haplotype since they do not undergo recombination. The haplotypes extracted from autosomal polymorphisms and from X-chromosomal polymorphisms in females are not linear and straightforward unless the two homologues chromosomes do not share the same haplotype (International HapMap Consortium, 2005) (same allele state at all the considered *loci*) being homozygous, hence two haplotypes are obtained in a normal case of heterozygote (Fig. 9).

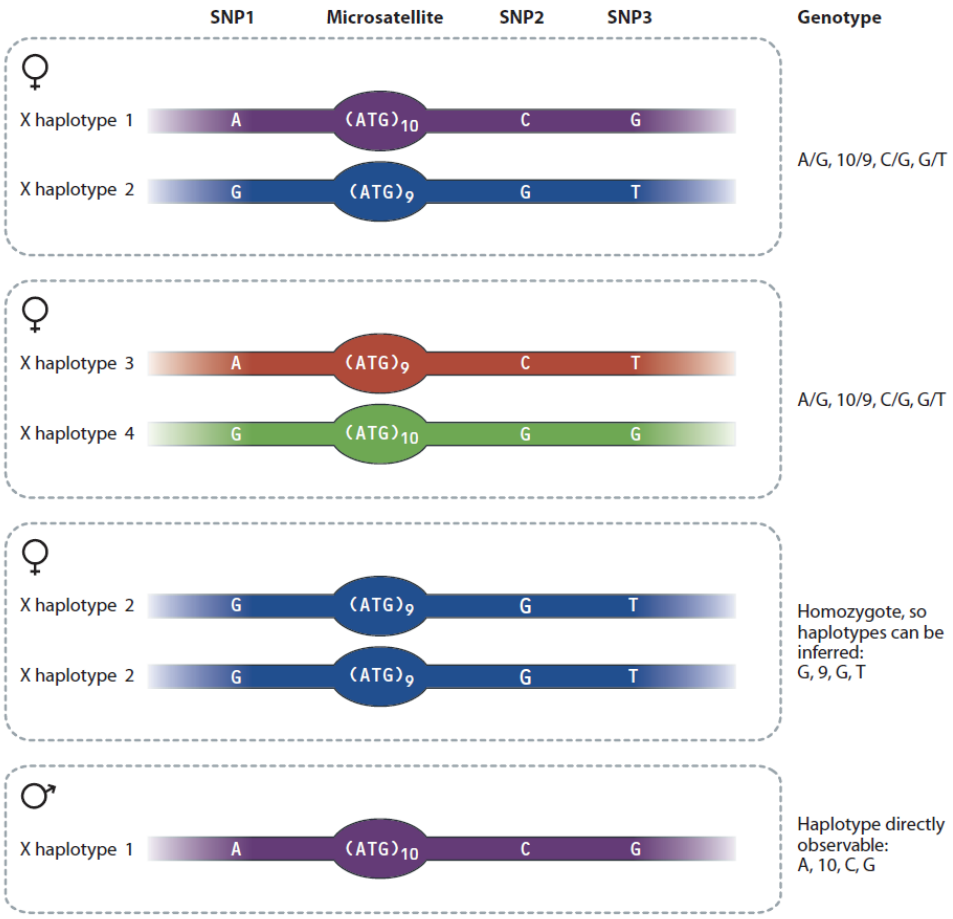


Figure 9. Representation of the genotypes and haplotypes of the X chromosomes in female and male. The first two females carry the same genotype but different haplotypes. The third example is a homozygous female from which it is easy deduct the haplotype. The final example is a male that being hemizygous, its haplotype corresponds to his genotype (from Jobling and Tyler-Smith 2014).

Recently, the determination and the deduction of the haplotypes starting from genotype frequencies (Browning and Browning, 2007; 2011) is widely used in population genetics, the technique, which is supposed to balance and minimise biases at SNP level, will be explained in the 4.4.1 section.

In the uniparental system, in which no recombination occurs, the haplotype will change only by the accumulation of mutations. Thus, from the comparison of several sequences, the ancestral haplotype will be the one carrying at each position the most represented allele while the derived haplotypes will differ from the ancestral one for the allelic state of at least one position. Instead, considering the remaining part of the genome, which is affected by recombination, haplotypes can change from generation to generation since the fragments can be broken up and lead to new haplotypes (Clark 1990). Recombination events can build new allele combinations from those of the

parents and these new haplotypes can be transmitted to the successive generations. These blocks of DNA have been considered genetic markers and can be used as predictor of the genetic history of populations (Lawson et al., 2012; Hellenthal et al., 2015).
In the human genome, recombination events tend to stick together closer *loci*, while separate the distant ones; this principle is not always applicable, indeed there are some regions of the genome that are more prone to recombine than others. The tendency of particular alleles at neighbouring *loci* to be co-inherited because of reduced recombination between them can lead to associations (correlations) between alleles in a population (Reich et al., 2001). This property is known as linkage disequilibrium (LD). For example, considering a case of random segregation, if two alleles have been found together many times, they are considered in linkage. A revolutionary study with the aim of dissecting LD patterns in genome-wide data of different populations started in the 2002 with the name of HapMap project (International HapMap Consortium, 2005). This project genotyped, with high-throughput genotyping methods, 269 samples with different ancestry. The study was concluded in the 2007; it was performed in three phases that generated haplotypes based on over 3.1 million of SNPs. Haplotypes were extracted from parent-child trios genotypes by using statistical methods. HapMap results can be summarized in few points: i) the identification of ~ 33,000 recombination spots, which are not distributed uniformly along the genome but they are organized in a block structure; ii) these haplotype blocks have a high frequency of alleles correlated to each other and iii) they show little differences in individuals with closer ancestry, mirroring few historical recombination events; iv) interruptions between hotspots are net, displaying the highest likelihood of recombination in these regions; v) sixty per cent of the events of recombination occur in these segments, even if their length accounts of only 6% of the genome. LD and haplotype composition is widely used in population genetic studies after conversion of genome-wide data in haplotypes.

1.5.5 Phylogeny

Phylogeny inferred from genome data is not too different than using uniparental markers. The tree is an intuitive method to represent the genetic diversity within populations. Phylogenetic trees can be used to represent sequence differences among species (i.e.: Human and Chimpanzee) as in the evolutionary genetics (Janečka et al., 2007); moreover, they can be employed as graphical tool of distance that measures the relationships among individuals of the same species. In this last example, it is necessary to use a model that will explain how this diversity arose, although it does not explain the mechanism behind it (Jobling et al. 2014). With genomic data, the interpretation of a tree has to be careful; this kind of tree separates individuals in groups, and fixes splits from a common ancestor although it does not show gene flows among groups. That is why trees constructed with genomic data should not be used as tools to reconstruct a common ancestor, but rather to define groups or “clusters” of samples (Lawson et al., 2012).

Only recently, a method called pairwise sequential Markovian coalescent (PSMC) (Li and Durbin 2011) tried to associate pattern of recombination along a genome to the population size history. It is utilized to reconstruct population size changes whereas multiple sequentially Markovian coalescent (MSMC) is employed to study the time course of population separation and is a good example of application of the phylogeny to genome data (Mallick et al., 2007).

1.5.6 Genome wide and genome whole studies

The purpose of this chapter is to summarize all the studies that used genome-wide or genome-whole data with the aim of investigating the history of the worldwide populations. Most of the data of these studies have been also assembled in the final dataset of this work of thesis (see dataset section 5.2.2).

The **Human Genome Diversity Project (HGDP)** is a project started in 2005 by the Morrison Institute at the Stanford University; it had the aim to create a collection of relevant and genetically diverse samples. Luigi Cavalli-Sforza was the mind behind this project that sequenced more than 1,000 individuals from 51 different populations (Cavalli-Sforza 2005). About 650,000 SNPs have been tested with the Illumina BeadChip technology in different facilities and with the collaboration of many scientists. Only in the 2008, Li et al., (2008) published a study aimed to characterize the most complete catalogue of the human genetic variation using the HGDP samples. The 51 populations considered were representative of sub-Saharan Africa, North Africa, Europe, the Middle East, South/Central Asia, East Asia, Oceania, and the Americas. Using ADMIXTURE analysis (Alexander et al., 2009) they interpreted mixed ancestry as a result of recent admixture between two founder populations (Fig. 10A).

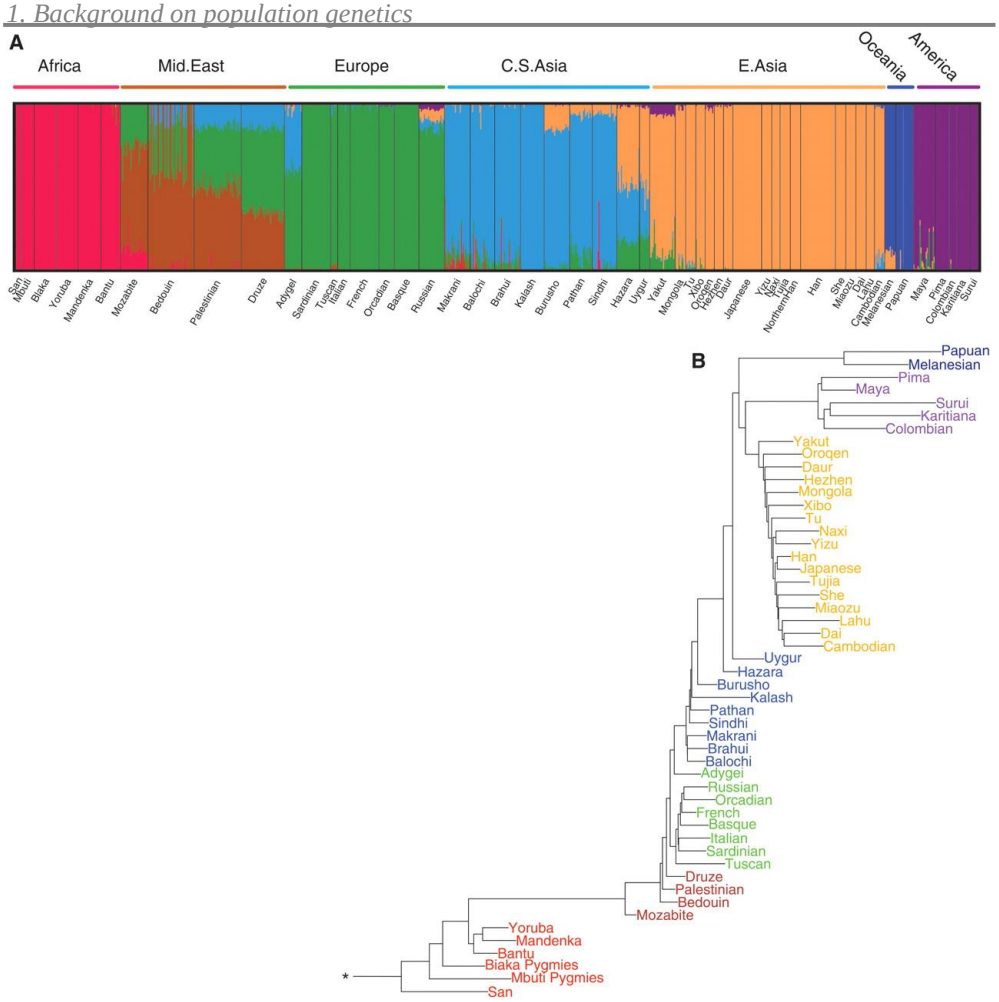


Figure 10. Ancestry and distance measures using genome-wide data. ADMIXTURE plot (A) and tree extracted from Fst analysis (B) from Li et al. 2008.

The Li et al. (2008) results pointed out that the Central Asian populations of Uygur and Hazara were the results of recent contact between Asia and Europe; differently, the Yakut and Russian ancestries discovered in the Native Americans might reflect an ancient contribution of the population which crossed the Bering Strait in the first peopling of the Americas. They also drawn a phylogenetic tree based on Fst genetic distance, using chimp as out-group (indicated with a star in Fig. 10B). The tips of the tree represent the different ancestries distributed per continent: the sub-Saharan samples are the ones closest to the root while the Native Americans with the Oceanian samples are the most distant. This study was one of the first examples of population genetics using genome-wide data on a large dataset with a large amount of SNPs.

The second example to be mentioned is the **1000 Genome Project**; it ran between 2008 and 2015. Its dataset, which was completed in three phases, became the largest

catalogue of human variation and genotype data. The aim of the project was to extract all the variants with frequency of at least 1 % in the population studied and to make it available to the scientific community. The final phase of the project counted 2,504 samples of 26 populations in which are present genome-wide as well as genome-whole data at low coverage (The 1000 Genomes Project Consortium 2015). In one of the last studies conducted in the 2015 the authors remarked the broad spectrum of variants, improving the catalogue published in 2010 (The 1000 Genome Project Consortium 2010). Moreover, they used a maximum likelihood approach (ADMIXTURE analysis cited above and in the Materials and Methods section) to infer the likely ancestries for each population examined (Fig. 11).

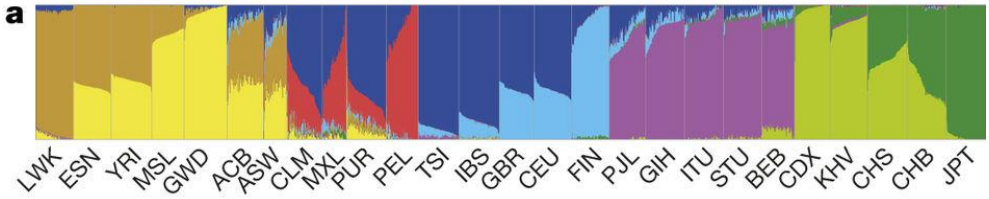


Figure 11. Admixture plot from 1000 Genome Project Consortium 2015. Eight ancestries (indicated by different colours) are displayed. The populations are as followed (LWK = Luhya in Webuye, Kenya; ESN = Esan in Nigeria; YRI = Yoruba in Ibadan, Nigeria; MSL = Mende in Sierra Leone; GWD = Gambian in Western Divisions in the Gambia; ACB = African Caribbeans in Barbados; ASW = Americans of African Ancestry in SW USA; CLM = Colombians from Medellin, Colombia; MXL = Mexican Ancestry from Los Angeles USA;PUR= Puerto Ricans from Puerto Rico; PEL = Peruvians from Lima, Peru; TSI = Toscani in Italia; IBS = Iberain Population in Spain; GBR = British in England and Scotland; CEU = Utah Residents (CEPH) with Northern and Western European Ancestry; FIN = Finnish in Finland; PJL = Punjabi from Lahore, Pakistan; GIH = Gujarati Indian from Houston, Texas; ITU = Indian Telugu from the UK; STU = Sri Lankan Tamil from the UK; BEB = Bengali from Bangladesh; CDX = Chinese Dai in Xishuangbanna, China; KHV = Kinh in Ho Chi Minh City, Vietnam; CHS = Southern Han Chinese; CHB = Han Chinese in Bejing, China; JPT = Japanese in Tokyo, Japan)

This approach clustered different continental groups, showing their internal sub-structure; from these results it is possible notice an east-west cline in Africa and East Asia, while a north-south cline is remarkable in Europe. American samples show some African and European ancestries. Using a statistical method applied to the genome-wide data the authors were also able to reconstruct high quality haplotypes with the help of the first-degree relatives whole genome data. The 1000 genomes dataset is widely used in almost all the recent population genomic projects as well as it is used as reference for haplotype-based studies.

Very recently, the **Simon Genome Diversity Project** (**SGDP**) made its entrance. It included 300 high quality whole genome sequences from 142 populations (some of them belonging to the 1000 Genome Project), provided at least 5.8 million bp not present in the human reference sequence and aimed to reveal key features of human genome variation landscape. The publication linked to this project (Mallick et al., 2016) assessed the structure of human genetic diversity and tried to date the time course of human population separation. More in details, the authors by analyzing the

Fst distribution, confirm that the highest diversity within human populations is found in sub-Saharan populations (Fig. 12).

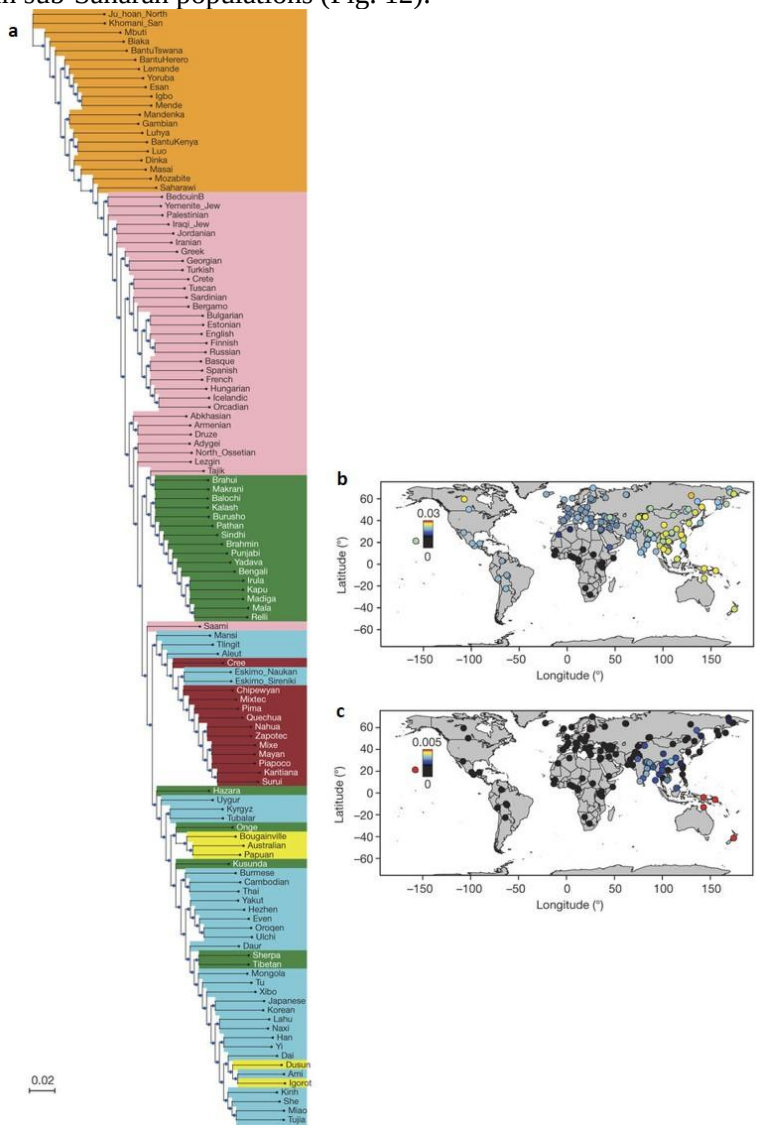


Figure 12. New approaches to measure genetic relationship and contributions. (a) Neighbour-joining tree based on pairwise divergence. (b) Neanderthal ancestry contribution. (c) Denisovan ancestry contribution. (Modified image from Mallick et al., 2016).

Moreover, statistical analyses confirmed that all non-African populations have Neanderthal ancestry with frequencies higher than previously thought in East Asia. Denisovan ancestry was demonstrated to be higher in South Asian than in Eurasian (Fig. 12). The results allowed to date back the appearance of the genetic substructure in the African continent, as find nowadays, to 200 kya. Considering all the other lineages, but Africans, they inferred a coalescent time of 50 kya, result in

concordance with the latest archaeological discoveries of the "out of Africa". This broad description about the projects conducted in the last years is just an example of how much interest there is behind the genetic history of human worldwide populations. Now that DNA can be extracted from fossil specimens (ancient bones, teeth, hair -Tobler et al., 2017-, eggshells, paleofeces and even soil -Slon et al., 2017-), analyzed with high-throughput sequencing and assembled with computational tools, ancient DNA (aDNA) is becoming a kind of molecular fossil (Orlando et al., 2016). Since it has been preserved for a long time, aDNA can shed new light on the evolution of genomes, epigenomes and, in general, on the Earth history. Ancient human DNA from between 7 and 45 kya has helped researchers to discover striking aspects of European population history (Fu et al., 2016). Thus, it would be extremely exciting assembling a database that encloses all the already published and the new aDNA from the whole globe.

2. Peopling of the world

2.1 The origin of the anatomically modern man

The origin of modern human was believed to happen sometime around 200 kya, either in a region of East Africa, as evidences from mtDNA data (Metspalu et al., 2004; Torroni et al., 2006) seem suggest, or in South Africa (Gronau et al., 2011; Pickrell et al., 2012; Schlebush et al., 2012; Veeramah et al., 2012) like some recent evidences from nuclear genome point out. Despite these clues, the origin of the AMH seems to be more complex and geographically diffuse than from a single area (Harding and McVean, 2004).

Few months ago a research (Schlebush et al., 2017) leaded by Matthias Jakobsson at Uppsala University in Sweden have just shaken population genetic scientific community. The study estimates a coalescence time of 260 kya between six ancient African genomes and a juvenile boy from Ballito Bay, who lived ~2,000 years ago. This discovery anticipates the previously dating of deep history of diversity among African populations of 60 kya (Schlebusch et al., 2012). This means that archaic humans with features similar to the ones of modern humans inhabited Africa around 300 kya. The theory fit precisely with the discovery of a new fossil specimen emerged from a recent excavation in Morocco carried out by Jean-Jacques Hublin and colleagues (Hublin et al., 2017): they attributed the found bone to the earliest-known AMH (Hublin et al., 2017, Richter et al., 2017), but unfortunately, until now no usable DNA has been recovered.

During times, several models have been proposed to explain the origin of AMH. The two most scientifically approved and debated are the "Multiregional" and the "Out of Africa" models (Fig. 13).

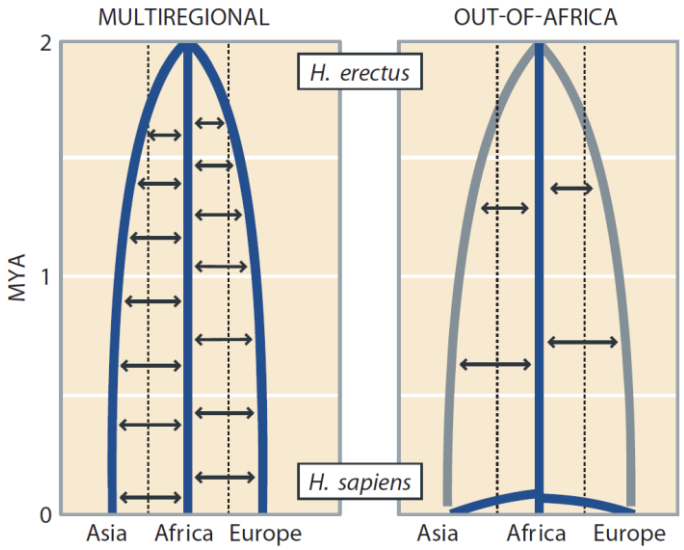


Figure 13. The two main models proposed for explaining the origin of AMH (from Jobling et al., 2014).

According to the **Multiregional model**, proposed by the anthropologist Weidenreich, the transition from *H. erectus* to AMH took place in several areas of the Old World, with many modern human characteristics arising at different times in different places (Weidenreich, 1940). Usually with the term "Multiregional" is also defined the Candelabra model, proposed by Coon in the 1962 and analysed by Templeton et al., 2007. The Candelabra model assumes an independent origin of AMH from separated groups of *H. erectus* that went out of Africa around one million of years ago. This assumption is in contrast with the "Multiregional" model that does not exclude a gene flow and a common evolution between the groups of *H. erectus* maybe driven by an interaction. Fossils like the Dali Man in China (Wu, 1981), that present anatomical characteristics in a midway between archaic and modern human, are the evidences supporting the Multiregional and Candelabra theory. That said, these fossils are poorly preserved and some authors have suggested that these characteristics are in fact shared by other Homo worldwide, and thus were not unique to Asia (Stringer and Andrews, 1988).

In contraposition with the Multiregional model, the "Out of Africa" model proposes a single and relatively recent transition from archaic hominins to AMH in Africa, followed by a later migration to the rest of the world replacing other extant hominin populations (Cavalli-Sforza and Feldman, 2003; Ingman et al., 2000; Relethford, 2008; Stringer and Andrews, 1988; Stringer, 2002; Tattersall, 2009).

The model implies a likely "bottleneck" or "founder effect" occurred sometimes in the lasts 50-200 kya in which only a subgroup of the population inhabiting the African continent contributed to the total of genetic variability present in the current non-African AMH. This is highlighted by the greatest amount of genetic diversity found in African population than in Non-African populations. Analysis of mtDNA phylogenetic tree massively contributed as first genetic evidences in the sustaining

of the "Out of Africa" model (Cann et al., 1987). These studies place the African mtDNAs closer to the root of the tree (Cann et al., 1987; Horai et al., 1995; Ingman et al., 2000; Vigilant et al., 1991); while all the other populations scatter from African individuals and belong to the mtDNA haplogroups M and N. It has been demonstrated that these haplogroups originated ~ 45-40 kya in South Asia from L3 haplogroup arisen in East Africa around 80 Kya (Metspalu et al., 2004). This hypothesis was supported by further studies on mtDNA (Caramelli et al., 2003; Relethford, 2001), Y chromosome (Hawks, 2001; Thomson et al., 2000; Underhill et al., 2000) and autosomal regions (Alonso and Armour, 2001; Rosenberg et al., 2002; Takahata et al., 2001; Tishkoff et al., 1996; Zhivotovsky et al., 2003). More recently, multi-*locus* studies of genome-wide data have demonstrated that the genetic distance between pairs of world populations (where one is African) increases proportionally moving away from Africa (isolation by distance or IBD); while linkage disequilibrium (LD), expressing how much genomes inside a particular population have shuffled their segments during the formation of gametes, is low in African populations. African genomes are characterized by the least linkage disequilibrium (smallest segments), in comparison with all the other worldwide populations (Prugnolle et al., 2005; Ramachandran et al., 2005). The impact of Neanderthals and Denisovans is overall rather small in modern subjects (1-2%) (Green et al., 2010; Meyer et al., 2012; Reich et al., 2010), but certain archaic *Homo* genes (e.g. immune genes) that are not present on African individuals, played a major role on the fitness of the expanding populations (Abi-Rached et al., 2011; Deschamps et al., 2016; Mendez et al., 2012).

2.2 Dispersal routes from Africa

Based on uniparental marker data, two principal routes of dispersal of modern humans out of Africa have been proposed: the northern and the southern costal route (Fig. 14).

Figure 14. The two routes of dispersal of modern humans out of Africa: the "Levantine corridor" is indicated by the dashed line while the "Coastal route" by the continuous line (Forster and Matsumura, 2005).

The first route, also referred as "the Levantine corridor" suggests an initial spread through the Nile Valley located in the Northern part of Africa and then eastward into Asia (Stringer and Andrew 1988; Lahr and Foley, 1998). Some fossils dated back 120 kya and belonging to hominins present in this area have been attributed to an early exit (Oppenheimer, 2012). This exit, however, was not considered successful (Armitage et al., 2011; Balter, 2011; Cruciani et al., 2011) since immediately before and during the last interglacial this region was hyper arid (Petraglia and Rose, 2009). The recent discovery of inhabited area in the Levant around 55 kya (Hershkovitz et al., 2015) led, however, to a resurgent of the Levantine corridor, with the addition of considerable break in the occupation of this region. In addition, a study on genome-wide data of Egyptian and Ethiopian populations (Pagani et al., 2015), revealed a higher similarity between Egyptians and non-Africans than between Egyptians and Ethiopians, suggesting a possible route out from Africa passing through Egypt.
The second route, also called "Southern or Costal route" sustains an early migration, before 70 kya, from the Horn of Africa into two directions: one through Arabia northward and the other eastward along coastlines of Arabia and India, arriving in Southeast Asia and Australia (Armitage et al., 2011; Boivin et al., 2013). Studies on mtDNA supported this theory suggesting that individuals assigned to haplogroup L3 migrated out of the continent via the Horn of Africa (Quintana-Murci et al., 1999; Macaulay et al., 2005; Soares et al., 2011; Torroni et al., 2006). These conclusions are largely debated, especially in the last years.
Another key debate has focused on the precise timing of the exit of the first humans out of Africa. Currently, two different proposals are reported in literature, differing each other for several tens of thousands of years and not mutually exclusive. The first claims that the out of Africa dispersal took place around 50-60 kya, reaching Australian continent by 45-50 kya (Mellars et al., 2013). The second suggests a much earlier exodus around 100-130 kya, prior to the eruption of Mount Toba (Northern Sumatra) dated 74 kya (Petraglia and Dennell, 2007). Genetic studies have been unable so far to settle the conflicting archaeological evidences for these different dates. The lack of aDNA data temporally and spatially in that region limited the genetic studies only to modern DNA. Studies based on mtDNA phylogenies have suggested a date for the African exit of modern humans between 40-60 kya (Underhill and Kivisild, 2007), STR analysis confirmed the previous results with an exit dated back around 50 kya (Shi et al., 2011; Zhivotovsky et al., 2004). Models applied to whole genome sequencing data have given different results, according to the choice of the model. For example, studies using the allele frequency spectrum (AFS), identity-by-state (IBS) or coalescent-based models suggested a divergence time of 50-60 kya (Gravel et al., 2011; Gronau et al., 2011; Harris and Nielsen, 2013). Analyses based on the pairwise sequencing Markovian coalescent model (PSMC) suggested a divergence time of 80-100 kya, with gene flow occurring until 20 kya (Li and Durbin, 2011).
Lately, two studies performed on high-coverage whole genome sequences from different worldwide populations provided a high-resolution portrait of human genetic diversity. They identified a genetic signature in the genomes of present-day Papuans that suggests human presence outside Africa before the main out of Africa

split time that involved other Eurasians (~75 kya) in line with a multiple dispersal model (Pagani et al., 2016).

2.3 Admixture with Neanderthal

In the 2010 Svante Pääbo and colleagues published a study in which they identified a Neanderthal contribution to the AMH genome (Green et al., 2010). Despite the great finds, some doubts remained about the quality of the single DNA sample used. Only in 2014 an amelioration of the techniques allowed Pääbo and colleagues to produce the first high coverage genome of this hominid Prüfer et al., (2014). This sequence opened a world of speculations and questions most of which have been answered by computational biologists. For example, many segments of recent Neanderthal genomes have been found in modern humans. A 40 kya Romanian skeleton resulted to be the product of a genetic admixture with a great-great-great-great-grandparent of a Neanderthal (Fu et al., 2015).
The breeding between Neanderthal and AMH has been dated back 50 kya even if there are evidences that the ancestors of Neanderthals from the Altai Mountains encountered early modern humans at last 100 kya, possibly in the Near East (Kuhlwilm et al., 2016).
The extinction of Neanderthal occurred around 40 kya, even if there are some proofs about the presence of this hominid in Asia about 24 kya. The decline of the population of Neanderthal coincides with the migration of AMH toward Europe and Asia (Higham et al., 2014). Many hypotheses have been advanced about the causes of their extinction, and among the others we can recall: i) the competition with the "new comers" for resources, ii) the introduction of some pathology for which the Neanderthal immune system was not prepared and iii) the high rate of genetic load caused by the high level of inbreeding due to the Neanderthal small groups. Recent analyses defined the structure of Neanderthal population formed by many groups characterized by a small population size and a high level of relatedness (Rogers et al., 2017).

2.4 Peopling of Europe

Three or more genetic components characterize all European populations mirroring the different events of colonization of the continent (Fig. 15) (Allentoft et al., 2015; Gamba et al., 2014; Günther et al., 2015; Haak et al., 2015; Lazaridis et al., 2014; Skoglund et al., 2012).

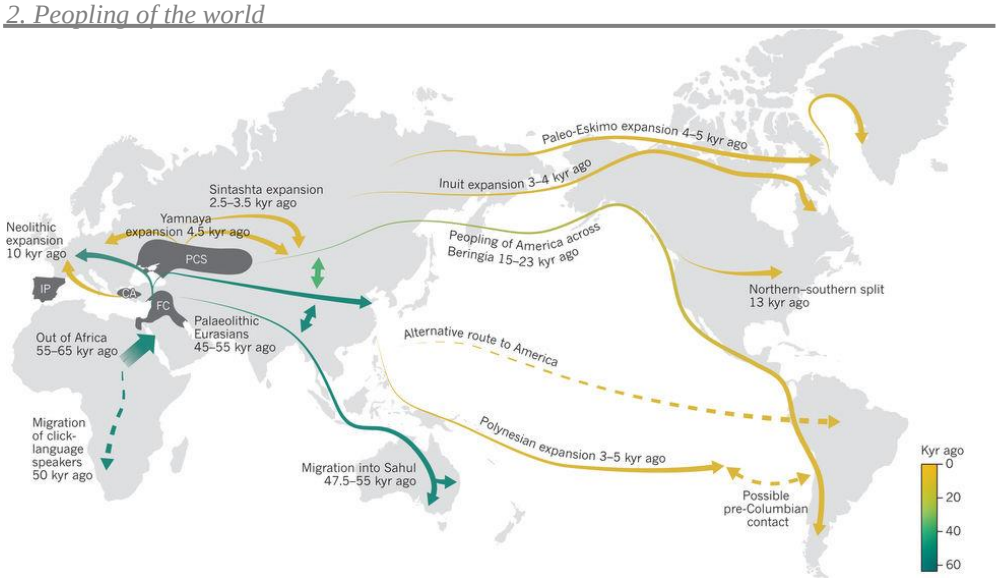


Figure 15. Major human migrations across the world inferred through analyses of genome-wide genome whole data (from Nielsen et al., 2017).

The first genetic component can be linked to the first peopling of Europe, likely occurred around 45-40 kya, from Asia or Far East Europe. Evidences of settlements in the Balkan area, as well as in Central West Europe dated around 40 kya, are clear examples of this rapid diffusion (Martini, 2008). These evidences show the presence of the first AMH in Europe as early as 43 kyr ago (Benazzi et al., 2011; Higham et al., 2014).

Palaeolithic Europeans, sharing the territory with other hominids as Neanderthal, acquired some signs of hybridization not only in their culture but also in their genome. Indeed, as previously reported about 1-2 % of our genome derives from the one of Neanderthal. Attributed either to climate oscillations or to a strong competition with Neanderthal, these early Palaeolithic Europeans have left a little genetic contribution to the European people of today (Günther and Jakobsson 2016).

The lifestyle of the first groups of AMH varied according to the environment. In Northern Europe this groups of hunter-gatherers survived hunting macro fauna, while in the Mediterranean coast they preferred fishing besides hunting. Movements of kilometres just because of the migratory routes of the animals have been documented by the presence of different stone tools.

After the LGM, a period of glaciation occurred 15-30 kya, followed a phase characterized by good climate, which contributed to form steppe landscape widespread in all the European continent. The oscillation of the temperatures created several microclimates in different regions with an increase of humidity. Frozen mountains became an obstacle for the communities of Central and Eastern Europe creating several regional or sub-regional cultural adaptations. An example is represented by the Alps; they formed a barrier between Italian and transalpine populations causing difficulties in the interaction and therefore determined a cultural differentiation. Pyrenees played a similar role to the Alps in the Iberian Peninsula.

Eastern steppe together with Western Atlantic coast represented the scenario in which intense migratory routes were followed by humans and animals. Several cultures developed in this period; among the others, the Epigravettian was widespread in Eastern Iberia, Southern France, Italy, Greece and Balkans. These areas were used as refuges during the LGM.
The second genetic component can be traced back to the expansion of the first inhabited centres after the LGM occurred around 11 kyr ago, when the wild landscape disappeared. A new way of life based on animal husbandry, agriculture and sedentary, known as Neolithic lifestyle, started to emerge in several sub-regions of the Fertile Crescent (Asouti et al., 2013). Ancient DNA showed that these first farmers expanded from Central Anatolia into Europe (Günther and Jakobsson 2016). They reached the Iberian Peninsula roughly 7 kya and arrived in Britain and Scandinavia ~ 6 kya (Günther and Jakobsson 2016). Ancient Neolithic DNAs have also demonstrated that this process was characterized by a massive migration of groups of farmers with a partial admixture with local hunter-gatherers carrying the first genetic component (Günther et al., 2015). This discovery once for all demonstrated that the Neolithisation (the process in which the Neolithic lifestyle spread across Europe) was a migration of people rather than solely as an idea or a culture. Neolithic increased the size of populations, as seen in the estimates of effective population sizes that were generated from genomic data (Jones et al., 2015). On the other hand, archaeological data suggests that the health of the individuals who lived as farmers was sometimes poor as there were ample signs of malnutrition and caries (Cohen et al., 2013; Skoglund et al., 2014).
After the Hunter-Gatherers and the Neolithic components, the third component arrived during the early Bronze Age. A strong wave of migration leaded by herders from the Pontic Caspian steppe belonging to the Yamnaya culture invested the Central Europe about 4.5 kya (Allentoft et al., 2015; Haak et al., 2015).
The genetic ancestry of these populations derived from various hunter-gatherers groups from (modern) Russia (Haak et al., 2015) and the Caucasus (Jones et al., 2015). This migration was likely facilitated by the introduction of new technology as horseback riding and the wheel, and may have spread Indo-European languages to Europe. Even if some linguist suggests that Neolithic farmers (Bouckaert et al., 2012) already spoke these languages, for sure this migration led to the spread of Bronze Age population genes in Western and northern Europe.
In conclusion the genetic composition of modern Europeans seems to be summarized from these three components, the Palaeolithic hunter-gathers component coming from refuges following the LGM; a Neolithic farmer component arrived with the introduction of agriculture from the Middle East following a Southern route and a Bronze Age migration to Europe from West Asia (Günther et al., 2014). The dissection of these components can reveal some crucial details on the migration routes or on the possible gene flows that shaped the present European genetic landscape. For example, the Neolithic genetic component seems to be most frequent in southern European populations such as the Sardinian people suggesting, a southern migration route followed by the early farmers (Lazaridis et al., 2014; Skoglund et al., 2012; Skoglund et al., 2014). The genetic variation observed in

Europe reveals a strong correlation with the latitude: Northern populations appear more genetically similar than Southern people (Auton et al., 2009). Thus, the actual European genetic gene pool is the result of multiple migration waves occurred during the centuries and followed by admixture events in a regional scale.

2.5 Peopling of America

2.5.1 First peopling

Americas was the last continent to be colonized. The first peopling occurred around 16-15 kya from Asia through the land bridge connecting northern Asia and Alaska. The existence of this land bridge was first proposed in 1590 by the Spanish missionary Fray Jos. de Acosta (Wilmsen, 1965) and subsequently confirm by different studies on the flora and fauna in both the Eurasian and American Arctic regions (Dall and Harris, 1892; Wallace, 1876; Heilprin, 1887). Evidences of widespread settlements were present only after 13 kya with the appearance of the Clovis culture, the first well-characterized prehistoric Native American culture (Jenkins et al., 2012). Before 13 kya two large glaciers (the Cordilleran in the west and the Laurentide in the east) covered all of Canada reaching the Missouri and Ohio Rivers, and eastward to New York City in what would be today's the United States. This barrier made it difficult the dispersal of people from Beringia (now northeastern Siberia and northwestern North America) to the southeastern parts of the Americas. The end of LGM and the improvement of climate conditions represented a critical point for the movement of humans toward North-West Asia and into Beringia. After the ice melted, the sea level increased, and America became an isolate until the arrival of Europeans (Manley, 2002). On the other hand, after 13 kya the melting of the ice opened a roughly 1,500 km interior ice-free corridor along North America, which could represent a possible route of migration towards southeast. Thus, the most plausible scenario about the American peopling considers two or more routes of dispersal: the costal route along the west coast of the continent, would be used by an early and rapid southward migration occurred before 16 kya; an internal route walkable only after 13 kya. These early Paleo-Americans soon spread throughout the Americas, diversifying into many hundreds of culturally distinct nations and tribes. The time frame and exact routes are still matters of debate, and the model faces continuous challenges (Fig. 15). This theory is also supported by some informative archaeological sites in South America, such as Pikimachay in Peru (dated ~ 12-14 kya) and Monte Verde in Chile (dated ~ 14-20 kya) (Dillehay et al., 2008, 2015; Gilbert et al., 2008). Following these routes, some dates of the earliest archaeological sites in the North America suggest an isolation of a sub-group of the first Native American ancestors in Siberia and Beringia until around 4 - 5 kya when their moved eastwards, colonizing the current North America and Canada.

