Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Biodiversità del Evoluzione

Ciclo XXVI

Settore Concorsuale di afferenza: 05/B1 - Zoologia e Antropologia

Settore Scientifico disciplinare: BIO/08 - Antropologia

THE GENETIC HISTORY OF ITALIANS: NEW INSIGHTS FROM UNIPARENTALLY-INHERITED MARKERS

Presentata da
Stefania Sarno

Coordinatore Dottorato

Prof.ssa Barbara Mantovani

Relatore

Prof. Davide Pettener

Esame Finale – Anno 2014

Table of contents

ABBREVIATIONS

ABO	**ABO** blood group system
A.D.	**A**nno **D**omini
aDNA	**a**ncient **DNA**
AMH	**A**natomically **M**odern **H**uman
ASD	**A**verage **S**quare **D**istance
B.C.	**B**efore **C**hrist
DNA	**D**ideoxyribo**n**ucleic **A**cid
HG	**H**aplo**g**roup
HLA	**H**istocompatibility Leucocyte **A**ntigen
HVS	**H**yper **V**ariable **S**egment
IBD	**I**dentical-**B**y-**D**escent
IBS	**I**dentity-**B**y-**S**tate
IW	**I**mpressed **W**are culture
KYBP	**K**ilo **Y**ears **B**efore **P**resent
LBK	**L**inear **B**and**K**eramik culture
LGM	**L**ast **G**lacial **M**aximum
mtDNA	**m**itochondrial **DNA**
NRY	**N**on-recombining **R**egion of the **Y**-chromosome
PC	**P**rincipal **C**omponent
SNP	**S**ingle **N**ucleotide **P**olymorphism
STR	**S**hort **T**andem **R**epeat
TMRCA	**T**ime of the **M**ost **R**ecent **C**ommon **A**ncestors
YBP	**Y**ear **B**efore **P**resent

GENERAL INTRODUCTION AND THESIS RATIONALE

1. Reconstructing human evolutionary history through genetic tools

Detecting and interpreting the existing patterns of genetic variation within and between human populations has long been one of the major aims of population genetic studies. In the last decades, the increasing availability of genetic data from worldwide sets of human populations has offered new powerful tools for population geneticists to ask new questions about our past (Wilkins, 2006). Studying the human populations genetic structure and the spatial distribution of genetic variation has helped increasing our knowledge on human population prehistory and proto-history (Ramachandran et al., 2010). It also allowed shedding new lights into the main demographic processes which led to the present-day patterns of genetic variability within and among human groups. The existence of various classes of molecular markers, differing from one another for both their patterns of inheritance and evolving mutation rates (Garrigan and Hammer, 2006), has made it possible not only to trace the genetic history of human populations separated either in ancient or more recent times, but also to investigate the evolutionary forces that have subsequently addressed their differentiation over time (Sazzini et al., 2014).

Two portions of our genome, commonly referred to as uniparental markers (namely Y-chromosome and mitochondrial DNA), have long been used in population genetic studies to reconstruct and compare the male and female genetic histories and demographic patterns at different geographical and cultural scales (Wilkins et al. 2006). Since they are uniparentally transmitted - hence free from the reshuffling effect of recombination - the subsequent occurrence of neutral mutations in these genetic systems defines maternal and paternal lineages whose origin (phylogenetic relationships) and diffusion pattern (phylogeographic distribution) are traceable back in time and space, actually reflecting the population past history and mobility (Jobling et al., 2004, Underhill and Kivisild 2007, Jobling et al., 2013). In addition, by largely escaping selective pressures, these markers offer the chance to specifically investigate the effect of genetic drift, admixture, geographic isolation and/or cultural practices on the present-day frequency and geographic distribution patterns of genetic variation (Sazzini et al., 2014).

More recently, genome-wide autosomal SNP data and X-chromosome genetic markers have offered new important tools to reveal population genetic sub-structures and to test detailed hypotheses about human demographic histories, despite being individually less reliable in reconstructing traceable genealogies (Wilkins et al., 2006; Henn et al., 2010).

Since the first pioneer studies by Cavalli-Sforza and colleagues (1994) on the “History and Geography of Human Genes”, the European country has attracted particular interest of population genetic studies. Classical, uniparental and autosomal markers (Cavalli-Sforza et al., 1994; Barbujani et al., 1994; Rosser et al. 2000; Semino et al, 2000; Chikhi et al. 2002; Lao et al., 2008; Novembre et al., 2008; Balaresque et al., 2010) agreed in revealing a principal South-East to North-West (SE-NW) distribution pattern of genetic variation within Europe, which has been mainly associated to the major prehistoric demographic events that affected the peopling history of the continent over time: namely the first colonisation by Anatomically Modern Humans (AMH), the post-glacial re-expansion from Southern-European refuges, and the Neolithic diffusion of agriculture from the Near East. Within this European genetic continuum, some internal population genetic sub-structures have been however invoked at more fine-grained regional scales (for instance within Finland and Italy, Lao et al., 2008; Nelis et al., 2009), and different geographic, cultural or linguistic isolates (such as Basques, Finns, Sardinians and Jewish; Bertranpetit et al., 1995; Thomas et al., 2002; Fraumene et al., 2006; Calò et al, 2008; Palo et al., 2009; Novembre and Ramachandran, 2011; Ostrer and Skorecki, 2013) confirmed to be genetic isolates as well. Therefore, if the genetic landscape of present-day European and Mediterranean populations has been widely explored under a continental point of view, on the other hand the extent to which continent-wide trends are re-shaped or modified at micro-geographical levels, and the measure of geographic, cultural or demographic impacts on local patterns of genetic variation, have been only marginally explored (Beleza et al. 2005, Sanchez-Faddeev et al. 2013).

2. A multi-level analysis approach in the study of Italian population history and genetic structure: from *macro-* to *micro-geographic* perspectives

Among the different European countries, the Italian Peninsula represents a key area of investigation to explore the genetic variability at different geographical levels, due to the valuable richness of its historical heritage and the remarkable cultural heterogeneity and environmental complexity (Destro-Bisol et al., 2008; Capocasa et al., 2014; Toso, 2014).

The territory of today’s Italy is one of the three major peninsulas located in the southern fringe of Europe (the other two being the Balkans and Iberia), particularly stretching for more than 1,000 km from the Alps mountain range in the north to the centre of the Mediterranean Sea in the south. Its strategic geographical location, connecting Central Europe with the Mediterranean Basin, has made Italy one of the most important access points to the European continent. Extensive population movements have occurred through, from and to the Peninsula during all the principal phases of the peopling of the continent, thus leading to multiple contacts and admixture events between different peoples and cultures over time. This makes the study of the Italian genetic structure and population variability an effective tool to understand the historic and demographic processes that have characterised the peopling history of Europe and the Mediterranean since Palaeolithic times up to the present.

Besides being one of the first areas populated by AMH approximately 43,000-45,000 YBP (Benazzi et al., 2011), during the Last Glacial Maximum (LGM) the Italian Peninsula was involved in the southward contraction of human groups from the Northern and Central Europe (Banks et al., 2008), a process which undoubtedly contributed to heavily reshape the early Palaeolithic inheritances of the pre-existing southern European populations (Sazzini et al., 2014). Italian Neolithic and Metal Ages subsequently represented an intricate tapestry of different population movements and cultural strata, which contributed further to complicate the pattern of genetic variation currently detectable in the Italian Peninsula (Pessina and Tinè, 2008). Before the Roman rule, Italy actually appeared as a patchwork of different peoples, either of unresolved origins (e.g. Etruscans, Ligurians, Veneti) or otherwise originating from specific migration processes (Celts in North-Western Italy; Greeks and Phoenicians in Southern Italy and Sicily) which left more or less evident genetic traces in the current Italian gene pool (Pesando, 2005). In more recent times, the Roman Empire and subsequent historical invasions (e.g. Barbaric and Arabic) may have further affected the present-day population genetic heterogeneity.

In addition to these multi-layered historical incomings, the environmental complexity of the Italian territory (ranging from the tundra biome in the Alps to the Mediterranean-climate regions in the south; Blasi et al., 2005) as well as its cultural heterogeneity (Capocasa et al., 2014, Toso et al., 2014), have represented additional sources of genetic variability, being potentially responsible for different paths of isolation or population mobility at more fine-grained local levels. Geographic constraints such as mountain chains (e.g. the Alps; Coia et al., 2013 and reference therein) or deep seas (e.g. Sardinia; Calò et al., 2008 and reference therein) could have acted as physical barriers to gene flow, representing limits to reproductive exchanges. At the same time, cultural factors (e.g. languages, socio-economic stratification, religion, etc.) might have additionally influenced human mating patterns, encouraging or avoiding preferential mate choices, and therefore constituting evolutionary forces able to generate significant genetic sub-structures within and among populations (Sazzini et al., 2014).

The extant patterns of Italian genetic variation can therefore be ascribed both i) to the broad-scale impact of pre-historical and historical migration processes, and ii) to the fine-scale extent to which local variability, in terms of cultural and geographic heterogeneity, intervened in preserving or confounding these historical genetic layers. Therefore, a multi-level analysis approach, aimed at combining both macro- and micro-geographic investigation of the Italian population variability, can represent a fruitful methodological framework to more deeply explore the historical and bio-demographic processes which have shaped human current biodiversity. At the same time, it might also help achieving a more complete characterisation of the micro-evolutionary effects of socio-cultural and geographic factors on the observed population genetic structures.

3. Work outline

In this work we exploit the genetic information provided by uniparental markers to address different aspects of the population history and genetic variability of Italy, by particularly adopting a "*multi-level analysis approach*". We start our analyses at a general level (thus looking at the "broad picture" that is the genetic landscape of Italy within Europe and the Mediterranean) and then we zoom on into specific case studies at micro-geographical local scales (thus focusing on specific population events that modified and/or reshaped the more general genetic landscape). Integrating different yet complementary perspectives of analysis, stems from the aim to achieve a more complete and accurate picture of the Italian population evolutionary history.

On one hand, addressing questions about the general composition and distribution pattern of genetic variation within the Italian Peninsula (*Article 1*) is important to understand the extent to which older or more recent migration processes, population introgressions and/or isolation events contributed to shape the current genetic landscape. In addition, the observed geographic patterns can provide further insights into the presence of discontinuities in the population genetic structure, potentially accounting for either different historical strata or sex-biased paternal and maternal contributions.

On the other hand, focusing on Italian specific local contexts (e.g. Sicily and Southern Italy, *Article 2*) or exploring the genetic variability within peculiar ethno-linguistic or socio-economic enclaves (e.g. the Albanian-speaking Arbereshe - *Article 3* and the Partecipanza of San Giovanni in Persiceto - *Article 4,* respectively), will offer the possibility to look more deeply into specific geographic or historical layers of the Italian variability and i) examine the different continental and within-continental contributions to the maternal and paternal gene pools, or ii) clarify the effects of socio-cultural and geographic factors on the current genetic composition.

Based on this approach, in this dissertation we are going to present the results of 4 papers, which are therefore intended to explore the Italian genetic variability at different geographical levels and time scales.

In the first part we used mtDNA and Y-chromosome genetic systems to update our knowledge about the population genetic history of Italy. To firstly assess the overall Italian genetic diversity, we analysed uniparentally-inherited markers in ~900 individuals from an extensive sampling across the Italian Peninsula, Sardinia and Sicily. In order to obtain a realistic and representative picture of the Italian genetic diversity, we paid special attention to two fundamental points of our research design, which are: i) the specificity of sampling strategy (built on a preliminary surname-based analysis); and ii) the unprecedented in-depth resolution level in the analysis of the uniparental molecular markers. Concerning the sampling strategy, a broad scientific consensus has been reached that adequate sampling coverage and accurate selecting criteria are essential elements to obtain representative results and to draw reliable conclusions about human population history. As for the molecular strategy, the combined use of slow- and fast-evolving genetic markers, together with innovative methods of analysis, has largely proved to be a suitable approach to explore and interpret the distribution patterns of genetic variation, even at fine-scale levels of analysis (Underhill and Kivisild 2007; Larmuseau et al., 2011).

Our major goal in this first study was to explore the genetic structure of Italy and try to identify the times and the population movements in the origin of the current genetic diversity. In addition, the joint analysis of maternal and paternal genetic components provides complementary information about female- and male-specific aspects of genetic variation. The results of this study are described in *Article 1*. Besides confirming the well-known outlier position of Sardinia, a major finding of our analyses is the identification of sex-biased patterns of genetic diversity within the Italian Peninsula, which reveal different historical and demographic dynamics for males and females. The NW-SE internal genetic sub-structure observed for the Y-chromosome could be linked with different Neolithisation processes and post-Neolithic incomings in continental Italy. On the contrary, the homogenous distribution of mtDNA genetic variation could reflect more ancient peopling events or a higher female mobility.

Based on the overall picture of the Italian genetic landscape, as a second step of our research we then directed our attention at exploring the genetic variation in Italian strategic local contexts, with the aim to test the different contributions of other European and Mediterranean populations to the current genetic variation, and to provide new clues about past population contacts and genetic interactions. The setting of this second investigation is Sicily and Southern Italy. In fact, due to its strategic geographical position between three different continents, this area has proved to be one of the most important Mediterranean crossroads in the peopling history of Europe, receiving the passage of various human groups both in pre-historical and historical times.

In *Article 2* we typed mtDNA and NRY genetic markers for a new set of accurately selected individuals coming from other provinces of Sicily than those already considered in the first study. These new data were integrated with the dataset of Sicilian and Southern Italian populations retrieved from the *Paper 1,* in such a way as to obtain the high sampling coverage requested to faithfully and exhaustively represent the whole variability of the area.

Driven by the contrasting results previously obtained in literature about the genetic structure of Sicily, this detailed dataset was first used to check the presence (Romano et al., 2003; Di Gaetano et al., 2009) or the absence (Rickards et al., 1998) of an east-west internal genetic heterogeneity, which potentially accounts for different historical contributions between the eastern and the western parts of the island.

Subsequently, our results were compared with selected published data from Central, Western and Southern Europe, as well as from North Africa and the Levant, in order to assess the temporal and geographic potential impacts of other European and/or Mediterranean populations on the current homogeneous Sicilian and Southern Italian maternal and paternal genetic pools. This also allowed us to test, on the wider geographical scale offered by the Mediterranean context, two important outcomes of the previous study (*Paper 1*) that are: i) the sex-biased genetic patterns observed within continental Italy, and ii) the differences between the North-Western and South-Eastern parts of the Italian Peninsula in the patterns of genetic similarities with Europe and the Mediterranean. While a more homogeneous genetic landscape has been confirmed for mtDNA, Y-chromosome results show a significant genetic differentiation between the North-Western and South-Eastern part of the Mediterranean Basin, the Italian Peninsula occupying an intermediate position therein. In particular, Sicily and Southern Italy revealed a paternal genetic background shared with the Balkan Peninsula, and the time estimates of the most important Y-chromosome lineages show paternal genetic traces of post-Neolithic population invasions of the Southern Italian territory (in particular attesting the importance of the Greek domination during the 8th and 9th centuries BC).

Based on these results, as a third step of our analysis we then explored some interesting aspects of the long-date relationship between Southern Italy and the Balkans, focusing on well-known historical events and trying to link them with the current patterns of genetic variation. Among the different migration processes and cultural exchanges connecting the two coasts of the Adriatic Sea, some population movements have indeed left evident signatures in the so-called ethno-linguistic minorities of Sicily and Southern Italy (e.g. the Greek-speaking Grecani and the Albanian-speaking Arbereshe).

The study of *marginal populations* (Soulè, 1973), such as the ethno-linguistic minorities, may provide unique opportunities to reconstruct recent patterns emerged from open-populations located in the same area and to investigate the evolutionary factors that shaped human cultural and genetic variation at relatively small spatial and temporal scales. Compared to open-populations, the ethno-linguistic groups (originated from a restricted number of founders and remained at least in part isolated from the subsequent events of admixture; Destro-Bisol et al. 2008), can in fact represent simplified models to specifically focus on the interplay between geographic and cultural factors, reducing the confounding noise otherwise ascribable to the multi-layered peopling history of a population. In addition, the condition of linguistic isolation (potentially acting as a genetic barrier between different ethnic groups) may have helped these communities to preserve a more direct genetic link with the populations of origin (Capocasa et al. 2014). As a consequence, such ethno-linguistic groups, besides giving an important contribution to the biodiversity of their recipient region, are also more likely to maintain genetic features otherwise lost (or diluted) in the surrounding open-population groups.

Among the various ethno-linguistic minorities established in the Italian territory (see Capocasa et al, 2014, on *Appendix I* for a more complete overview), the Albanian speaking Arbereshe are one of the most numerous, having also shown an exceptional resilience in preserving their distinct cultural traits (Arberisht language and Greek Orthodox religion; Tagarelli et al., 2007; Fiorini et al., 2007).
In *Article 3* we investigated the paternal genetic composition of two Arbereshe ethno-linguistic groups, of Sicily and Calabria (Southern Italy) respectively, with a dual-level purpose: i) to obtain a more fine-grained description of Y-chromosomal diversity within the Arbereshe ethno-linguistic minority, and ii) to distinguish between homogeneous or differential patterns of genetic variation among the different Arbereshe groups. Despite the common origin and the unifying cultural and linguistic traits, different founder events and bio-demographic histories, as well as different degrees of isolation or admixture with the local populations once established in the Italian Peninsula, could actually have addressed the evolutionary history of each Arbereshe community differently. Moreover, by directly comparing linguistic isolates with both their geographic neighbours and source populations, we intended to explore the effects of cultural isolation on population genetic variability and to examine the degrees of genetic continuity or local admixture with the Balkan source and Italian recipient populations respectively.

Besides geographic constraints and cultural traits (e.g. language as explored in Paper 3), also socio-economic factors may strongly influence the way in which unions between individuals (within or among population groups) took place over time (Destro-Bisol et al., 2004), especially as far as the more recent periods of human history are concerned. However, recent population events are often rarely detectable in population genetic studies, because they mainly resulted in only limited modifications of the pre-existing genetic background. Moreover, if we consider that the history of Italy – as indeed of the whole Europe – is composed by multiple layers through which different genetic ancestries deeply mixed over time (Ralph and Coop, 2013), it becomes even more evident that trying to connect genetic signals with specific (relatively recent) historical events necessarily requires the choice of peculiar case studies.

In *Article 4*, we propose a case study in which a well-conceived sampling strategy (by carefully selecting the investigated population and the individuals to be sampled therein), coupled with i) the in-depth high-resolution analysis of Y-chromosome markers, ii) the reconstruction of deep-rooted paternal pedigrees; and iii) the occurrence of elite socio-economic conditions, have constituted an unique experimental field to shed new lights on some aspects of the recent genetic history of the Italian Peninsula. Our target population is particularly represented by the *Partecipanza* of San Giovanni in Persiceto, an idiosyncratic socio-economic institution of Northern Italy (Padana Plain) whose origin traced back to the Middle Ages and are related to a collective ownership of the land. The privilege to *participate* (i.e. to share the leased assets) is inherited following a gene-like pattern along the paternal line (male descents of a group of founder families) and moreover requires the maintenance of the place of residence in the municipality of S. Giovanni in Persiceto (highly specific local ancestry). In this study we explored the paternal (Y-chromosome) and maternal (mtDNA) genetic variability of the *Partecipanza* of S. Giovanni in Persiceto by sampling at least one individual for each surname of all the still living founding families. Results were compared to those of a set of 'control' individuals, sharing the same geographic location and cultural environment (San Giovanni in Persiceto) but not the affiliation to the *Partecipanza*. Subsequently we interpreted the genetic patterns obtained in a wider perspective by including other 'control' populations from Italy and Central Europe.

In this way we particularly aim i) to check whether socio-economic stratification within the same population (San Giovanni in Persiceto) could have induced internal genetic structures (*Partecipanza* vs. *Non-Partecipanza*), and ii) to reconstruct the genetic history of the *Partecipanza,* verifying if its idiosyncratic socio-economic system may have helped preserving specific traces of the more recent genetic history of Italy. In addition the association of deep-rooted paternal pedigrees to Y-chromosome genetic profiles has also allowed us to address some still highly-debated issues, that transcend the regional setting of this case-study, such as the estimation of evolutionary parameters like Y-STRs mutation rate and population average generation time.

PART 1

The genetic history and population structure of Italy: a macro-geographic overview from the uniparental markers

1.1. Introduction

The need to deepen our knowledge on the Italian genetic landscape resides in the complex mosaic of people incomings and cultural exchanges which have characterised the history of the area since Palaeolithic times. Since the first expansion of Anatomically Modern Humans (AMH) out of Africa to the more recent historical migrations, the Italian Peninsula has indeed represented a strategic landing and starting point for several human migration processes, thus acting as a major melting pot for populations of different geographic origins and cultural matrices. Both paleo-anthropological studies and archaeological records agree in highlighting the key role of Italy in the whole range of pre-historical and historical population dynamics occurred in Europe and the Mediterranean. Each of these events may potentially have left footprints in the gene pool of modern Italian populations. Integrating archaeological records with the perspective achievable through the study of genetic diversity patterns (each of them compensating and overcoming the relative limitations of the other) will represent a fruitful approach to investigate and reconstruct the human population past history.

In the next paragraphs we briefly resume the main historical and archaeological evidences on the Italian (pre-)history within the context of the European and Mediterranean peopling dynamics, and we then review the present-state of knowledge on the Italian genetic variability and population history, which obviously represents the necessary background of our investigation.

1.1.1 Pre-historical and historical population dynamics

Some of the information included in this section were extracted from the book chapter *"The Mediterranean human population: an Anthropological Genetics perspective"* by Sazzini M, Sarno S, Luiselli D (2014), as it appears in Goffredo S, Baader H, Dubinsky Z (Editors). The Mediterranean Sea: Its History and Present Challenges. Berlin: Springer, pp. 529-551.

1.1.1.1 Palaeolithic occupation and post-glacial re-expansion

Both archaeological and genetic evidences indicate that Anatomically Modern Humans (AMH) evolved in Africa and then expanded out of Africa, gradually colonising the other continents by means of subsequent founder events ("*serial founder model*"; Ramachandran et al., 2010).

Concerning the European continent, first groups of modern humans are thought to have entered Europe through the Levantine corridor (as suggested by all the current genetic evidences, Mellars, 2011), during the period coinciding with the Middle to Upper Palaeolithic transition. Despite the still-open questions about biological and cultural interactions - with replacement or admixture - of modern humans with local populations of Neanderthals (Krause et al., 2010; Prat et al., 2011; Currat and Excoffier, 2011), it is now largely accepted that the appearance of AMHs in Europe has coincided with the beginning of Neanderthal decline (String and Davies 2001). Recent re-examinations of findings from two site of North-Western and South-Eastern Europe have contributed to cast new lights on the long-date debate whether the arrival of AMHs in Europe could be associated with the widespread Aurignacian culture, or may additionally be linked to other Early Upper Palaeolithic industries (Mellars, 2011). Recent reassessments of human fossils from the Kent's Cavern (UK) have attested the presence of modern humans at the extreme north-western tip of Europe by at least ~42,000-43,000 years before present (YBP) (Hinghman et al. 2011). These findings corroborate previous radiocarbon datings of several archaeological sites of western and central Europe belonging to modern humans and associated to the Aurignacian culture (Mellars, 2011). If the dispersal of modern human groups in Western Europe by at least 42,000 YBP is relatively well-established, on the other hand evidences of early settlements in Southern Europe have long been comparatively more scarce (Mellars, 2011). By contrast, recent re-examination of human teeth from the cave site of Grotta del Cavallo, in the extreme southern heel of Italy (Apulia), suggested a previously-unknown dispersal route of modern humans into the southern-fringe of Europe by around 43,000-45,000 YBP **(**Figure 1.1.1.1.1), thus making the human remains of Italy the oldest known European AMHs (Benazzi et al, 2011). Interestingly, similar dating results of human remains from South-Eastern European sites (Bulgaria and Romania) moreover suggested a possible pre-Aurignacian population dispersion along the Mediterranean coast into some parts of Southern and Central Europe (Mellars, 2011 and references therein), thus confirming a rapid diffusion of AMHs across the continent before the Aurignacian culture and the disappearance of Neanderthals (Figure 1.1.1.1.1). The Cavallo human remains were associated to a highly distinctive early human culture, known as Uluzzian industry, which is currently confined only to sites in Italy and Greece (Kozlowski, 2007). The attribution of these remains to modern humans rather than Neanderthals has casted new lights also on the still-debated belief according to whom the Uluzzian technology was the result of acculturation between indigenous Neanderthal populations in Italy and intrusive early Proto-Aurignacian populations of modern humans from the adjacent Adriatic coast (Mellars, 2011 and references therein).

Approximately 25,000 YBP another major event involved the Italian Peninsula, in response to dramatic worldwide climatic changes. During the period known as the Last Glacial Maximum (LGM), the expansion of large ice sheets on most of the north-central hemisphere (Mithen, 2006) forced the abandonment of northern Europe and the contraction of human range in the Southern Peninsulas located along the Mediterranean coastlines (Figure 1.1.1.1.2), that actually acted as refugee areas for human populations (Banks et al. 2008).

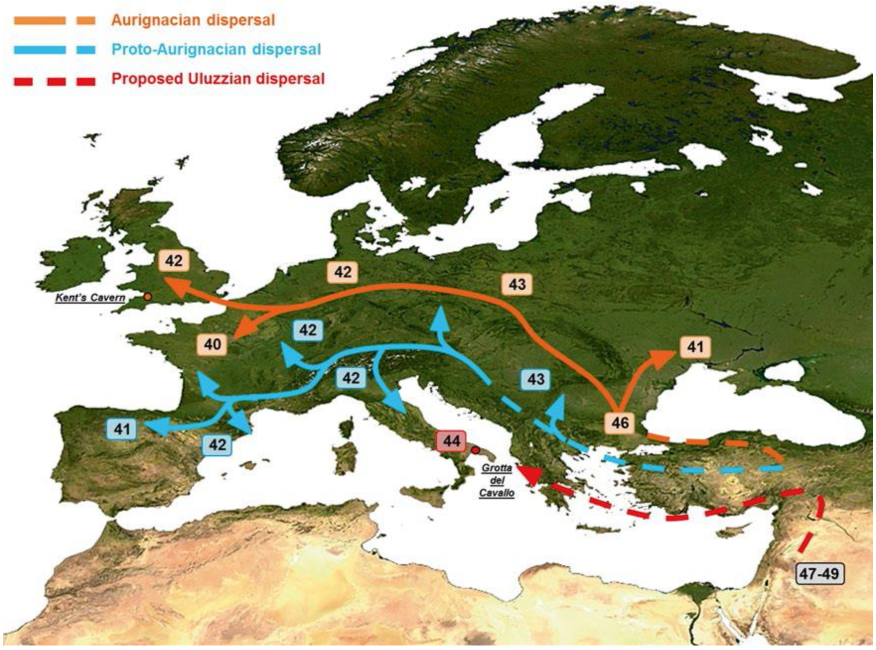


Figure 1.1.1.1.1. Palaeolithic main dispersal routes of anatomically modern humans (AMH) across the Mediterranean area. *Numbers* indicate kilo years before the present (KYBP). From Sazzini et al., 2014.

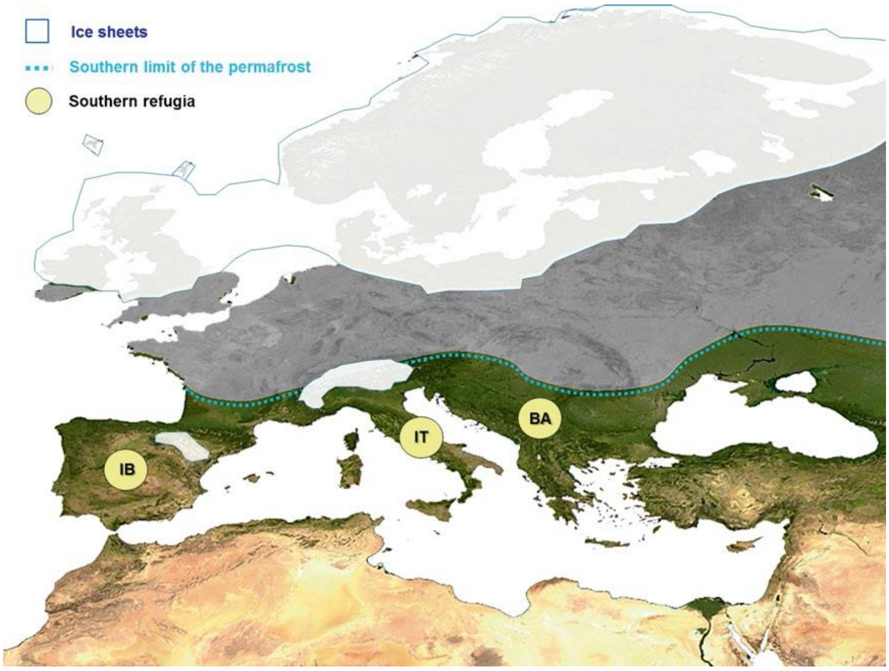


Figure 1.1.1.1.2. Locations of the three main refugia during the last glacial maximum (LGM). IB stands for Iberian Peninsula, IT stands for Italian Peninsula, BA stands for Balkan Peninsula. From Sazzini et al., 2014.

Accordingly with this view, gaps in archaeological records were observed in Northern and Central Europe. On the contrary, the distribution of both Solutrean and Epigravettian sites in Southern Europe testified the presence of a clear boundary for the potential human range during the LGM, moreover suggesting different geographic distributions of human ecological niches between the Iberian Peninsula on one side, and Italy and the Balkans on the other side (Banks et al. 2008). The Solutrean culture appeared mainly present in Southern-Western France, Northern Iberia, North-Western Italy, and sporadically in the Balkans (Figure 1.1.1.1.3). On the other hand, the Epigravettian culture revealed its presence mainly in the Balkan and Italian Peninsulas (Figure 1.1.1.1.3). These differences probably resulted in more effective restrictions between the Italian and Iberian Peninsulas, while facilitating exchanges between Italy and the Balkans (as also attested by the broad Adriatic plain created during the LGM by the lowering of Mediterranean sea level; Antonioli et al., 2004; Lambeck et al., 2004).

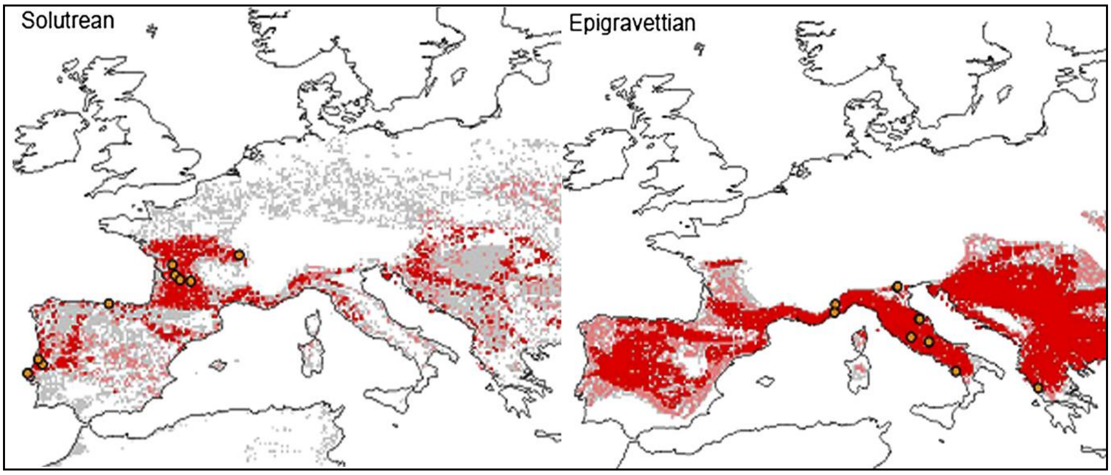


Figure 1.1.1.1.3. Distribution of human eco-cultural niches for the Solutrean and Epigravettian technologies during the Last Glacial Maximum (LGM). Archaeological site locations are indicated by yellow circles. From Bank et al., 2008

The contraction of human range and the resulting demographic reduction are known to have caused a bottleneck in human genetic diversity, with obvious consequences on the gene pool composition of modern Mediterranean and European populations. Early human populations escaping from the northern Europe came into contact with the pre-existing Southern human groups at least until the subsequent population re-expansion, which started approximately 16,000-13,000 YBP with the improvement of weather conditions. The re-peopling process from these refugee areas and the consequent geographic re-distribution of admixed and reduced populations has contributed deeply to reshape and relocate the early Palaeolithic pattern of genetic variation (Sazzini et al., 2014).

After the end of the Ice Age, the Mesolithic in Europe was marked by much warmer climatic conditions and by a new way of life in which hunting and gathering (and possibly fishing) became the main modes of subsistence, at least until the subsequent massive spread of agriculture technology all over the Europe during the Neolithic revolution (Soares et al., 2010).

Although very little is known about the Upper-Palaeolithic and Mesolithic hunter-gathers human groups of southern Europe and the Mediterranean, two notable exceptions come once again from the Italian Peninsula. The recent discoveries of Upper Palaeolithic human remains from Grotta del Paglicci (in the South-Eastern heel of Italy, Apulia; Caramelli et al., 2003) and Grotta d'Oriente (in the island of Favignana, Sicily; Mannino et al., 2012) - both assigned to mtDNA haplogroup HV - were interpreted in favour of the descendance of the early-Holocene hunter-gatherers of Sicily from the human groups occupying Southern Italy before (Gravettian) and after (Epigravettian) the LGM, thus casting new lights on the first peopling of Sicily by AMHs (Mannino et al., 2012).

Pre-Neolithic genetic inheritance of European populations

Many genetic evidences have long suggested that a large amount of modern maternal (mtDNA) and paternal (Y-chromsome) European genetic pools, can be traced back to the re-expansion and re-location of human Palaeolithic inheritance during the warming phases that followed the LGM, with only marginal contributions (quantified at ~20-22%) from the subsequent Neolithic incoming lineages from the Near East (Soares et al. 2010).

The most ancient maternal lineage in Europe belonged to haplogroup U5, which is thought to have locally-originated approximately 30,000-37,000 YBP (Soares et al., 2010; Malyarchuck et al., 2010). Analogously, Y-chromosome haplogroup I, which is mainly, if not exclusively, present in Europe (being almost absent elsewhere; Rootsi et al., 2004) is supposed to represent the paternal counterpart of mtDNA haplogroup U5 in marking the early Upper Palaeolithic inheritance of modern Europeans (Soares et al., 2010). The age estimates and the geographic distributions of their major sub-clades, have however suggested that both mtDNA U5 and Y-chromosome I actually experienced deep post-glacial re-expansions and relocations after the LGM (Rootsi et al., 2004; Pereira et al., 2010; Malyarchuk et al., 2010; Soareas et al., 2010). Recent phylogeographic analyses revealed that expansions of the U5 sub-clusters (namely U5a and U5b, respectively) started earlier in southern Europe. In addition, after the LGM, Central Europe apparently represented an area of intermingling between human flows coming from the refugee Peninsulas located in the Balkans, the Mediterranean coastline and the Pyrenees (Malyarchuk et al., 2010). Similarly, several sub-clades of haplogroup I were proposed to have expanded from the Franco-Iberian (I1-M253) and Balkan (I2a-P37.2) refugia (Rootsi et al., 2004; Marjanovic et al., 2005). With respect to haplogroup I2a-

P37.2, which is highly predominant in Eastern Europe and particularly in the Balkans, its sub-clade I2a1-M26 instead represents the dominant paternal lineage of Sardinia. Outside of Sardinia, this haplogroup has been found at moderate frequencies only in Western Europe, and particularly in Iberia and Southern France, where is thought to have originated during the LGM (Lopez-Parra et al., 2009). The age estimate and the distribution pattern of Y-chromosome I2a1-M26, exactly matches the one of the Sardinian-specific female counterpart, the mtDNA-haplogroup U5b3. Except for this rare U5b3 lineage, which presumptively expanded from the Italian Peninsula after the LGM (Pala et al., 2009), most of the other maternal lineages commonly found in Western and Central Europe are instead postulated to have expanded from the Franco-Cantabrian area (Soares et al., 2010). On the other hand, the Balkans or eastern Europe (Ukraine) were invoked as the source expansion areas for mtDNA lineages today prominent in the eastern part of Europe (e.g. mtDNA haplogroup U4; Malyarchuk et al., 2008; Soares et al., 2010). The most frequent European maternal lineage, the so-called haplogroup H, is supposed to have expanded in Europe from the Near East during the diffusion of the Gravettian culture (~25,000-20,000 YBP; Richards et al., 2000) and then to have been deeply involved in post-glacial re-expansion processes. Accordingly, it's major subclades, namely H1, H3 (11,000-15,000 YBP) and H5 (~13,900 YBP), together with haplogroups V and U5b1, are suggested to have expanded from the Franco-Cantabria area to the Western, Northern and Central Europe through different dispersal routes (Torroni et al., 2001; Achilli et al., 2004; Pereira et al., 2005). Y-chromosome haplogroup R1b, paralleling mtDNA lineages H and V, is thought to have arrived in the European continent from the East just after the LGM (Rosser et al., 2000; Semino et al., 2000) and then to have locally originated the R1b1b2-M269, which currently represents the main paternal lineage throughout Western Europe (Myres et al. 2011 and references therein). The North-West to South-East decreasing frequency pattern of haplogoup R1b-M269 in Europe has long been ascribed to the survival of the Upper-Palaeolithic substrate to the subsequent Neolithic demic diffusion from Near-East (Rosser et al., 2000; Semino et al., 2000). However, more recent results have postulated the haplogroup R1b-M269 to be younger and likely associated with Neolithic and Post-Neolithic invasions of the European continent (Balaresque et al., 2010). The distribution of its microsatellite (Y-STRs) variance in a large sample of R1b-M269 European Y-Chromosomes (Figure 1.1.1.1.4) and its age estimate at around ~6.500 YBP (95% confidence intervals 4,600-9,100 YBP) has indeed suggested a rapid expansion of the haplogroup during the Neolithic period (Balaresque et al., 2010). These results stress the complexity of interpreting processes of formation and expansion of human populations in Europe during the last 10,000 years. Most of them have indeed occurred in more than one migration events, being moreover separated by narrow temporal windows (Myres et al., 2011; Busby et al., 2011).

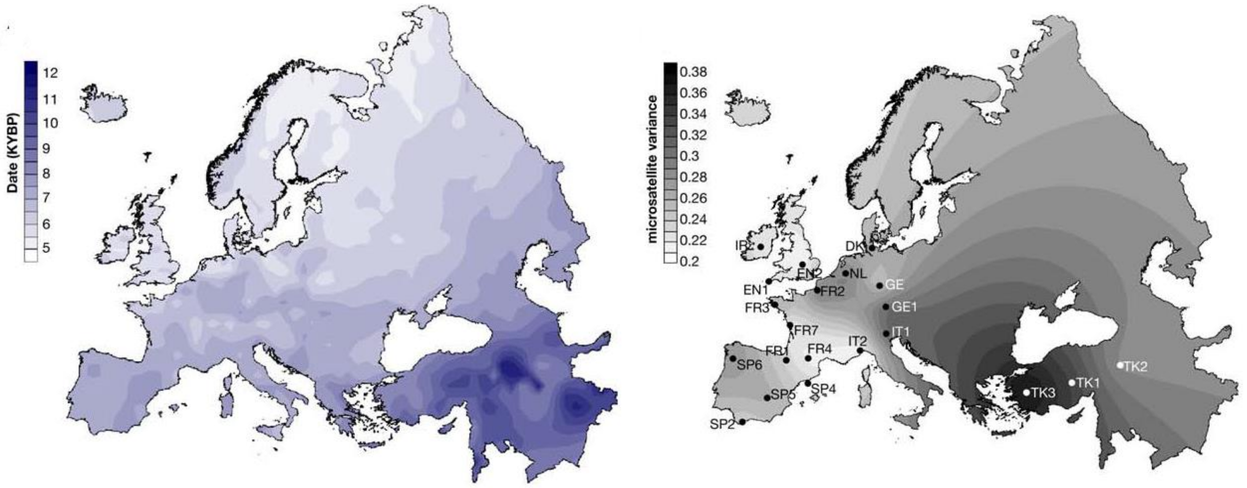


Figure 1.1.1.1.4. Isochronal maps showing dates of Neolithic sites in Europe (left) - based on data of Pinhasi et al. (2005) – and distribution map in Europe of the microsatellite variance of haplogroup R-M269 (right). From Balaresque et al., 2010.

1.1.1.2 Main migration patterns associated to the Neolithic transition

The Neolithic transition from hunter-gathers to farming societies has surely marked one of the most drastic cultural changes in European pre-history. However, the impact of Neolithic revolution on the European and Mediterranean populations is still highly debated among archaeologists, linguists, anthropologists and geneticists, in terms of both dispersal routes and extents of *cultural* or *demic* diffusions (Sazzini et al., 2014 and reference therein). Two main hypotheses were historically formulated to account for the spread of Neolithic culture across Europe (Figure 1.1.1.2.1). The first hypothesis (referred to as *demic diffusion model*) suggests population movements of farmer human groups from the Near East into Europe, with the replacement of most of the pre-Neolithic hunter-gatherer populations (Ammerman and Cavalli-Sforza, 1984). According to this model, an appreciable Near Eastern genetic signature is therefore expected to be found in the gene pool of current European and Mediterranean populations. As a consequence we would expect strong affinities between gene pools of modern European and Near Eastern populations, with little genetic differentiations within the European population except for those arose because of drift (Jobling et al., 2013). On the other hand, the second hypothesis (referred to as *cultural transition model*) suggests that cultural transmission was the main factor of the Neolithic transition, thus assuming a selective adoption of specific aspects of the Neolithic technology by the local pre-existing populations, with no or little genetic input by the new incoming farming groups (Whittle, 1996). This obviously would prevent us to found any appreciable influence of Neolithic revolution on the gene pool of modern Mediterranean and European populations, with a sharp distinction between European and Near Eastern genetic pools.

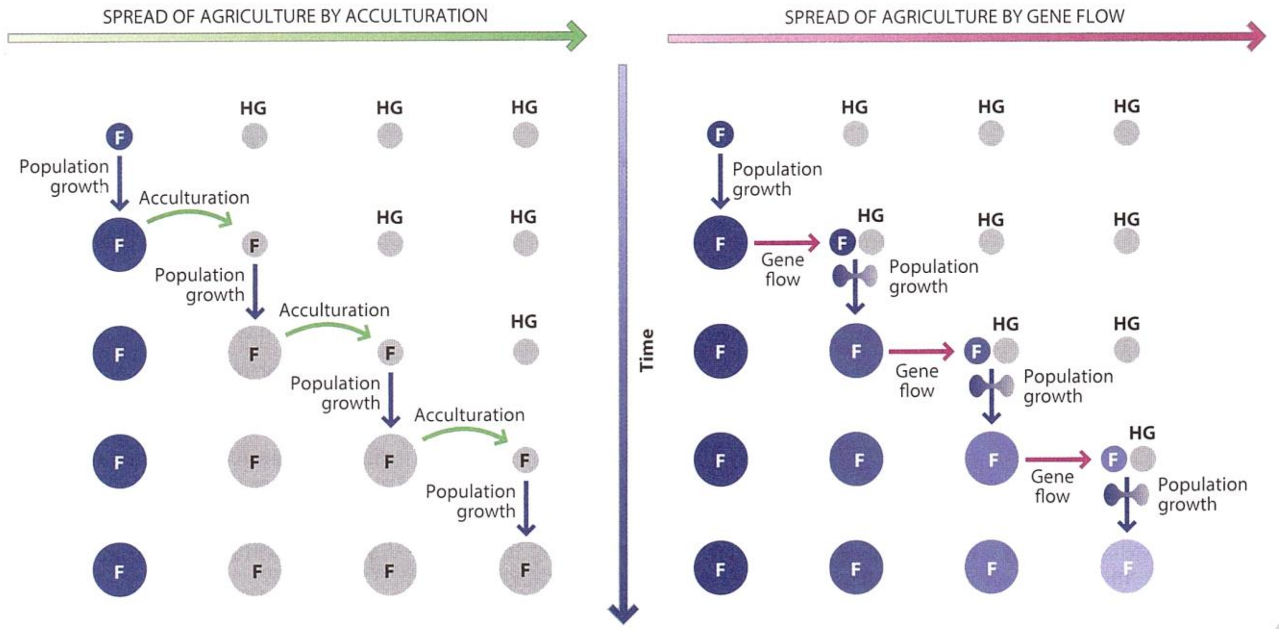

SPREAD OF AGRICULTURE BY ACCULTURATION
SPREAD OF AGRICULTURE BY GENE FLOW
HG
F
Population growth
Acculturation
Gene flow
Time

Figure 1.1.1.2.1. Acculturation and gene-flow models for the diffusion of agriculture during the Neolithic transition. From Jobling et al., 2013.

Despite the initial tendency to consider the Neolithic transition as the result of only one single homogeneous process, several sources of evidence have recently outlined a more complex picture, implying outstanding regional variations in the process of agricultural spread both in the times of diffusion – involving rapid expansions followed by period of stasis (Bocquet-Appela et al., 2009) - as well as in the extents of interactions between residents and newcomers - with differential and intersecting levels of demic and cultural processes (Pinhasi et al., 2005).

Archaeological evidences suggest that agriculture societies, from their origins in Western Asia by around 11,000 YBP, rapidly spread reaching South-Eastern Europe approximately 8,500-9,000 YBP. From South-Eastern Europe, the so-called Neolithic package was disseminated by following two main routes: 1) the Vardar-Danube-Rhine corridor, and 2) the Adriatic and Mediterranean Sea coasts (Tresset and Vigne, 2011). These routes are associated with the two main cultures that co-existed in Early-Neolithic: the Linear BandKeramik culture (LBK) and the Impressed Ware culture (IW) respectively. After its first appearance in Southern Europe, the diffusion of agriculture within the continent took place relatively quickly. If approximately 2,000 years is thought to be the time required to Neolithic farmers for moving from Cyprus to the Aegean Sea, ~500 years were then needed to reach the Italian Peninsula, whereas only another 600-800 years were required to arrive at the most western Atlantic fringe (Figure 1.1.1.2.2).

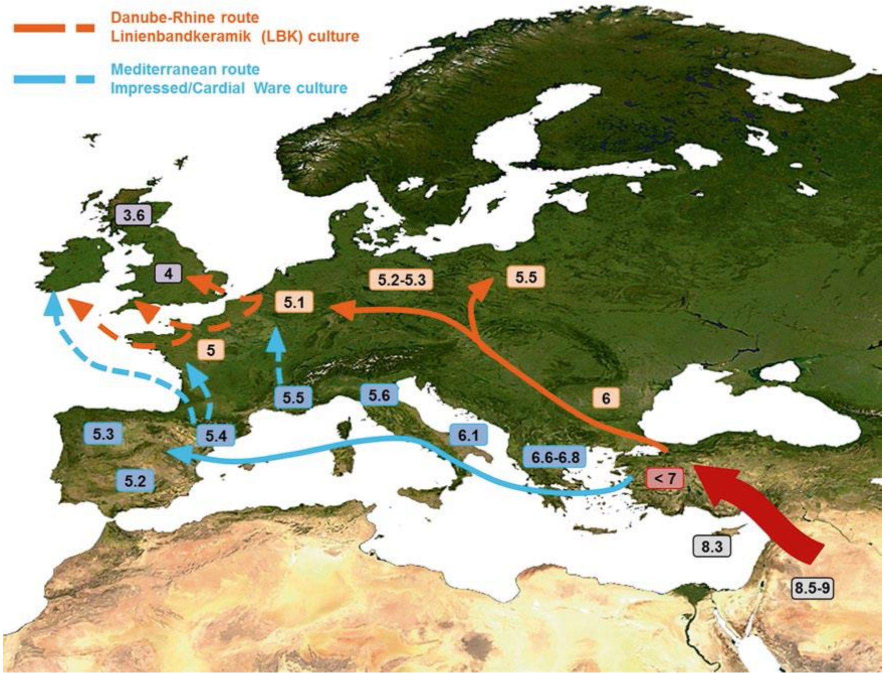


Figure 1.1.1.2.2. Chronology and main routes of Neolithic spread across the Mediterranean area. Number indicate years BC. From Sazzini et al., 2014

Neolithic in the Italian Peninsula

The Italian Peninsula was involved since the very beginning in the process of Neolithic expansion starting from the opposite Balkan coast. The presence of the first Neolithic groups (mainly expanded from the northern Greece and southern Albania) in the extreme south-eastern heel of Italy (Apulia) date back approximately at 6,100-5,800 BC, and is associated to the oldest phase of the Italian Neolithic Period, which is characterized by the *Archaic Impressed Ware* culture (Figure 1.1.1.2.3; Pessina and Tinѐ 2008). From this region, agriculture (and the connected Neolithic human groups) expanded westward to Southern Calabria and Eastern Sicily, where traces of material cultures belonging to the subsequent phase of the Upper Neolithic Period (*Stentellino Impressed Wares*) date roughly between 5,800 and 5,300 BC. However the Neolithic pottery (*pre-Stentinello Impressed Wares*) uncovered in western Sicily (Uzzo and Kronio caves) is suggested to be approximately coeval (6000-5750 BC) with the earliest occurrence of Neolithic materials in the more South-Eastern portion of the Italian Peninsula, thus potentially suggesting different processes of colonisation between western and eastern parts of Sicily (Figure 1.1.1.2.4; Pessina and Tinѐ 2008), albeit this hypothesis would need more data to be confirmed.

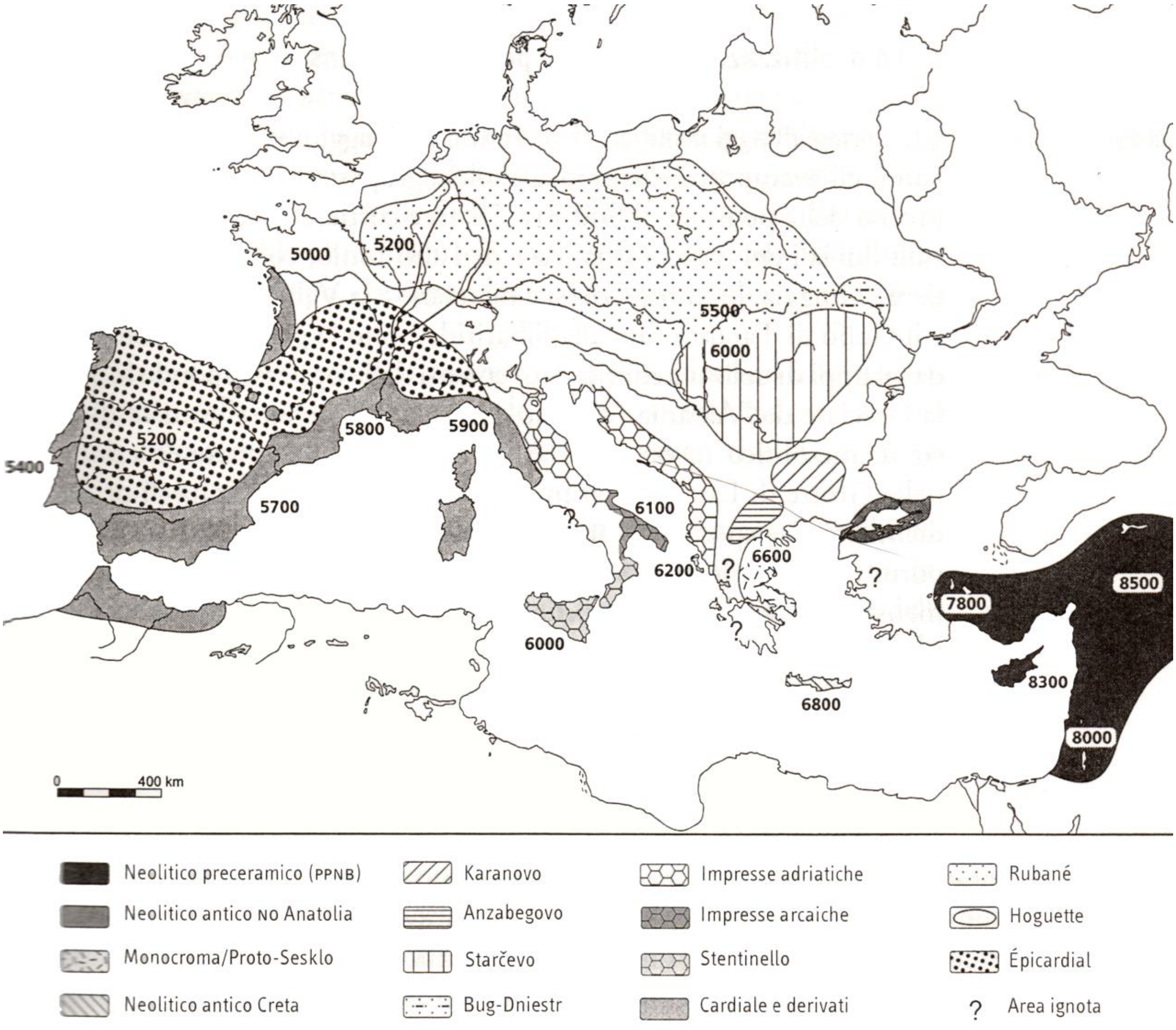


Figure 1.1.1.2.3. Patterns of diffusion of Neolithic cultures in Europe and the Mediterranean. Number indicate years BC. From Pessina and Tiné, 2008.

Farming technology appeared in Central Italy on both sides of the Apennines, but the diffusion of the Neolithic package seems to have proceeded differently between the east Adriatic and the west Tyrrhenian coasts. From Southern Italy, farming communities expanded steadily northwards along the Adriatic coast, as testified by the documented link between the Adriatic pottery of the Early Middle Neolithic (*Adriatic impressed wares*) and the Southern impressed pottery of the same period (*Guadone facies*) (Figure 1.1.1.2.4; Pessina and Tinè 2008). More complex appeared instead the settlement of the western Tyrrhenian coast. Recent re-examinations of Early-Neolithic ceramic materials particularly suggested two major episodes of diffusion: an older one (referable to 5,800-5,500 BC) associated with the first colonization of the Ligurian-Provencal area; and a more recent one (5,400-5,100 BC) linked with the appearance and diffusion of the Franco-Iberian Cardial culture. In North-Western Italy (as well as in South-Eastern France) the Impressed Ware tradition first appeared in Liguria by around 6,000-5,800 BC, and then spread quickly along the Tyrrhenian coast (*Tyrrhenian impressed wares*), thus suggesting a chronologically parallel but culturally independent process of Neolithization along the Tyrrhenian side relative to the Ionic-Adriatic one (Figure 1.1.1.2.3 and Figure 1.1.1.2.4; Pessina and Tinè 2008). The second phase of expansion, dating approximately to 5,400 BC (or perhaps slightly earlier), witnessed the appearance and the diffusion of the Cardial material, from its origins in the western Mediterranean regions of France and South-Eastern Iberia. The spread of Cardial Neolithic is thought to have included both colonisation of uninhabited regions and local adoption of farming – and other aspect of Neolithic package – by the pre-existing populations (Price, 2000).

The remaining areas of the Italian Peninsula (Po Valley and North-Eastern Italy) experienced a later arrival of the Neolithic culture, being also characterized by a more marked continuity with the earlier Mesolithic groups (Cunliffe, 2001). In fact, indigenous communities tended to select specific aspects of the new technology and to integrate them with their pre-existing ways of life (Cunliffe, 2001). Consequently, most of the cultural *facies* (*Fiorano, Vhò, Gruppi Friulani, Gaban* and *Isolino*) belonging to the Po-Alpine Early Neolithic Period (Figure 1.1.1.2.4; Pessina and Tinè 2008), have been long interpreted as the result of cultural transfer from the Neolithic incoming groups to the local pre-existing Mesolithic communities. However, due to the environmental and population heterogeneity of the area, the process of Neolithization in Northern Italy should not be interpreted exclusively as the one-way result of single processes of acculturation or demic diffusion, but rather as a more complex and geographically articulated mosaic of situations, with different regional variations.

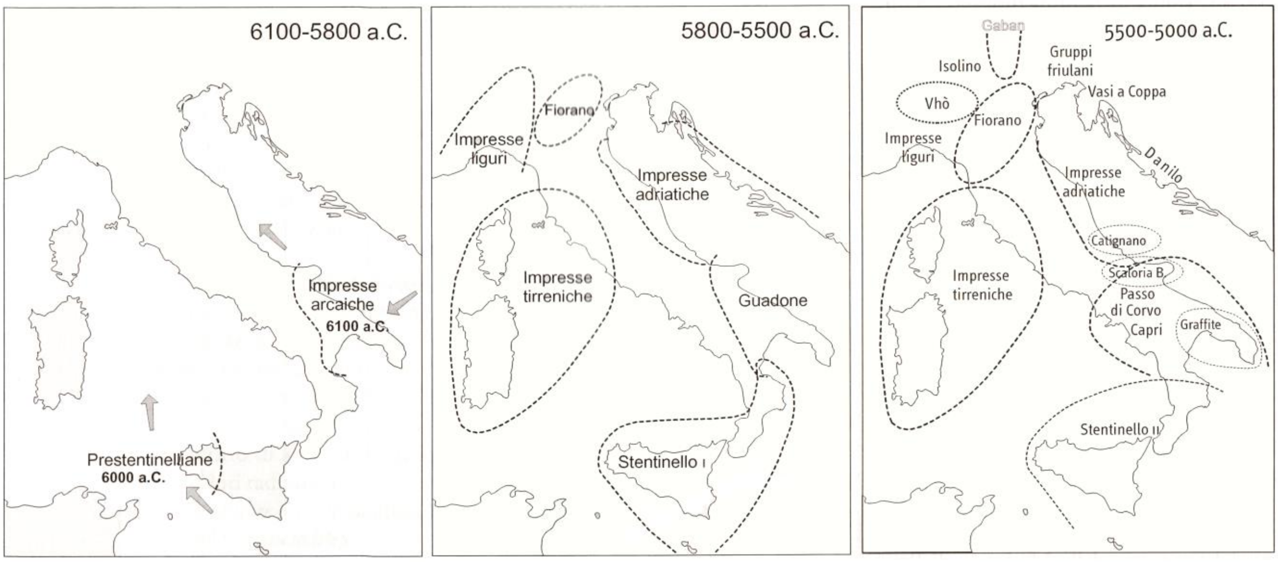


Figure 1.1.1.2.4. Main cultural facies during the subsequent phases of Neolithic diffusion within the Italian Peninsula. From Pessina and Tinè, 2008.

Despite these difference at local levels, particularly concerning intersecting patterns of demic and/or cultural diffusion processes, it is possible to summarize the Neolithization of Italy by hypothesizing the presence of two main and independent farming-diffusion directions: a North Italian-Tyrrhenian and a South Italian-Adriatic one respectively (Cunliff, 2001).

Neolithic genetic inheritance of European populations

Since the pioneering work by Cavalli-Sforza, Menozzi and Piazza (1994) to the more recent researches on uniparental markers, the study of the complex interplay between demic dispersal and cultural diffusion processes of Neolithization has been one of the central topic of population genetics studies (Sazzini et al., 2014). Investigations on diversity patterns of both classical nuclear and molecular uniparental (mainly Y-chromosome) markers often revealed the presence of allele frequency clines along the SE-NW axis of Europe (Cavalli-Sforza et al., 1994, Barbujani et al., 1994, Rosser et al. 2000, Semino et al, 2000 Chikhi et al. 2002; Balaresque et al., 2010). Although some authors argued that such clinal patterns may have also originated from other processes than the Neolithic transition – for instance by repeated founding events during the initial Palaeolithic colonisation of Europe (Bocquet-Appel and Demars, 2000; Barbujani and Bertorelle, 2001) – the presence of SE-NW major frequency clines within the European continent has been long interpreted in favour of the demic model of agriculture diffusion.

Accordingly, the first studies performed on Y-chromosome markers (Rosser et al., 2000; Semino et al., 2000) yielded gradients similar to the classical demic diffusion model, thus supporting the *Wave-of-advance* hypothesis. In particular, based on both clinal patterns of distribution and TMRCAs of specific lineages, these first studies identified Y-chromosome haplogroups J2-M172, E1b-M35 and G-M201 as the putative markers of Neolithic expansion into Europe (Rosser et al., 2000; Semino et al., 2000). More recently, the increased resolution in Y-chromosome phylogeny further suggested the possibility to distinguish the European Neolithic Y-chromosome lineages from the pre-existing Mesolithic ones (Battaglia et al., 2009). Several J2-subclades were confirmed to be linked with the demic expansion of farmers, relying upon the fact that their distribution patterns from Anatolia to the Mediterranean as well as their diffusion time estimates (~7,000-8,000 YBP) are both compatible with the archaeological evidences of the early appearance of Neolithic pottery in the Adriatic region. In a similar way, J2a-M410 and G2a-P15 lineages were hypothesized to have marked the Neolithic colonization of mainland Greece, Crete and Southern Italy (Di Giacomo et al., 2004; King et al., 2011). On the contrary the distribution pattern of the Balkan autochthonous I2a-M423 clade (derived from I2a-P215 excluding the typical Sardinian lineage I2a-

M26) appeared consistent with a late Mesolithic time frame, thus suggesting that also cultural adoption of agriculture techniques took place in the Balkan Peninsula (Battaglia et al., 2009).
A clinal distribution pattern of genetic variation has been recently suggested also for the Y-chromosome haplogroup R-M269 (Balaresque et al., 2010), previously thought to be the marker (together with the mtDNA haplogroup H) of the Palaeolithic inheritance in Europe. By showing that microsatellite variation associated to R-M269 followed an opposite direction to the NW-SE decreasing frequency cline of the same haplogroup (Figure 1.1.1.2.5), Balaresque and colleagues proposed the current distribution pattern of R-M269 in Europe to be associated with the Neolithic diffusion from the Near East. In particular, by considering that alleles arising on the edge of a wave of expansion could "surf" upon it by rapidly increasing their frequency (Jobling et al., 2013), the authors explained the higher frequency of R-M269 in North-Western Europe as the result of the Neolithic wave of advance (Balaresque et al., 2010).

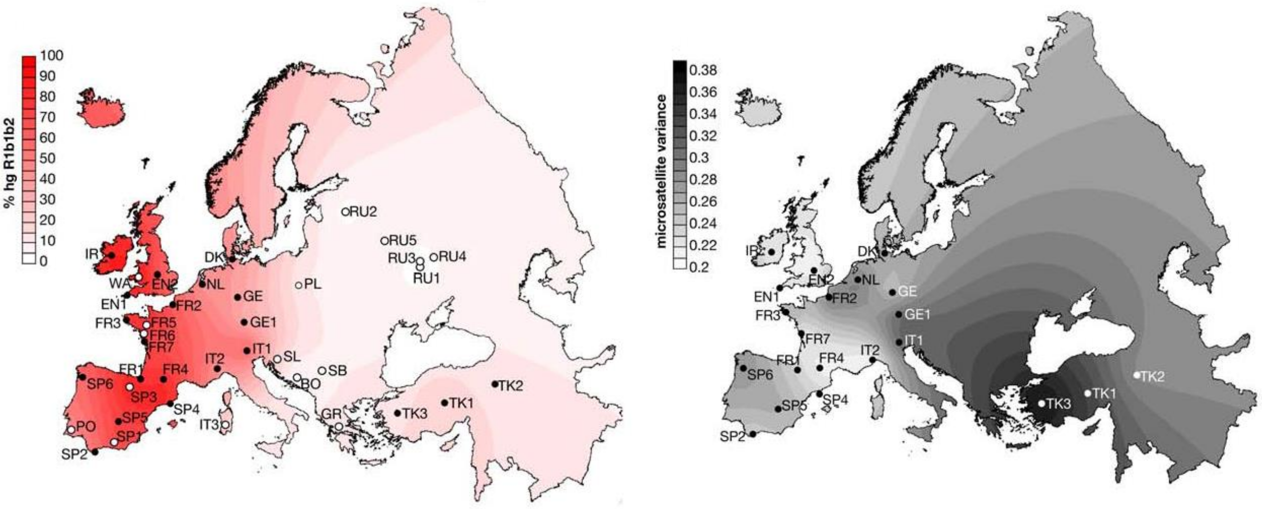


Figure 1.1.1.2.5. Geographical distribution of haplogroup frequency of R1b-M269 (in the left) and mean microsatellite variance (in the right). From Balaresque et al., 2010.

In the last years, new insights on the European genetic pre-history were provided by studies based on ancient DNA (aDNA). Although obtaining reliable aDNA sequences from Y-chromosome is generally more difficult than from mtDNA due to the worse preservation of nuclear DNA in ancient samples (Lacan et al., 2011b), different studies have recently offered the possibility to directly investigate Y-chromosome genetic patterns in Neolithic human remains from Central Europe (Germany, n=3; Haak et al., 2010), Western Europe (France, n=22; Lacan et al., 2011a and Iberia, n=6; Lacan et al., 2011b) and Southern Europe (Italian Alps, n=1 i.e. Otzi Tyrolean Iceman; Keller et al., 2012). Although more data are required to obtain a reliable picture of the paternal

Neolithic genetic pool, it is however noteworthy that all these studies highlighted the importance of haplogroup G2a in the genetic make-up of Neolithic specimens of both Central and Mediterranean Europe, while underlying the complete absence of the most common haplogroup R-M269 in all the Neolithic samples analysed. If confirmed, these results would at least in part suggest the discontinuity between Neolithic Y-lineages and those of modern Europeans, thus postulating an important role of men in the Neolithic dissemination that is no longer – or only partially – visible today (Lacan et al., 2011b; Jobling et al., 2013). This would also imply that intra-European migration processes occurred in Post-Neolithic times (and presumptively related to the Metal Ages) may have strongly affected the modern European gene pool and partially involved R-M269 sublineages, that could therefore have expanded throughout Europe more recently than the arrival of agriculture (Lacan et al., 2011b; Jobling et al., 2013).

Contrarily to Y-chromosome results, first investigations on the mtDNA diversity pattern in Europe and the finding that most of the mtDNA haplogroups coalesce in pre-Neolithic times, have long suggested the association of current European and Mediterranean maternal genetic pool to the Upper Palaeolithic and Mesolithic post-glacial re-peopling events, while indicating the Neolithic contribution at approximately 15-20% (Richards et al., 2000; Soares et al, 2010). However, more recent mtDNA analysis of ancient skeleton from both pre-Neolithic hunter-gatherers and Early-Neolithic farmers suggested the mtDNA distribution pattern in modern Europeans to be more similar to that of early farmers, thus pointing to the replacement of hunter-gatherers by the new incoming Neolithic groups (Figure 1.1.1.2.6). In particular two main mtDNA lineages - namely H and U - have been proposed as markers of the early-Neolithic farmers and pre-Neolithic hunter-gatherers respectively (Fu et al., 2012). Estimation of past population sizes based on a Bayesian Coalescent approach indicated hunter-gatherers U-types to have expanded between 15,000 and 10,000 YBP, in accordance with the end of LGM in Europe and the subsequent northwards post-glacial re-expansion; on the contrary population increase for early farmers carrying the H-types dated approximately to 9,000 YBP, consistently with the beginning of Neolithic transition. After 4,000 YBP, both lineages showed similar patterns of population growth, thus probably suggesting the merging between the two populations after their primarily different histories (Fu et al. 2012). This picture contrasts with the earlier views on the European maternal inheritance (Jobling et al., 2013), especially for what concerns the position of mtDNA haplogroups H which has been initially proposed as the marker of the Palaeolithic/Mesolithic substrate and subsequently associated to the Neolithic expansion.

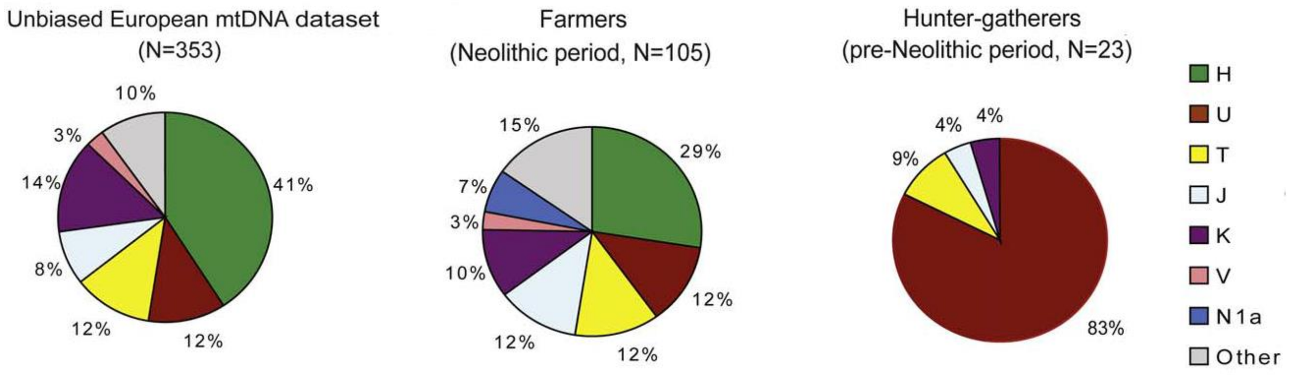


Figure 1.1.1.2.6. Haplogroup frequencies of contemporary European (based on complete mtDNA dataset) as well as of early farmer and hunter-gatherer populations (based on short segments of the mtDNA). From Fu et al., 2012.

Within the traditional dichotomy between demic diffusion and acculturation process, some recent aDNA-based studies on European mtDNA diversity have suggestd a more complex scenario, implying differential impacts of Neolithic process through the different parts of Europe. A first genetic study, based on individuals (n=24) belonging to the early Neolithic Linear Band Keramik (LBK; 5,500-4,900 calibrated BC) of Central Europe, observed a relatively high frequency (25%) of the currently rare (0.2%) mtDNA lineage N1a (Haak et al, 2005). This lineage appeared completely absent in the neighbouring Mesolithic groups (Bramanti et al., 2009; Malmstrom et al., 2009), that instead were almost exclusively (80%) composed of U-derived lineages (particularly U4 and U5; Haak et al., 2010). This suggested N1a to be a candidate signature for the early Neolithic LBK populations (Haak et al. 2005), although with a relatively small impact on the genetic background of present-day Central Europeans (0.2%). A subsequent study, that increased the LBK dataset (n=42; Haak et al., 2010), highlighted the general affinity of LBK Neolithic groups with the modern-day Near-Eastern and Anatolian populations. These results led to the hypothesis of a "*leap-frog*" or "*individual-pioneer*" colonization model, with an initial input of Neolithic farmers from the Near East to the LBK populations. However further events after this early Neolithic process, for instance the cultural adoption by surrounding hunter-gatherers of Neolithic technologies (Sampietro et al. 2007) have been also invoked to account for the differences observed between LBK and the modern Central European genetic pools (e.g. N1a frequency; Haak et al, 2005; Haak et al, 2010).

On the contrary, a parallel study based on Middle/Late Neolithic remains from the Iberian Peninsula belonging to the Impressed Ware Culture, failed to reveal any mtDNA differences in the genetic make-up of Neolithic groups relative to present-day Iberian populations. The absence of N1a lineages and the haplogroup composition of Neolithic groups which is extremely similar to that of modern Iberians, suggested a long-date genetic continuity compatible with the demic diffusion model of Neolithic culture in the Mediterranean part of Europe (Sampietro et al. 2007). Similarly, the more recent mtDNA analysis of Upper Paleolithic and Early Neolithic individuals from Northern Spain (Hervella et al. 2012), reaffirmed the above mentioned differential patterns between Central-Europe and the Western-Mediterranean, providing further support to a *random dispersion model* of diffusion, according to which Neolithic transition affected differently the various regions of Europe (Hervella et al. 2012). All these results therefore highlight that Neolithic revolution was neither genetically nor geographically a perfectly homogeneous process across the Mediterranean and European areas; on the contrary it involved differential extents of demographic and/or cultural diffusions, depending on the European region considered (Sampietro et al. 2007).

1.1.1.3 Population dynamics in historical times and their genetic impacts on the Italian genetic pool

In addition to the complex interplay between Palaeolithic and Neolithic inheritances, a large number of more recent events contributed to complicate further the patterns of human genetic variation in the European continent – and hence within the Italian Peninsula – leaving more or less evident genetic signatures in the gene pool of current human populations. Nevertheless, the attempt to disentangle the different contributions of each of these historical events is often complicated by the fact that most of these migrations actually resulted in overlapping and short-time separated processes of diffusion, being moreover responsible for only limited modifications in the pre-existing genetic backgrounds.

The period included between the Neolithic Age and the "dawn of history", is commonly referred to as Metal Ages and is traditionally divided into different intervals, whose names substantially reflect the most advanced metallurgical knowledge of that period: Copper Age (c. 3200–2300 BCE), Bronze Age (2300–700 BCE), and Iron Age (700–1 BCE). Simultaneously with technological innovations, during these phases the European societies witnessed drastic changes in settlement organization, ritual life, and interactions with other peoples, with obvious consequences on the material cultures referable to that periods (Forsythe, 2005). Accordingly, the Bronze Age of Italy (roughly coinciding with the second millennium BC - 1800-1100 BC) exhibits not only a more advanced metallurgical technology, but also continuous contacts with Bronze-Age peoples and cultures of both Central Europe and the South-Eastern part of Mediterranean Basin. The Bronze-Age sites of the Po-Valley in Northern Italy (*Terramare Culture*), mirroring the patterns of the previous Copper Age, appeared particularly influenced by the cultures of Central Europe. This is testified for instance by the replacement of the standard inhumation with the cremation funerary custom typical of the *Urnfield Culture* of Danube area, or by the new practice of erecting human habitations upon wooden platforms as the lakeside settlements of the northern latitudes (Forsythe, 2005).

While the material culture of the Po Valley developed mainly in response to influences from Central Europe, the culture of central and southern Italy (*Apennine Culture*) was relatively uniform at that time and characterised by standard inhumation funerary practices and large agricultural and pastoral settlements (Forsythe, 2005).

At the beginning of the 12th century BC, the collapse of Late-Bronze-Age societies of the Eastern Mediterranean marked the appearance and the spread of ironworking. In Italy, Iron metallurgy has not arrived before the 9th century BC. The approximately 200-years-time interval needed for iron to replace bronze as the most commonly used metal, is generally referred to as the Final Bronze Age (1100-900 BC). During this period, the funerary practice of cremation, previously confined exclusively in Northern Italy, spread southward of the Po Valley, characterising what is generally known as the *Proto-Villanovian Culture* (so labelled since this culture subsequently developed into the *Villanovian Culture*, that prevailed in Etruria and much of the Po Valley during the Iron Age, approximately around 900-700 BC). Whether the development of this culture was the result of a) cultural interaction between the *Terramare Culture* of Northern Italy and the *Appennine Culture* of Southern Italy, b) the simple extension of the *Terramare Culture* and relative peoples southward to the Po Valley, or c) the consequence of new incoming peoples from the Danubian *Urnfield Culture*, is not completely clarified yet (Forsythe, 2005). In any case, with the advent of the Early Iron Age, from the 9th century BC onwards, regional differences progressively appeared in the archaeological records of Italy, presumptively reflecting the linguistic and population diversity that later characterized the pre-Roman Italian peoples in historical times (Forsythe, 2005).

During the first millennium BC, Italy actually appeared as a patchwork of different peoples (Figure 1.1.1.3.1), either whose origin remain still largely unresolved (e.g. Etruscans, Ligures, Veneti), or that are known to have originated from specific migration processes (e.g. Celts in North-Western Italy; Greeks and Phoenicians in Southern Italy and Sicily). Unlikely to have completely deleted precedent genetic structures, such migration processes may have resulted in partially overlapping patterns of diffusion within the Italian Peninsula (Pesando, 2005).

By the collapse of Eastern Mediterranean Late-Bronze-Age societies (around 3,000 YBP), Southern Europe and the Mediterranean Basin underwent important demographic and socio-economic changes, peaking with the expansion of Phoenicians from the coastal Levant and Greeks from the Aegean Sea (Cunliffe 2001).

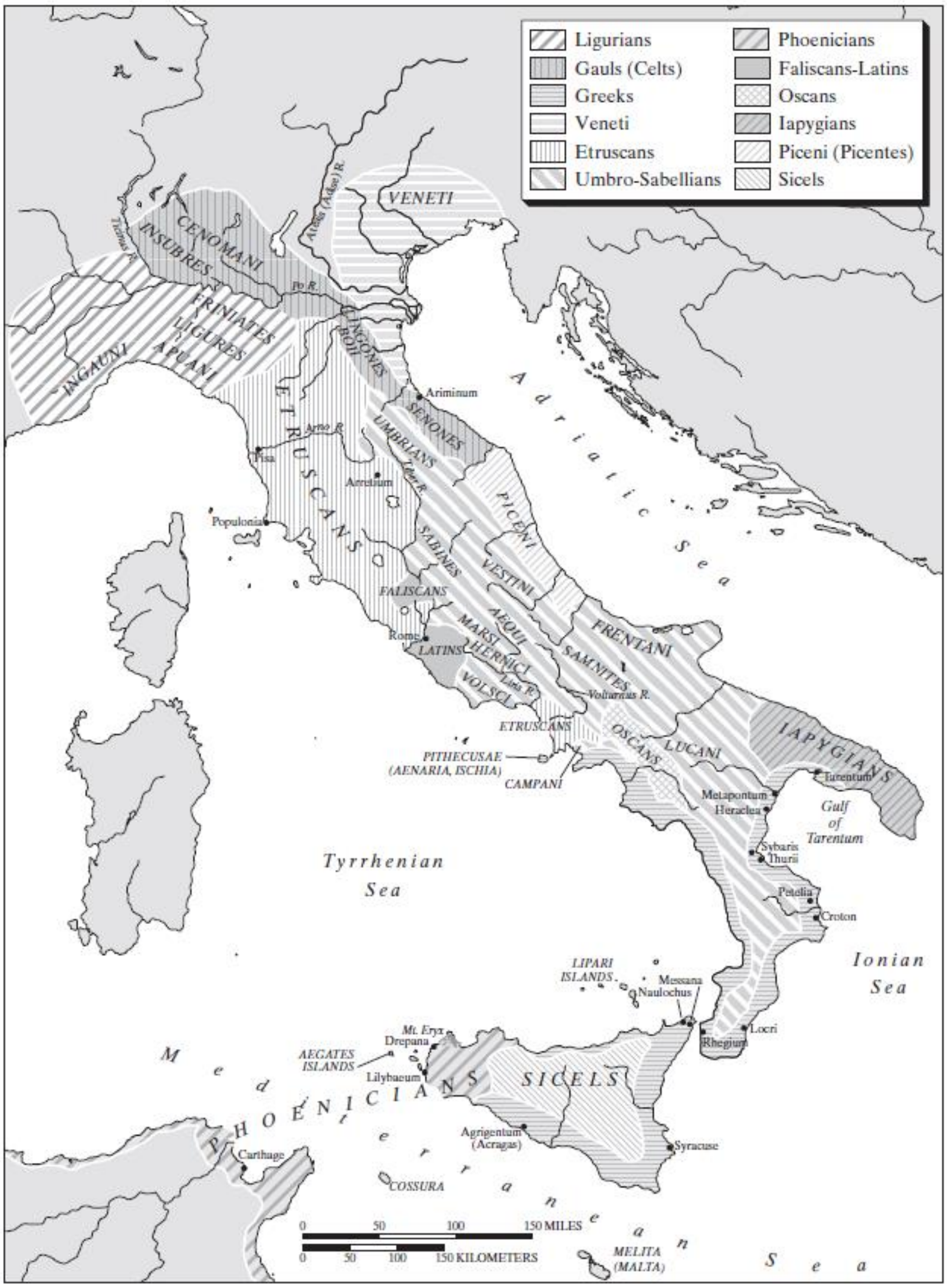


Figure 1.1.1.3.1. Map of the pre-Roman peoples of Early Italy. From Ward et al., 2009.

Ancient Phoenicia, located in the South-Eastern extreme of the Mediterranean Basin (between the present-day Syria and Palestine) appeared as a narrow strip of land enclosed between the Mediterranean and the mountains of Lebanon. From this area, the seafaring tradition of Phoenicians allowed them to established a trading empire throughout the whole Mediterranean area, as early as the Bronze Age (Zalloua et al., 2008 and references therein). The apogee of Phoenician colonization (referable approximately to 2,850 YBP) was characterized by the foundation of new cities (i.e. Carthage in 814 BC) along the North-African and Southern-Iberian coastlines, as well as in western Sicily (the latter particularly representing a strategic trading post for controlling the African coastal route to Iberia and Sardinia). Although diffusion patterns with origins in the Near East (or the Levant) and westwards decreasing gradients largely overlap those of different events in human pre-history (primarily the Palaeolithic and Neolithic colonisation processes), recent studies based on Y-chromosome markers managed to identify Phoenician genetic traces in modern Mediterranean populations by comparing paternal genetic profiles of historically documented Phoenician colonies with neighbouring non-Phoenician areas. These comparisons demonstrated that haplogroup J2 in general and six haplotypes in particular exhibited a Phoenician signature contributing to more than 6% to the current Y-chromosome variation in the Phoenician-influenced populations (Zalloua et al., 2008).

After the Phoenician colonization, the Greek expansion from the Aegean Sea surely played a major rule in the population dynamics of Southern Europe and the Mediterranean, with remarkable effects also on the Italian population variability. From the second half of the 8th century BC onwards, many Greek city-states established new communities along the coast of the Black Sea and northern Aegean, in Sicily and Southern Italy (giving rise to the so-called *Magna Graecia*), as well as in Southern France and North-Eastern Spain (Forsythe, 2005). While the Greek colonies of *Magna Graecia* in Sicily and Southern Italy are thought to be established mainly from a mixture of predominantly Dorian cities of the Aegean, Peloponnesus and central Greece, on the other hand the western Mediterranean coastal regions of Provence, Spain and Corsica are supposed to have received predominant influences from Ionian cities (e.g. Phocaea; King et al., 2011 and references therein). A recent study on Y-chromosome diversity in Sicily has estimated at ~37% the amount of current genetic variation of the island which can be traced back to the Greek immigration (using the E-V13 Balkan/Greek specific Y-chromosome genetic marker; Di Gaetano et al., 2009). In a similar way, a Y-chromosome genetic study on the Greek colonies of central and western Mediterranean, evaluate that ~17% of the current Y-chromosomes of Provence may be linked to the Greek colonization (King et al., 2011), therefore suggesting that subsequent and potentially confounding demographic events, such as the Roman dominion, actually did not completely erase the previous influences or however had a minor impact on the current genetic background of the area.

The period encompassing the Roman rule in Europe and the Mediterranean (spanning from 3rd century BC to 5th century AD) was characterized by even more complex and wide-ranging colonization processes and expansion dynamics. If during the first phase of colonization the new Roman colonies were mainly limited to the Italian Peninsula, in its apogee the Roman Empire actually acquired the complete control of the entire Mediterranean Sea and Southern Europe, assuming a pivotal role for the wide commercial routes and contacts among populations, at least until the Empire decline in 476 AD with the period of the Germanic invasions. In fact, between the

5th and 6th centuries AD the Italian Peninsula, as a reflection of the whole Europe, witnessed a series of invasions by Germanic peoples (the most notably for Italy being for example Ostrogoth and Lombard peoples) that possibly introduced a German/Eastern European component into the population of Italy. However, the exact genetic impact of such migration processes on the current Italian genetic pool remain still largely unknown (but see *Paper 4* in the second part of this thesis for further insights on this topic).

After the collapse of the Roman Empire, the Arab expansion across the Mediterranean Basin and Southern Europe represented one of the most important historical events. Starting from the Arabian Peninsula and then moving through the Near East, at the beginning of 7th century AD the Arab peoples succeeded in occupying North Africa, colonising the native African Berber groups. In less than 100 years, they then expanded in most of the Iberian Peninsula as well as in Sicily, ruling over these regions until the 15th and 13th centuries AD respectively (Capelli et al., 2009 and references therein). The North-African Y-chromosome haplogroup E-M81, as well as a subset of J1-M267 derived lineages, have been proposed as good candidates for the Arab genetic signature in Southern Europe. Relatively high frequencies of such lineages were indeed found in the Iberian Peninsula, as well as in the Southern part of Italy - mainly Sicily – in accordance with the long-term Arab rule in these areas of the Mediterranean Basin (Capelli et al., 2009).

Eventually, other more recent and localized migration processes, such as the massive movements of Albanians that between the 15th-16th centuries AD gave rise to the Arbereshe linguistic minority of Sicily and Southern Italy, contributed further to increase the ethno-linguistic heterogeneity of current Italian population (but a more complete description and discussion of this topic will be put off in the second part of this thesis, particularly referring to the *Paper 3*).

1.1.2 A genetic overview on the Italian population variability

As reviewed in the previous paragraph, the history of the Italian Peninsula is marked by multi-layered episodes of population migrations, differing from each other for both the time and place of their origin, as well as for the patterns of diffusion and the extent of their impact on current population variability. Besides the still-open debates about the exact dynamics of peopling, inferring the genetic contributions of each population event to the current Italian genetic pool is moreover complicated by the fact that - as clearly demonstrated in previous paragraph – most of these migration processes actually followed similar or partially-overlapping directions of dispersion from or to the Italian Peninsula, being also separated by short-evolutionary time frames.

From a genetic point of view, the pioneering work by Cavalli-Sforza, Menozzi and Piazza (1994) on human protein polymorphisms and geographic distribution of human genes (ABO, HLA, etc.), has opened the way to different population genetic studies which progressively addressed questions about Italian population history and that tried to use the current distribution patterns of genetic variation to enhance our understanding on human past migrations. While several studies have focused their attention on specific regions of Italy (e.g. Sardinia, Calò et al., 2008; Central Italy, Onofri et al., 2007; Southern Italy, Ottoni et al., 2009; and Sicily, Di Gaetano et al., 2009) or have otherwise considered particularly interesting Italian peoples (e.g. Etruscans, due to their unsolved origins; Achilli et al., 2007; Brisighelli et al., 2009; Tassi et al., 2013), only few researches have specifically addressed the whole Italian genetic diversity. In the next paragraphs we will particularly concentrate on them, in order to obtain the general picture of the current state of the art on the Italian genetic history and population variability.

1.1.2.1 Classical genetic markers

Over 30 years ago, Cavalli-Sforza et al. (1994) used Principal component analysis (PCA) to create synthetic maps of European (and Italian) genetic variation, summarizing the spatial allele frequency distribution of the so-called *classical genetic markers* (so labelled because these genes actually represented the first typology of markers used in population genetic studies). By collecting count data for many variants of human genes (such as ABO, HLA, etc.) from populations at different geographic locations, they first reconstructed maps representing variation in allele frequencies across space. Subsequently, they used PCA to reduce these many allele-frequency maps into a smaller number of "*synthetic maps*" able to summarizing (in descending order) the most important contributions to the current distribution patterns of genetic variation (Pinhasi et al., 2012). The synthetic map of the first principal component (PC1) in Europe displayed a South-East to North-West genetic cline, which has been interpreted as the reflection of the "*wave of advance*" or "*demic diffusion*" of Near-Eastern farmers into Europe during the Neolithic transition. Since this first overview on the European genetic landscape, some populations clearly appeared as "genetic outliers" within the continental distribution pattern of genetic variation. Among them, the Sardinian people turned out to particularly represent an example of genetic isolate not only within the Italian Peninsula, but also within the whole European and Mediterranean context - as subsequently attested by several population genetic studies bases on classical, uniparental and autosomal markers which specifically addressed the genetic history and variability of Sardinia (see Calò et al., 2008 for a genetic review). Accordingly, Sardinia was not included in the PC analyses performed for building synthetic maps of the Italian genetic variation, since it would have otherwise accounted for most of the information resumed by the first PC (Piazza et al., 1988, Cavalli-Sforza et al., 1994).

The first three PCs of Italy (on the whole accounting for ~60% of the total variability) highlighted some specific patterns within the Peninsula, that have been explained as reflecting peculiar population dynamics occurred during the peopling history of Italy. The first synthetic map (27%) showed a decreasing gradient from the south (with peaks in Southern Calabria and Eastern Sicily) to the north-central part of the Italian Peninsula, moreover highlighting the presence of differences between the eastern and western parts of Sicily (Figure 1.1.2.1.1a; Cavalli-Sfroza et al., 1994). The striking dark areas in Eastern Sicily and Southern Italy were particularly interpreted as mirroring the regions of Greek dominion during the 8th and 9th centuries BC, actually corresponding to the so-called *Magna Graecia*. The demographic impact of Greek colonization on Sicilian and Southern Italian populations is thought to have been remarkable. Consistently, the current distribution of Greek surnames in the area showed a PC1-like pattern with frequency peaks in Southern Calabria (Cavalli-Sfroza et al., 1994). Analogously, it is reported that the Greek language has continued to be spoken in the area at least until the 12th or 13th AD (Cavalli-Sfroza et al., 1994 and reference therein). Interestingly, "Greek-speaking islands" – represented by the so-called ethno-linguistic minorities of Griko-Bovesi (Greek-speakers of Southern Calabria) and Griko-Salentini (Greek-speakers of Southern Apulia) – still currently persist in the extreme southern parts of Calabria and Apulia. Actually, the exact origin of these ethno-linguistic groups remains still uncertain and different hypotheses either linked with i) the foundation of *Magna Grecia*, ii) the Byzantine dominations or iii) the input of Byzantine elements within a pre-existing *Magna Graecia* matrix, have been suggested (Brisighelli et al. 2012 and references therein). More recently, genetic studies based on uniparental markers, suggested little or negligible genetic differences between these Greek-speaking enclaves (at least for the Griko-Salentini of Apulia) and the surrounding Italian populations (Brisighelli et al., 2012). However further studies are requested to more deeply address this point.

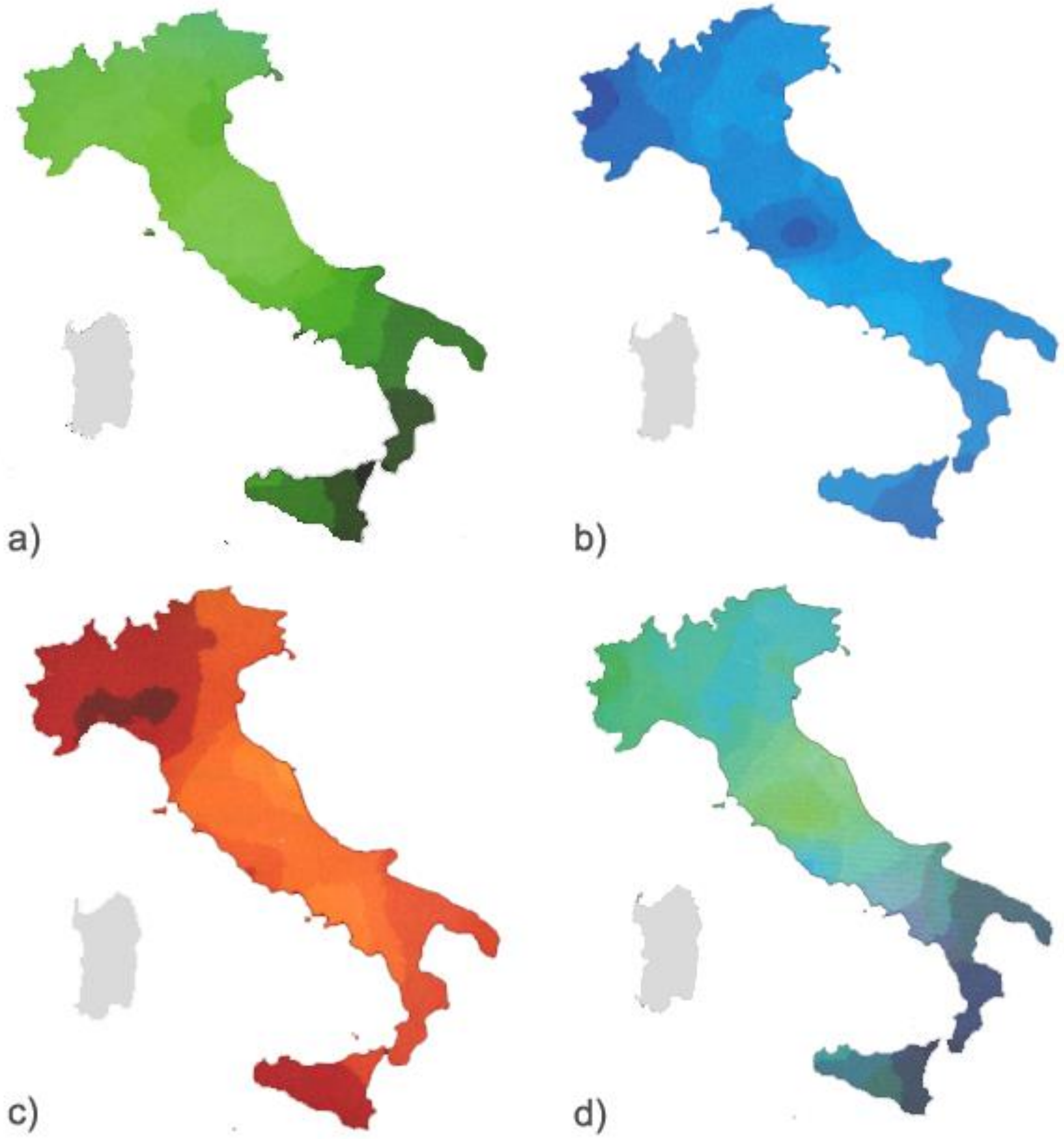


Figure 1.1.2.1.1. Synthetic maps of Italy showing a) the first PC, 27% ; b) the second PC,18%; c) the third PC, 14%; d) the superimposition among the three PCs. Modified from Cavalli-Sforza et al., 1994.

The distribution of genetic variation along other two principal components, respectively PC2 (18%) and PC3 (14%), has been interpreted as accounting for traces of specific Italian pre-Roman populations. In particular, the area of high frequency in the second PC, ranging from southern Tuscany to northern Latium (Figure 1.1.2.1.1b; Cavalli-Sforza et al., 1994), has been observed to almost exactly correspond with the area of diffusion of the Etruscan civilization - documented in central Italy from the 8th century BC to the 1st century AD and defined by a material culture, a non-Indo-European language, and a Greek-derived alphabet (Tassi et al., 2013 and reference therein). Despite being well-documented on both linguistic and archaeological point of views, the exact origin of Etruscan people is still largely debated. Two contrasting hypotheses have been particularly proposed: i) an Anatolian derivation (*Herodotus' view*) and ii) an autochthonous origin in the Italian Peninsula (*Dionysus' view*). First mtDNA-based studies suggested a genetic continuity between Etruscans and modern inhabitants of Tuscany, and consequently interpreted the observed genetic similarities between the latter and the current Turks as supporting the Anatolian origin of the Etruscan people (Achilli et al., 2007). On the contrary, more recent investigations, performed by directly comparing modern and ancient DNA samples, seem to agree in excluding a biological origin of the Etruscans out of Italy, instead interpreting the similarities between Anatolian and Tuscan genetic pools as the effect of older pre-historical contacts (possibly but not exclusively related to the spread of farmers during the Neolithic period), but in any case not related to the later development of the Etruscan culture (Tassi et al., 2013).

The third principal component has been finally associated with the inheritance of those ancient peoples that occupied the Apuan Alps and the Northern Apennines area in pre-Roman times, also known as *Ligurians* or *Ligures* (Figure 1.1.2.1.1c). Historical sources suggest that *Ligures* were not completely eliminated neither by the Roman military campaigns of the 2nd century BC nor by the subsequent massive deportation of more than 45,000 people in the Samnium. On the contrary, they are supposed to have survived in the hilly and mountainous areas of the Appennines, that actually served as places of refuge for these popolations. Similarly, from a linguistic point of view, most of the communities that remained settled in the Apuan Alps and the Northern Apennines are thought to have demonstrated a sort of resistance to the widespread process of "tuscanization" that instead affected the surrounding valley areas (Capocasa et al., 2014 and references therein). Following another line of evidence, recent genetic studies have hypothesized the genetic continuity of the human (isolated) groups today inhabiting the Apuan area (as well as the Samnium) with the ancient *Ligures* tribes (Bertoncini et al., 2012; Capocasa et al., 2014).