The comparison of Native American languages led Greenberg to group the Native Americans into three groups: Eskimo-Aleut in the Artic, Na-Dene in Canada and south-western US and Amerinds, which groups all the other Native Americans in much of North America and all South America (Grennberg, 1987). Analysis on

genome whole sequences suggested a unique event of migration from Siberia after the LGM followed by a gene flow from East Asia into both Amerindian and Na-Dene populations (Raghavan et al., 2015). Therefore, whether the division between the Native American branches took place either in Siberia or in the north or south of the American ice sheets is still under debate, and the analysis of further ancient genomes will be necessary to resolve this matter. Raghavan et al., (2015), from the analysis of 31 present-day whole genome sequences from the Americas, Siberia, and Oceania and 23 ancient samples from the Americas, were able to infer a divergence time of 23 kya between Siberians and Native Americans. The genome of a 24 ky-old Mal'ta Siberian skeleton (Raghavan et al., 2014), suggests that Native Americans are the result of an admixture between an ancient population related to Mal'ta, West Eurasian as well as one or more East-Asian lineages. Because of the non-clear affinities between Mal'lta and East Asian populations the most likely scenario is an admixture event between the East Asian population and Mal'ta related population happened more than 12.6 kya but before 24 kya either in or outside America.
Different origin seems to be attribute at the Inuit of the American Arctic populations (Gilbert et al., 2008; Reich et al., 2012); long has been debated about a possible origin of Inuit attributed to the now extinct Paleo-Eskimo culture, which appeared about 5 kyr ago. Raghavan et al., (2014) compared the genomes of the two Artic populations (Rasmussen et al., 2010) and have been able to confirm that the Paleo-Eskimo populations had migrated from Siberia to the North American Arctic independently from the Inuit populations.

2.5.2 Late peopling of Americas

At the time of America discovery, the population size was much lower than today, but no precise information is available. At that period, there was a considerable population density only in Mexico and in the North and Central Andes; after 1492, important demographic changes occurred, and the indigenous population decreased almost everywhere. With the arrival of Europeans, three major changes affected the native populations: i) the number of indigenous people decreased practically everywhere and accounts today for less than 5% (much less in North America) of the total population; ii) individuals with European ancestry became the absolute majority in northern North America and in the southern part of South America. Elsewhere, high level of admixture with local indigenous groups took place (i.e: up to 80% of the total Mexican population are Mestizos, that means people of combined European and Amerindian or Pacific Islander descent); iii) starting from 1650 began the trafficking of Africans as slaves in the plantations and hence their number increased, especially in Brazil. Today individuals with African ancestry living in the American continent account for approximately 15-20% of the total population (Bryc et al., 2015; Cavalli Sforza et al., 1994). However, one of the major components that strongly affected our current ability to reconstruct the original genetic landscape of the Western Hemisphere from the modern-day population is the total loss of several tribal groups particularly in North America and along the Atlantic coastline of the double-continent because of extermination and epidemics.

3. Aims of the research

Y-chromosome NGS in South American populations

The maternally transmitted mtDNA has provided and is providing new insights on the possible routes followed by the AMHs inside the Americas and their arrival on the Southern Cone, but this is only a part of the story. A more complete picture of the peopling of America would be provided by the analysis of the oloandric transmitted Y chromosome whose variation has been less explored. Q-M242 is the only main Y-chromosome haplogroup present all across the Americas. Its major founding sub-haplogroup Q-L54, accounts for the majority of the North Native American and virtually all the South Native American Y chromosomes. Two main haplogroup Q Native American founding sub-lineages have been identified: Q-L54(xM3) and Q-M3 (Dulik et al., 2012; Battaglia et al., 2013). However, the incomplete phylogeny and the absence of phylogeographic data make this information not yet suitable to investigate the history and demography of Native American populations.

Aim of this project was to provide, through a fine dissection of the Pan-American haplogroup Q, new Y-chromosome markers to investigate from a male perspective the genetic history of South America. Therefore, to shade light on America's first colonizers, particularly regarding the timing of their arrival in South America and the routes followed. I carried out a comprehensive and detailed reconstruction of the phylogeography of its main sub-lineages.

Genome-wide haplotype analysis in Italian populations

In the past, many population geneticists have investigated the Italian population. Only in the recent years, their attention was centred on genome-wide data. Among them only one study used haplotype dense data to infer a fine genetic structure in Italy (Fiorito et al., 2015), but unfortunately the dataset assembled did not represent all the area of Italy and explored only in part its population structure. One aim of this work was, therefore, to investigate the genome structure of Italian population at a finer level of resolution and to compare its variation to the European and non-European groups. With this purpose, I merged data obtained from newly genotyped Italian samples and published data from Italian and worldwide samples and I used allele frequencies and haplotype-dense data to explore the degree of population structure and provide a preliminary chronological scale of the episodes of admixture that shaped the Italian genome variation.

4. Materials and methods

4.1 NGS Y-chromosome Analysis

On the whole, five regions of the X-degenerate portion of the MSY (3,768,982 bp) were sequenced and 5,274 fragments as in Scozzari et al. (2014), for 1,495,512 bp, were considered. Library preparation, targeting, sequencing and alignment were carried out by BGI-Tech (Shenzhen, Guangdong, China) (https://www.bgi.com/).

4.1.2 Variants calling

BGI-Tech provided us with BAM files with the suffix .bam, the binary version of a Sequence Alignment Map (SAM) files. It is a tab-delimited text file that contains large nucleotide sequence alignments, including a header section (optional, starts with "@") and an alignment section (Li et al., 2009). It is widely used for storing data, such as nucleotide sequences, generated by Next Generation Sequencing technologies. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position and variable number.
With the aim of an interactive exploration of the new generated sequences, the BAM files were first loaded into Integrative Genomics Viewer (IGV, Robinson et al., 2011; Thorvaldsdóttir et al., 2013), a software that enables the visualization of large, integrated genomic datasets. The files are aligned to a reference sequence and it is possible with a graphic interface to see how many times and which base it is read at each corresponding position.
SAMtools (Li et al., 2009) was the tool used to index, sort and merge the BAM files. In order to accelerate the analyses, the sequences were shortened to the region of our interest with the command.

```
samtools view -bh input.bam chrY: 2,689,001-19,550,033 >
output.bam
```

Then variant calling was performed with the following command lines,

```
samtools mpileup –g –f chrY.fa input.bam > output.bcf
```

in which chrY.fa is the reference sequence of the Y chromosome in FASTA file. By the comparison within the reference and the query sequences some outputs are generated they resulted in binary format files called BCF containing all the variants. The final conversion to the VCF file, was operated with BCFtools (Narasimhan et al., 2016) with the following command:

```
bcftools view –I –l positions.bed variants.bcf >
variants.vcf
```

where "positions.bed" is a BED (Browser Extensible Data) file that contains the coordinates for all the 5,274 fragments of Scozzari et al., (2014). The VCF file is instead a text file format containing meta-information lines, a header line, and then data lines each containing information about all the variants at a specific position in the genome. It also displays for each candidate variant the nucleotide position and a single Phred quality score (accuracy of consensus calling, named also Quality Score of consensus); it identifies the quality of the identification of the nucleobases generated.
The number of reads (coverage or depth) for each position ≥ 4 was used to confirm the validity of candidate mutations with a Phred quality score (QS) and a mapping quality score >90. In addition, the difference between the depth and the total number of reads for the two best bases for position was introduced and a cut-off of 4 was taken in consideration.
Finally, variant calls with QS = 99, depth ≥ 4 and difference < 1 were considered true mutations. Variant calls with QS ≤ 90 or QS > 90 but difference > 4 were indicated as Not Available (NA) to discard potential false SNP calls, while calls with 90 ≤ QS ≤ 99 and 1 ≤ difference ≤ 4 were manually inspected by visual examination of BAM files through the Integrative Genomics Viewer (IGV) software (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

4.1.3 Dataset merging

Several public available samples have been merged for different purposes. To establish the ancestral allele, for each of the positions found as variants in the new samples, it was compared with the allelic state carried by sample belonging to the deepest branches of the Y-chromosome phylogeny, A00, reported by a recent high-coverage MSY resequencing studies (Scozzari et al., 2014).
To increase the number of haplogroup Q samples we considered sequences from different works: Balanovsky et al. (2016); Hallast et al. (2014), Karmin et al., (2015), Mallick et al. (2016), Raghavan et al. (2014), The 1000 Genomes Project Consortium, (2015) and Zhou et al. (2013). Moreover, to better define the phylogenetic structure of the tree two haplogroup R1b samples were added (Trombetta et al., 2015; Underhill et al., 2015). Furthermore, with the purpose of calibrating the age of haplogroup Q sub-lineages some ancient Y-chromosome data were added (Rasmussen et al., 2010, 2014, 2015).
Q samples from Karmin et al., (2015) were available as Complete Genomics Master Variation (masterVarBeta-[Sample-ID].tsv.bz2) zipped files at the data repository of the Estonian Biocentre (http://evolbio.ut.ee/chrY/); they were converted to VCF format files using the C compiled tool masterVar2VCFv41_rev1.c. The variants not included in the regions of interest were filtered out using tabix (Li, 2011) with the following command:

```
tabix -fh input.vcf positions.bed > output.vcf
```

in which, positions.bed is the file described in the previous section.

R1b samples and samples from Mallick et al., (2016) were available as BAM files from which the resulting VCF files were extracted using the same pipeline applied to our samples (see section 4.1.2).
The merging of all the VCF files was obtained with the use of VCFtool:

```
vcf -merge output_1.vcf.gz -outputfile2.vcf.gz > output_tot.vcf
```

Tabix was used again to call new putative variants along the regions of interest.

```
tabix -R positions.bed -h output_total.vcf.gz > finaloutput.vcf
```

VCF files were available for the samples published by Hallast et al., (2014), Raghavan et al., (2014) and Zhou et al., (2013); the new variants were retrieved with the procedure mentioned above.

A final data set of 154 samples was composed of which 34 individuals are newly reported here (for details on the composition see 5.1.2 section).
A total of 1,563 variant positions were identified, among with, 1,550 were filtered out when the recurrent variants were discarded and 1,328 belonged to the haplogroup Q.

4.1.4 Parsimony tree and Network

Phylogenetic tree can be inferred using a Maximum Parsimony (MP) principle. Most parsimonious trees have the minimum tree length (fewest genetic changes from the common ancestor) needed to explain the observed distributions of all the characters. The branch lengths are the numbers of individual genetic changes. When two (or more) trees are equally parsimonious, there is no criterion for choosing between them, and no unique tree can be inferred, in the opposite case, the best tree is the one that requires the smallest number of mutations to account for the sequences. The optimality criterion used is that the chosen tree should be the most likely one. Under a given evolutionary model, the best tree is that which has the maximum likelihood (ML) of representing the real data. Moreover, alternative evolutionary models can be compared by applying the likelihood ratio test. Both tree topology and branch lengths should be inferred at the same time using a ML approach, as both affect the likelihood of the tree. Although maximum parsimony was originally proposed as an approximation to maximum likelihood methods, it has been shown that the shortest tree (the MP tree) is not always the most likely tree (the ML tree), although in practice they are often very similar. In order to cope the issue mentioned above, the new Subtree-Pruning-Regrafting (SPR) algorithm of the software MEGA6 (Tamura et al., 2013) was used. It searches for the optimal tree under the maximum likelihood (ML) and maximum parsimony (MP) criteria (Nei and Kumar, 2000; Swofford, 1998) generating the best hypothesis of relationships among the individuals. A contingency table in which alternative bases by subject represent rows and

chromosome position are columns was prepared and converted to the input file (.meg). Furthermore the graphical used interface (GUI) introduced in the MEGA6 version (Tamura et al., 2013) was used to run this program, therefore no command lines were required.

Gene flow between populations can result in their sharing young alleles despite a more ancient fission from a common ancestor. These kinds of processes generate loops within phylogenies known as reticulations or cycles. Such phylogenies are called networks. A single network contains within it several trees. Thus, if two trees are similarly well supported by the data, we do not have to choose one over the other, but can summarize them in a network. This network represents more of the information present in the data than does either tree alone. As with trees, networks can be constructed from distance or character data (such us SNPs or STRs), and there are a number of alternative methods of construction. A method for constructing networks with limited levels of reticulation is known as median joining (Bandelt et al., 1995). The algorithm used is based on the limited introduction of likely ancestral sequences/haplotypes into a minimum spanning network of the observed sequences. The median-joining algorithm is fast, has the advantage of being applicable to multiallelic polymorphisms as STRs, and is useful for large dataset.

In order to build a network useful to infer dates of splits for different lineages and solve their distribution, the same contingency matrix used for MEGA6 as well as STRs data were converted into the Network 4.6.1.5 (Bandelt et al., 1999) input file format .rdf. Network analysis uses a median joining method; it combines features of Kruskal's algorithm for finding minimum spanning trees favoured by short connections and maximum-parsimony (MP) algorithm, which sequentially adds new vertices called 'median vectors' (Fig. 16).

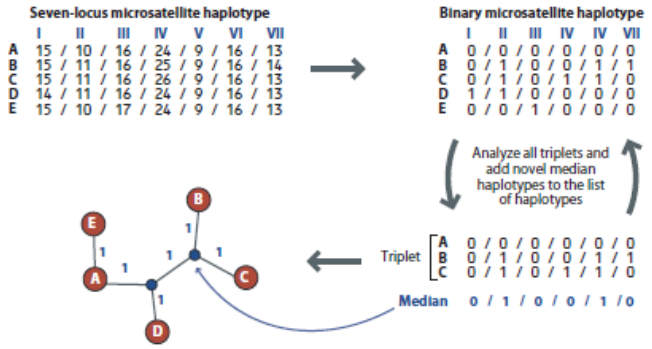


Figure 16: Network construction. The data are five haplotypes (A-E) with seven linked microsatellites. The haplotypes are converted in a binary format. Invariant loci (V and VI) are ignored. The smallest allele at a microsatellite is designated 0. An allele one repeat larger is designated 1. For microsatellite IV, where three alleles are present, the smallest is 00, the allele one repeat unit longer is 01, and the allele two
repeats longer than the shortest is 11. This takes account of a single-step mutational mechanism. Ancestral nodes not present in the sampled data are represented by small filled circles. A single most parsimonious tree is produced by the median network algorithm in this case.

Namely, MEGA6 was used to obtain the phylogenetic tree of the sequences sequences listed above (all sequences, Hg A00 sample excluded), through the maximum parsimony (MP) method. 10 bootstraps were used to generate a condensed tree. All positions containing gaps and missing data were previously eliminated. The exclusion of the Hg A00 sample was for tree topology reasons, since introducing a too different sequence would have flattered too much the tree.
Subtree-Pruning-Regrafting (SPR) algorithm (Nei and Kumar, 2000) with search level 0 was used in which the initial trees were obtained by the random addition of sequences with software MEGA6.

MEGA6 was used to obtain the phylogenetic tree of the 154 sequences listed above (all sequences, Hg A00 sample excluded), through the maximum parsimony (MP) method. 10 bootstraps were used to generate a condensed tree. All positions containing gaps and missing data were previously eliminated. The exclusion of the Hg A00 sample was for tree topology reasons, since introducing a too different sequence would have flattered too much the tree.
Subtree-Pruning-Regrafting (SPR) algorithm (Nei and Kumar, 2000) with search level 0 was used in which the initial trees were obtained by the random addition of sequences with software MEGA6.

4.1.5 Time estimation (Rho-statistic)

Rho statistic is a method that averages the number of differing sites between a set of sequences and a specified common ancestor (which needs not be among the sampled sequences), estimating a coalescent time using the Network 4.6.1.5 (Bandelt et al., 1999) output (Forster et al., 1996). This statistic, rho= ρ, is linearly related to mutation rate and time, assuming constancy of the rate across the tree branches. This requires a haplotype phylogeny, which must contain the root haplotype, and so the median joining network generated was used, as it allowed the reconstruction of ancestral haplotypes even if it was not observed within the sample. These mutational changes were counted from the network previously generated, rather than by estimation from the observed number of differences between two haplotypes. Confidence intervals for ρ were also calculated directly from the network.

Rho statistic analysis and associated confidence interval were computed with the program Network (Bandelt et al., 1999), a mutation of 0.82 x 10-9 (Poznik et al., 2013) was applied.

4.1.6 Time estimation (Bayesian methos)

Another coalescent approach for time estimation can be applied using a Bayesian method which requires prior estimates of the parameters, usually based upon existing diversity data. Recently bayesian methods are widely used due to the increase of computational power. They are based on Markov chain Monte-Carlo (MCMC) simulations, a class of algorithms that sample a probability distribution after some

sequence data are given. If an ML approach gives the probability of the data (such as the sequence alignment) given the hypothesis (the tree), a Bayesian approach determines the probability of the hypothesis (the tree) given the data (the sequence alignment). So, given a prior distribution of parameters, and a likelihood of the tree will be generated a posterior distribution of the data and a tree. Bayesian approaches do not generate a single tree, but a posterior distribution of trees that can be used to infer the timing of the most important demographic events that shaped groups of individuals.
BEAST (Drummond and Rambaut 2007) is a cross-platform program for Bayesian analysis of molecular sequences. It is orientated towards rooted, time-measured phylogenies inferred using molecular clock models. It can be used as a method of reconstructing phylogenies but is also a framework for testing evolutionary hypotheses without conditioning on a single tree topology. BEAST uses MCMC to average over tree space, so that each tree is weighted proportional to its posterior probability. Usually the input file required to run BEAST is generated using BEAUti (Bayesian Evolutionary Analysis Utility) (Drummond et al., 2012), a GUI application used after the conversion of the MEGA file in NEXUS file. A strict clock (fixed mutation rate across branches) was preferred to a relaxed molecular clock model after the inspection of the distribution of the tree likelihoods (Nylander et al., 2004).
Trees generated running two times BEAST with 20 million of steps and sampling every 10,000 steps using a coalescent expansion growth model were combined after discarding the first 2,000 generations of each replicate using LogCombiner v.1.7.5 (a GUI tool that combine the output of multiple runs of BEAST). The trees were then summarized using TreeAnnotator v.1.7.5, a program that summarizes the information from a sample of trees produced by BEAST onto a single “target” tree. The information obtained includes the posterior probabilities of the nodes, the posterior estimates and High Posterior Density (HPD) limits of the node heights. The results of divergence times applied to a phylogenetic tree were visualized on maximum clade credibility (MCC) tree using FigTree v.1.4.2

The Bayesian skyline plot (BSP) model uses standard MCMC sampling procedures to estimate a posterior distribution of effective breeding population size (Ne) through time. Tracer v1.6 (Rambaut et al., 2014) was used to produce the BSP starting from the information included in output file with suffixes .tree and .log.

4.2 Genotyping of South-American Y chromosomes

The genotyping was performed by RFLP analysis or Sanger sequencing. For each variant position the sequence surrounding was downloaded from the UCSC DAS server. The primers for amplification of the pertinent fragment were either designed with Primer3 software and checked with Primer-BLAST or chosen from YSEQ DNA shop (www.yseq.net). The signature markers characterizing the most important sub-clades of haplogroup Q identified were tested in all the samples of our collection belonging to haplogroup Q (positive for M242 marker). After having amplified the pertinent fragments the amplicons were analysed with different methods depending on the type of mutation and its surrounding sequence. Restriction analysis (RFLP) and electrophoresis of the digested amplicons were chosen when the mutation under analysis created or abolished a site for a specific restriction enzyme. The markers not affecting any restriction enzyme site were analysed through Sanger sequencing.

Four hundred and twenty-six (426) Native American and 87 Eurasian samples belonging to haplogroup Q were genotyped for markers (Table S3) defining the identified sub-haplogroups.

Table 1. Biallelic markers analysed in the present work.

ID SNP	Nucleotide change	Forward Primer 5'→3'	Reverse Primer 5'→3'	Amplicon size (bp)	Amplicon SNP position (bp)	Analysis used	Primer Refs
B37	A->G	gaaagaggctttaagtcaacaacag	tttatttcattgtagaccaagaagttg	433	249	RFLP	This work
B43	A->G	tgcacactgtgtagcccttc	ggacaaatgcgtgtccttaaa	212	128	RFLP	This work
B50	C->A	tatagcatggcggggagtag	ccactgagcatctgtggaag	356	178	SEQ	This work
CTS1002	G->A	agcaaagctgagatagtggaaa	tgtagaggggagaacaaatgat	300	115	RFLP	This work
CTS2731	A->G	caggatcttgccatgttgct	aaataaacgacctcagtcattg	363	187	RFLP	This work
CTS4000	C->T	ctgctgatgctctctgttgc	tgggccatatagtgagactctg	380	199	RFLP	This work
CTS748	C->A	tgcaaaccctctccttctgg	ttgctggcctaagttcctca	186	87	RFLP	This work
L330	T->C	tgggtggtagagaggaatgg	tgatttgttccggacacttg	519	340	SEQ	familytreeDNA
L713	A->C	cttgggctgtgggacaac	tgacaatgaaacactgggattt	177	126	RFLP	This work
M1107	G->A	tttttgagacagtttcctcttg	tgactaagaacaggttatgatagg	442	257	RFLP	Ybrowse
M378	A->G	tatgcatttgttgagtatatgtc	gttctgaatgaaagttcaaacg	326	114	SEQ	Sengupta et al. 2006
M848	G->A	atcactatcatggccgcagt	agcataattcccctgctcct	226	112	RFLP	This work
M925	G->A	gccttgaagacaatttgact(c)t	tgctatttttgaagtaactatgtgca	197	22	RFLP	This work
L301	T->C	ttggcatcccttaactcttg	aattcggaggtctttggttg	449	283	RFLP	Gurianov et al., 2014
L245	G->T	tcaaacaggaatattaagtccacc	tttctcagccaagagaaggg	391	278	SEQ	Gurianov et al., 2014
Y2250	A->T	atttgtgtcacgaggtcatgg	ccctggccagtcagaaaag	385	188	RFLP	Gurianov et al., 2014
Y26467	C->T	gcagatggtctagtggctct	gcatcaagtcccagcatcag	387	249	SEQ	This work
Y12421	C->T	ccaacgaggctgcaggtc	atgttacttatcagagttaaggtctgtgtg	291	254	RFLP	This work
Y26547	G->C	aagcatgagccaccatatcc	cctggtgccttacttaatttgg	219	173	SEQ	This work
Y4273	C->A	actcagtttctcgcaacagc	gggtctgatttcctggggta	276	134	SEQ	This work
Y4303	G->A	aacccagacaaggaattcaga	cccaatgtgagaacctggat	217	104	RFLP	This work
Y780	C->A	tcacaatgaccctttgtgc	tttatgaacgcctagctgacac	598	258	RFLP	This work
Y805	G->A	aaaataaagggatggaggaacatac	tgcagatagcattgttttcttaatg	443	119	SEQ	This work
YP1102	T->C	tcaattgcaaaaggatatttcaac	cccagaacagctctccacac	362	311	RFLP	This work
YP910	T->C	gcctgcccatctcttacat	cattgtctcaccatggcaac	360	184	RFLP	This work
YP919a	G->C	tccacaaaaaggttgggaga	cctcaattacagtcccaggaa	376	181	RFLP	This work
YP919b	T->A	ccaacgttgtacccaaattttc	aaacgttttaggttgattttgtttg	442	225	SEQ	This work

ID SNP	Nucleotide change	Forward Primer 5'→3'	Reverse Primer 5'→3'	Amplicon size (bp)	Amplicon SNP position (bp)	Analysis used	Primer Refs
YP919	G->A	gagcaacatcat ctgggtactg	gtgatgtggctttt tcaacctg	304	158	RFLP	This work
Z5906	A->C	tgatgtccctttg ggctatc	aatggcacaaaa ggagttgc	395	288	SEQ	This work
Z5906a	G->T	gtcttctctcag ccccatct	agcagctttgga tggggata	222	33	SEQ	This work
Z5907	G->A	ctattccttccat cgtctgcc	acaaacagcatg cattctgg	372	111	SEQ	This work
Z5908	T->C	agtaatttgcct gcctcagc	ggaatccttcca atatctcatgg	436	333	RFLP	This work
Z5910	T->C	ggcagagaagtt ttgttcttgg	aggggatggaca aatcacac	407	256	RFLP	This work
Z5910a	G->T	acttggggtgga aagagacc	aaccatggggaa cacaggt	280	166	RFLP	This work
Z5910b	C->A	tgcttccaatttt atccttgc	tgggaagccttt agtgcaaa	297	151	SEQ	This work
Z5911	G->A	ctaacatggatt cagataacttcc tg	tcagctgatctg tgttttcctc	402	58	SEQ	This work
Z5915	C->A	aaaatgtgctta ttccctttctg	ccctgctctaca aggattgc	420	228	RFLP	This work
Z35921	T->C	tttcagtgaactt agtttgcagtatt	tgaatgtgtttca gtttttctgg	357	168	SEQ	This work
Z780	C->T	cttcagtgaggt tggcttcc	cccctcaaagtg taaatctgc	403	187	RFLP	This work
Z781	C->T	ggagtacaagca atgttcaatgag	cactgtcaggtt ttaatgactcga	419	348	SEQ	This work
Z782	A->G	gagtttccaaaa tggttcctg	cataggttgtttg ggccatt	437	170	RFLP	This work

4.2.1 DNA quantification

The correct quantitation of DNA is a major step because the subsequent amplification reactions strictly depend on the concentration of each reagent involved.

DNA concentrations were determined using Quantus™ Fluorometer (Promega). The Quantus™ Fluorometer is a dual-channel fluorometer designed to provide highly sensitive fluorescent detection when quantifying nucleic acids and proteins. The instrument is designed for use with Promega QuantiFluor® Dye capable of excitation and emission in the proper wavelength range.

For the quantitation 1 µl of each sample was mixed well (pipetting or vortexing) with 199 µl of QuantiFluor® ONE dsDNA Dye in a 0.5 ml PCR tube (provided by Promega). Then, it was placed into the tube holder, and closes the lid. The instrument software automatically measures fluorescence and the calculated concentration is displayed.

4.2.2 Polymerase Chain Reaction (PCR)

PCR is one of the fundamental techniques that improved the development of molecular biology and genetics during the last 30 years. Kary Mullis idealized it in 1983, who ten years later, thanks to this revolutionary invention, won the Nobel Prize for Chemistry. The methodology consists in the direct amplification of a fragment of interest, called target DNA, of which only the flanking sequences (primers) need to be known. Starting from a single DNA molecule, this technique allows generating, thanks to DNA polymerase action, a key enzyme for DNA replication, thousands to millions of copies of a specific fragment. It consists of three cyclic repetitions, DNA denaturation, primer annealing and filament extension, which occur at different temperatures in the same tube: i) Denaturation: it occurs by raising the temperature to 94-95°C. Two single-stranded DNA filaments are obtained. ii) Primer annealing: at this stage, the primers anneal to complementary sequences on the fragment of interest. The melting temperature, also known as annealing, may vary from 50°C to 70°C according to the sequence of primers used. iii) Synthesis: after primers pair with the two complement filaments, the DNA polymerase begin to synthesize copies of the fragment in the 5'→ 3' direction. This step occurs at 72 ° C, the temperature at which Taq polymerase works optimally. PCR requires a specific enzyme that is Taq polymerase (extracted from Thermus aquaticus a thermophilic bacterium), which being heat-resistant does not inactivate by the elevated temperatures of the denaturing stage, thus it is available for the entire process.
(Table 1) summarizes concentrations and volumes of reagents used in the GoTaq amplification reactions.

Table 2. Solutions and their concentration in PCR reactions performed with GoTaq® DNA polymerase (Promega).

Stocks solutions	Reaction volume (final volume of 25 µl)	Final Concentration
Green Buffer®[a]	5.0 µl	1X
dNTP MIX	2.0 µl	100 µM
BSA[a]	0.25 µl	10 ng/µl
Forward Primer	0.05 µl	0.20 µM
Reverse Primer	0.05 µl	0.20 µM
Go Taq®	0.1 µl	0.50 U/µl
DNA	1.0 µl	2 ng/µl

GeneAmp® PCR System 9700 thermocycler was the tool used for the PCR reaction. For all the markers a touch down procedure was used, so called because in the first 14 cycles the annealing temperature drops by 0.5 °C per cycle; this program increases primer annealing specificity to complementary sequences. Table 2 reports the detailed conditions of the reaction.

Table 3. Scheme of the basic touch down PCR method used for reactions performed with GoTaq® DNA polymerase (Promega).

	Hold	Time	Temperature
Initial phase	Activation	2 min (GoTaq®)	95°C
Touch-down 14 cycles	Denaturation	20 sec	94 °C
	Annealing	1 min	63°C →56 °C
	Extention	1 min	72°C
35 cycles	Denaturation	20 sec	94°C
	Annealing	45 sec	56°C
	Extention	90 sec	72°C
Final phase	Final extension	10 min	72°C
	Maintenance	forever	15°C

4.2.3 Restriction Length Polymorphism (RFLP) analysis

When a SNP either creates or removes a recognition site for a restriction enzyme, the RFLP analysis was performed. The specific enzyme cuts the amplified fragment, producing an expected restriction pattern. Enzymes have to work at 100 % of their activity in order to avoid partial digestions; therefore all reaction conditions were properly chosen according to the enzyme manufacturer indication. The scheme of digestion reaction is reported in Table 3.

Table4. **Digestion reaction scheme.**

Stock solutions	Reaction volume (final volume of 30 µl)	Final concentration
10X PCR Gold Buffer	3.0 µl	1X
10X BSA (10 mg/ml)	0.3 µl	1X (3 ng)
Enzyme (10 U/µl)	0.15 µl	1.5 U/reaction
DNA (~200 ng/µl)	10.0 µl	~50 ng/µl

The acquisition/loss of a restriction site, caused by the mutation, introduces a variation of the electrophoretic pattern, making possible its detection. An electrophoresis run at 2,5 V/cm for about 1 hour, at the condition previously described, followed the digestion reaction and allowed to identify the different electrophoresis patterns.
A 3 or 4% agarose-gel was used depending on the expected sizes of the fragments.

4.2.4 Electrophoretic analysis

The fragments obtained by amplification and enzymatic digestion (RFLP) were analyzed by gel electrophoresis, a technique in which molecules are induced to migrate into a gel matrix subjected to an electric field.

The agarose gel is mainly used for the study of nucleic acids, as it forms a less dimensional three-dimensional grating and allows the separation of fragments that differ from many base pairs. Agarose concentration within the gel determines the size of the pores through which DNA fragments pass. Gel with low agarose concentrations (e.g.1%) has large pores, in which very long fragments (up to 10 kb) can migrate; gel with higher agarose concentrations have smaller pores and are appropriate to study shorter DNA molecules (few hundred pairs of bases).
The usual agarose concentration ranges from 1% to 4%; for PCR amplicons was used a gel with 2% of agarose, while in order to separate fragments obtained from restriction analysis a more concentrated gel (4%) was used.
The agarose was melted in a TBE (Tris-Amino-Methane 89 mM, Boric Acid 89 mM and Na2EDTA 20 mM) buffer solution, using a microwave.

Table 5. Buffers and solutions used in electrophoretic analysis

Solutions	Compositions
6X gel-loading buffer	0.2% bromophenol blue
	0.2% xylene cianol
	60% glycerol
	60 mM Na_2EDTA[a]
10X TBE	0.89 M Tris[b]
	0.89 M Boric acid
	20 mM Na_2EDTA
10 mg/ml ethidium bromide	

[a] Tris (hydroxymethyl) aminomethane
[b] Ethylenediaminetetraacetic acid.

When the mixture was completely melted, it was added ethidium bromide at the final concentration of 0.5 μg/ml. Once polymerized, the gel was inserted into the electrophoretic cell and so immersed in the TBE buffer. Samples were then loaded into the wells: 5 μl for amplification analysis and 15 μl for restriction analysis; parallel to the samples a known molecular weight marker (ladder) was loaded. For PCR reactions performed in buffer without the green dye, 5 μl of the amplified product were mixed with 1 μl of 6X gel loading dye before loading.
The electrophoretic run was performed at 4V/cm for 30-50 minutes. After that, the gel was exposed to ultraviolet rays and photographed using the Uvitec UVIdoc HD2 Gel Documentation System. The pictures were analysed with a dedicated computer program (Uvidoc 1D software), able to quantify the concentration and the size of the DNA loaded.

4.2.5 Sanger sequencing

When RFLP method could not be used it was necessary to perform Sanger sequencing.
The Amplified fragments were purified using the ExoSAP-IT® enzymatic system (Exonuclease I and Shrimp Alcaline Phosphatase, GE Healthcare) in order to degrade every single- strand DNA (i.e. not-incorporated primers). The enzyme is activated and inactivated with 15 minutes of incubation at 37°C and at 80°C,

respectively. The enzyme itself was degrades as well. The amount of DNA requested for sequencing reactions depends on the length of the fragment to be analysed: 1-2 ng/100 bp, thus for a 500 bp fragments 5-10 ng of amplified DNA is needed. An approximate quantification of the sample was obtained on agarose gel through comparison with the DNA ladder.

Once purified, one of the two amplification primers (3.2 pmoles) was added to the reaction, paying attention to the localization of the mutated site inside the fragment, because the quality of the sequencing reaction for the first 70 bps after the primer is low.

The final product was dried and sent to BMR Genomics (http://www.bmr-genomics.it). The analysis was carried out using a 96-capillary 3730XL and 3130XL Analyser systems, in accordance with Sanger's method.

4.3 Analysis on the allele frequencies of Italian populations

A clear distinction needs to be made between analysis in which variants (SNP) in genome-wide data are considered as a single entity and haplotype-based analysis in which chunks or segments of DNA, built starting from combination of SNPs, are taken in consideration. In this section it will be illustrated an overview of the most common techniques used when the SNPs are displayed as allele frequencies in different populations. These techniques can be different and include tools for quality control and data exploration to genetic admixture and demographic event detection. Analyses like principal components analysis (PCA) analysis (Price et al., 2006, 2010), FST (Excoffier et al., 2005; Goudet, 2002) belong to the first group, while ADMIXTURE (Alexander et al., 2009) and other inference based approaches including TreeMix (Pickrell et al., 2012) are of the second.

4.3.1 PLINK 1.9 format and merging the data

Genome-wide data coming from the SNP typing methods need to be merged, filtered and processed; all this tasks in this thesis have been operated with PLINK 1.9 (Purcell et al., 2007; Chang et al., 2015). It is an implemented and updated version of the free, open-source command-line program developed by Purcell and colleagues in the 2007.

Usually all the information of genome-wide data is contained in three files with specific suffix that can be processed by PLINK 1.9; important is consider that many formats can be handle by PLINK1.9 but the following are the ones used in this thesis. File with the suffix ".bed" need to be used together with ".bim" and ".fam" files. The first is a binary file containing the genotype information; the second, is a text file with no header and six separated fields, containing: -1 Population ID, -2. Individual ID, -3. Individual ID of father (if it is not present in the dataset "0"), -4. Individual ID of mother (if it is not present in the dataset "0"), -5. Sex code ("1" male, 2 "female", "0" unknown), -6. Phenotypic case ("1" control, "2" case, "-9"/"0"/non-numeric missing data if case/control); the third file is another text file but containing the variant (SNP) information as: -1.Chromosome, -2. Variant ID, -3. Centimorgan position, -4. bp coordinates, -5. Allele with minor frequency on the dataset, -6. Allele with the major frequency on the dataset.

The process that allows the combination of different group of samples in a unique dataset is called merging. During this procedure, the three file formats need to be considered as a single unit and they are read by PLINK 1.9 through the command line flag `--bfile`.

The command line used for merging is the following one:

```
plink --bfile filetomerge1 --merge filetomerge2 --make-
bed --out mergedfilename
```

in which the flag `--merge` allows the combination of each of file format with the names of `filetomerge1` and `filetomerge2`.

The output, will be three files named `mergedfilename`. This operation was repeated as many time as the number of dataset to be merged.

4.3.2 Filtering and pruning

PLINK1.9 was also used to convert data to the UCSC build37 and to filter for missing data for both genotypes (using the flag `--geno 0.02`) and individuals (`--mind 0.02`).

Subsequently through a pairwise comparison of all the individuals a pi-hat value (proportion of SNPs that were IBD) was estimated in order to discard samples with high degree of relationship. The acronym IBD in population genetics, might assume two meanings: isolation by distance or identical by descend, as is this case; in other words, IBD are segments or chunks of DNA that have been inherited from a common ancestor without recombination. A matrix with all the comparisons was generated (`--genome` flag) and a pi-hat value of 0.2 was used as threshold for the exclusion of highly related individuals (Busby et al. 2015).

According to the number of SNPs retained with the aforementioned methodology, two dataset defined as Low Density Dataset (LDD) and High Density Dataset (HDD) and containing 218,725 (4,852 subjects) and 591,217 (1,651 individuals) SNPs respectively, were defined and extracted.

To avoid capturing too much variance from SNPs in linkage disequilibrium (LD) regions in the analyses based on single SNP allele frequencies (PCA, F_{st} and F3) a recursively removing of SNPs with a threshold of R2 (square correlation coefficient) higher than 0.2 and using 50 kb sliding windows (using flag `--indep-pairwise 50 5 0.2)` was carried out (Busby et al. 2016). This lead to 83,079 and 135,628 SNPs in LDD and HDD, respectively.

4.3.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised learning method and is similar to clustering: it finds patterns without prior knowledge about whether the samples differ between them either genetically or due to some external factors (identification of outlier). This method simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, called principal components (PCs) or eigenvectors, which act as summaries of features. The spatial separation based on the genetic distances of a dataset can be plotted using the values (eingenvalues) for the first two PCs, of which figure 17 is an example.

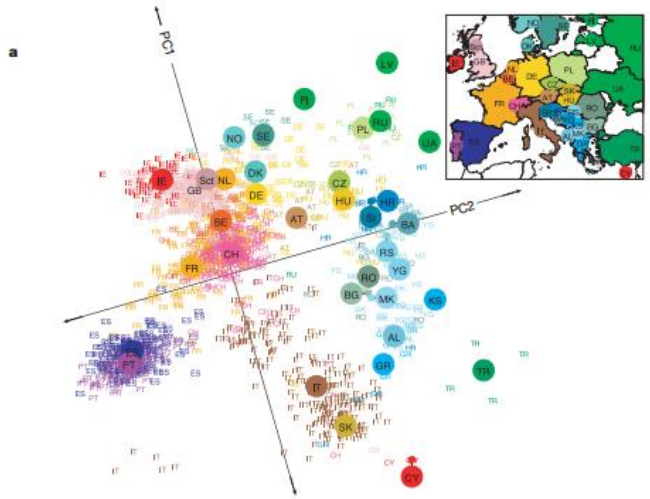


Figure 17. PCA analysis of European population from Novembre et al., (2008). The colours and letters are according the countries of birth; the big circles represent the median of the eigenvalues of each sample for the two eginvectors.