The pole opposite to the *Ligurians* in the third PC of Cavalli-Sforza et al. (1994), was supposed to roughly reflect the Early-Iron-Age population of *Picenes*, a people mainly settled along the mid-Adriatic coast and linguistically belonging to the eastern group of the Italic languages, i.e. Osco-Umbro-Sabellic (Cavalli-Sforza et al., 1994).

Despite the PCA approach can provide useful summaries of the data and can help revealing specific hypothesis related to particular aspects of past population history, its results should however be interpreted cautiously and integrated with those provided by other analysis as well as by additional genetic markers, since they only explain a limited portion of the overall genetic variation (Pinhasi et al., 2012). In addition, it has been argued that, in some circumstances, PC1 component may produce gradient perpendicular to the direction of population expansion (Francois et al., 2010) whereas other PC patterns may arise as mathematical artefacts (Novembre and Stephens, 2008).

1.1.2.2 Uniparental molecular markers

After the landmark study by Cavalli-Sforza et al. (1994), a significant part of our present understanding on the Italian genetic variation and population structure has been provided by genetic investigations of the uniparental markers: namely the mitochondrial DNA (mtDNA) and the non-recombining region of Y-chromosome (NRY). Due to their uniparental way of inheritance and the lack of recombination, maternal and paternal genealogies can be traced back both in time and space, providing a temporal framework for mutation accumulation, which can be linked with differential patterns of geographic distribution. In addition, the comparison of maternal and paternal inheritances can provide complementary information about female- and male-mediated aspects of population history and genetic variability (Wilkins, 2006).

After the pioneering effort carried out by Cavalli-Sforza and colleagues (1994), a mtDNA-based study on the Italian population followed in 1995 (Barbujani et al., 1995). Geographic patterns of genetic variation have been investigated in 12 populations from continental Italy, Sicily and Sardinia (Figure 1.1.2.2.1a) totally accounting for 1072 individuals. The analysis of frequencies of 12 common restriction morphs showed non-random spatial patterns in Italy, identifying a North-South major gradient within the Peninsula and underlying the genetic differentiation between Sardinia and the mainland (Barbujani et al. 1995). These patterns have been generally confirmed also when spatial genetic variation was analysed by directly comparing individual haplotype sequences - instead of population morph frequencies - at different geographical distances (autocorrelation index for DNA analysis – AIDA; Bertorelle and Barbujani, 1995). In fact, Sardinian samples proved again to be significantly differentiated from the mainland and, along peninsular Italy, the differences between haplotype pairs tended to increase with their geographical distances, thus confirming the presence of clinal trends. However, the genetic cline observed within the mainland appeared to be less smooth, with genetic differences between Northern and Southern populations becoming significant beyond 800 km (Figure 1.1.2.2.1b, Barbujani et al., 1995).

The distribution of pairwise sequences difference (mismatch distribution) in Peninsular Italy suggested a demographic expansion between 8,000 and 20,500 YBP compatible with the archaeological evidences of population processes occurred both during the LGM and the Neolithic transition. On the contrary, the mismatch distribution in Sardinia reflected the one expected for isolated and demographically stable populations (Barbujani et al., 1995).

More recent mtDNA-based studies, but focused on specific Italian regions (particularly of Central - Onofri et al., 2007 - and Southern Italy - Ottoni et al., 2009), revealed homogeneous patterns of distribution for maternal lineages (mtDNA haplogroups), thus pointing towards a substantial homogeneity for the maternal genetic pool within the different areas of the Peninsula.

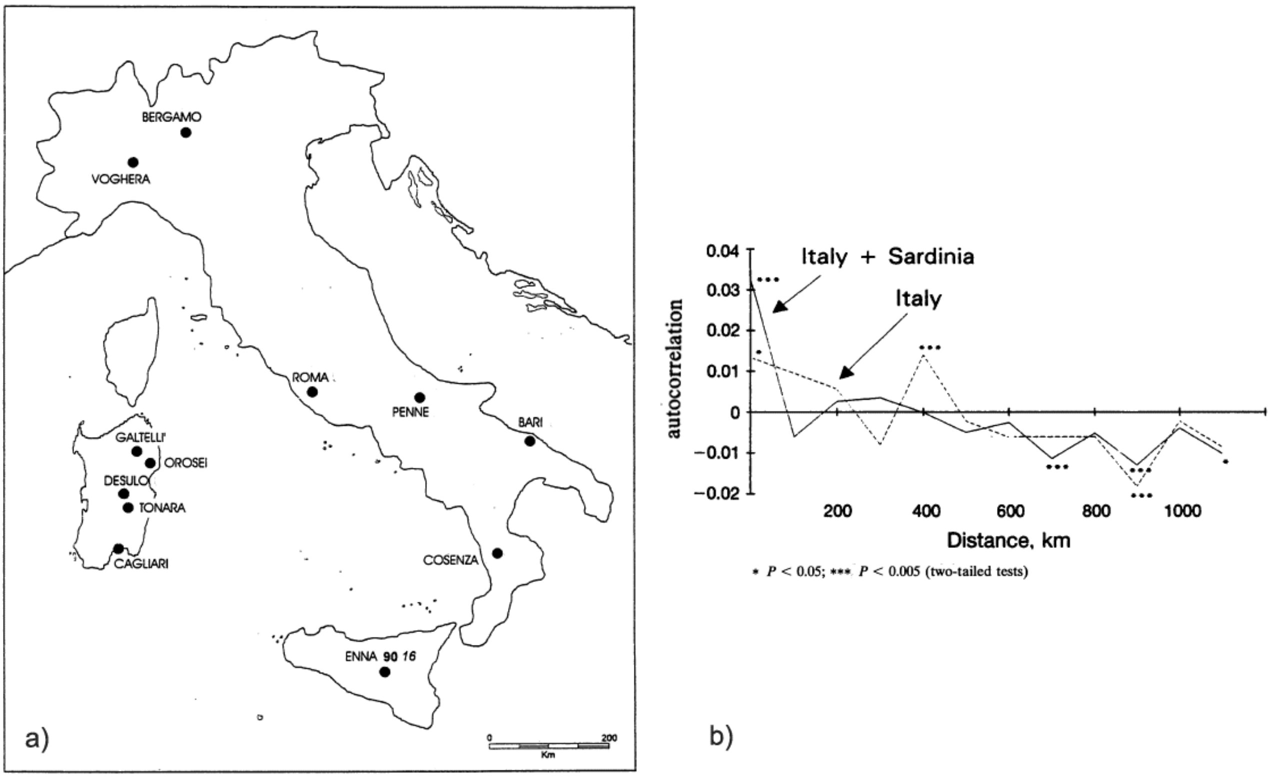


Figure 1.1.2.2.1. a) Geographical location of the samples analysed by Barbujani et al., 1995. b) Coefficients of spatial autocorrelation obtained by AIDA in Italy (dashed line) and Italy + Sardinia (solid line). Modified from Barbujani et al., 1995.

On the paternal perspective, Di Giacomo et al. (2003) carried out an investigation of Y-chromosome diversity in 524 individuals from 17 populations of continental Italy (excluding Sicily and Sardinia; Figure 1.1.2.2.2a). Spatial autocorrelation analysis (Bertorelle and Barbujani, 1995) showed a progressively decreasing pattern in between-samples genetic similarities with the increasing of geographic distances, with genetic differences becoming sharper among populations separated by more than 800 km (i.e. northern vs. southern samples; Figure 1.1.2.2.2b). A single North-West/South-East decreasing major cline was observed within the mainland and associated to the Y-chromosome haplogroups P*(xR1a) (Figure 1.1.2.2.2a) - that today accounts for the most common Western European haplogroup R1b-M269. However local founder or drift effects have been also invoked as factors playing a fundamental role in shaping the micro-geographic Y-chromosomal diversity among the different Italian populations (Di Giacomo et al., 2003).

The detection of one single NW-SE cline within the Peninsula, without any opposite gradients, was suggested to reflect the incomplete colonisation of the Italian territory by different lineages (such as DE, G, J) that entered Southern Italy on the existing P*(xR1a) Palaeolithic background (Di Giacomo et al., 2003), but remained localized at higher frequencies exclusively in some areas of the Peninsula (e.g. reflecting the Greek colonization of Southern-Italy already postulated by Cavalli-Sforza et al., 1994 in the first PC of Italian genetic variation based on classical genetic markers).

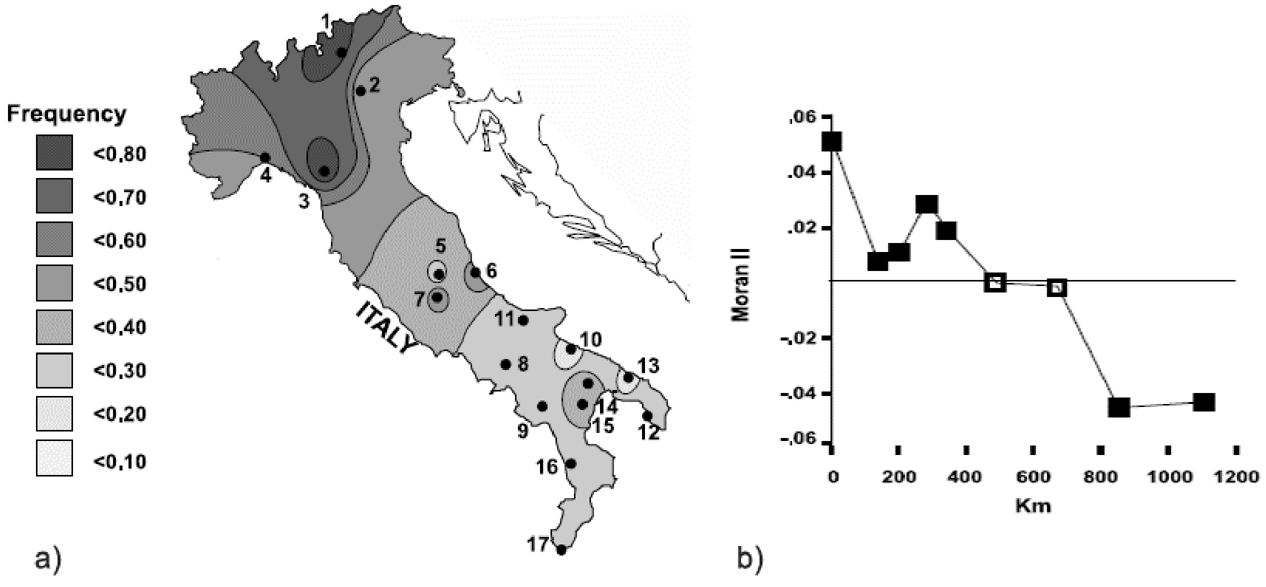


Figure 1.1.2.2.2. a) Geographical location of the samples analysed by Di Giacomo et al., 2003 over-imposed on the interpolated frequency surface map for haplogroup P*(xR1a). Increasing intensity of the haplogroup are represented by the colour-legend at the bottom left. b) Moran II coefficients of spatial autocorrelation obtained by AIDA for Italians. Values significantly different from 0 are represented by filled symbols; open symbols reflect not significantly values. Modified from Di Giacomo et al., 2003.

This first investigation of the Italian paternal genetic composition, was subsequently replicated by Capelli et al. (2007) with a wider set of genetic markers and a higher sampling coverage (699 samples from 12 different locations; Figure 1.1.2.2.3a). Y-chromosome genetic variation confirmed to be not randomly-distributed within the Italian Peninsula, with more than 70% of the observed paternal diversity that appeared to be distributed along latitude-related gradients. Opposite frequency clines were particularly observed for haplogroups R1*(xR1a1), J2 and E3b1 - with northwards increasing frequencies for the former, and decreasing clines along the same direction for the latters (Figure 1.1.2.2.3b). However, contrarily to J2 and E3b1 lineages, no diversity gradient was found to be associated to the R1*(xR1a1) clinal distribution of frequency (Figure 1.1.2.2.3b). This result was interpreted as reflecting the interaction between the Mesolithic – R1*(xR1a1) – and Neolithic – E3b1 and J2 – components of the Italian paternal genetic pool. In particular, being the Mesolithic groups differentially present on the Italian territory depending on the different local demographic histories, they were not expected to generate frequency and diversity gradients for the haplogroup R1*(xR1a1). On the contrary, the subsequent expansion of Neolithic newcomers presumptively resulted in frequency and diversity clines for the haplogroups E3b1 and J2. This consequently produced, along the direction of dispersion, the observed opposite frequency, but not diversity, gradient for Y-chromosome haplogroup R1*(xR1a).

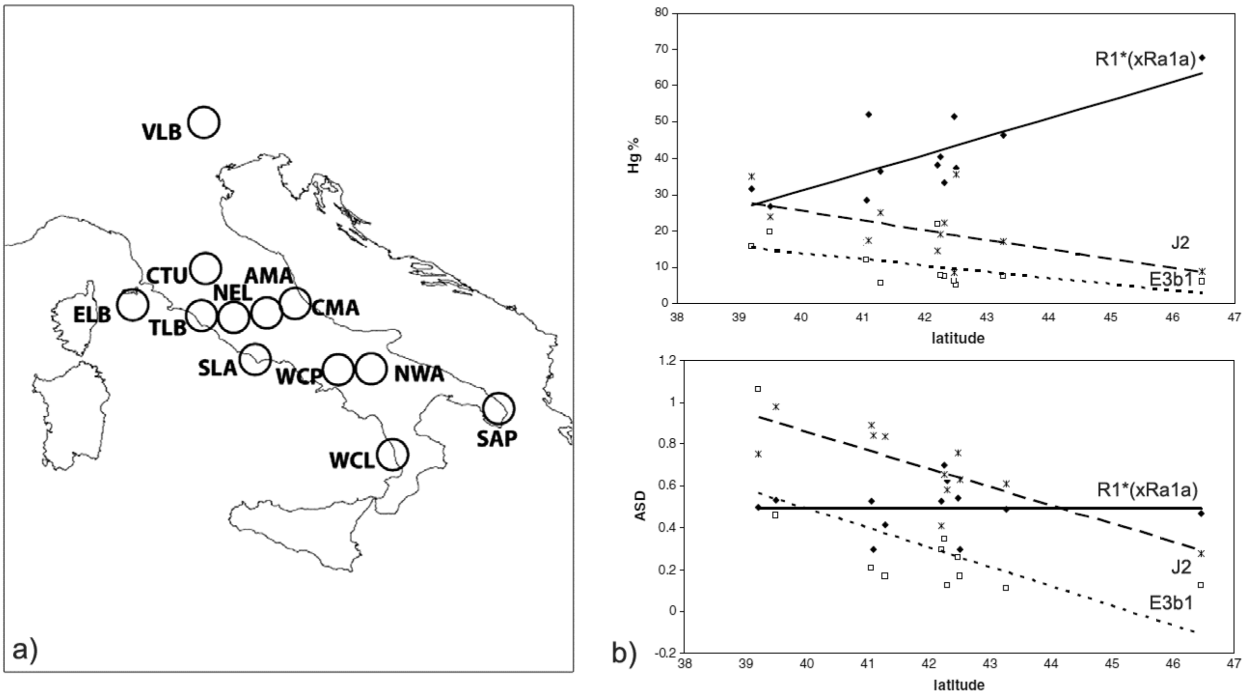


Figure 1.1.2.2.3. a) Geographical location of the samples analysed by Capelli et al., 2007. b) Regression lines for the haplogroup frequencies (top) and average square distance (bottom) versus latitudes. Thick line, hg R1* (xR1a1); dashed line, hg J2; dotted line, hg E3b1. Modified from Capelli et al., 2007.

Consistently with this view, when compared with other European samples, Italian populations clearly fell within the SE-NW continental genetic distribution pattern associated to the admixture between the Neolithic incoming farmers and the pre-existing Mesolithic groups. However some discontinuities between Northern and Southern parts of the Italian Peninsula were also observed and imputed to differential Neolithic/Mesolithic contributes, with Southern samples that presumptively experienced higher Anatolian introgression than Northern ones (Capelli et al., 2007). In particular, the impact of Neolithic transition on Southern Italy was suggested to be greater than the previously postulated contribution of the Greek colonisers (Cavalli-Sforza et al., 1994). Similarly, also North-African genetic contribution to the Southern Italian genetic pool was reported to be negligible (Capelli et al., 2007).

Contrarily to this view, genetic traces of Greek and North-African inputs to the current paternal gene pool of Southern Italy has been detected by a more recent genetic study, which specifically focused on the Sicilian Y-chromosomal variability (Di Gaetano et al., 2009). Although Sicily was generally observed to share a common genetic history with Southern Italy, an internal genetic sub-substructure between the eastern and the western parts of the island was observed and interpreted as the result of different contributions from specific pre-historical and historical population events (Di Gaetano et al., 2009). The differential presence of specific Y-chromosome lineages between the Eastern (G2-P15, J2-M172) and Western (R1b-M269) parts of Sicily, was particularly advocated as another evidence in favour of such a heterogeneity (Figure 1.1.2.2.4). On the other hand, the homogeneous distribution in both the parts of the island of the Balkan-specific haplogroup E-V13, along with its time estimate at around 2380 YBP, suggested the impressive impact (quantified at approximately 37%) of Greek colonization on the Sicilian gene pool, being instead ~6% the contribution estimated for the North-African specific male haplogroup E-M81.

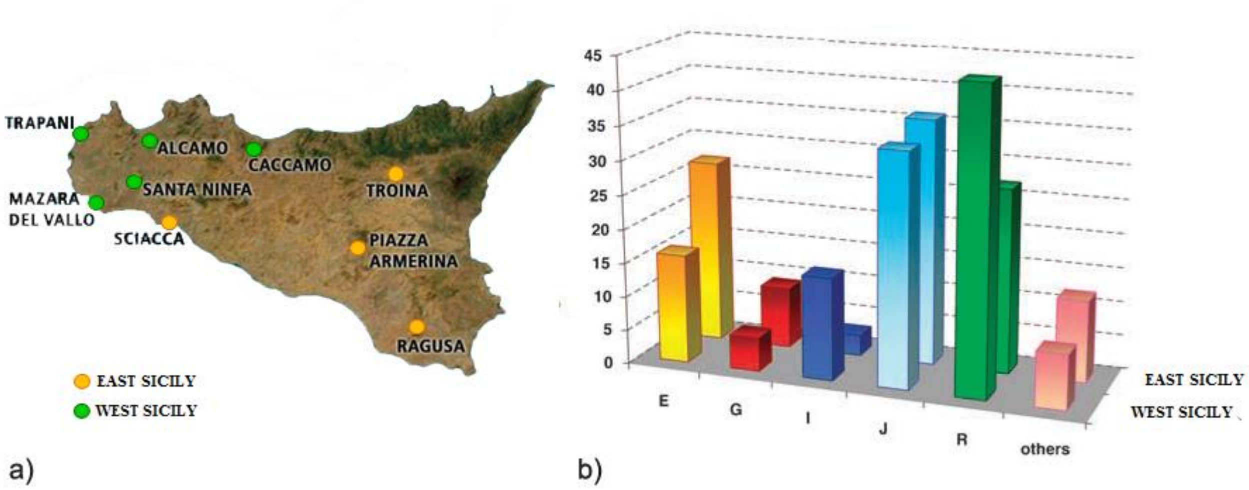


Figure 1.1.2.2.4. a) Geographical location of the Sicilian samples analysed by Di Gaetano et al., 2009. b) Histograms plot of the frequencies of the main haplogroups in the eastern and western sides of the island. Modified from Di Gaetano et al., 2009.

More recently, a first attempt to integrate the maternal and paternal perspectives about Italian genetic variation confirmed the presence of North-South latitudinal clines for both Y-chromosome and mtDNA genetic variability, with Y-chromosome that however revealed more marked regional differences between Northern, Central and Southern Italy, than mtDNA (Brisighelli et al. 2012).

1.1.2.3 Genome-wide based studies

In the last years, the advances in high-throughput SNP-genotyping technologies have greatly refined the description of the overall European genetic variation, providing new details about the genetic relationships and/or similarities between closely-related human populations, both within and among continental regions. The first overall pictures of European genetic structure, emerging by the analyses of genome-wide data, confirmed the strong correlation between genetic and geographic population distances (Lao et al., 2008; Nelis et al., 2009). In fact, the first two PCs showed a SNP-based grouping of European populations that exactly reflected the geographic map of Europe, mirroring the distribution along the northwest-southeast axis already detected by classical and uniparental markers. Within this extremely detailed picture, genetic barriers between Southern Italians and the other European populations (Lao et al. 2008) as well as fine-scale internal genetic differentiation between northern and southern Italian groups (Nelis et al. 2009) have been also invoked, therefore suggesting a more complex and heterogeneous genetic landscape for the Italian Peninsula than what generally observed for the whole Europe.

Moving from these results, a recent genome-wide based study has specifically investigated the genetic structure of the Italian population on a finer geographical scale (Di Gaetano et al., 2012). Besides the general correlation between PC and geography that appeared when Italian populations were compared with other European and reference groups, the Sardinian sample confirmed its "outlier position" within the European genetic landscape, whereas Northern and Southern Italy revealed different patterns of genetic similarities (Figure 1.1.2.2.5, Di Gaetano et al., 2012). More precisely, Northern Italian populations appeared genetically closer to the North-Western European populations (i.e. French and CEU), whereas Southern Italians revealed higher similarities with the South-Eastern Mediterranean groups (actually the Middle Eastern ones, due to the lack of reference populations from the Balkan Peninsula). Interestingly, Sardinia showed a shared ancestral component with both European and North-African (Mozabite) populations, that reveals however to be higher in frequency in the Italian island (70.4%) than in the other comparison populations (Figure 1.1.2.2.5, Di Gaetano et al., 2012).

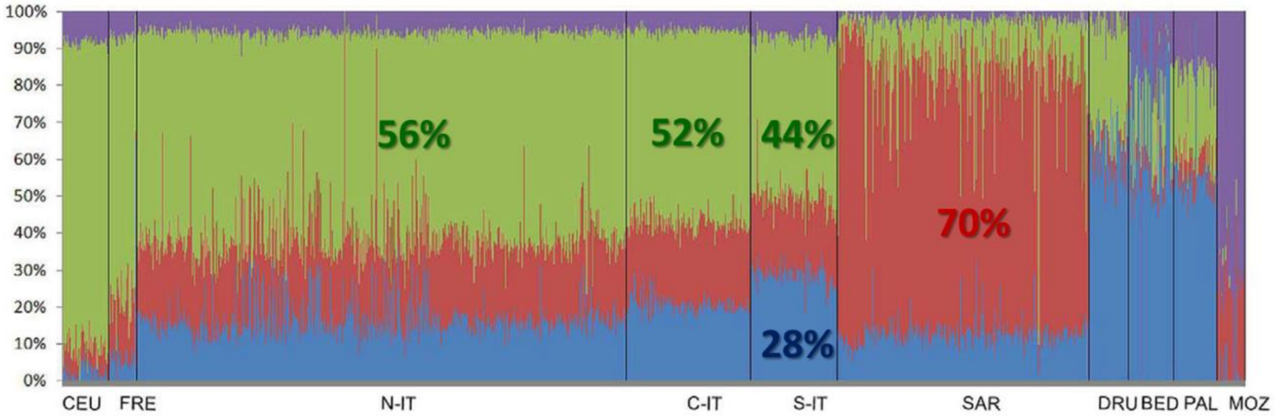


Figure 1.1.2.2.5. Density estimates for empirical distributions of genome-wide mean proportions of alleles sharing identity-by-state (IBS) between subjects from different population or within the same populations. Abbreviations: CEU, Utah Residents with Northern and Western European Ancestry; FRE, French; BED, Bedouin; PAL, Palestinian; DRU, Druze; MOZ, Mozambite; N-IT, North-Italy; C-IT, Central-Italy; S-IT, South-Italy; SAR, Sardinia. Modified from Di Gaetano et al., 2012.

The internal genetic sub-structure observed within the Italian Peninsula, was confirmed also when PCA analysis was repeated by considering exclusively the genetic relationships between Sardinia, Northern-, Central- and Southern-Italian population samples (**Figure 1.1.2.2.6a**). In fact, the first PC detached Sardinia from the Italian mainland, thus stressing further the outlier genetic position of the island within the Italian genetic landscape. On the other hand, the second PC (as well as the PCA analysis performed excluding Sardinia, Figure 1.1.2.2.6b) suggested the partial separation between Southern Italy and an overlapped North-Central Italian cluster. The overlapping pattern between North and Central Italy could be partially explained by the fact that around 79% and 87% of the Northern and Central Italian analysed samples came from Piedmont and Tuscany respectively, both of them actually belonging to the North-Western part of the Italian Peninsula.

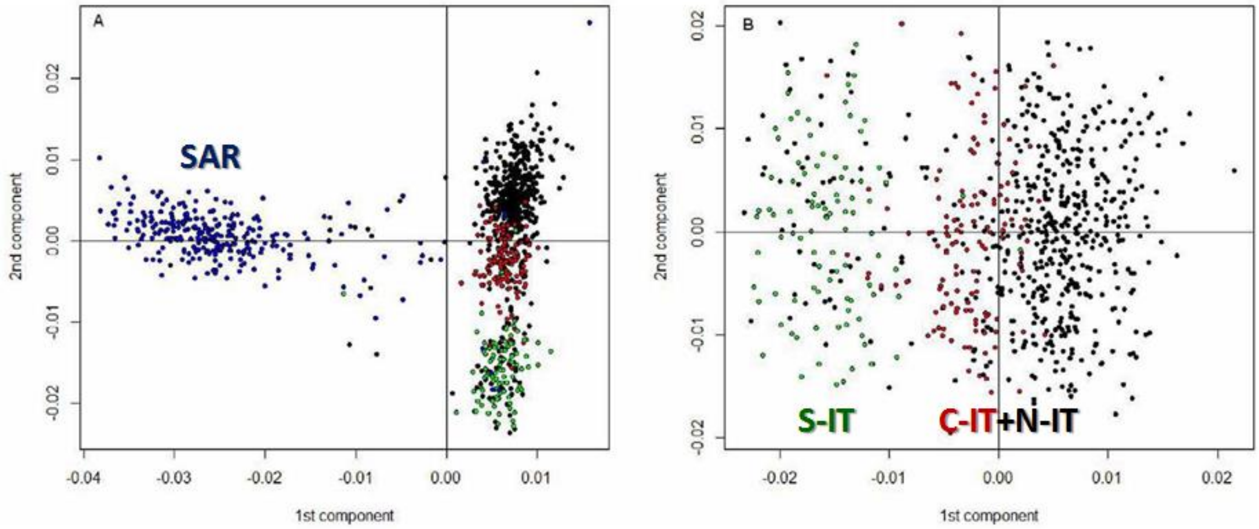


Figure 1.1.2.2.5. Scatter plots of the first two principal component analysis for Italian samples both including (A) and excluding (B) Sardinia. Both analyses were based on 125,799 autosomal SNPs analysed in a total of 1,014 individuals for the Italian dataset and 746 individuals for the Italian dataset without Sardinia. Colour code: Northern Italy (N-IT), black dots; Central Italy (C-IT), red dots; Southern Italy (S-IT), green dots; Sardinian (SAR), blue dots. Modified from Di Gaetano et al., 2012.

1.2. Specific aims of the studies

Previous investigations on the Italian genetic variation based on uniparentally-inherited markers that have tried to reconstruct the population structure and genetic history of Italy, mainly agreed in revealing i) the outlier genetic position of Sardinia and ii) the presence of a major North-South genetic cline within the mainland, that is compatible with the historically and archeologically well-documented interaction between pre-existing Palaeolithic groups and subsequent Neolithic incomers (Barbujani et al., 1995, Capelli et al., 2007). From a paternal perspective (Y-chromosome) genetic traces of that processes have been particularly associated to the differential presence between the Northern and the Southern parts of Italy of particular male lineages. More precisely, the detection of opposite latitude-related gradients for haplogroup R1-M173*(xR1a1) – more frequent in the North – and E3b1-M35 and J2-M172 lineages – instead prevalent in the South – was interpreted as the result of different degrees of Neolithic admixture with Mesolithic inhabitants among the different parts of Italy (Capelli et al., 2007). Similar patterns have been more recently suggested also for the Upper-Palaeolithic (H,K,T*,T2, W and X) and Neolithic (J and T1) maternal (mtDNA) haplogroups, in a first attempt to integrate the maternal and paternal genetic perspectives (Brisighelli et al., 2012).

Despite these clinal patterns in the distribution of maternal and paternal genetic variation, some local and regional differences have been however perceived within the Italian Peninsula, especially for what concerns Y-chromosome markers. Local drift and founder effects were invoked to explain the detected distribution of genetic variation (Di Giacomo et al., 2003). Analogously, a certain degree of discontinuity between Northern and Southern Italy has been also suggested (Capelli et al., 2007; Brisighelli et al., 2012), in line with the results of more recent genome-wide studies at both European continental (Lao et al, 2008, Nelis et al., 2009) and Italian-specific regional (Di Gaetano et al., 2012) scales.

The two Italian major islands, namely Sardinia and Sicily, revealed different demographic histories. In fact, while Sardinia confirmed to be a genetic outlier within Europe and the Mediterranean, with clear signals of founder effects (Barbujani et al., 1995), on the other hand Sicily showed a shared genetic history with Southern Italy, but enriched of Greek and North African influences (Di Gaetano et al., 2009) and characterized by more complex peopling dynamics possibly resulting in east-west genetic heterogeneous pattern in the distribution of genetic variation within the island (Romano et al., 2003; Di Gaetano et al., 2009).

Despite having contributed greatly to the present-state of knowledge on the Italian genetic landscape, most of the above mentioned studies have actually represented either regional studies with a relative high-sampling coverage (Rickards et al., 1998; Romano et al., 2003; Di Gaetano et al., 2009), or broad-scale investigations of the whole Italian genetic landscape with high number of samples and sampling points but a relative unbalanced distribution within the mainland or the major islands (Barbujani et al., 1995; Di Giacomo et al., 2003; Capelli et al., 2007). In addition, most of these studies have mainly considered the maternal (mtDNA) and paternal (Y-chromosome) perspectives separately, the only exception being a recent effort to integrate the two genetic landscapes (Brisighelli et al., 2012), that however reached different levels of resolution in the sampling coverage and molecular analysis between the two uniparental markers. Concerning the sampling strategy, a broad scientific consensus has been reached that adequate sampling coverage and accurate selecting criteria are essential elements to obtain representative results and to draw reliable conclusions about human population history. Similarly, the contemporaneous investigation of both uniparental genetic systems, combined with the use of slow- and fast-evolving genetic markers, has largely proved to be a suitable approach to explore and interpret the different patterns of genetic variation, even at fine-scale levels of analysis (Underhill and Kivisild 2007; Larmuseau et al., 2011). Each of the two uniparental systems may actually be responsible for sex-specific spatial and temporal paths of genetic introgression, an important aspect that might has been ignored in the more recent whole-genome investigations.

In this thesis we therefore aim to update our knowledge on the Italian population genetic history by increasing the specificity of sampling strategy and the resolution level of the molecular markers analysed, in such a way as to achieve a comprehensive in-depth high-resolution evaluation of both maternal and paternal genetic landscapes at the same time.

As a first step, we specifically focused on the Italian genetic structure in order to distinguish between clinal and/or discontinuous patterns of genetic variation, and seeking to identify the main times and population movements in the origin of the observed genetic diversity. While doing this moreover looked for the presence of any sex-biased genetic pattern in the Italian population variability. The results of this study are presented in *Article 1*.

Subsequently we looked at a particularly strategic Italian local context, that is Sicily and Southern Italy, in order to provide new clues about past population contacts and genetic interactions, and aiming to interpret the patterns observed in the Italian Peninsula within the wider context offered by the European and Mediterranean genetic relationships. In addition we specifically addressed some unsolved questions at finer regional scale such as the presence or the absence of an east-west internal genetic differentiation within Sicily. The results of this study are presented in *Article 2*.

A more detailed discussion of some aspects of the results, for both the abovementioned articles, are presented afterwards as a commentary.

1.3. Results and Discussion

Article 1

Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, Sanz P, Yang-Yao D, Manry J, Ciani G, Luiselli D, Quintana-Murci L, Comas D, Pettener D; Genographic Consortium (2013) *Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata*. PLoS One. 8:e65441. doi:10.1371/journal.pone.0065441.

OPEN ACCESS Freely available online

PLOS ONE

Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata

Alessio Boattini[1]◑, Begoña Martinez-Cruz[2]◑, Stefania Sarno[1], Christine Harmant[3,4], Antonella Useli[5], Paula Sanz[2], Daniele Yang-Yao[1], Jeremy Manry[3,4], Graziella Ciani[1], Donata Luiselli[1], Lluis Quintana-Murci[3,4], David Comas[2]* , Davide Pettener[1]* , the Genographic Consortium¶

1 Laboratorio di Antropologia Molecolare, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, Bologna, Italy, **2** Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Spain, **3** Institut Pasteur, Human Evolutionary Genetics Unit, Department of Genomes and Genetics, Paris, France, **4** Centre National de la Recherche Scientifique, Paris, France, **5** Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, Sassari, Italy

Abstract

Located in the center of the Mediterranean landscape and with an extensive coastal line, the territory of what is today Italy has played an important role in the history of human settlements and movements of Southern Europe and the Mediterranean Basin. Populated since Paleolithic times, the complexity of human movements during the Neolithic, the Metal Ages and the most recent history of the two last millennia (involving the overlapping of different cultural and demic strata) has shaped the pattern of the modern Italian genetic structure. With the aim of disentangling this pattern and understanding which processes more importantly shaped the distribution of diversity, we have analyzed the uniparentally-inherited markers in ~900 individuals from an extensive sampling across the Italian peninsula, Sardinia and Sicily. Spatial PCAs and DAPCs revealed a sex-biased pattern indicating different demographic histories for males and females. Besides the genetic outlier position of Sardinians, a North West–South East Y-chromosome structure is found in continental Italy. Such structure is in agreement with recent archeological syntheses indicating two independent and parallel processes of Neolithisation. In addition, date estimates pinpoint the importance of the cultural and demographic events during the late Neolithic and Metal Ages. On the other hand, mitochondrial diversity is distributed more homogeneously in agreement with older population events that might be related to the presence of an Italian Refugium during the last glacial period in Europe.

Citation: Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, et al. (2013) Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata. PLoS ONE 8(5): e65441. doi:10.1371/journal.pone.0065441

Editor: David Caramelli, University of Florence, Italy

Received January 8, 2013; **Accepted** April 24, 2013; **Published** May 29, 2013

Funding: This study was supported by Strategic Project 2006-09 from the University of Bologna to DP and from MIUR PRIN 2007 and 2009 Grants to DP. The project was also supported by the Spanish Government grant CGL2010-14944/BOS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.comas@upf.edu (DC); davide.pettener@unibo.it (DP)

◑ These authors contributed equally to this work.

¶ Membership of the Genographic Consortium is provided in the Acknowledgments.

Introduction

Due to its central position and to the extension of its coastal line (~7,460 Km), the modern Republic of Italy – e.g. the Italian Peninsula and the two major islands of Sicily and Sardinia – has been one of the focal points in the settlement history of Southern Europe and the Mediterranean Basin.

Populated by early modern humans since approximately 30,000–40,000 years before present (YBP) [1] during the LGM (~25,000 YBP) it was involved in the southward contraction of human groups from Central Europe that rapidly retreated to the Mediterranean coastlines, occupying refuge areas, such as in the well-known cases of Iberia and the Balkans [2–5]. After contributing to the substantial re-shaping of the early Paleolithic genetic composition of glacial Refugia, northward re-peopling processes started approximately 16,000–13,000 YBP [3], [6–9].

Subsequently Italy has received the passage of multiple human groups in prehistoric and historic times, acting both as a step point and an area of expansion during the different major migratory events following the early Paleolithic colonization.

The most recent archaeological syntheses [10] describe the early Neolithisation of Italy as the result of two independent and parallel processes, involving respectively the Adriatic and the Tyrrhenian coasts and dating as early as 8,100 YBP (Apulia, South-Eastern Italy) and 7,900 YBP (Liguria, North-Western Italy).

Italian Late Neolithic and the Metal Ages revealed to be a complicated tapestry of different cultural strata, potentially associated with population movements. During the first millennium BC, Italy hosted a vast set of different peoples whose origins in some cases remain unknown (e.g. Etruscans, Ligurians, Veneti), while in other cases are the result of specific migration processes (Celts in North-Western Italy; Greeks in Southern Italy and Sicily) [11].

In addition, independent and/or intersecting subsequent historic events (related with the trade and expansion of different populations in our era: Phoenician, Greek, Carthaginian, Roman, Arabic and Barbaric) also contributed to the present genetic composition of Italy. Unlikely to have completely deleted precedent genetic structures, such migrations may have resulted in partially overlapping patterns of diffusion within Italy.

At present, only few studies addressed the reconstruction of the genetic structure and history of Italian populations. Barbujani and colleagues (1995), in a study based on mtDNA variability [12], identified a North-South gradient within the peninsula, confirming what was previously revealed by classical genetic markers [13], while underlying the genetic differentiation between Sardinia and the mainland [12]. More recent studies focused only on specific regions of Italy and revealed a homogeneous pattern of distribution for mtDNA haplogroups. These findings point towards a substantial homogeneity of the mtDNA gene pool within the different areas of the Peninsula [14], [15].

On the paternal perspective, Di Giacomo et al. (2003) carried out an investigation of Y-chromosome diversity in continental Italy [16]. They identified a single decreasing North-South major cline within the Peninsula, while local drift and founder effects were invoked to explain the observed distribution of genetic variation. The study was replicated by Capelli et al. (2007) with a much larger set of genetic markers and a more specific sampling strategy [17]. They observed that more than 70% of the detected diversity was distributed along latitude-related gradients. A certain level of discontinuity was suggested between Northern and Southern portions of the Italian peninsula that, according to the authors, may be related to differential Neolithic/Mesolithic contributes in the two regions [17]. These results – North-South clinal patterns related to differential Neolithic contributes – were largely confirmed in a recent update of the same study adding more populations and including mtDNA information [18]. Some discontinuity between Northern and Southern Italy was apparent also in genome-wide studies at the European geographical scale [19], [20] and in a specific analysis on Italian samples [21].

Although a common north-south cline has been described for maternal and paternal lineages in Italy, recent data on the Neolithisation of southern Europe [22], [23] suggest a sex-biased Neolithic migration that might account for an asymmetrical pattern of structure in Italy. Eventually more recent migrations could have magnified these sex-biased patterns. For example, this seems to be the case for the first Greek groups in Southern Italy and Sicily, reportedly biased towards a low number of females [11]. Such differential sex-specific demographic events could therefore have affected the genetic structure of Italy in a way that might have been ignored in recent whole-genome analyses.

The present research aims to update our knowledge about Italian population genetic history, by increasing the specificity of sampling strategy and the resolution power of uniparental molecular markers. For the first time, we present an extensive study of both mitochondrial DNA and Y-chromosomal variation in the Italian Peninsula, Sicily and Sardinia. Almost 900 individuals from eight sampling macro-areas have been deeply typed for 136 SNPs and 19 STRs of Y-chromosome, as well as for the whole control region and 39 coding SNPs of mtDNA. We use this detailed and complete dataset to address the following issues. First, we seek to describe the genetic structure of Italy and compare it with the patterns obtained before, in order to distinguish between a clinal and a discontinuous pattern of genetic variation. Second, we want to investigate whether the structure observed is sex-biased and which factors could account for any differential contributes from paternal and maternal lineages. Third, we seek to identify which population movements mostly could be in the origin of the current genetic diversity of the Italian populations.

Materials and Methods

Ethics Statement

For all subjects, a written informed consent was obtained, and Ethics Committees at the Universitat Pompeu Fabra of Barcelona (Spain), and at the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna (Italy), approved all procedures.

Sample collection

A total of 884 unrelated individuals from continental Italy, Sicily and Sardinia were collected according to the following sampling strategy. Firstly, based on the results of a precedent reconstruction of the surname structure of Italy [24], we defined lists of monophyletic surnames for each of the 96 Italian provinces. Secondly, monophyletic surnames frequencies were used to define eight clusters of homogeneous Italian provinces (sampling macro-areas, Figure S1). Within each sampling macro-area, we selected a set of provinces (sampling points) from a minimum of one to a maximum of three, depending on the geographical extension of the macro-area as well as their historical background. This was done in order to depict a sampling grid able to capture as much genetic variability as possible (given the number of planned samples/sampling points). Within each sampling point, individuals were finally sampled according to the standard 'grandparents' criterion, thus considering as eligible for our study only those individuals whose four grandparents were born in the same sampling macro-area. It is important to underline that individuals within sampling points were not selected by surnames. That way 1) our data are consistent with those from other similar studies; 2) we avoid to introduce a bias between Y-chromosome and mtDNA results.

DNA was extracted from fresh blood by a Salting Out modified protocol [25].

Y-chromosome genotyping

A total of 884 samples were successfully typed for Y-chromosome markers. 121 SNPs in the non-recombining region of the Y chromosome were genotyped using the OpenArray® Real-Time PCR System (Applied Biosystems) as described previously [26]. Six additional SNPs (M91, M139, M60, M186, M175, and M17) were genotyped in a single multiplex, Multiplex2 [27]. Nine additional single SNPs (M227, L22, M458, L48, L2, L20, M320, P77) were typed with individual TaqMan assays. Nomenclature of the haplogroups is in accordance with the Y-Chromosome Consortium [28]. Detailed phylogeny may be found at Y-DNA SNP Index - 2009 (http://isogg.org/tree/ISOGG_YDNA_SNP_Index09.html). For simplicity reasons, we will use asterisks (*) to indicate those chromosomes that are derived at a certain SNP, but ancestral at all the tested downstream SNPs.

All individuals were additionally typed for a set of 19 STRs: 17 using the Yfiler kit (Applied Biosystems) and two (DYS388, DYS426) included in the Multiplex2. As the Yfiler kit amplifies DYS385a/b simultaneously avoiding the determination of each of the two alleles (a or b), DYS385a/b were excluded from all the analyses performed. DYS389b was obtained by subtracting DYS389I from DYS389II [29].

Mitochondrial DNA genotyping

865 samples were successfully sequenced for the whole control region as in Behar et al. (2007) [30], and typed using a 22 coding region SNPs multiplex as described previously [27], [31]. Variable positions throughout the control region were determined between positions 16,001 and 573. Sequences were deposited in the GenBank nucleotide database under accession numbers

KC806300-KC807164. In addition, for haplogroup H, the most frequent in Western Europe [2], [6], we used a specifically designed multiplex (named HPLEX17) in order to resolve 17 distinct sub-lineages [27]. Based on combined HVS sequence and coding region SNP data, individuals were assigned to the major haplogroups of the mtDNA phylogeny with the software Haplogrep [32] that uses Phylotree version 13 [33]. Due to their phylogenetic uncertainty, indels at nucleotide positions 309, 315, and 16193 were not taken into account.

Statistical Analyses

Population structure and genetic variability. Haplogroup frequencies were estimated by direct counting. Standard diversity parameters (haplogroup diversity, number of observed STR haplotypes, sequence diversity values, and mean number of pairwise differences) were calculated with Arlequin 3.5 [34]. FST and RST results were corrected with Bonferroni test for multiple comparisons ($p<0.05$).

The relationships between geographical distances and genetic diversity were investigated by using several spatial analyses. The correlation between geographical distances and genetic distances (Reynolds distance), based on haplogroup frequencies, was evaluated by means of a Mantel test (10,000 replications). In order to distinguish any clinal pattern (Isolation-by-Distance pattern) from any discontinuous genetic structure (both of them can result in significant correlations with geography), geographical distances were plotted against genetic ones. A 2-dimensional kernel density estimation layer [35] was added to the plot in order to highlight the presence of discontinuities in the cloud of points. The analysis was performed with all the samples and then removing the Sardinian ones, given their outlier status previously described in literature [7], [13], [21], [36–38].

To further explore spatial patterns of variation a spatial principal component analysis (sPCA) based on haplogroup frequencies was performed using the R software package *adegenet* [39–41]. Additional information about the sPCA method is provided in Methods S1.

To further test the significance of the structure found with the sPCA analysis, we carried out a series of hierarchical analyses of molecular variance (AMOVA) pooling populations according to the sPCA results. We used haplogroup frequencies (both Y-chromosome and mtDNA), RST distances (Y-STRs) and number of pairwise differences (HVRI-HVRII mtDNA sequences). In order to explore genetic variability within the most frequent haplogroups, and in particular within those identified by sPCA loadings, we applied a Discriminant Analysis of Principal Components (DAPC) to Y-STR haplotypes and mtDNA sequences (see Methods S1 for more details). Analyses were performed using the R software *adegenet* package [39–41]. In addition, for comparison purposes we calculated a Network representation of haplogroup G2a using a Median Joining (MJ) algorithm as implemented in the Network 4.6.1.1 software (http://www.fluxus-engineering.com, [42]), weighting STR loci according to the variance method.

DAPC was first performed using Italian haplotypes only. As a second step, in order to investigate the origin of the genetic diversity for the most common haplogroups in Italy, additional individuals from selected European populations were incorporated into the DAPC of major haplogroups. Unpublished 194 Y-chromosome data from Iberia, Germany and the Balkans were provided by the Genographic Project, while data for Causasus and Western Anatolia were extracted from literature [43], [44]. Comparison data for mtDNA was generated using additional information from Basque [45], Austrian [46] and Balkan samples [46], [47].

Y-chromosome and mtDNA dating. In order to minimize the biasing effect of STRs saturation through time (especially important for rapidly evolving STRs as some of those included in the Yfiler kit, [48]), all Y-chromosome age estimations were calculated selecting the eight markers (DYS448, DYS388, DYS392, DYS426, DYS438, DYS390, DYS393, DYS439) with the highest values of duration of linearity D approximated as in Busby et al. (2011) [49].

Splitting time between the sPCA-identified regions (NWI and SEI, see Results) was estimated with BATWING [50] under a model of exponential growth and splitting from a constant size ancestral population. Two samples (Treviso, Foligno/PG) were excluded from the analysis according to a 5% quantile threshold of the sPC1 scores. Two chains with different starting points were run with a total of 3.5×10[6] samples with an initial burn in of 1.5×10[6] samples and a thinning interval of 10×20. The outfiles were treated with the R package [41] to get the posterior distributions of the parameters of interest. We checked that results were equivalent for both runs and reported the mean values of both analyses for every parameter. We used a prior distribution for mutation rates as proposed by Xue et al. (2006) [51] based on Zhivotovsky et al. (2004) [52]. Such distribution is wide enough to encompass all mutation rates for each of the eight considered Y-STRs. A generation time of 25 years was used [52]. Priors and further information about the BATWING procedure are shown in the Methods S1.

The age of Y-chromosome DAPC clusters exhibiting peaks of frequency higher than 70% in any of the sPCA-identified populations (NWI, SEI, and SAR) – with the exception of haplogroup G2a due to its particular relevance in our populations (see Results) – and composed by at least ten individuals, as well as the age of the entire haplogroups, were estimated with the standard deviation (SD) estimator [53]. Differently from BATWING, this method does not estimate the population split time, but the amount of time needed to evolve the observed STRs variation within haplotype clusters (or whole haplogroups) at each population. As for mutation rates, we adopted locus-specific rates for each of the eight considered loci as estimated by Ballantyne et al. (2010) [48]. These rates were preferred to the 'evolutionary' one [52] for the following reasons: 1) 'germline' rates are locus-specific and based on the direct observation of transmission between father-son pairs; 2) 'germline' rates share the same magnitude with genealogy based estimates [54] while the 'evolutionary' rate is a magnitude lower; 3) a recent study [43] suggested that family based rates (germline, genealogies) provide a better fit with history and linguistics. The 95% confidence intervals of time estimates were calculated based on the standard error (SE). Only individuals with a membership >99% in their corresponding DAPC clusters were considered. Given that moments like mean and variance – hence time estimates based on variance – are very sensitive to the presence of outliers (e.g. non-robust), we designed a "jackknife-like" procedure in order to detect possible outlier individuals that could be significantly biasing our estimates (see Methods S1 for details).

TMRCA for the most common mtDNA haplogroups was estimated by means of the ρ (rho) statistic with the calculator proposed by Soares et al. (2009) [55] for the entire control region (that considers a mutation rate corrected for purifying selection of one mutation every 9,058 years).

However, results have to be taken with caution, given that molecular date estimates with ρ can be affected by past demography. Simulations show that error rates tend to increase with effective size, bottleneck and growth effects [56]. In order to avoid sampling errors, the estimates were calculated only for those haplogroups with absolute frequencies higher than 30 individuals.

Results

Y-chromosome lineages in Italy

Haplogroup frequencies. A total of 884 unrelated individuals from 23 Italian locations (Figure S1) were successfully genotyped for 19 STRs and 136 SNPs, and classified in 46 different haplogroups (including sub-lineages) whose phylogeny ([28]; ISOGG Y-DNA SNP Index – 2009) and frequencies for the whole dataset are detailed in Table S1; Y-STR haplotypes of each individual are provided in Table S2.

The haplotype and haplogroup diversity (h), STR diversity (π_n) and mean number of pairwise differences (π) of the population samples are listed in Table S3. The lowest values for haplogroup diversity (h) are observed in Sardinia, while the Italian peninsula is characterized by a negative correlation between haplogroup diversity and latitude, resulting in a south-north decreasing pattern of variation (Spearman's rho $= -0.463$, p-value $= 0.036$). The most frequent haplogroups in Italy are R-U152* (12.1%), G-P15 (11.1%), E-V13 (7.8%) and J-M410* (7.6%). They are followed by three R1b-lineages (R-M269*, R-P312* and R-L2*), whose frequencies ranged from 6.9% to 5.7%; and finally from I-M26, which embraced more than the 4% of total variability. On the whole these haplogroups encompass ~62% of Y-chromosomes lineages, while the remaining 38 haplogroups show frequencies lower or equal to 3.3%. Haplogroups distribution in the considered eight sampling areas is detailed in Table S1.

Paternal population structure. In order to explore the relationship between geographical and paternal genetic distances among the 23 investigated Italian populations a Mantel test was performed. A significant correlation was found (observed value $= 0.26$, p-value $= 0.006$), even after removing Sardinian samples (observed value $= 0.19$, p-value $= 0.03$). However, a non-homogeneous distribution of points is apparent when plotting geographical distances against genetic ones (Figure S2), indicating that the genetic structure of Italy is better characterised by discontinuities than by clinal patterns.

These general spatial patterns were further explored by means of sPCA based on haplogroup frequencies. The analysis showed that the Italian genetic structure is characterised by two significant global components (positive eigenvalues) with similar variance values, being sPC1 characterized by a higher spatial autocorrelation (Moran's I) (Figure S3). These observations are further assessed by means of a significant Global test (observed value $= 0.08$, p-value $= 0.015$) and a non-significant Local test (observed value $= 0.06$, p-value $= 0.677$).

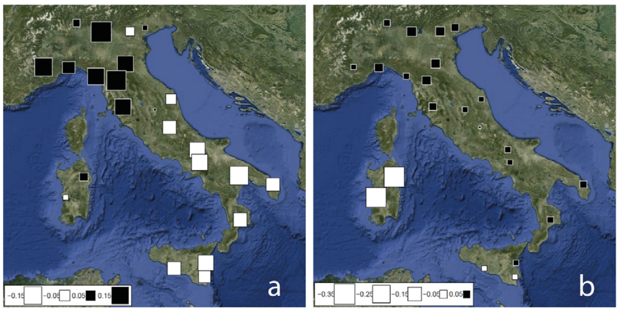


Figure 1. Spatial Principal Component Analysis (sPCA) based on frequencies of Y-chromosome haplogroups. The first two global components, sPC1 (a) and sPC2 (b), are depicted. Positive values are represented by black square; negative values are represented by white squares; the size of the square is proportional to the absolute value of sPC scores.
doi:10.1371/journal.pone.0065441.g001

Geographical patterns of sPC1 and sPC2 are plotted in Figure 1. sPC1 identifies two main groups of populations separated by an almost longitudinal line (Figure 1a). The first group (black squares) is represented by populations from North-Western Italy, including most of the Padana plain and Tuscany. The second group (white squares) includes locations from South-Eastern Italy and the whole Adriatic coast, being represented also in North-Eastern Italy. Nonetheless, these two groups are not separated by a sharp discontinuity, but by some sort of gradient, as it is represented by a few samples from North-Eastern and Central Italy that show very low absolute values of sPC1 scores. However, sPC2 scores differentiate Sardinia from the rest of Italy (Figure 1b). Indeed, scores from these populations show the highest absolute values, while those from the other Italian locations (especially in the South) are much lower. In summary, sPC1 and sPC2 depict a three-partitioned structure of Italian population: 1) North-Western Italy (from now on NWI), 2) South-Eastern Italy (from now on SEI), and 3) Sardinia (from now on SAR).

When we tested the reliability of these three groups (NWI, SEI, SAR), by means of AMOVA based both on haplogroup frequencies and STR variability, the proportion of variation between groups (haplogroup frequencies: 3.71%; haplotypes: 4.48%; both p-values <0.001; Table S4) was 1.5 times higher than the variation explained when grouping according to the eight sampling macro-areas (2.62%, p-value <0.001, and 3.11%, p-value <0.001, respectively, Table S4). Interestingly, there is a partial congruence between sPCA-based groups and sampling macro-areas (Figure S1). In particular, SAR coincides with macro-area 8, while macro-areas 1, 3 and 4 are grouped in NWI and macro-areas 6 and 7 are grouped in SWI; macro-areas 2 and 5 are crossed by the sharp gradient that separate NWI from SEI.

To further test the reliability of the mentioned structure, for each of the considered populations we calculated DAPC-based posterior membership probabilities to the considered three groups. Results (Table S5) show that all the populations are characterised by high congruence (membership probability $= \sim 9\%$ or higher) to the given SPCA-group, the only exception being a single population from Central Italy (Foligno/PG), whose intermediate position between NWI and SEI has been already revealed by sPCA.

Interestingly, NWI revealed a high and significant degree of internal differentiation, while SEI is a fairly homogeneous group (Fst = 0.014, p-value <0.001 and Fst = 0.002, p-value >0.05, respectively; both estimates are based on haplogroup frequencies).

In order to quantify the contribution of each haplogroup to the genetic structure detected, the loadings values of the sPC1 and sPC2 were calculated and plotted in Figure S4. Lineages contributing more to the differentiation along the first sPC were R-U152*, G-P15 and, with lower loadings values, R-L2* and R-P312* (Figure S4a). On the contrary, sPC2 is influenced primarily and almost exclusively by the haplogroup I-M26 (Figure S4b).

Haplogroup DAPC analysis. DAPC was performed within the most frequent haplogroups (E-V13, G-P15, I-M26, J-M410*, R-P312*, R-U152*, R-L2*). Results (Table 1, Figure 2, Figure S5) show how the seven considered haplogroups disaggregate in 25 clusters, ranging from a minimum of two (I2a-M26) to a maximum of five (E-V13, G2a-P15). Considering a 70% threshold, 13 out of 25 are mostly frequent in one of the sPCA-identified areas (NWI: 7, SEI: 4, SAR: 2) (Table 1).

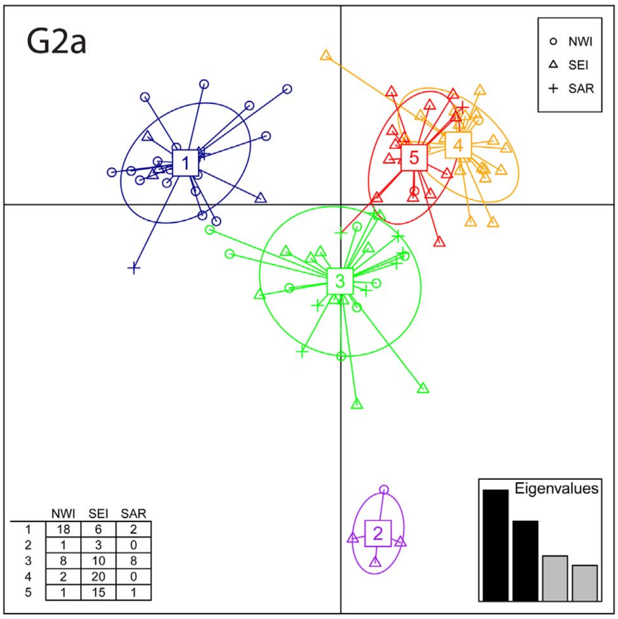


Figure 2. Discriminant Analysis of Principal Components (DAPC) for G2a-P15 haplotypes. Samples are grouped according to their affiliation at the sPCA-identified groups (NWI; SEI; SAR; symbols in the top right table). The table in the bottom left shows the number of haplotypes in each of the five G2a clusters and their geographical distribution in the three Italian areas. DAPC eigenvalues are depicted in the enclosed barplot.

doi:10.1371/journal.pone.0065441.g002

It is noteworthy the structure shown by haplogroup G2a-P15 (Figure 2), which includes clusters with very different spatial distribution: cluster 1 is mostly frequent in NWI, while clusters 4 and 5 – partially overlapping in the DAPC plot – are found in SEI. For comparison purposes, we calculated a Median Joining Network (Figure S6) based on the same haplotypes. While results from both methods are largely overlapping, DAPC offers some advantages compared to the network, namely 1) it outputs clear-cut clusters (while in Network the definition of clusters is in some way arbitrary), 2) it gives probability memberships for each individual. Networks for other haplogroups are not shown.

DAPC comparisons with additional samples (Table S6, Figure S7) suggest differential affinities for some of the considered haplogroups and clusters of haplotypes. Most notably, G2a-P15 haplotypes from NWI cluster mainly with German ones, while haplotypes from SEI seem to indicate wider relationships, going from Iberia to the Balkans and the Caucasus. On the contrary, I2-M26 samples from Sardinia (SAR) cluster in a separate group than Iberians, suggesting a geographical neat separation between continental and Sardinian I2-M26 lineages.

Date estimates for paternal variation. BATWING was used to estimate the age of split between the Italian regions identified by the first sPCA (NWI and SEI, excluding SAR). BATWING modelled population growth starting at 12,890 YBP (95% CI: 3,700–83,070), with a rate of 0.00429 (95% CI: 0.00254–0.01219) per year. Our results suggest that the split happened around 5,490 YBP (95% CI: 1,620–26,830). Since BATWING does not consider migration, admixture between NWI and SEI would likely bias the split time estimate towards more recent dates.

Concerning Y-chromosome lineages, STR variation within the 13 clusters mentioned above suggests that most of them date back to relatively recent times (Table 2). In fact, the ages of the considered clusters (with a peak in one of the considered sPCA groups) fall roughly within the interval from the time of split estimated with BATWING between NWI and SEI and the present. This is consistent with the fact that group-specific clusters of haplotypes (NWI, SEI) are very likely to have emerged after the split within the Italian 'ancestral' population or later. No different patterns of timing are detected between both regions. The time estimates were similar for whole haplogroups with the notable exception of G2-P15, which showed older ages. These results suggest that most of the Y-chromosomal diversity present in modern day Italians was originated from few common ancestors living during late Neolithic times and the Early Metal Ages. However, if we would take into account evolutionary rates, we would observe results three times higher than those above mentioned, meaning that most dates would shift to late Paleolithic.

Mitochondrial DNA lineages in Italy

Haplogroup frequencies. The maternal genetic ancestry of Italian populations was explored by characterizing coding region SNPs and control region sequences from 865 individuals, which yielded to 79 distinct mtDNA haplogroups (including sublineages). Haplogroup frequencies and within-population diversity parameters are shown in Table S7 and Table S3 respectively.

The haplogroup distribution in Italy reflects the typical pattern of mtDNA variability of Western Europe. As described for other European and Italian populations [2], [6], [14], [15], [57] most of the sequences belong to the super-haplogroup H, which includes 44.4% of the Italian mtDNA lineages. In particular, H1 turned out to represent a large proportion of H samples, encompassing the 13.8% of the total variability (10.4% excluding sub-lineages). Compared to H1, sub-haplogroups H3 and H5 represent much smaller fractions of H composition, reaching however noteworthy frequencies (3.9% and 4.3% respectively). Most of the remaining samples belong to haplogroups frequently found in western Eurasia, including U5, K1, J1, J2, T1, T2, and HV. Among the U5 lineages, U5a is the most frequent (3.70%). Haplogroups K1a, HV and J1c take into account respectively the 4.39%, 4.05% and the 3.93% of the total mtDNA variability. The remaining lineages reach frequencies that do not exceed a 3.5% threshold.

Maternal population structure. In contrast to paternal lineages, correlation between geographical and genetic distances was non-significant (Mantel Test: observed value = 0.011, p-value = 0.45). These results point to a strong homogeneity within the Italian Peninsula for the mtDNA gene pool composition. In order to extract further insights into the distribution of mtDNA lineages, a sPCA was performed using haplogroup frequencies. The highest absolute eigenvalues (Figure S8) correspond to the first two positive components (global structure). According to the Global test of significance, the geographical distribution of the genetic variability observed with sPCA was found to be marginally significant (observed value = 0.061, p-value = 0.046).

Scores of the sPC1 and sPC2 are plotted in Figure 3. Both sPC1 and sPC2 highlight the extreme position of Sardinia (large white squares). In addition, sPC1 identifies a North-East centred group that spreads southwards along the Apennines (including most of populations from central Italy), while sPC2 highlights the same East-West pattern observed for Y-chromosome. Loadings of sPC1 and sPC2 (Figure S9) identify lineages H1 and H3 respectively as the haplogroups affecting more the spatial genetic differentiation of Italian populations.

Haplogroup DAPC analysis. DAPC was performed within the eight most frequent haplogroups (H*, H1, H3, H5, HV, J1c, K1a, U5a). They disaggregate in 24 haplotype clusters (Table S8,

Table 1. Frequencies of Y-Chromosome DAPC cluster for each Italian sPCA-identified group.

HG	DAPC CLUSTER	N. HAPLOTYPES				N. INDIVIDUALS				MAX% (GROUP)
		NWI	SEI	SAR	TOT	NWI	SEI	SAR	TOT	
E-V13	1	8	10	1	19	8	10	1	19	53% (SEI)
	2	6	6	0	12	6	6	0	12	50% (NWI, SEI)
	3	3	11	1	15	3	**11**	1	15	**73% (SEI)**
	4	5	6	0	11	5	6	0	11	55% (SEI)
	5	6	6	0	12	6	6	0	12	50% (NWI, SEI)
G2a-P15	1	18	6	2	26	**20**	6	2	28	**71% (NWI)**
	2	1	3	0	4	1	3	0	4	75% (SEI)*
	3	8	10	8	26	8	10	8	26	38% (SEI)
	4	2	20	0	22	2	**20**	0	22	**91% (SEI)**
	5	1	15	1	17	1	**16**	1	18	**89% (SEI)**
I2a-M26	**1**	0	1	18	19	0	1	**19**	20	**95% (SAR)**
	2	2	1	12	15	2	1	**13**	16	**81% (SAR)**
J2a-M410	1	7	9	3	19	7	9	3	19	47% (SEI)
	2	8	18	2	28	8	19	2	29	66% (SEI)
	3	7	11	0	18	7	12	0	19	63% (SEI)
R-P312	**1**	11	4	1	16	**12**	4	1	17	**71% (NWI)**
	2	13	8	0	21	13	9	0	22	59% (NWI)
	3	6	5	0	11	6	5	0	11	55% (NWI)
R-U152	1	16	7	2	25	16	7	2	25	64% (NWI)
	2	21	1	0	22	**21**	1	0	22	**95% (NWI)**
	3	23	8	2	33	24	10	2	36	67% (NWI)
	4	16	4	2	22	**17**	5	2	24	**71% (NWI)**
R-L2	**1**	18	1	1	20	**18**	1	1	20	**90% (NWI)**
	2	18	6	1	25	**18**	6	1	25	**72% (NWI)**
	3	10	4	0	14	**10**	4	0	14	**71% (NWI)**

*Number of individuals <10

The absolute number of haplotypes and individuals are shown for each DAPC-cluster, and the maximum frequency for each cluster is expressed in percentage (max%). NWI: North-Western Italy; SEI: Southern and Eastern Italy; SAR: Sardinia.

doi:10.1371/journal.pone.0065441.t001

Figure S10), ranging from a minimum of two (K1a) to a maximum of four (U5a). Most of them are widespread in the whole of Italy, in fact, if we consider a 70% threshold, only nine clusters show traces of geography-related distributions (but six of them are composed by less than 10 individuals). Haplogroup HV is the most important exception, including two clusters located in NWI and SEI, respectively. It is noteworthy a cluster from haplogroup H3 that is almost exclusive of SAR.

Comparisons with other European samples (Table S9, Figure S11) confirm that great part of Italian mtDNA haplotypes share a wide range of affinities spanning from Iberia to Eastern Europe, but haplotypes from H1 and H3 appear to be related mostly with Western and Central Europe.

Date estimates for maternal variation. TMRCA estimates for the most frequent haplogroups (Table 2) could be classified in two groups: "old" haplogroups, predating the Last Glacial Maximum, LGM (~31,600 YBP for HV, ~28,300 YBP for U5a and ~19,500 YBP for J1c), and haplogroups dating after the LGM (~16,200 YBP for H*, ~15,600 YBP for H1, ~15,500 YBP for H3, ~14,700 YBP for H5, ~16,700 YBP for K1a). Estimates for H1 and H3 haplogroups are slightly older than estimates in Western Eurasia for the same haplogroups [2], [4], [5], [55]. These results are in agreement with what has been shown for the Basque region in Iberia [27] and may be related to the length of the mitochondrial region used.

Additionally, we calculated TMRCA for the two DAPC clusters within HV haplogroup (2 and 3), given that they show a clear spatial polarity within continental Italy and Sicily. Their ages fall between the time estimate for the whole haplogroup (~31,600 YBP) and the LGM, suggesting that their differentiation happened during this time frame (Table 2).

Discussion

Previous reconstructions of the genetic structure of Italy agreed on two points: the peculiarity of the population of Sardinia – due to a distinct background and a high degree of isolation [58], [59] – and the clinal pattern of variation in the Italian Peninsula, which has been explained by differential migration patterns [17], [18] although some genetic discontinuity due to local drift and founder effects have been described [16], [19], [20]. This study represents a significant upgrade on the knowledge of the genetic structure of Italy for the following reasons: the wide sampling coverage (coupled to a detailed sampling strategy), the high number of typed

Table 2. Age estimates (in YBP) of STR and HVS variation for the most common haplogroups in the Italian data set.

Y Chromosome Haplogroups	**SD**	**SE**	**Age estimate**	**SE**
E-V13	146.46	51.78	3662	1295
Cluster3 (SEI 70.3%)	139.52	49.33	3488	1233
G-P15	600.79	212.41	15020	5310
Cluster1 (NWI 71.4%)	144.31	51.02	3608	1276
Cluster3	505.72	178.80	12643	4470
Cluster4 (SEI 90.9%)	111.40	39.39	2785	985
Cluster5 (SEI 88.9%)	240.62	85.07	6016	2127
I-M26	206.11	72.87	5153	1822
Cluster 1 (SAR 95.0%)	48.26	17.06	1207	427
Cluster 2 (SAR 81.3%)	227.81	80.54	5695	2014
R-U152	137.29	48.54	3432	1214
Cluster2 (NWI 95.5%)	199.16	70.41	4979	1760
Cluster4 (NWI 70.8%)	184.29	65.16	4607	1629
R-L2	129.67	45.85	3242	1146
Cluster1 (NWI 90.0%)	250.32	88.50	6258	2213
Cluster2 (NWI 72.0%)	185.52	65.59	4638	1640
Cluster3 (NWI 71.4%)	148.55	52.52	3714	1313
R-P312	302.55	106.97	7564	2674
Cluster1 (NWI 70.6%)	130.05	45.98	3251	1149
mtDNA Haplogroups	**Rho**	**SE**	**Age estimate**	**SE**
H*	1.79	0.16	16229	2889
H1_whole (including all H1 derivates)	1.72	0.15	15604	2588
H1*	1.43	0.14	12983	2549
H3	1.71	0.28	15452	4954
H5	1.62	0.23	14689	4015
HV	3.49	0.33	31574	5872
Cluster 2 (NWI 75%)	2.00	0.42	18116	7476
Cluster 3 (SEI 85%)	2.33	0.39	21135	7002
U5a	3.13	0.35	28306	6128
K1a	1.84	0.25	16686	4383
Cluster 2 (NWI 71%)	1.33	0.28	12077	4929
J1c	2.15	0.27	19448	4757

Standard deviation (SD) estimator (Sengupta et al. 2006) and ñ statistic calculator (Soares et al. 2009) were used for Y-chromosome and mtDNA haplogroups respectively. Ages were estimated for the entire haplogroups as well as for each DAPC cluster with at least 10 individuals and frequencies >70% in NWI, SEI, or SAR (excepted for G-P15, cluster 2, see Methods).

doi:10.1371/journal.pone.0065441.t002

markers and the innovative methodological approach. Our results show that the Y-chromosomal genetic diversity of Italy is not clinal but structured in three geographical areas: North-Western Italy (NWI), South-Eastern Italy (SEI) and Sardinia (SAR). The outlier position of SAR described in previous studies [21], [58–61] is mainly due to the high frequency of I-M26 haplogroup, that in turn is almost completely absent in continental Italy. In addition, it is noteworthy the scanty haplotype affinities with other European I-M26 lineages as DAPC results seem to indicate (Figure S7, Table S6). However, the structure observed for paternal lineages in continental Italy and Sicily was not characterised by North-South gradients as previously described: our results show a NWI-SEI clustering (Figure 1a), suggesting a shared genetic background between Southern Italy and the Adriatic coast from one side, and between Northern Italy and Tuscany from the other side. Actually, the most accurate description of the discontinuity between NWI and SEI is that of a "belt", that is a restricted portion of territory in which haplogroup frequencies tend to change more rapidly than in the rest of the Italian peninsula. This model was suggested by the presence of a few populations from North-Eastern and Central Italy (Treviso, Foligno/PG) that reveal an intermediate position between the two main groups.

The discontinuous Y-chromosomal structure of continental Italy is also confirmed by the distribution of DAPC haplotype clusters identified for the most frequent haplogroups (Table 1). Haplogroup G2a provides the most compelling case, being widespread in the whole region, but revealing different clusters in NWI and SEI (Figure 2). This is in agreement with a recent G haplogroup

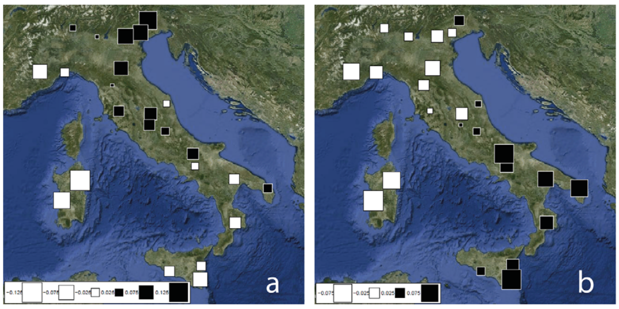


Figure 3. Spatial Principal Component Analysis (sPCA) based on frequencies of mtDNA haplogroups. The first two global components sPC1 (a) and sPC2 (b) are depicted. Positive values are represented by black squares; negative values are represented by white squares; the size of the square is proportional to the absolute value of sPC scores.
doi:10.1371/journal.pone.0065441.g003

survey that revealed the presence of different G2a sub-clades in Italy [62]. Nevertheless, we are not identifying the whole Italian population history with a single haplogroup. In fact, comparisons with other populations taking into account the whole haplogroup spectrum suggest differential patterns of haplotype similarity, implying different genetic histories for the identified sPCA-groups. In particular, NWI is mostly related with Western and Central Europe, while SEI seems to indicate more affinities with the Balkans. In addition, NWI and SEI are characterised by different distributions of genetic variance, the latter showing higher intra-population and lower (not significant) inter-population variability, while the opposite is true for NWI, where significant variation between populations was detected. On the whole, these patterns may be explained by a higher degree of population mobility in SEI, while in NWI local drift effects may have had a greater impact.

In contrast to the results obtained for Y-chromosome, the mtDNA diversity in Italy is characterised by a high degree of homogeneity: the only exception (a marginally significant sPCA global test based on haplogroup frequencies) is due to significant differentiation found in the Sardinian samples compared to continental Italy and Sicily (AMOVA difference between groups = 1.02%, $p<0.05$, Table S4). These results (in agreement with Y chromosome) suggest at least partially different demographic histories for SEI-NWI populations on one hand and SAR on the other hand, the latter being less affected to the gene flow of different migrations occurred in the Italian Peninsula and Sicily. Traces of such processes are visible in sPCA results (Figure 3) and in particular in sPC2, reflecting the same NWI-SEI pattern shown by Y-chromosomal sPC1. Anyway, such differentiation was not significant in the case of mtDNA (AMOVA difference between groups = 0.10%, $p = 0.08$). Analogously, DAPC clusters of mtDNA haplotypes do not show any geographic structure even when compared with other European samples, with clusters of similar haplotypes spanning from Iberia to the Balkans. However, not only uniparental differences in the genetic structure but also in time estimates are shown in the present dataset: our age estimates for the Y-chromosome and the mtDNA haplogroups (as well as the corresponding clusters of haplotypes) highlight significantly different time periods (Table 2), which could reflect multi-layered histories in Italy. Age estimates for mtDNA haplogroups - even if past demographic events affecting error rates cannot be excluded - point almost unanimously to pre-Neolithic times, ranging approximately from ~13,000 (H1*) to ~31,600 (HV) YBP. Although such estimates might reflect the haplogroups pre-existent diversity previous to their establishment in Italy (which could be the case of HV, that includes two DAPC clusters with different geographical distributions and whose ages largely post-date that of the whole haplogroup; Table 2), this does not seem to hold for most of the mtDNA haplogroups analysed. Indeed, most of our mtDNA time estimates are consistent with the hypothesis of the existence of a Glacial Refugium in the Italian Peninsula and its probable role in subsequent post-glacial expansions.

Actually, the role of Italy as a Southern European Glacial Refugium – together with the Iberian and Balkan peninsulas – is demonstrated for a high number of animal and plant species [63–69]. The presence of numerous Epigravettian sites suggests strongly that Italy could have acted as such also for humans [70]. Nevertheless, molecular evidences going in the same direction are still scarce, the only exception being mitochondrial haplogroup U5b3 [8], [9] whose frequency in Italy is relatively low (U5b lineages account for 1.73% in our data). Our results suggest that most of Italian mitochondrial diversity originated during and immediately after LGM. In particular, estimates for H1 and H3 are even older in Italy than in the Franco-Cantabrian area [27] where these clades have been postulated to originate [4]. Furthermore, DAPC comparisons with a wide set of European haplotypes (Table S9) show that Italy, in most cases, is characterised by the highest number of different haplotypes. On the whole, these observations not only are in agreement with the existence of a human Glacial Refugium in Italy, but also suggest that its relevance has been until now largely underrated.

The use of STR variation for dating Y-chromosome lineages or population splits, is a controversial issue, due to the effect that both mutation rates and STR choice has on the temporal scale of age estimates. Following the most recent studies our estimates are based on those STRs that show the highest duration of linearity [49] and by using locus-specific mutation rates (Ballantyne et al. 2010). This is one of the reasons that led us to exclude 'evolutionary' mutation rates (see Methods for details). In addition, we removed 'outlier' haplotypes (see Methods S1), since their presence could inflate significantly the ages of haplogroups and DAPC clusters. However, these results have to be taken with great caution, keeping in mind that 'evolutionary' rates (applied to the same data) would yield time estimates around three times greater. Nonetheless, we observe that two independent methods applied to our data – BATWING and SD-based estimates – yield consistent results. In fact, in contrast to mtDNA age estimates, almost all Y-chromosome estimates fall between late Neolithic and the Bronze Age. This finding supports the hypothesis that group-specific clusters of haplotypes did originate after the split between NWI and SEI (dated with BATWING), even if the confidence interval for BATWING estimate is not tight enough to exclude alternative hypotheses. Interestingly, the NWI and SEI structure detected (Figure 1, Table S4) might be traced back around 5,500 YBP indicating relevant demographic events within continental Italy in this period. Anyway, this value has to be considered as a lower bound, given that the model used does not account for migration that would bias the split time towards recent dates. In fact, given a specific level of populations differentiation, the separation time estimated between these populations has necessarily to be higher (i.e. more ancient) as migration is considered.

According to the most recent syntheses, the Neolithic revolution diffused in Italy following two independent routes along the Adriatic (Eastern) and the Tyrrhenian (Western) coasts. Furthermore, archaeological sites from NWI are characterized by a deeper continuity with earlier Mesolithic cultures and a higher degree of local variability than SEI, while this last area, besides being culturally more homogeneous, shows clear links with the Southern Balkans [10]. Our Y-chromosome results – showing discontinuity between NWI and SEI, higher inter-population variability in NWI, higher homogeneity in SEI coupled with relevant contributes from the Balkans – are quite consistent with this model. Thus, we can hypothesize that the NWI-SEI structure detected with paternal lineages could have its origins after these different Neolithic processes. Indeed, comparisons with other European and Near-Eastern populations (Table S6) suggest a stronger affinity between NWI with Iberia and Central Europe, while SEI is more related to the Balkans and Anatolia. The emergence of population structures during the Neolithic has been recently shown in two different studies using Y-chromosome markers, in Near East [71] and in Western Europe [27]. Our results confirm these findings and emphasize the role of demographic expansions and cultural advances related to the Neolithic revolution in shaping human genetic diversity, at least for male lineages. Nonetheless, such pattern might have been further influenced and/or re-shaped also by more recent events.

For instance, the dates of several DAPC clusters fall within the range of the Metal Ages (Table 2). During this long period (third and second millennia BC) Italy underwent important technological and social transformations finally leading to the ethnogenesis of the most important proto-historic Italic peoples. On the whole, our results indicate that these transformations, far from being exclusively cultural phenomena, actually involved relevant population events.

It is worth noting the older age estimate obtained for Y-haplogroup G2-P15 (15,020 YBP) that, coupled with its high frequency (11.09%), makes it the most probable candidate for a continuity with Italian Mesolithic populations (although a Neolithic origin for G2-P15 is discussed, [22], [23]). The most frequent G2-P15 cluster (12,643 YBP, Table 2), besides being evenly diffused in NWI and SEI, it encompasses almost all Sardinian G2-P15 individuals (Figure 2, Table 1). These facts, together with the higher degree of isolation of Sardinia to Neolithic and Post-Neolithic migration processes, support the antiquity of this haplogroup in Italy. Despite obtaining similar time estimates for G2a in Italy (12,899 YBP), Rootsi et al. (2012) [62] explain the diffusion of its main sub-lineages in this country solely as a consequence of Neolithic and Post-Neolithic events.

Conclusions

This study depicts the most complete picture of Italian genetic variability from the point of view of uniparental markers to date. Our analyses revealed that the Y-chromosomal genetic structure of Italy is characterised by discontinuities. Such a structure is defined by three different and well-defined groups of populations: the Sardinia island (SAR), North-Western Italy (NWI) and South-Eastern Italy (SEI). Furthermore, we observed that NWI and SEI are not separated according to latitude but following a longitudinal line. Such discontinuity may date at the Neolithic revolution in Italy, which was characterised by (at least) two independent diffusion processes involving the Western and Eastern coasts, respectively. Mitochondrial DNA, despite showing some correspondence with Y-chromosome results, depicts a substantially homogeneous genetic landscape for the Italian peninsula. Significantly different ages were estimated for mtDNA and Y-chromosome systems. mtDNA variability dates back to Paleolithic and supports the existence of an Italian human Refugium during the last glacial maximum whereas Y-chromosome points to the importance that the demographic events happened during the Neolithic and the Metal Ages had in the male Italian patterns of diversity and distribution.

Supporting Information

Figure S1 Map showing the geographical location of populations sampled in the present study. Colors indicate the eight clusters of homogeneous Italian provinces (sampling macro-areas) identified after a preliminary surname-based analysis [24]. The set of provinces (sampling points) and the number of samples successfully typed for Y-chromosome and mtDNA markers are detailed for each sampling macro-area (table on the left).
(TIF)

Figure S2 Plot of geographical distances against genetic distances (based on frequencies of Y-chromosome haplogroups). A 2-dimensional kernel density estimation layer (Venables and Ripley 2002) was added to the plot. The analysis was performed including (a) and excluding (b) the Sardinian samples.
(TIF)

Figure S3 Eigenvalues of Y-chromosome-based sPCA analysis (A) with their decomposition in spatial and variance components (B). Eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index). They are both positive and negative depending from Moran's I positive or negative values. Large positive components correspond to global structures (cline-like structures); large negative components correspond to local structures (marked genetic differentiation among neighbours).
(TIF)

Figure S4 Loadings of the most informative components (a: sPC1, b: sPC2). These values identify Y-chromosome haplogroups that mostly affect the genetic structure of Italian populations.
(TIF)

Figure S5 DAPC analysis of STRs variation for the most frequent Italian Y-chromosome haplogroups (E-V13, I-M26, J-M410, R-P312*, R-U152*, R-L2). Samples are grouped according to their affiliation to sPCA-identified areas (NWI, SEI, SAR; symbols in the top right legend of each plot). For each plot, the number of different haplotypes per cluster and their geographic distribution in the above areas are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot. Haplogroup I-M26, including two clusters only, is represented by a single discriminant function (no eigenvalues barplot).
(TIF)

Figure S6 Median joining network for Italian G2a-P15 haplotypes. Individuals have been assigned and colored according to the correspondent DAPC-based clusters (Figure 2).
(TIF)

Figure S7 DAPC analysis of STRs variation for the most frequent Y-chromosome haplogroups. Results are based on Italian data and additional comparison samples (NWI; SEI; SAR; IBE: Iberian Peninsula; BAL: Balkan Peninsula; GER: Central-Europe (Germany); CAU: Caucasus; WAN: Western Anatolia; symbols in the legend of each plot). For each plot, the number of different haplotypes per cluster and their geographical distribution are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot.
(TIF)

Figure S8 Eigenvalues of mtDNA-based sPCA analysis (A) with their decomposition in spatial and variance components (B). Eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index), and are both positive and negative, depending from Moran's I positive or negative values. Large positive components correspond to global structures; large negative components correspond to local structures (marked genetic differentiation among neighbours).
(TIF)

Figure S9 Loadings of the most informative components (a: sPC1, b: sPC2). These values identify mtDNA haplogroups that mostly influence the genetic structure of Italian populations.
(TIF)

Figure S10 DAPC analysis of HVS variation for the most frequent mtDNA haplogroups (H*, H1, H3, H5, HV, J1c, K1a, U5a) in the Italian data set. Results have been grouped geographically using the same categories as for Y-Chromosome (NWI; SEI; SAR); "0" codes were attributed to those populations for which Y-chromosome information was not available and whose geographical position lies along the boundary between NWI and SEI (Aviano, Terni). For each plot, the number of different haplotypes per cluster and their geographical distribution are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot. Haplogroup K1a, including two clusters only, is represented by a single discriminant function (no eigenvalues barplot).
(TIF)

Figure S11 DAPC analysis of HVS variation for the most frequent mtDNAhaplogroups. Results are based on Italian data and comparison European populations (ITA: Continental Italy; SAR: Sardinia; BASQ: Iberian Peninsula (Basques); AUST: Central Europe (Austria); MAC: Macedonians; ROM: Romanians; BALK: Balkan Peninsula; symbols in the legend of each plot). For each plot, the number of different haplotypes per cluster and their geographical distribution are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot.
(TIF)

Table S1 Frequencies of Y-chromosome haplogroups. Absolute values are reported for the whole Italian data set, while the frequencies within the eight sampling areas (from I to VIII) are expressed in percentage (%).
(XLS)

Table S2 Y-Chromosome STRs haplotypes in the 884 Italian samples of the present study.
(XLS)

Table S3 Diversity indices computed for the different Italian sampling points. Standard diversity parameters were calculated for both Y-chromosome and mtDNA based on haplotype/sequence data and haplogroup frequencies.
(XLS)

Table S4 Analyses of the molecular variance (AMOVA). Apportionment of the variance in %. Samples were grouped according to the geographic clusters (eight macro-areas) and to the sPCA results.
(XLS)

Table S5 DAPC membership probabilities to the SPCA-identified groups.
(XLS)

Table S6 Frequencies of Y-Chromosome DAPC clusters based on Italian data and comparison to other populations. The absolute number of haplotypes and individuals are shown for each population (NWI: sPCA North-Western Italy; SEI: sPCA Southern and Eastern Italy; SAR: Sardinia; IBE: Iberian Peninsula; BAL: Balkan Peninsula; GER: Central-Europe (Germany); CAU: Caucasus; WAN: Western Anatolia).
(XLS)

Table S7 Frequencies of mtDNA haplogroups. Absolute values are reported for the whole Italian data set, while the frequencies within the eight sampling areas (from I to VIII) are expressed in percentage (%).
(XLS)

Table S8 Frequencies of mtDNA DAPC clusters in Italy. Values were calculated both grouping according to the geographical clusters identified with Y-Chromosome sPCA (NWI: Y-sPCA North-Western Italy; SEI: Y-sPCA Southern and Eastern Italy; SAR: Sardinia) as well as considering the continental Italy (including Sicily) altogether (ITA). The absolute number of haplotypes and individuals are shown for each DAPC-cluster, and the maximum frequency for each cluster is expressed in percentage (max%).
(XLS)

Table S9 Frequencies of mtDNA DAPC clusters based on Italian data and comparison to other populations. The absolute number of haplotypes and individuals are shown for each population (ITA: Continental Italy and Sicily; SAR: Sardinia; BASQ: Iberia Peninsula (Basques); AUST: Central Europe (Austria); MAC: Macedonians; ROM: Romanians; BALK: Balkan Peninsula).
(XLS)

Methods S1 Spatial Principal Component Analysis (sPCA). Discriminant Analysis of Principal Components. Batwing analysis. "Jackknife-like" procedure for outliers identification.
(DOC)

Acknowledgments

We are indebted to the Personnel of the Italian Blood Centers and Hospital Centers from: Agrigento, Alessandria, Ancona, Ascoli Piceno, Aviano, Barletta, Bassano del Grappa, Belluno, Benevento, Brescia, Campobasso, Castrovillari, Catania, Catanzaro, Chianciano Terme, Chioggia, Chiusi, Como, Crotone, Cuneo, Enna, Foligno, Genova, Imperia, L'Aquila, La Spezia, Lecce, Macerata, Matera, Montalcino, Novara, Olbia, Oristano, Padova, Pesaro, Pescara, Pistoia, Policoro, Potenza, Ragusa, Sarteano, Sassari, Savona, Teramo, Terni, Trapani, Treviso, Trieste, Varese, Vercelli, Vicenza, Villa d'Agri. Their participation to the Project has been of invaluable help to perform the sampling campaign. We thank Dr. Tosca Corti, Dr. Luisa Stella Dolci and Mr. Carlo Pascucci for helping us in sample collection. We thank all the volunteers who kindly accepted to participate to the study. We are very grateful to CESGA (Centro de Supercomputación de Galicia), where BATWING computational analyses were performed.

The Genographic Consortium includes: Syama Adhikarla[1], Christina J. Adler[2], Elena Balanovska[3], Oleg Balanovsky[3], Jaume Bertranpetit[4], Andrew C. Clarke[5], Alan Cooper[2], Clio S. I. Der Sarkissian[2], Matthew C. Dulik[6], Jill B. Gaieski[6], ArunKumar GaneshPrasad[1], Wolfgang Haak[2], Marc Haber[4,7], Li Jin[8], Matthew E. Kaplan[9], Hui Li[8], Shilin Li[8], Elizabeth A. Matisoo-Smith[5], Nirav C. Merchant[9], R. John Mitchell[10], Amanda C. Owings[6], Laxmi Parida[11], Ramasamy Pitchappan[1], Daniel E. Platt[11], Colin Renfrew[12], Daniela R. Lacerda[13], Ajay K. Royyuru[11], Fabrício R. Santos[13], Theodore G. Schurr[6], Himla Soodyall[14], David F. Soria Hernanz[15], Pandikumar Swamikrishnan[16], Chris Tyler-Smith[17], Arun Varatharajan Santhakumari[1], Pedro Paulo Vieira[18], Miguel G. Vilar[6], R. Spencer Wells[15], Pierre A. Zalloua[7], Janet S. Ziegle[19].

Affiliations for participants: [1]Madurai Kamaraj University, Madurai, Tamil Nadu, India; [2]University of Adelaide, South Australia, Australia; [3]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia; [4]Universitat Pompeu Fabra, Barcelona, Spain; [5]University of Otago, Dunedin, New Zealand; [6]University of Pennsylvania, Philadelphia, PA, USA; [7]Lebanese American University, Chouran, Beirut, Lebanon; [8]Fudan University, Shanghai, China; [9]University of Arizona, Tucson, AZ, USA; [10]La Trobe University, Melbourne, Victoria, Australia; [11]IBM, Yorktown Heights, NY, USA; [12]University of Cambridge, Cambridge, UK; [13]Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [14]National Health Laboratory Service, Johannesburg, South Africa; [15]National Geographic Society, Washington, DC, USA; [16]IBM, Somers, NY, USA; [17]The Wellcome Trust Sanger Institute, Hinxton, UK; [18]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil; [19] Vitapath Genetics, Foster City, CA, USA.

Author Contributions

Conceived and designed the experiments: DP DL DC. Performed the experiments: BMC AU DYY SS GC CH JM LQM PS. Analyzed the data: AB SS BMC AU. Contributed reagents/materials/analysis tools: DL DP DC LQM. Wrote the paper: AB BMC SS DC AU. Performed field work, sampling design and collection: AB DYY AU DL DP.

References

1. Cunliffe B (2001) The Oxford Illustrated History of PreHistoric Europe. Oxford: Oxford University Press. 544.
2. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910–918.
3. Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, et al. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. Am J Hum Genet 75: 128–137.
4. Pereira L, Richards M, Goios A, Alonso A, Albarrán C, et al. (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res 15: 19–24.
5. Soares P, Achilli A, Semino O, Davies W, Macaulays V, et al. (2010) The Archaeogenetics of Europe Curr Biol 20: 174–183.
6. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276.
7. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, et al. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science. 290: 1155–1559.
8. Pala M, Achilli A, Olivieri A, Hooshiar Kashani B, Perego UA, et al. (2009) Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 84: 814–821.
9. Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, et al. (2012) Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. Am J Hum Genet 90: 915–924.
10. Pessina A, Tinè V (2008) Archeologia del Neolitico. L'Italia tra il Vi e il IV millennio a.C. Roma: Carrocci editore. 375.
11. Pesando F (2005) L'Italia antica. Culture e forme del popolamento nel I millennio a. C. Roma: Carocci editore. 326.
12. Barbujani G, Bertorelle G, Capitani G, Scozzari R (1995) Geographical structuring in the mtDNA of Italians. Proc Natl Acad Sci U S A 92: 9171–9175.
13. Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton: Princeton University Press. 1088.
14. Turchi C, Buscemi L, Previderè C, Grignani P, Brandstätter A, et al. (2008) Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int J Legal Med 122: 199–204.
15. Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, et al. (2009) Human mitochondrial DNA variation in Southern Italy. Ann Hum Biol 36: 785–811.
16. Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, et al. (2003) Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol Phylogenet Evol 28: 387–395.
17. Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia' A, et al. (2007) Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phylogenet Evol 44: 228–239.
18. Brisighelli F, Alvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo A, et al. (2012) Uniparental Markers of Contemporary Italian Population Reveals Details on Its Pre-Roman Heritage. PLoS ONE 7: e50794.
19. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18: 1241–1248.
20. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, et al. (2009) Genetic structure of Europeans: a view from the North-East. PLoS One 4: e5472.
21. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, et al. (2012) An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data. PLoS One 7: e43759.
22. Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, et al. (2011a) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci U S A 108: 9788–9791.
23. Lacan M, Keyser C, Ricaut FX, Brucato N, Tarrús J, et al. (2011b) Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc Natl Acad Sci U S A 108: 18255–18259.
24. Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F (2012) General method to unravel ancient population structures through surnames. Final validation on Italian data. Hum Biol 84: 235–270.
25. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16: 1215.
26. Martínez-Cruz B, Ziegle J, Sanz P, Sotelo G, Anglada R, et al. (2011) Multiplex single-nucleotide polymorphism typing of the human Y chromosome using TaqMan probes. Investig Genet 2: 13.
27. Martínez-Cruz B, Harmant C, Platt DE, Haak W, Manry J, et al. (2012) Evidence of Pre-Roman Tribal Genetic Structure in Basques from Uniparentally Inherited Markers. Mol Biol Evol 29: 2211–2222.
28. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830–838.
29. Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, et al. (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci Int 157:187–197.
30. Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, et al. (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3: e104.
31. Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, et al. (2010) Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PloS Biol 8: e1000536.
32. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2010) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32: 25–32.
33. Van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: 386–394.
34. Excoffier L, Laval G, Schneider S (2007) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1: 47–50.
35. Venables WN, Ripley BD (2002) Modern Applied Statistics with S. New York: Springer495.
36. Caramelli D, Vernesi C, Sanna S, Sampietro L, Lari M, et al. (2007) Genetic variation in prehistoric Sardinia. Hum Genet 122: 327–336.
37. Calò CM, Melis A, Vona G, Piras I (2008) Sardinian population (Italy): a genetic review. International Journal of Modern Anthropology 1: 39–64.
38. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
39. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403–1405.
40. Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101: 92–103.
41. R Development Core Team (2008) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.
42. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48.
43. Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, et al. (2011). Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol 28: 2905–2920.
44. King RJ, Di Cristofaro J, Kouvatsi A, Triantaphyllidis C, Scheidel W, et al. (2011) The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. BMC Evol Biol 11: 69.
45. Behar DM, Harmant C, Manry J, van Oven M, Haak W, et al. (2012) The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. Am J Hum Genet 90: 486–493.
46. Brandstätter A, Zimmermann B, Wagner J, Göbel T, Röck AW, et al. (2008) Timing and deciphering mitochondrial DNA macro-haplogroup R0 variability in Central Europe and Middle East. BMC Evol Biol 8: 191.
47. Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Drobnic K, Miścicka-Sliwka D (2003) Mitochondrial DNA variability in Bosnians and Slovenians. Ann Hum Genet 67: 412–425.

48. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, et al. (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet 87: 341–353.
49. Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, et al. (2011). The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279: 884–892.
50. Wilson I, Weale M, Balding D (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J Roy Stat Soc A 166:155–188.
51. Xue YL, Zejal T, Bao WD, Zhu S, Shu Q, et al. (2006) Male demography in East Asia: A north-south contrast in human population expansion times. Genetics 172: 2431–2439.
52. Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, et al. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74: 50–61.
53. Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, et al. (2006) Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian Pastoralists. Am J Hum Genet 78: 202–221.
54. King TE, Jobling MA (2009) Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26:1093–1102.
55. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am J Hum Genet 84: 740–759.
56. Cox MP (2008) Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. Hum Biol 80:335–357.
57. Babalini C, Martínez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, et al. (2005) The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. Eur J Hum Genet 13: 902–912.
58. Contu D, Morelli L, Santoni F, Foster JW, Francalacci P, Cucca F (2008) Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. PLoS One 3: e1430.
59. Pardo LM, Piras G, Asproni R, van der Gaag KJ, Gabbas A, et al. (2012) Dissecting the genetic make-up of North-East Sardinia using a large set of haploid and autosomal markers. Eur J Hum Genet 20: 956–964.
60. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38:556–560.
61. Chiò A, Borghero G, Pugliatti M, Ticca A, Calvo A, et al. (2011) Large proportion of amyotrophic lateral sclerosis cases in Sardinia due to a single founder mutation of the TARDBP gene. Arch Neurol 68:594–598.
62. Rootsi S, Myres NM, Lin AA, Järve M, King RJ, et al. (2012) Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. Eur J Hum Genet 20: 1275–1282.
63. Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol 7: 453–464.
64. Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, et al. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300: 1563–1565.
65. Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. Philos Trans Ser B 359: 183–195.
66. Randi E (2007) Phylogeography of South European Mammals. In: Weiss S, Ferrand N, editors. Phylogeography of Southern European Refugia. Amsterdam: Kluwer Academic Publishers. 101–126.
67. Grassi F, De Mattia F, Zecca G, Sala F, Labra M (2008) Historical isolation and Quaternary range expansion of divergent lineages in wild grapevine. Biological Journal of the Linnean Society 95: 611–619.
68. Grassi F, Minuto L, Casazza G, Labra M, Sala F (2009) Haplotype richness in refugial areas: phylogeographical structure of Saxifraga callosa. Journal of Plant Research 122: 377–387.
69. Zecca G, Casazza G, Labra M, Minuto L, Grassi F (2011) Allopatric divergence and secondary contacts in Euphorbia spinosa L: Influence of climate change on the split of the species. Organisms Diversity and Evolution 11: 357–372.
70. Banks WE, d'Errico F, Peterson AT, Vanhaeren M, Kageyama M, et al (2008) Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. Journal of Archaeological Science 35: 481–491.
71. Haber M, Platt DE, Ashrafian Bonab M, Youhanna SC, Soria-Hernanz DF, et al. (2012) Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. PLoS One 7: e34288.