The generation of the covariance matrix, based on LD-pruned dataset (see section 4.3.2), in this work of thesis have been performed using PLINK 1.9 (Purcell et al., 2007) with the command flag `--pca`.

4.3.4 Population cluster analysis

Genetic clustering methods can be used to identify ancestry-based groups In the recent years the most commonly used methods for this purpose have been based on Bayesian clustering algorithms, such as STRUCTURE, initially introduced by Pritchard et al., (2000) and more recently improved in the software ADMIXTURE (Alexander et al., 2009; Alexander and Lange, 2011). These algorithms have the peculiarity of grouping genetic samples into K ancestral populations, sometimes called ancestries or components, on the basis ofgenetic affinities. Individuals belonging to the same group would be genetically more similar to each other than with samples of other components. This approach clusters individuals in order to maximise differences among groups. In addition, individual can be assigned as fractions of their genome to different clusters (Fig. 18).
ADMIXTURE input files used in this thesis, are the PLINK1.9 binary files cited above. Several runs characterized by several Ks are computed by ADMIXTURE, each run gives a probability value (cross-validation error) displaying the fit of each ancestral populations (K) to the dataset. The number K of ancestral populations with the best fit will be the one that better represents the analysed dataset. Among the output files generated by ADMIXTURE a matrix file with the suffix ".Q" is obtained, here each row represents a subject while the columns indicate the fractions of the genome representative for each K ancestral population. The matrixes can be represented as stacked bar plot in which the ancestries can be distinguished by a

number of colours equal to the one of the K ancestral populations (fig. 18). Clustering methods can take multi-locus genotypes of individuals from several populations and separate them into well-resolved Ks.

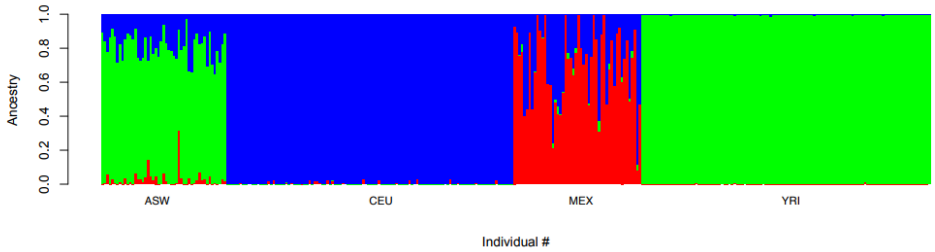


Figure 18. Barplot representing 1000 Genome Project subjects. Barplot generated by R (R Core Team 2012) (from ADMIXTURE manual).

In this thesis ADMIXTURE run have been computed for K = 2-18 on the LD-prunded LDD dataset, and the lowest cross validation values (called using the flag `--cv`) was reached at K = 15.

4.3.5 F3 statistic

F-statistic, developed in the 50's (Wright, 1949), is a statistical framework designed to test if an admixture event took place in the past. This statistic has been implemented to infer models on the basis of population frequencies that can be represented in the shape of a tree. This leads to an interpretation of the structure with the aim to predict pattern of admixture in populations. F-statistic methods can be divided according to the number of populations compared in the analysis: f4 when four populations are considered while f3 when the comparison of three populations is carried out. In this thesis, f3 analysis using treepop (Reich et al., 2009) implemented in the Treemix (Pickrell and Pritchard, 2012) software package, was used. Namely, the model, tests a possible admixture event considering three populations: A, B and a target population X (Fig. 19).

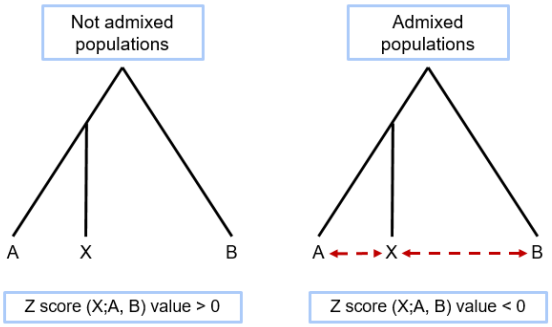


Figure 19. F3 model. The red arrows indicate the possibility of target population X to be either closer to A or B.

Under a null model it is assumed that if X underwent to an instantaneous divergence from A and B and it has been genetically isolated, any changes in allele frequency in the lineage of X due to genetic drift, should be independent of the allele frequency changes in A or B. Contrarily, if X is the result of an admixture between A and B the allele frequencies changes in the two source populations would not allow X to fit in a simple tree model with A and B (Fig. 19).
F3 analysis was performed on the LD-pruned LDD, with the following command line:

```
threepop -i input.frq.strat.gz -k 500 -o f3_Low.txt
```

The input file was obtained by the conversion of PLINK1.9 format files with a Python script implemented in treemix. The flag `-k 500` indicates a grouping of 500 SNPs that will be performed in order to account the fact the nearby SNPs are not independent. The resulting f3 statistic standard output is a four column file in which is present: populations used in the calculation (X;A, B), f3 statistic, f3 statistic standard error and Z-score. A negative Z-score indicates that population X is the result of an admixture between A and B; in this work, a threshold of -5 was considered as sign of admixture, and results with values smaller this threshold.

4.3.6 F_{ST} analysis

The degree of relatedness between populations by comparing DNA molecules can be estimated by different measures generally indicated as genetic distances. Considering four populations (A, B, C and D), there is a linear correlation between the values that calculated the genetic distances and their actual genetical similarity. For example, if the genetic distance between A and B has a value lower than between C and D, it is possible to assume that populations A and B are genetically closer than population C and D. Moreover, by comparing all the possible combination of populations (pairwise comparison) it is possible to identify possible genetic drift or

selective pressure that causes the differentiation is some population present high values of genetic differentiation.

Many ways have been implemented to measure the genetic distance, some of them are used according to the kind of data or depending on the purpose of the study. In this thesis I have used the so called F_{ST} statistic.

$F_{ST,}$ is a statistic method developed independently by Sewall Wright and Gustave Malécot in the 40s and 50s and it is possibly the most widely used statistic in population genetics to measure genetic distances. Belonging to the family of the fixation indexes, it is able to provide a value for the deviation of observed heterozygote frequencies supposed to be under a Hardy - Weinberg equilibrium (Holsinger and Weir, 2009; Wright, 1931); which states that the allele frequencies will remain constant from generation to generation in the absence of other evolutionary influences. The measure given by the F_{ST} statistic is a comparison between the diversity found within subpopulation to the genetic diversity of the total population. This rule can be applied, as in this thesis, as the measure of the allele frequencies differences between groups of sub-population.

F_{ST} value ranges from 0 to 1 and it is estimated from genetic diversity data, for example comparing two populations (pairwise F_{ST}):

$$Fst = \frac{\mathrm{Vp}}{\mathrm{p}(1-\mathrm{p})}$$

In which p and Vp are the mean and variance of allele frequencies between the two populations respectively. An F_{ST} equal to 0 means that there is no variance among the frequencies of the sub-populations therefore that the populations are similar or very close genetically. On the contray, high values of F_{ST} indicate population genetically diverse and differentiated. The software used in this thesis to calculate F_{ST} is EIGENESOFT (Price et al., 2006) under the implementation of smartpca. It accepts the output file of PLINK 1.9 (Purcell et al., 2007) and requires the preparation of a parameter text file after which the following command is given

```
smartpca -p parfile.txt
```

The output generated is a table that contains the F_{ST} values for all the possible sub-population comparison.

4.4 Haplotype-based analysis on the Italian populations

The use of methods based only on allele frequencies have unfortunately some limitations, for examples they cannot detect important information contained in variants that have been inherited together (see haplotype chapter 1.5.3) and they can be biased by the populations from which the SNPs have been extracted to build the array used to genotype; for this reason analyses on haplotypes and LD patterns has demonstrated to be a good method to improve the resolution of population structure at finer scale than analysing SNPs individually (Busby et al., 2015, 2016; Conrad et al., 2006; Gattepaille and Jakobsson, 2012; Jakobsson et al., 2008; Lawson et al., 2012; Leslie et al., 2015; Loh et al., 2013; Montinaro et al., 2015). A section of this thesis will be focused on the application of haplotype-based methods as CHROMOPAINTER (Lawson et al. 2012), with the aim of exploring genetic structure at a resolution higher than that provided by the analysis of SNPs frequencies.

4.4.1 Phasing

The information to be provided in the input file for linked methods, as for example CHROMOPAINTER, are haplotypes. In a diploid individual, the gametic phase or simply phase, represents the original allelic combination that an individual received from its parents. Generation of haplotypes (phasing) from genome-wide data in the form of allele frequencies needs the use of an algorithm; the one used in this thesis is SHAPEIT (Delaneau et al., 2012). It is an implementation of two previous software Impute2 (Howie at al., 2009) and MaCH (Li et al., 2010) and it is based on the collapse and the segmentation of known haplotypes in a graph. This shorter displayed haplotypes will be then used as template to know if each allele present in genome-wide data comes from the maternally or paternally inherited copies of each chromosome.
The input files for SHAPEIT software are PLINK binary files (see section 4.3.1) and a reference genetic map. A run for each chromosome is processed by SHAPEIT, therefore, if we consider the chromosome 22 the command line is as follow:

```
shapeit --input-bed gwas_chr22.bed gwas_chr22.bim
gwas_chr22.fam --input-map genetic_map_chr22.txt --
output-max gwas_chr22.phased.haps
gwas_chr22.phased.sample
```

in which are present the PLINK1.9 binary files and the genetic reference map (.txt) for the chromosome 22.
The output files contain respectively the phased haplotypes (.haps) and the included individuals with their missing data proportion.
Conversion of phased files (.haps) in input files for CHROMOPAINTER algorithm will be carried out using the perl-made script impute2chromopainter.pl

4.4.2 CHROMOPAINTERv2 analysis

Li and Stephens (2003) ideated the CHROMOPAINTER "painting" algortithm, which accounts for patterns of recombination along the chromosomes using the position of the available SNPs. As in the example illustrated in figure 20, we can observe that CHROMOPAINTER paints a recipient haploid genome (*h) as a combination of segments (chunks) that match two haploid genomes (h1 and h2) identified among all available donors (h1, h2, h3) through the use of a Hidden Markov Model (HMM) approach (Lawson et al., 2012). In this system, we will define as donors, individuals from which the algorithm will take segments of DNA to reconstruct the halotype of the recipients, in a process that is informally referred to as "painting".
Generally speaking it can be said that larger chunks from a specific donor means a more recent common ancestor between recipient and that donor than with another donor with whom smaller chunks are shared.
The process of painting is done along the entire genome and the result is a matrix in which the recipients are rows and appear as a mosaic of the donors (columns). Each row takes the name of copying vector or painting profile.

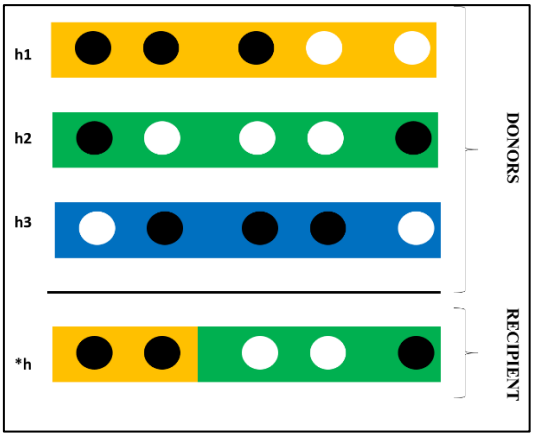


Figure 20. CHROMOPAINTER method. Donors haplotypes are indicated with h1, h2 and h3. The only recipient individual is named *h. The circles are the SNPs while the colours (black, white) is the allele type. The recipient individual has an ancestry shared with h1 and h2 as the allelic state shows, therefore it will be painted as a combination of these two haplotypes ("coping vector"). [modified ofVan Dorp (2017) PhD thesis]

In order to run CHROMOPAINTER there is the needs to specify two parameters, N_e and θ. The first one is the recombination or switch parameter while the second one is the mutation parameter. These values are calculated from an expectation maximisation algorithm (EM) that has to be run on a subset of the main dataset and chromosomes before the main run on the entire dataset and genome. Typically, ten iterations of the EM algorithm are run on chromosomes 1, 4, 15 and 22 to represent chromosomes of different sizes across the genome and for populations representative

of the genetic variation across the dataset. In this work, analysis on chromosomes 2, 5, 13 and 17 and on a subset of 382 subjects representative of all the genetic/geographic variation along the HDD (see section 5.2.2), was run.
The command used to estimate N_e and θ parameters is as follow:

```
ChromoPainterv2 -g outputshapeitChr22.haps -r
outputshapeitChr22.rec -t indfile.txt -a 0 0 -i 10 -in
-iM -o outputnameCP
```

in which the file .haps and .rec are the conversion of the outputfile of SHAPEIT done with the impute2chromopainter.pl pearl script mentioned above. The indfile(.txt) is a text separated file with three columns in which are indicated sample ID, population ID and inclusion or exclusion from the analysis ("0" excluded and "1" included). The flag `-a` indicates from which individual start and end the run, in this case `0 0` means all the individuals has to be considered. Flag `-i` indicates the iteration while `-in -iM` are the signals for the estimation of the N_e and θ parameters that will be present in the output file (EM.probs.out). Estimate values for these two parameters resulted to be $N_e = 364.578$ and $\theta = 0.00051$
The same command except for the flags `-in -iM` but with the flags `-n` and `-M`, in the form `-n` [N_e] `-M` [θ], will be employed for the main run of CHROMOPAINTER. Since computationally CHROMOPAINTER is quite costly the analysis need to be divided in groups of samples until all the dataset will be analysed. The output generated will consist in two files for all the chromosomes and all the groups of samples in which the analysis was divided. The output files will contain the following suffixes: .chunkcounts.out (chunkcount files) and .chunklenghts.out (chunklenght files).
The chunkcount files are matrixes in which in each row there is a recipient individual divided by as many columns as it is the number of the donors individual or populations. Each column will assign to each row a value representative to the number of chunks or segments shared with that specific recipients (name of the row). The chunkleghtsfiles are as the chunkcounts with the difference that each copying vector, instead of being formed by number of chunks, is composed by lengths of the segments shared with the specific recipients.
The totality of chunkcount and chunklenght files combined with a software called CHROMOCOMBINE will generate the final co-ancestry matrix used as input file for fineSTRUCTURE analysis (.chunkcounts.out) (Lawson et al.,2012).

4.4.2 FineSTRUCTURE analysis

The software fineSTRUCTURE (Lawson et al., 2012) is a Bayesian model-based markov chain monte carlo (MCMC) algorithm that cluster groups of samples; it is based on the co-ancestry matrix generated with CHROMOPAINTER.
FineSTRUCTURE is widely used nowadays, to infer the fine and detailed genetic on the basis of haplotype data. The groups generated by the analysis are called clusters or genetic groups, this nomenclature is widely used in this work of thesis

especially in the results and discussion section (see chapter 5.2.5). Clusters can be used as units for different analyses, instead of grouping samples o the basis of ethnicity or geographical origin. Sometimes, as in the 5.2.6 section they can be more reliable than the only geographic origin of each dataset (section 5.2.7).
FineSTRUCTURE software, in this work, was run in three subsequent modalities: the first, also called "greedy", infers in a fast way a rough structure of the dataset, and it is used when the number of samples is large (> 5000 individuals); the second, starting from the greedy tree of the first run, generates a MCMC file (.xml) that is used, by the third run, to build the tree structure analysed and visualized with R (R Core Team 2012). The following script is referring to the second run:

```
fs finestructure -X -Y -x 1000000 -y 1000000 -z 10000 -
t 100000 coancestrymatrix.chunkcounts.out
greedytree.xml MCMCfile.xml
```

Here, it has been set a number of "burn-in" iterations (`-x`), followed by sample iterations (`-y`), for the MCMC algorithm. These steps are followed by the number of sampled values (-z) and by the additional hill-climbing moves to reach its final inferred state.
Usually multiple runs of the above command are performed and only the MCMC file with the lower posterior likelihood is used as input file in the script of the third run:

```
fs finestructure -X -Y -m T -k 2 -T 1 -t 100000
coancestrymatrix.chunkcounts.out MCMCfile.xml
outputtree.xml
```

This command generates a tree (t) under the flag method (`-m`). Flags `-k` and `-T`, indicate the algorithm (`-k`) to be used in order to build the tree and the initialization (`-T`) to be performed.
FineSTRUCTURE inferred tree can contain many clusters which are difficult to refine or interpret. For this reason, it is possible to filter the clusters by cutting the tree at different heights and at different parts after visual inspecting the cluster organization. When considering a large dataset dendrogram of which only a subset of samples contains the populations of interest, by using this approach it possible to obtain finer or rougher structure by cutting the tree toward the leaves or the root respectively. The "cuttree()" function in the stats package (version 3.30) in R (R Core Team, 2012) helped on this task.

CHROMOPAINTER program and subsequently fineSTRUCTURE analysis, in this project, was performed using two separated analyses on both LDD and HDD (see section 5.2.2): i) in creating copying vector of Italians as a mosaic of all worldwide samples included in the datasets and ii) in creating the copying vectors of Italian individuals as a mosaic of only Italian samples. The first analysis was then used to

build the principal genetic structure and cluster composition displayed in the results and discussions section 5.2.5; while the second helped to build genetic groups used for the F_{st} analysis (see section 5.2.7)

5. Results and Discussions

5.1 Y-chromosome NGS in South American populations

5.1.1 Overview

There is a general agreement that AMH peopled the American continent across the Bering Strait between 20 and 15 kya from Asia. Into America these people followed two dispersal routes, one coastal and one inland. The first route would have brought a rapid human expansion through the double continent while the second, through the Cordillera corridor, would have contributed to the population of North America.
The same scenario is supported by genomic data (O'Rourke and Raff, 2010; Raghavan et al., 2014; 2015; Rasmussen et al., 2010; 2014; 2015; Reich et al. 2012; Skoglund et al., 2015) even if the majority of the information on the Americas colonization largely derives from the maternally transmitted mitochondrial DNA (mtDNA), which, from the very first studies, has shown that few different mtDNA haplogroups (hgs: A2, B2, C1, D1, D4h3, and X2a [Bandelt et al., 2003]), nested into Asian clades, characterize all present Native Americans. These haplogroups, with the only exception of X2a, entered North America from Beringia along the Pacific coast (Achilli et al. 2013; Perego et al. 2009; Tamm et al. 2007) and they have been observed also in South America (de Saint-Pierre et al., 2012; Brandini et al., 2017). Differently, the distribution of the founding Native American haplogroup X2a, for which the ice-free corridor migration route has been proposed (Perego et al. 2009; O'Rourke and Raff 2010; Hooshiar-Kashani et al. 2012) seems to be limited to the North-East America. Unfortunately, the identification of Y-chromosome Native American founding lineages has been complicated by the post-Columbian uneven male/female native population decline and the high historical rate of male mediated admixture into Native American communities. Nevertheless, two founding lineages with Asian ancestry, Hg C and Hg Q, were described in previous studies (Karafet et al. 1997, 1999; Underhill et al. 1996). Hg C is virtually limited to North America; differently, Hg Q-M242 is present as Q-L54 all over the double continent with two main Native American founding sub-lineages: Q-M3 and Q-L54*(xM3, L330) (Battaglia et al., 2013; Dulik et al., 2012). Little information is available about their distribution, but recent papers support the concomitant arrival of both lineages in Mesoamerica, where Mexico acted as recipient for the first wave of migration, followed by a rapid southward spread into the Southern continent (Battaglia et al., 2013; Grugni et al., 2015). In the last couple of years, thanks to advances in DNA sequencing technology, which allowed large-scale analyses of complete Y-chromosome sequences, and the increasing interest of citizens in participating in genealogical projects (ISOGG - http://www.isogg.org/tree-; Thomas Krahn (http://ytree.ftdna.com), new L54 sub-lineages have been identified (Jota et al., 2016; Karmin et al., 2015; Poznik et al., 2016; Wong et al., 2017). However, at present, the current level of resolution of this haplogroup, as well as its

phylogeography, are still inadequate to explore the history and demography of Native American populations.

5.1.2 The dataset and the variants

To increase the chance of identifying new phylogenetically informative SNPs, 34 unrelated males belonging to Native American haplogroup Q were selected, on the basis of their geographic origin and Y-STR haplotypes from Battaglia et al. (2013). Among these, 25 subjects were from Meso- and South-America (Battaglia et al., 2013) while nine from Asia, Middle East and Europe (Battaglia et al., 2009; Fornarino et al., 2009; Grugni et al., 2012 and unpublished data) (Fig. 21).

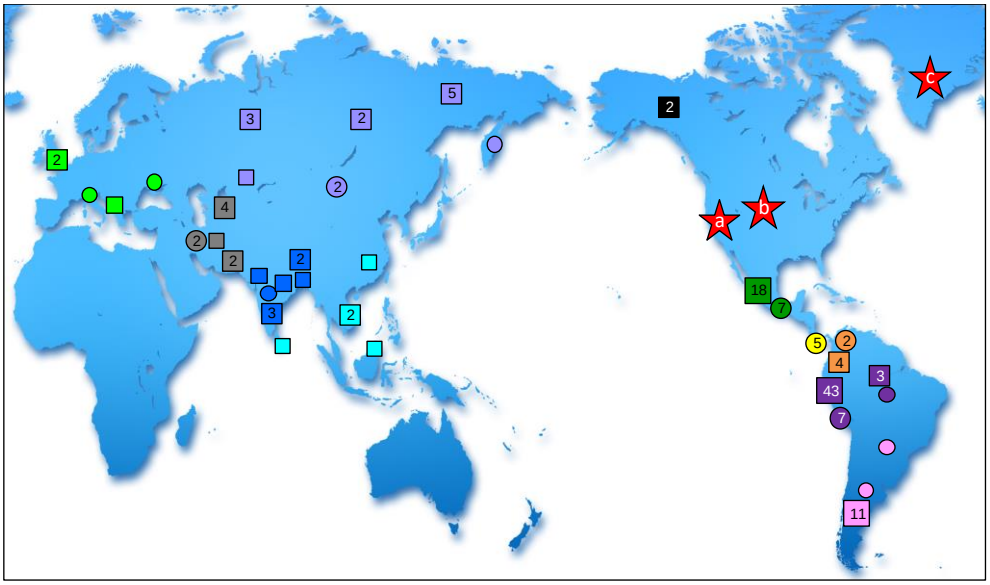


Figure 21. Geographic origin of the subjects included in the analysis. When more than one subject is from the same area, their number is reported inside symbols. (a) Rasmussen et al., 2015; (b) Rasmussen et al., 2014; (c) Rasmussen et al., 2010.

For all the 34 individuals, high-depth re-sequencing of a large MSY portion (3.7 Mb) was carried out by Beijing Genomics Institute (BGI) in Shenzhen (Guangdong, China), using the same protocol of Scozzari et al. (2014).
In order to capture the worldwide haplogroup Q variability a dataset including 113 modern samples (Balanovsky et al., 2017; Hallast et al., 2014; Karmin et al., 2015; Mallick et al., 2016; Raghavan et al., 2014; The 1000 Genomes Project Consortium, 2015; Zhou et al., 2013) and three ancient Y chromosomes ("Anzick", "Saqqaq" and "Kennewick") (Rasmussen et al, 2010; 2014; 2015) available from the literature, were merged (see chapter 4.1.3) with all the newly genotyped samples (Table S1). Two sequences belonging to Hg R1b, the sister clade of haplogroup Q (Scozzari et al., 2014; Underhill et al., 2015) were also included in our dataset.

From the analysis of the 152 Y-chromosome sequences in comparison with the hg A00 sequence (Karmin et al., 2015, Trombetta et al., 2015) defining the deepest branch of the Y-chromosome phylogeny (see chapter 1.3.2): 1,550 nucleotide positions (1,563 included recurrent positions; 1,328 inside haplogroup Q) carrying a derived allelic state were identified. Out of them, 937 (72.0%) variant positions were not annotated and 527 also not described in ISOGG, YFull or in Karmin et al. (2015).

5.1.3 Phylogeny of the haplogroup Q

The relationships occurring among the 154 (152 belonging to Hg Q and 2 to Hg R) sequences under study are illustrated in the phylogenetic tree of Figure 22 that has been obtained through Network and MEGA analyses. For its construction, informative SNPs outside the studied regions, which have become available from literature (Karmin et al., 2015, Poznik et al., 2016) and/or from genealogy web sites (ISOGG tree; YFull tree), were verified in our samples and considered in the analysis. On the phylogeny they are reported in Italics. The signature markers of the sub-haplogroups identified are listed for both the newly genotyped and already published Y-chromosome sequences.

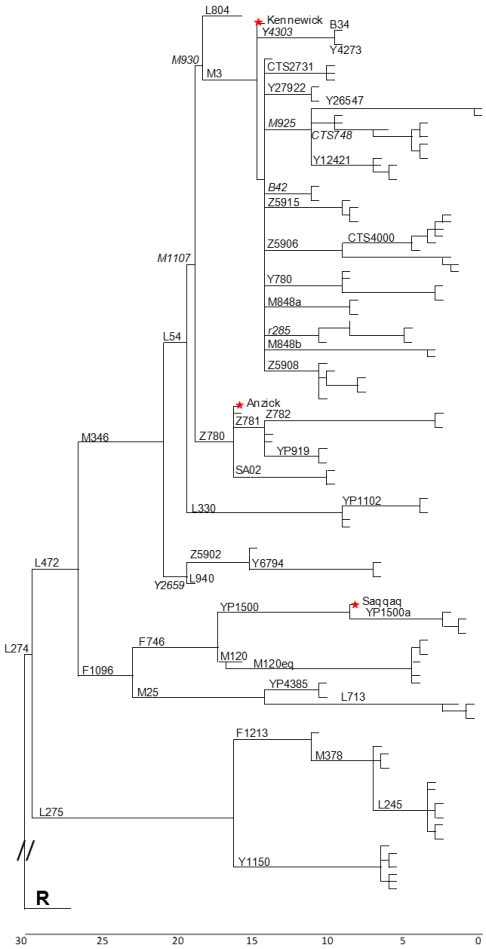


Figure 22. Condensed version of the MP tree. The cladogram was obtained considering 1,550 variable positions in 154 modern DNAs (this work; Zhou et al., 2013; Poznik et al., 2013; Hallast et al., 2014; Raghavan et al., 2014; Karmin et al., 2015; The 1000 Genome Project Consortium, 2015; Underhill et al., 2015; Trombetta et al., 2015; Mallick et al., 2016; Balanovsky et al., 2017) and 3 ancient DNAs (Rasmussen et al., 2010, 2014, 2015). The name of the markers defining each branch is shown above it. Markers reported in italics are outside the regions considered. Stars indicate ancient DNAs as in Figure 21.

On the whole eight main branches, Q-L275, Q-F1096, Q-Y2659, Q-L330, Q-Z780, Q-M848, Q-Y4303 and Q-L804, were identified by the phylogny: the first four include virtually only Eurasia Y chromosomes, instead the other branches, marked by M1107 mutation, harbour virtually only Native Americans. The only exception is represented by Q-L804, which in the tree is represented by an English Y chromosome (Hallast et al., 2014).
Divergence datings for each of the haplogroup Q sub-lineages were obtained with BEAST analysis; they are reported in the supplementary table 3.

Based on BEAST results and cladogram topology, a detailed analysis on the main splits characterizing the haplogroup Q phylogeny is provided in the following paragraphs.

5.1.3 Refinement of haplogroup Q phylogeny structure

The first split, occurred before 26 kya, distinguishes the Eurasian Q-L275 branch. Despite the early phylogenetic separation of this Eurasian branch, the age of its Most Recent Common Ancestor (MRCA) has been evaluated around 15 ky (14.0 ± 2.2 ky, this study and 15.1 ± 1.2, Balanovsky et al., 2017). In particular, in our dataset Q-L275 includes two groups of chromosomes with MRCA relatively young (Q-M378: 6.6 ± 1.3 ky; Q-Y1150: 6.1 ± 1.2 ky). The second subdivision generates Q-F1096 with a MRCA dated 19.3 ± 2.6 ky. It harbours two main branches distributed across Eurasia and the Middle East: Q-M25 (12.4 ± 2.2 ky) and Q-F746 (14.9 ± 2.2 ky). Both branches are further sub-divided. Q-M25 splits into Q-YP4385 (6.9 ± 1.5) represented by samples from India and Pakistan and Q-L713 (2.2 ± 0.8) by samples from Uzbekistan and Iran; Q-F746 is subdivided into Q-M120 (14.2 ± 2.1 ky), mainly represented by samples from East Asia, and into Q-YP1500 (8.4 ± 1.3 ky), comprising Siberian Y chromosomes and the Greenland ancient DNA of Saqqaq (4 kya, C14 dated). The observation inside the East Asian lineage Q-M120 of HG01944 and Tsimshian samples looks in disagreement with their geographic origin, being the first from the Andean region and the second from Alaska. While the single Andean sample likely reflects the origin of the carrier, rather than more general population history, the position of the Tsimshian chromosome is peculiar: it does not display any variant positions observed in the other M120 chromosomes, but shares with them three mutations that are outside the regions considered in this analysis. It could therefore represent a Northern sub-branch of Q-M120 early differentiated from the M120 ancestor. Should this be true, this Y chromosome would represent a descendent of a further, rare, Native American founding lineage. The third bifurcation, occurred around 17 kya, distinguishes the Q-Y2659 branch. This clade has been dated 16.8 ± 2.1 ky. It includes the lineage Q-L940 observed in one Ukrainian sample (this study) and in one Dutch of Karmin et al. (2015) (sample not included in this study in that not available) and the clade Q-Z5902 (13.4 ± 1.9 ky), which, with the exception of one Croat Q-YP1600, harbours mainly chromosomes of Asian origin. The forth branching, arisen about 16 kya, splits Q-L54 into Q-L330 (MRCA 8.3±1.5 ky), diffused in Asia, and Q-M1107 (MRCA 15.2±1.7 ky), observed in North Eurasia and predominant in the Americas. The latter splits into the Native American branch Q-Z780, also characterizing the Anzick aDNA (C14 dated 12.6 kya), and into Q-M930. This branch, in turn, is divided into Q-L804, spread in North Europe, and into Q-M3, the main Native American lineage. Finally, Q-M3, which characterizes the Kennewik aDNA (C^{14} dated 9 kya), differentiated sometimes after 14 kya into Q-Y4303 and Q-M848. While the first branch harbours both Siberian and Native American Y chromosomes with a MRCA dated 9.3±1.2 ky, Q-M848 is Native American specific. It is dated 12.5±1.6 kya and includes 11 main sub-branches and 10 singletons.

5.1.4 Phylogeography of the haplogroup Q

The distribution of the main clusters identified with the phylogenetic analysis was investigated by combining literature data (The 1000 Genome Project Consortium, 2015; Balanovsky et al., 2017; Battaglia et al., 2013; Di Cristofaro et al., 2013; Dulick et al., 2012; Fornarino et al., 2009; Grugni et al., 2012; Hallast et al., 2014; Jota et al. 2016; Karachanak et al., 2013; Karmin et al., 2015; Lippold et al. 2014; Malyarchuk 2011; Nonak et al., 2007; Poznik et al., 2013; Raghavan et al., 2015; Rasmussen et al., 2010; Regueiro et al., 2013; Sengupta et al., 2006; Mallick et al., 2016; Zei et al., 2003; Zhong et al., 2013) with the classification of more than 500 (513) samples of our dataset, mainly Native Americans (N=426) but also from Eurasia (N=87), obtained by hierarchical genotyping of the main haplogroup-defining-markers (Table 2). Haplogroup classification is summarized in supplementary tables 2 while the geographic distribution of the most diffused haplogroups are illustrated in figures 23, 24, 25, 26 and 27.

The Eurasian-specific branches of haplogroup Q

The deepest branch of the tree, the **Q-L275,** is considered the result of the oldest split of haplogroup Q (within 27.8 to 32.5, Poznik et al., 2016). It includes the sub-lineages Q-Y1150 and Q-M378. Q-Y1150 has been found mainly in West Asia, confirming the results of Balanovsky et al., (2017), the only exception is represented by an Italian sample. It could represent the result of gene flow of a member of a rare European lineage that can be visible at a finer scale.

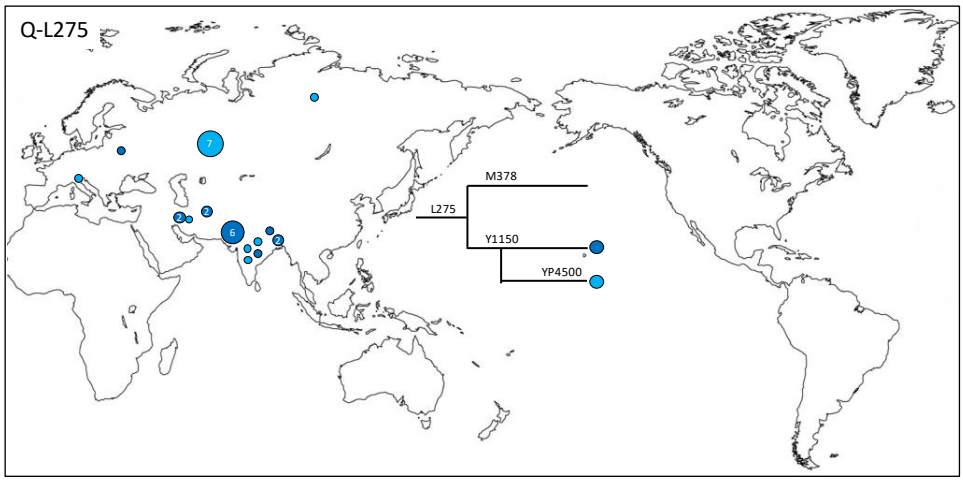


Figure 23. Phylogeography of the Q-Y4303 sub-clade. The figure illustrates the phylogenetic relationships of the markers investigated and their pattern of distribution. Circles without any number refer to one subject.

Q-M378, sub-clade of Q-F1213, has been recently dissected in the frame of a collaboration with citizen scientists (Balanovsky et al., 2017); it includes four branches, which spread across West, Central and parts of South Asia, harbouring mainly Middle Eastern Y chromosomes, with one branch typical of Ashkenazi Jews,

as well as European samples. In this work of thesis this sub-haplogroup is represented by three subjects: an Iranian, which in the Balanovsky tree (Balanovsky et al., 2017; Figure 3) would groups together with Azeri and Hazara into the Q3a7-BZ310 clade, an English which do not fall in any described Q-M378 branch, and a Panamanian sample, which shares the BZ15 marker with one Polish. The distribution of the main M378 branches is illustrated in figure 24. At the light of this information, the Panamanian sample has to be interpreted as the result of a post Colombian arrival from Eurasia, as previously hypothesized (Battaglia et al., 2013).

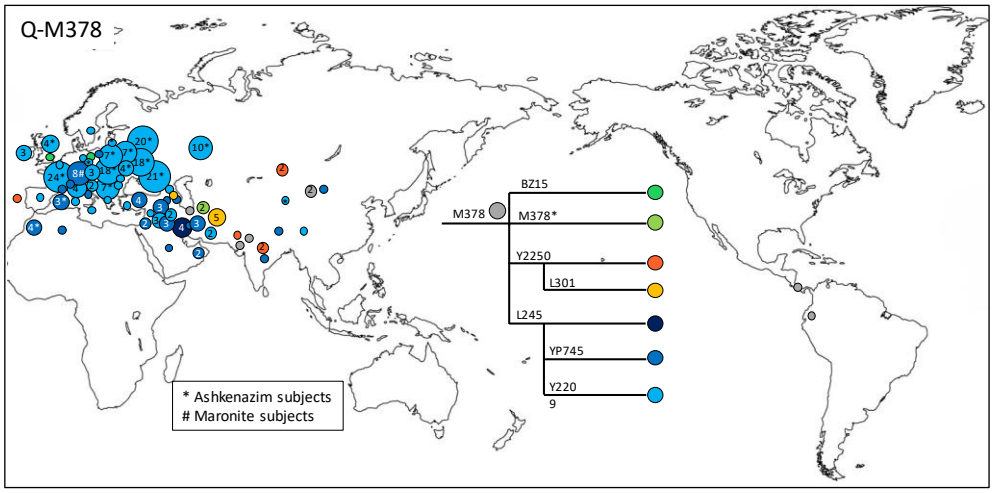


Figure 24. Phylogeography of the Q-M378 sub-clade. The figure illustrates the phylogenetic relationships of the markers (in different colours) investigated and their pattern of distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols.

Q-F1096, which has been generated by a second bifurcation, includes Q-M25 and Q-F746 branches. The first is observed in six South West Asian Y chromosomes whose MRCA has been evaluated to be 12.4 ± 2.2 kilo years (ky) old (Table S3) (Fig. 25); the second can be subdivided into Q-YP1500 observed in five Siberians and in the Saqqaq ancient DNA and Q-M120 observed in five South East Asians, in one sample from Alaska and in one South American subject. Y chromosome belonging to Q-NWT01 (Q-F746), not better sub-classified, have been also described in Northern Canada (Dulik et al., 2012) but not southwards, thus, taking in account the distribution of these sub-lineages it is lakely that these Canadian samples belong to the Q-YP1500 observed in Siberia and characterizing the Saqqaq ancient DNA. On the other hand, considering the phylogeography of haplogroup Q-M120 (Zhong et al., 2011), we can hypothesize that its presence in South America is ascribable to recent gene flow.

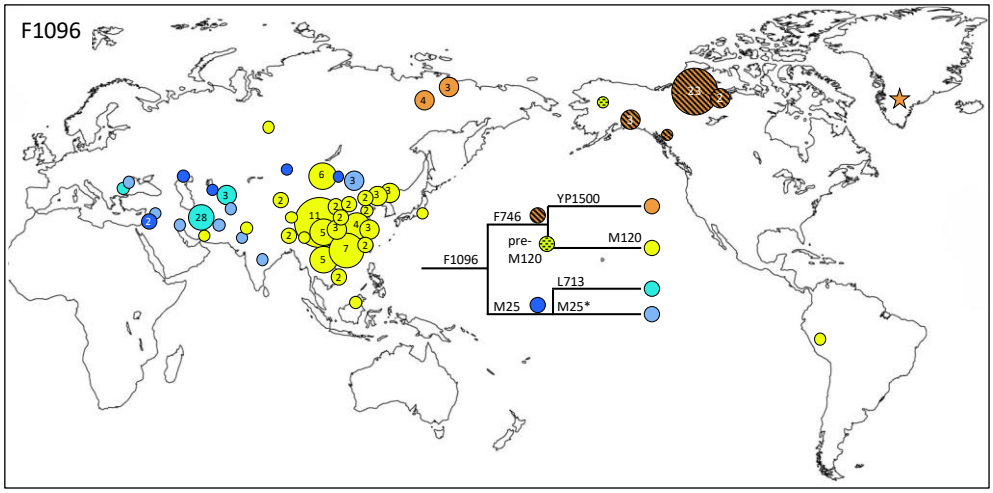


Figure 25. Phylogeography of the Q-F1096 sub-clades. The figure illustrates the phylogenetic relationships of the markers (in different colours) investigated and their pattern of distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols.

The third bifurcation generates **Q-Y2659**, it has been dated 16.8 ± 2.1 kya and includes Q-Z5902 branch; observed in four samples from the Indian sub-continent and in one European from Bosnia, and a single lineage (Q-L940) observed in one Ukraine.