Article 2

Sarno S, Boattini A, Carta M, Ferri G, Alù M, Yang Yao D, Ciani G, Pettener D, Luiselli D (2014) *An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of Sicily and Southern Italy*. PLoS One (*submitted*).

An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of Sicily and Southern Italy.

Short title.
Uniparental genetic structures in Sicily and Southern Italy

Stefania Sarno[1], Alessio Boattini[1], Marilisa Carta[1], Gianmarco Ferri[2], Milena Alù[2], Daniele Yang Yao[1], Graziella Ciani[1], Davide Pettener[1], Donata Luiselli[1]

[1] Laboratorio di Antropologia Molecolare, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, 40126 Bologna, Italy
[2] Dipartimento di Medicina Diagnostica, Clinica e di Sanità Pubblica, Università degli Studi di Modena e Reggio Emilia, 41124 Modena, Italy

Corresponding Author: Alessio Boattini (alessio.boattini2@unibo.it)

Author Contributions.
Authors contributions are listed in alphabetical order. Conceived and designed the experiments: DL DP. Performed field work, sampling design and collection: DL DP DYY. Performed the experiments: GC GF MA MC SS. Analyzed the data: AB MC SS. Contributed reagents/ materials/analysis tools: DL DP GF MA. Wrote the paper: AB SS.

Keywords
Sicily; Southern Italy; Mediterranean; Y-chromosome; mtDNA; migration; admixture; TMRCA.

ABSTRACT

Due to their strategic geographic location between three different continents, Sicily and Southern Italy have long represented a major Mediterranean crossroad where different peoples and cultures came together over time. However, this multi-layered history of migration pathways and cultural exchanges, has made the reconstruction of the genetic history and population structure of the area extremely controversial and widely debated. To address this debate, we surveyed the genetic variability of 326 accurately selected individuals from 8 different provinces of Sicily and Southern Italy, through a comprehensive evaluation of both Y-chromosome and mtDNA genomes. The main goal was to investigate the structuring of maternal and paternal genetic pools within Sicily and Southern Italy, and to examine their degrees of interaction with other Mediterranean populations.

Our findings show high levels of within-population variability, coupled with the lack of significant genetic sub-structures both within Sicily, as well as between Sicily and Southern Italy. When Sicilian and Southern Italian populations were contextualized within the Euro-Mediterranean genetic space, we observed different historical dynamics for maternal and paternal inheritances. Y-chromosome results highlight a significant genetic differentiation between the North-Western and South-Eastern part of the Mediterranean, the Italian Peninsula occupying an intermediate position therein. In particular, Sicily and Southern Italy reveal a shared paternal genetic background with the Balkan Peninsula and the time estimates of main Y-chromosome lineages signal paternal genetic traces of Neolithic and post-Neolithic migration events. On the contrary, despite showing some correspondence with its paternal counterpart, mtDNA reveals a substantially homogeneous genetic landscape, which may reflect older population events or different demographic dynamics between males and females. Overall, both uniparental genetic structures and TMRCA estimates confirm the role of Sicily and Southern Italy as an ancient Mediterranean melting pot for genes and cultures.

INTRODUCTION

Due to their central geographic location in the Mediterranean domain, Sicily and Southern Italy hosted various human groups in both prehistoric and historic times [1], acting as an important crossroad for different population movements involving Europe, North-Africa and the Levant.

The first unquestioned colonization of Sicily has been linked to the Palaeolithic, and in particular to Epigravettian human groups coming from the mainland and entering Sicily through the present-day Strait of Messina [2-3]. Human remains, referable to the Upper Palaeolithic, recently discovered in Southern Italy (Grotta of Paglicci, Puglia [4]) and Sicily (Grotta d'Oriente in the island of Favignana, [5]), have been attributed to the mtDNA haplogroup HV and tentatively interpreted as descendants of the early-Holocene hunter-gatherers of Sicily and Southern Italy, who occupied this area before (Gravettian) and after (Epigravettian) the Last Glacial Maximum [5]. The transition to agriculture with the Neolithic revolution, occurred in the South-Eastern heel of Italy between 6000-5700 years BCE, then moving west towards Southern Calabria and Eastern Sicily, where traces of the same material cultures (*imprinted ceramics stentinelliane*) have been dated roughly to 5800-5400 BCE [6]. However the Neolithic pottery (*imprinted ceramics prestentinelliane*) uncovered in western Sicily (Uzzo and Kronio) are coeval (6000-5750 BCE) with the earliest occurrence of Neolithic materials in the more South-Eastern heel of the Italian Peninsula, thus suggesting potentially parallel and culturally independent processes of colonization between the eastern and western parts of the island [6].

In addition to Upper-Palaeolithic and Neolithic material cultures, historical and archaeological data offer a detailed and reliable understanding of the more recent population influences on Sicily and Southern Italy. Among the well-documented historical events, at least four main migration processes could potentially have affected the current genetic variability of the area: i) the massive occupation of Greeks (giving rise to the "*Magna-Graecia*") started in the 8th century BC from the Southern Balkans; ii) the Phoenician and Carthaginian colonization of the western part of Sicily occurred since the first millennium BC from the Levant through North Africa; iii) the Roman and post-Roman (Germanic) invasions from continental Italy and Central-Western Europe between the 300 BC and 500 AD; and iv) the more recent Muslim and Norman conquests of Sicily and Southern Italy in 8th-9th and 11th-12th centuries AD respectively. If on one hand the Greek colonisation of the south-eastern regions vs. the Phoenician occupation of western Sicily could have caused internal east-west cultural differentiation, on the other hand the later conquests (such as Germanic, Islamic and Norman occupations) may have contributed to reshape at different levels the genetic landscape of one of the largest Mediterranean islands, albeit their relative impacts remain still questioned.

Such a deep and complex historical stratification made the reconstruction of the genetic history and population structure of the area open to debate. Previous investigations on the genetic structure of Sicily, based on both classical, autosomal and uniparental markers, have indeed shown contrasting results about the presence [7-8] or the absence [9] of an east-west geographically heterogeneous distribution of genetic variation within the island [8]. By contrast, a substantial homogeneity in genetic variation emerged from more recent mtDNA-based studies, that focused on specific regions of Southern Italy [10-11]. To the best of our knowledge, all previous studies that specifically addressed the reconstruction of the genetic structure and population history of Sicily and Southern Italy, have been mostly focused on only one of the two areas at a time, moreover considering the maternal (mtDNA) and the paternal (Y-chromosome) perspectives separately.

In this study we present an high-resolution analysis of the uniparental genetic variability of Sicily and Southern Italy, by using a new accurately selected set of samples and, for the first time, by jointly analysing both paternal and maternal genetic systems at the same time. More than 300 individuals from 8 different Sicilian and Southern Italian provinces have been deeply typed for 42 Y-SNPs and 17 Y-STRs, as well as for the HVS-I and HVS-II regions and 22 coding SNPs of mtDNA. These data have been used to compare and contrast Y-chromosome and mtDNA genetic patterns within Sicily and Southern Italy, and then to investigate their affinities within the overall Mediterranean genetic landscape by further comparing our data with those of reference populations selected from Central, Western and Southern Europe, as well as from North Africa and the Levant. In this way we particularly seek to address the following questions: i) Is the genetic diversity of Sicily structured along its east-west axis and how is it patterned compared to Southern Italy? ii) Are the observed genetic patterns stratified temporally or geographically in terms of more ancient or recent peopling events, and are there any differences between maternal and paternal perspectives? iii) How is the genetic variability of Sicily and Southern Italy related to the wider Euro-Mediterranean genetic space and what are the main contributes to the current genetic pool? Since Sicily and Southern Italy have long played an important key role in the history of demic and cultural transitions occurred in Southern Europe and the Mediterranean, the clarification of these points will be of great relevance for the understanding of the different population, cultural and linguistic dynamics occurred within the whole Mediterranean area.

MATERIALS AND METHOD

Ethics Statement

All donors provided a written informed consent to this study according to the ethical standards of the institutions involved. The Ethics Committee at the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna (Italy) approved all procedures.

Population sample

The genetic structure of Sicily and Southern Italy (SSI) was investigated by means of a high resolution analysis of 326 Y-chromosomes and 313 mtDNAs representing eight different SSI provinces (Figure S1). Five of these (Agrigento, Catania, Ragusa-Siracusa, Matera, Lecce) were previously published in Boattini et al. (2013) [12], whereas the remaining three (Trapani, Enna, Cosenza) were typed and analysed here for the first time. Individual samples were collected according to the standard ‘grandparents criterion’ (i.e. three generations of ancestry in the sampled province). In addition, a subsample of 129 Y-chromosomes has been selected on the basis of surnames, thanks to the availability of Italian-province-specific lists of founder surnames [13]. Due to their link with Y-chromosomes, the selection of males bearing surnames which unequivocally belong to specific places can be used to select autochthonous participants in regional population genetic studies and to obtain an “older” picture of Y-chromosomal diversity [14]. That way, we were able to simulate a putative Late-Middle-Ages sample, that is the period during which surnames spread in Italy, thus allowing to verify the effects of very recent admixture events on population genetic structure.

Blood samples (3-5cc) were processed to extract the whole genome DNA by using a Salting Out modified protocol [15].

Y-chromosome genotyping

PCR amplification of 17 Y-STR loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393,DYS385a/b, DYS437, DYS438, DYS439, DYS448, DYS456,DYS458, DYS635, and GATAH4) was carried out by using the AmpFlSTR Yfiler PCR Amplification Kit (Applied Biosystems, Foster City, CA) following the manufacturer’s recommendations [16] in a final volume of 5µl. The PCR reaction consisted of denaturation at 95°C for 11 min, followed by 30 denaturation cycles at 94°C for 1 min, annealing at 61°C for 1 min, extension at 72°C for 1 min, and a final extension at 60°C for 80 min. Products were sized on an ABI Prism 310 Genetic Analyzer by using the GeneScan 3.7 software (Applied Biosystems, Foster City, CA). As the Yfiler kit amplifies DYS385a/b simultaneously, avoiding the determination of each of the two alleles (a or b), these two loci were excluded from all the analyses performed. The DYS389b locus was obtained by subtracting DYS389I from DYS389II [17]. Basal haplogroups were assigned by typing the 7 SNPs (R-M173, J-M172 , I-M170, E-M35, K-M9, P-M45, F-M89) implemented in the MY1 Multiplex PCR by Onofri et al. (2006) [18]. Subsequently, we explored Y-chromosome genetic variability by further typing 35 Y-SNPs. 33 of them (E-M78, E-V12, E-V13, E-V22, G-P15, G-P16, G-M286, G-U8, G-U13, I-M253, I-M227, I-L22, I-P215, I-M26, I-M223, J-M410, J-L27, J-M67, J-M92, J-M12, R-M17, R-M343, R-M18, R-M269, R-L51/S167, R-L11/S127, R-S21/U106, R-S116/P312, R-SRY2627/ M167, R-S28/U152, R-M126, R-M160, R-L2/S139, R-L21/S145) were typed by using six haplogroup-specific multiplexes [19] aimed at deeply investigating the Y-markers downstream of all the major European clades (namely E1b1b1*, G*, I*, J2* and R1*). The SNP genotyping was carried out by means of PCR Multiplex amplification, followed by Minisequencing reaction based on dideoxy Single Base Extension (SBE), which was performed with the SNaPshot multiplex kit (Applied Biosystem). SBE products were analysed with capillary electrophoresis on an ABI Prism 310 Genetic Analyser. Two more SNPs (E-M81, E-M123) were finally tested with RFLP analysis, by using *HpyCH4IV* [20] and *DdeI* [21] enzymes respectively.

Mitochondrial DNA genotyping

MtDNA genetic markers were successfully typed for 313 out of the 326 total samples. Variation at the mtDNA HVS-I and HVS-II regions was investigated by sequencing a total of 750 base pairs (bp) encompassing nucleotide positions from 15975 to 155. Polymerase chain reaction (PCR) of the HVSI/II regions was carried out in a T-Gradient Thermocycler (Whatman Biometra, Gottingen, Germany) with the following amplification profile: initial denaturation 95 °C for 5 min, 35 cycles of 95 °C for 30 sec, 58 °C for 30 sec, 72 °C for 5 min and final extension at 72 °C for 15 min.

PCR products were purified by ExoSap-IT1 (USB Corporation, Cleveland, OH) and sequenced on an ABI Prism 3730 Genetic Analyzer by using a Big-Dye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) according to the manufacturer’s instructions. To reduce ambiguities in sequence determination the forward and reverse primers were used to sequence both strands of HVS-I and HVS-II regions. The CHROMAS 2.33 software was used to read the obtained electropherograms. Sequences were finally aligned to both the Revised Cambridge reference sequence - rCRS [22-23] and the new Reconstructed Sapiens Reference Sequence – RSRS [24] by using the DNA Alignment software 1.3.1.1 (http://www.fluxusengineering.com/align.htm).

MtDNA haplogroups were determined on the basis of diagnostic sites in the D-loop region following Phylotree mtDNA phylogeny (http://www.phylotree.org/) and confirmed with the analysis of 22 SNPs in the mtDNA-coding region by means of two PCR and one SNaPshot minisequencing reactions [25]. 17 SNPs (3010L, 3915H, 3992L, 4216L, 4336L, 4529L, 4580L, 4769H, 4793H, 6776H, 7028L, 10398L, 10400H, 10873H, 12308L, 12705L, 14766L) were those implemented in the multiplexes by Quintans et al. (2004) [26], whereas five further SNPs (3936H, 4310L, 4745L, 13708L, 13759L) were added in order to reach a finer resolution level of analysis in the mtDNA genotyping.

Statistical Analyses

Haplogroup frequencies were estimated by direct counting. Standard diversity parameters were calculated with Arlequin 3.5.1.2 [27]. The proportion of genetic variance due to differences within or between populations was hierarchically apportioned through the analysis of molecular variance (AMOVA) implemented in the Arlequin software.

In order to set the observed genetic patterns within the Mediterranean and Southern European genetic landscape, we compared our samples with additional populations extracted from the literature (Table S1). Comparison samples were selected for representing the following key areas: North-Central Italy, Iberian Peninsula, Central Europe, the Balkans, the Levant and North Africa. As for North-African groups, literature data come mainly from urban areas, which presumptively include both Arab and Berber elements. Within each of these areas, we sought for Y-chromosome and mtDNA data (preferably but not necessarily from the same populations) that showed an in-depth resolution level comparable to our data. Sub-haplogroups were concatenated when needed for comparison purposes reaching a common level of 21 paternal and 16 maternal lineages. The number of samples bearing mtDNA and Y-chromosome reduced haplogroups within each Mediterranean population was estimated by mere counting, and relative haplogroup frequencies were computed by using the R software [28].

The correlation between geographic distances and genetic distances (Reynolds distance) based on haplogroup frequencies, was evaluated by means of a Mantel test (10,000 replications). To investigate the distribution of genetic variability within the Mediterranean Basin, Principal Component Analysis (PCA) and Spatial Principal Component Analysis (sPCA) were performed on HGs frequencies, by using the R software package *adegenet* [29-30]. Contrary to classic PCA where eigenvalues are calculated by maximizing variance of the data, in sPCA eigenvalues are obtained by maximizing the product of variance and spatial autocorrelation (Moran's I index) [30]. To evaluate the consistency of the sPCA-detected geographical structures versus a random spatial distribution of genetic variability, the Global and Local random tests implemented in the *adegenet* package have been applied [29-30]. Subsequently, to further test the significance of the genetic clusters identified by sPCA, we performed a Discriminant Analysis of Principal Components (DAPC), by using the *adegenet* package [29-31]. The DAPC method is aimed at describing the diversity among pre-defined groups of observations, by maximizing the between-group variance and minimizing the within-group variance. Moreover, based on the retained discriminant functions, it provides group membership probabilities of each population, which can be interpreted in order to assess how clear-cut or admixed the detected clusters are [31].

Fisher exact tests were performed on haplogroup frequencies among Mediterranean population groups, in order to determine significantly over- or under-represented HGs in any of the geographic areas considered. These tests were first performed against a background of all the Mediterranean populations by using the reduced common level of HGs resolution, and then by comparing single haplogroup frequencies of Sicily and Southern Italy with those of each comparison Mediterranean group, this time exploiting the deepest HG level available for each pairwise comparison.

The age of haplogroups (TMRCA) was estimated for those lineages found to be significantly differentiated between pairs of Mediterranean population groups, as well as focusing on the most frequent haplogroups of our dataset, due to their peculiar relevance in the genetic composition of the studied area. As for Y-chromosome time estimates, the standard deviation (SD) estimator from Sengupta et al. (2006) [32] has been used and the 95% confidence intervals were calculated based on the standard error (SE). This method does not estimate the population split time, but the amount of time needed to evolve the observed STRs genetic variation within a given haplogroup. In order to minimize the biasing effect of STRs saturation through time, all Y-chromosome age estimates were calculated selecting the eight markers with the highest duration of linearity D with time [33] and corrected for the presence of outliers as in Boattini et al. (2013) [12]. As for mutation rates, we adopted locus-specific mutation rates for each of the eight considered loci as estimated by Ballantyne et al. (2010) [34]. TMRCA for the most frequent mtDNA haplogroups was estimated by means of the ρ (rho) statistic with the calculator proposed by Soares et al. (2009) for the HVS-I region [35]. Being the molecular date estimates with ρ statistic potentially affected by past demography [36], these dates should however be interpreted cautiously. In order to avoid sampling errors, time estimates were calculated only for those haplogroups with absolute frequencies of at least 10 individuals.

The maternal and paternal genetic relationships of Sicily and Southern Italy with the other Mediterranean populations, were further addressed and compared by means of admixture-like plots based on Fst (HVS-I) and Rst (STRs) genetic distances among Mediterranean groups. Population groups were first clustered by using a non-hierarchical algorithm based on Gaussian mixture models (*mclust* R package, [37-38]), and then the posterior membership probabilities (for each population group to belong at each identified cluster) were calculated by using DAPC method (*adegenet* R package, [29,31]) and graphically represented with barplots.

Finally, to formally assess on a large geographic scale, the impact of the various continental and within-continental contributions to the current Sicilian and Southern Italian (SSI) genetic variation, admixture analysis was carried out by using the mY estimator implemented in the software Admix 2.0 [39-40]. A special attention was paid to the selection of parental populations, due to its critical rule in obtaining appropriate estimate of admixture proportions [41-43]. By taking the historical and archaeological records into account, we considered the Balkans, the Levant and the North-Central Italy as putative source regions for migration processes (the latter being representative of the North-Western Mediterranean cluster identified in the Results). North Africa was excluded from the model given its negligible contribution to the current SSI genetic pool (see Results). A try-hybrid model of parental populations was therefore used to estimate the admixture rates: i) average haplogroup frequencies of North-Central Italy (SVGE, TV, BO and GRSN) for both Y-chromosome and mtDNA markers were taken as representative of the North-Central Italian parental population [NCI]; ii) data of Anatolian Greeks (PHO and SMY) and Northern Greece (NGRE) were taken as proxies for the Balkan parental population [BALK], respectively for Y-chromosome and mtDNA markers; iii) data from Lebanon (respectively LBEI, LBEK, LMOU, LNOR, LSOU for Y-chromosome and LEB for mtDNA markers) were finally taken for the Levantine parental population [LEV]. Additional information about the selected comparison populations are provided in Table S1. Finally, in order to promote reliable analysis and minimize sampling components of variance, subsets of 50 individuals were randomly selected for each putative parental group.

RESULTS

Y-Chromosome perspective

The 326 unrelated individuals from 8 different locations of SSI have been assigned to 33 different haplogroups whose frequencies, for both the whole dataset as well as for each of the 8 sampling points, are detailed in Table S2. Y-STR haplotypes for the 119 newly-typed individuals are provided in Table S3. Haplogroups G-P15 (12.3%), E-V13 and J-M410* (both 9.5%), together with R-M269* (7.4%) represent the most frequent lineages found in Sicily and Southern Italy (SSI). These are followed by five R1-sublineages (R-M17, R-L2, R-P312, R-U152, R-U106), whose frequencies range from 5.2% to 3.7%, and by J-M267 which embraces almost the 5% of total variability. All these paternal lineages reportedly originated in Europe or in the Near East, whereas much lower it seems to be the African paternal contribution, mainly represented by haplogroups belonging to HG-E sub-lineages (E-V12, 2.76%; E-V22, 2.15%; E-M81, 1.53%). Contrarily to what previously reported in literature [8], no differential distribution of Y-chromosome lineages has been found in our dataset. Fisher exact tests performed on HG frequencies between Southern Italy and Sicily (P-value: 0.4765), as well as between Eastern and Westerm Sicily (P-value: 0.2998), indeed do not reveal any significant differentiation. No significant percentage of variance among groups of populations (F_{CT}) has been detected by regional AMOVAs (Table S4). In the same way, when our

Sicilian populations were grouped with those of Di Gaetano et al. (2009) [8] following their East-West subdivision scheme and by using the same HG resolution level, both AMOVA (variation among groups 0.30% , P-value 0.091) and Fst index (P-value 0.094), failed to reveal any significant difference in Y-chromosome HGs composition, thus pointing out a substantial homogeneous pattern of genetic variation within the island.

Moreover, when the distribution of Y-chromosome lineages in the present-day Sicilian and Southern-Italian population has been compared with the one of the surname-based selected subset, no significant differentiation appeared (P-value: 0.9551).

High levels of within-population variability have been observed for all the 8 populations analysed, as well as for the whole dataset (Table S5), thus suggesting a high genetic heterogeneity at a micro-geographical level among the considered Sicilian and Southern-Italian populations, as confirmed also by the presence of 312 out of 326 unique STRs haplotypes. In addition, all shared haplotypes involve at most two individuals.

In order to more deeply explore the genetic relationships among Mediterranean groups, our samples were then compared with the 29 Euro-Mediterranean, Levantine and North-African populations extracted from the literature (Table S1), by using a common level of Y-HGs resolution. A significant positive correlation between geographical and paternal genetic distances has been observed (Mantel Test: observed value = 0.591, P-value < 0.001), but no clear-cut discontinuous genetic structure was found when plotting geographical distances against the genetic ones (data not shown). However, when this general pattern of Y-chromosome HG distribution has been more deeply investigated by means of a spatial Analysis of Principal Components (sPCA), a highly significant global structure appeared (Gtest: obs=0.146, P-value < 0.001), clearly differentiating the North-Western from the Central and South-Eastern Euro-Mediterranean genetic pools (Figure 1). More precisely, the first sPC (Figure 1a) separates the Iberian, Central-European and North-Western Italian populations on one hand (black squares), from the Balkans and the Levant on the other hand (white squares). Sicily and Southern Italy particularly reveal to be well set in the genetic context of the Central and South-Eastern Mediterranean group, the only exception being Catania (CT), which instead shows a stronger affinity to the North-Western cluster (Iberian Peninsula, Germany and Northern Italy). A significant positive correlation was found between sPC1 scores and the corresponding longitudinal coordinates (R2 = 0.663, P-value < 0.001), the correlation with latitudes instead being R2 = 0.440 (P-value < 0.001).These facts confirm the observed North-West vs. Central/South-East pattern of HGs distribution within the Mediterranean domain.

Interestingly, the second sPC (Figure 1b), despite being much less representative compared to the first one in terms of both variance and spatial autocorrelation, identifies a subdivision between the two Mediterranean coastlines, which seems to involve the Eastern and Western parts of Sicily. The first group (black squares) is indeed represented by populations from the South-Eastern Mediterranean shore (Levant and North-Africa), including also the most western Sicilian provinces (Trapani and Agrigento) and the Iberian populations. Conversely, the second cluster (white squares) is mainly a North-Eastern Mediterranean centred group, encompassing the Balkans, South-Italy and East-Sicily, together with the other central European populations. When the reliability of the sPCA-identified structures was tested by means of AMOVA based on haplogroup frequencies, the proportion of genetic variation between groups (F_{CT}) results however two times higher when grouping according to the sPC1 (8.31 %, P-value < 0.001) than sPC2 (4.31%, P-value = 0.004). The sPCA-suggested pattern of genetic relationships among the different Mediterranean populations, has been confirmed in the classical PCA plots reported in Figure S2a.

The two high-structured Mediterranean clusters identified with sPC1, were further tested by means of DAPC analysis. Membership probabilities, represented with a structure-like plot (Figure 2), highlight the intermediate position of the Italian samples between the two Mediterranean clusters. In this context, Sicily and Southern Italy show clearly their stronger affinity with the populations from the South-Eastern Mediterranean side (with the partial exception of Catania - CT).

Fisher exact tests were carried out among groups of populations in order to identify significantly over- or under-represented HGs in any of the geographic areas analysed, against a background of all the other Mediterranean populations (Table S6). Haplogroup G-M201 appears significantly over-represented in the SSI genetic pool. Haplogroup R-M269, has been found significantly over-represented in Western-Mediterranean populations (IBE, GER and NCI), and under-represented in the South-Eastern Mediterranean ones (BALK, LEV and NAFR). By contrast, haplogroup J-M304(xM172) is significantly over-represented in the non-European Mediterranean shore (LEV and NAFR), being instead under-represented in European Mediterranean populations. In order to investigate further, we then performed a set of Bonferroni-corrected Chi-square tests by comparing frequencies of single lineages in SSI with those of each reference Mediterranean population group, this time exploiting the highest Y-SNP level of resolution available for each pairwise populations comparison (and considering only those lineages with absolute frequency of at least 10 individuals in SSI). Being aware that migration processes cannot be linked only with single specific haplogroups, it is however known that signals of migration should be more easily detected in more highly differentiated lineages [44]. Different haplogroups have shown significantly higher frequency in specific comparison groups than in SSI: R1b-sublineages in the western European samples (R-U152 for North-Central Italy, P-value < 0.001; R-P312 for Iberian Peninsula, P-value < 0.001; and R-U106 for German region, P-value < 0.001), R-M17 in the Balkan Peninsula and

Germany (both P-values < 0.05), and J1-M267 in both Levant and North-Africa (both P-values < 0.001).

As for TMRCA estimates, STR variation within the most frequent haplogroups of SSI suggests that most of them (with the exception of haplogroup G2a-P15: 9339 ± 3302 YBP) date back to relatively recent times (Table 1), in some cases falling into time periods compatible with specific documented historical events occurred in SSI. Despite the fact that these time estimates must be taken with caution, as they might be affected by the choice of both STRs markers and their mutation rates, overall our results agree in suggesting that most of the Y-chromosomal diversity in modern day Southern Italians originated during late Neolithic and Post-Neolithic times (~2,300 YBP for E-V13; from ~3,200 to ~3,700 YBP for J sub-lineages; ~4,300 YBP for R-M17 and R-P312; and ~ 2,000 YBP for R-U106 and R-U152).

Mitochondrial DNA perspective

The maternal genetic ancestry of SSI population was explored by successfully typing both coding region SNPs and HVSI-HVSII sequences in 313 out of the 326 samples. Overall, the polymorphic sites observed in the D-loop and coding region allowed assignment of subjects to 40 mtDNA HGs (including sub-lineages), whose frequencies for both the whole dataset as well as for each of the 8 sampling points are reported in Table S2. In order to ensure the easiest access to the data [45], mtDNA sequences were deposited in the GenBank nucleotide database, under accession numbers KJ522492-KJ522611.

The observed mtDNA HGs distribution reflects the typical maternal variability pattern documented for Mediterranean Europe. In fact, most of the individuals belong to super-haplogroup H, that on the whole accounts for the 38% of the total mtDNA lineages detected in our dataset. Within H, H1 represents the most frequent sub-lineage (10.9%), followed by H5 (3.2%) and H3 (2.6%). Noteworthy is also haplogroup HV, that has been found at relatively high frequencies (4.8%). Most of the remaining samples belong to haplogroups U5, K1, J1, J2, T1, T2, thus confirming prevalent European and Middle-Eastern genetic ancestries. MtDNA haplotypes of African origin are instead represented by few haplogroups at low frequencies, namely M1 (1.3%), U6a (0.6%) and L3 (0.6%).

Within-population diversity indices reveal that, in the context of our dataset, Sicily (and particularly Western Sicily) shows slightly lower diversity values than Southern Italy (Table S5). Nevertheless, the diversity parameters observed for all the 8 populations analysed as well as for the whole dataset, fall within the range of values commonly reported in literature for both Italian and Southern European populations [11]. Similarly to Y-chromosome, mtDNA does not reveal any kind of population sub-structure both within Sicily (East vs. West Sicily) as well as between Sicily and Southern Italy, neither considering haplogroups nor haplotypes (sequences). AMOVA results show low and non-significant F_{CT} values when population samples were grouped according to geography (Table S4). Analogously, Fisher exact tests reveal no significantly different HG composition in any of the geographic regions considered (South Italy vs, Sicily, P-value: 0.5019; East Sicily vs. West Sicily, P-value: 0.0698). In the same way, both AMOVA (variation among groups 0.52% , P-value 0.082) and Fst (P-value 0.076) based on HG frequencies show the absence of significant genetic differentiation along the east-west axis of Sicily.

The mtDNA HGs geographic distribution within the Mediterranean domain was investigated by comparing our sample with 26 Euro-Mediterranean, Levantine and North-African populations selected from the literature (Table S1). A Mantel test shows a low correlation between geographic and genetic distances (observed value = 0.279, P-value = 0.016). In order to further explore the relationships between geography and mtDNA genetic variability, we performed a sPCA (using HG frequencies). The highest eigenvalue obtained is the most positive one (sPC1) associated with the presence of a global structure. As previously emerged for Y-chromosome, sPC1 plot reveals a North-West/South-East (NW-SE) distribution of mtDNA genetic variation (Figure 3a). Nearly all of the Mediterranean populations (with some exceptions, i.e. AG, TV, BUR) appear indeed distributed along a longitudinal transect running from North African and Near Eastern countries (large white squares) to the Iberian Peninsula (large black squares), with the bulk of the South-Eastern European populations (including Balkans and Italy) roughly occupying an intermediate position therein (see also Figure S2b). Among them, Sicily and Southern-Italy appear linked to the South-Eastern Mediterranean coast. When the reliability of this sPC1-identified structure has been tested by means of AMOVA, the proportion of genetic variation between groups (F_{CT}) results lower than in the case of Y-chromosome (2.45%) but still significant (P-value < 0.001).

The second sPC (Figure 3b) highlights the position of Italy within the Mediterranean context and particularly of its South-Eastern part (large white squares). However, when tested with AMOVA, the proportion of variation between groups (F_{CT}) explained by sPC2 revealed to be not significant (0.48%, P-value = 0.212). On the whole, the lack of statistical support for the global structure observed in the mtDNA sPCA (Gtest: obs=0.165, P-value = 0.065), suggests a higher homogeneity in Mediterranean genetic variability for maternal than paternal genetic pools. Nevertheless, both uniparental markers show a similar NW-SE distribution pattern of genetic variation.

Fisher exact tests were applied to determine if differences in HG frequencies among population groups were statistically significant (Table S6). As expected, haplogroup H is found to be over-represented in Euro-Mediterranean populations and under-represented in North-African ones, while the opposite has been observed for haplogroup L. Haplogroup K is over-represented in Levantine populations, and haplogroup M in North-Africa. However, when the deepest level of HG resolution has been exploited for single pairwise comparisons between SSI and Mediterranean reference populations, we do not found any HG whose frequency is significantly higher than in our dataset. The only exception is a slightly significant (P-value: 0.045) over-representation of H1 haplotypes in the Iberian Peninsula.

Differently from Y-chromosome results, TMRCA estimates for the most frequent mtDNA haplogroups of Sicily and Southern Italy (Table 1) date back to pre-Neolithic times and could be mainly classified in lineages pre-dating the Last Glacial Maximum - LGM (~32,200 YBP for HV; ~31,100 YBP for J2; ~28,900 and ~28,600 YBP for T1 and T2; ~27,300 for U5; and ~25,000 YBP for J1) or dating immediately after it (~16,700 YBP for H5 and ~15,700 YBP for H1).

Comparative analysis of maternal and paternal genetic pools

The admixture-like plot represented in Figure 4 summarizes the genetic relationships between SSI and the chosen Mediterranean populations by directly comparing Y-chromosome and mtDNA genetic results.

From a Y-chromosome point of view, SSI form a fairly coherent group with the Levantine and the Balkan populations (cluster 2), despite showing some minor contribution (black component) also from the North-Western Mediterranean group (cluster 3). From a mtDNA point of view, our results show the differentiation between European and non-European Mediterranean populations, with North Africa and the Levant clustering in separate and different groups (1 and 2). However – and differently from the other European populations – SSI shows a noteworthy contribution (grey component) from the Levantine cluster. Both genetic systems reveal a negligible contribution from North Africa (white component).

The extent of different contributions to the current SSI genetic variation was further assessed by means of an admixture analysis performed (on HG-frequencies) with the coalescent-based mY estimator implemented in the software Admix 2.0 [39-40]. We used a tri-hybrid admixture model, considering as source populations North-Western Italy, the Balkans and the Levant (see Materials and Methods for more details). While keeping in mind that selection of parental populations can potentially misrepresent the real estimate of admixture proportions [41-43], our admixture rates (Figure S3) are however quite consistent with the above-mentioned results (despite the high standard errors values). Y-chromosome admixture proportions to the current SSI genetic pool indeed confirm an high paternal contribution from the South-Eastern Mediterranean populations, and particularly from the Balkan Peninsula (~60%), whereas about 25% of SSI Y-chromosomes can be traced back to North-Western European group. Analogously, although the present-day SSI mtDNA genetic pool is largely shared with the other South-Eastern European populations of the

Mediterranean Basin (respectively Balkan and Italian Peninsulas), a remarkable proportion of maternal ancestry (especially if compared with its paternal counterpart) derives from the Levant.

DISCUSSION AND CONCLUSIONS

Sicily and Southern Italy have long represented a natural hub for the expansion of human genes and cultures within the Mediterranean Basin [1]. Accordingly, the genetic pool of current populations inhabiting this area can be interpreted as the result of complex interplays and superimpositions between different prehistoric and more recent demographic events, ranging from the Neolithic expansion and the proto-historic Greek and Phoenician colonisations, to the post-Roman invasions by Byzantines, Arabs and Normans. The real demographic impacts of these settlements on the population structure remain still largely uncertain based on the study of material culture and the available historical sources, and different hypotheses about the relative contributions of these events to the current gene pool composition have been proposed from a genetic point of view [7-9].

As a contribution to the human history of such a key area of the Mediterranean we surveyed, by means of a comprehensive evaluation of both maternal and paternal genetic landscapes, the genetic variability of a wide number of populations settled in a broad transect encompassing Sicily and Southern Italy (Figure S1). Previous reconstructions of the genetic structure of Sicily [7-9] focused their attention mainly on two points in the attempt to clarify its genetic history: a) the presence or absence of an internal genetic differentiation along the east-west axis, and b) the extent of the genetic relationships with other populations of the Mediterranean Basin.

Population structure and genetic history of Sicily and Southern-Italy

In contrast with previous investigations on the distribution pattern of genetic variation in Sicily [7-8], our results point to a substantially homogeneous composition of maternal and paternal genetic pools both within Sicily (East vs. West) as well as between Sicily and Southern Italy (Table S4). The absence of significant differences in the distribution of HG frequencies along the east-west axis of the island, as observed not only among our Sicilian populations, but also when including the samples from Di Gaetano et al. (2009) [8], provides further support to these conclusions. The comparison of the whole SSI dataset with a subset based on founder surnames, moreover suggests that the observed homogeneity in Y-chromosome composition is not the result of recent events (e.g. increased population mobility related to the social and economic changes of the 19th and 20th centuries); on the contrary it has been preserved at least since the initial founding and spreading of surnames in Italy. In addition, and consistently with the complex history of migration pathways and cultural exchanges that have characterized the peopling history of the area, high levels of Y-chromosome and mtDNA genetic variability at both SNP and haplotype (STRs or sequence) data, have been observed in all the SSI populations here examined (Table S5).

Altogether, the high levels of within-population variability and the lack of significant genetic sub-structures fit well with the historic role of Sicily and Southern Italy as a major migration crossroad within the Mediterranean Basin. Anyway, differential contributions from the considered Euro-Mediterranean areas were observed. For instance, if the Near East, the Balkans, and – at a lesser extent – North-Western Italy probably had a relevant role in the genetic make-up of SSI, Northern African contributions seem to be almost negligible. As for the Iberian Peninsula, at present its specific genetic contribution cannot be distinguished from that of North-Western Italy, given their observed genetic similarity. These multiple migration events have probably favoured the reduction of genetic differentiation across the region, by increasing the rates of gene flows between different ethnic groups and in some cases mixing up the different genetic strata. Interestingly, the presence of massive migratory phenomena not necessarily yields genetic homogeneity in a given region. For instance, recent studies [46-47] showed how ethno-linguistic minorities from Sicily and Southern Italy - such as the Albanian-speaking Arbereshe - may conserve a significant genetic diversification from the rest of the population. In general, such features are more easily observed in isolated populations, thanks to their reduced population size and their cultural distinctiveness, if compared to open populations.

The patterns of genetic variability observed in our SSI sample are in agreement with the general statement that Southern European populations tend to show higher levels of genetic diversity when compared with those located at more northern latitudes [48] by virtue of the several past demographic events that affected their genetic composition over time. Additionally to the postglacial re-expansion and the demic diffusion of agriculture from Near East, also more recent events (e.g. gene flows from North Africa [48]) have been recently advocated as other possible explanations for the increased genetic diversity in the Southern European populations. Among the several historical occupations of Sicily and Southern Italy, the Pre-Roman colonisation by Greeks and Phoenicians as well as the subsequent invasions from North Africa (including the Muslim conquest, that, at least in part, was conducted by Berber forces) have been previously suggested as putative contributors to the gene pool of current Sicilian population (at least from a male perspective [8]). At this respect, the distribution of Y-chromosome haplogroup E-M81 is widely associated in literature with recent gene flows from North-Africa [49]. Besides the low frequency (1.5%) of E-M81 lineages in general observed in our SSI dataset, the typical Maghrebin core haplotype 13-14-30-24-9-11-13 [8] has been found in only two out of the five E-M81 individuals. These results, along with the negligible contribution from North-African populations revealed by the admixture-like plot analysis, suggest only a marginal impact of trans-Mediterranean gene flows on the current SSI genetic pool. Together with the Berber E-M81, the occurrence of the Near-Eastern J1-M267 in Southern-European populations has been linked to population movements from the Near East through North-Africa, and particularly as a marker of the Islamic expansion over Southern-Europe (started approximately in the 8th century AD and lasted for more than 500 years). Fisher exact tests based on HGs frequencies have revealed the presence of haplogroup J1-M267 at significantly higher frequencies in both North-Africa and the Levant than in Sicily and Southern Italy (both P-values < 0.001). However, the estimated age for Sicilian and Southern-Italian J1 haplotypes refers to the end of the Bronze Age (3261 ± 1345 YBP), thus suggesting more ancient contributions from the East. Nevertheless, our time estimate does not necessarily coincide with the time of arrival of J1 in SSI; in fact a pre-existing differentiation could potentially backdate the time estimate here obtained.

By the collapse of the Late Bronze Age societies (approximately 3200 YBP), the Mediterranean Basin underwent different waves of invasion, particularly by the Greeks of the Aegean Sea and, to a lower extent, by Levantine (Phoenicians) groups [50]. Both of them established a set of different colonies along the Mediterranean coasts of Southern Europe and North Africa. The Phoenician colony of Carthage (present-day Tunisia), given its geographic proximity to Sicily, may have played an important role in the colonization of this region. Previous Y-chromosome genetic studies on the Phoenician colonization demonstrated that haplogroup J2 in general, and six haplotypes in particular (PCS1+ through PCS6+), may potentially have represented lineages linked with the spread of the Phoenicians (‘‘Phoenician Colonization Signal’’) into the Mediterranean [51]. At this respect, it is worth noting the presence of 4 PCS+ haplotypes (namely PCS1+, PCS2+, PCS4+, PCS5+; [51]) in 9 samples of our Sicilian and Southern Italian dataset, particularly belonging to haplogroups J1-M267 (n=2), J2-M410* (n=1), J2-M67 (n=5), and J2-M12 (n=2). However, sub-lineages of haplogroup J2 have been also associated with the Neolithic colonization of mainland Greece, Crete and Southern Italy [52], and our TMRCA estimates for J2-subhaplogroups (ranging from 3271 ± 1157 YBP to 3767 ± 1332 YBP) cannot exclude an earlier arrival of at least some of the J2 chromosomes in Sicily and Southern-Italy during Neolithic times.

On the other hand, Y-chromosome lineage E-V13 is thought to have originated in southern Balkans [53-54] and then to have spread in Sicily at high frequencies with the Greek colonization of the island [8]. The E-V13 core haplotype 13-13-30-24-10-11-13 (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393), which define the southern Balkan Modal Haplotype and reaches frequencies of ~12% in continental Greece [52], has been found in 10 out of the 31 E-V13 samples of Sicily and Southern Italy. This result, along with the high frequency of E-V13 lineages generally observed in our dataset (the second most frequent haplogroup after G2a), confirms the presence of gene flows into Sicily from the Balkans as previously observed by Di Gaetano et al. (2009) [8]. Accordingly, our TMRCA estimate for E-V13 (2354 ± 832 YBP) agrees with the results previously reported in literature for the Sicilian population (2380 YBP, [8]). Altogether, these results do not exclude the possible introduction of some of these Y-lineages with migration processes originated in the Balkans and particularly associated with the Greek colonisation of Southern Italy.

Y-chromosome haplogroup G2a-P15 turned out to be of particular interest in the paternal genetic make-up of Sicily and Southern Italy. Its older age estimate (9339 ± 3302 YBP) – if compared to those of other haplogroups – along with its significantly over-represented frequency in SSI, are consistent with the hypothesis recently suggested by Boattini et al. (2013) [12] according to whom this lineage could be a possible candidate for a pre-Neolithic ancestry in Italy. However the CIs of our time estimate cannot exclude alternative hypotheses such as a diffusion of its major sub-clades during Neolithic and Post-Neolithic times, as recently discussed by Rootsi et al. 2012 [55].

Contrarily to Y-chromosome results, age estimates for mtDNA haplogroups suggest that most of the maternal diversity of the current Sicilian and Southern Italian population is composed by lineages present in Europe as early as the LGM (Table 1). The Late Glacial and Postglacial re-occupation of Europe from refugial areas located in the Mediterranean Peninsulas, has played a major role in shaping the gene pool of modern Europeans [56] and some of the differences in genetic diversity of current European populations have been attributed also to this process [48]. Consistently, the geographic distribution and ages of some mtDNA haplogroups, such as V, H1 and H3, have been associated to events of postglacial re-colonisation from Southern European glacial refugia, and particularly from the Franco-Cantabrian area [57-60]. Further evidences of post-glacial resettlement from Southern refugia have been recently suggested also for the mtDNA haplogroup H5 (the third most common European H-sublineage after H1 and H3), by considering its higher occurrence in southern European populations (particularly Italy) and its evolutionary age ranging approximately between 11,500 and 16,000 YBP [61].

Together with the Iberian and Balkan Peninsulas, also Italy and particularly SSI might have played an important role during the post-glacial re-expansion, as widely attested by several animal and plant species [62-68]. As in the case of Iberia and the Balkans, the presence of numerous Epigravettian sites suggests that Italy could have acted as such also for humans [69], despite the fact that strong genetic evidences are still missing (except for mtDNA haplogroup U5b3 [70]).

Haplogroups H1 and H5 appeared to represent the most frequent H-sublineages in SSI, and their age estimates (Table 1) are consistent with post-glacial time periods, as previously observed for both Southern Italy [11] and the entire Peninsula [12]. Nevertheless, a significant (P-value 0.045) over-representation of H1 haplotypes and an older age (17295 ± 5119 YBP) has been obtained for the Iberian population (as represented by the considered reference samples) than in our SSI datatset, thus suggesting, at least for H1, a post-glacial re-expansion presumptively originated in the Franco-Cantabrian area.

Interestingly, mtDNA haplogroup HV confirmed to be the most ancient lineage in Sicily and Southern Italy, predating the LGM (32242 ± 12595 YBP) and thus representing a possible candidate for the Palaeolithic ancestry of Southern Italy, even though possible post-LGM expansions of its major sub-branches should be taken into account as potentially affecting the time estimates here obtained. Further analyses, involving the complete sequencing of mtDNA genomes and the analysis of ancient DNA samples, are therefore needed in order to more deeply address this point and to confirm the relevance of this haplogroup in the first peopling of Sicily by moderns humans, as recently suggested by some Palaeogenetic researches [5].

Patterns of genetic relationships within the Mediterranean Basin

When comparing SSI with Mediterranean reference populations, Y-chromosome results (Figure 1 and Figure S2) revealed a clear-cut genetic differentiation between the North-Western vs. the Central- and South-Eastern Mediterranean genetic pools (as confirmed by both sPCA G-test and AMOVA F_{CT} statistically significant tests). These results are consistent with our previous study about Italy [12], in which we detected a discontinuous paternal genetic structure, clearly separating the South-Eastern and the North-Western parts of the Italian Peninsula. Here this pattern appears extended to the whole Mediterranean Basin, particularly suggesting a shared genetic background between South-Eastern Italy and the South-Eastern Mediterranean cluster from one side, and between North-Western Italy and the Western Europe from the other side (Figure 2).

Y-chromosome results however contrast with the lack of statistical support to the sPCA global structure observed for mtDNA diversity, excepted for a similar NW-SE genetic pattern identified by sPC1 (Figure 3). The common South-East to North-West pattern in the distribution of genetic variation across the European and Mediterranean domain, could be interpreted as reflecting the same SE to NW genetic cline extensively reported in literature for the whole of Europe [71-74]. However, the general lack of statistical support to the global structure observed for mtDNA markers suggests a higher homogeneity for maternal than paternal genetic pools in the Mediterranean genetic landscape. These results could be ascribed to older population events and/or different demographic and historical dynamics for females than males. The differential income of male genes into a population has been indeed advocated as one of the possible reasons why matrilines tend to be more stable over time than patrilines. Such a male-biased pattern has been suggested for the

Neolithisation of Southern Europe [75-76] and proposed also in the case of the first Greek incoming groups in Sicily and Southern Italy [77]. As a consequence of such kind of sex-biased dynamics, male lineages could be better suited to detect more recent population events than the female ones, which instead trace back to more ancient time periods [49]. Accordingly, while the time estimates for Sicilian and Southern Italian mtDNA haplogroups date almost unanimously to Pre-Neolithic times, Y-chromosome results highlight the importance of Neolithic and Post-Neolithic (Metal Ages) demographic events in shaping the current paternal diversity composition (Table 1). Differences between the two uniparental genetic systems also appeared when the genetic relationships among Mediterranean population groups were more deeply addressed in admixture analyses (Figure 4 and Figure S3). In fact, whereas the different continental and within continental contributions to the current SSI genetic pool appeared to be more equally distributed on the maternal side (despite a noteworthy contribution of Levantine females), the paternal counterpart appeared to be clearly affected by South-Eastern Mediterranean, mainly Balkan, males.