The Q-L54 branch

The discovery of **Q-L54** and **Q-L330** markers (Dulik et al., 2012) clearly separated the Asian Y chromosomes, carrying the derived allele for both markers, from the American ones, derivatives only for Q-L54 (Q-L54(xM3, L330)).
Chromosomes Q-L53*(xL54) have been reported only in Central Asia (15 Chelkans and in 10 Tubalars from Altai Rep. – Dulik et al., 2012- and in a Mongolian sample of our dataset). These chromosomes could belong to the sub-clade Q-YP4004 recently reported in four subjects from Russia and one from Poland (YFull tree). If this scenario Q-YP4004 would represent a new sister clade of Q-L54.
Q-L330 branch in the tree of figure 22 includes a cluster observed in Siberia, Kazakhstan and Mongolia and two single lineages observed in Mongolia and in Uzbekistan (distribution showed in Fig. 26). Its MRCA has been evaluated 8.3+1.5 ky old, thus much younger than the bifurcation (occurred at least 15.2+1.7 kya) that separated this lineage from Q-M1107. In addition, L330 is much less represented than its sister clade L54. This observation indicates that, before spreading in Asia, L330 had a long persistence in some region of Central/North East Asia at low frequency or underwent a bottleneck.

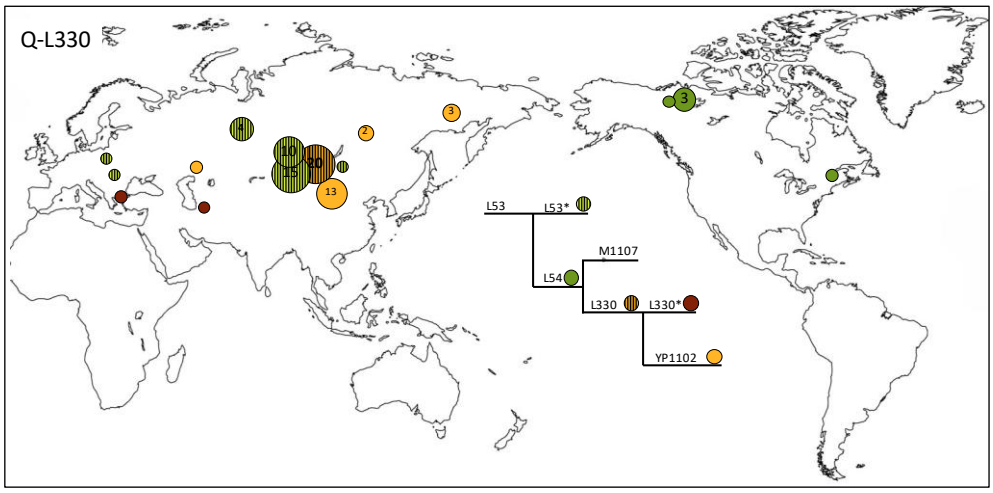


Figure 26. Phylogeography of the Q-L330 sub-clades. The figure illustrates the phylogenetic relationships of the markers (in different colours) investigated and their pattern of distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols.

The identification of the markers **M1107, Z780** and **M930** allows a better discrimination of the Native American Y chromosomes.
M1107 characterizes all the Q-L54(xL330) while the Z780 and M930 sub-lineages identify the two main Native American founding lineages, the first detecting all the Native American Y chromosomes previously classified as L54(xM3, L330) plus the Anzick aDNA (C^{14} dated 12.6 kya, Rasmussen et al., 2014), the second harbouring all the M3 Y chromosomes. However, Q-M930 includes also a rare North European lineage, the Q-L804 suggesting that the marker M1107 likely originated in small population groups during the Beringia standstill where it differentiated into Z780 and M930 and the last, in turn, into M3 and L804. From here these sub-lineages started to diffuse and, probably for genetic drift, Z780 and M3 toward the Americas, while L804 toward Northern Europe. Q-M3, which characterizes the Kennewik aDNA (C^{14} dated 9 kya, Rasmussen et al., 2015), after 13.4 kya differentiated into Q-Y4303 and Q-M848. While the first branch harbours both Siberian and Native American Y chromosomes with a MRCA dated 9.3±1.2 kya, Q-M848 is Native American specific. It is dated 12.5±1.6 kya, includes 11 main sub-branches and nine singletons.
Q-L804 is represented in our phylogeny by one English Y chromosome previously classified as Q-M3 (Hallast et al., 2014). Conversely, citizen scientists of the Q Nordic project of FTDNA (http://hoijen.se/2016/01/29/q-l804-current-status-2016-01-29/) have extensively investigated this rare European lineage in some England, Norwegian and French participants (Fig. 27). Although their results are not representative of its distribution, they support an Asian origin of M930, the upstream

marker of L804 and M3, indicating a North Eurasian route of dispersal and recent dissemination in some North European populations.

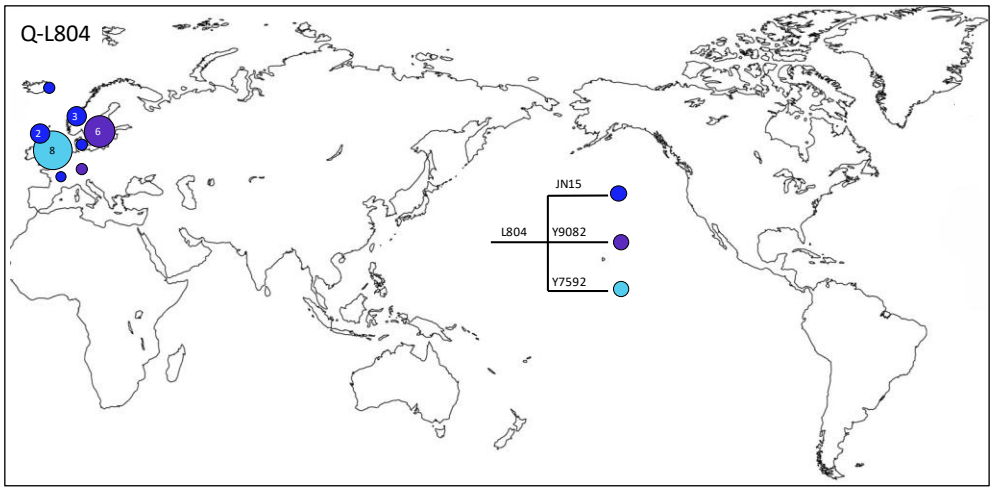


Figure 27. Phylogeography of the Q-L804 sub-clades. The figure illustrates the phylogenetic relationships of the markers (in different colours) investigated for this clade and their pattern of distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols.

The Native American-specific branches: Q-Z780 and Q-M3 and their sub-clades

All the Native American Hg Q Y chromosomes fell into two main classes on the basis of their M3 status: Q-M3 virtually exclusive of Native Americans and Q-M242(xM3) found also in Eurasia.

Q-M242 down-stream markers L54 and L330 clearly separated the Asian Y chromosomes, carrying the derived allele for both markers, from the American ones, derivatives only for L54 (Q-L54(xM3, L330). Now, the identification of the markers M1107, Z780 and M930 further distinguishes Native American Y chromosomes. Indeed, M1107 further characterizes the Q-L54(x L330), while Z780 and M930 identify the two main American founding lineages, the first detecting all the Native American Y chromosomes previously classified L54(xM3, L330), the second harbouring all the M3 Y chromosomes. Inside M3, two branches, Y4303 and M848, are distinguishable. Y4303, despite present in the phylogeny with only three chromosomes, is observed both in Asia (one Siberian) and in America (one Alaskan and one Mexican). M848 is the most represented M3 branch and is American specific. It harbours all the remaining M3 chromosomes organized into 11 structured clades but still including 10 singletons. Q-Z780 is less represented and structured than Q-M3. It includes only two clades, one of which, the SA02 previously identified as DYS391 = 6 (Battaglia et al., 2013) and five singletons.

Q-Y4303 (Fig. 28) is present in the phylogeny with only two lineages, one, B34, shared between a Siberian and an Alaskan Native and the other, Y4273, represented by only one Mexican. Although scarcely represented in our dataset, this clade

displays the widest geographic distribution, being observed from Siberia to South America. At present, it is the only clade found in the United States of America (in Virginia, Carolina and Georgia but also in the South-Western regions) where it seems to be associated to subjects speaking the Algonquian language (Shurr et al., 2004). The Algonquian are one of the most populous and widespread North American native language groups. Historically, these people were prominent along the Atlantic Coast and into the interior along the St. Lawrence River and around the Great Lakes and the Rocky Mountains. In North America, the distribution of this Y-chromosome lineage parallels that of mtDNA haplogroups X2a and C4c (Smith et al., 1999), which 10-8 kya entered from Beringia into North America through the ice-free corridor between the Laurentide and Cordilleran ice sheets (Hooshiar Kashani et al., 2012; Perego et al., 2009; 2010). Thus, it is likely that Q-Y4303 could have been diffused in North America by the same groups carrying the X2a and C4c mtDNAs. In addition, the observation of Q-Y4303 in Brazil but not in the numerous samples from the western regions suggests that this lineage entered South America not along the Pacific costal route but following the continental route. The MRCA of this clade has been dated 9.3±1.2 kya, thus more recent than the estimates for the peopling of the America. Thus, it is likely that the American age of Y4303 falls into the range of the American peopling and that Siberian-Alaskan B34 lineage, that dates 5.4±1.2 kya, represents a later back-migration to Asia.

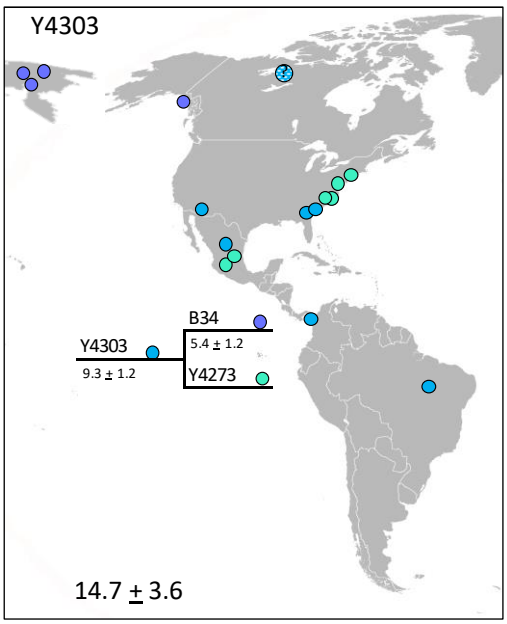


Figure 28. Phylogeography of the Q-Y4303 sub-clade. The figure illustrates the phylogenetic relationships of the markers investigated for this clade and their pattern of distribution. Circles without any number refer to one subject. Datings refer to Bayesian estimation of node ages and of the whole lineage.

Q-M925 (Fig. 29) includes three branches: Y26547 found in 2 Brazilian samples, Y12421 in Colombian and Panamanian samples and CTS748 observed in Mexico. In particular, this latter branch harbours almost all the Mexican M3 Y chromosomes of our dataset, half of them further characterized by the CTS1002 marker. Although less represented, the other two lineages, are both observed in Panama and in South America where, however, seemed to follow two different routes being the Y26547 encountered in Brazil and Y12421 in Colombia and Peru. The age of Q-M925 has been evaluated at 9.8±1.4 kya. Among the three branches, the Mexican CTS748 turned out the most ancient (8.5±1.4 kya) followed by its sub-clade CTS1002 (6.8±1.2 kya). The other two branches seem much more recent (Y12421: 5.3±1.0 kya; Y26547: 1.2±0.6 kya) but these values could be affected by their small sample size.

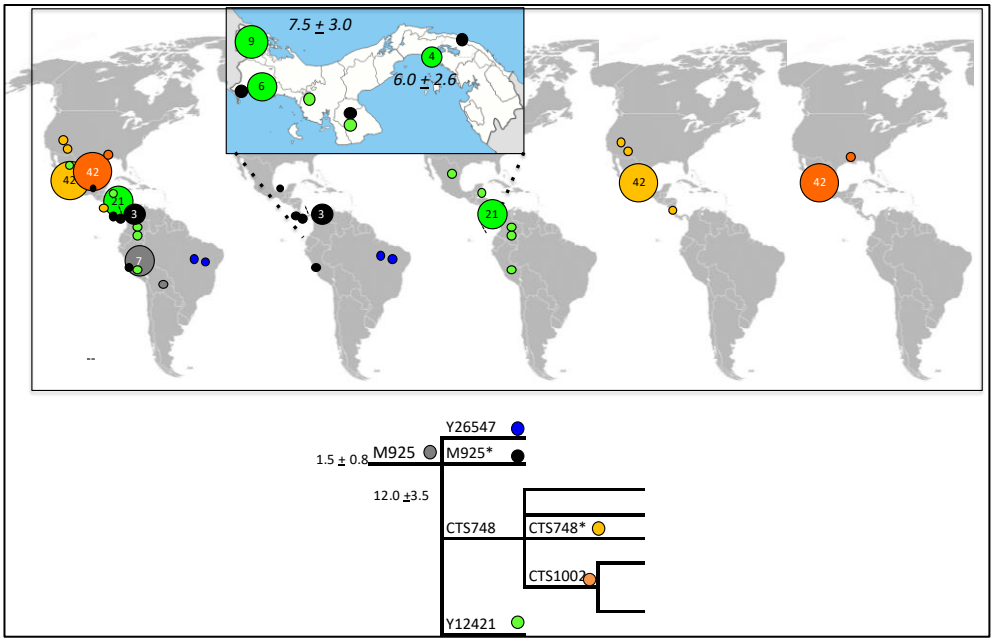


Figure 29. Phylogeography of the Q-M925 sub-clade. The figure illustrates the phylogenetic relationships of the markers (in different colours) investigated per each clade and their pattern of distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols.

Q-Z5906 and **Q-Z5908** (Fig. 30) display similar pattern of distribution from Mexico to Argentina with maximum frequencies in Peru where both of them show the greatest diversification including a local specific sub-clade: Q-M557 and Q-SA01, respectively. Z5906 is almost completely represented by CTS4000 observed at high frequency also in Bolivia. Although the clear expansion of both clades is in Peru, their origin should have occurred before enetering South America.

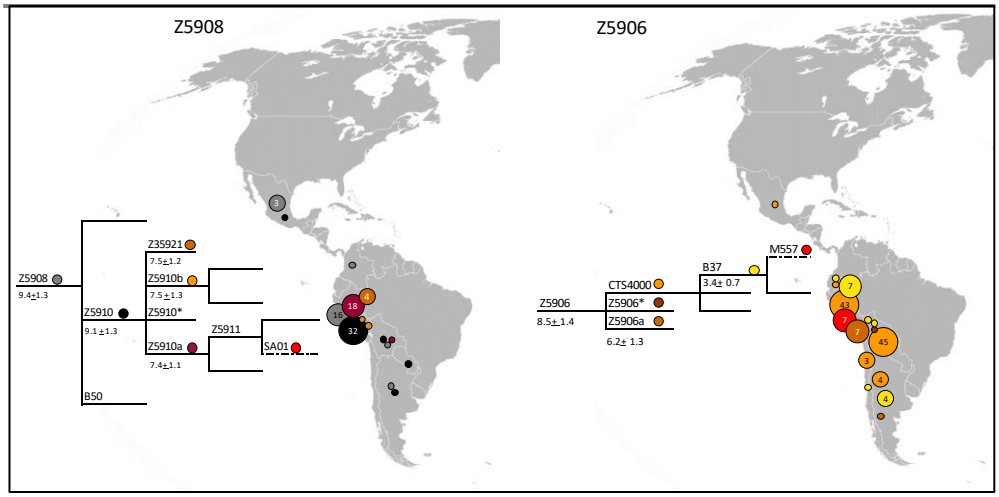


Figure 30. Phylogeography of the Q-Z5906 (A) and Q-Z5908 (B) sub-clades. The figure illustrates the phylogenetic relationships of the markers (in different colours) investigated and their pattern of frequency distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols. Dating refer to Bayesian estimation of node ages.

Within **Q-Z780** (Fig 31), the present level of resolution distinguishes three main groups of Y chromosomes: Q-Z781, Q-SA02 and Q-Z780*(x Z781, SA02). The first is the most represented, ancient (12.5±1.5 kya) and structured including two sub-lineages, Q-Z782 and Q-YP919 dated 3.1±1.1 kya and 9.6±1.4kya, respectively; Q-SA02, which seems to be restricted to the Isthmo-Colombian Area, is represented by few samples and dates back to 9.3±1.5. Chromosomes Q-Z780*(x Z781, SA03.2) are observed both in Meso-America and the Andean region and the analysis of their STR haplotypes (data not shown) highlights a wide variation, which suggests an ancient origin. The network of the STR haplotypes (data not shown) identifies at least two main high variable sub-lineages whose structure is far to be resolved.

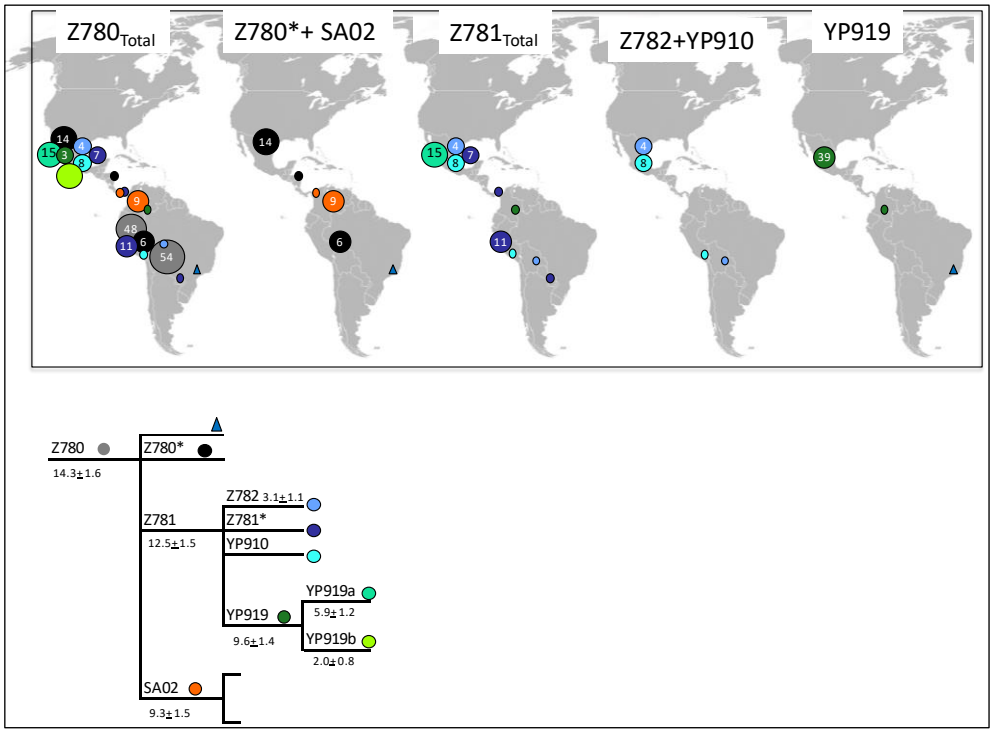


Figure 31. Phylogeography of the Q-Z780 sub-clade. The figure illustrates the phylogenetic relationships of the markers (in different colors) investigated per each clade and their pattern of distribution. Circles without any number refer to one subject. When more than one subject is from the same area, their number is reported inside symbols. Dating refer to Bayesian estimation of node ages

5.1.5 Bayesian analysis

Through a Bayesian method, the posterior distribution of the effective population size through time was estimated for the entire sample of Native Americans (all Q-M1107 samples) and for the samples belonging to the most significant sub-haplogroups described above. The overlap of the resulting curves is shown in figure 32. The trend of the dark blue curve, which represents all Native American Q-M1107 samples, shows a major phase of population growth after 15 kya, after the first entrance into the Americas, well represented by the orange curve (Z780 samples), followed by a period of constant population size from 8 to 4 kya and a little, even if not significant, sign of population growth from 3 kya. The second period of growth could be the one marked by M925 (brown) and Z5906 (green), while the curve of Z5908 (light blue), rapidly growing at 8 kya, seems not to have influenced the global trend of Native American populations. Interestingly, archaeological data found in South America recently published by Goldberg and collaborators show a similar trend: a first signal of growth linked to a resource-limited (megafauna extinction)

growth over time, then 9 kya domestication slowly started in NW South America until 3 kya when there was a shift to a predominantly sedentary and agricultural subsistence. To that time the second period of growth started, not linked to the climatic change but rather to a cultural and technological change. However, the presence in South America of divergent environments, geographic barriers to gene flow and low population density did not lead to the diffusion of a single cohesive culture. As evident from our analyses it seems that the cultural and technological revolution reached at that time was isolated and different for every population and, for this reason, not able to cause an evident expansion of the South American populations.

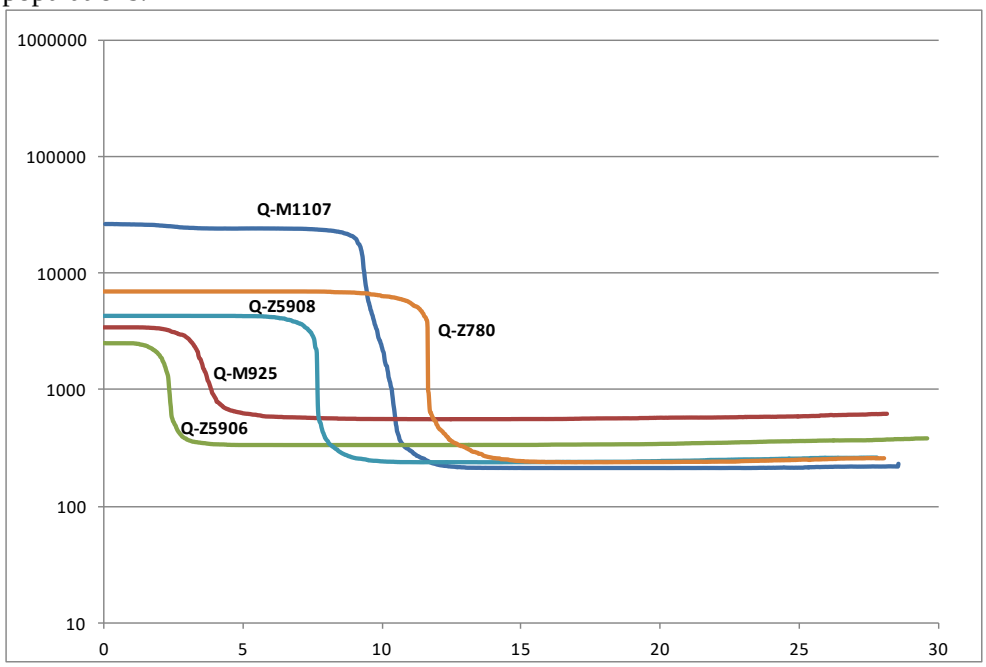


Figure 32. Bayesian skyline plots (BSPs) of Native American Q-M1107 samples. Hypothetical effective population size (N_e), Y axis, through time, X axis (it is limited to 25 ky, beyond which the curve remains flat) of the different sub-haplogroups.

5.2 Genome-wide haplotype analysis in Italian populations

5.2.1 Overview

AMHs inhabited the Italian territory since the Upper Palaeolithic, as attested by archaeological and anthropological remains in caves as the Grotta di Fumane (Cave of Fumane) in the Pre-Alps of Veneto region (Benazzi et al., 2015) and Grotta del Cavallo (Horse Cave) in Apulia, where among the oldest (dated ~ 45-43 kya) European skeletal remains were found (Benazzi et al. 2011; 2015). During LGM, around 30-19 kya, when large part of Europe was covered with thick ice sheets, Italy and the Balkans were partly forested and probably acted as refugia for Europeans escaping from the North, until the end of the LGM around 11 kya (Martini, 2013).
Agricultural revolution in Neolithic period (~ 6 kya) spread all over Italy, introducing substantial differences (especially between North and South) in ceramics industry and archaeological remains. These first influences from Levantine area started in South Italy in the regions of Apulia, Calabria and Sicily. Likely two migratory events along the Italian peninsula introduced the agriculture: one started from the south-east (Apulia) and followed an eastern coastal route, while another from East Sicily spreads along the Tyrrenian Sea reaching Sardinia, Corsica and Tuscany (Pessina and Tinè 2008).
Many culture successions characterized the Post-Neolithic ages and were added to the Palaeolithic and Upper-Palaeolithic layers of the previous periods.
The use of metals from Caucasus and Anatolia region became quite common in Italy as the Remedello Culture in North Italy confirms (Pessina and Tinè 2008).
During the Bronze Age, groups of Yamnaya elders introduced horse domestication and wheeled vehicles in Europe and Italy, this allowed to shorten the distances between peoples and ease the connections between cultures (Anthony, 2007). The impact of these groups on the past European and Italian cultural layers was strong and lead to the development of many widespread small communities with their own identity (Pessina and Tinè 2008). In Italy, signs of post-Neolithic settlements are present in different areas for example, the Nuraghe stones in Sardinia, the Polada Culture in Southern-Piedmont dated to the early Bronze Age, and the Villanovan Culture, mainly spread in the Tuscan-Emilian Apennine area during the Iron Age (Bietti Sestieri, 2010). All these cultures predated the long period of Roman Empire (27 BCE–476 CE), that left almost unchanged the Italian cultural background of the Bronze and Iron Age; only subsequently a series of migratory waves as the Barbarian Invasion (300 CE – 800 CE) and Arabic rule (827 CE – 1091 CE) reduced the population size of Italy and introduced new factors that left the signs still in the present Italian populations (Taylor J, 2003).
The first attempts to interpret the pattern of genetic variation of the Italian Peninsula were made by using classical polymorphisms, including the ABO blood group. Piazza et al. (Piazza et al. 1988) studied the frequencies of 34 “independent” alleles (ABO, MNS, KELL, RH, HP and 24 HLA alleles) and their Principal Component (PC) analyses. and synthetic maps of the first three PCs are generally considered the foundations for the Italian “genetic history” reconstruction (Fig. 33). These initial

results revealed a very distinct pattern for Sardinia, which is located far away from all other Italian regions, and represents an outlier in the European genetic landscape. When Sardinia was excluded from the PC analysis, the first principal component synthetic map, accounting for the 27% of the total variation, highlighted a North to South gradient, with similarities between Northern Italy and Central Europe, in contrast to the affinities between Central and Southern Italy with Greece and other Mediterranean populations. In this analysis Italy was placed in a peripheral position in the major European clusters, like the Iberian Peninsula.

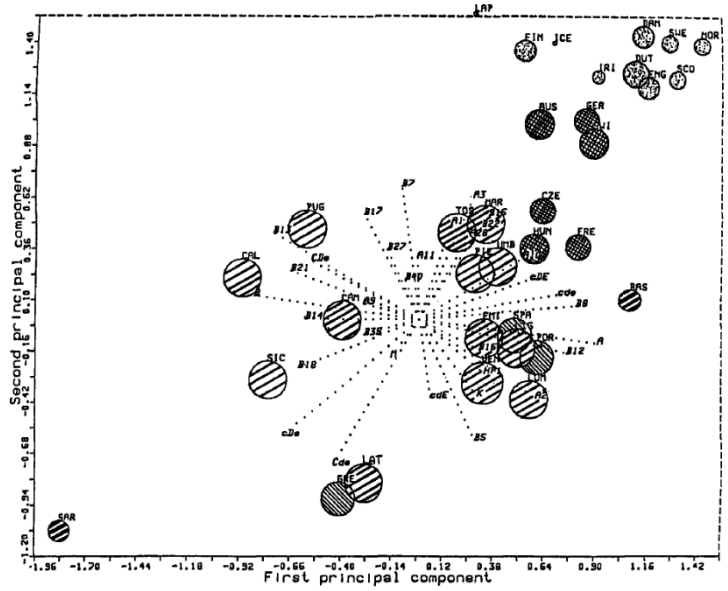


Figure 33. **A synthetic view of how the gene frequencies differ within Italy and Europe.** (From Piazza et al., 1988).

Later on, with the introduction of uniparental markers in population genetics, numerous studies tried to investigate the genetic structure of modern Italian populations from either the maternal or paternal side (Barbujani et al. 1995; Capelli et al. 2007; Destro Bisol et al. 2008; Di Giacomo et al. 2003) and, in some cases, comparatively (Boattini et al. 2013; Brisighelli et al. 2012a; Brisighelli et al. 2012b). Overall, these studies suggested that the genetic structure of modern Italy reflects, at least in part, the ethnic stratification of pre-Roman times.

In the last years, four relevant studies analysed the whole genome structure of the Italian populations (Fiorito et al., 2015; Parolo et al., 2015; Sarno et al., 2017; Sazzini et al., 2016). They all presented data covering many but not all of the administrative regions of Italy. Parolo et al., (2015) combined genetic analysis with biological insights and provided an evaluation of complex disease risk at the population level; the analyses were done using allele frequencies on a vast dataset that comprise only

Italian samples. Subsequently, on the wicker of Parolo et al., (2015), Sazzini et al., (2016) published a research that had the aim of discover peculiar patterns of population structure and local adaptations responsible for heterogeneous genomic background of present-day Italians. Both Parolo et al., (2015) o and Sazzini et al. (2016) identified a North-South cline possibly shaped by immune and inflammatory functions.
Some deeper insights into the history of the Italian population have been provided by Fiorito et al. (2015) that, through the implementation of a haplotype-based method as the one used in this thesis, identified five main Italian clusters on a dataset of 300 individuals. Two years later, Sarno et al. (2017), published a research focused on a big data set of Sicilian and Southern Italian populations, including also ethno-linguistic minorities; the results identified two Post-Neolithic ancestries linked to Caucasian and Levantine regions, compatible with maritime Bronze-Age migrations. Overall, these investigations provided evidence for within country variation and an indication of demographic events which might have contributed to such diversity. However, an extensive survey of Italy variation and a detailed description of its structure is still missing.

5.2.2 The dataset

To compile a dataset sampling in an exhaustive way the great majority of the Italian population variation, I started considering the available data from previous studies (Fiorito et al., 2015; Parolo et al., 2015). Since not all the administrative regions (Fig. S1) were equally represented in such datasets, we combined samples from the Pavia DNA repository with that avaibale from other collaborators (Dr. Brisighelli F, Prof. Capelli C, Prof. Lancioni H, Dr. Montinaro F, Prof. Pascali V) for a final set of 166 samples with four grand-parents born in 16 out of 20 administrative regions of Italy (Fig. S1). These samples, together with six DNAs from Albania were newly genotyped by using two versions (1.2 and 1.3) of the Infinium Omni2.5-8 Illumina beadchip. Out of the 172 samples, 135 (129 Italian and 6 Albanian) passed the quality control (call rate > 0.99).
Data from these newly genotyped samples were assembled with other Italian individuals present in the literature and genotyped on the Illumina chip (Fiorito et al., 2015; Li et al., 2008; Parolo et al., 2015; The 1000 Genome Project Consortium 2015; Metspalu unpublished). As for comparison, data from samples representative of worldwide population genotyped with Illumina chip (Behar et al., 2010, 2013; Busby et al., 2015; Chaubey et al., 2012; Di Cristofaro et al., 2013; Hellenthal et al.,2014; Rasmussen et al.; 2010; Haber et al.,2013, 2015; Hodoglugil et al., 2012; Kovacevic et al., 2014; Kushniarevich et al., 2015; Li et al., 2008; Metspalu et al., 2011; Pagani et al., 2015; Paschou et al.,2015; Raghavan et al., 2013; Yunusbayev et al.,2011, 2015) were added. The initial dataset of 5,048 samples was cleaned for missing data for both genotypes (missing data higher than 0.02) and individuals (higher than 0.01) and for degree of relationship lower than third, obtaining a Low Density Dataset (LDD) of 4,852 (1,589 Italians) and 218,725 markers (Table S4). Out of the 1,589 Italian samples, 1,365 were assigned to one of the 20 administrative

regions according to the best information available in this order: the birthplace of the four grandparents (indicated by the suffix "_G") or the birthplace of the two parents ("_P") or the birthplace of the samples ("_B") (Fig. 34). The remaining 224 were either unassigned ("_U") or with mixed parents ("_M-F_").

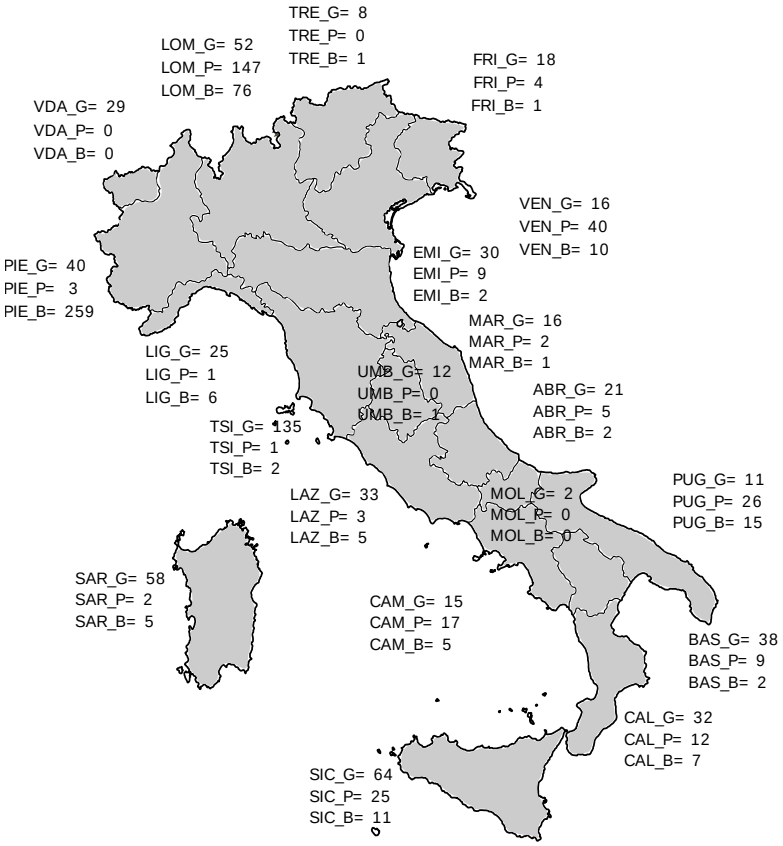


Figure 34. **Italian administrative regions with sample sizes indicated.** The labels at the end of each region represent the different birthplace information for group of samples (_G = grandparents; _P = parents or _B= sample birthplace)**.** The inset shows the name and location of the administrative regions.

A second dataset named High Density Dataset (HDD) was built with samples genotyped only with Illumina chip family "Omni" (Behar et al., 2013; Fiorito et al., 2015; Pagani et al.,2015; Paschou et al.,2014; The 1000 Genome Project; Yunusbayev et al., 2015). After the cleaning procedures it counted 1,651 worldwide samples (including 524 Italians covering all 20 administrative regions) and 591,217 markers (Table S4).

5.2.3 PCA analysis of allele frequency in Italian and European context

Two PCAs were drawn using the LDD allele frequencies, one of all the Italian samples and one including also Western Eurasian samples, respectively (Fig. 35 and Fig. 36).

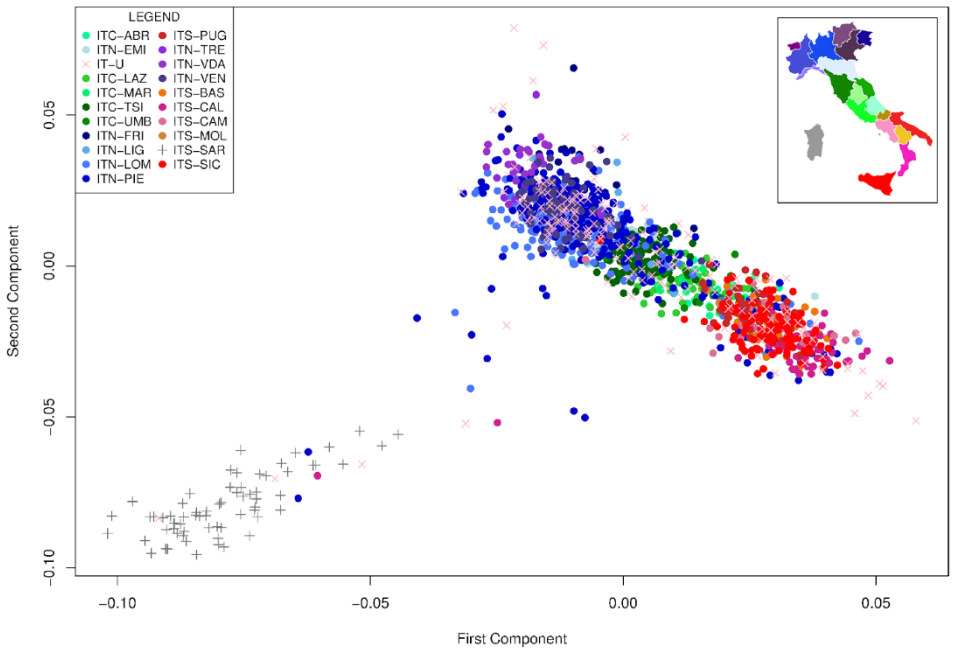


Figure 35. Genome-wide principal component analysis of the Italian samples. Each individual is depicted by a symbol and a colour representing the administrative region of origin (inset map). The inset map provides a key to the labels.

The distribution of the Italian samples in figure 35 is in line with those by recent investigations (Fiorito et al., 2015; Parolo et al., 2015; Sazzini et al., 2016): a well-defined North-South cline is evident comprising all Sicilian and continental samples; Sardinians are separated from the other Italians, possibly as the result of long term isolation (Cavalli-Sforza et al., 1994).
Discordances between geographic origin and position in the plot are observed for a few samples. For example, some samples occupy intermediate positions between the peninsula and the Sardinians, whereas some North-Italian (ITN-) individuals cluster together with South-Italians (ITS-). These discrepancies could be easily explained by recent admixture event (see the section 5.2.2 and Fig. 44) and by south-north migrations occurred in the recent years. It is worth noting here that comparing between information about birthplaces of individuals, of parents and grandparents is particular informative to highlight the possible presence of recent event of admixture and migration; for example, discordant clustering of samples with recent origin

information (birthplace of parents and samples) in groups of subjects for which the birthplace of the grandparents all born in the same region was known, could be possibly attributed to a recent event of migration of the first subjects; therefore implying their real origin as the same of the individulas for which the most ancient origin information was known (birthplace of the grandparents).
When I analysed Italian individuals in a wider context (Fig. 36), it was possible to appreciate the genetic diversity of the Italians compared to the other Eurasian regions.

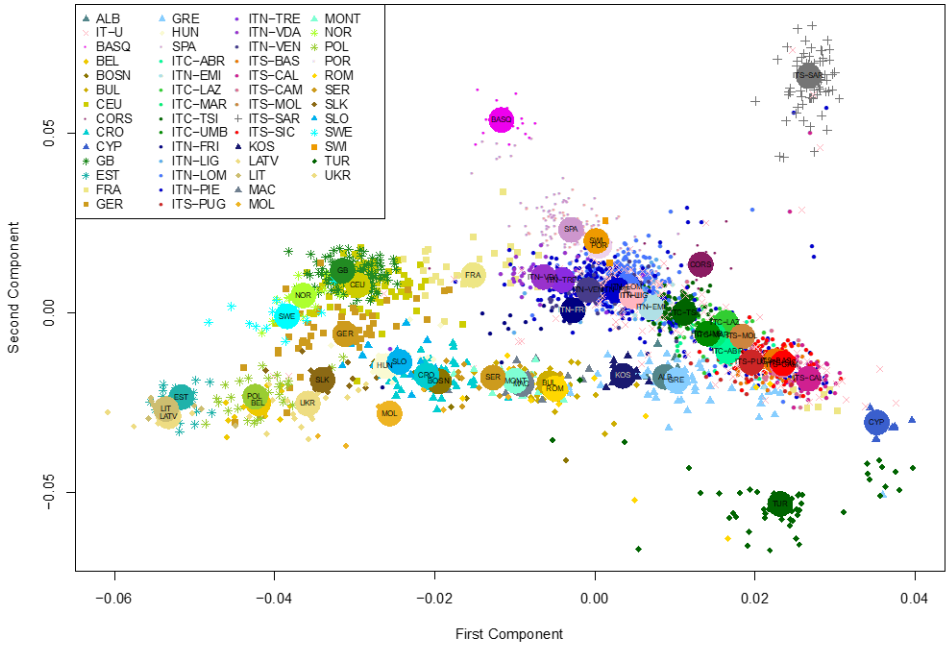


Figure 36. Genome-wide principal component analysis of 1,589 Italians and 983 Eurasians. Small coloured symbols represent individuals whereas large coloured circles represent median PC1 and PC2 values for each country. ALB, Albania; BASQ, Basque; BEL, Belgium; BOSN, Bosnia-Herzegovina; BUL, Bulgaria; CEU, Central Europe; CORS, Corsica; CRO; Croatia; CYP, Cyprus; GB, United Kingdom; EST, Estonia; FRA, France; GER, Germany; GRE, Greece; HUN, Hungary; SPA, Spain; KOS, Kosovo; LATV, Latvia; LIT, Lithuania; MAC, Macedonia; MOL, Moldova; MONT, Montenegro; NOR, Norway; POL, Poland; POR, Portugal; ROM, Romania; SER, Serbia; SLK, Slovakia; SLO, Slovenia; SWE, Sweden; SWI, Switzerland; TUR, Turkey; UKR, Ukraine; for the Italian populations the same symbols and colours as in figure 35 are used.

Continental Italy is stretched between the Mediterranean area (Cyprus “CYP”) and West/Central Europe (Spain “SPA” and France “FRA”) with high affinity with Portugal (POR), Switzerland (SWI) and Corsica (COR). Interesting is the partial overlap between France (FRA) and Aosta Valley (ITN-VDA), as well as the spread of Friuli Venezia Giulia (ITN-FRI) individuals toward the Central Balkan area.

Sardinians (ITS-SAR) and Basques (BASQ) are well-known examples of genetically differentiated populations (Cavalli-Sforza et al., 1994; Novembre et al.,2008) and clearly diverge from the rest of the continental samples in our plot.
As observed in the PCA of figure 35 for Italian individuals, along the first PC a North-South cline is also observed for the Balkans with southern and northern populations closer to the Southern and Central Europe, respectively.

5.2.4 ADMIXTURE analysis of allele frequencies in a worldwide context

A clustering model for K = 2-18 using ADMIXTURE on Italian individuals combined with ~ 300 worldwide populations of the LDD was carried out. The results obtained for K=15 are shown in figure 37.

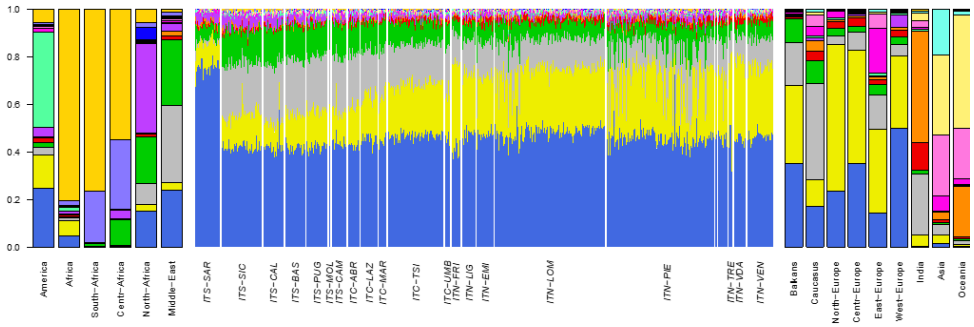


Figure 37. ADMIXTURE plot of individuals and populations at K = 15. For the Italian populations results for each individual (represented as a thin vertical bar) are displayed grouped for reference region; for the other populations the average across individuals is presented. The Italian samples are grouped in administrative regions as in figure 34.

Focusing on the Italian populations, it is possible to identify five major ancestral components, identified in the plot by the colours blue, yellow, grey, green and violet. These components are present in all the Italian regions, some of which showing some clinal North-South distributions. The distinctiveness of Sardinian population is confirmed. The blue component is present in all the Italian regions; the highest frequencies observed in Sardinia bing a remark of its genetic peculiarity, possibly related to its unique ancestry compostion and isolation history. The yellow component is modal in Northern and Central Europe. In Italy, it displays a north-south decreasing frequency distribution, reflecting the genetic affinity of this region with neighbouring Central-North European populations, as revealed also by the f3 statistics analysis (section 5.2.5). An opposite distribution is evident for the violet component, in which the highest frequencies are observed in the South, with decreasing values towards Northern Italy. This component, which is found at high frequencies in North Africa, might represent the legacy of the Arab rule in Southern Italy (Busby et al., 2015; Capelli et al, 2009; Fiorito et al., 2015; Sazzini et al. 2016). Finally, the grey and the green components are modal in Caucasian and Middle

Eastern groups. In Italy, they are slightly more represented in the Southern Italian regions, where they could represent results of recent interactions with populations from Eastern Europe or Western Asia.

5.2.5 Inferring signal of admixture on the Italian populations

Considering the f3 statistics analysis applied to the LDD, virtually all the Italian populations show signature of gene flow involving at least one extra European population. The most significant admixture signals are shown in figure 38 and involve five main sources. Three of them are from outside Europe, the Caucasus (Armenia), Middle East (Palestine) and North Africa (Morocco), while the other two are Europeans (Great Britain and Sardinia). These two are possibly associated to the pre-Neolithic and Neolithic European components, respectively (see chapter 2.4 "Peopling of Europe"). Great Britain (GBR) is the farthest contributing population for the first Middle Eastern (Palestine) and North Africa (Morocco) admixture sources whereas Sardinia, well known to be an Italian and European genetic outlier (Cavalli-Sforza et al., 1994) with a strong Neolithic affinity (Haak et al., 2015; Lazaridis et al., 2015), turned out to be the strongest contributing population for the Caucasus and the second Middle Eastern admixture sources.

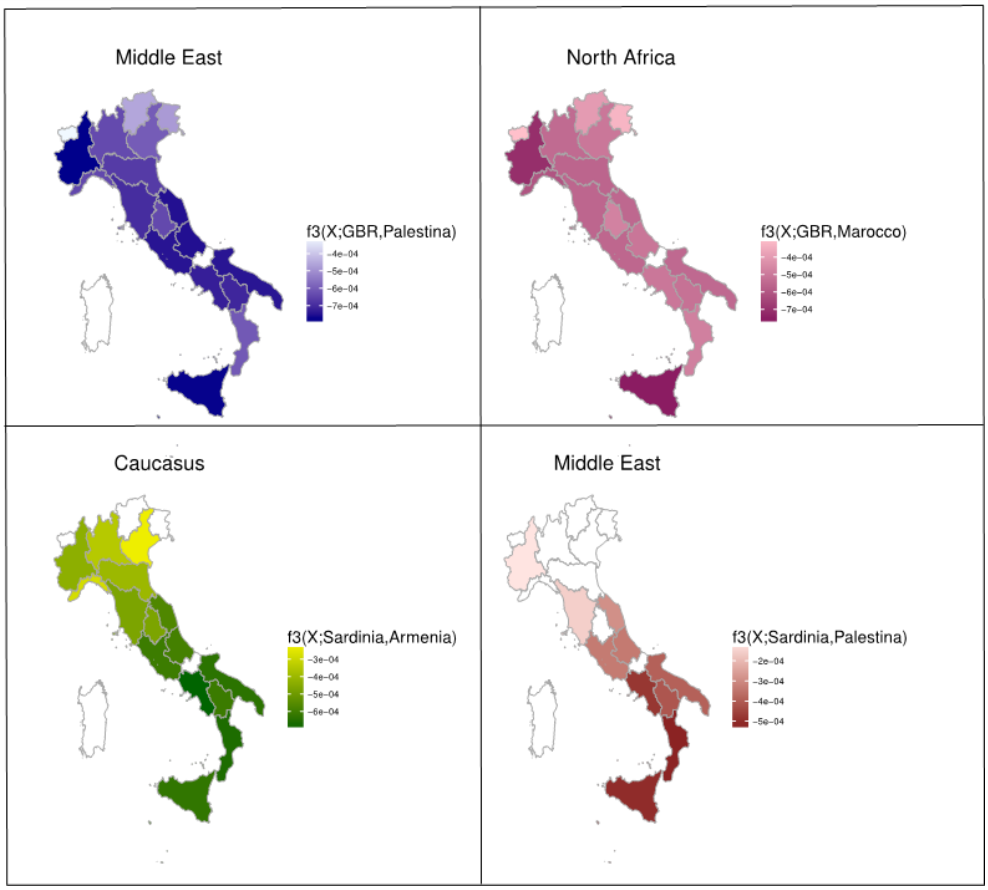
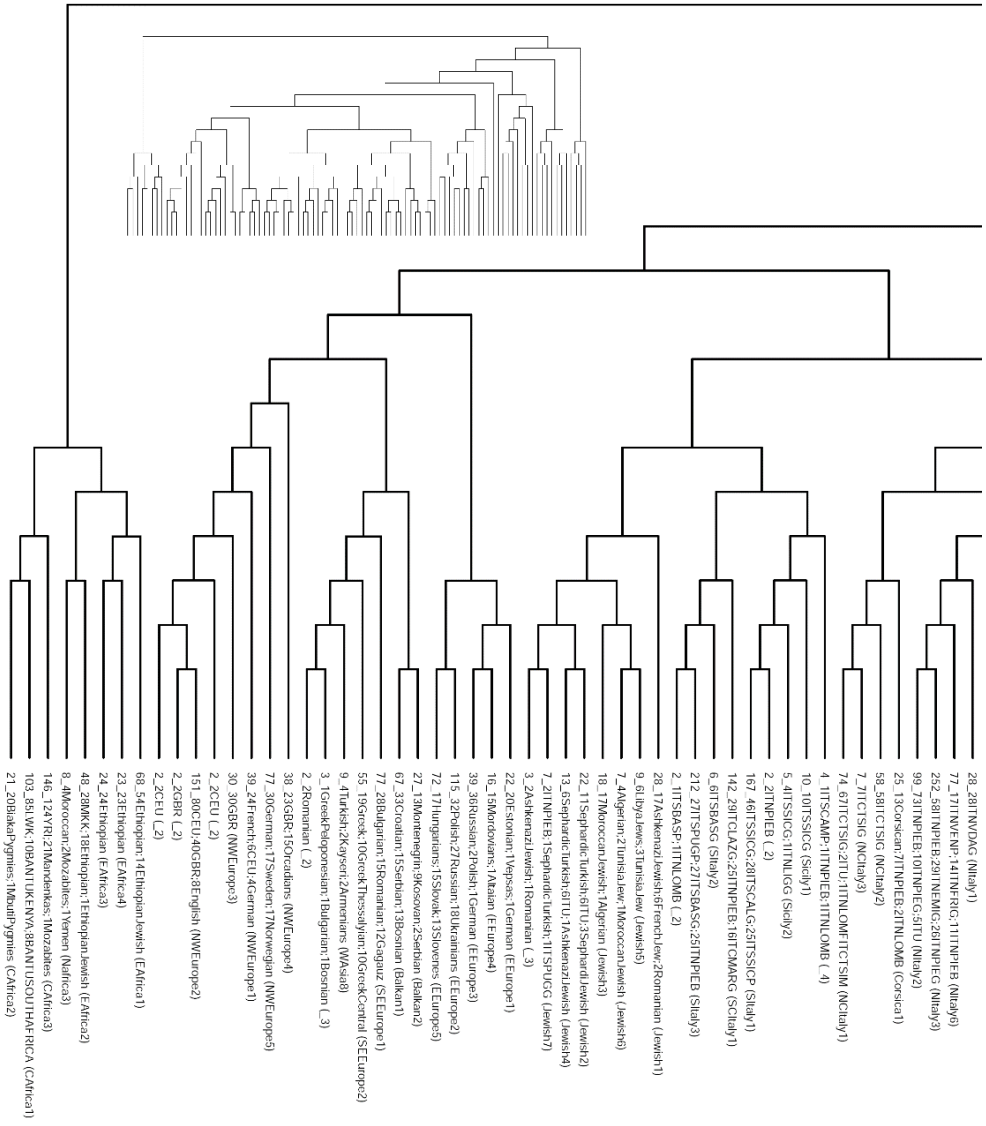


Figure 38. Distribution of the f3 statistic in Italy. The density of the colours displays the degree of admixture of each Italian population (X) with two source populations (A and B). For each panel the main title indicates the most likely source region while the legend the most likely admixture event for the corresponding populations.

The results align along a North-South cline. This is particularly evident when the Caucasus and the Middle Eastern f3 value distributions are considered, possibly reflecting a similar temporal and geographic source related to both for these events. On the other hand, the signal of admixture involving North African populations is stronger in Sicily. For this source, Piedmont is characterized by negative f3-values, comparable to those of Sicily and in striking contrast with neighboring populations. This is probably due to the fact that Piedmont was one of the regions most affected by migration from Southern Italy in the second half of the 20th century and only for a subset of the Piedmont samples the origin of the grandparents of the Piedmont samples was known (Fig. 34). This recent migration could also explain the low f3 value in Piedmont for the first Middle East admixture source. The large number of individuals from Piedmont included in the analyses has allowed me to further assess the genetic impact of this important socio-cultural phenomenon (Fig. 41).

5.2.6 Inferring clusters using CHROMOPAINTER and fineSTRUCTURE on the worldwide populations

The CHROMOPAINTER algorithm was used to paint all the individuals in the LDD and HDD previously converted in haplotypes with SHAPEIT software. Two haplotype-based co-ancestry matrixes were generated. They have been used for two separated fineSTRUCTURE analyses in order to evaluate the structure of the Italian population at different resolutions, the purpose in this case was the possibility to see more details by analysing the genetic structure with a higher number of variants and as well as compare the clusters of samples generated with the two datasets.
The dendrogram obtained with the LDD fineSTUCTURE haplotype based co-ancestry matrix is illustrated in supplementary figure 2.
On the whole, from the 4,853 samples of the LDD, 300 worldwide clusters were obtained. The cluster organization of the tree was modified to provide specific information on the Italian population. For clusters containing Italians and populations relevant to understand its structure (ADMIXTURE and f3 analyses) a finer resolution was maintained (splits closer to the leaves); and for the remaining samples a lower resolution approach was adopted was adopted (splits closer to the root). With this procedure I reduced the number of cluster to 95. The obtained simplified dendrogram is shown in figure 39.


28_28ITNVDAG (NItaly1)
77_17ITNVENP;14ITNFRTG;11ITNFHEB (NItaly6)
252_58ITNPIEB;29ITNEMIG;26ITNPIEG (NItaly3)
99_73ITNPIEB;10ITNPIEG;5ITU (NItaly2)
25_13CorsicanI7ITNPIEB;2ITNLOMB (Corsica1)
58_58ITCTSIC (NCItaly2)
7_7ITCTSIG (NCItaly3)
74_67ITCTSIG;2ITU;1ITNLOMFITCTSIM (NCItaly1)
4_1ITSCAMP;1ITNPIEB;1ITNLOMB (_4)
10_10ITSSICG (SIcily1)
5_4ITSSICG;1ITNLGG (SIcily2)
2_2ITNPIEB (_2)
167_46ITSSICG;28ITSCALG;25ITSSICP (SItaly1)
142_29ITCLAZG;25ITNPIEB;16ITCMARG (SCItaly1)
6_6ITSBASG (SItaly2)
212_27ITSPUGP;27ITSBASG;25ITNPIEB (SItaly3)
2_1ITSBASP;1ITNLOMB (_2)
28_17AshkenaziJewish;6FrenchJew;2Romanian (Jewish1)
9_6LibyaJews;3TunisiaJew (Jewish5)
7_4Algerian;2TunisiaJew;1MoroccanJewish (Jewish6)
18_17MoroccanJewish;1Algerian (Jewish3)
22_11SephardicTurkish;6ITU;3SephardiJewish (Jewish2)
13_6SephardicTurkish;6ITU;1AshkenaziJewish (Jewish4)
7_2ITNPIEB;1SephardicTurkish;1ITSPUGG (Jewish7)
3_2AshkenaziJewish;1Romanian (_3)
22_20Estonian;1Vepsas;1German (EEurope1)
16_15Mordovians;1Ataaln (FEurope4)
39_36Russian;2Polish;1German (EEurope3)
115_32Polish;27Russian;18Ukrainians (EEurope2)
72_17HungarIans;15Slovak;13Slovenes (EEurope5)
27_13MontenegrIn;9Kosovan;2Serbian (Balkan2)
67_33Croatian;15Serbian;13Bosnian (Balkan1)
77_28Bulgarian;15Romanian;12Csaguz (SEEurope1)
55_19Greek;10GreekThessalyian;10GreekCentral (SEEurope2)
9_4Turkish;2Kayseri;2Armenians (WAsia8)
3_1GreekPeloponnesian;1Bulgarian;1Bosnian (_3)
2_2Romanian (_2)
38_23GBR;15Orcadians (NWEurope4)
77_30German;17Sweden;17Norwegian (NWEurope5)
39_24French;5CEU;4German (NWEurope1)
30_30GBR (NWEurope3)
2_2CEU (_2)
151_80CEU;40GBR;8English (NWEurope2)
2_2GBR (_2)
2_2CEU (_2)
68_54Ethiopian;14Ethiopian Jewish (EAfrica1)
23_23Ethiopian (EAfrica4)
24_24Ethiopian (EAfrica3)
48_28MKK;18Ethiopian;1Ethiopian Jewish (EAfrica2)
8_4Moroccan;2Mozabites;1Yemen (NAfrica3)
146_124YRI;21Mandenkas;1Mozabites (CAfrica3)
103_85LWK;10BANTUKENYA;8BANTUSOUTHAFRICA (CAfrica1)
21_20BiakaPygmies;1MbutiPygmies (CAfrica2)

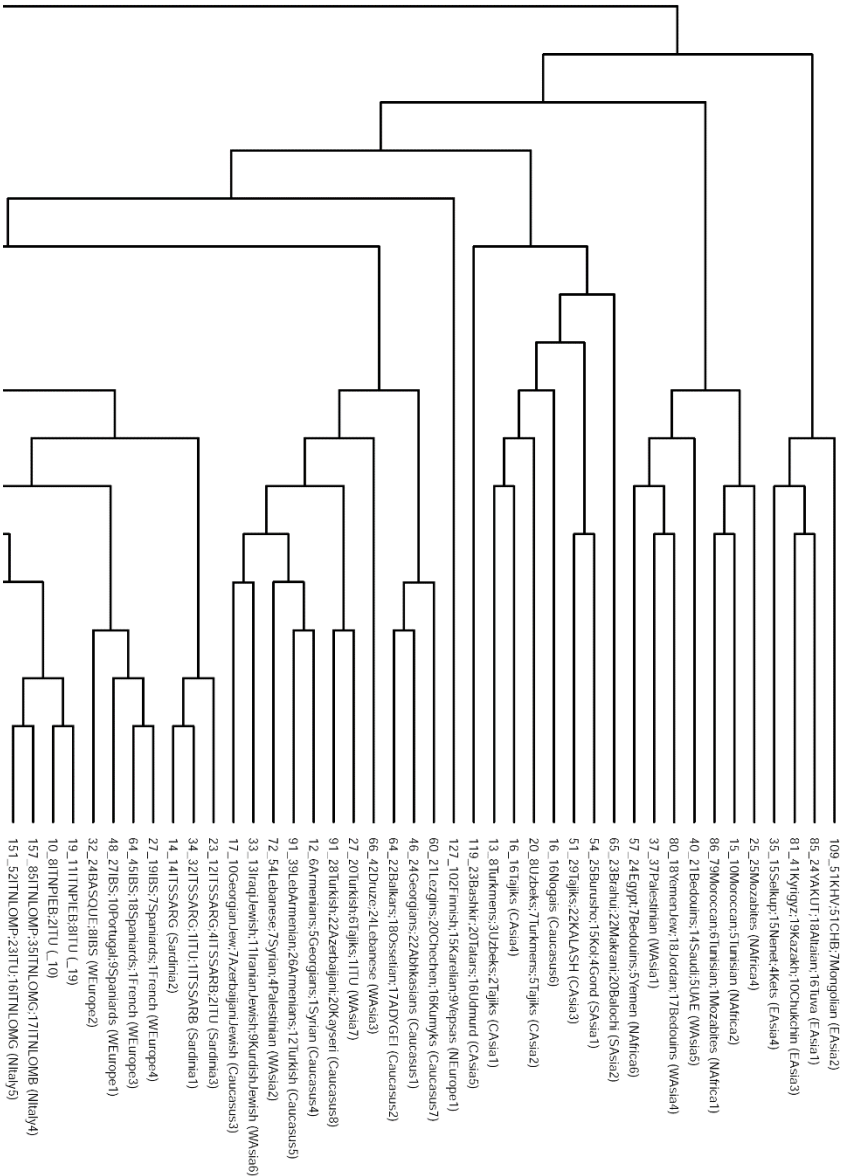


Figure 39. Simplified LDD dendrogram after visual inspection. Each tip of the dendrogram is a group of individuals with similar copying vectors. Tip labels as follows: total number of samples "_" the name of the three most representative geographical-assigned population with their respective number of samples. In the parenthesis the name assigned to the cluster.

The same simplification was applied on the HDD original tree. The obtained dendrogram contains 60 clusters out of the original 100, representative of 1,651 individuals (fig. 40).

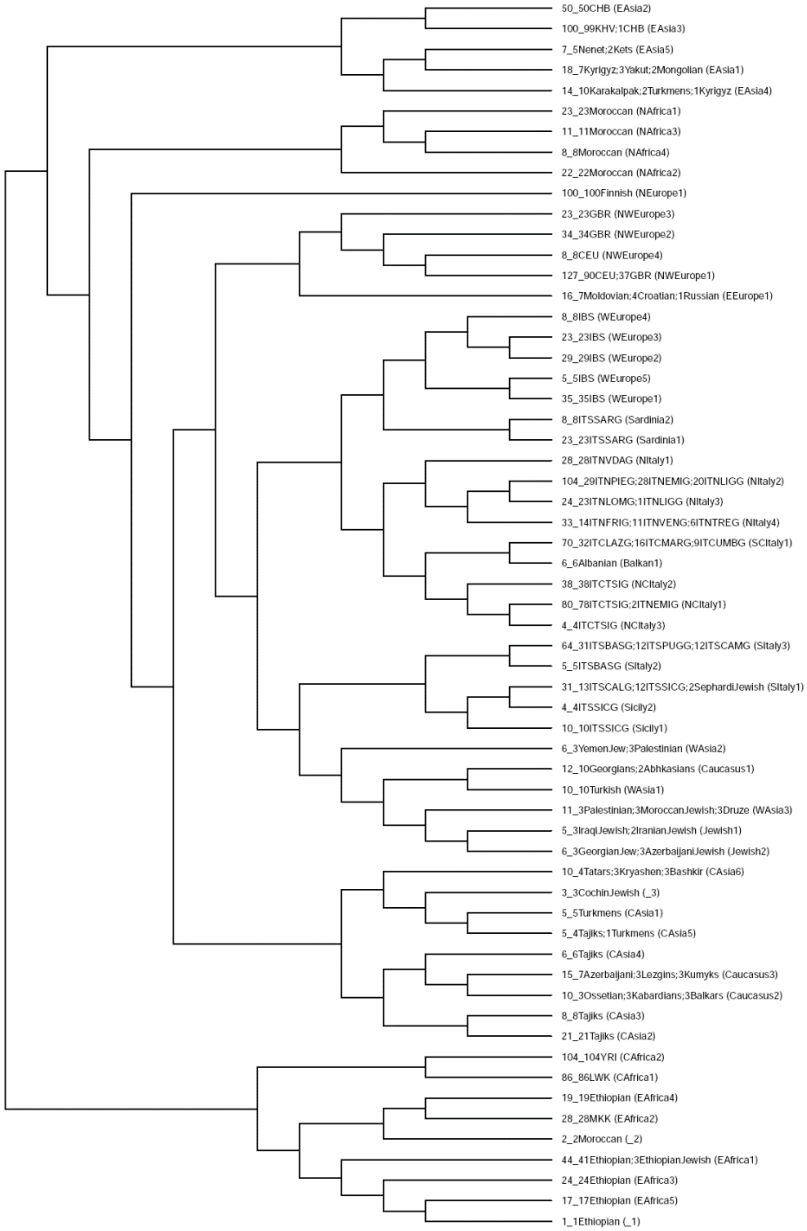


Figure 40. Simplified HDD dendrogram after visual inspection. Each tip of the dendrogram is a group of individuals with similar copying vectors. Tip labels as follows: total number of samples "_" the name of the three most representative geographical-assigned populations with their respective number of samples. In the parenthesis the name assigned to the cluster.

In order to have consistent genetic components, clusters with less than five individuals were disregarded (Busby et al., 2015). In addition, clusters containing only Italian samples with no ascertained origin (samples with the birth place of the parents "_P", the birth place of themselves "_B" or no information "_U") were discarded as well. In the simplified dendrograms (Fig. 39 and Fig. 40), these clusters are indicated between brackets by a number preceded by an underscore. Clusters with more than five individuals, instead, are indicated with the name of the macro-area from most of the samples are from and numbered sequentially from West to East. This scheme was applied also to the HDD.

After this pruning procedure, 84 of the 95 LDD and 57 out of 60 HDD simplified clusters were considered for further analyses.

Italian samples included in the LDD comprise subjects with known mixed ancestry (different regions of birth for the parents). To evaluate how mixed ancestry could influence the affiliation of a sample to a specific cluster, I compared the copying-vectors of individuals belonging to the same cluster, by focusing on the amount of copying vector that each individual copy within the same cluster. I named this analysis "cluster self-copy" (Fig. 41)

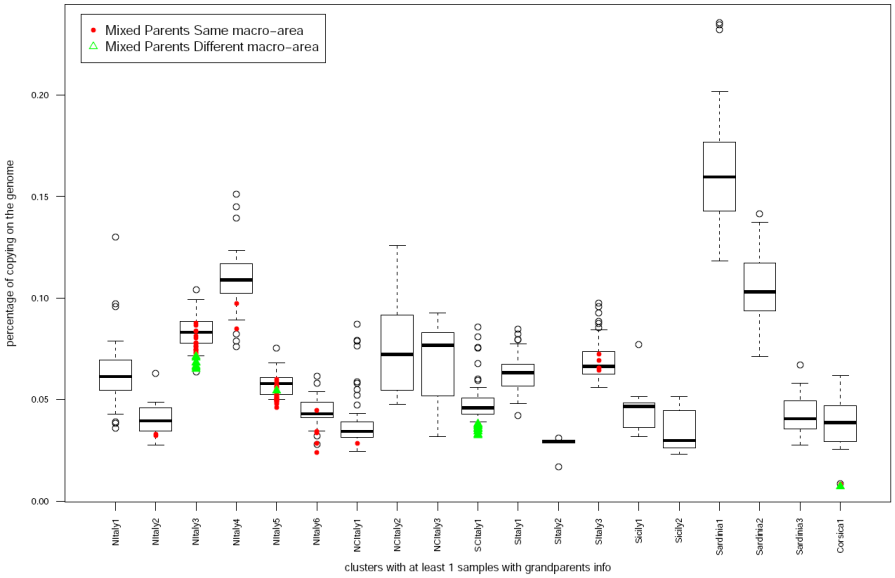


Figure 41. "Cluster self-copy" analysis. The boxplots represent, for each cluster, the distribution of the self-copying vectors of samples with ascertained origin (same birthplace region for the grandparents). Red and green symbols represent the copying vectors of samples with mixed ancestry (parents born in different adxministrative region). Triangles identified individuals with parents born in different macro-area (North Italy "ITN-", Central Italy "ITC-", South Italy "ITS-" present as suffix in each Italian population Table S2), while circles refer to samples with parents born in the same macro-area.

The boxplots show that (i) more the 90 % (21 out of 22) of mixed ancestry samples with parents born in different macro-area (North Italy "ITN-", Central Italy "ITC-", South Italy "ITS-") were outliers of the distribution; (ii) all the mixed ancestry

samples in the Corsica1 cluster turned out to be outlier. Since these results validated the efficacy of the approach, I applied this to the whole set of samples to identify individuals showing similar outlier behaviour when compared to samples with the four grandparents born in the same region. The boxplots distributions obtained for the different cases are illustrated in figure 42.

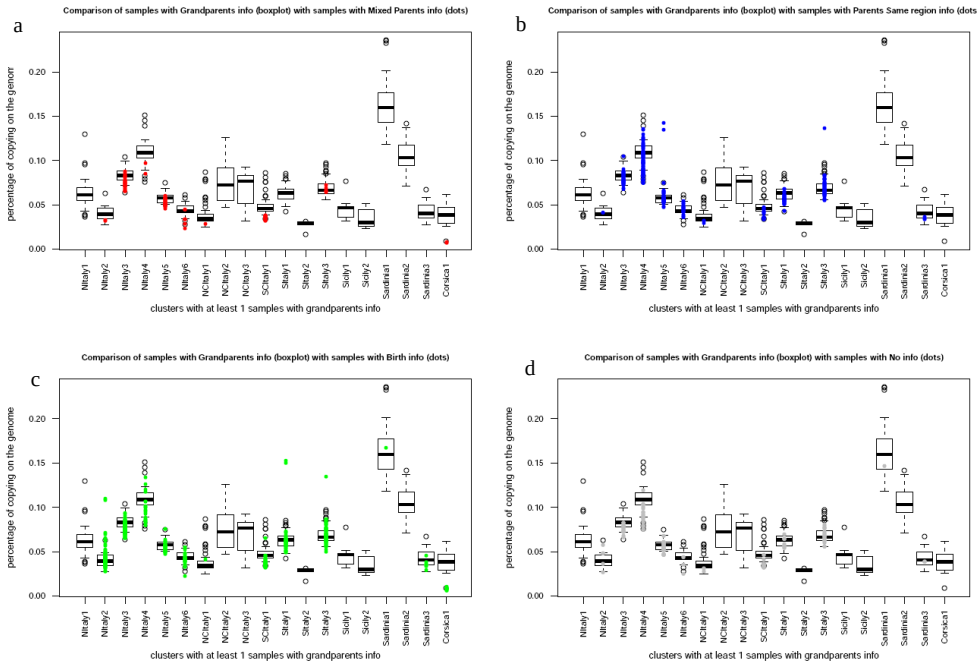


Figure 42. Self-copy analysis for all the Italians in the LDD. Boxplots representing, for each cluster, the distribution of the self-copying vectors of samples with ascertained origin (same birthplace region for the grandparents). (a) Red symbols, subjects with "mixed" parental ancestry (as described in figure 41); (b) Blue symbols, subjects with parent birthplace information; (c) Green symbols, subjects with their own birthplace information; (d) Grey symbols, subjects with no information available.

On the totality of the samples with mixed ancestry (two parents from different regions) 49 % were discarded (42 out of 86 subjects, of which 22 were from different marco-areas); predictably the percentage of subjects discarded when the parents were born in the same regions (12 %; 35 out of 304) is lower than the percentage of individuals discarded when the parents are born in different regions (which is 49% correct). Interestingly, when only place of birth of the samples was known, (21 %; 79 on 383) of the samples were removed. This reflects the fact that shallower in the genealogy information on the geographic origin is, less robust is the confidence in the resulting cluster composition. These observations are relevant for future sampling strategies as they suggest that, at least in certain areas of Italy, not having information on the parents place of birth might lead to spurios clustering assignation if no additional genetic analysis is conducted. The impact of recent historical events is also shown in figure 45. Of the 643 samples for which the four grandparents place

of birth was available, 14 (2%) were removed. Of the total of unknown samples 24 % (25 out of 104) were removed.

Considering the results plotted in figure 42, a total of 195 LDD samples were discarded from the further analyses, and are not considered anymore when referring to that cluster. Among these 195 subjects, 14 were part of the HDD and they were also from this dataset.

After having applied the above described filtering methods, a total of 21 LDD and 15 HDD clusters with Italian ancestry were identified (Fig. 43).

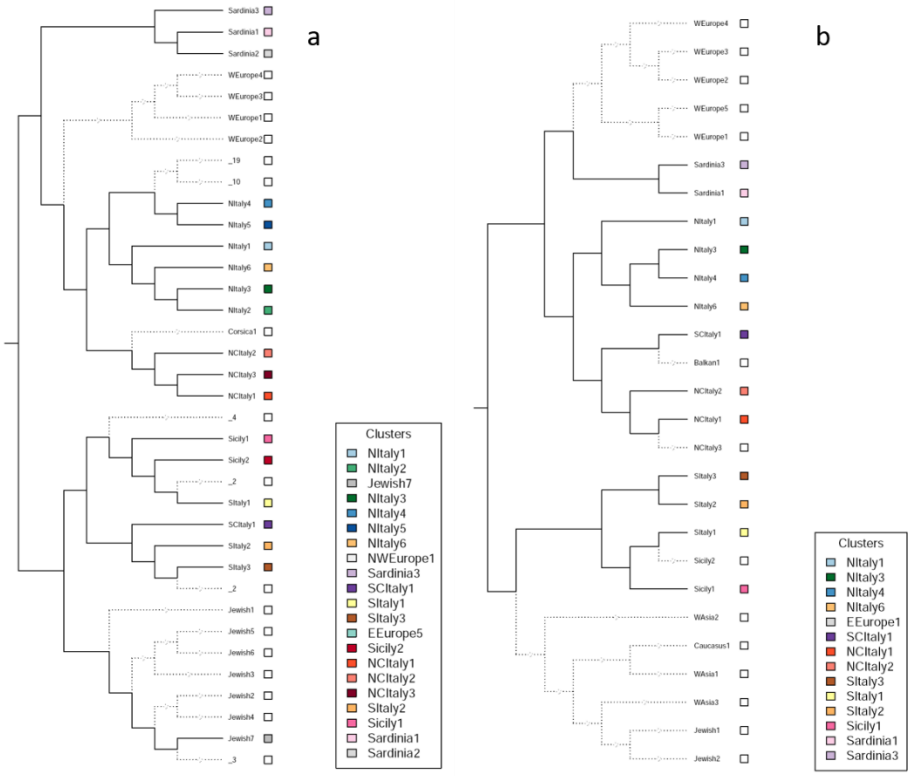


Figure 43. **Dendrograms of the main LDD and HDD genetic-based clusters**. Two LDD (NWEuope1 and EEurope5) (a) and one HDD (EEurope1) (b) cluster containing few subjects were not illustrated in the dendrograms due to representation limits.

The two LDD and HDD trees show a similar structure: Sardinian clusters grouped together but closer to the northern that to the southern Italian clusters; the latter grouped together and are well separated from the northern ones. The only exception is represented by the SCItaly1 cluster (purple): it is part of the southern Italian LDD clusters but of the northern Italian HDD clusters, where it groups with NCItaly 1 (brick red) and NCItaly2 (dark salmon). The presence of samples with "mixed ancestry" in the LDD fineSTRUCTURE analysis but not in the HDD fineSTRUCTURE analysis could have influenced the observed LDD tree topology.

However, this shift could also indicate a continuity of the peninsula genetic background or the meeting point between the two main Italian genetic components.
To test if the composition of the Italian clusters was consistent along the two datasets, the cluster distribution of individuals shared between the LDD and HDD was compared (Fig. 44).

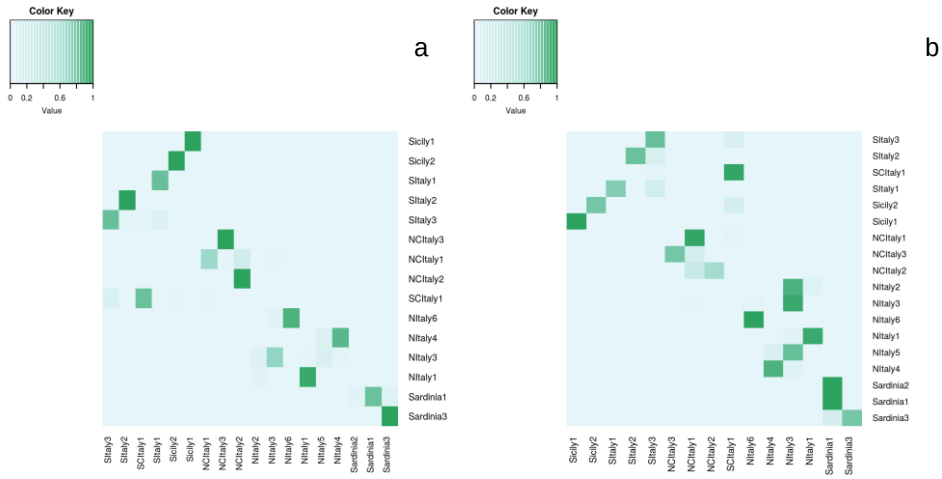


Figure 44. **Consistensy matrix**. The heatmaps represent the percentage of individuals shared within clusters of HDD (rows) on LDD (columns)(a) and LDD (rows) on HDD (columns) (b) datasets.

The heatmaps in figure 44 confirms the high efficency of the cluster analysis (fineSTRUTCTURE) by showing high level of similarity along the two datasets. The larger number of samples, and therefore clusters, on the LDD allows to see more splits than in the clusters of HDD (Sardinia 1, Sardinia 2, Sardinia 3 of LDD in Sardinia 1 and Sardinia 3 of HDD in Fig. 44-b).
Additionaly the LDD and HDD cluster distributions on the 20 administrative Italian regions are illustrated in figure 45.

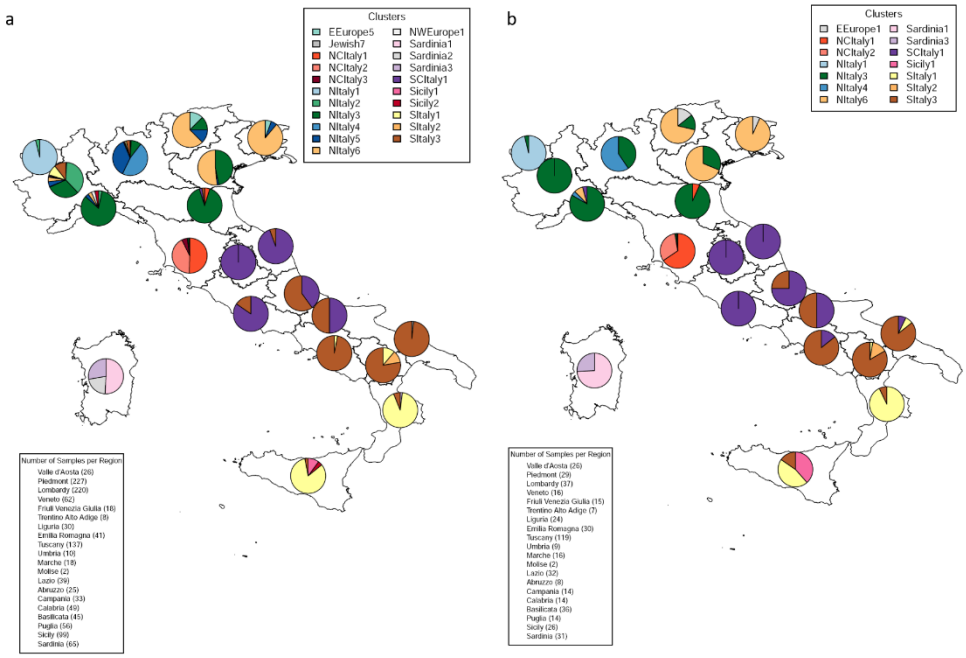


Figure 45. Pie charts summarizing the relative proportions of LDD (a) HDD (b) inferred genetic clusters. Cluster names are detailed in the inset at the top right. The numbers of samples according to their regional ancestry are in the inset at the bottom right

Considering only the main clusters the two distributions display high similarity. Both clearly recognise specific macro-areas clusters.