In summary, Sicilian genetic diversity revealed to be not structured along the east-west axis of the island; on the contrary both maternal and paternal genetic markers suggest an homogeneous genetic composition both within Sicily, as well as between Sicily and Southern Italy. These results are consistent with the largely shared genetic histories of the Southern Italian populations, and reflect their historical and archaeological role as a major Mediterranean ‘melting pot’ where different peoples and cultures came together over time, albeit with different contributes depending from the source area.

When Sicilian and Southern Italian population were contextualized within the Mediterranean domain, the observed homogeneous pattern of genetic variation, however revealed different temporal dynamics and spatial genetic contributions to the maternal and paternal inheritances.

Besides a common SE-NW distribution pattern of genetic variation, the mtDNA suggests an homogeneous genetic landscape related to older populations events and/or a higher female mobility. On the contrary, Y-chromosomal genetic diversity appears significantly differentiated between a Central/South-Eastern and a North-Western Mediterranean group, the Italian Peninsula occupying an intermediate position between them. In particular, and consistently with the most recent syntheses on the Italian genetic structure based on both uniparental markers [12] and genome wide data [78], Sicily and Southern Italy exhibit predominant influences from the Central and South-Eastern Mediterranean regions, especially the Balkans. If contacts between SSI and the Balkans date back at least to the Neolithic, the Greek dominion of the late Metal Ages seems to have played a particularly important role, accounting at least in part for the observed shared genetic background between SSI and the Balkan Peninsula. Further studies involving model-like populations such as ethno-linguistic minorities, together with wide-genome analyses, will provide a complementary overview to the perspectives offered by uniparentally-inherited markers, thus allowing to more deeply test specific hypotheses related to the peopling history of Sicily and Southern Italy. In addition, this study will represent the starting point for future explorations aimed at specifically investigating the impact of different historical, geographical and linguistic factors on the population genetic substratum, within specific macro- and micro-geographic contexts of the Euro-Mediterranean genetic landscape.

ACKNOWLEDGEMENTS

This study was supported by the ERC Langelin Project grant (FP7-Ideas-ERC2011-AdG295733) to DP and DL. We are indebted to all the Personnel of the Local Blood Centres and Hospital Centres of Sicily and Southern Italy for their invaluable help in performing the sampling campaign. We thank Dr. Serafina Salimbeni for helping us in the collection of samples from Cosenza (Corigliano Calabro). We thank all the volunteers who kindly agreed to participate in this study. We are very grateful to Dr. Eugenio Bortolini for his valuable suggestions to the manuscript and for the language revision. We would like to thank the two reviewers for their insightful and constructive comments which helped to improve the quality of the manuscript.

REFERENCES

1. Sazzini M, Sarno S, Luiselli D (2013) The Mediterranean human population: an Anthropological Genetics perspective. In: Goffredo S, Baader H, Dubinsky Z, editors. The Mediterranean Sea: Its History and Present Challenges. Berlin: Springer, pp. 529-551.
2. Mannino MA, Thomas KD (2007) New radiocarbon dates for hunter-gatherers and early farmers in Sicily. Accordia Research Papers 10: 13–34.
3. Mannino MA, Di Salvo R, Schimmenti V, Di Patti C, Incarbona A, Sineo L, Richards MP (2011) Upper Palaeolithic hunter-gatherer subsistence in Mediterranean coastal environments: an isotopic study of the diets of the oldest directly-dated humans from Sicily. J Archaeol Sci 38: 3094–3100. doi: 10.1016/j.jas.2011.07.009.
4. Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, Chiarelli B, Dupanloup I, Bertranpetit J, Barbujani G, Bertorelle G (2003) Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci U.S.A. 100: 6593–6597. doi: 10.1073/pnas.1130343100.
5. Mannino MA, Catalano G, Talamo S, Mannino G, Di Salvo R, Schimmenti V, Lalueza-Fox C, Messina A, Petruso D, Caramelli D, Richards MP, Sineo L (2012) Origin and diet of the prehistoric hunter-gatherers on the mediterranean island of Favignana (Ègadi Islands, Sicily). PLoS One. 7:e49802. doi: 10.1371/journal.pone.0049802.
6. Pessina A, Tinè V (2008) Archeologia del Neolitico. L'Italia tra il Vi e il IV millennio a.C. Roma: Carrocci editore. 375 p.
7. Romano V, Cali F, Ragalmuto A, D'Anna RP, Flugy A, De Leo G, Giambalvo O, Lisa A, Fiorani O, Di Gaetano C, Salerno A, Tamouza R, Charron D, Zei G (2003). Autosomal microsatellite and mtDNA genetic analysis in Sicily (Italy). Ann Hum Genet 67:42–53.
8. Di Gaetano C, Cerutti N, Crobu F, Robino C, Inturri S, Gino S, Guarrera S, Underhill PA, King RJ, Romano V, Cali F, Gasparini M, Matullo G, Salerno A, Torre C, Piazza A (2009). Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. Eur J Hum Genet. 17:91-99.
9. Rickards O, Martinez-Labarga C, Scano G, De Stefano GF, Biondi G, Pacaci M, Walter H. 1998. Genetic history of the population of Sicily. Hum Biol 70:699–714.
10. Turchi C, Buscemi L, Previderè C, Grignani P, Brandstätter A, Achilli A, Parson W, Tagliabracci A and Ge.F.I. Group (2008). Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int J Legal Med. 122:199-204.
11. Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, Contini I, Biondi G, Rickards O (2009). Human mitochondrial DNA variation in Southern Italy. Ann Hum Biol. 36:785-811. doi: 10.3109/03014460903198509.
12. Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, Sanz P, Yang-Yao D, Manry J, Ciani G, Luiselli D, Quintana-Murci L, Comas D, Pettener D, the Genographic Consortium (2013). Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. PLoS One. 8:e65441. doi: 10.1371/journal.pone.0065441.
13. Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F (2012) General method to unravel ancient population structures through surnames. Final validation on Italian data. Hum Biol 84: 235-270.
14. Larmuseau MH1, Vanoverbeke J, Gielis G, Vanderheyden N, Larmuseau HF, Decorte R (2012) In the name of the migrant father--analysis of surname origins identifies genetic admixture events undetectable from genealogical records. Heredity. 109:90-95. doi: 10.1038/hdy.2012.17.
15. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 16:1215.
16. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK (2006) Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci. 51:64-75.
17. Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, Morling N, Prinz M, Roewer L, Tyler-Smith C, Schneider PM; DNA Commission of the International Society of Forensic Genetics (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci Int. 157:187-197.
18. Onofri V, Alessandrini F, Turchi C, Pesaresi M, Buscemi L, Tagliabracci A (2006) Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. Forensic Sci Int. 57:23-35.
19. Ferri G, Alù M (2012) Development of six-Y-SNPs assay for forensic analysis in European population. DNA in Forensics 2012, 5th International EMPOP Meeting- 8th International Forensic Y-User Workshop, Innsbruck.
20. Neto D, Montiel R, Bettencourt C, Santos C, Prata MJ, Lima M (2007) The African contribution to the present-day population of the Azores Islands (Portugal): analysis of the Y chromosome haplogroup E. Am J Hum Biol. 19:854-860.
21. Gayden T, Regueiro M, Martinez L, Cadenas AM, Herrera RJ (2008) Human Y-chromosome haplotyping by allele-specific polymerase chain reaction. Electrophoresis. 29:2419-2423. doi: 10.1002/elps.200700702.

22. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature. 290:457-465.
23. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 23:147.
24. Behar DM, Van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R (2012a) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet. 90:675-684. doi: 10.1016/j.ajhg.2012.03.002.
25. Bertoncini S, Bulayeva K, Ferri G, Pagani L, Caciagli L, Taglioli L, Semyonov I, Bulayev O, Paoli G, Tofanelli S (2012) The dual origin of Tati-speakers from Dagestan as written in the genealogy of uniparental variants. Am J Hum Biol. 24:391-399. doi: 10.1002/ajhb.22220.
26. Quintáns B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int. 140:251-257.
27. Excoffier L, Laval G, Schneider S (2007) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 1:47-50.
28. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.
29. Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 24:1403-1405.
30. Jombart T, Devillard S, Dufour AB, Pontier D (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101:92-103.
31. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.
32. Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet. 78:202-221.
33. Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, Bradley DG, Gusmão L, Winney B, Bodmer W, Vennemann M, Coia V, Scarnicci F, Tofanelli S, Vona G, Ploski R, Vecchiotti C, Zemunik T, Rudan I, Karachanak S, Toncheva D, Anagnostou P, Ferri G, Rapone C, Hervig T, Moen T, Wilson JF, Capelli C (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci. 279:884-892. doi: 10.1098/rspb.2011.1044.
34. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, Choi Y, van Duijn K, Vermeulen M, Brauer S, Decorte R, Poetsch M, von Wurmb-Schwark N, de Knijff P, Labuda D, Vézina H, Knoblauch H, Lessig R, Roewer L, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet. 87:341-353. doi: 10.1016/j.ajhg.2010.08.006.
35. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am. J. Hum. Genet. 84:740-759.
36. Cox MP (2008) Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. Hum Biol 80:335-357.
37. Fraley C and Raftery AE (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97:611:631.
38. Fraley C and Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. Journal of Classification 24:155-181.
39. Bertorelle G and Excoffier L (1998) Inferring admixture proportions from molecular data. Mol Biol Evol. 15:1298-1311.
40. Dupanloup I and Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol. 18:672-675.
41. Chakraborty R (1986) Gene admixture in human populations: models and predictions. Yearb Phys Anthropol 29:1–43
42. Sans M, Salzano FM, Chakraborty R (1997) Historical genetics in Uruguay: estimates of biological origins and their problems. Hum Biol 69:161–170
43. Wen B, Xie X, Gao S, Li H, Shi H, Song X, Qian T, Xiao C, Jin J, Su B, Lu D, Chakraborty R, Jin L (2004) Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet. 74:856-865.
44. Zalloua PA, Xue Y, Khalife J, Makhoul N, Debiane L, Platt DE, Royyuru AK, Herrera RJ, Hernanz DF, Blue-Smith J, Wells RS, Comas D, Bertranpetit J, Tyler-Smith C; Genographic Consortium (2008a) Y-chromosomal

diversity in Lebanon is structured by recent historical events. Am J Hum Genet. 82:873-882. doi: 10.1016/j.ajhg.2008.01.020.

45. Congiu A1, Anagnostou P, Milia N, Capocasa M, Montinaro F, Destro Bisol G (2012) Online databases for mtDNA and Y chromosome polymorphisms in human populations. J Anthropol Sci. 90:201-215. doi: 10.4436/jass.90020.
46. Boattini A1, Luiselli D, Sazzini M, Useli A, Tagarelli G, Pettener D (2010) Linking Italy and the Balkans. A Y-chromosome perspective from the Arbereshe of Calabria.Ann Hum Biol. 38:59-68. doi: 10.3109/03014460.2010.491837.
47. Capocasa M, Anagnostou P, Bachis V, Battaggia C, Bertoncini S, Biondi G, Boattini B, Boschi I, Brisighelli F, Calò CM, Carta M, Coia V, Corrias L, Crivellaro F, Ferri G, Francalacci P, Franceschi ZA, Luiselli D, Morelli L, Rickards O, Robledo R, Sanna D, Sanna E, Sarno S, Tofanelli S, Vona G, Pettener D and Destro Bisol G (2014) Linguistic, geographic and genetic isolation: a collaborative study on Italian populations. J Anthropol. Sci. 92:1-32 doi:10.4436/JASS.92001
48. Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, Bertranpetit J, Comas D, Bustamante CD (2013) Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci U S A. 110:11791-6. doi: 10.1073/pnas.1306223110.
49. Bekada A, Fregel R, Cabrera VM, Larruga JM, Pestano J, Benhamamouch S, González AM. (2013) Introducing the Algerian mitochondrial DNA and Y-chromosome profiles into the North African landscape. PLoS One. 8:e56775. doi: 10.1371/journal.pone.0056775.
50. Murray O (1993) Early Greece. 2 edition. Cambridge: Harvard University Press.
51. Zalloua PA, Platt DE, El Sibai M, Khalife J, Makhoul N, Haber M, Xue Y, Izaabel H, Bosch E, Adams SM, Arroyo E, López-Parra AM, Aler M, Picornell A, Ramon M, Jobling MA, Comas D, Bertranpetit J, Wells RS, Tyler-Smith C; Genographic Consortium (2008b) Identifying genetic traces of historical expansions: Phoenician footprints in the Mediterranean. Am J Hum Genet. 83:633-642. doi: 10.1016/j.ajhg.2008.10.012.
52. King RJ, Ozcan SS, Carter T, Kalfoğlu E, Atasoy S, Triantaphyllidis C, Kouvatsi A, Lin AA, Chow CE, Zhivotovsky LA, Michalodimitrakis M, Underhill PA. (2008) Differential Y-chromosome Anatolian influences on the Greek and Cretan Neolithic. Ann Hum Genet. 72:205-214. doi: 10.1111/j.1469-1809.2007.00414.x.
53. Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon JM, Crivellaro F, Benincasa T, Pascone R, Moral P, Watson E, Melegh B, Barbujani G, Fuselli S, Vona G, Zagradisnik B, Assum G, Brdicka R, Kozlov AI, Efremov GD, Coppa A, Novelletto A, Scozzari R (2007) 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. 24:1300-1311.
54. Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, King RJ, Rootsi S, Marjanovic D, Primorac D, Hadziselimovic R, Vidovic S, Drobnic K, Durmishi N, Torroni A, Santachiara-Benerecetti AS, Underhill PA, Semino O (2008) Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. Eur J Hum Genet. 17:820-830. doi: 10.1038/ejhg.2008.249.
55. Rootsi S, Myres NM, Lin AA, Järve M, King RJ, Kutuev I, Cabrera VM, Khusnutdinova EK, Varendi K, Sahakyan H, Behar DM, Khusainova R, Balanovsky O, Balanovska E, Rudan P, Yepiskoposyan L, Bahmanimehr A, Farjadian S, Kushniarevich A, Herrera RJ, Grugni V, Battaglia V, Nici C, Crobu F, Karachanak S, Hooshiar Kashani B, Houshmand M, Sanati MH, Toncheva D, Lisa A, Semino O, Chiaroni J, Di Cristofaro J, Villems R, Kivisild T, Underhill PA (2012) Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. Eur J Hum Genet. 20:1275-1282. doi: 10.1038/ejhg.2012.86.
56. Behar DM, Harmant C, Manry J, van Oven M, Haak W, Martinez-Cruz B, Salaberria J, Oyharçabal B, Bauduer F, Comas D, Quintana-Murci L; Genographic Consortium (2012) The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. Am J Hum Genet 90: 486-493. doi: 10.1016/j.ajhg.2012.01.002.
57. Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB (2010) The Archaeogenetics of Europe. Curr Biol 20: R174-183. doi: 10.1016/j.cub.2009.11.054.
58. Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumäki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet. 69:844-852.
59. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogväli EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet. 75:910-918.
60. Pereira L, Richards M, Goios A, Alonso A, Albarrán C, Garcia O, Behar DM, Gölge M, Hatina J, Al-Gazali L, Bradley DG, Macaulay V, Amorim A (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. 15:19-24.

61. Mielnik-Sikorska M, Daca P, Malyarchuk B, Derenko M, Skonieczna K, Perkova M, Dobosz T, Grzybowski T (2013) The history of Slavs inferred from complete mitochondrial genome sequences. 8:e54360. doi: 10.1371/journal.pone.0054360.
62. Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol 7: 453-464.
63. Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, et al. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300: 1563-1565.
64. Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. Philos Trans Ser B 359: 183-195.
65. Randi E (2007) Phylogeography of South European Mammals. In: Weiss S, Ferrand N, editors. Phylogeography of Southern European Refugia. Amsterdam: Kluwer Academic Publishers. pp. 101-126.
66. Grassi F, De Mattia F, Zecca G, Sala F, Labra M (2008) Historical isolation and Quaternary range expansion of divergent lineages in wild grapevine. Biological Journal of the Linnean Society 95: 611-619.
67. Grassi F, Minuto L, Casazza G, Labra M, Sala F (2009) Haplotype richness in refugial areas: phylogeographical structure of Saxifraga callosa. Journal of Plant Research 122: 377–387.
68. Zecca G, Casazza G, Labra M, Minuto L, Grassi F (2011) Allopatric divergence and secondary contacts in Euphorbia spinosa L: Influence of climate change on the split of the species. Organisms Diversity and Evolution 11: 357-372.
69. Banks WE, d’Errico F, Peterson AT, Vanhaeren M, Kageyama M, Sepulchre P, Ramstein G, Jost A, Lunt D (2008) Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. J Archaeol Sci, 35:481–491.
70. Pala M, Achilli A, Olivieri A, Hooshiar Kashani B, Perego UA, Sanna D, Metspalu E, Tambets K, Tamm E, Accetturo M, Carossa V, Lancioni H, Panara F, Zimmermann B, Huber G, Al-Zahery N, Brisighelli F, Woodward SR, Francalacci P, Parson W, Salas A, Behar DM, Villems R, Semino O, Bandelt HJ, Torroni A (2009) Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet. 84:814-821. doi: 10.1016/j.ajhg.2009.05.004.
71. Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton: Princeton University Press.
72. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol. 18:1241-1248. doi: 10.1016/j.cub.2008.07.049.
73. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. 456:98-101. doi: 10.1038/nature07331.
74. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskácková T, Balascák I, Peltonen L, Jakkula E, Rehnström K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgár N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julià A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A (2009) Genetic structure of Europeans: a view from the North-East. PLoS One 4:e5472. doi: 10.1371/journal.pone.0005472.
75. Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, Guilaine J, Crubézy E, Ludes B (2011a) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci U S A. 108:9788-9791. doi: 10.1073/pnas.1100723108.
76. Lacan M, Keyser C, Ricaut FX, Brucato N, Tarrús J, Bosch A, Guilaine J, Crubézy E, Ludes B (2011b) Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc Natl Acad Sci U S A. 108:18255-18259. doi: 10.1073/pnas.1113061108.
77. Pesando F (2005) L’Italia antica. Culture e forme del popolamento nel I millennio a. C. Roma: Carocci editore. 326 p.
78. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, Manzini P, Dianzani I, Betti M, Cusi D, Frau F, Barlassina C, Mirabelli D, Magnani C, Glorioso N, Bonassi S, Piazza A, Matullo G (2012) An overview of the genetic structure within the Italian population from genome-wide data. PLoS One. 7:e43759. doi: 10.1371/journal.pone.0043759.

FIGURES

.

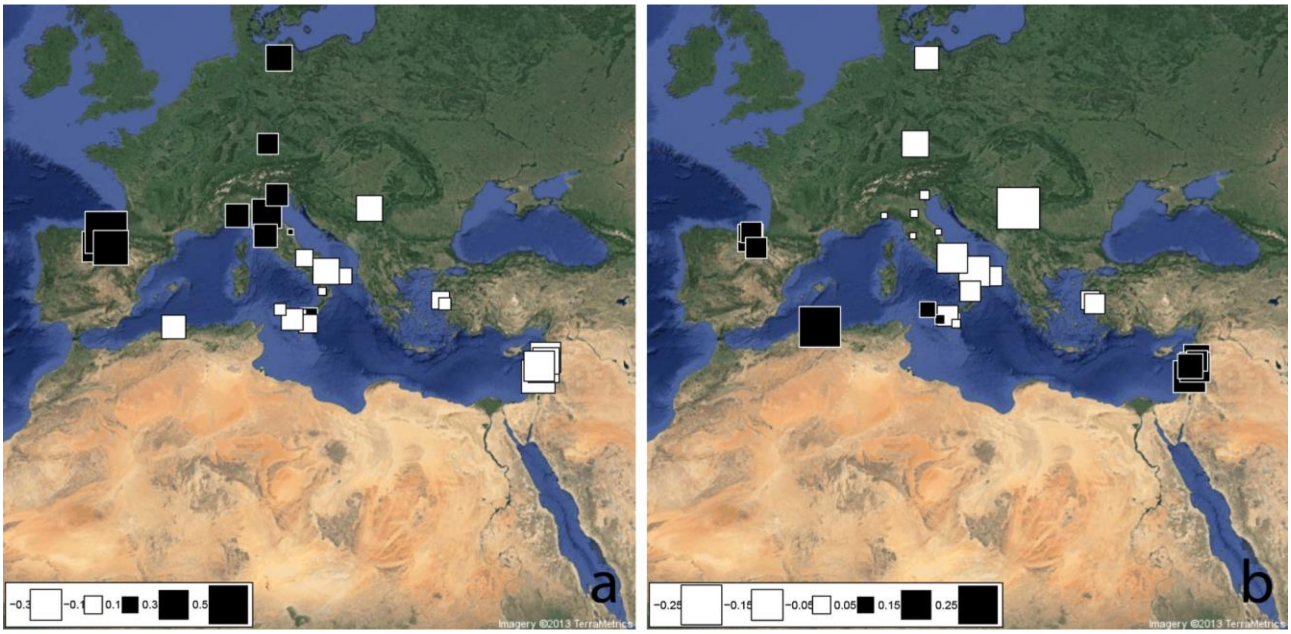


Figure 1. Spatial Principal Component Analysis (sPCA) based on Y-chromosome haplogroups frequencies. The first two global components, sPC1 (a) and sPC2 (b), are depicted. Positive values are represented by black squares; negative values are represented by white squares; the size of the square is proportional to the absolute value of sPC scores

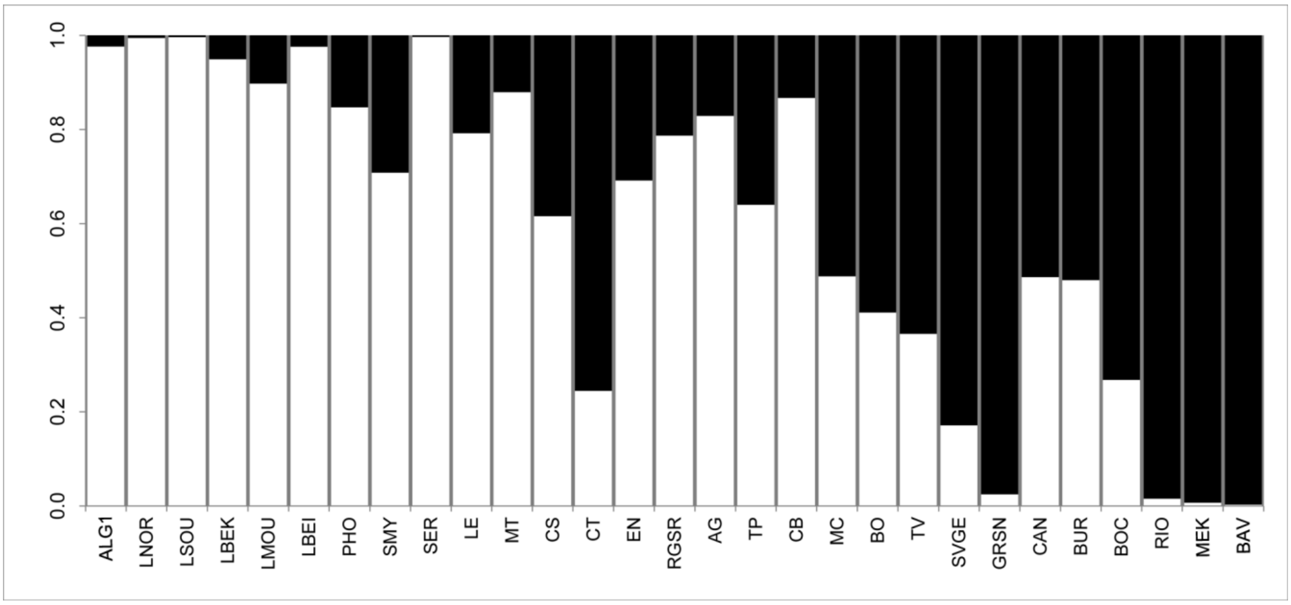


Figure 2. Discriminant Analysis of Principal Components (DAPC) based on Y-chromosome sPC1-identified structure. The barplot represents DAPC-based posterior membership probabilities for each of the considered populations to belong at each of the two sPC1-identified groups (white = South-Eastern Mediterranean; black = North-Western Mediterranean). Population codes as in Table S1.

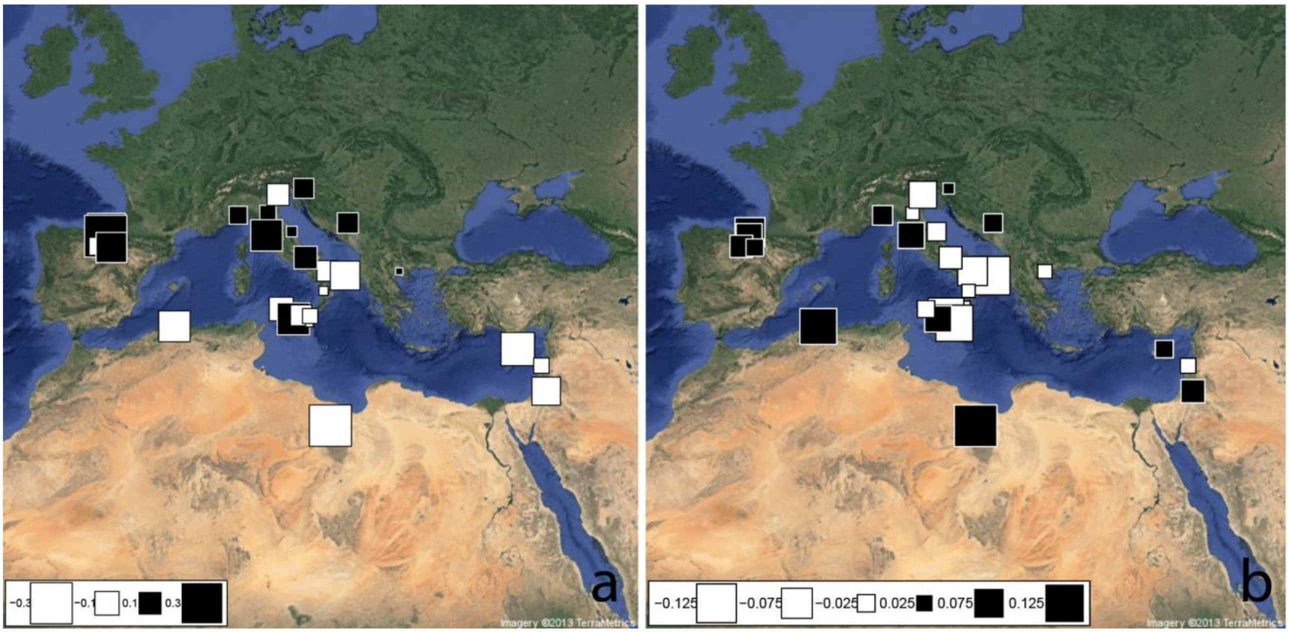


Figure 3. Spatial Principal Component Analysis (sPCA) based on mtDNA haplogroups frequencies. The first two global components sPC1 (a) and sPC2 (b) are depicted. Positive values are represented by black squares; negative values are represented by white squares; the size of the square is proportional to the absolute value of sPC scores.

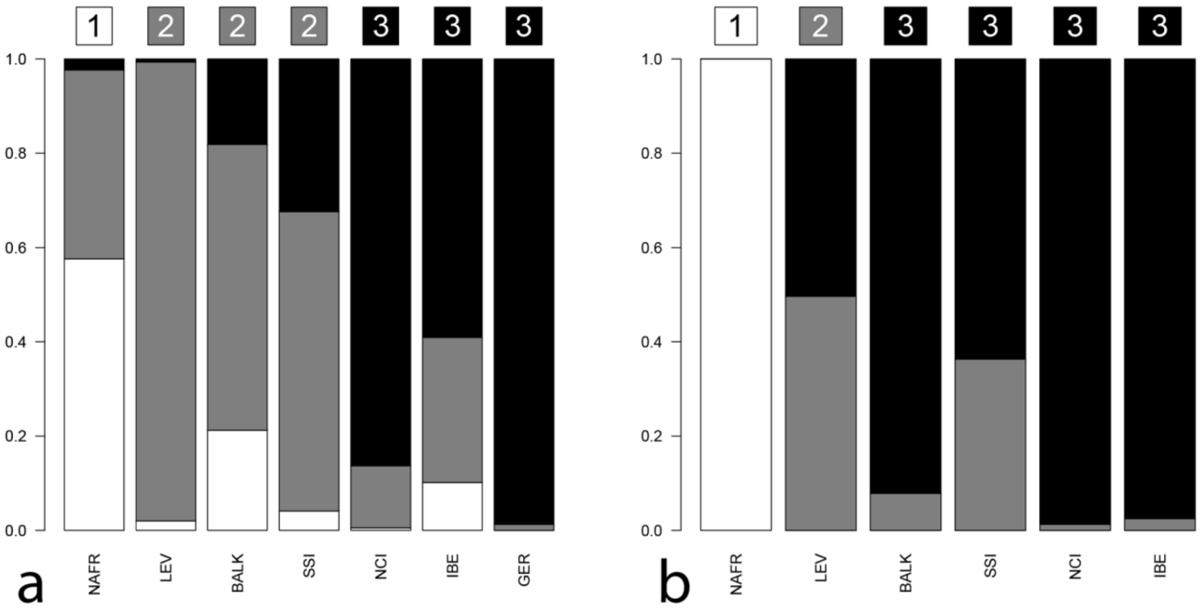


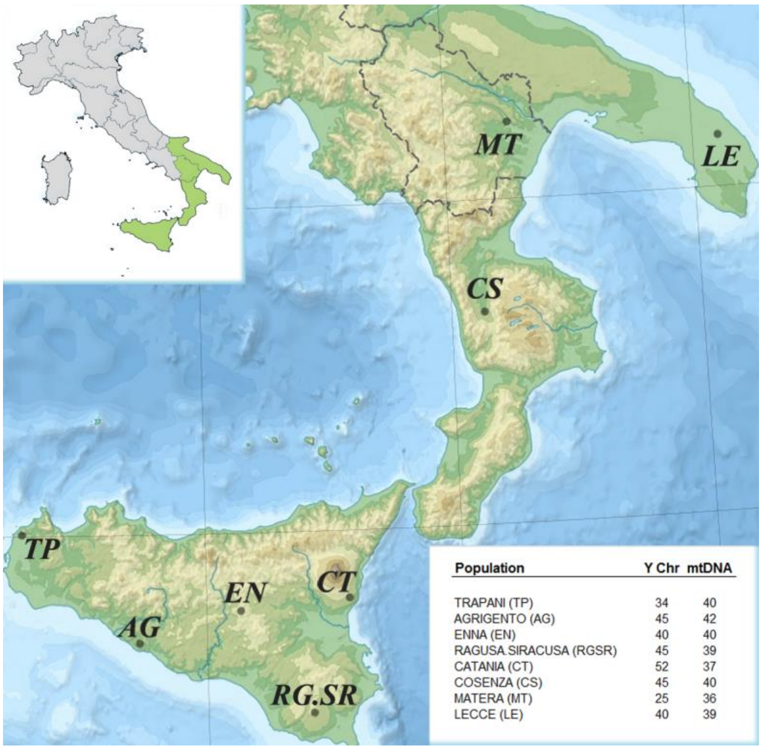
Figure 4. Admixture-like barplots for Y-chromosome (a) and mtDNA (b). The barplots represent DAPC-based posterior membership probabilities for each of the considered populations and for each inferred cluster (*mclust* algorithm). The affiliation of each population to a given cluster and its corresponding colour code are represented by letters (within coloured squares) on the top of each bar. Labels: NAFR: North-Africa, LEV: Levant, BALK: Balkans, SSI: Sicily and South-Italy, NCI: North-Central Italy, IBE: Iberian Peninsula, GER: Germany.

TABLES

Table 1. **Age estimates (in YBP) of STR and HVS variation for the most frequent haplogroups in Sicily and Southern Italy.** Standard deviation (SD) estimator (Sengupta et al. 2006) and ρ statistic calculator (Soares et al. 2009) were used for Y-chromosome and mtDNA haplogroups respectively.

Y-chromosome HG	N	%	SD	SE	TMRCA	SE
G-P15	40	12.3	373.6	132.1	9339	3302
E-V13	31	9.5	94.2	33.3	2354	832
J-M410(xM67,M92)	31	9.5	150.7	53.3	3767	1332
R-M17	17	5.2	172.2	60.9	4305	1522
J-M267	16	4.9	130.4	53.8	3261	1345
R-P312	15	4.6	175.2	61.9	4380	1549
R-U152	14	4.3	80.1	28.3	2002	708
R-U106	12	3.7	82.6	29.2	2066	730
J-M92	11	3.4	146.3	55.3	3658	1382
J-M12	11	3.4	148.6	52.6	3716	1314
J-M67	10	3.1	130.8	46.3	3271	1157
MtDNA HG	**N**	**%**	**Rho**	**SE**	**TMRCA**	**SE**
H	43	13.7	0.93	0.17	15513	5586
H1	34	10.9	0.94	0.18	15696	5768
T2	28	8.9	1.71	0.30	28589	9905
J1	16	5.1	1.50	0.38	25016	12258
HV	15	4.8	1.93	0.39	32242	12595
J2	15	4.8	1.87	0.38	31130	12434
T1	11	3.5	1.73	0.39	28806	12626
U5	11	3.5	1.64	0.39	27290	12734
H5	10	3.2	1.00	0.30	16677	9806

SUPPLEMENTARY MATERIALS



Population	Y Chr	mtDNA
TRAPANI (TP)	34	40
AGRIGENTO (AG)	45	42
ENNA (EN)	40	40
RAGUSA.SIRACUSA (RGSR)	45	39
CATANIA (CT)	52	37
COSENZA (CS)	45	40
MATERA (MT)	25	36
LECCE (LE)	40	39

Figure S1. Geographic map showing the location of the eight populations analysed in the present study. The table at the bottom right details the set of provinces (sampling points) and the number of samples successfully typed for both Y-chromosome and mtDNA markers.

(Map modified from Wikipedia, http://en.wikipedia.org/wiki/File:Southern_Italy_topographic_map-blank.png).

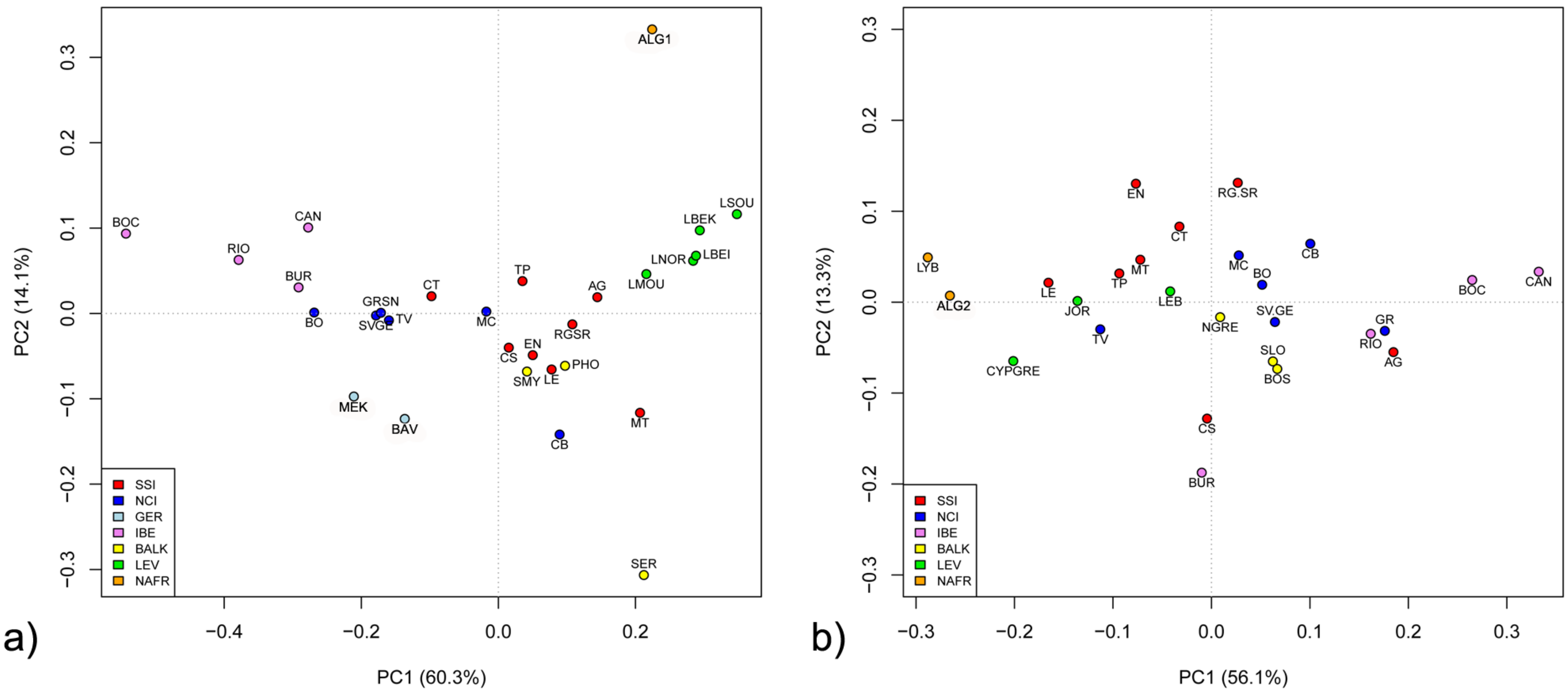


Figure S2. Principal Component Analysis (PCA) based on haplogroup frequencies for Y-chromosome (a) and mtDNA (b). Population codes as in Table S1. Colour codes for geographic affiliations as in the legends at the bottom-left of each plot. Legend abbreviations: NAFR: North-Africa, LEV: Levant, BALK: Balkans, SSI: Sicily and South-Italy, NCI: North-Central Italy, IBE: Iberian Peninsula, GER: Germany.

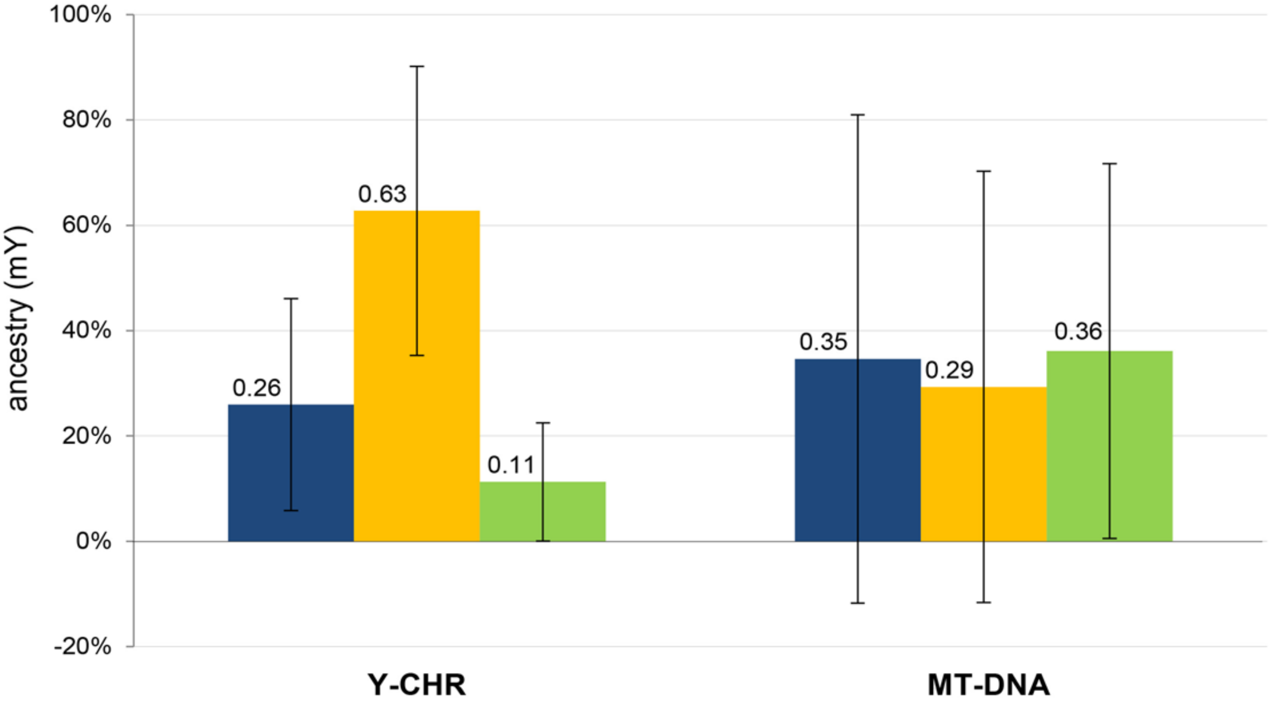


Figure S3. Estimated admixture contributions (mY estimator) from three parental populations to the current population of Sicily and Southern Italy for Y-chromosome (left) and mtDNA (right). Color codes: South-Western Europe (blue), the Balkans (yellow) and the Levant (green). Error bars represent standard deviations calculated on the basis of 10,000 bootstraps.

Table S1. List of the selected Mediterranean populations used for Y-chromosome and mtDNA comparative analyses.

Y-Chr	Geographic Area	Area Code	Populations	Pop Code	N	Longitude	Latitude	Reference
	Sicily and Southern Italy	SSI	Trapani	TP	34	12.52	38.02	This study
			Agrigento	AG	45	13.59	37.32	Boattini et al. 2013
			Enna	EN	40	14.27	37.57	This study
			Ragusa-Siracusa	RGSR	45	15.01	37.01	Boattini et al. 2013
			Catania	CT	52	15.07	37.52	Boattini et al. 2013
			Cosenza	CS	45	16.25	39.30	This study
			Matera	MT	25	16.60	40.67	Boattini et al. 2013
			Lecce	LE	40	18.17	40.35	Boattini et al. 2013
	North-Central Italy	NCI	Savona-Genova	SVGE	51	8.71	44.36	Boattini et al. 2013
			Treviso	TV	33	12.24	45.67	Boattini et al. 2013
			Bologna	BO	29	11.34	44.49	Boattini et al. 2013
			Grosseto-Siena	GRSN	86	11.22	43.05	Boattini et al. 2013
			Macerata	MC	40	13.45	43.3	Boattini et al. 2013
			Campobasso	CB	30	14.67	41.57	Boattini et al. 2013
	Iberian Peninsula	IBE	Western Bizkaia	BOC	19	-2.93	43.27	Martiinez-Cruz et al. 2012
			Cantabria	CAN	18	-3.00	43.00	Martiinez-Cruz et al. 2012
			Burgos	BUR	20	-3.68	42.33	Martiinez-Cruz et al. 2012
			La Rioja	RIO	54	-2.50	42.25	Martiinez-Cruz et al. 2012
	Germany	GER	Mecklenburg	MEK	131	12.43	53.61	Rebala et al. 2012
			Bavaria	BAV	218	11.43	48.78	Rebala et al. 2012
	Balkan Peninsula	BALK	Serbia	SER	102	20.46	44.82	Regueiro et al. 2012
			Phokaia	PHO	30	26.76	38.67	King et al. 2011
			Smyrna	SMY	53	27.14	38.42	King et al. 2011
	Levant	LEV	Beirut	LBEI	61	35.52	33.9	Zalloua et al. 2008
			Bekaa Valley	LBEK	71	35.92	33.83	Zalloua et al. 2008
			Mount Lebanon	LMOU	98	36.12	34.3	Zalloua et al. 2008
			North Lebanon	LNOR	218	36.08	34.56	Zalloua et al. 2008
			South Lebanon	LSOU	129	35.44	33.12	Zalloua et al. 2008
	North Africa	NAFR	Algeria	ALG	102	3.07	36.78	Robino et al. 2012

mtDNA	Geographic Area	Area Code	Populations	PopCod	N	Longitude	Latitude	Reference
	Sicily and Southern Italy	SSI	Trapani	TP	40	12.52	38.02	This study
			Agrigento	AG	42	13.59	37.32	Boattini et al. 2013
			Enna	EN	40	14.27	37.57	This study
			Ragusa-Siracusa	RGSR	39	15.01	37.01	Boattini et al. 2013
			Catania	CT	37	15.07	37.52	Boattini et al. 2013
			Cosenza	CS	40	16.25	39.30	This study
			Matera	MT	36	16.60	40.67	Boattini et al. 2013
			Lecce	LE	39	18.17	40.35	Boattini et al. 2013
	North-Central Italy	NCI	Savona-Genova	SVGE	43	8.71	44.36	Boattini et al. 2013
			Treviso	TV	39	12.24	45.67	Boattini et al. 2013
			Bologna	BO	35	11.34	44.49	Boattini et al. 2013
			Grosseto-Siena	GRSN	36	11.22	43.05	Boattini et al. 2013
			Macerata	MC	39	13.45	43.3	Boattini et al. 2013
			Campobasso	CB	37	14.67	41.57	Boattini et al. 2013
	Iberian Peninsula	IBE	Western Bizkaia	BOC	21	-2.93	43.27	Martiinez-Cruz et al. 2012
			Cantabria	CAN	19	-3.00	43.00	Martiinez-Cruz et al. 2012
			Burgos	BUR	24	-3.68	42.33	Martiinez-Cruz et al. 2012
			La Rioja	RIO	52	-2.50	42.25	Martiinez-Cruz et al. 2012
	Balkan Peninsula	BALK	Bosnia	BOS	144	18.41	43.86	Malyarchuk et al 2003
			Slovenia	SLO	104	14.51	46.06	Malyarchuk et al 2003
			Northrn Greece	NGRE	318	22.95	40.64	Irwin et al. 2008
	Levant	LEV	Cypro Greeks	CYPGRE	85	33.43	35.13	Irwin et al. 2008
			Lebanon	LEB	363	35.52	33.90	Haber et al. 2011
			Jordania	JOR	101	35.94	31.97	González et al. 2008
	North Africa	NAFR	Lybia	LYB	268	16.87	29.3	Fadhlaoui-Zid et al. 2011
			Algeria	ALG	240	3.07	36.78	Bekada et al. 2013

Table S2. Y-chromosome and mtDNA haplogroup frequencies. The absolute number of individuals and the percentage frequency (between brackets) are reported for both the whole Sicialian and Southern Italian dataset and for each population analyzed.