Namely, in North Italy it is possible to distinguish the following components: (i) the NItaly1 (light blue) characterizing genetically isolated individuals from the small region of Valle d'Aosta; (ii) the NItaly3 (dark green) shared by Piedmont, Liguria, Emilia Romagna and at lower frequencies by the north-eastern regions; (iii) NItaly6 (salmon) frequent in the north-eastern regions but not found in the neighbouring Lombardy and in the southern part of the Po valley. The single observation in Liguria likely reflects the origin of the carrier rather than more general population history.

Central Italy is characterized by a complex scenario: (i) NCItaly1 (brick red) and NCItaly2 (dark salmon), closely related (Fig. 43) turned out to be Tuscany-specific being observed outside this region at low frequency only in Emilia Romagna (ii) SCItaly2 (purple) represents the totality of the Umbria sample, the great majority of the Marche and Lazio samples and spreads southwards at decreasing frequencies in Molise, Campania and Apulia.

In contrast with the previous finds (Fiorito et al., 2015 and Sazzini et al., 2016), South Italy displays a clear genetic structure: (i) SItaly3 (brown) shows an opposite distribution pattern to the SCItaly1 (purple), presenting the highest frequencies in Campania, Basilicata and Apulia; (ii) SItaly1 (yellow) characterizes Sicily and

Calabria; (iii) SItaly2 (orange) and Sicily1 (fuchsia) are encountered only in Basilicata and Sicily, respectively.
Sardinia, once again, shows its distinctiveness being represented by specific clusters (Sardinia1 and Sardinia3), very well separated from all the other Italian clusters (Fig. 43-a and 43-b) but closer to the northern ones.
It is interesting to note here that the presence of individulas belonging to Southern Italian clusters in Northern Italian regions in the LDD disappear in the HDD. The samples that behave in this way are 58 (46 in Piedmont and 11 in Lombady and 1 in Emilia Romagna). Of these the totality have only the place of birth of the participant, which once again stress the relevance of collecting deeper genealogical informations.
Some considerations could be done after these preliminary results. First, the genetic isolation of regions as Aosta Valley and Sardinia clearly reflects the geographic location: one embraced by Alps and the other in the middle of the Tyrrhenian sea, Other geographical barriers could be at the basis of the observed cluster subdivision. For example, the Po River could have restricted the diffusion of NItaly6 (salmon) to the northern part of the Po valley and as well as could have played a major role also in the Tuscany cluster differentiation. Indeed, NCItaly1 (brick red) and NCItaly2 (dark salmon) clearly differentiated both in the LDD and HDD trees but fall into the northern genetic sub-structure. Similarly, the specific-small clusters of Basilicata (SItaly2, orange) and Sicily (Sicily1, fuchsia) likely are the result of genetic drifts caused by either geographical location, specific ancestries and historical contributions.

5.2.7 Fst Analysis

CHROMOPAINTER algorithm was run using as donors and recipients only Italian individuals who passed the filtering and cleaning procedures in both datasets.
FineSTRUCTURE was run on the co-ancestry matrixes obtained as output of CHROMOPAINTER. Genetic clusters generated by fineSTRUCTURE analyses with less than five individuals were discarded and a total of 26 clusters for the LDD and 13 for the HDD were identified. These clusters were used to infer pairwise Fst with smartpca software implemented in the EIGENSOFT package using a set of LD-pruned SNPs as explained in the 4.3 Method section.
The Fst distribution obtained for Italian clusters (blue) in comparison with those obtained with same procedure for United Kingdom (yellow; Leslie et al., 2015) and Spanish (pink, Bycroft et al., 2017 in preparation) populations, are illustrated in figure 46.

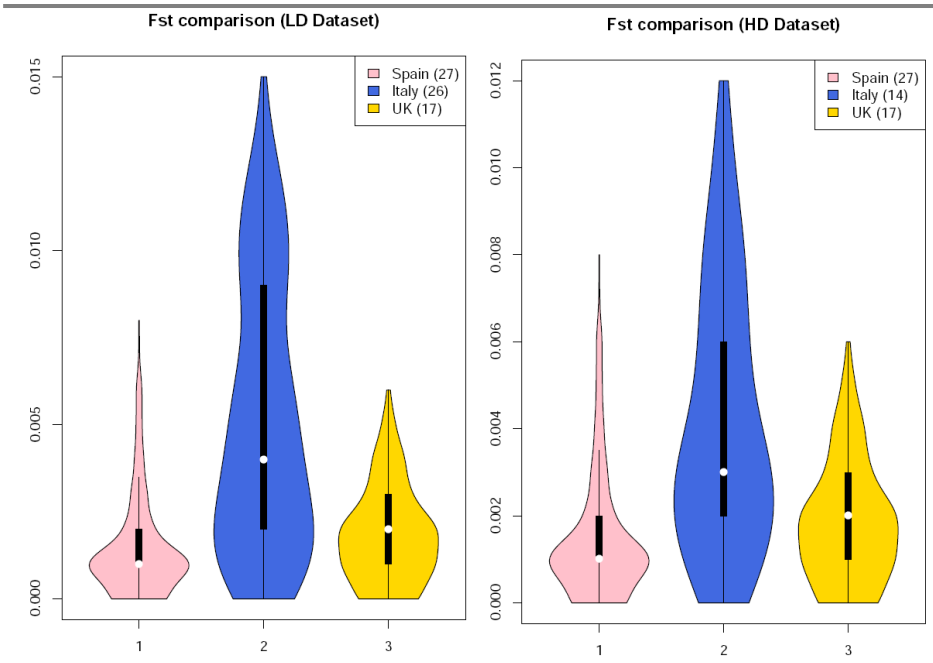


Figure 46. Italian (Italy) Fst values for LDD and HDD, compared with those of United Kingdom (UK) and Spanish (Spain) populations. The insets report the number of cluster in each population; white dots are median of the values and black bars are confidence intervals (CI).

The pairwise Fst estimates for the Italian populations (mean-LDD = 0.0051; mean-HDD = 0.0042) are more differentiated than those of the other countries (mean-UK = 0.0018; mean-Spain = 0.0014). The Wilcoxon rank sum test was considered to identify significative differences in the distributions, and the p-values for LDD Italians compared to UK and Spain was of $2.2e^{-16}$; the same test was also applied to the HDD Italians and I obtained a value of $1.597e^{-10}$ and $2.2e^{-16}$ for UK and Spain populations. This result is also confirmed by the high level of biodiversity recorded in Italy for plant and animal species in comparison with other European countries (UNEP-WCMC, 2004; Froese and Pauly, 2007; AmphibiaWeb. 2017; IUCN, Conservation International, 2008; Uetz et al., 2016).
Since the level of genetic diversity could be attributed to outliers included in the analysis, the pairwise Fst values that involved Sardinia, Orkney and Basque clusters were removed from the distributions (Fig. 47).

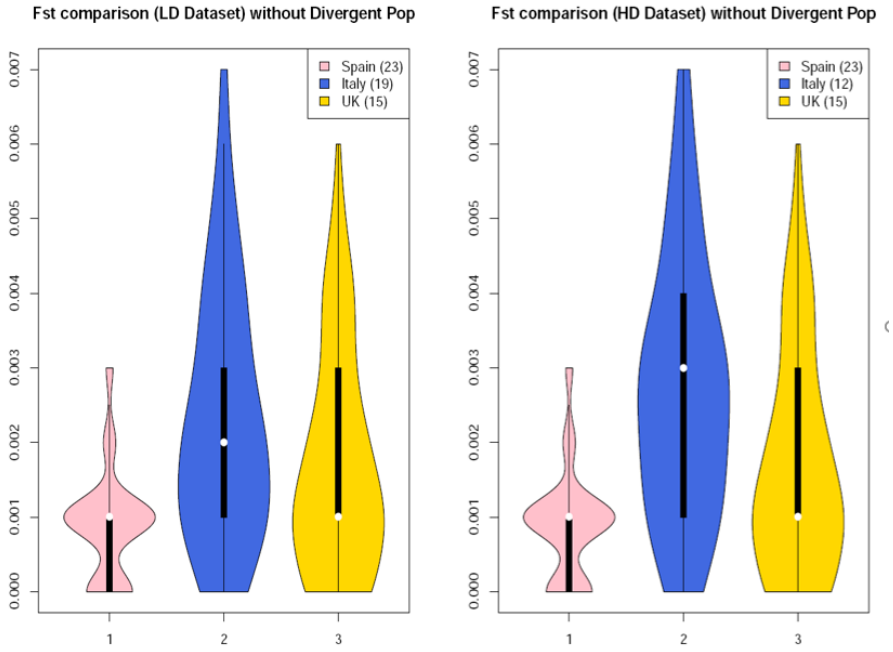


Figure 47. Italian (Italy), United Kingdom (UK) and Spanish (Spain) Fst distributions without divergent populations. The insets report the number of cluster in each population; white dots are median of the values and black bars are confidence intervals (CI).

Overall the results are consistent with the one of figure 46: greater variation is found in Italian populations with p-values of 0.00081 and $2.2e^{-16}$ for LDD samples in UK and Spain comparisons respectively and p-values of $3.986e^{-05}$ and $2.2e^{-16}$ in HDD. A subtler differentiation is present in UK and Spanish populations.
When considering Italian clusters on the LDD, besides Sardinian cluster comparisons, the highest Fst values are those involving the comparison between northern and southern clusters (Table S4).
To verify if the observed Italian F_{ST} genetic variation was in the range of the European one, preliminary comparisons were performed using Italians and European clusters generated by the simplified LDD dendogram from the first fineSTRUCTURE analysis. The obtained violin plots are illustrated in figure 48.

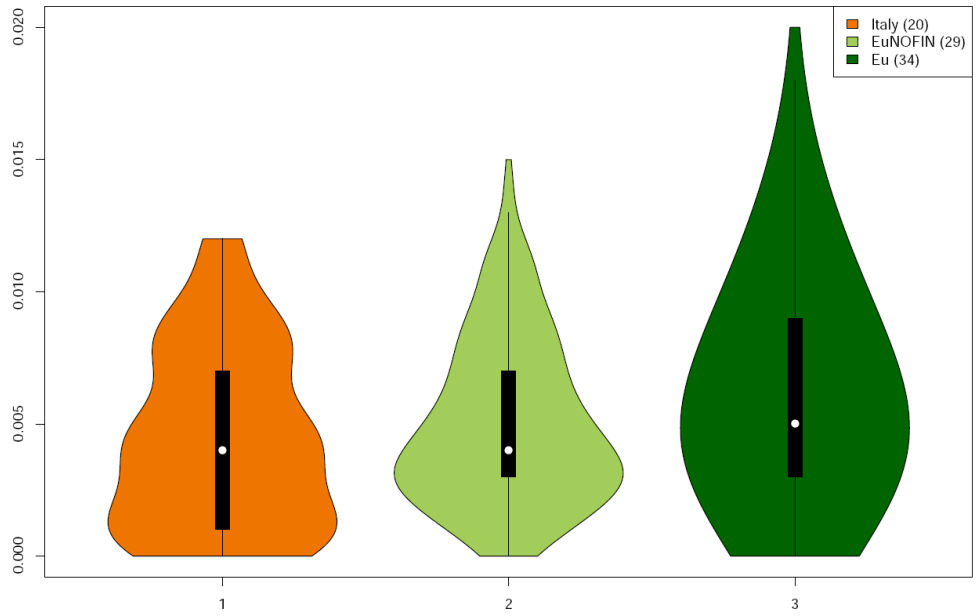


Figure 48. Pairwise Fst values distributions for European and Italian clusters. The clusters considered in this analysis are those generated by the simplified LDD tree of the first fineSTRUCTURE analysis. The inset represents the number of clusters considered and EuNOFIN includes all the European clusters but Finnish (considered outlyers, see simplified LDD dendrogram in figure 39).

Interstingly the distributions, when comparing the Italians and Europeans without the extreme outlier of the Finnish populations, remain statistically different (p-vale = 0.01097 after W-test) but strongly less different than the previously comparison.

6. Conclusion and perspective

6.1 Analyses of the NGS data on the South American populations

In this thesis work, a large portion of the MSY region was sequenced in 34 subjects belonging to haplogroup Q. These sequences were analysed together with other 120 available from literature reaching a total of 154, distributed all-over Europe, the Middle East, Asia and the Americas and 1,550 variant positions were identified. The tree, that illustrates the phylogenetic relationships among the 154 sequences based on the identified markers, clearly distinguishes the Y chromosomes of the New World from those of the old one. This is due to Q-M1107, which, together with Z780 and Q-M930, identify the two main American haplogroup Q lineages. However, unexpectedly, Q-M930, downstream of Q-M1107, reveals a closer relation than previously thought between the North European L804 and the most diffused Native American Q-M3 branch. On the whole, two Q-Z780 and eleven Q-M3 main sub-clades have been identified and their distribution investigated.

In conclusion, this work provides: (a) the most updated Hg-Q phylogeny, which clearly distinguishes Native American lineages from the Eurasian ones; (b) the phylogeographic distribution of all the new sub-lineages, which identifies the diffusion of the main Z780 and M3 lineages from North to the Southern Cone of America but also region-specific distributions of their sub-lineages; (c) a precisely estimate of the ages of the newly defined tree branches.

All this information is essential to compare genetic results with archaeological data and, possibly, also with linguistic and historical data.

6.2 Analysis based on allele frequencies and haplotypes of Italian population

These preliminary analyses confirmed the distinctiveness of Sardinia compared to the rest of Italy, while a more continuum distribution of genetic variation along a North-South axis was observed for the remaining samples. The results confirm the observations of previous works and highlighted a substantial degree of population structure within Italy, higher than other European populations and comparable to the overall variation reported across the continent, (Fiorito et al. 2014; Parolo et al, 2015; Sarno et al. 2017; Sazzini et al. 2016). The main components identified in the Italian nuclear genome include: (i) North-West and a North-East European similarities with North-West and North-East Italy, respectively; (ii) two Middle Eastern contributions, one likely due to the Neolithic demic expansion and the second to more recent migrations, mainly affecting the Southern and central Italy.

A strong component from Caucasus along the entire Italian peninsula was found starting from the ADMIXTURE plot. This component was previously identified by Sazzini et al., (2016), and it was explained by Sarno et al., (2017) as a post-Neolithic introgression in South Italy independent from the Yamnaya Culture (Late Bronze

Age) that instead influenced part of the Balkans. Therefore, these results, suggested different admixture histories for the Italian and Balkanic peninsula when compared with the rest of Central and Eastern Europe genetic history. These involved many events of migration carrying Caucasian and Levantine Mediterranean components, as the high level of Middle East recorded in my analyses suggest, which were likely associated to populations that occupied those regions during the period of the Bronze Age. It could be speculated here, given the current set of results that this contribution might be attribute to "Sea peoples", a culture, whose origin is considered to be in Asia Minor, that conquered and invaded many regions of the East Europe and Egypt in the Bronze Age period (Van De Mieroop et al., 2010).
F3 statistic analysis on the population of the LDD added new information regarding the sources and the possible admixture occurred in Italian populations, the observations, which are at the basis of the inference of the most likely source regions, could also suggest a temporal scale hypothesis of the observed admixture events (X;GBR,First Middle East before X; Sardinia, Caucasus/Second Middle East) to be evaluated with further analyses. Moreover, the high densitiy of color (high z-values) found in Southern Italy and associated with North Africa can be attributed to the Arab occupation of the region that occurred between the IX and XI centuries (Ahmad, 2010).

High order branching in the first fineSTRUCTURE clustering revealed a good correlation between genetics and geography. Additionaly the final clustering presented the South-North cline already observed by PCA and ADMIXTURE analyses with the distinctiveness of Sardinia associated with northern clusters. Four main components dissect northern regions; while in central regions, except for Tuscany, there are clusters found with high frequencies shared wtih southern regions. The lack of significant genetic sub-structures among Southern Italian groups (Sazzini et al., 2016) was confuted by the presence of different clusters that connect individuals from Calabria to individuals of Sicily and as well as Basilicata with Apulia. In addition, the presence, at low frequencies, of Southern Italian clusters in LDD north-western regions is worth of notice. This could be the result of the migratory events occurred in the '50s and '60s decades of the last century, which involved many family units. The same phenomena it is not found in the HDD map, which is based only on samples with ascertained origin.

The introduction of "self-copy" analyses allowed to estimate a confidency in assigning a sample to a genetic cluster. Indeed, when only the most recent information about the origin of the samples is available (place of birth of the samples themselves), the percentage of subjects that fall in the right clusters (clusters containing subjects with the most ancient origin information) was of the 79 %, while the remaining 21 % toghether with the others outlyers possibly show a cluster affiliation due to an extremely recent admixture events (1-2 generations). These preliminary results, might introduce an innovative approach for the acceptance of samples in population genetics and maybe a new method for sampling.

I compared Fst estimate distributions among haplotpe-based fineSTRUCTURE clusters considering only Italian, Spanish and United Kingdom subjects: the results clearly delineated a connection between the genetic diversity and the great variation found in plant and animal species of Italy if compared with other European countries.

This work assembled a large and exhaustive dataset comprising Italian and non-Italian subjects, which has been investigated to identify the genetic structure and that can be usefull to understand and identify migration and gene-flow events occurred in the last few thousand years. My future work will focus on the identfication and characterisation of these eevents to generate a map of the admixture history of the Italian population This additional work is expected to test some of the hypothesis here proposed and clarify the role of historical events in the generation of Italian genetic variation.

.

Appendix

Table S1. Classification of samples sequenced (chapter 5.1.2)

Sample ID	Source	Main Hg	Hg classification*	Macroarea	Country	Reference*
NA2	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274	Andes	Bolivia	Present study
NA1	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780, Y805, r289	Andes	Peru	Present study
NA15	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274	Andes	Peru	Present study
NA61	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M848b	Andes	Peru	Present study
NA59	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780	Andes	Peru	Present study
NA3	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, Y12421, Y12421a	C America	Panama	Present study
NA35	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, Y12421, Y12421a	C America	Panama	Present study
NA34	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	C America	Panama	Present study
NA58	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, CTS2731	Mexico	Mexico	Present study
NA12	modern DNA	Hg Q	M242, L472, M346, L53, M3, Y4303, Y4273	Mexico	Mexico	Present study
NA30	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, Z5906a, Z5906a1	SE America	Argentina	Present study
NA11	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, B42, B43	SE America	Brazil	Present study
NA23	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, B42, B43, B43b	SE America	Paraguay	Present study
NA13	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274	Andes	Peru	Present study
NA33	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5911, Z5912, Z5912a1	Andes	Peru	Present study
NA45	modern DNA	Hg Q	M242, L472, M346, L53, L330	Mongolia	Myangad, Khovd	Present study
NA46	modern DNA	Hg Q	M242, L472, M346, L53, L330, B287, B30	Mongolia	Myangad, Khovd	Present study
NA25	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, Z782, YP919, YP919a	Mexico	Mexico	Present study
NA31	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, Z782, YP919, YP919b	Andes	Colombia	Present study
NA5	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, SA02	Andes	Colombia	Present study
NA4	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780	Andes	Peru	Present study
NA62	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, SA02	C America	Panama	Present study
NA42	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780	Mexico	Valley Zapotecs	Present study
NA22	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, YP910	Mexico	Mexico	Present study
NA52	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780	Mexico	Mexico	Present study
NA53	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, Z782, YP919, YP919a	Mexico	Mexico	Present study
NA63	modern DNA	Hg Q	M242, L472, M346, B28, L940	Europe	Ukraine	Present study
EA38	modern DNA	Hg Q	M242, L472, F1096, F746, YP1500, YP1500a, YP1500a1	Siberia	Kamchatka-Koriaks	Present study
EA49	modern DNA	Hg Q	M242, L472, F1096, M25, L713, YP1677	Middle East	Iran-Lorestan	Present study
EA43	modern DNA	Hg Q	M242, L472, F1096, F746, M120	South East Asia	China	Present study
EA47	modern DNA	Hg Q	P36.2, L275, Y1150, YP4500	Europe	Bergamo, Lombardia	Present study
EA50	modern DNA	Hg Q	P36.2, L275, Y1150, YP4500	South Asia	India-Hindù	Present study
EA56	modern DNA	Hg Q	M242, L275, M378, BZ386	C America	Panama	Present study
EA51	modern DNA	Hg Q	M242, L275, M378, L245	Middle East	Iran-Golestan	Present study
GRC12133613	modern DNA	Hg Q	M242, L472, M346, L53, L330, YP1102	Asia	Kazakhstan	Balanovsky et al., 2017
bhu-1564	modern DNA	Hg Q	M242, L472, F1096, F1202, M120	South Asia	Bhutan	Hallast et al., 2014
bhu-1813	modern DNA	Hg Q	M242, L472, F1096, F1202, M120	South Asia	Bhutan	Hallast et al., 2014
eng-hgQ-1	modern DNA	Hg Q	M242, L472, M346, L53, L804	Europe	England	Hallast et al., 2014
eng-hgQ-2	modern DNA	Hg Q	M242, L275, M378	Europe	England	Hallast et al., 2014
GS18389	modern DNA	Hg Q	M242, L472, M346, L53, L330, B287, B30, B32	Asia	West Siberia	Karmin et al., 2015
GS18390	modern DNA	Hg Q	M242, L472, M346, L53, L330, B287, B30, B33	Asia	West Siberia	Karmin et al., 2015
GS18391	modern DNA	Hg Q	M242, L472, M346, L53, L330, B287, B31	Asia	West Siberia	Karmin et al., 2015
GS20363	modern DNA	Hg Q	M242, L472, M346, B28, Z5902, B285, B29	Europe	Bosnia-Herzegovina	Karmin et al., 2015
GS21412	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5910a	South America	Argentina	Karmin et al., 2015
GS21413	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, B47	South America	Argentina	Karmin et al., 2015
GS21414	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274	South America	Argentina	Karmin et al., 2015
GS21417	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, B42, B43, B44	South America	Argentina	Karmin et al., 2015
GS21418	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, B42, B43, B45	South America	Argentina	Karmin et al., 2015
GS21423	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274, B36	South America	Argentina	Karmin et al., 2015
GS27499	modern DNA	Hg Q	M242, L472, F1096, F746, M120, F745	South East Asia	Brunei	Karmin et al., 2015
GS27507	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, B42, B46	South America	Argentina	Karmin et al., 2015
GS27508	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780, Y805, r289	South America	Argentina	Karmin et al., 2015
GS27510	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, B50	South America	Argentina	Karmin et al., 2015
GS27511	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274, B41	South America	Argentina	Karmin et al., 2015
GS27516	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000, Z5907, r274, B40	South America	Argentina	Karmin et al., 2015
GS27530	modern DNA	Hg Q	M242, L472, M346, L53, M3, Y4303, B34	Asia	Northeast Siberia	Karmin et al., 2015
GS32361	modern DNA	Hg Q	M242, L472, F1096, F746, YP1500, B280	Asia	Northeast Siberia	Karmin et al., 2015
GS32362	modern DNA	Hg Q	M242, L472, F1096, F746, YP1500, B280	Asia	Northeast Siberia	Karmin et al., 2015
GS32363	modern DNA	Hg Q	M242, L472, F1096, F746, YP1500, B284	Asia	Northeast Siberia	Karmin et al., 2015
GS32368	modern DNA	Hg Q	M242, L472, F1096, F746, YP1500, B280	Asia	Northeast Siberia	Karmin et al., 2015
GS35460	modern DNA	Hg Q	M242, L472, M346, L53, L330, YP771	Asia	Uzbekistan	Karmin et al., 2015
GS35468	modern DNA	Hg Q	M242, L472, M346, B28, Z5902, B285, L717, L717a	South Asia	Nepal	Karmin et al., 2015
GS35483	modern DNA	Hg Q	M242, L472, F1096, M25, L713, YP1677a	Asia	Uzbekistan	Karmin et al., 2015
GS35484	modern DNA	Hg Q	M242, L472, F1096, M25, L713, YP1677a, B279	Asia	Uzbekistan	Karmin et al., 2015
GS35485	modern DNA	Hg Q	M242, L472, F1096, M25, L713	Asia	Uzbekistan	Karmin et al., 2015
HGDP00856	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, CTS1002	South America	Peru	Poznik et al., 2013
HGDP00877	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780	Meso America	Mexico (Maya)	Poznik et al., 2013; Raghavan et al., 2015
Athabascan_2	modern DNA	Hg Q	M242, L472, M346, L53, M3, Y4303, B34	Alaska	Alaska (Athabascan)	Raghavan et al., 2015
CEPH_11_D12	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, Z782, YP919, YP919b	Meso America	Mexico (Pima)	Raghavan et al., 2015
Ket1	modern DNA	Hg Q	M242, L472, M346, L53, L330, B287, B30, B33	Asia	Central Siberia (Ket)	Raghavan et al., 2015
Ket2	modern DNA	Hg Q	M242, L472, M346, L53, L330, B287, B31	Asia	Central Siberia (Ket)	Raghavan et al., 2015
TA6	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000	South America	Peru/Bolivia	Raghavan et al., 2015
Tsimshian	modern DNA	Hg Q	M242, L472, F1096, F746, M120	Alaska	Alaska (Tsimshian)	Raghavan et al., 2015
Y2040	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Colombia (Yukpa)	Raghavan et al., 2015
HGDP00998	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, Y26547	South America	Brazil (Karitiana)	Raghavan et al., 2015 (PRUFER 2014)
Saqqaq	ancient DNA	Hg Q	M242, L472, F1096, F746, YP1500	Greenland	Greenland	Rasmussen et al., 2010; Karmin et al., 2015

Anzick	ancient DNA	Hg Q	M242, L472, M346, L53, L54, Z780	North America	Montana	Rasmussen et al., 2014; Karmin et al., 2015
Kennewick	ancient DNA	Hg Q	M242, L472, M346, L53, M3	North America	Washington	Rasmussen et al., 2015
R1b-BT	modern DNA	Hg R1b	M343	Europe		Scozzari et al., 2014?
HGDP01012	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Brazil	Simons Genome Diversity Project Consortium, 2016
HGDP01015	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, Y26547	South America	Brazil	Simons Genome Diversity Project Consortium, 2016
HGDP01047	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, CTS1002, YP4722	Meso America	Mexico	Simons Genome Diversity Project Consortium, 2016
IRAN17	modern DNA	Hg Q	M242, L275, M378, L245	Middle East	Iran	Simons Genome Diversity Project Consortium, 2016
Mex02	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, CTS1002, Z768, Z768a	Meso America	Mexico	Simons Genome Diversity Project Consortium, 2016
Mex20	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781	Meso America	Mexico	Simons Genome Diversity Project Consortium, 2016
MIXA0099	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	Meso America	Mexico	Simons Genome Diversity Project Consortium, 2016
NA11200	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910	South America	Peru	Simons Genome Diversity Project Consortium, 2016
SS6004481	modern DNA	Hg Q	M242, L472, M346, B28, Z5902, B285, L717	South Asia	India	Simons Genome Diversity Project Consortium, 2016
SS6004486	modern DNA	Hg Q	M242, L472, F1096, M25, YP4385	South Asia	India	Simons Genome Diversity Project Consortium, 2016
TGBS21	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848,Y27992	South America	Argentina	Simons Genome Diversity Project Consortium, 2016
ZAPO0098	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, CTS2731, Y26491, Y26467	Meso America	Mexico	Simons Genome Diversity Project Consortium, 2016
ZAPO0099	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, CTS2731, Y26491, Y26467	Meso America	Mexico	Simons Genome Diversity Project Consortium, 2016
HG01124	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Colombia	The 1000 Genomes Project Consortium, 2015
HG01139	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, Y12421	South America	Colombia	The 1000 Genomes Project Consortium, 2015
HG01142	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, Y12421	South America	Colombia	The 1000 Genomes Project Consortium, 2015
HG01565	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, r285, r283, r281	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01892	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01920	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M848a	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01923	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, Z5907	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01926	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5915	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01938	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780, Y805	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01944	modern DNA	Hg Q	M242, L472, F1096, F1202, M120	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01950	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01961	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01967	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5915, Z5916	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01974	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, r285, r284	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01977	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780, Y805, r289	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG01979	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z35921	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02090	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02104	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02116	modern DNA	Hg Q	M242, L472, F1096, F1202, M120, Y515	South East Asia	Vietnam	The 1000 Genomes Project Consortium, 2015
HG02134	modern DNA	Hg Q	M242, L472, F1096, F1202, M120, Y515	South East Asia	Vietnam	The 1000 Genomes Project Consortium, 2015
HG02146	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, CTS4000	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02259	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02265	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5915, Z5916	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02271	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, r285, r283, r282	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02277	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5911	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02285	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5911, Z5912	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02291	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, Z5907, r274	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02299	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5911, Z5912, Z5912a	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02304	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, Z5906a, Z5906a1	South America	Peru	The 1000 Genomes Project Consortium, 2015
HG02696	modern DNA	Hg Q	M242, L472, F1096, M25, Y16849	Middle East	Pakistan	The 1000 Genomes Project Consortium, 2015
HG03652	modern DNA	Hg Q	P36.2, L275, Y1150, YP3943	Middle East	Pakistan	The 1000 Genomes Project Consortium, 2015
HG03681	modern DNA	Hg Q	M242, L472, M346, B28, Z5902	South Asia	Pakistan	The 1000 Genomes Project Consortium, 2015
HG03864	modern DNA	Hg Q	P36.2, L275, Y1150, YP4500	South Asia	India	The 1000 Genomes Project Consortium, 2015
HG03914	modern DNA	Hg Q	M242, L275, Y1150	South Asia	Bangladesh	The 1000 Genomes Project Consortium, 2015
HG03943	modern DNA	Hg Q	M242, L472, M346, Y2659, Z5902, B285, L717, L717a	South East Asia	Sri Lanka	The 1000 Genomes Project Consortium, 2015
NA19664	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19682	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, CTS1002, Z768	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19729	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, CTS2731	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19732	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, CTS1002, YP4722	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19735	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, Y10781	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19771	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, Z782	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19774	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, CTS2731, Y26491	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19783	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, YP4673	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19786	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M925, CTS748, CTS1002, Z768, Z768a	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
NA19795	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781, Z782	Meso America	Mexico	The 1000 Genomes Project Consortium, 2015
R1b-SUFG	modern DNA	Hg R1b	M343, M412, S116	Europe		Underhill et al., 2015
103-RQ	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910	South America	Peru	Zhou et al., 2013
109-EP	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5906, Z5906a	South America	Peru	Zhou et al., 2013
16-EJ	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781	South America	Peru	Zhou et al., 2013
31-CE	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5910b	South America	Peru	Zhou et al., 2013
40-JI	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5911, Z5912, Z5912a1a	South America	Peru	Zhou et al., 2013
43-AM	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5910b	South America	Peru	Zhou et al., 2013
50-JB	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Y780, Y805, r289	South America	Peru	Zhou et al., 2013
53-JR	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M848a	South America	Peru	Zhou et al., 2013
55-EC	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, M848b	South America	Peru	Zhou et al., 2013
56-AM	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910, Z5911, Z5912, Z5912a1a	South America	Peru	Zhou et al., 2013
72-DA	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780, Z781	South America	Peru	Zhou et al., 2013
76-AS	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910	South America	Peru	Zhou et al., 2013
77-JN	modern DNA	Hg Q	M242, L472, M346, L53, L54, Z780	South America	Peru	Zhou et al., 2013
7-PCE	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, r285, r283, r282	South America	Peru	Zhou et al., 2013
93-BS	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910	South America	Peru	Zhou et al., 2013
96-JP	modern DNA	Hg Q	M242, L472, M346, L53, M3, M848, Z5908, Z5910	South America	Peru	Zhou et al., 2013

Table S2. Classification of samples in the literature (The 1000 Genome Project Consortium, 2015; Balanovsky et al., 2017; Battaglia et al., 2013; Di Cristofaro et al., 2013; Dulick et al., 2012; Fornarino et al., 2009; Grugni et al., 2012; Hallast et al., 2014; Jota et al. 2016; Karachanak et al., 2013; Karmin et al., 2015; Lippold et al. 2014; Malyarchuk 2011; Nonak et al., 2007; Poznik et al., 2013; Raghavan et al., 2015; Rasmussen et al., 2010; Regueiro et al., 2013; Sengupta et al., 2006; Mallick et al., 2016; Zei e t al., 2003; Zhong et al., 2013) (chapter 5.1.2)

	CTS2731	Q-B28	Q-B34	Q-B37	Q-CTS1002	Q-CTS4000	Q-CTS748	Q-F746	Q-L245	Q-L301	Q-L330	Q-L330*	Q-L53*	Q-L713	Q-M1107(xM3 Z780)	Q-M120	Q-M25	Q-M25*	Q-M378	Q-M378*	Q-M557	Q-M925	Q-M925*	Q-preM120	Q-S401	Q-S402	Q-Y1150	Q-Y1150*	Q-Y12421	Q-Y2269
Artic Area																														
Greenland																														
North America			1		1		2	29																1						
Canada								29																						
Alaska			1																					1						
Georgia																														
Luisiana					1																									
Montana																														
NE United States																														
North Carolina																														
SW United States																														
United States							1																							
Utah							1																							
Wicocomico																														
Mesoamerica	6				41	1	43								1							5			1			23		
Costa Rica																						1								
Guatemala							1															1								
Mexico	6				41	1	42																					1		
Nicaragua															1													1		
Panama																						3			1			21		
South Americ	2			15	1	82									2	1					7	8	1	19	1			3		
Argentina				4																										
Bolivia				2		43																1		1						
Brazil																														
Chile				1		2																								
Colombia																									1			2		
Ecuador				1																										
Paraguay	1																													
Peru	1			7	1	36									2	1					7	7	1	18				1		
Peru/Bolivia						1																								
North East Asia			3																											
Siberia			3																											
East Asia																1														
Japan																1														
South East Asia																61			2											2
China																61			2											2
China, Gansu																														
Central Asia										20	2	26			6	2	3													
Altai Republic										20		25				1														
Mongolia											2				5		3													
Mongolia												1			1	1														
South Asia		4													4		1										3	1		
Bangladesh																											1	1		
Bhutan															2															
India		1															1										1			
Nepal		1																									1			
Sri Lanka		2																												
Vietnam															2															
South West Asia										1		1		3	1	2	1	2									7	1		2
Afghanistan																											2			
Armenia																														1
Caucaso										1																				
Caucasus																1														
Kazakhstan																														
Pakistan															1		1	2									4	1		
Pakistan																											1			
Saudi Arabia																														
Turkey																														1
United Arab Emirates																														
Uzbekistan												1		3		1														
South West Asia/Middle East									4	5				27	1	2	2		2								2			3
Iran									4	5				27	1		1		2								2			
Iraq																	1													2
Jordan																														1
Lebanon																2														
East Europe		2										1	5	1	1		1										1			100
Belarus																														7
Bosnia-Herzegovir		1																												
Bulgaria												1		1					1											
Bulgaria																														2
Croatia																														1
Croatia, Hvar																														1
Estonia																														1
Hungary																														4
Latvia																														7
Lithuania																														18
Macedonia																														1
Moldova																														1
Poland													1														1			19
Poland																														1
Romania																														6
Russia													4			1														10
Ukranie		1																												21
Indonesia																1														
Brunei																1														
Central Europe		1																												54
Austria																														3
Czech Republic																														1
Czech Republic																														1
Germany																														20
Italy																														6
Netherlands																														22
Netherlands		1																												
Sweden																														1
Switzerland																														
West Europe																				1										8
France																														
Ireland																														3
Portugal																														
Spain																														1
United Kingdom																				1										4
North Africa																														
Algeria																														
Morocco																														
Unknown Origin																														14
Total	8	7	4	15	43	83	45	29	4	6	20	4	31	31	3	77	6	8	4	3	7	8	6	1	19	2	13	2	26	176

	Cl-Y2250	Cl-Y2647	Cl-Y4273	Cl-Y4303	Cl-Y780*	Cl-Y805	Cl-Y91102	Cl-Y91500	Cl-Y94500	Cl-Y9745	Cl-Y9910	Cl-Y9919	Cl-Y9919a	Cl-Y9919b	Cl-Z35921	Cl-Z5906	Cl-Z5906*	Cl-Z5906a	Cl-Z5908	Cl-Z5908*	Cl-Z5910	Cl-Z5910*	Cl-Z5910a	Cl-Z5910b	Cl-Z780	Cl-Z780*	Cl-Z781*	Cl-Z782	Y16467	Y16491	Total
Artic Area									1																						1
Greenland									1																						1
North America			4	3																					1						42
Canada																															29
Alaska																															2
Georgia			1	1																											2
Luisiana																															1
Montana																									1						1
NE United States			1																												1
North Carolina			1	1																											2
SW United States				1																											1
United States																															1
Utah																															1
Wicocomico			1																												1
Mesoamerica			1	3							8	3	16	15						3			1		2	15	8	4	2	1	203
Costa Rica																															1
Guatemala																															2
Mexico			1	2							8	3	16	15						3			1		2	14	7	4	2	1	170
Nicaragua																										1					3
Panama				1																							1				27
South America		2		1	10	7					1			1	4	3	1	8	3	16	10	5	14	2	1	6	12	1			250
Argentina						1												1		1			1								8
Bolivia																	1			1	1							1			51
Brazil		2		1																											3
Chile																															3
Colombia														1						1											5
Ecuador																															1
Paraguay																							1				1				3
Peru					10	6					1				4	3		7	3	13	9	5	12	2	1	6	11				175
Peru/Bolivia																															1
North East Asia							5	7	1																						16
Siberia							5	7	1																						16
East Asia																															1
Japan																															1
South East Asia										1																					66
China																															65
China, Gansu										1																					1
Central Asia							13																								72
Altai Republic																															46
Mongolia							13																								23
Mongolia																															3
South Asia	2								3	1																					19
Bangladesh																															2
Bhutan																															2
India	2								3	1																					9
Nepal																															2
Sri Lanka																															2
Vietnam																															2
South West Asia	3									8																					32
Afghanistan																															2
Armenia																															1
Caucaso										1																					2
Caucasus																															1
Kazakhstan	2																														2
Pakistan	1																														10
Pakistan																															1
Saudi Arabia										1																					1
Turkey										4																					5
United Arab Emirates										2																					2
Uzbekistan																															5
South West Asia/Middle East									1	8																					57
Iran									1	3																					46
Iraq										3																					6
Jordan										2																					3
Lebanon																															2
East Europe									7																						119
Belarus																															7
Bosnia-Herzegovina																															1
Bulgaria																															3
Bulgaria																															2
Croatia																															1
Croatia, Hvar																															1
Estonia																															1
Hungary																															4
Latvia																															7
Lithuania																															18
Macedonia																															1
Moldova																															1
Poland																															21
Poland																															1
Romania																															6
Russia										7																					22
Ukranie																															22
Indonesia																															1
Brunei																															1
Central Europe										1	11																				67
Austria																															3
Czech Republic																															1
Czech Republic																															1
Germany											1																				21
Italy										1	2																				9
Netherlands																															22
Netherlands																															1
Sweden																															1
Switzerland										8																					8
West Europe	1									4																					14
France										1																					1
Ireland																															3
Portugal	1																														1
Spain										3																					4
United Kingdom																															5
North Africa										5																					5
Algeria										1																					1
Morocco										4																					4
Unknown Origin																															14
Total	6	2	5	7	10	7	18	8	13	38	9	3	16	16	4	3	1	8	3	19	10	5	15	2	4	21	20	5	2	1	972

Table S3. Divergence dating of haplogroup Q sub-lineages obtained with BEAST. (chapther 5.1.5)

Lineages	TMRCA Mean (kya)	TMRCA Standard Deviation (kya)	Lineages	TMRCA Mean (kya)	TMRCA Standard Deviation (kya)
B280	0.7	0.4	Y515	1.9	0.7
B34	5.4	1.2	Y6794	6.5	1.3
B38	0.5	0.4	Y780	8.1	1.4
B42	10.1	1.7	Y805	2.9	0.9
B43	5.6	1.2	YP1102	3.9	0.9
B43b	2.7	0.8	YP11241	1.7	0.7
B43b2	1.5	0.6	YP1500	8.4	1.3
CTS1002	6.8	1.2	YP1500a	2.8	0.9
CTS1002a	4.7	0.9	YP1500a1	1.9	0.7
CTS2731	9.2	1.5	YP1677	1.5	0.5
CTS4000	4.3	0.9	YP1691	2.6	0.7
CTS748	8.5	1.4	YP4385	6.9	1.5
F1096	19.3	2.6	YP4500	3.3	0.9
F746	14.9	2.2	YP4722	4.3	0.9
L245	2.5	0.8	YP919	9.6	1.4
L274	26.4	3.6	YP919a	5.9	1.2
L275	14.0	2.2	YP919b	2.0	0.8
L330	8.3	1.5	Z35921	7.5	1.2
L472	23.2	2.9	Z5902	13.4	1.9
L54	15.6	1.8	Z5906	8.5	1.4
L713	2.2	0.8	Z5906a	6.2	1.3
L717	3.3	0.9	Z5906a1	2.4	0.8
L717a	1.8	0.6	Z5907	3.6	0.7
M1107	15.2	1.7	Z5908	9.4	1.3
M120	14.2	2.1	Z5910	9.1	1.3
M120eq	4.6	0.9	Z5910a	7.4	1.1
M25	12.4	2.2	Z5910b	7.5	1.3
M3	12.9	1.6	Z5911	6.9	1.1
M346	17.5	2.1	Z5912	5.8	1.0
M378	6.6	1.3	Z5912a	3.6	0.7
M848	12.5	1.6	Z5912a1	3.2	0.7
M848a	7.6	1.6	Z5912a1a	1.4	0.7
M848b	3.5	1.1	Z5915	7.7	1.5
M925	9.8	1.4	Z5916	7.1	1.4
M930	14.2	1.8	Z768	3.8	0.8
SA02	9.3	1.5	Z768a	2.7	0.8
Y1150	6.1	1.2	Z780	14.3	1.6
Y12421	5.3	1.0	Z781	12.5	1.5
Y12421a	4.2	0.9	Z782	3.1	1.1
Y26491	7.5	1.4	r274	3.4	0.7
Y26467	0.5	0.3	r282	4.9	1.2
Y26547	1.2	0.6	r283	7.8	1.5
Y2659	16.8	2.1	r285	9.5	1.6
Y27992	9.6	1.6	r289	2.5	0.8
Y4303	9.3	1.2	OUTGROUP	4.8	1.4

Table S4. Dataset of samples in chapter 5.2.2

Population	Continent	Geographic Region	Country	n(pre-QC)	n(post-QC) (LDD)	n(post-QC) (HDD)	Platform	Source
Abkhasians	Caucasus	Eurasia	Georgia	23	23	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2013 - Yunusbayev et al. 2011
Adygey	Caucasus	Eurasia	Russia	17	17	0	OmniExpress 610 – 660	Li et al. 2008
Albanian	Balkan	Europe	Albania	6	6	6	Omni2.5	This study
Algerian	NAfrica	Africa	Algeria	5	5	0	OmniExpress 610 – 660	Behar et al. 2013
Altaian	Siberia	Asia	Altai Republic	19	19	2	OmniExpress 610 – 660 / Omni 1M	Rasmussen et al. 2010 - Raghavan et al., 2013 - Yunusbayev et al. 2015
Armenians	Caucasus	Eurasia	Armenia	35	35	0	OmniExpress 610 – 660	Yunusbayev et al., 2011 - Behar et al. 2010
Ashkenazilewish	CEurope	Europe	France	20	20	2	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
Azerbaijani	Caucasus	Eurasia	Daghestan	2	2	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Azerbaijani	Caucasus	Eurasia	Iran	21	21	7	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2015
AzerbaijaniJewish	Caucasus	Eurasia	Azerbaijan	7	7	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
Balkars	Caucasus	Eurasia	Kabardino-Balkaria	22	22	3	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2015
Balochi	India	Asia	Pakistan	24	24	0	OmniExpress 610 – 660	Li et al. 2008
BantuKenya	EAfrica	Africa	Kenya	11	10	0	OmniExpress 610 – 660	Li et al. 2008
BantuSouthAfrica	SAfrica	Africa	South Africa	8	8	0	OmniExpress 610 – 660	Li et al. 2008
Bashkir	CAsia	Asia	Russia	23	23	3	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2015
Basque	WEurope	Europe	Spain	24	24	0	OmniExpress 610 – 660	Li et al. 2008
Bedouins	NAfrica	Africa	Egypt	46	46	0	OmniExpress 610 – 660	Li et al. 2008
Belarusians	EEurope	Europe	Belarus	16	16	0	OmniExpress 610 – 660	Behar et al. 2010 - Behar et al. 2013
BiakaPygmies	CAfrica	Africa	Central African Republic	21	20	0	OmniExpress 610 – 660	Li et al. 2008
Bosnian	Balkan	Europe	Bosnia andHerzegovina	15	15	0	OmniExpress 610 – 660	Kovacevic et al. 2014
Brahui	SAsia	Asia	Pakistan	25	25	0	OmniExpress 610 – 660	Li et al. 2008
British	UnitedKingdom	Europe	United Kingdom	101	95	94	Omni2.5	1000 Genome
Bulgarian	EEurope	Europe	Bulgaria	31	31	0	OmniExpress 610 – 660	Yunusbayev et al. 2011 - Hellenthal et al. 2014
Burusho	SAsia	Asia	Pakistan	25	25	0	OmniExpress 610 – 660	Li et al. 2008
CEU	CEurope	Europe	Utah	104	99	99	Omni2.5	1000 Genome
CHB	CAsia	Asia	Cina	100	51	51	Omni2.5	1000 Genome
Chechen	Caucasus	Eurasia	Chechen	24	24	0	OmniExpress 610 – 660	Metspalu et al. 2011 - Chaubey et al.2012 - Yunusbayev et al. 2011
Chukchin	Siberia	Asia	Chukotka Autonomous Okrug (Russia)	15	10	1	OmniExpress 610 – 660 / Omni 1M	Rasmussen et al. 2010 - Yunusbayev et al. 2015
CochinJewish	MiddleEast	Asia	Israel	3	3	3	Omni 1M	Behar et al. 2013
Corsican	SEurope	Europe	Corsica	16	16	0	OmniExpress 610 – 660	Metspalu unpublished data
Croatian	SEEurope	Europe	Croatia	43	43	4	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2013- Hellenthal et al. 2014
Cypriots	SEEurope	Europe	Cyprus	12	12	0	OmniExpress 610 – 660	Behar et al. 2010
Druze	MiddleEast	Asia	Lebanon	45	44	3	OmniExpress 610 – 660 / Omni 1M	Li et al. 2008 - Behar et al. 2013
Egypt	NAfrica	Africa	Egypt	112	24	0	OmniExpress 610-660 / Omni2.5	Behar et al. 2010 - Pagani et al. 2015
English	UnitedKingdom	Europe	England	8	8	0	OmniExpress 610 – 660	Hellenthal et al. 2014
Estonian	NEurope	Europe	Estonia	21	21	0	OmniExpress 610 – 660	Raghavan et al., 2013 - Kushniarevic et al. 2015
Ethiopian	EAfrica	Africa	Ethiopia	143	120	102	OmniExpress 610 - 660 / Omni2.5	Behar et al. 2010 - Pagani et al. 2015
EthiopianJewish	EAfrica	Africa	Ethiopia	16	15	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
Finnish	Scandinavia	Europe	Finland	102	102	100	OmniExpress 610 - 660 / Omni2.5	Hellenthal et al. 2014 - 1000 Genome
French	CEurope	Europe	France	28	28	0	OmniExpress 610 – 660	Li et al. 2008
French	CEurope	Europe	Provance (France)	5	5	0	OmniExpress 610 – 660	Li et al. 2008
FrenchJew	CEurope	Europe	France	6	6	0	OmniExpress 610 – 660	Behar et al. 2013
Gagauz	EEurope	Europe	Moldova	12	12	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
GeorgianJew	Caucasus	Eurasia	Georgia	11	11	3	OmniExpress 610 – 660 / Omni 1M	Behar et al.2013
Georgians	Caucasus	Eurasia	Georgia	30	30	10	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
German	CEurope	Europe	Germany	44	44	0	OmniExpress 610 – 660	Hellenthal et al. 2014 - Busta / Parob et al. 2014 - Yunusbayev et al. 2015
GermanAustrian	SEurope	Europe	Germany	4	4	0	OmniExpress 610 – 660	Hellenthal et al. 2014
Gond	SAsia	Asia	India	4	4	0	OmniExpress 610 – 660	Metspalu et al. 2011
Greek	SEEurope	Europe	Greek	0	20	0	OmniExpress 610 – 660	Hellenthal et al. 2014

Population	Continent	Georgraphic Region	Country	n(pre-QC)	n(post-QC) (LDD)	n(post-QC) (HDD)	Platform	Source
GreekCentral	SEEurope	Europe	Greek	10	10	0	OmniExpress 610 – 660	Behar et al. 2013
GreekMacedonian	SEEurope	Europe	Greek	7	7	0	OmniExpress 610 – 660	Kushniarevic et al. 2015
GreekPeloponesian	SEEurope	Europe	Greek	9	9	0	OmniExpress 610 – 660	Kushniarevic et al. 2015
GreekThessalyian	SEEurope	Europe	Greek	10	10	0	OmniExpress 610 – 660	Behar et al. 2013
Hungarians	EEurope	Europe	Hungaria	20	19	0	OmniExpress 610 – 660	Behar et al. 2010
IranianJewish	MiddleEast	Asia	Iran	12	11	2	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
Iranians	MiddleEast	Asia	Iran	20	19	0	OmniExpress 610 – 660	Behar et al. 2010
IraqiJewish	MiddleEast	Asia	Iraq	13	13	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
Irish	United Kingdom	Europe	Irland	7	7	0	OmniExpress 610 – 660	Hellenthal et al. 2014
IT_U	SEurope	Europe	Italy	144	138	0	OmniExpress 610 – 660	Behar et al. 2013 - Parolo et al. 2014
ITC-ABR_B	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-ABR_G	SEurope	Europe	Italy	22	19	8	OmniExpress 610 - 660 / Omni2.5	This study - Metspalu Unpublished data
ITC-ABR_P	SEurope	Europe	Italy	5	5	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-LAZ_B	SEurope	Europe	Italy	5	5	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-LAZ_F-ITS-PUG_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-LAZ_G	SEurope	Europe	Italy	35	32	32	OmniExpress 610 - 660 / Omni2.5 / Omni1M	Fiorito et al. 2015 - This study
ITC-LAZ_P	SEurope	Europe	Italy	3	3	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-MAR_B	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-MAR_F-ITN-LOM_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-MAR_G	SEurope	Europe	Italy	16	16	16	Omni2.5	This study
ITC-MAR_P	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-TSI_B	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-TSI_F-ITC-ABR_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-TSI_F-ITN-LOM_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-TSI_F-ITN-VEN_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-TSI_F-ITS-SIC_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-TSI_G	SEurope	Europe	Italy	140	135	124	OmniExpress 610 - 660 / Omni2.5 / Omni1M	1000 Genome - Fiorito et al. 2015 - Li et al. 2008 - Metspalu Unpublished - This study
ITC-TSI_P	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-UMB_B	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITC-UMB_G	SEurope	Europe	Italy	12	9	9	Omni2.5	This study
ITN-EMI_B	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-EMI_F-ITC-ABR_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-EMI_F-ITN-LOM_M	SEurope	Europe	Italy	4	4	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-EMI_F-ITN-VEN_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-EMI_G	SEurope	Europe	Italy	30	30	30	Omni2.5 / Omni1M	Fiorito et al. 2015
ITN-EMI_P	SEurope	Europe	Italy	9	9	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-FRI_B	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-FRI_F-ITC-EMI_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-FRI_F-ITN-LOM_M	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-FRI_F-ITN-PIE_M	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-FRI_F-ITS-CAL_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-FRI_G	SEurope	Europe	Italy	18	15	15	Omni2.5	This study
ITN-FRI_P	SEurope	Europe	Italy	4	4	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-LIG_B	SEurope	Europe	Italy	6	6	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-LIG_F-ITN-LOM_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-LIG_G	SEurope	Europe	Italy	25	25	25	Omni2.5 / Omni1M	Fiorito et al. 2015 - This study
ITN-LIG_P	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-LOM_B	SEurope	Europe	Italy	76	76	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITC-EMI_M	SEurope	Europe	Italy	6	6	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015

Population	Continent	Georgraphic Region	Country	n(pre-QC)	n(post-QC) (LDD)	n(post-QC) (HDD)	Platform	Source
ITN-LOM_F-ITC-LAZ_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITC-TSL_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITC-UMB_M	SEurope	Europe	Italy	2	2	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITN-ALT_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITN-FRL_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITN-LIG_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITN-PIE_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITN-TRE_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITN-VEN_M	SEurope	Europe	Italy	12	12	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_F-ITS-SIC_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-LOM_G	SEurope	Europe	Italy	52	52	40	OmniExpress 6.0 - 660 / Omni2.5 / Omni1M	Li et al. 2008 - Fiorito et al. 2015 - This study
ITN-LOM_P	SEurope	Europe	Italy	147	147	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-PIE_B	SEurope	Europe	Italy	239	239	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-PIE_F-ITN-LOM_M	SEurope	Europe	Italy	5	5	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-PIE_F-ITN-VDA_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-PIE_G	SEurope	Europe	Italy	40	40	29	OmniExpress 6.0 - 660 / Omni2.5 / Omni1M	Fiorito et al. 2015 - Metspalu Unpublished
ITN-PIE_P	SEurope	Europe	Italy	3	3	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-TRE_B	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-TRE_F-ITN-LOM_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-TRE_F-ITN-PIE_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-TRE_F-ITN-VEN_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-TRE_G	SEurope	Europe	Italy	8	8	8	Omni2.5	This study
ITN-VDA_G	SEurope	Europe	Italy	29	29	29	Omni2.5 / Omni1M	Fiorito et al. 2015 - This study
ITN-VEN_B	SEurope	Europe	Italy	10	10	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-VEN_F-ITC-EMI_M	SEurope	Europe	Italy	2	2	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-VEN_F-ITN-LOM_M	SEurope	Europe	Italy	9	9	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-VEN_F-ITN-PIE_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-VEN_F-ITS-SAR_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITN-VEN_G	SEurope	Europe	Italy	16	16	16	Omni2.5	This study
ITN-VEN_P	SEurope	Europe	Italy	40	40	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-BAS_B	SEurope	Europe	Italy	2	2	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-BAS_G	SEurope	Europe	Italy	38	37	37	Omni2.5 / Omni1M	Fiorito et al. 2015 - This Study
ITS-BAS_P	SEurope	Europe	Italy	9	9	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAL_B	SEurope	Europe	Italy	7	7	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAL_F-ITS-SIC_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAL_G	SEurope	Europe	Italy	32	32	14	OmniExpress 6.0 - 660 / Omni2.5 / Omni1M	Hellenthal et al. 2014 - Fiorito et al. 2014 - This study
ITS-CAL_P	SEurope	Europe	Italy	12	12	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAM_B	SEurope	Europe	Italy	5	5	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAM_F-ITC-LAZ_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAM_F-ITN-LOM_M	SEurope	Europe	Italy	2	2	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAM_F-ITN-VEN_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-CAM_G	SEurope	Europe	Italy	15	14	14	Omni2.5	This study
ITS-CAM_P	SEurope	Europe	Italy	17	17	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-MOL_G	SEurope	Europe	Italy	2	2	2	Omni2.5	This study
ITS-PUG_B	SEurope	Europe	Italy	15	15	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-PUG_F-ITN-LOM_M	SEurope	Europe	Italy	3	3	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-PUG_F-ITN-VEN_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-PUG_F-ITS-CAM_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015
ITS-PUG_F-ITS-SIC_M	SEurope	Europe	Italy	1	1	0	OmniExpress 6.0 – 660	Besta / Parolo et al. 2015

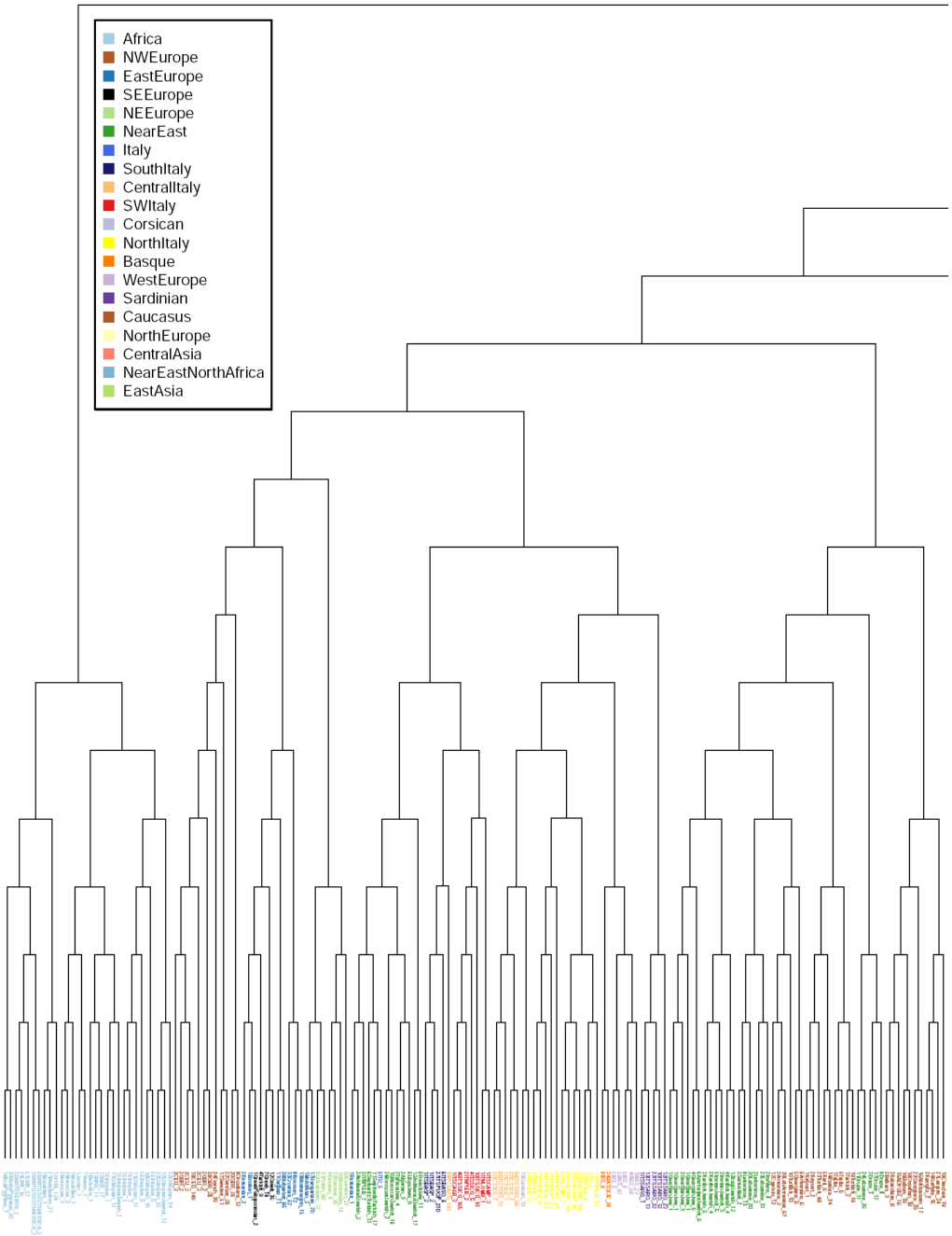
Population	Continent	Georgraphic Region	Country	n(pre-QC)	n(post-QC) (LDD)	n(post-QC) (HDD)	Platform	Source
ITS-PUG_G	SEurope	Europe	Italy	15	14	14	Omni2.5	This study
ITS-PUG_P	SEurope	Europe	Italy	27	27	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SAR_B	SEurope	Europe	Italy	5	5	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SAR_F-ITS-CAM_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SAR_G	SEurope	Europe	Italy	59	58	31	OmniExpress 610 - 660 / Omni2.5 / Omni1M	Li et al. 2008 - Fiorito et al. 2015
ITS-SAR_P	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SIC_B	SEurope	Europe	Italy	11	11	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SIC_F-ITC-EMI_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SIC_F-ITN-FRL_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SIC_F-ITN-LOM_M	SEurope	Europe	Italy	2	2	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SIC_F-ITN-VEN_M	SEurope	Europe	Italy	1	1	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
ITS-SIC_G	SEurope	Europe	Italy	64	64	31	OmniExpress 610 - 660 / Omni2.5 / Omni1M	Behar et al. 2013 - Hellenthal et al. 2014 - Fiorito et al. 2015
ITS-SIC_P	SEurope	Europe	Italy	25	25	0	OmniExpress 610 – 660	Besta / Parolo et al. 2015
Jordan	MiddleEast	Asia	Jordan	20	20	0	OmniExpress 610 – 660	Behar et al. 2010
Kabardians	Caucasus	Eurasia	Kabardino-Balkaria	3	3	3	Omni1M	Yunusbayev et al. 2015
Kalash	Caucasus	Asia	Pakistan	23	22	0	OmniExpress 610 – 660	Li et al. 2008
Kalmyk	Caucasus	Eurasia	Kalmykia	14	14	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Karakalpak	CAsia	Asia	Uzbekistan	10	10	10	Omni1M	Yunusbayev et al. 2015
Karelian	NAsia	Asia	Republic of Karelia	15	15	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Kayseri	WAsia	Eurasia	Turkey	23	23	0	OmniExpress 610 – 660	Hodoˇglugil et al 2012
Kazakh	CAsia	Asia	Kazakhstan	20	20	2	OmniExpress 610 – 660 / Omni 1M	Raghavan et al., 2013 - Yunusbayev et al. 2015
Kenyan	EAfrica	Africa	Kenya	31	28	28	Omni2.5	1000 Genome
Kets	Siberia	Asia	Krasnoyarsk Krai (Russia)	4	4	2	OmniExpress 610 – 660 / Omni 1M	Rasmussen et al. 2010 - Yunusbayev et al. 2015
kol	SAsia	Asia	India	16	15	0	OmniExpress 610 – 660	Metspalu et al., 2011
komi	NAsia	Asia	Komi Republic	16	16	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Kosovan	Balkan	Europe	Serbia	9	9	0	OmniExpress 610 – 660	Kovacevic et al. 2014
Kryashen	CAsia	Asia	Kazakhstan	3	3	3	Omni1M	Yunusbayev et al. 2015
Kumyks	WAsia	Asia	Kalmykia	17	17	3	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2011 - Yunusbayev et al. 2015
Kurd	MiddleEast	Asia	Turkey	6	6	0	OmniExpress 610 – 660	Yunusbayev et al. 2011
Kurdish-Jewish	MiddleEast	Asia	Turkey	10	9	1	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2013
Kyrgyz	CentralAsia	Asia	Kyrgyzstan	44	41	8	OmniExpress 610 – 660 / Omni 1M	Hodoˇglugil et al 2012 - Raghavan et al., 2013 - Yunusbayev et al. 2015
Latvian	EEurope	Europe	Latvia	6	6	0	OmniExpress 610 – 660	Kushniarevic et al. 2015
Lebanese	MiddleEast	Asia	Lebanon	82	79	0	OmniExpress 610 – 660	Behar et al. 2010 - Haber et al. 2013
LebArmenian	MiddleEast	Asia	Lebanon	39	39	0	OmniExpress 610 – 660	Haber et al. 2015
Lezgins	Caucasus	Eurasia	Dagestan	21	21	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013
Libyalews	NAfrica	Africa	Lybia	6	6	0	OmniExpress 610 – 660	Behar et al. 2013
Lithuanians	EEurope	Europe	Lithuania	10	10	0	OmniExpress 610 – 660	Behar et al. 2010
LWK	EAfrica	Africa	Kenya	100	85	86	Omni2.5	1000 Genome
Macedonian	Balkan	Europe	Macedonia	14	14	0	OmniExpress 610 – 660	Kovacevic et al. 2014
Makrani	SAsia	Asia	Pakistan	25	25	0	OmniExpress 610 – 660	Li et al. 2008
Mandenkas	WAfrica	Africa	Senegal	22	21	0	OmniExpress 610 – 660	Li et al. 2008
Maris	CAsia	Asia	Mari El Republic (Russia)	15	15	0	OmniExpress 610 – 660	Raghavan et al., 2013
MbutPygmies	SAfrica	Africa	Democratic Republic of Congo	13	1	0	OmniExpress 610 – 660	Li et al. 2008
Moldovian	EEurope	Europe	Moldova	7	7	7	Omni1M	Behar et al. 2013
Mongolian	EAsia	Asia	Mongolia	21	21	2	OmniExpress 610 – 660 / Omni 1M	Li et al. 2008 - Rasmussen et al. 2010 - Yunusbayev et al. 2015
Montenegrin	Balkan	Europe	Montenegro	14	14	0	OmniExpress 610 – 660	Kovacevic et al. 2014
Mordovians	CAsia	Asia	Mordovia	15	15	0	OmniExpress 610 – 660	Yunusbayev et al. 2011
Moroccan	NAfrica	Africa	Morocco	72	93	66	OmniExpress 610 - 660 / Omni2.5	Behar et al. 2010 - Hellenthal et al. 2014 - Kivisild Unpublished data
MoroccanJewish	NAfrica	Africa	Morocco	18	18	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010 - Behar et al. 2013

Population	Continent	Geographic Region	Country	n(pre-QC)	n(post-QC) (LDD)	n(post-QC) (HDD)	Platform	Source
Mozabites	NAfrica	Africa	Algeria	29	29	0	OmniExpress 610 – 660	Li et al. 2008
Nenet	Siberia	Asia	Russia	16	15	5	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2015
Nogais	CAsia	Asia	Russia	16	16	0	OmniExpress 610 – 660	Yunusbayev et al. 2011
Norwegian	Scandinavia	Europe	Norway	0	18	0	OmniExpress 610 – 660	Hellenthal et al. 2014
Orcadians	United Kindgom	Europe	Orkneyislands	15	15	0	OmniExpress 610 – 660	Li et al. 2008
Ossetian	CAsia	Asia	Ossetia	18	18	3	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2011 - Behar et al. 2013
Palestinian	MiddleEast	Asia	Palestine	52	52	6	OmniExpress 610 – 660 / Omni 1M	Li et al. 2008; Behar et al. 2013
Polish	EEurope	Europe	Poland	19	36	1	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2013 - Hellenthal et al. 2014
Portugal	WEurope	Europe	Portugal	10	10	0	OmniExpress 610 – 660	Metspalu unpublished data
Romanian	EEurope	Europe	Romania	20	20	0	OmniExpress 610 – 660	Behar et al. 2010 - Besta / Parcelo et al. 2015
Russians	CAsia	Asia	Arch	1	1	0	OmniExpress 610 – 660	Li et al. 2008
Russians	CAsia	Asia	Pinega	4	4	0	OmniExpress 610 – 660	Li et al. 2008
Russians	CAsia	Asia	Russia	58	58	0	OmniExpress 610 – 660	Li et al. 2008 - Behar et al. 2013 - Kushniarevic et al. 2015 - Yunusbayev et al. 2015
Samaritan	MiddleEast	Asia	Israel	3	2	0	OmniExpress 610 – 660	Behar et al. 2010
Saudi	MiddleEast	Asia	Saudi Arabian	20	20	0	OmniExpress 610 – 660	Behar et al. 2010
Scottish	United Kindgom	Europe	Scotland	12	6	0	OmniExpress 610 – 660	Hellenthal et al. 2014
Selkups	CAsia	Asia	Siberia	17	17	0	OmniExpress 610 – 660	Raghavan et al., 2013 - Rasmussen et al. 2010
SephardicJewish	MiddleEast	Asia	Israel	3	3	0	OmniExpress 610 – 660	Behar et al. 2013
SephardicTurkish	MiddleEast	Asia	Israel	19	19	0	OmniExpress 610 – 660	Behar et al. 2010
Serbian	Balkan region	Europe	Serbia	18	18	0	OmniExpress 610 – 660	Kovacevic et al. 2014
Slovak	CEurope	Europe	Slovakia	15	15	0	OmniExpress 610 – 660	Kushniarevic et al. 2015
Slovenes	Balkan region	Europe	Slovenia	15	15	0	OmniExpress 610 – 660	Kushniarevic et al. 2015
Spaniards	WEurope	Europe	Spain	0	34	0	OmniExpress 610 – 660	Hellenthal et a. 2014 - Behar et al. 2010
Spanish	WEurope	Europe	Spain	150	100	100	Omni2.5	1000 Genome
Sweden	Neurope	Europe	Sweden	18	17	0	OmniExpress 610 – 660	Behar et al. 2013
SyriaJew	MiddleEast	Asia	Syria	2	2	0	OmniExpress 610 – 660	Behar et al. 2013
Syrian	MiddleEast	Asia	Syria	16	16	0	OmniExpress 610 – 660	Behar et al. 2010
Tahassetan	Caucasus	Eurasia	Dagestan (Russia)	3	2	0	OmniExpress 610 – 660	Behar et al. 2013
Tajiks	CAsia	Asia	Tajikistan	60	60	40	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2011
Tatar	CAsia	Asia	Tatar	20	20	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Tunisia	NAfrica	Africa	Tunisia	16	12	0	OmniExpress 610 – 660	Hellenthal et al. 2014
TunisiaJew	NAfrica	Africa	Tunisia	6	6	0	OmniExpress 610 – 660	Behar et al. 2013
Turkish	Capadoccia	Asia	Turkey	49	49	10	OmniExpress 610 - 660 / Omni2.5	Behar et al. 2010 - Hodo˘glugil et al 2012 - Paschou et al. 2015
Turkish	MiddleEast	Asia	Turkey	20	20	0	OmniExpress 610 – 660	Hodo˘glugil et al 2012
Turkmens	CAsia	Asia	Turkmenistan	27	23	8	OmniExpress 610 – 660 / Omni 1M	Yunusbayev et al. 2011 - Di Cristoforo et al. 2012 - Yubusbayev et al. 2015
Tuva	CAsia	Asia	Tuva	19	16	1	OmniExpress 610 – 660 / Omni 1M	Rasmussen et al. 2010 - Yunusbayev et al. 2015
UAE	MiddleEast	Asia	uae	19	14	0	OmniExpress 610 – 660	Hellenthal et al. 2014
Udmurd	CAsia	Asia	Kirov Oblast	16	16	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Ukrainians	EEurope	Europe	Ukraine	20	20	0	OmniExpress 610 – 660	Yunusbayev et al. 2011
Uzbeks	CAsia	Asia	Uzbekistan	24	24	0	OmniExpress 610 – 660	Behar et al. 2010 - Di Cristoforo et al. 2012 - Raghavan et al. 2015
Vespas	CAsia	Asia	Republic of Karelia	11	10	0	OmniExpress 610 – 660	Yunusbayev et al. 2015
Vietnamese	EAsia	Asia	Vietnam	121	51	99	Omni2.5	1000 Genome
Welsh	United Kindgom	Europe	Wales	0	4	0	OmniExpress 610 – 660	Hellenthal et al. 2014
Yakut	Siberia	Asia	Yakutia	25	25	3	OmniExpress 610 – 660 / Omni 1M	Li et al. 2008; Yunusbayev et al. 2015
Yemen	MiddleEast	Asia	Yemen	10	8	0	OmniExpress 610 – 660	Behar et al. 2010
Yemen_Jew	MiddleEast	Asia	Yemen	18	18	3	OmniExpress 610 – 660 / Omni 1M	Behar et al. 2010; Behar et al. 2013
Yoruba	Africa	Africa	Nigeria	179	124	104	OmniExpress 610 - 660 / Omni2.5	Li et al. 2008 - 100 Genome
			Total	**5058**	**4052**	**1641**		

Table S4. Fst estimates according to fineSTRUCTURE Italian genetic clusters. Clusters are named as follow: tot number of samples in the cluster "_" number and origin of the most representate subjects (5.2.7).

55_17IT NVENP	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
123_47IT NLOMP	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
313_111IT NPIEB	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
132_24IT SSICG	0,0021	0,0015	0,0015	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
186_25IT SBASG	0,0015	0,0010	0,0010	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
13_4IT SSARG	0,0049	0,0042	0,0044	0,0048	0,0046	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
26_8IT SPUGG	0,0018	0,0015	0,0014	0,0005	0,0003	0,0053	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
51_31IT NLOMG	0,0007	0,0004	0,0006	0,0022	0,0016	0,0046	0,0022	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
76_28IT CLAZG	0,0009	0,0005	0,0005	0,0004	0,0002	0,0043	0,0005	0,0011	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
58_46IT NLOMP	0,0005	0,0002	0,0004	0,0017	0,0013	0,0044	0,0019	0,0005	0,0007	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
9_7IT SSARG	0,0060	0,0055	0,0056	0,0063	0,0061	0,0009	0,0066	0,0057	0,0055	0,0056	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
61_54IT CTSIG	0,0006	0,0003	0,0003	0,0009	0,0005	0,0042	0,0009	0,0009	0,0002	0,0005	0,0054	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16_13IT NFRIG	0,0005	0,0009	0,0007	0,0027	0,0020	0,0058	0,0026	0,0013	0,0013	0,0010	0,0072	0,0012	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
5_4IT SSARG	0,0105	0,0097	0,0101	0,0110	0,0107	0,0042	0,0112	0,0099	0,0102	0,0100	0,0044	0,0099	0,0114	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
10_10IT NVDAG	0,0022	0,0021	0,0021	0,0041	0,0035	0,0069	0,0039	0,0028	0,0028	0,0024	0,0079	0,0025	0,0029	0,0115	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
13_13IT NVDAG	0,0016	0,0018	0,0016	0,0040	0,0032	0,0066	0,0037	0,0024	0,0025	0,0018	0,0078	0,0021	0,0020	0,0126	0,0024	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
13_13IT SSARG	0,0066	0,0058	0,0060	0,0066	0,0065	0,0014	0,0070	0,0062	0,0060	0,0061	0,0008	0,0059	0,0075	0,0021	0,0083	0,0086	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
6_6IT SSARG	0,0092	0,0087	0,0090	0,0095	0,0095	0,0039	0,0099	0,0090	0,0088	0,0090	0,0034	0,0088	0,0107	0,0067	0,0114	0,0114	0,0024	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
10_10IT SSICG	0,0039	0,0036	0,0034	0,0023	0,0022	0,0068	0,0027	0,0041	0,0025	0,0038	0,0077	0,0028	0,0046	0,0133	0,0058	0,0057	0,0088	0,0117	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
5_5IT SCALG	0,0043	0,0037	0,0037	0,0019	0,0018	0,0072	0,0025	0,0044	0,0022	0,0038	0,0088	0,0031	0,0047	0,0138	0,0060	0,0065	0,0087	0,0122	0,0034	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
5_4IT SSICG	0,0051	0,0049	0,0048	0,0045	0,0043	0,0083	0,0048	0,0054	0,0041	0,0050	0,0093	0,0045	0,0053	0,0131	0,0072	0,0070	0,0093	0,0120	0,0070	0,0062	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
7_7IT SSICG	0,0025	0,0021	0,0019	0,0015	0,0013	0,0062	0,0016	0,0029	0,0013	0,0024	0,0071	0,0017	0,0033	0,0119	0,0044	0,0039	0,0080	0,0109	0,0032	0,0031	0,0052	0,0000	0,0000	0,0000	0,0000	0,0000
10_10IT SSARG	0,0100	0,0092	0,0094	0,0102	0,0100	0,0042	0,0105	0,0094	0,0095	0,0094	0,0037	0,0094	0,0111	0,0063	0,0125	0,0119	0,0042	0,0077	0,0123	0,0123	0,0130	0,0112	0,0000	0,0000	0,0000	0,0000
7_7IT SSARG	0,0112	0,0106	0,0108	0,0117	0,0115	0,0062	0,0119	0,0109	0,0109	0,0111	0,0063	0,0110	0,0121	0,0086	0,0125	0,0135	0,0033	0,0055	0,0138	0,0143	0,0151	0,0126	0,0096	0,0000	0,0000	0,0000
45_45IT CTSIG	0,0008	0,0005	0,0005	0,0011	0,0007	0,0044	0,0011	0,0011	0,0004	0,0008	0,0056	0,0002	0,0015	0,0099	0,0026	0,0023	0,0062	0,0093	0,0030	0,0034	0,0044	0,0018	0,0096	0,0110	0,0000	0,0000
19_19IT CTSIG	0,0015	0,0012	0,0012	0,0019	0,0015	0,0051	0,0017	0,0017	0,0011	0,0015	0,0061	0,0009	0,0021	0,0105	0,0033	0,0032	0,0069	0,0096	0,0038	0,0041	0,0054	0,0028	0,0103	0,0119	0,0005	0,0000

Figure S1. **Italian administrative regions.** The labels for the regions in this thesis are as follows: ITC-ABR, Abruzzo; ITN-EMI, Emilia Romagna; IT-U, Unassigned (samples with mixed/unknown origins); ITC-LAZ, Lazio; ITC-MAR, Marche; ITC-TSI, Tuscany; ITC-UMB, Umbria; ITN-FRI, Friuli Venezia Giulia; ITN-LIG, Liguria; ITN-LOM, Lombardy; ITN-PIE, Piedmont ITS-PUG, Apulia; ITN-TRE, Trentino Alto Adige; ITN-VDA, Valle D'Aosta; ITN-VEN, Veneto; ITS-BAS, Basilicata; ITS-CAL, Calabria; ITS-CAM, Campania; ITS-MOL, Molise; ITS-SAR, Sardinia; ITS-SIC, Sicily. The ITN-, North-Italy; ITC-, Central-Italy; ITS-, South Italy; prefixes are referring to the macro-area in which this regions are located.