Y-Chr	SSI	ITAS	SIC	ESIC	WSIC	TP	AG	EN	RGSR	CT	CS	MT	LE
	326	110	216	137	79	34	45	40	45	52	45	25	40
E1a-M33	1 (0.31)	0 (0)	1 (0.46)	0 (0)	1 (1.27)	0 (0)	1 (2.22)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
E1b-M2	1 (0.31)	0 (0)	1 (0.46)	0 (0)	1 (1.27)	0 (0)	1 (2.22)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
E1b-M35	3 (0.92)	0 (0)	3 (1.39)	1 (0.73)	2 (2.53)	0 (0)	2 (4.44)	0 (0)	0 (0)	1 (1.92)	0 (0)	0 (0)	0 (0)
E1b-V12	9 (2.76)	5 (4.55)	4 (1.85)	3 (2.19)	1 (1.27)	0 (0)	1 (2.22)	2 (5)	0 (0)	1 (1.92)	4 (8.89)	1 (4)	0 (0)
E1b-V13	31 (9.51)	13 (11.82)	18 (8.33)	9 (6.57)	9 (11.39)	5 (14.71)	4 (8.89)	3 (7.5)	3 (6.67)	3 (5.77)	5 (11.11)	2 (8)	6 (15)
E1b-V22	7 (2.15)	2 (1.82)	5 (2.31)	4 (2.92)	1 (1.27)	0 (0)	1 (2.22)	2 (5)	1 (2.22)	1 (1.92)	0 (0)	2 (8)	0 (0)
E1b-M81	5 (1.53)	1 (0.91)	4 (1.85)	3 (2.19)	1 (1.27)	0 (0)	1 (2.22)	0 (0)	2 (4.44)	1 (1.92)	1 (2.22)	0 (0)	0 (0)
E1b-M123	7 (2.15)	3 (2.73)	4 (1.85)	2 (1.46)	2 (2.53)	0 (0)	2 (4.44)	1 (2.5)	0 (0)	1 (1.92)	1 (2.22)	1 (4)	1 (2.5)
G1-M285	2 (0.61)	1 (0.91)	1 (0.46)	0 (0)	1 (1.27)	0 (0)	1 (2.22)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)
G2a-P15	40 (12.27)	14 (12.73)	26 (12.04)	17 (12.41)	9 (11.39)	3 (8.82)	6 (13.33)	7 (17.5)	5 (11.11)	5 (9.62)	5 (11.11)	6 (24)	3 (7.5)
I1-M253	1 (0.31)	0 (0)	1 (0.46)	1 (0.73)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.22)	0 (0)	0 (0)	0 (0)	0 (0)
I1a-L22	3 (0.92)	1 (0.91)	2 (0.93)	1 (0.73)	1 (1.27)	0 (0)	1 (2.22)	0 (0)	1 (2.22)	0 (0)	0 (0)	0 (0)	1 (2.5)
I2-P215	7 (2.15)	5 (4.55)	2 (0.93)	2 (1.46)	0 (0)	0 (0)	0 (0)	1 (2.5)	1 (2.22)	0 (0)	1 (2.22)	1 (4)	3 (7.5)
I2a-M26	2 (0.61)	0 (0)	2 (0.93)	2 (1.46)	0 (0)	0 (0)	0 (0)	0 (0)	2 (4.44)	0 (0)	0 (0)	0 (0)	0 (0)
I2a-M223	6 (1.84)	3 (2.73)	3 (1.39)	1 (0.73)	2 (2.53)	1 (2.94)	1 (2.22)	0 (0)	0 (0)	1 (1.92)	3 (6.67)	0 (0)	0 (0)
J1-M267	16 (4.91)	3 (2.73)	13 (6.02)	5 (3.65)	8 (10.13)	3 (8.82)	5 (11.11)	1 (2.5)	2 (4.44)	2 (3.85)	1 (2.22)	0 (0)	2 (5)
J2-M172	2 (0.61)	2 (1.82)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.22)	0 (0)	1 (2.5)
J2a-M410	31 (9.51)	7 (6.36)	24 (11.11)	15 (10.95)	9 (11.39)	5 (14.71)	4 (8.89)	3 (7.5)	7 (15.56)	5 (9.62)	4 (8.89)	1 (4)	2 (5)
J2a-M67	10 (3.07)	4 (3.64)	6 (2.78)	4 (2.92)	2 (2.53)	1 (2.94)	1 (2.22)	1 (2.5)	2 (4.44)	1 (1.92)	2 (4.44)	1 (4)	1 (2.5)
J2a-M92	11 (3.37)	4 (3.64)	7 (3.24)	2 (1.46)	5 (6.33)	3 (8.82)	2 (4.44)	0 (0)	1 (2.22)	1 (1.92)	0 (0)	1 (4)	3 (7.5)
J2b-M12	11 (3.37)	5 (4.55)	6 (2.78)	5 (3.65)	1 (1.27)	1 (2.94)	0 (0)	2 (5)	1 (2.22)	2 (3.85)	0 (0)	2 (8)	3 (7.5)
L-M20	1 (0.31)	1 (0.91)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.5)
P-M45	1 (0.31)	0 (0)	1 (0.46)	1 (0.73)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
R1a-M17	17 (5.21)	7 (6.36)	10 (4.63)	8 (5.84)	2 (2.53)	1 (2.94)	1 (2.22)	1 (2.5)	4 (8.89)	3 (5.77)	3 (6.67)	1 (4)	3 (7.5)
R1b-M343	3 (0.92)	0 (0)	3 (1.39)	3 (2.19)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	2 (3.85)	0 (0)	0 (0)	0 (0)
R1b-M269	24 (7.36)	8 (7.27)	16 (7.41)	11 (8.03)	5 (6.33)	2 (5.88)	3 (6.67)	1 (2.5)	1 (2.22)	9 (17.31)	3 (6.67)	2 (8)	3 (7.5)
R1b-U106	12 (3.68)	1 (0.91)	11 (5.09)	10 (7.3)	1 (1.27)	1 (2.94)	0 (0)	2 (5)	2 (4.44)	6 (11.54)	0 (0)	0 (0)	1 (2.5)
R1b-P312	15 (4.6)	4 (3.64)	11 (5.09)	7 (5.11)	4 (5.06)	3 (8.82)	1 (2.22)	3 (7.5)	1 (2.22)	3 (5.77)	2 (4.44)	0 (0)	2 (5)
R1b-SRY2627	2 (0.61)	0 (0)	2 (0.93)	2 (1.46)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	1 (1.92)	0 (0)	0 (0)	0 (0)
R1b-U152	14 (4.29)	6 (5.45)	8 (3.7)	5 (3.65)	3 (3.8)	0 (0)	3 (6.67)	0 (0)	4 (8.89)	1 (1.92)	3 (6.67)	1 (4)	2 (5)
R1b-L2	17 (5.21)	7 (6.36)	10 (4.63)	4 (2.92)	6 (7.59)	5 (14.71)	1 (2.22)	3 (7.5)	1 (2.22)	0 (0)	6 (13.33)	0 (0)	1 (2.5)
R1b-L21	6 (1.84)	1 (0.91)	5 (2.31)	4 (2.92)	1 (1.27)	0 (0)	1 (2.22)	1 (2.5)	1 (2.22)	2 (3.85)	0 (0)	0 (0)	1 (2.5)
T1a-M70	8 (2.45)	2 (1.82)	6 (2.78)	5 (3.65)	1 (1.27)	0 (0)	1 (2.22)	3 (7.5)	2 (4.44)	0 (0)	0 (0)	2 (8)	0 (0)

mtDNA	SSI	ITAS	SIC	ESIC	WSIC	TP	AG	EN	RGSR	CT	CS	MT	LE
	313	115	198	116	82	40	42	40	39	37	40	36	39
L3	2 (0.64)	0 (0)	2 (1.01)	0 (0)	2 (2.44)	2 (5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M1	4 (1.28)	0 (0)	4 (2.02)	0 (0)	4 (4.88)	3 (7.5)	1 (2.38)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
C	1 (0.32)	0 (0)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.7)	0 (0)	0 (0)	0 (0)
D	1 (0.32)	0 (0)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.7)	0 (0)	0 (0)	0 (0)
A	2 (0.64)	0 (0)	2 (1.01)	1 (0.86)	1 (1.22)	0 (0)	1 (2.38)	1 (2.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
N1	9 (2.88)	2 (1.74)	7 (3.54)	7 (6.03)	0 (0)	0 (0)	0 (0)	4 (10)	1 (2.56)	2 (5.41)	0 (0)	1 (2.78)	1 (2.56)
I	8 (2.56)	5 (4.35)	3 (1.52)	1 (0.86)	2 (2.44)	2 (5)	0 (0)	0 (0)	0 (0)	1 (2.7)	0 (0)	1 (2.78)	4 (10.26)
W	3 (0.96)	1 (0.87)	2 (1.01)	1 (0.86)	1 (1.22)	0 (0)	1 (2.38)	0 (0)	0 (0)	1 (2.7)	0 (0)	0 (0)	1 (2.56)
X	14 (4.47)	7 (6.09)	7 (3.54)	4 (3.45)	3 (3.66)	3 (7.5)	0 (0)	0 (0)	3 (7.69)	1 (2.7)	1 (2.5)	3 (8.33)	3 (7.69)
R	1 (0.32)	1 (0.87)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.78)	0 (0)
R0	3 (0.96)	2 (1.74)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.56)	0 (0)	1 (2.5)	1 (2.78)	0 (0)
HV	15 (4.79)	4 (3.48)	11 (5.56)	9 (7.76)	2 (2.44)	1 (2.5)	1 (2.38)	5 (12.5)	2 (5.13)	2 (5.41)	0 (0)	2 (5.56)	2 (5.13)
H	43 (13.74)	14 (12.17)	29 (14.65)	13 (11.21)	16 (19.51)	4 (10)	12 (28.57)	2 (5)	7 (17.95)	4 (10.81)	4 (10)	4 (11.11)	6 (15.38)
H1	34 (10.86)	12 (10.43)	22 (11.11)	13 (11.21)	9 (10.98)	5 (12.5)	4 (9.52)	2 (5)	7 (17.95)	4 (10.81)	4 (10)	5 (13.89)	3 (7.69)
H2	6 (1.92)	1 (0.87)	5 (2.53)	3 (2.59)	2 (2.44)	1 (2.5)	1 (2.38)	1 (2.5)	0 (0)	2 (5.41)	1 (2.5)	0 (0)	0 (0)
H3	8 (2.56)	1 (0.87)	7 (3.54)	3 (2.59)	4 (4.88)	2 (5)	2 (4.76)	0 (0)	1 (2.56)	2 (5.41)	1 (2.5)	0 (0)	0 (0)
H4	2 (0.64)	0 (0)	2 (1.01)	2 (1.72)	0 (0)	0 (0)	0 (0)	2 (5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
H5	10 (3.19)	4 (3.48)	6 (3.03)	4 (3.45)	2 (2.44)	1 (2.5)	1 (2.38)	3 (7.5)	1 (2.56)	0 (0)	3 (7.5)	0 (0)	1 (2.56)
H6	1 (0.32)	0 (0)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
H7	5 (1.6)	1 (0.87)	4 (2.02)	1 (0.86)	3 (3.66)	0 (0)	3 (7.14)	0 (0)	1 (2.56)	0 (0)	0 (0)	1 (2.78)	0 (0)
H8	4 (1.28)	1 (0.87)	3 (1.52)	3 (2.59)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	2 (5.41)	0 (0)	1 (2.78)	0 (0)
H12	1 (0.32)	1 (0.87)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	0 (0)
H13	5 (1.6)	2 (1.74)	3 (1.52)	1 (0.86)	2 (2.44)	0 (0)	2 (4.76)	1 (2.5)	0 (0)	0 (0)	1 (2.5)	1 (2.78)	0 (0)
V	5 (1.6)	3 (2.61)	2 (1.01)	1 (0.86)	1 (1.22)	1 (2.5)	0 (0)	1 (2.5)	0 (0)	0 (0)	2 (5)	1 (2.78)	0 (0)
T	2 (0.64)	0 (0)	2 (1.01)	1 (0.86)	1 (1.22)	1 (2.5)	0 (0)	0 (0)	0 (0)	1 (2.7)	0 (0)	0 (0)	0 (0)
T1	11 (3.51)	4 (3.48)	7 (3.54)	6 (5.17)	1 (1.22)	1 (2.5)	0 (0)	0 (0)	4 (10.26)	2 (5.41)	1 (2.5)	2 (5.56)	1 (2.56)
T2	28 (8.95)	12 (10.43)	16 (8.08)	12 (10.34)	4 (4.88)	2 (5)	2 (4.76)	6 (15)	3 (7.69)	3 (8.11)	3 (7.5)	2 (5.56)	7 (17.95)
J1	16 (5.11)	5 (4.35)	11 (5.56)	8 (6.9)	3 (3.66)	2 (5)	1 (2.38)	4 (10)	2 (5.13)	2 (5.41)	2 (5)	2 (5.56)	1 (2.56)
J2	15 (4.79)	5 (4.35)	10 (5.05)	6 (5.17)	4 (4.88)	3 (7.5)	1 (2.38)	1 (2.5)	4 (10.26)	1 (2.7)	2 (5)	2 (5.56)	1 (2.56)
U	2 (0.64)	1 (0.87)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.56)	0 (0)	0 (0)	0 (0)	1 (2.56)
U1	8 (2.56)	4 (3.48)	4 (2.02)	1 (0.86)	3 (3.66)	0 (0)	3 (7.14)	0 (0)	1 (2.56)	0 (0)	2 (5)	0 (0)	2 (5.13)
U2	4 (1.28)	1 (0.87)	3 (1.52)	1 (0.86)	2 (2.44)	0 (0)	2 (4.76)	1 (2.5)	0 (0)	0 (0)	0 (0)	1 (2.78)	0 (0)
U3	6 (1.92)	5 (4.35)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	0 (0)	4 (10)	0 (0)	1 (2.56)
U4	1 (0.32)	0 (0)	1 (0.51)	1 (0.86)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.7)	0 (0)	0 (0)	0 (0)
U5	11 (3.51)	4 (3.48)	7 (3.54)	2 (1.72)	5 (6.1)	3 (7.5)	2 (4.76)	0 (0)	0 (0)	2 (5.41)	2 (5)	1 (2.78)	1 (2.56)
U6	2 (0.64)	2 (1.74)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.5)	0 (0)	1 (2.56)
U7	4 (1.28)	2 (1.74)	2 (1.01)	0 (0)	2 (2.44)	2 (5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (5.56)	0 (0)
U8	3 (0.96)	3 (2.61)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (5)	0 (0)	1 (2.56)
K	4 (1.28)	2 (1.74)	2 (1.01)	2 (1.72)	0 (0)	0 (0)	0 (0)	2 (5)	0 (0)	0 (0)	1 (2.5)	0 (0)	1 (2.56)
K1	9 (2.88)	3 (2.61)	6 (3.03)	3 (2.59)	3 (3.66)	1 (2.5)	2 (4.76)	1 (2.5)	0 (0)	2 (5.41)	1 (2.5)	2 (5.56)	0 (0)

Abbreviations: SSI, Sicily and South-Italy; ITAS, South-Italy; SIC, Sicily; ESIC, East Sicily; WSIC, West Sicily; TP, Trapani; AG, Agrigento; EN, Enna; RGSR, Ragusa-Siracusa; CT, Catania; CS, Cosenza, MT, Matera; LE, Lecce

Table S3. Y-Chromosome STRs haplotypes and SNPs analysis results for the newly-typed samples of the present study (N=119). This table will be provided in the online version of the manuscript once published.

Table S4. Analyses of the molecular variance (AMOVA). Apportionment of the variance are in percentage (%) and based on both haplogroup frequencies (SNPs) and haplotype data (STRs or sequences).

Y-chromosome Grouping	N° of Groups	N° of Pop	Proportion of variation (%)					
			Among Groups		Among population Within Group		Within Population	
			Y-SNPs	Y-STRs	Y-SNPs	Y-STRs	Y-SNPs	Y-STRs
ALL	1	8			0.16	0.36	99.84	99.64
ITAS/SIC	2	8	-0.13	-0.42	0.23	0.58	99.90	99.85
ITAS/ESIC/WSIC	3	8	-0.09	0.10	0.23	0.28	99.86	99.62
ESIC/WSIC	2	5	0.01	0.55	0.20	0.54	99.79	98.91
Mitochondrial DNA Grouping	N° of Groups	N° of Pop	Proportion of variation (%)					
			Among Groups		Among population Within Group		Within Population	
			Mt-SNPs	Mt-SEQs	Mt-SNPs	Mt-SEQs	Mt-SNPs	Mt-SEQs
ALL	1	8			0.41	1.02**	99.59	98.98**
ITAS/SIC	2	8	-0.40	-0.51	0.62	1.29***	99.78	99.22**
ITAS/ESIC/WSIC	3	8	-0.11	-0.58	0.49	1.45***	99.62	99.13**
ESIC/WSIC	2	5	0.11	-0.46	0.98*	1.71**	98.81*	98.75**

ITAS (CS, MT, LE); **SIC** (TP, AG, EN, RGSR, CT); **ESIC** (EN, RGSR, CT); **WSIC** (AG, TP)
******* P-value < 0,001; ****** P-value < 0,01; ***** P-value < 0,05

Table S5. Diversity indices computed for the different sampling points. Standard diversity parameters were calculated for both Y-chromosome and mtDNA based on haplotype/sequence data and haplogroup frequencies

Y-Chromosome		**STRs**			**SNPs**
Population	**N**	**STR div**	**MNPD**	**Nucleotyde div**	**Haplogroup div**
ALL	326	0.9997 +/- 0.0003	9.8284 +/- 4.5107	0.6143 +/- 0.3118	0.9438 +/- 0.0041
ITAS	110	0.9997 +/- 0.0013	9.1465 +/- 4.2394	0.6098 +/- 0.3130	0.9441 +/- 0.0077
SIC	216	0.9994 +/- 0.0005	9.8941 +/- 4.5447	0.6184 +/- 0.3143	0.9441 +/- 0.0052
WSIC	79	0.9984 +/- 0.0025	10.0951 +/- 4.6620	0.6309 +/- 0.3229	0.9400 +/- 0.0099
ESIC	137	0.9992 +/- 0.0010	9.2676 +/- 4.2852	0.6178 +/- 0.3163	0.9459 +/- 0.0068
TP	34	0.9929 +/- 0.0099	9.9840 +/- 4.6804	0.6240 +/- 0.3253	0.9234 +/- 0.0194
AG	45	0.9990 +/- 0.0050	10.1434 +/- 4.7210	0.6340 +/- 0.3276	0.9535 +/- 0.0142
EN	40	0.9987 +/- 0.0060	9.7269 +/- 4.5504	0.6079 +/- 0.3159	0.9487 +/- 0.0182
RGSR	45	0.9980 +/- 0.0052	9.3919 +/- 4.3936	0.6261 +/- 0.3252	0.9475 +/- 0.0157
CT	52	0.9977 +/- 0.0043	9.2232 +/- 4.3092	0.6149 +/- 0.3188	0.9367 +/- 0.0162
CS	45	1.0000 +/- 0.0047	9.6586 +/- 4.5098	0.6037 +/- 0.3129	0.9384 +/- 0.0132
MT	25	1.0000 +/- 0.0113	10.0133 +/- 4.7370	0.6258 +/- 0.3299	0.9333 +/- 0.0354
LE	40	0.9974 +/- 0.0063	9.0949 +/- 4.2744	0.6063 +/- 0.3166	0.9526 +/- 0.0151
mtDNA		**SEQs**			**SNPs**
Population	**N**	**SEQ div**	**MNPD**	**Nucleotyde div**	**Haplogroup div**
ALL	313	0.9899 +/- 0.0026	6.8780 +/- 3.2453	0.0102 +/- 0.0053	0.9452 +/- 0.0051
ITAS	115	0.9969 +/- 0.0017	7.3143 +/- 3.4484	0.0109 +/- 0.0057	0.9504 +/- 0.0079
SIC	198	0.9828 +/- 0.0049	6.6343 +/- 3.1454	0.0099 +/- 0.0052	0.9439 +/- 0.0068
WSIC	82	0.9759 +/- 0.0098	6.3359 +/- 3.0349	0.0094 +/- 0.0050	0.9359 +/- 0.0148
ESIC	116	0.9862 +/- 0.0050	6.8311 +/- 3.2397	0.0102 +/- 0.0053	0.9454 +/- 0.0077
TP	40	0.9795 +/- 0.0109	7.2368 +/- 3.4623	0.0108 +/- 0.0057	0.9024 +/- 0.0345
AG	42	0.9640 +/- 0.0191	5.3917 +/- 2.6514	0.0080 +/- 0.0044	0.9024 +/- 0.0345
EN	40	0.9897 +/- 0.0076	6.8792 +/- 3.3059	0.0102 +/- 0.0055	0.9436 +/- 0.0170
RGSR	39	0.9798 +/- 0.0114	6.2853 +/- 3.0475	0.0093 +/- 0.0050	0.9163 +/- 0.0220
CT	37	0.9760 +/- 0.0182	7.0905 +/- 3.4042	0.0105 +/- 0.0056	0.9640 +/- 0.0130
CS	40	0.9949 +/- 0.0069	7.0106 +/- 3.3633	0.0104 +/- 0.0056	0.9654 +/- 0.0119
MT	36	0.9984 +/- 0.0070	7.3811 +/- 3.5339	0.0110 +/- 0.0058	0.9587 +/- 0.0160
LE	39	0.9946 +/- 0.0071	7.4118 +/- 3.5408	0.0110 +/- 0.0058	0.9325 +/- 0.0216

Table S6. Fisher exact test for Y-chromosome and mtDNA HG frequencies among the Mediterraean population groups

Y-chr	PopCod	Freq	C-M130	E-M96	E-M35	E-M78	E-M81	E-M123	F-M89	G-M201	I-M170	I-M26	I-M223	J-M304	J-M172	J-M410	J-M12	K-M9	R-M207	R-SRY10831.2	R-M343	R-M269	T-M70
	SSI	N	0	2	3	47	5	7	0	42	11	2	6	16	2	52	11	2	0	17	3	90	8
		%	0.00	0.61	0.92	14.42	1.53	2.15	0.00	12.88	3.37	0.61	1.84	4.91	0.61	15.95	3.37	0.61	0.00	5.21	0.92	27.61	2.45
		Pvalue								1.68E-04	8.44E-04			4.68E-02				1.43E-02					
	NCI	N	0	0	1	28	2	7	4	25	18	1	3	9	0	30	4	4	0	8	1	122	2
		%	0.00	0.00	0.37	10.41	0.74	2.60	1.49	9.29	6.69	0.37	1.12	3.35	0.00	11.15	1.49	1.49	0.00	2.97	0.37	45.35	0.74
		Pvalue							7.95E-03					2.05E-03								7.35E-10	
	IBE	N	0	0	0	2	3	0	0	2	3	8	0	2	0	6	2	2	0	2	0	75	4
		%	0.00	0.00	0.00	1.80	2.70	0.00	0.00	1.80	2.70	7.21	0.00	1.80	0.00	5.41	1.80	1.80	0.00	1.80	0.00	67.57	3.60
		Pvalue				3.04E-02						2.94E-07		4.42E-02								5.02E-18	
	GER	N	1	0	1	23	1	0	0	9	66	0	13	2	0	15	3	6	0	45	0	161	3
		%	0.29	0.00	0.29	6.59	0.29	0.00	0.00	2.58	18.91	0.00	3.72	0.57	0.00	4.30	0.86	1.72	0.00	12.89	0.00	46.13	0.86
		Pvalue					1.04E-03	7.31E-03		6.92E-03	6.65E-10		4.27E-03	1.90E-12		5.63E-08				1.05E-06		9.94E-15	
	BALK	N	0	0	0	30	0	2	0	13	47	0	4	6	1	14	8	2	0	26	0	30	2
		%	0.00	0.00	0.00	16.22	0.00	1.08	0.00	7.03	25.41	0.00	2.16	3.24	0.54	7.57	4.32	1.08	0.00	14.05	0.00	16.22	1.08
		Pvalue					4.41E-02				2.39E-11			3.97E-02						4.12E-04		2.86E-03	
	LEV	N	1	9	2	53	7	25	0	38	26	0	0	120	0	132	17	49	12	15	4	44	23
		%	0.17	1.56	0.35	9.19	1.21	4.33	0.00	6.59	4.51	0.00	0.00	20.80	0.00	22.88	2.95	8.49	2.08	2.60	0.69	7.63	3.99
		Pvalue					8.90E-03	8.44E-04			9.32E-05		2.60E-03	6.09E-26		8.40E-14		1.28E-12	9.99E-05	4.25E-04		5.69E-43	1.96E-02
	NAFR	N	0	8	0	6	46	0	0	0	0	0	0	23	0	5	0	1	1	1	0	11	0
		%	0.00	7.84	0.00	5.88	45.10	0.00	0.00	0.00	0.00	0.00	0.00	22.55	0.00	4.90	0.00	0.98	0.98	0.98	0.00	10.78	0.00
		Pvalue		4.74E-05			9.80E-48			3.31E-02	2.15E-03			5.84E-04								6.96E-04	

mtDNA	PopCod	Freq	L	M	N	N1	I	W	X	R*	R0	HV	H	V	T	J	U	K
	SSI	N	2	6	2	9	8	3	14	2	2	15	119	5	41	31	41	13
		%	0.64	1.92	0.64	2.88	2.56	0.96	4.47	0.64	0.64	4.79	38.02	1.60	13.10	9.90	13.10	4.15
		Pvalue	2.19E-06															
	NCI	N	0	2	2	5	3	2	3	0	3	9	105	3	24	19	28	21
		%	0.00	0.87	0.87	2.18	1.31	0.87	1.31	0.00	1.31	3.93	45.85	1.31	10.48	8.30	12.23	9.17
		Pvalue	3.28E-06										4.81E-02					
	IBE	N	1	0	0	0	1	0	1	0	1	3	66	1	8	8	19	7
		%	0.86	0.00	0.00	0.00	0.86	0.00	0.86	0.00	0.86	2.59	56.90	0.86	6.90	6.90	16.38	6.03
		Pvalue											1.08E-04					
	BALK	N	1	6	0	4	14	15	15	2	9	19	251	11	40	53	100	26
		%	0.18	1.06	0.00	0.71	2.47	2.65	2.65	0.35	1.59	3.36	44.35	1.94	7.07	9.36	17.67	4.59
		Pvalue	5.10E-16					2.84E-02					3.00E-04					
	LEV	N	18	13	8	15	12	6	10	6	20	24	182	8	56	39	78	54
		%	3.28	2.37	1.46	2.73	2.19	1.09	1.82	1.09	3.64	4.37	33.15	1.46	10.20	7.10	14.21	9.84
		Pvalue	8.94E-03		3.33E-02													9.29E-04
	NAFR	N	122	26	0	5	3	3	9	0	21	37	115	9	38	34	68	18
		%	24.02	5.12	0.00	0.98	0.59	0.59	1.77	0.00	4.13	7.28	22.64	1.77	7.48	6.69	13.39	3.54
		Pvalue	1.00E-60	2.26E-04									2.90E-13					

Abbreviations: SSI, Sicily and South-Italy; NCI, North-Central Italy; IBE, Iberian Peninsula; GER, Germany; BALK, Balkan Peninsula; LEV, Levant; NAFR, North-Africa

All non-significant p-values after Bonferroni correction have been removed

1.4. Comments

1.4.1 Clines or discontinuous structures in the Italian genetic landscape

The presence of clinal or discontinuous patterns of genetic variation within the Italian Peninsula has been one of the main issues of several genetic studies since the pioneering work by Cavalli-Sforza, Menozzi and Piazza in 1994. Consistently with the first evidences provided by classical genetic markers (Piazza et al., 1988, Cavalli-Sforza et al., 1994), most of the subsequent studies based on uniparental molecular markers (Barbujani et al., 1995, Capelli et al, 2007, Brisighelli et al., 2012) have mainly described the Italian genetic landscape as characterised by the outlier position of Sardinia and the presence of a North-South major cline within the mainland, moreover suggesting the higher similarity of Southern Italian populations with South-Eastern European and Mediterranean groups (Piazza et al., 1988, Capelli et al., 2007). While the South-North clinal distribution pattern of genetic variation was initially thought to be the signature of Greek domination in Southern Italy (Cavalli-Sforza et al., 1994, Di Giacomo et al., 2003), it has been subsequently explained in favor of the demic model of Neolithic diffusion into Europe (Capelli et al., 2007). Latitude-related gradients of specific genetic lineages – i.e. Y-haplogroups R1*(xR1a1), J2 and E3b1 – have however suggested (at least from a paternal perspective) the presence of different Neolithic/Mesolithic contributions between the Northern and the Southern parts of the Peninsula (Capelli et al., 2007), indicating that population replacement by new incoming farmers was neither complete nor homogeneous along the whole Italy. More recently, genome-wide studies have also postulated some genetic discontinuities between Southern Italy and both European (Lao et al., 2008, Nelis et al., 2009) and North-Central Italian (Di Gaetano et al., 2012) populations, thus pointing out a more complex picture of the Italian genetic structuring than previously thought.

In our analyses, if Sardinia attested again to be a genetic outlier for both Y-chromosome and mtDNA uniparental markers and the maternal genetic variation confirmed to be homogeneously distributed within the mainland (*Paper 1*), on the other hand Y-chromosomal genetic diversity in Italy revealed to be not clinal but highly structured in three well-detached geographic clusters of populations: namely Sardinia (SAR), North-Western Italy (NWI) and South-Eastern Italy (SEI) (Figure 1, *Paper 1*). Most interestingly, NWI and SEI appeared to be separated not by a latitudinal demarcation line (as it could be expected if considering some preliminary suggestions of previous studies) but rather by a longitudinal "belt" (Figure 1a, *Paper 1*), that particularly outlines a shared genetic background between the South and the Adriatic coast on one side, and between the North and Tuscany on the other side. Previous observation that samples located to the east of the Apennines (Adriatic side) tended to cluster with southern Italian samples also at higher latitudes than those settled on the western (Tyrrhenian) side (Capelli et al., 2007), could have at least partially suggested and anticipated the NW-SE discontinuity shown by our results, although this hypothesis was not further discussed by previous authors. The differences between our results and those obtained in earlier Y-chromosome-based studies (Di Giacomo et al., 2003 and Capelli et al., 2007) could be explained if considering a) the increased sampling coverage (coupled with the detailed sampling strategy), b) the higher genotyping resolution for the uniparental lineages analysed, and c) the different methodological approach exploited. These improvements might have allowed us to unmask genetic patterns not fully detected in the previous analyses (Capelli et al 2007), thus underlying the importance of a well-conceived research design to achieve a fine-grained dissection of population genetic structure.

Besides the discontinuous genetic structure observed, the identified NWI and SEI Italian clusters appeared also characterised by different distributions of genetic variance (*Paper 1*) and revealed different genetic relationships with other comparison populations once contextualised within the Euro-Mediterranean genetic landscape (*Paper 2*). A higher degree of inter-population variability was observed for the Northern-Western Italian populations; on the contrary Southern Italy revealed higher homogeneity among populations, moreover showing a shared paternal genetic background with the Balkan Peninsula (Figure 4 and Figure S3, *Paper 2*). When compared with European and the Mediterranean populations, the NW-SE pattern observed in continental Italy appeared extended to the whole Mediterranean Basin (Figure 1, *Paper 2*), particularly suggesting differential Eastern (Balkans, Levant) and Western (Iberia, Germany) contributions to the SEI and NWI genetic pools respectively (Figure 2 and Figure 4, *Paper 2*). On the whole, the resulting picture is quite consistent with the most recent archaeological evidences on the Neolithisation of the Italian Peninsula, that suggest two parallel but opposite ways of diffusion along the Adriatic and the Tyrrhenian coasts (see Figure 1.1.1.2.3) and postulate higher degrees of local heterogeneity and regional complexity for Northern Italian sites compared to the higher homogeneity instead observed in central and southern Italy.

1.4.2 Origin and structuring of the Italian Y-chromosomal gene pool

The Y-chromosomal discontinuous genetic structure observed within the Italian Peninsula appeared particularly related to the differential presence of specific lineages within the three detected Italian geographic groups (I2a-M26 for SAR; G2a-P15 – besides the classical E-V13 and J2a-M410 – for SEI; and R-U152 and R-L2 for NWI; Figure 1.4.2.1).

Although we are aware that the whole history of a population cannot be represented with only individual haplogroups, population groups differing in their HG composition can actually retain specific signatures of past populations processes. In fact, during the formation and differentiation of human populations, haplotype clusters within particular lineages could have emerged in response of specific demographic events. As a consequence population-specific cluster of haplotypes are more likely to contain traces of the past genetic history of a population (Balanosvky et al., 2011).

In our analyses, the (STRs) variation within those haplogroups that were found to significantly differentiate the Italian specific population groups (NWI, SEI and SAR) was explored (by means of DAPC analysis) with the aim to define population-specific clusters of haplotypes (Table 1, *Paper 1* and Figure 1.4.2.2).

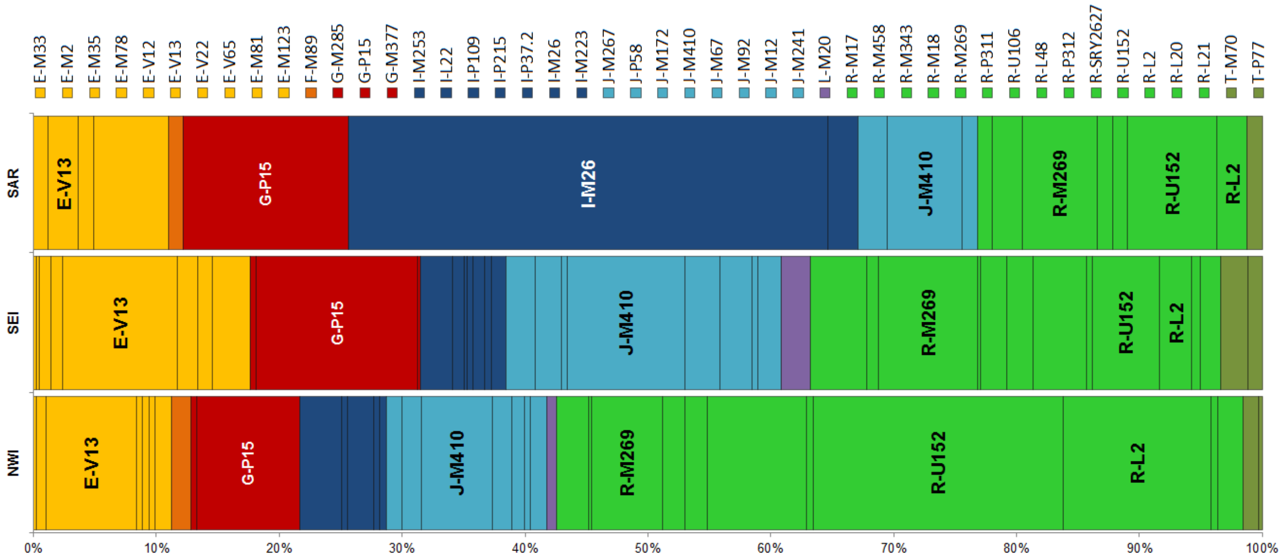
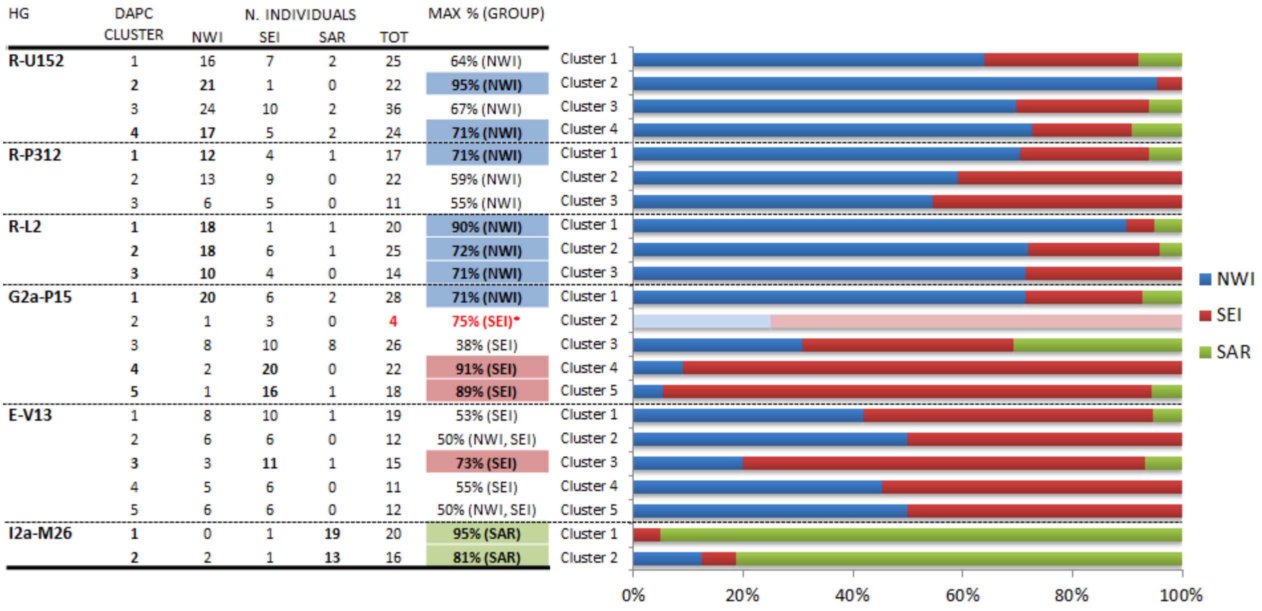


Figure 1.4.2.1. Y-chromosome haplogroups composition within the three Italian specific clusters identified by sPCA analysis (*Paper 1*). The most important lineages are labeled according to Karafet et al., 2008.



HG	DAPC CLUSTER	N. INDIVIDUALS NWI	SEI	SAR	TOT	MAX % (GROUP)
R-U152	1	16	7	2	25	64% (NWI)
	2	21	1	0	22	95% (NWI)
	3	24	10	2	36	67% (NWI)
	4	17	5	2	24	71% (NWI)
R-P312	1	12	4	1	17	71% (NWI)
	2	13	9	0	22	59% (NWI)
	3	6	5	0	11	55% (NWI)
R-L2	1	18	1	1	20	90% (NWI)
	2	18	6	1	25	72% (NWI)
	3	10	4	0	14	71% (NWI)
G2a-P15	1	20	6	2	28	71% (NWI)
	2	1	3	0	4	75% (SEI)*
	3	8	10	8	26	38% (SEI)
	4	2	20	0	22	91% (SEI)
	5	1	16	1	18	89% (SEI)
E-V13	1	8	10	1	19	53% (SEI)
	2	6	6	0	12	50% (NWI, SEI)
	3	3	11	1	15	73% (SEI)
	4	5	6	0	11	55% (SEI)
	5	6	6	0	12	50% (NWI, SEI)
I2a-M26	1	0	1	19	20	95% (SAR)
	2	2	1	13	16	81% (SAR)

Figure 1.4.2.2. Frequencies of Y-Chromosome DAPC main clusters for each Italian sPCA-identified group. The absolute number of individuals and the maximum frequency for each DAPC-cluster are detailed in the table on the left. The relative (%) proportion of each DAPC-cluster in NWI (blue), SEI (red) and SAR (green) are represented by barplot on the right.

Implications for the Sardinian genetic pool

Consistently with previous results, the outlier position of Sardinia (SAR) is mainly ascribable to Y-chromosome haplogroup I-M26 (Figure S4 and Table 1, *Paper 1*). Despite being generally present at low frequencies in most of the European continent – with the only exception of moderate occurrence in Iberia – in our dataset the I-M26 haplogroup was found at significantly higher frequencies in SAR (39%, p-value 3.06 E-30), being instead significantly under-represented, not to say almost completely absent, in continental Italy (NWI: 0.52%, p-value 1.93 E-05 and SEI: 0.48% p-value 1.92 E-06). Its time estimate at around 5,153 YBP (Table 2, *Paper 1*), consistently with archeological data, points to an expansion within the island compatible with the arrival and the achievement of Neolithic technologies. However, when compared to the Iberian I-M26 haplotypes – being the Franco-Cantabrian area the postulated place of origin for this haplogroup (Lopez-Parra et al., 2009) – Sardinian haplotypes appeared clearly distinguishable from Iberian ones (Figure S5, *Paper 1*). This presumtively suggests a more ancient presence of I-M26 haplotypes within the island, possibly resulting from past founding events. These findings are in agreement with the hypothesis of a largely pre-Neolithic settlement of Sardinia, with little subsequent gene flows from external populations (Contu et al. 2008). This scenario is also consistent with that recently proposed by Pala and collaborators (2009) for the female Sardinian-specific counterpart, the mtDNA haplogroup U5b3a1, thus suggesting an intriguing case of parallel founding events for both maternal and paternal lineages within the European context (Pala et al., 2009). Unlike the majority of modern mtDNA haplogroups, which are supposed to have expanded from the Franco-Cantabria refuge after the LGM, the most likely homeland for the maternal U5b3 haplogroup was postulated in the Italian Peninsula. From there, this lineage is thought to have then proceeded along the Tyrrhenian coast moving westwards through Provence (southern France), the northwards expansion being instead limited by the presence of the Alps. From the Mediterranean coast of southern France, the U5b3a1 sub-lineage seems to have subsequently expanded in Sardinia sometime between 7,000-9,000 YBP, where it gave rise by founder event to the mtDNA U5b3a1a clade, that today distinctively marks the present-day people of the island (Pala et al., 2009).

Implications for the mainland genetic interactions

Within the mainland, two main R-M269 sub-lineages, namely R-U152 and R-L2, were found at significantly higher frequencies in NWI (21%, p-value 3.16 E-09 and 12% p-value 1.57 E-06, respectively) compared to the lower values instead observed in SEI (5.5%, p-value 3.09 E-07 and 2.6%, p-value 1.26 E-04, respectively). DAPC analysis confirmed the specificity of these haplogroups for the NWI Italian group, by showing a high percentage (5 out of 7, accounting for more than 70%) of haplotype clusters that specifically belonged to NWI group (Figure 1.4.2.2). Consistently with the results of previous studies (Capelli et al., 2007), R-M269 sub-lineages showed decreasing latitude-related gradients of frequency moving southwards. On the other hand, opposite frequency clines have been observed for haplogroups G2a-P15 and E-V13, with higher frequencies in SEI (13.1% and 9.3%), despite being well-presented also in NWI (8.4% and 7.3%).

If the Neolithic revolution is the most probable candidate for explaining the paternal structure observed within the Italian Peninsula, the age estimates of NWI- or SEI-specific haplotype clusters (Table 2, *Paper 1*) and their differential relationships with other European populations (Figure S7, *Paper 1*), pointed to the relevant impact also of subsequent population events on the observed genetic pattern, thus suggesting that the spatial structure of Italian genetic variation was not fixed in the Neolithic, but continued to be reshaped by post-Neolithic demographic processes. Ages of NWI clusters within haplogroups R-U152 and R-L2 (ranging from 3,714 to 4,979 YBP; Table 2, *Paper 1*) and their relationships with Germanic and Iberian haplotypes actually agree with post-Neolithic incomings from Central-Europe. On the contrary, the age of SEI-specific cluster within haplogroup E-V13 (cluster 2; 3,488 YBP) suggests migrations processes from the Balkans referable to the Bronze Age.

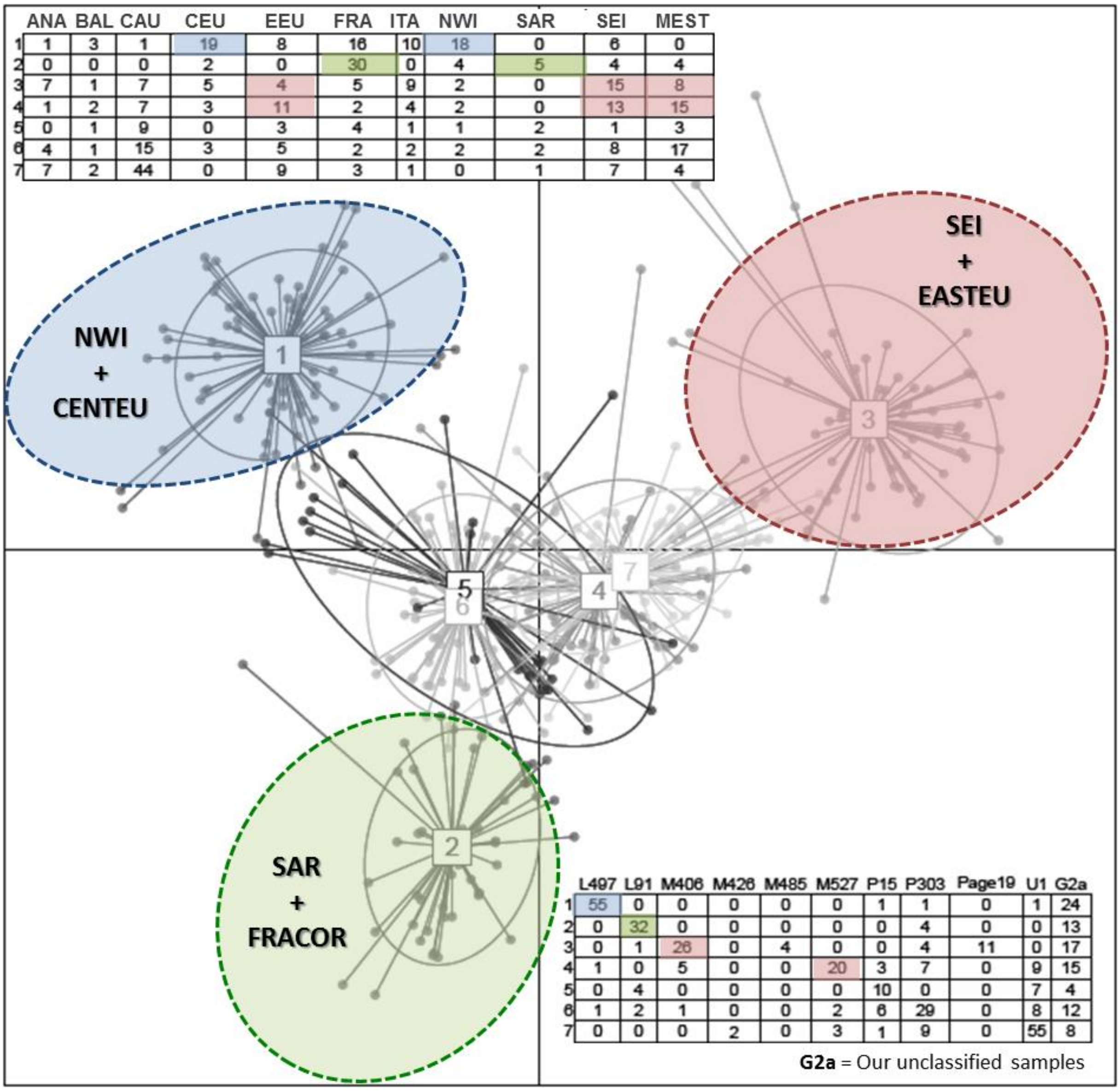
Implications for the Sicilian genetic pool

After the collapse of the Late Bronze Age societies (approximately 3,200 YBP) the populations of the Mediterranean Basin experienced different waves of invasion - particularly by the Greeks of the Aegean Sea and the Phoenicians of the Levant. These processes, unlikely to have completely masked the previous Neolithic structures, could however have left significant genetic signatures in the gene pool of present-day populations. While the exact contributions of Phoenicians and North-African (Arabs) populations to the current Sicilian and Southern Italian (SSI) paternal genetic pool remain still debated (though supposed to be low, *Paper 2*), our results outlined the impressive demographic and cultural impacts of Greek dominion during the 8th and 9th centuries BC (*Paper 2*). Accordingly with previous studies (Di Gaetano et al., 2009), the high frequency of Balkan E-V13 haplotypes observed in our SSI dataset (one third of which belonged to the southern Balkan Modal Haplotype) as well as the time estimate obtained for this haplogroup in our population (*Paper 2*), confirm the introduction of at least some of the current paternal lineages in Sicily during the migration processes referable to the Greek colonisation of Southern Italy. Contrarily to what postulated in previous studies (Di Gaetano et al., 2009), these migration processes seem however to not have resulted in an east-west genetic differentiation within Sicily. On the contrary, our analyses of both maternal and paternal genetic markers revealed a homogeneous genetic composition not only within Sicily, but also between Sicily and Southern Italy (*Paper 2*), which is in accordance with the largely shared genetic histories of the Southern Italian populations during Neolithic and post-Neolithic times.

Implications for the Italian genetic layers

Within the paternal genetic landscape of both Sardinia and the mainland, particularly interesting revealed to be the case of G2a-P15 lineage. This haplogroup, despite being present in the whole region, showed different geographic-specific patterns of clustering within Italy - cluster 1 in NWI, clusters 4 and 5 in SEI, and cluster 3 in both mainland and Sardinia (Figure 2, *Paper 1*). In addition, when compared with other European populations, the Italian-specific clusters exhibited different patterns of genetic similarities. The NWI-cluster appeared mostly related to Germans, while SEI-clusters showed affinities mainly with the Balkans (Figure S7, *Paper 1*). In accordance with our results, a recent G-haplogroup survey by Rootsi et al. 2012 revealed the presence of different G2a sub-clades within the Italian Peninsula. In particular the occurrence in Italy of G2a-L497 sub-lineages has been associated to migratory flows from Central Europe (Germany), while the presence of G2a-M406 and G2a-M527 derived-clades appeared to be related to events of Neolithic and post-Neolithic colonisation from the Balkans (Rootsi et al. 2012). To explore more in details the G2a clustering pattern found in our Italian dataset and to further investigate the origins of such geographic differentiation, we incorporated into the DAPC analysis of our Italian G2a haplotypes, also the additional European G2a samples extracted from Rootsi et al. (2012) dataset - for which a deeper typing for G2a sub-lineages was available. The resulting DAPC plot (Figure 1.4.2.3) shows clearly the connection between NWI and the Central Europe (particularly Germany), while SEI reveals higher affinities with South-Eastern European populations. Furthermore the main lineages associated with this pattern confirm to be G2a-L497 on one side, and G2a-M406 and G-M527 on the other side.