Africa
NWEurope
EastEurope
SEEurope
NEEurope
NearEast
Italy
SouthItaly
CentralItaly
SWItaly
Corsican
NorthItaly
Basque
WestEurope
Sardinian
Caucasus
NorthEurope
CentralAsia
NearEastNorthAfrica
EastAsia

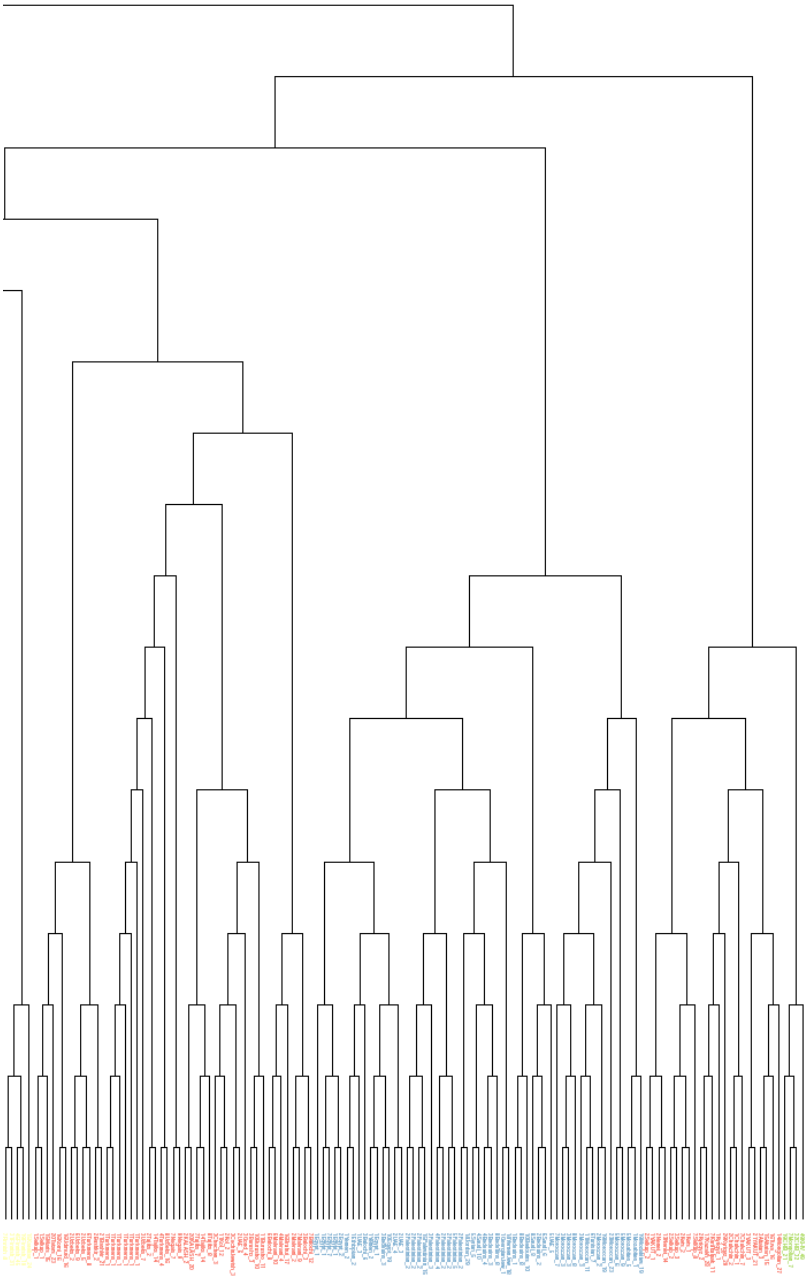


Figure S2. **FineSTRUCTURE original cladogram of the LDD.** (chapther 5.2.5)

References

1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526: 68–74.

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–1073.

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science. 2011;334: 89–94.

Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, Woodward SR et al. The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS One. 2008;3: e1764.

Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Kashani BH, et al.. Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proc Natl Acad Sci U S A. 2013;110: 14308–14313.

Ahmad A. History of Islamic Sicily. 2010. Columbia Univ Pr. United States of America

Ajay SS, Parker SCJ, Abaan HO, Fajardo KVF, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res. 2011;21: 1498–1505

Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12: 246.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664.

Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12: 363–376.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522: 167–172.

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522: 167–172.

Alonso S, Armour JA. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outstide Africa. Proc Natl Acad Sci U S A. 2001;98: 864–869.

Anderson S, Bankier AT, Barrell BG, De Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290: 457–465.

AnphibiaWeb. 2017. (https://amphibiaweb.org)

Anthony D. The horse, the wheel and language. How Bronze-Age riders from the Eurasian steppes shaped the modern world. Princeton University Press. Princeton and Oxford. 2007.

Armitage SJ, Jasim SA, Marks AE, Parker AG, Usik VI, Uerpmann H-P. The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. Science. 2011;331: 453–456.

Asouti E, Fuller DQ. A contextual approach to the emergence of agriculture in Southwest Asia. Curr Anthropol. 2013;54: 299–345.

Balanovsky O, Gurianov V, Zaporozhchenko V, Balaganskaya O, Urasin V, Zhabagin M, et al. Phylogeography of human Y-chromosome haplogroup Q3-L275 from an academic/citizen science collaboration. BMC Evol Biol. 2017;17

Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al.. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet. 2010;87: 341–353.

Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16: 37–48.

Bandelt HJ, Forster P, Sykes BC, Richards MB. Mitochondrial portraits of human populations using median networks. Genetics. 1995;141: 743–753.

Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. Ancient substructure in early mtDNA lineages of southern Africa. Am J Hum Genet. 2013;92: 285–292.

Barbujani G, Sokal RR, Oden NL. Indo-European origins: a computer-simulation test of five hypotheses. Am J Phys Anthropol. 1995;96: 109–132.

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, et al. Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. Eur J Hum Genet. 2009;17: 820–830.

Battaglia V, Grugni V, Perego UA, Angerhofer N, Gomez-Palmieri JE, Woodward SR, et al. The First Peopling of South America: New Evidence from Y-Chromosome Haplogroup Q. PLoS One. 2013;8: e71390.

Behar DM, Metspalu M, Baran Y, Kopelman NM, Yunusbayev B, Gladstein A, et al. No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. Hum Biol. 2013;85: 859–900.

Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, et al. The genome-wide structure of the Jewish people. Nature. 2010;466: 238–242.

Benazzi S, Bailey SE, Peresani M, Mannino MA, Romandini M, Richards MP, et al. Middle Paleolithic and Uluzzian human remains from Fumane Cave, Italy. J Hum Evol. 2014;70: 61–68.

Benazzi S, Douka K, Fornai C, Bauer CC, Kullmer O, Svoboda J, et al. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature. 2011;479: 525–528.

Benazzi S, Slon V, Talamo S, Negrino F, Peresani M, Bailey SE, et al. Archaeology. The makers of the Protoaurignacian and implications for Neandertal extinction. Science. 2015;348: 793–796.

Bergström A, Nagle N, Chen Y, McCarthy S, Pollard MO, Ayub Q, et al.. Deep Roots for Aboriginal Australian Y Chromosomes. Curr Biol. 2016;26: 809–813.

Bianchi NO, Catanesi CI, Bailliet G, Martinez-Marignac VL, Bravi CM, Vidal-Rioja LB, et al. Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet. 1998;6: 1862-71.

Bietti Sestieri AM. L'Italia nell'età del Bronzo e del Ferro. Carocci Editore. Roma. 2010.

Boattini A, Castrì L, Sarno S, Useli A, Cioffi M, Sazzini M, et al. mtDNA variation in East Africa unravels the history of Afro-Asiatic groups. Am J Phys Anthropol. 2013;150: 375–385.

Boivin N, Fuller DQ, Dennell R, Allaby R, Petraglia MD. Human dispersal across diverse environments of Asia during the Upper Pleistocene. Quat Int. 2013;300: 32–47.

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, et al. Mapping the origins and expansion of the Indo-European language family. Science. 2012;337: 957–960.

Brandini S, Bergamaschi P, Cerna MF, Gandini F, Bastaroli F, Bertolini E et al. The Paleo-Indian Entry into South America According to Mitogenomes. Mol Biol Evol. 2017.

Brandstätter A, Niederstätter H, Parson W. Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. Int J Legal Med. 2004;118: 47–54.

Brandt G, Haak W, Adler CJ, Roth C, Szécsényi-Nagy A, Karimnia S, et al. Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. Science. 2013;342: 257–261.

Brisighelli F, Álvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo A, Pascali VL, et al. Uniparental markers of contemporary Italian population reveal details on its pre-Roman heritage. PLoS One. 2012;7: e50794.

Brisighelli F, Blanco-Verea A, Boschi I, Garagnani P, Pascali VL, Carracedo A, et al. Patterns of Y-STR variation in Italy. Forensic Sci Int Genet. 2012;6: 834–839.

Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, et al. A early Neolithic genomes from the eastern Fertile Crescent. Science. 2016;353: 499–503.

Bryc K, Durand EY, Michael Macpherson J, Reich D, Mountain JL. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015;96: 37–53.

Burgarella C, Navascués M. Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. Eur J Hum Genet. 2011;19: 70–75.

Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. Elife. 2016;5. doi:10.7554/eLife.15266

Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, et al. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc R Soc B. 2012; 279:884-892.

Busby GBJ, Hellenthal G, Montinaro F, Tofanelli S, Bulayeva K, Rudan I, et al. The role of recent admixture in forming the contemporary west Eurasian genomic landscape. Curr Biol. 2015;25: 2518–2526.

Calafell F, Larmuseau MH. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. Hum Genet. 2016; 1:1-5

Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987;325: 31–36.

Cann RL. mtDNA and Native Americans: a Southern perspective. Am J Hum Genet. 1994;55: 7–11.

Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia' A, Vetrugno G, et al. Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic–Neolithic encounter. Mol Phylogenet Evol. 2007;44: 228–239.

Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, et al. Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci U S A. 2003;100: 6593–6597.

Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, et al. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science. 1985;230: 1403–1406.

Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, et al. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. Proc Natl Acad Sci U S A. 2016;113: 368–373.

Castellano S, Parra G, Sánchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, et al. Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A. 2014;111: 6666–6671.

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003;33 Suppl: 266–275.

Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton University Press; 1994.

Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 2005;6: 333–340.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4. doi:10.1186/s13742-015-0047-8

Chaubey G, Metspalu M, Choi Y, Mägi R, Romero IG, Soares P, et al. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol Biol Evol. 2011;28: 1013–1024.

Chaubey G, Metspalu M, Kivisild T, Villems R. Peopling of South Asia: investigating the caste-tribe continuum in India. Bioessays. 2007;29: 91–100.

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet. 1995;57: 133–149.

Chiaroni J, Underhill PA, Cavalli-Sforza LL. Y chromosome diversity, human expansion, drift, and cultural evolution. Proc Natl Acad Sci U S A. 2009;106: 20174-20179.

Cinnioğlu C, King R, Kivisild T, Kalfoğlu E, Atasoy S, Cavalleri GL, et al. Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet. 2004;114: 127–148.

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

Clarkson C, Shipton C, Weisler M. Front, back and sides: experimental replication and archaeological analysis of Hawaiian adzes and associated debitage. Archaeol Oceania. 2015;50: 71–84.

Cohen MN, Armelagos GJ. Paleopathology at the Origins of Agriculture. 2013.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006;38: 1251–1260.

Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, et al. Tracing past human male movements in Northern/Eastern Africa and Western Eurasia: new clues from Y-Chromosomal haplogroups E-M78 and J-M12. Mol Biol Evol. 2007;24: 1300–1311.

Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R. A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa. Am J Hum Genet. 2011;88: 814–818.

Dall WH, Harris GD. Correlation Papers, Neogene. Bulletin No. 84. Washington, DC. Geological Survey. 1892.

de Saint Pierre M, Bravi CM, Motti JMB, Fuku N, Tanaka M, Llop E, et al. An alternative model for the early peopling of southern South America revealed by analyses of three mitochondrial DNA haplogroups. PLoS One. 2012;7: e43486.

Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2012;10: 5–6.

Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli-Sforza LL, et al. Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. Proc Natl Acad Sci U S A. 1981;78: 5768–5772.

Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova J-L, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. Am J Hum Genet. 2016;98: 5–21.

Destro Bisol G, Anagnostou P, Batini C, Battaggia C, Bertoncini S, Boattini A, et al. Italian isolates today: geographic and linguistic factors shaping human biodiversity. J Anthropol Sci. 2008;86: 179–188.

Di Cristofaro J, Pennarun E, Mazières S, Myres NM, Lin AA, Temori SA, et al. Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. PLoS One. 2013;8: e76748.

Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, Cresta M, et al. Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol Phylogenet Evol. 2003;28: 387–395.

Dictionary OE. Oxford English dictionary online.

Dillehay TD, Ocampo C, Saavedra J, Sawakuchi AO, Vega RM, Pino M, et al. New Archaeological Evidence for an Early Human Presence at Monte Verde, Chile. PLoS One. 2015;10: e0141923.

Dillehay TD, Ramírez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD. Monte Verde: seaweed, food, medicine, and the peopling of South America. Science. 2008;320: 784–786.

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7: 214.

Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29: 1969–1973.

Dulik MC, Owings AC, Gaieski JB, Vilar MG, Andre A, Lennie C, et al. Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan- and Eskimoan-speaking populations. Proc Natl Acad Sci U S A. 2012;109: 8471–8476.

Dulik MC, Zhadanov SI, Osipova LP, Askapuli A, Gau L, Gokcumen O, et al. Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. Am J Hum Genet. 2012;90: 229-246.

Dupuy BM, Andreassen R, Flønes AG, Tomassen K, Egeland T, Brion M, et al. Y-chromosome variation in a Norwegian population sample. Forensic Sci Int. 2001;117: 163–173.

Excoffier L, Estoup A, Cornuet J-M. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics. 2005;169: 1727–1738.

Fiorito G, Di Gaetano C, Guarrera S, Rosa F, Feldman MW, Piazza A, et al. The Italian genome reflects the history of Europe and the Mediterranean basin. Eur J Hum Genet. 2016;24: 1056–1062.

Fornarino S, Pala M, Battaglia V, Maranta R, Achilli A, Modiano G, et al. Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. BMC Evol Biol. 2009;9: 154.

Forster P, Harding R, Torroni A, Bandelt HJ. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet. 1996;59: 935–945.

Forster P, Matsumura S. Evolution. Did early humans go north or south? Science. 2005;308: 965–966.

Forster P, Röhl A, Lünnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, et al. A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet. 2000;67: 182–196.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science. 2013;341: 565–569.

Francalacci P, Sanna D, Useli A, Berutti R, Barbato M, Whalen MB, et al. Detection of phylogenetically informative polymorphisms in the entire euchromatic portion of human Y chromosome from a Sardinian sample. BMC Res Notes. 2015;8: 174.

Froese R, Pauly D. Fishbase as a tool for comparing the life history patterns of flatfish. Neth J Sea Res. 1994;32: 235–239.

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, et al. An early modern human from Romania with a recent Neanderthal ancestor. Nature. 2015;524: 216–219.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al.. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514: 445–449.

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013;110: 2223–2227.

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534: 200–205.

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, et al. Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun. 2014;5: 5257.

Gattepaille LM, Jakobsson M. Combining markers into haplotypes can improve population structure inference. Genetics. 2012;190: 159–174.

Gilbert MTP, Jenkins DL, Götherstrom A, Naveran N, Sanchez JJ, Hofreiter M, et al. DNA from pre-Clovis human coprolites in Oregon, North America. Science. 2008;320: 786–789.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci U S A. 1995;92: 6723–6727.

Goudet J, Perrin N, Waser P. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. Mol Ecol. 2002;11: 1103–1114.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A. 2011;108: 11983–11988.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328: 710–722.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011;43: 1031–1034.

Grugni V, Battaglia V, Perego UA, Raveane A, Lancioni H, Olivieri A, et al. Exploring the Y chromosomal ancestry of modern Panamanians. PLoS One. 2015;10: e0144223.

Grugni V, Raveane A, Ongaro L, Battaglia V, Trombetta B, Olivieri A, et al. The first peopling of Americas: new insights on the Y chromosome haplogroup Q. 2017; (in preparation).

Gunnarsdóttir ED, Nandineni MR, Li M, Myles S, Gil D, Pakendorf B, et al. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nat Commun. 2011;2: 228.

Günther T, Jakobsson M. Genes mirror migrations and cultures in prehistoric Europe — a population genomic perspective. Curr Opin Genet Dev. 2016;41: 115–123.

Günther T, Valdiosera C, Malmström H, Ureña I, Rodriguez-Varela R, Sverrisdóttir ÓO, et al. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. Proc Natl Acad Sci U S A. 2015;112: 11917–11922.

Gusmão L, Sánchez-Diz P, Calafell F, Martin P, Alonso CA, Álvarez-Fernández F, et al. Mutation rates at Y chromosome specific microsatellites. Hum Mutat. 2005;26: 520–528.

Ha NT, Freytag S, Bickeboeller H. Coverage and efficiency in current SNP chips. European Journal of Human Genetics. 2014;22: 1124-1130.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522: 207–211.

Haber M, Gauguier D, Youhanna S, Patterson N, Moorjani P, Botigué LR, et al. Genome-wide diversity in the levant reveals recent structuring by culture. PLoS Genet. 2013;9: e1003316.

Haber M, Mezzavilla M, Xue Y, Comas D, Gasparini P, Zalloua P, et al. Genetic evidence for an origin of the Armenians from Bronze Age mixing of multiple populations. Eur J Hum Genet. 2016;24: 931–936.

Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, et al. The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol. 2014;32: 661-73.

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, et al. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol. 1998;15: 427–441.

Hammer MF, Zegura SL. The role of the Y chromosome in human evolutionary studies. Evol Anthr. 1996;5: 116–134.

Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 2013;9: e1003521.

Hawks J. The Y chromosome and the replacement hypothesis. Science. 2001;293: 567.

Hebsgaard MB, Phillips MJ, Willerslev E. Geologically ancient DNA: fact or artefact? Trends Microbiol. 2005;13: 212–220.

Heilprin A. The geographical and geological distribution of animals. Appleton,

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A Genetic Atlas of Human Admixture History. Science. 2014;343: 747–751.

Hershkovitz I, Marder O, Ayalon A, Bar-Matthews M, Yasur G, Boaretto E, et al. Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans. Nature. 2015;520: 216–219.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet. 1997;6: 799–803.

Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am J Hum Genet. 2001;69: 1113-1126

Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, et al. The timing and spatiotemporal patterning of Neanderthal disappearance. Nature. 2014;512: 306–309.

Hodoğlugil U, Mahley RW. Turkish Population Structure and Genetic Ancestry Reveal Relatedness among Eurasian Populations. Ann Hum Genet. 2012;76: 128–141.

Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-del-Molino D, et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc Natl Acad Sci U S A. 2016;113: 6886–6891.

Hooshiar Kashani B, Perego UA, Olivieri A, Angerhofer N, Gandini F, Carossa V, et al. Mitochondrial haplogroup C4c: a rare lineage entering America through the ice-free corridor? Am J Phys Anthropol. 2012;147: 35–39.

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci U S A. 1995;92: 532–536.

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am J Hum Genet. 2003;72: 659-670.

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5: e1000529.

Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature. 2017;546: 289–292.

Hughes JF, Page DC. The biology and evolution of mammalian Y chromosomes. Ann Rev Genet. 2015;49: 507-27.

Ilumäe AM, Reidla M, Chukhryaeva M, Järve M, Post H, Karmin M, et al. Human Y chromosome haplogroup N: a non-trivial time-resolved phylogeography that cuts across language families. Am J Hum Genet. 2016;99: 163–173.

Ingman M, Gyllensten U. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J Hered. 2001;92: 454–461.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000;408: 708–713.

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

International Society of Genetic Genealogy. Y-DNA Haplogroup Tree 2017

Janečka JE, Grassman LI, Honeycutt RL, Tewes ME. Whole Genome Amplification for Sequencing and Applications in Conservation Genetics. J Wildl Manage. 2007;71: 1357–1360

Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29: 217–222.

Jenkins DL, Davis LG, Stafford TW Jr, Campos PF, Hockett B, Jones GT, et al. Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves. Science. 2012;337: 223–228.

Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. Human Evolutionary Genetics, Garland Science.

Jobling M. Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet. 1998;7: 643–653.

Jobling MA, Pandya AR, Tyler-Smith CH. The Y chromosome in forensic analysis and paternity testing. Int J Leg Med. 1997;110: 118-124.

Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet. 2017;18: 485–497.

Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003;4: 598–612.

Johnson NA, Lachance J. The genetics of sex chromosomes: evolution and implications for hybrid incompatibility. Ann N Y Acad Sci. 2012;1256: 1–22.

Jones ER, Gonzalez-Fortes G, Connell S, Siska V, Eriksson A, Martiniano R, et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat Commun. 2015;6: 8912.

Jota MS, Lacerda DR, Sandoval JR, Vieira PPR, Ohasi D, Santos-Júnior JE, et al. New native South American Y chromosome lineages. J Hum Genet. 2016;61: 593–603.

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet. 2001;69: 615–628.

Karafet T, Zegura SL, Vuturo-Brady J, Posukh O, Osipova L, Wiebe V, et al. Y chromosome markers and Trans-Bering Strait dispersals. Am J Phys Anthropol. 1997;102: 301–314.

Karafet TM, Osipova LP, Gubina MA, Posukh OL, Zegura SL, Hammer MF. High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum Biol. 2002;74: 761–789.

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, et al. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am J Hum Genet. 1999;64: 817–831.

Karmin M, Saag L, Vicente M, Sayres MA, Järve M, Talas UG, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25: 459–466.

Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, et al. A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet. 2004;74: 1183–1197.

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, et al. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet. 2000;66: 1580–1588.

Kayser M. The human genetic history of Oceania: near and remote views of dispersal. Curr Biol. 2010;20: 194–201.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 2012;3: 698.

Kemp BM, Malhi RS, McDonough J, Bolnick DA, Eshleman JA, Rickards O, et al. Genetic analysis of early holocene skeletal remains from Alaska and its implications for the settlement of the Americas. Am J Phys Anthropol. 2007;132: 605–621.

Kivisild T, Metspalu M, Bandelt H-J, Richards M, Villems R. The World mtDNA Phylogeny. Human mitochondrial DNA and the evolution of Homo sapiens;2015: 149-179.

Kivisild T. The study of human Y chromosome variation through ancient DNA. Hum Genet. 2017;136: 529–546.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488: 471–475

Kovacevic L, Tambets K, Ilumäe A-M, Kushniarevich A, Yunusbayev B, Solnik A, et al. Standing at the gateway to Europe--the genetic structure of Western balkan populations based on autosomal and haploid markers. PLoS One. 2014;9: e105090.

Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, et al. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature. 2016;530: 429–433

Kushniarevich A, Utevska O, Chuhryaeva M, Agdzhoyan A, Dibirova K, Uktveryte I, et al. Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. PLoS One. 2015;10: e0135820.

Lahr MM, Foley RA. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. Am J Phys Anthropol. 1998;27: 137–176.

Landsteiner K. Ueber Agglutinationserscheinungen normalen menschlichen Blutes. 1901.

Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8: e1002453.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536: 419-424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513: 409–413.

Lell JT, Sukernik RI, Starikovskaya YB, Su B, Jin L, Schurr TG, et al. The dual origin and Siberian affinities of Native American Y chromosomes. Am J Hum Genet. 2002;70: 192–206.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015;519: 309–314.

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475: 493–496.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319: 1100–1104.

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

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, et al. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig Genet. 2014;5: 13.

Liu EY, Li M, Wang W, Li Y. MaCH-admix: genotype imputation for admixed populations. Genet Epidemiol. 2013;37: 25–37.

Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Sci Adv. 2016;2: e1501385.

Llorente MG, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, et al. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science. 2015;350: 820–822.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193: 1233–1254.

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

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. Nature. 2016;538: 207–214.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538: 201–206.

Manley WF. Postglacial flooding of the Bering land bridge: a geospatial

Martini F. Archeologia del Paleolitico: storia e culture dei popoli cacciatori-raccoglitori. Carocci Editore. Milan. 2013.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528: 499–503.

McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005;433: 733–736.

Mellars P, Gori KC, Carr M, Soares PA, Richards MB. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc Natl Acad Sci U S A. 2013;110: 10699–10704.

Mendez FL, Krahn T, Schrack B, Krahn AM, Veeramah KR, Woerner AE, et al. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet. 2013;92: 454–459.

Mendez FL, Poznik GD, Castellano S, Bustamante CD. The Divergence of Neandertal and Modern Human Y Chromosomes. Am J Hum Genet. 2016;98: 728–734.

Mendez FL, Watkins JC, Hammer MF. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. Mol Biol Evol. 2012;29: 1513–1520.

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, et al. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet. 2004;5: 26.

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet. 2011;89: 731–744.

Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga JL, et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature. 2013;505: 403–406.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338: 222–226.

Montinaro F, Busby GBJ, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling the hidden ancestry of American admixed populations. Nat Commun. 2015;6: 6596.

Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in Enzymology. 1987;155: 335–350.

Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J. A DNA probe detecting multiple haplotypes of the human Y chromosome. Am J Hum Genet. 1986;38: 407–418.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12: 443–451.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456: 98–101.

Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey J, Buckley T. Bayesian phylogenetic analysis of combined data. Syst Biol. 2004;53: 47–67.

O'Connell JF, Allen J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. J Archaeol Sci. 2015;56: 73–84.

O'Rourke DH, Raff JA. The human genetic history of the Americas: the final frontier. Curr Biol. 2010;20: R202–7.

Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CW, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature. 2014;507: 225–228.

Olivieri A, Sidore C, Achilli A, Angius A, Posth C, Furtwängler A, et al. Mitogenome diversity in Sardinians: a genetic window onto an island's past. Mol Biol Evol. 2017;34: 1230–1239.

Oppenheimer S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Philos Trans R Soc Lond B Biol Sci. 2012;367: 770–784.

Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 2015;16: 395–408.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538: 238–242.

Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015;96: 986–991.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, et al. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. Am J Hum Genet. 2012;90: 915–924.

Pala, M., Soares, P., Chaubey, G., & Richards, M. (2015). Archaeogenetics. In G. Barker & C. Goucher (Eds.), The Cambridge World History (The Cambridge World History, pp. 26-54). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511978807.003.

Parolo S, Lisa A, Gentilini D, Di Blasio AM, Barlera S, Nicolis EB, et al. Characterization of the biological processes shaping the genetic structure of the Italian population. BMC Genet. 2015;16: 132.

Paschou P, Drineas P, Yannaki E, Razou A, Kanaki K, Tsetsos F, et al. Maritime route of colonization of Europe. Proc Natl Acad Sci U S A. 2014;111: 9211–9216.

Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A et al. Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol. 2009;19: 1–8.

Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Hooshiar Kashani B, et al. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Res. 2010;20: 1174–1179.

Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Kashani BH et al. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Res. 2010;20: 1174–1179.

Pessina A, Tiné V. Archeologia del neolitico: l'Italia tra VI e IV millennio a.C. 2008. Carocci Editore. Milano. 2008

Petraglia MD, Dennell R. Global Expansion 300,000-8000 years ago, Asia. Encyclopedia of quaternary science. 2007. 107–118.

Piazza A, Cappello N, Olivietti E, Rendine S. A genetic history of Italy. Ann Hum Genet. 1988;52: 203-13.

Pickrell J, Reich D. Towards a new history and geography of human genes informed by ancient DNA. Trends Genet. 2014;30: 377-389.

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nat Commun. 2012;3: 1143.

Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8: e1002967.

Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. Curr Biol. 2016;26: 827–833.

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science. 2013;341: 562–565.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 2016;48: 593-599.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499: 471–475.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5: e1000519.

Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11: 459–463.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155: 945–959.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505: 43–49.

Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005;15: R159–60.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

Quintana-Murci L, Semino O, Minch E, Passarimo G, Brega A, Santachiara-Benerecetti AS. Further characteristics of proto-European y chromosomes. Eur J Hum Genet. 1999;7: 603–608.

Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS et al. The genetic prehistory of the New World Arctic. Science. 2014;345: 1255832.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014;505: 87–91.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 2015;349: aab3884.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005;102: 15942–15947.

Rambaut A, Suchard MA, Xie D & Drummond AJ. Tracer v1.6. 2014. (http://tree.bio.ed.ac.uk/software/tracer)

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506: 225-229

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science. 2011;334: 94–98.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463: 757–762.

Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD et al. The ancestry and affiliations of Kennewick Man. Nature. 2015;523: 455-458.

Rebolledo-Jaramillo B, Su MS, Stoler N, McElhoe JA, Dickins B, Blankenberg D, et al. Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A. 2014;111: 15474–15479.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468: 1053–1060.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012;488: 370–374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461: 489–494.

Relethford JH. Ancient DNA and the origin of modern humans. Proc Natl Acad Sci U S A. 2001;98: 390–391.

Relethford JH. Geostatistics and spatial analysis in biological anthropology. Am J Phys Anthropol. 2008;136: 1–10.

Richards MB, Soares P, Torroni A. Palaeogenomics: mitogenomes and migrations in Europe's past. Curr Biol. 2016;26: 243–246.

Richter D, Grün R, Joannes-Boyau R, Steele TE, Amani F, Rué M, et al. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. Nature. 2017;546: 293–296.

Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, et al. Improved calibration of the human mitochondrial clock using ancient genomes. Mol Biol and Evol. 2014;31: 2780-2792.

Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, et al. Improved calibration of the human mitochondrial clock using ancient genomes. Mol Biol and Evol. 2014;31: 2780-2792.

Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, Soares P. The first modern human dispersals across Africa. PLoS One. 2013;8: e80031.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26.

Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, den Velde Boots PM, et al. Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. PLoS One. 2012;7: e41634.

Rogers AR, Bohlender RJ, Huff CD. Early history of Neanderthals and Denisovans. Proc Natl Acad Sci U S A. 2017;114: 9859–9863.

Rosenberg NA. Genetic Structure of Human Populations. Science. 2002;298: 2381–2385.

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, et al. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet. 2000;67: 1526–1543.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409: 928–933.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. Am J Hum Genet. 2008;82: 290–303.

Santos C. Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: a model using families from the Azores Islands (Portugal). Mol Biol Evol. 2005;22: 1490–1505.

Sarno S, Boattini A, Pagani L, Sazzini M, De Fanti S, Quagliariello A, et al. Ancient and recent admixture layers in Sicily and Southern Italy trace multiple migration routes along the Mediterranean. Sci Rep. 2017;7: 1984.

Sazzini M, Gnecchi Ruscone GA, Giuliani C, Sarno S, Quagliariello A, De Fanti S, et al. Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. Sci Rep. 2016;6: 32513.

Schadt EE, Turner S, Kasarskis A. A window into third generation sequencing. Hum Mol Genet. 2010;20: 853–853.

Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, et al. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nat Commun. 2016;7: 10408.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017.

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;338: 374–379.

Schurr TG, Sherry ST. Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am J Hum Biol. 2004;16: 420–439.

Scozzari R, Massaia A, Trombetta B, Bellusci G, Myres NM, Novelletto A, et al. An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res. 2014;24: 535–544.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346: 1113–1118.

Semino O, Passarino G, Quintana-Murci L, Liu A, Béres J, Czeizel A, et al. MtDNA and Y chromosome polymorphisms in Hungary: inferences from the palaeolithic, neolithic and Uralic influences on the modern Hungarian gene pool. Eur J Hum Genet. 2000;8: 339–346.

Shi M, Bai R, Bai L, Yu X. Population genetics for Y-chromosomal STRs haplotypes of Chinese Xibe ethnic group. Forensic Sci Int Genet. 2011;5: e119–21.

Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sánchez-Quinto F, et al. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. PLoS genetics. 2014;10: e1004353.

Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, et al. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. Science. 2014;344: 747–750.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science. 2012;336: 466–469.

Slon V, Hopfe C, Weiß CL, Mafessoni F, de la Rasilla M, Lalueza-Fox C, et al. Neandertal and Denisovan DNA from Pleistocene sediments. Science. 2017;356: 605–608.

Smith DG, Malhi RS, Eshleman J, Lorenz JG, Kaestle FA. Distribution of mtDNA haplogroup X among Native North Americans. Am J Phys Anthropol. 1999;110: 271–284.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, et al. The Archaeogenetics of Europe. Curr Biol. 2010;20: 174–183.

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, et al. Ancient voyaging and Polynesian origins. Am J Hum Genet. 2011;88: 239–247.

Stoneking M, Delfin F. The human genetic history of East Asia: weaving a complex tapestry. Curr Biol. 2010;20: 188–193.

Stringer C. Modern human origins: progress and prospects. Philos Trans R Soc Lond B Biol Sci. 2002;357: 563–579.

Stringer CB, Andrews P. Genetic and fossil evidence for the origin of modern humans. Science. 1988;239: 1263–1268.

Takahata N, Lee SH, Satta Y. Testing multiregionality of modern human origins. Mol Biol Evol. 2001;18: 172–183.

Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ et al. Beringian standstill and spread of Native American founders. PLoS One. 2007;2: e829.

Tang T, Lu J, Huang J, He J, McCouch SR, Shen Y, et al. Genomic variation in rice: genesis of highly polymorphic linkage blocks during domestication. PLoS Genet. 2006;2: e199.

Tattersall I. Out of Africa: modern human origins special feature: human origins: out of Africa. Proc Natl Acad Sci U S A. 2009;106: 16018–16021.

Taylor J. Muslims in medieval Italy: The colony at Lucera. Rowman & Littlefield Inc. Lanhyam, Maryland. 2003

Templeton AR. Gene flow, haplotype patterns and modern human origins. eLS. 2007.

The Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 2002;12: 339–348.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A. 2000;97: 7360–7365.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, et al. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science. 1996;271: 1380–1387.

Tobler R, Rohrlach A, Soubrier J, Bover P, Llamas B, Tuke J, et al. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. Nature. 2017;544: 180–184.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. Harvesting the fruit of the human mtDNA tree. Trends Genet. 2006;22: 339–345.

Torroni A, Brown MD, Lott MT, Newman NJ, Wallace DC, The Cuba Neuropathy Field Investigation Team. African, Native American, and European mitochondrial DNAs in Cubans from Pinar del Rio Province and implications for the recent epidemic neuropathy in Cuba. Hum Mutat. 1995;5: 310–317.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, et al. Classification of European mtDNAs from an analysis of three European populations. Genetics. 1996;144: 1835-1850.

Torroni A, Neel JV, Barrantes R, Schurr TG, Wallace DC. Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc Natl Acad Sci U S A. 1994;91: 1158–1162.

Torroni A, Petrozzi M, D'Urbano L, Sellitto D, Zeviani M, Carrara F, et al. Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. Am J Hum Genet. 1997 60:1107.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, et al. Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet. 1993;53: 563–590.

Trombetta B, D'Atanasio E, Massaia A, Ippoliti M, Coppa A, Candilio F, et al. Phylogeographic refinement and large scale genotyping of human Y chromosome haplogroup E provide new insights into the dispersal of early pastoralists in the African continent. Genome Biol Evol. 2015;7: 1940–1950.

Uetz, P, Freed, P, Jirí Hošek. The Reptile Database. 2016. (http://www.reptile-database.org)

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci U S A. 1996;93: 196–200.

Underhill PA, Kivisild T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 2007;41: 539–564.

Underhill PA, Poznik GD, Rootsi S, Järve M, Lin AA, Wang J, et al. The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. Eur J Hum Genet. 2015;23: 124–131.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, et al. Y chromosome sequence variation and the history of human populations. Nat Genet. 2000;26: 358–361.

United Nations Environment Programme. 2004. (http://www.unenvironment.org/)

Van De Mieroop M. A History of Ancient Egypt. 2010. Wiley-Blackwell. New jersey, United Sates of America.

van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30: 386–394.

van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MHD. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat. 2014;35: 187–191.

Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, Destro-Bisol G, et al. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol Biol Evol. 2012;29: 617–630.

Veerappa AM, Padakannaya P, Ramachandra NB. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Funct Integr Genomics. 2013;13: 285-293

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. African populations and the evolution of human mitochondrial DNA. Science. 1991;253: 1503–1507.

Wallace AR. 1823-1913. The geographical distribution of animals: with a study of the relations of living and extinct faunas as elucidating the past changes of the earth's surface Volume V. 2. Arkose Press; 2015.

Weber JL, Wong C. Mutation of human short tandem repeats. Hum Mol Genet. 1993;2: 1123–1128.

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res. 2013;23: 388–395.

Wei W, Luo HB, Yan J, Hou YP. Exploring of new Y-chromosome SNP loci using pyrosequencing and the SNaPshot methods. Int J Legal Med. 2012;126: 825–833.

Weidenreich F. Some problems dealing with ancient man. Am Anthropol. 1940;42: 375–383.

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, et al. The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci U S A. 2001;98: 10244–10249.

Wilder JA, Mobasher Z, Hammer MF. Genetic evidence for unequal effective population sizes of human females and males. Mol Biol Evol. 2004;21: 2047–2057.

Willuweit S, Roewer L. International Forensic Y chromosome user group. Y chromosome haplotype reference database (YHRD): update. Forensic Sci Int Genet. 2007;1: 83–87.

Wilmsen EN. An outline of early man studies in the United States. Am Antiq. 1965;31: 172–192.

Wu XZ. A well-preserved cranium of an archaic type of early Homo sapiens from Dali, China. Sci Sin. 1981;24: 530–541.

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science. 2015;348: 242–245.

Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, et al. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol. 2009;19: 1453–1457.

Yunusbayev B, Metspalu M, Järve M, Kutuev I, Rootsi S, Metspalu E, et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol. 2012;29: 359–365.

Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, et al. The genetic legacy of the expansion of Turkic-Speaking nomads across Eurasia. 2014. doi:10.1101/005850

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. Mol Biol Evol. 2004;21: 164–175.

Zhivotovsky LA, Rosenberg NA, Feldman MW. Features of evolution and expansion of modern humans, inferred from genome-wide microsatellite markers. Am J Hum Genet. 2003;72: 1171–1186.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet. 2004;74:50-61.

Zhivotovsky LA, Underhill PA. On the evolutionary mutation rate at Y-chromosome STRs: comments on paper by Di Giacomo et al. (2004). Hum Genet. 2005;116: 529–532.

Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L,et al. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Mol Biol Evol. 2011;28: 717–727.

Zhou D, Udpa N, Ronen R, Stobdan T, Liang J, Appenzeller O, et al. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. Am J Hum Genet. 2013;93: 452–462

List of original manuscripts

Raveane A[#], Montinaro F[#], Aneli S[#], Lancioni H, Mulas A, Grugni V, Cardinali I, Zoledziewska M, Baali A, Barlera S, Boncoraglio G, Brisighelli F, Di Blasio AM, Cherkaoui M., Di Gaetano C., Dugoujon JM, Guerrera S, Kivisild T, Melhaoui M, Pagani L, Parolo S, Paschou P, Piazza A, Pascali V, Peyret-Guzzon M, Ricaut FX, Stamatoyannopoulos G, Cucca F, Angius A, Torroni A, Metspalu M, Semino O, Hellenthal G, Matullo G*, Achilli A*, Olivieri A*, Capelli C*. Ancient and recent genomic ancestries of the Italian population. (In preparation)

Grugni V, **Raveane A**, Mattioli F, Battaglia V, Sala C, Toniolo D, Ferretti L, Gardella R, Achilli A, Olivieri A, Torroni A, Passarino G, Semino O (2017). Reconstructing the genetic history of Italians: new insights from a male (Y-chromosome) perspective. Ann Hum Biol (in press).

Balanovsky O, Gurianov V, Zaporozhchenko V, Balaganskaya O, Urasin V, Zhabagin M, Grugni V, Al-Zahery N, **Raveane A**, Wen O, Yan S, Wang X, Marafi P, Koshel S, Semino O, Tyler-Smith C, Balanovska E (2017). Phylogeography of human Y-chromosome haplogroup Q3-L275 from an academic/citizen science collaboration. BMC Evol Biol 17:18.

Grugni V, Battaglia V, Perego UA, Raveane A, Lancioni H, Olivieri A, Ferretti L, Woodward SR, Pascale J M, Cooke R, Myres N, Motta J, Torroni A, Achilli A, Semino O (2015). Exploring the Y chromosomal ancestry of modern Panamanians. PLoS One. 10: e0144223.