Interestingly, Sardinian haplotypes were found to cluster mainly with the French/Corsican ones and seem particularly associated with the G2a-L91 sub-lineage. At this respect, it is notable that Otzi, the 5300-year-old Alpine mummy, was derived for this Y-chromosome L91-SNP. Analogously, it is noteworthy that whole-genome sequencing of the Tyrolean Iceman (Keller et al., 2012) revealed its closest affinity to modern Sardinians. This presumptively suggests that at least continental Europeans living 5,300 YBP were more similar to the current population of the island than to the modern Northern Italian groups. In line with this view, a recent genome-wide investigation of the Italian genetic structure (Di Gaetano et al., 2012) highlighted the Sardinian population to be characterised by a proportion of genetic ancestry (Figure 1.1.2.2.5) shared with other European and non-European populations, but highest in frequency in the Italian island.



	ANA	BAL	CAU	CEU	EEU	FRA	ITA	NWI	SAR	SEI	MEST
1	1	3	1	19	8	16	10	18	0	6	0
2	0	0	0	2	0	30	0	4	5	4	4
3	7	1	7	5	4	5	9	2	0	15	8
4	1	2	7	3	11	2	4	2	0	13	15
5	0	1	9	0	3	4	1	1	2	1	3
6	4	1	15	3	5	2	2	2	2	8	17
7	7	2	44	0	9	3	1	0	1	7	4

	L497	L91	M406	M426	M485	M527	P15	P303	Page19	U1	G2a
1	55	0	0	0	0	0	1	1	0	1	24
2	0	32	0	0	0	0	0	4	0	0	13
3	0	1	26	0	4	0	0	4	11	0	17
4	1	0	5	0	0	20	3	7	0	9	15
5	0	4	0	0	0	0	10	0	0	7	4
6	1	2	1	0	0	2	6	29	0	8	12
7	0	0	0	2	0	3	1	9	0	55	8

G2a = Our unclassified samples

Figure 1.4.2.3. Discriminant Analysis of Principal Components (DAPC) for G2a-P15 haplotypes. Haplotypes from the sPCA-identified Italian groups (NWI, SEI, SAR) were integrated and compared with those by Rootsi et al., 2012 from Anatolia (ANA), the Balkans (BAL), Caucasus (CAU), Central Europe (CEU), Eastern Europe (EEU), France and Corsica (FRA), Italy (ITA) and Middle East (MEST). The table on the top left shows the number of haplotypes in each of the seven DAPC-identified clusters and their geographical distribution in the different groups considered. The table on the bottom right details for each DAPC cluster the number of haplotypes belonging to each G2a-sublineage (as defined by Rootsi et al., 2012). Our sample - for which G2a-sublineages were not genotyped - were generally referred to as G2a without any further specification (last column of the table).

On the whole these findings suggest the presence of a common genetic background to all the continental European populations, that was probably better preserved in Sardinia due to its condition of higher isolation or impermeability relative to the subsequent Neolithic and Post-

Neolithic migration processes which instead occurred in the Italian Peninsula – as well as throughout Europe. This hypothesis could also explain the reason why the most frequent G2a-cluster found in our Italian dataset – which is present in both NWI and SEI and encompasses almost all the SAR individuals (cluster 3) – actually resulted in more ancient time estimate compared to the others G2a-Italian clusters and closer to the age of the whole haplogroup (Table 2, *Paper 1*).

1.4.3 Contrasting maternal and paternal histories

The clear pattern of structure observed both in Italy (*Paper 1*) and the Mediterranean (*Paper 2*) for Y-chromosome, contrasts with the higher homogeneity found for the mtDNA, that showed no significant internal genetic subdivision or geography-related clustering patterns. Once again, the only exception is the different maternal genetic history suggested for the Sardinian population compared to the mainland (Figure 3 and Table S4, *Paper 1*). Differences between mtDNA and Y-chromosome patterns may be explained if considering different demographic and historical dynamics for females than males. Sex-biased migration events, implying the differential income of male genes into a population, have been recently demonstrated for the Neolithisation of Southern Europe (Lacan et al., 2011a; Lacan et al., 2011b) as well as for some post-Neolithic processes such as the Greek colonisation of the Italian Peninsula (Pesando, 2005). Furthermore, the documented increasing of patrilocality in post-marital residential systems with the transition from hunter-gatherers to farmers societies (Rastiero et al. 2012; Heyer et al., 2012) could provide an additional – but not necessarily alternative – reason why matrilines tend to be more stable over the time than patrilines, thus resulting in more ancient time estimates for mtDNA than Y-chromosome haplogroups (Table 2, *Paper 1* and Table 1, *Paper 2*). Differential sex-specific demographic behaviours could therefore have affected differently the male and female migration patterns during and after the Neolithic transition in Europe. Accordingly, ages of Y-chromosome and mtDNA haplogroups or haplotype clusters point to significantly different historical periods (Table 2, *Paper 1*), therefore highlighting different phases of the making of the genetic structure of Italy within the Mediterranean and European genetic landscape.

PART 2

Insights into the Italian micro-geographic genetic variability: two case studies from ethno-linguistic and socio-economic enclaves of Italy

2.1. Introduction

The genetic structure of current human populations is deeply influenced by different historical, geographic, cultural and bio-demographic processes that affected the human genetic history at global, regional and local scales (Beleza et al., 2005; Sanchez-Faddeev et al., 2013). Investigating and understanding the effects of these events allows us to interpret the present-day diversity patterns of genetic variation by virtue of the past demographic history, and consequently to get new insights into the population processes - such as migration, admixture and/or divergence - that addressed the human population evolution over time (Jobling et al., 2013).

Historical processes and geographical factors proved to have deeply affected the broad-scale patterns of genetic variation in the current European populations. On a continental perspective in fact, a strong correlation between geographic and genetic distances has been extensively reported in literature (Lao et al., 2008; Novembre et al., 2008, Nelis et al., 2009). As revised in the first part of this thesis, the current distribution pattern of genetic variation within Europe – as well as within most of European populations - has been particularly interpreted in the light of the major prehistoric demographic events that affected the peopling history of the continent over time, in particular: i) the first Palaeolithic colonisation by AMHs, ii) the post-glacial re-expansion from Southern-European refuges, and iii) the Neolithic diffusion of agriculture from Near East. However, regional population structures might have been also affected by additional and different processes – such as more recent historical dynamics and/or socio-cultural and environmental factors – which may have led to further population differentiations and/or stratifications.

Genetic differences currently existing among human groups can therefore be ascribed both i) to the broad-scale impact of pre-historical and historical migration processes, and ii) to the fine-scale extent to which local variability, in terms of socio-cultural and geographic heterogeneity, intervened in preserving or confounding these ancient or more recent historical genetic layers. Nevertheless, the complex interaction of these factors both in space and time has often made the attempt to disentangle their relative contributions to the current genetic variation a challenging issue in population genetics studies. In this context, the study of geographically or culturally marginal populations can provide a simplified observatory for testing the composite interplay among such forces. In fact, compared to large human groups, marginal populations show some features, such as a reduced genetic complexity, a higher environmental and cultural homogeneity, and a greater availability of historical and genealogical records, that may facilitate our insights into the human genetic diversity (Destro-Bisol et al., 2008).

In this section, after having summarised the processes potentially affecting population genetic structures at both broad and fine scales, we will direct our attention on particular case studies, represented by Italian “marginal” populations, in order to assess the role of demographic history and socio-cultural factors in shaping the current patterns of genetic variation, and to test whether a recent history of socio-cultural isolation may have helped preserving otherwise hidden genetic signatures of past (especially more recent) population events.

2.1.1 Interaction between historic, geographic and cultural factors: how culture, space and time shape the current genetic legacy

When we examine the genetic structure of modern populations, we do not simply detect the impact of single migration processes or admixture events, but rather the complex results of cumulative gene flows occurred from the population origin to the present (Jobling et al., 2013). As a result, the relative contributions of each single event on human population variability, in terms of both genetic and/or cultural impact, are often difficultly detectable. In addition, the signatures of past contacts on the gene pool of present-day populations can be confounded by micro-evolutionary forces, such as genetic drift, isolation and/or mutation, that interact at local level in modifying the existing patterns of genetic variation (Jobling et al., 2013).

Most of the studies which tried to address and estimate the impact of different human expansions and population admixtures on the European genetic landscape, basically focused their attention on comparing the Palaeolithic and Neolithic contributions, often reaching different conclusions about their relative extents. Besides these pre-historical genetic strata, the more recent history of Europe represented a complex tapestry of large-scale population movements, individual migration processes and/or other demographic events, which have added further confounding layers to the modern European genetic landscape by affecting the between-population genetic relatedness (Ralph and Coop, 2013). The formation (ethnogenesis) of the first proto-historic peoples (from the Metal Ages onwards), accompanied by the development of trade relationships and commercial exchanges, have actually brought different populations from different regions into contact among each other, thus mixing up the genetic diversity of the extant European populations. However, our current understanding of these events is often only deduced from archaeological, linguistic, cultural and/or historical evidences, whereas their genetic extents remain mostly uncertain, since the majority of these processes actually brought only partial modifications to pre-existing genetic backgrounds, while large population substitutions could be considered comparatively more rare.

Additionally to this multi-layered mosaic of pre-historical and historical incomings, socio-cultural and environmental heterogeneity represents another source of within- and among-population variability, being potentially responsible for different paths of isolation and/or population-interaction at more fine-grained local scales. Geographic constraints such as mountain chains and/or deep seas could not only have acted as physical barriers to gene flow limiting the reproductive exchanges (Jobling et al., 2013), but may also have represented survival areas for culturally-distinct or genetically-isolated populations (Coia et al., 2013). Another way that gene-culture interaction may have affected the human population variability, is when cultural factors - such as languages, socio-economic stratification, religion or ethnicity – influence the selection of mates within a population, thus becoming evolutionary forces able to generate significant genetic sub-structures within and among human groups (Laland et al., 2010). Among the different aspects of human culture that deeply condition the contacts and relationships among human populations, language is probably one of the most relevant, being directly involved in processes of communication and cultural transmission (Destro-Bisol et al., 2008). In addition, socio-cultural factors have recently demonstrated to play a critical role also in determining patterns of sex-biased genetic flows (Oota et al., 2001; Destro-Bisol et al., 2004; Marchani et al., 2008).

2.1.2 Marginal populations as a simplified observatory in our understanding of human genetic diversity and population history

Due to the complexity and heterogeneity of (pre-)historic and demographic processes which have characterised the evolutionary history of human populations, the impact of single population and/or admixture events on the present-day pattern of genetic diversity can vary quite substantially among and within the different human groups (Jobling et al., 2013). However, by looking at the current distribution of human populations in greater details, it is possible to notice the presence of numerous geographically and/or culturally marginal populations, which may represent windows into our understanding of European genetic variability and complexity. *Marginal populations* can be referred to as sub-populations that, by virtue of their geography, history or culture, have experienced reduced gene flows with the surrounding human groups (Arcos-Burgos and Muenke, 2002). These characteristics, when combined with low population sizes – as a result of founder events or bottlenecks – can give rise to *isolates* (Boattini et al., 2011). According to the classical definition provided by James Neel (1992), *genetic isolates* are marginal populations derived from a small population size, which experienced slow - if any - demographic expansion and very little recruitment from outside human groups. The descent from a limited number of founders few generations ago, the restricted geographical distribution and the pronounced socio-cultural uniformity, coupled with the presence of exhaustive and detailed historical records, have made human isolates a powerful tool to investigate peopling events - particularly concerning the human more recent history - the demographic behaviours, micro-geographic forces (like assortative mating, genetic drift, bottleneck, etc.) as well as the effects of historical, socio-economic and linguistic factors on the current human population genetic structures (Destro-Bisol et al., 2008; Robledo et al., 2014). As a result of their demography (small population size, recent shared genetic ancestry and low immigration rate), genetic isolates have become particularly important also in medical genetic studies, for mapping genes involved in rare diseases or identifying alleles involved in complex traits (Heutink and Oostra, 2002; Kristiansson et al., 2008; Veeramah et al., 2011; Esko et al., 2012).

In addition to geographical constraints, cultural features and particularly language, are among the most significant sources of isolation. This is particularly true for the so-called ethno-linguistic minorities, representing small groups of people speaking a language variety which is different from the official language of the surrounding area (Toso et al., 2014). Generally, differences in linguistic traits are also associated with differences in cultural practices (e.g. religion, traditions, mytho-history), on the whole defining the ethnic identity of each group. Ethno-linguistic minorities can provide an important opportunity to explore the roles of language (culture) vs. geography in shaping the genetic relationships among different human populations at micro-geographical levels (Nasidze et al., 2007). Moreover, the condition of cultural distinctiveness (potentially acting as a genetic barrier between different ethnic groups) may have helped them preserving a more direct genetic thread with the populations of origin, reducing the possibility of subsequent confounding events of admixture (Capocasa et al. 2014). As a consequence, culturally marginal populations can represent a unique tool to explore the different layers of human history, since they are more likely to have maintained genetic features otherwise lost (or diluted) in the surrounding open populations (Boattini et al., 2011).

Accordingly with this view, different ethno-linguistic groups and socio-cultural enclaves from different parts of the European continent, have largely attracted the attention of several population genetic studies, in the attempt to obtain a more fine-grained description of the European genetic variation and to consider the effects of geographic and cultural (especially linguistic) heterogeneity on the current patterns of genetic diversity. Since the first investigations on the European distribution patterns of genetic diversity, populations such as Finns, Sardinians and Basques (Cavalli-Sforza et al., 1994; Bertranpetit et al., 1995; Fraumene et al., 2006; Calò et al, 2008; Palo et al., 2009; Novembre and Ramachandran, 2011) proved to represent notable exceptions in the common South-East to North-West (SE-NW) major cline observed within the continent. Their geographical and/or cultural isolation has kept their genetic and demographic history preserved over a long period of time. As a consequence these populations have offered powerful tools to gain news insight into the ancient migratory paths and to investigate the human evolution and population structuring processes.

On a micro-geographic perspective, studies of geographic and cultural factors of genetic isolation have mainly focused their attention on European specific population context, at regional and local scales. Due to their geographic and reproductive isolation, island populations demonstrated to be among the most suitable populations for studying the effects of geographic isolation on human genetic structure (e.g. Croatian Cres islanders, Jeran et al., 2009; or Nias Island in Indonesia, Van Oven et al., 2011); on the other hand, ethno-linguistic minorities confirmed to be particularly useful to clarify the micro-evolutionary effects of socio-cultural vs. geographic forces on the between-population genetic relatedness. In some cases the geographic proximity of culturally-distinct groups better explained the genetic pattern observed, thus suggesting that the linguistic differences do not acted as genetic barriers in limiting the gene-flow between different ethnic-groups (e.g. Indo-European-speaking Armenians and Turkish-speaking Azerbaijani in the Caucasus, Nasidze and Stoneking, 2001; or the Turkish-speaking Gagauzes from Eastern-Europe, Nasidze et al., 2007 – but see also Verzari et al., 2009). On the contrary, in other cases the linguistic similarity of geographically distant groups proved to have served as genetic barrier between the linguistic enclaves and their culturally-different geographic neighbours, helping the former to preserve common cultural and genetic origins (e.g. the Kalmyks who speak a Mongolian language and are surrounded by Russian-speaking groups, Nasidze et al., 2005; the Sorbs who speak a west-Slavic language in an area with a majority of Germanic speakers, Veeramah et al., 2011; the Aromuns who speak a Romance language in the Balkans, Bosch et al., 2006). In some peculiar cases, the condition of isolation of culturally-linked but geographically-distant ethnic groups, have limited the genetic exchanges among the different linguistic “islands”, giving rise to small population groups that progressively diverged through drift from a common genetic substrate (e.g. Aromuns of the Balkan Peninsula, Bosch et al., 2006; or the Eastern Alps linguistic islands, Capocasa et al., 2013).

On the whole, this picture highlights the usefulness to combine data from large populations with the results obtained from “marginal” populations such as the ethno-linguistic minorities and the geographic isolates, in order to achieve a more complete picture of the genetic diversity of human populations (Capocasa et al., 2014).

2.1.3 Italian micro-geographic genetic variability: the role of socio-cultural factors

Among the different European countries, the Italian Peninsula represents a key area of investigation to explore the genetic variability at different geographical levels, due to the valuable richness of its historical heritage as well as to its environmental and cultural complexity (Destro-Bisol et al., 2008; Capocasa et al., 2014; Toso, 2014).

The composite peopling history and the environmental heterogeneity of Italy have been invoked to explain the genetic structure observed among the extant Italian populations. Both uniparental markers (*Paper 1*) and genome-wide autosomal data (Nelis et al. 2009, Di Gaetano et al. 2012) revealed a remarkable population sub-structure between the Northern and Southern Italian genetic pools, and suggested a shared genetic background between North-Western Italy and the North-Western Europe on one hand, and between South-Eastern Italy and the Balkan Peninsula on the other hand (*Paper 1*, *Paper 2*). These patterns, partly reflecting the topographic complexity of the Italian territory which is responsible for different paths of local drift or population mobility (*Paper 1*), have been mainly ascribed to the long history of population settlements that has affected the Italian Peninsula since pre-historical time, but particularly throughout the Neolithic transition and the subsequent Metal-Ages (*Paper 1*). During this period Italy underwent important demographic and socio-cultural transformations – which led to the ethnogenesis of the most important proto-historic Italic peoples – that, likely to have not been only cultural phenomena, contributed at least in part to affect the Italian population genetic structure (*Paper 1*).

The complexity of historical processes and environmental features that affected the Italian genetic structure, has also enriched the current genetic variability of a large number of populations with a history of geographic or socio-cultural isolation, which actually constitute another important component of the overall Italian human biodiversity (Destro-Bisol et al., 2008; Capocasa et al., 2014). As a consequence, the achievement of a comprehensive understanding of Italian population genetic variability cannot abstract from the analysis of these human groups, whose scientific value resides in both population genetics and archaeogenetics points of view, since these minorities actually offer the possibility to clarify the interaction among different micro-evolutionary forces as well as to find traces of past migration peopling events (Boattini et al., 2011).

A first comprehensive overview of the micro-evolutionary patterns of genetic variation in some of the main Italian ethno-linguistic minorities and geographic isolates is provided in *Appendix I* (Capocasa et al, 2014) as a part of a collaborative study performed within the framework of the MIUR- Projects PRIN2007 (*Isolating the isolates: geographic and cultural factors of human genetic variation*) and PRIN2009 (*Human biodiversity in Italy: micro-evolutionary pattern*).

Within this framework, in the second part of this thesis we have particularly considered two cultural enclaves of the Italian Peninsula, namely the Albanian-speaking Arbereshe and the *Partecipanza* of San Giovanni in Persiceto, that differ from one another for both the place of their present-day location - Southern Italy and Sicily for the former and the Padana Plain in Northern Italy for the latter - as well as for the cultural mechanism which is responsible for their marginality (and eventually isolation) – i.e. linguistic and socio-economic factors respectively. By analysing the uniparental genetic structures of these populations we especially aim to reach a dual-level purpose: i) to assess the importance of cultural factors (language or socio-economic condition) in shaping the local population genetic variability and ii) to reconstruct the genetic history of each isolate, verifying whether the recent conditions of socio-cultural isolation may have contributed to maintain signatures of migration and demographic processes occurred in Southern and Northern Italy during their more recent genetic history.

2.1.3.1 Ethno-linguistic factors: the case of the Albanian-speaking Arbereshe (Sicily and Southern Italy)

The Italian Peninsula is an ideal place to explore the relationships between language and genetics, since it hosts twelve officially-recognised ethnic-linguistic minorities, altogether representing about 5% of the overall Italian people: Albanian, Catalan, Croatian, French, Franco-Provençal, Friulian, German, Greek, Ladin, Occitan, Slovenian and Sardinian (Capocasa et al., 2014; Toso, 2014). Among them, the Albanian-speaking Arbereshe are one of the largest, today persisting in 41 municipalities (Figure 2.1.3.1.1) scattered through different regions of Southern Italy (including Sicily). They originated from massive immigrations of Albanian-speaking groups, leaving the Balkan Peninsula as a consequence of the Ottoman Empire invasions, approximately between the 15th and 16th centuries AD. The settlement of these communities in the Italian territory, has been the complex results of several waves of invasion, involving different migration routes as well as different numbers of intermediate steps both in the Balkan (through Greece) and/or Italian Peninsulas (Giunta, 2003).

The recent history of cultural isolation and the availability of well-documented records about their peopling history, have made the Arbereshe communities the object of many bio-demographic and genetic studies (Tagarelli et al., 2007; Fiorini et al., 2007; Boattini et al., 2010; Capocasa et al., 2014). The Arbereshe of the Cosenza Province (Calabria, Southern Italy) provides one on the most compelling cases due to the unique occurrence of particularly interesting conditions, that are i) the high number of villages, accounting for ~43% of the entire Albanian-speaking community of Italy (Tagarelli, 2004); ii) the exceptional preservation of ethnic identity and cultural traits at least until the second half of the 20th century (Fiorini et al., 2007), and iii) the historical settlement in the geographically isolated area of the Pollino massif, which forced the different Arbereshe communities to remain at least partially separated from each other and interspersed among Italian non-Arbereshe villages (Tagarelli et al., 2007). These factors constituted challenging features to specifically test the effects of historic, cultural and geographic factors on human population genetic variability, throughout relatively small spatial and temporal scales.

N	Municipality	Province
1	Acquaformosa	Cosenza (CS)
2	Castoregio	
3	Cerzeto	
4	Civita	
5	Falconara Albanese	
6	Firmo	
7	Frascineto	
8	Lungro	
9	Plataci	
10	San Basile	
11	San Benedetto Ullano	
12	San Cosmo Albanese	
13	San Demetrio Corone	
14	San Giorgio Albanese	
15	San Martino di Finita	
16	San Caterina Albanese	
17	Santa Sofia d'Epiro	
18	Spezzano Albanese	
19	Vaccarizzo Albanese	
20	Andali	Catanzaro (CZ)
21	Caraffa di Catanzaro	
22	Marcedusa	
23	Carfizzi	Crotone (KR)
24	Pallagorio	
25	San Nicola dell'Alto	
26	Barile	Potenza (PZ)
27	Ginestra	
28	Maschito	
29	San Costantino Albanese	
30	San Paolo Albanese	
31	Campomarino	Campobasso (CB)
32	Montecilfone	
33	Portocannone	
34	Ururi	
35	Casalvecchio di Puglia	Foggia (FG)
36	Chieuti	
37	Contessa Entellina	Palermo (PA)
38	Piana degli Albanesi	
39	Santa Cristina Gela	
40	Greci	Avellino (AV)
41	San Marzano di San Giuseppe	Taranto (TA)

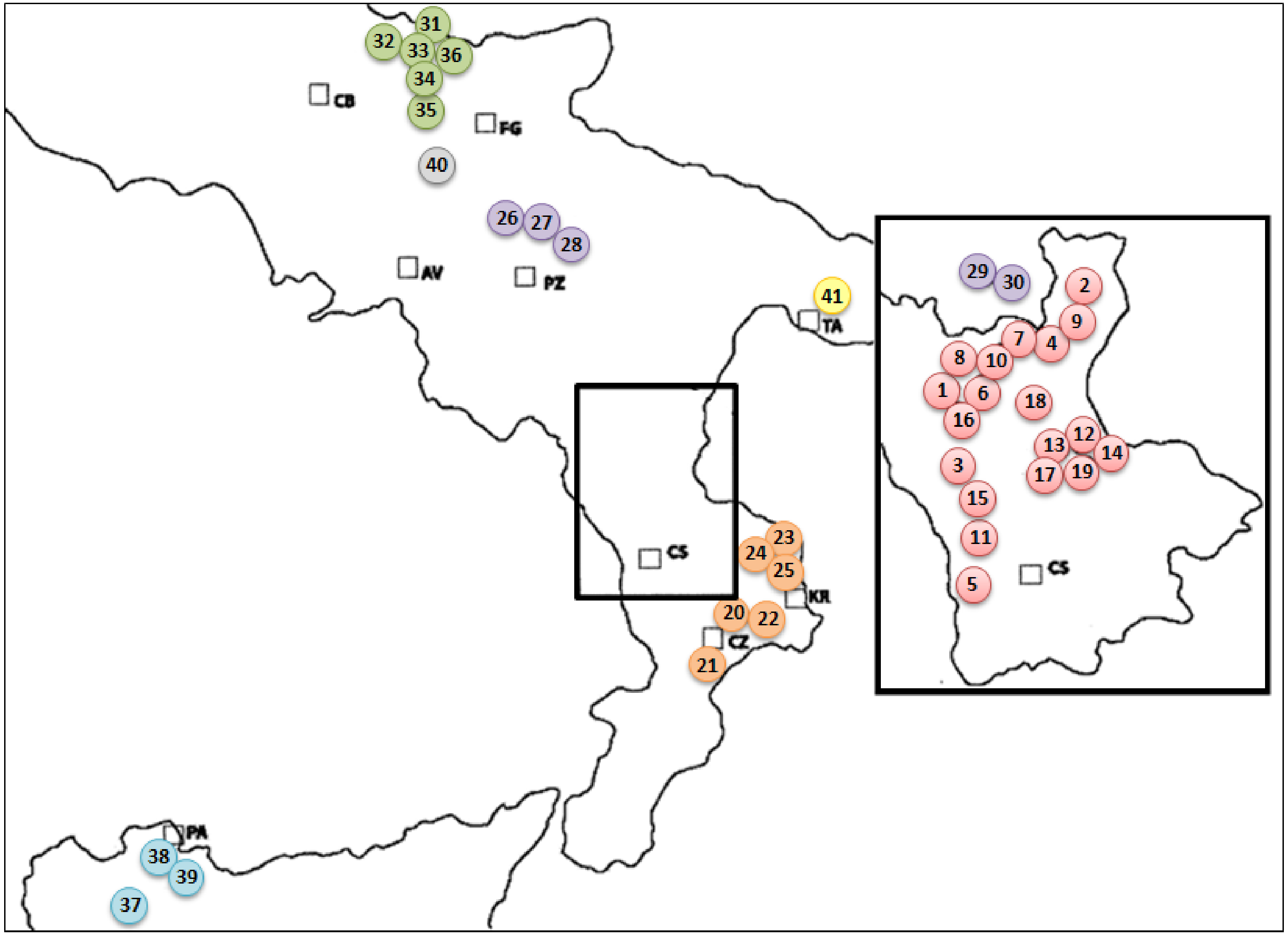


Figure 2.1.3.1.1. List and geographic location of the 41 Arbereshe municipalities of the Italian Peninsula. Modify from Tagarelli, 2004

Previous bio-demographic studies (Tagarelli et al., 2007; Fiorini et al., 2007) demonstrated that the Arbereshe communities of Calabria kept a clear genetic differentiation from the surrounding Italian non-Arbereshe groups by virtue of their linguistic and cultural isolation. However, environmental elements were shown to have strongly influenced the genetic relationships between the different Arbereshe communities, thus suggesting a double effect of cultural and geographic isolation in shaping the genetic variability between and within ethnic groups (Figure 2.1.3.1.2; Tagarelli et al., 2007). Notably, a diachronic-analysis approach revealed that such isolation features declined drastically during the second half of the 20th century, as a consequence of the so-called "*breakdown of isolates*" (Fiorini et al., 2007).

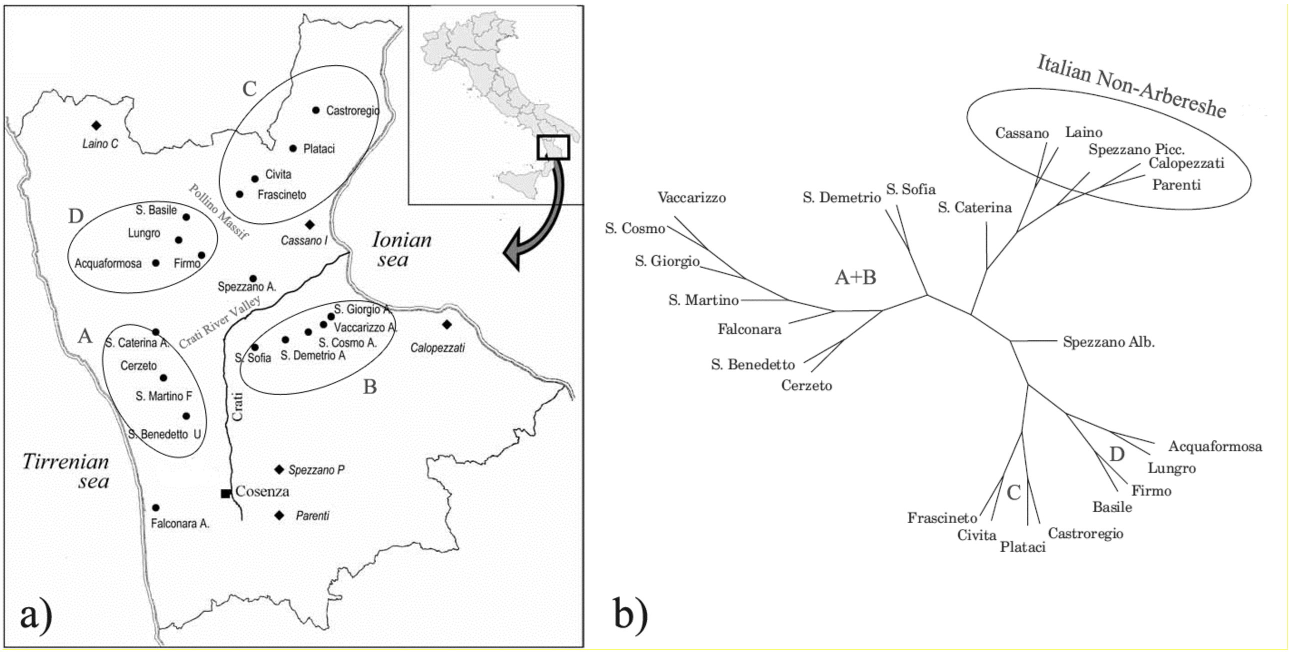


Figure 2.1.3.1.2. a) Location of the 19 Calabrian Arbëreshe (dots) and the 5 Italian Non-Arbëreshe (diamonds) populations investigated by Tagarelli et al., 2007. Geographic groups: A, East of the Crati River; B, West of the Crati River; C) Pollino Area; D) Pollino South-West b) Neighbor Joining Tree of isonymic relationships among the analysed populations. Modify from Tagarelli et al., 2007.

From a genetic point of view, a first Y-chromosome-based study (Boattini et al., 2010) confirmed the discontinuity of Calabrian Arbereshe from the surrounding Italian genetic space, revealing their shared genetic variation with modern Balkan populations. More recently, a first attempt to compare the uniparental variability of the Calabrian Arbereshe with the Sicilian Albanian-speaking groups, preliminarily hypothesised divergent micro-evolutionary histories between the two Arbereshe groups (Figure 2.1.3.1.3; Capocasa et al., 2014), thus suggesting a more complex interplay between historical, cultural and geographic factors, even among culturally and historically related populations.

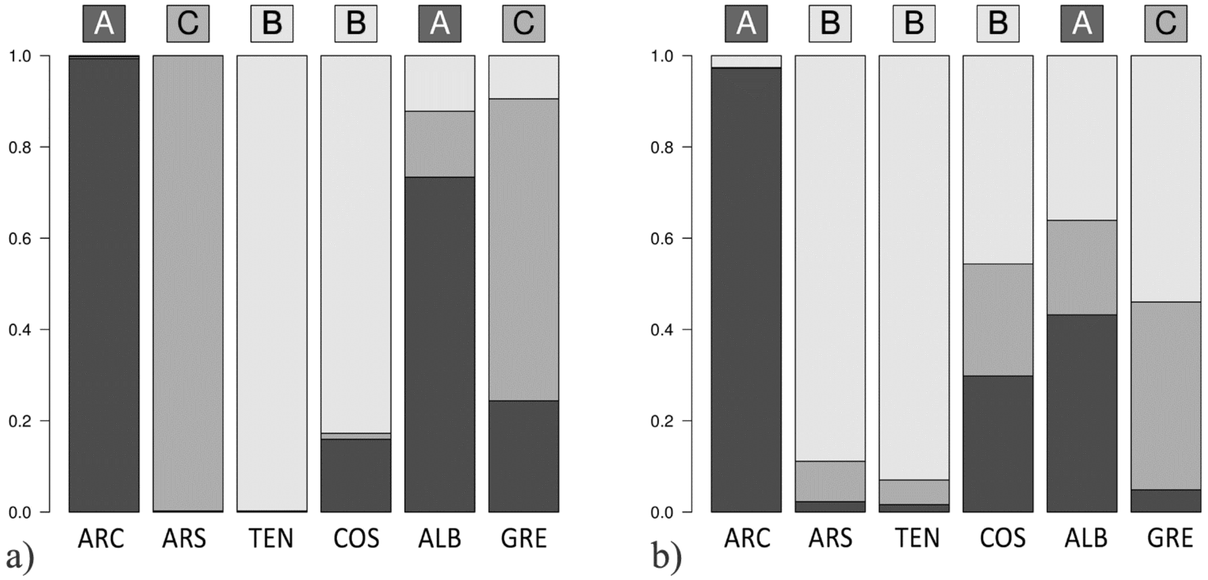


Figure 2.1.3.1.3. Admixture-like barplots for Y-chromosome (a) and mtDNA (b). The barplots represent DAPC-based posterior membership probabilities for each of the considered populations and for each inferred cluster (*mclust* algorithm). The affiliation of each population to a given cluster and its corresponding colour code are represented by letters (within coloured squares) on the top of each bar. Abbreviations: Arbereshe from Calabria (ARC), Arbereshe from Sicily (ARS), Trapani-Enna (TEN), Cosenza (COS), Albania (ALB), Greece (GRE).

2.1.3.2 Socio-economic factors: the case of the Partecipanza of San Giovanni in Persiceto (Northern Italy)

In addition to geographic factors (physical constraints) and/or linguistic diversity (cultural barriers), also socio-economic practices have recently demonstrated to potentially affect the genetic structure of human populations, in some cases inducing sex-biased patterns of genetic structuring (Destro-Bisol et al., 2004). By influencing the way in which unions between individuals took place, the study of the relationships between genetic variation and socio-economic forces can improve our understanding of the role of culture in determining and differentiating the current genetic variability within and among populations (Destro-Bisol et al., 2004). However, this phenomenon has been only partially explored in population genetic studies, probably due to the difficulty to define stable categories of populations. In fact, socio-economic features tend to rapidly change over the time, with an increasing trend especially in the last centuries of the human history.

In this context, the *Partecipanza* of San Giovanni in Persiceto (a community settled in Northern Italy) is one of the rare case-studies able to provide an ideal framework of investigation to specifically address the relationships between genetics and socio-economic forces.

The *Partecipanze* are an idiosyncratic way of sharing and devolving collective lands of Medieval origins, that still survive in some areas of the Padana Plain (Northern Italy). The privileged status of *Partecipante* (i.e. the ability to participate in the sharing of leased assets) is conditioned by two fundamental rules: i) to be a male legitimate descent of a *Partecipante* founder family (i.e. gene-like patrilineal inheritance), and ii) to maintain the residence in the municipality of San Giovanni in Persiceto throughout the different generations (i.e. highly specific local ancestry). The membership to an economically advantaged élite group of Medieval origins, and the peculiar way of inheritance of the status of *Partecipante* (strictly and exclusively male-mediated), are both conditions able to have generated population-specific and sex-biased patterns of genetic structuring, thus offering the unique chance to test whether a relatively short history of "socio-cultural élite environment" has left detectable signatures in the genetic diversity pattern of current population.

In addition, the nine-year periodic availability of census-like descriptions of the population composition (made possible by the cyclic redistribution of the shared lands among the heads of the *Partecipante* families, Figure 2.1.3.1.4), offers the rare possibility to reconstruct deep-rooted paternal pedigrees, which can be associated to paternal genetic profiles. Thanks to the unique link between traceable culturally inherited elements - patrilineal surnames of founder families - and genetically inherited markers - male specific Y-chromosomes - (Manni et al., 2005; Winney et al., 2012), it is therefore possible to select samples that are more likely to be representative or to contain signatures of the ancestral population, as it appeared before the reshuffling effects of more recent migration and expansion events - especially since the Industrial Revolution which has progressively blurred the picture of fine-scaled historical population structure and genetic variation (Bowden et al., 2008; Larmuseau et al., 2012).

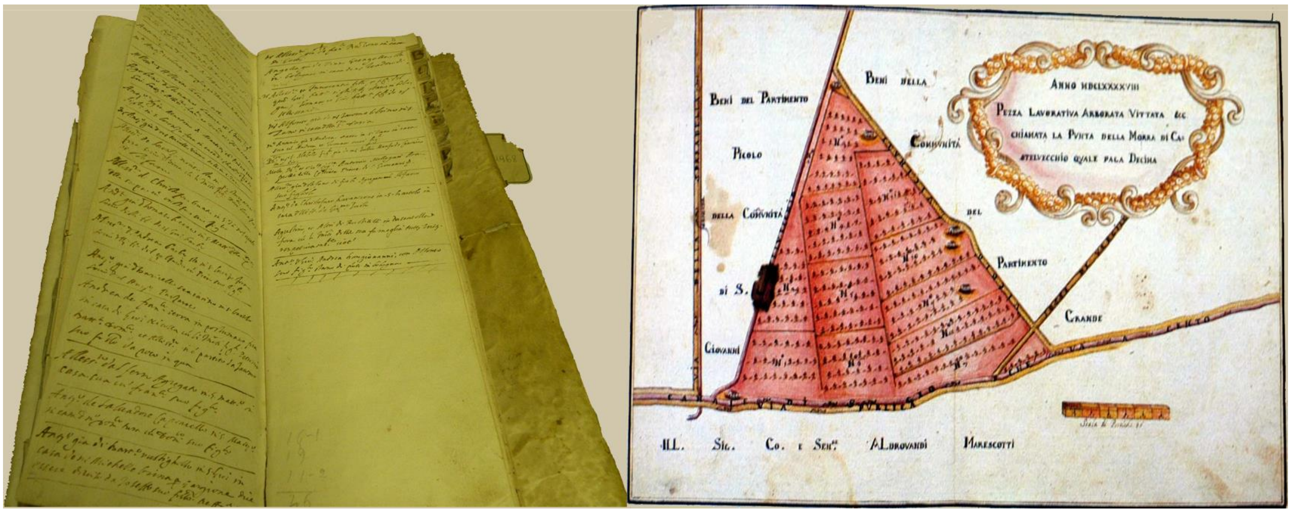

Anno MDCLXXXXVIII
Pezza Lavorativa Arborata Vittata &c
chiamata la Punta della Morra di Ca-
stelvecchio quale paga Decima
Beni del Partimento
Picolo
della Comunità
di S.
Giovanni
Beni della
Comunità
del
Partimento
Grande
Ill. Sig. Co. e Sen.re Aldrovandi Marescotti

Figure 2.1.3.1.4. (Left) Example of registers of the Enrolments ("*Registri delle iscrizioni"*) preserved at the historic archives of the *Partecipanza* (photo by Medoro C). The *right to participate* to the sharing of the leased assets is based on these registers which are compiled every nine years on the occasion of the subdivision of the lands (an example in the map on the right) among the heads of the *Partecipante* families.

2.2. Specific aims of the studies

As briefly reviewed in the previous paragraphs, the complex patterns of genetic diversity in extant human populations are the product of different strata of historical, demographic and evolutionary events acting on different timescales, that include colonisation, migrations, splits, population expansions or contractions, admixture, isolation, mutations and genetic drift. In particular, if we consider the multi-layered history, the geographic complexity and the cultural heterogeneity that characterise the Italian Peninsula, it becomes even more evident that trying to connect genetic signals with particular (relatively recent) historical events and/or with specific geographic or socio-cultural factors, necessarily requires the choice of peculiar case-studies and the planning of well-conceived sampling strategies and research designs. In this respect, the study of geographically and/or culturally marginal populations may provide unique opportunities to clarify the genetic patterns emerged from open-populations located in the same area as well as to investigate the evolutionary forces that shaped human genetic variation at relatively short spatial and temporal scales. Within the framework of a comprehensive investigation of the Italian micro-evolutionary variability (MIUR-Project PRIN2007 and PRIN2009), in the second part of this thesis we have particularly considered two peculiar case studies.

In the first case-study we compared the paternal genetic variation of two Arbereshe ethno-linguistic groups –from Sicily and Southern Italy (Calabria) respectively – both among each other as well as with their putative Balkan source and Italian sink populations. That way, by exploiting the available historical data and by considering the different conditions of cultural and geographic "marginality", we particularly intended to explore how different demographic histories, geographic influences and cultural forces may have intervened in shaping the within-ethnic group genetic variability and/or in maintaining the genetic continuity with the populations of origin. The results of this study are presented in *Article 3*.

In the second case-study we focused our attention on the *Partecipanza* of San Giovanni in Persiceto, to verify whether its peculiar socio-economic condition might have induced (sex-biased) population genetic structures and may have helped preserving specific traces of the more recent genetic history of Northern Italy. In addition, the association of deep-rooted paternal pedigrees to Y-chromosome genetic profiles (see the previous paragraph for more details), allowed us to address some more generally debated questions concerning the assessment of evolutionary parameters, such as Y-STRs mutation rates and population average generation time. The results of this study are presented in *Article 4.*

A more detailed discussion of some aspects of the results, for both the above mentioned articles, is presented afterwards as a commentary.

1.3. Results and Discussion

Article 3

Sarno S, Carta M, Boattini A, Tofanelli S, Ferri G, Alù M, Motta V, Anagnostou P, Sineo L, Tagarelli G, Luiselli D, Pettener D. *Same language, different genetic histories: high resolution analysis of Y-chromosome variability in the Arbereshe populations of Calabria and Sicily* (in preparation).

Same language, different genetic histories: high resolution analysis of Y-chromosome variability in the Arbereshe populations of Calabria (Southern-Italy) and Sicily.

Stefania Sarno[1], Marilisa Carta[1], Alessio Boattini[1], Sergio Tofanelli[2], Gianmarco Ferri[3], Milena Alù[3], Vincenzo Motta[1], Paolo Anagnostou[4], Luca Sineo[5], Giuseppe Tagarelli[6], Donata Luiselli[1], Davide Pettener[1]

1) Laboratorio di Antropologia Molecolare, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, 40126 Bologna, Italia
2) Dipartimento di Biologia, Università di Pisa, 56126 Pisa, Italia
3) Dipartimento di Medicina Diagnostica, Clinica e di Sanità Pubblica, Università degli Studi di Modena e Reggio Emilia, 41124 Modena, Italia
4) Dipartimento Biologia Ambientale, Sapienza Università di Roma, 00185 Roma, Italia
5) Dipartimento di Biologia Ambientale e Biodiversità, Università degli Studi di Palermo, 90133 Palermo, Italia
6) Istituto di Scienze Neurologiche-CNR, 87050 Mangone (CS), Italia

Number of text pages: 20. Number of figures: 3. Number of Supplementary Figures: 8. Number of Tables: 2. Number of Supplementary Tables: 5.
Abbreviated Title: Y-chromosome in Sicilian and Calabrian Arbereshe
Key Words: linguistic minorities, culture, genetic ancestry, admixture, Balkan, micro-evolution

Corresponding Author
Stefania Sarno, Laboratorio di Antropologia Molecolare - Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Via Selmi 3, 40126, Bologna, Italy. Tel. +39 051 2094105. E-mail: stefania.sarno2@unibo.it

Grant
This study was supported by the ERC Langelin Project grant (FP7-Ideas-ERC2011-AdG295733) to DP and DL.

ABSTRACT

Albanian-speaking minorities of Sicily and Southern Italy proved to have maintained a clear genetic link with the Balkan-source populations, by virtue of their linguistic isolation from the surrounding Italian groups. However, preliminary results of both bio-demographic and population genetics studies suggested differences in the ethnogenesis processes between the Arbereshe groups today settled in the Italian Peninsula. In this study we address the genetic history of Arbereshe people with a high-resolution analysis of Y-chromosome variability. A large set of slow- and fast-evolving molecular markers was typed in different Arbereshe communities of Sicily and Southern Italy. By combining this dataset with newly-generated Y-chromosome data from both Balkan-source and Italian-sink populations, we particularly explored the role of geographic, demographic and cultural factors in shaping the current genetic diversity of Arbereshe population. Simulations of admixture dynamics reveal different scenarios for the two ethno-linguistic groups. The Albanian genetic ancestry appears clerly conserved in Calabrian Arbereshe, the Sicilian ones instead showing a more complex inheritance pattern, implying partial contributions from Greek populations and higher levels of local admixture. The dissection of haplogroup composition highlights the differential presence of specific haplotype clusters both between the two Arbershe groups, as well as among the different communities within them. Accordingly with historical evidences, these results signal traces of diverging histories between the Calabrian and Sicilian Arbereshe, suggesting different founding events and different patterns of population contact or cultural isolation *in situ*, which resulted in the differential acquisition or preservation of specific paternal lineages by the two Arbereshe groups.

INTRODUCTION

The term "linguistic minority" generally refers to a group of people (or a small number of geographically close communities) speaking a language variety which is markedly different from the official language of the surrounding geographic area (Toso, 2014). Usually, these populations have originated from a restricted number of founders and have remained at least partially isolated from the subsequent events of admixture (Destro-Bisol et al. 2008). Compared to open populations, linguistic minorities can therefore represent a simplified model of investigation to focus on the interplay between geographic and cultural factors, reducing the confounding noise otherwise ascribable to the multi-layered history of a population. In addition, their condition of linguistic isolation (potentially acting as a genetic barrier between different ethnic groups) might have helped these communities to preserve a more direct genetic link with the populations of origin (Capocasa et al. 2014). Besides increasing the cultural and genetic variability of their recipient populations (Capocasa et al., 2014), ethno-linguistic groups can therefore represent important "tools" to address two main issues in population genetics studies, that are i) the influence of cultural (linguistic) traits on population genetic structures, and ii) the extent to which specific genetic signatures of population past history are traceable and detectable in the gene pool of the extant individuals. In addition, isolated populations have recently attracted the increasing interest of medical genetics studies for mapping genes involved in rare diseases or identifying alleles involved in complex traits (Destro-Bisol et al., 2008; Veeramah et al., 2011; Esko et al., 2012). Due to their demographic features (small population size, recent shared genetic ancestry and low immigration rate), isolated populations are indeed expected to show lower levels of genetic heterogeneity and internal population structure, as well as increased levels of Linkage Disequilibrium (LD) compared to the large outbred populations (Heutink and Oostra, 2002; Kristiansson et al., 2008).

Among the numerous ethno-linguistic groups today established in the Italian territory (see Capocasa et al, 2014 and Toso et al. 2014 for a full overview), the Albanian-speaking Arbereshe are one of the largest. Unlike most of the Italian ethno-linguistic minorities, the Arbereshe case is well-documented from the historical point of view. The available historical data provide a quite reliable and precise understanding on the place and time of their origin, as well as on the main geographic and chronological patterns of diffusion within Southern Italy (including Sicily). Their origins are generally referable to massive movements of Albanians, occurred between the 15th and the 16th centuries, mainly in response to the invasion of the Balkans by the Ottoman Empire (Zangari, 1941). However, their presence in Italy is actually the result of several migration waves, either coming directly from Toskeria (Southern Albania) or arrived in Italy after intermediate steps in

Greece (Peloponnese). Today the Arbereshe people survive in 41 municipalities, distributed among different regions of Sicily and Southern Italy (Tagarelli, 2004).

The Arbereshe of Calabria (Southern Italy) particularly persist in 19 municipalities of the province of Cosenza scattered around the highland area of the Pollino massif and along the Crati River valley. Besides sharing a condition of geographic marginality, these communities have demonstrated a deep sense of group identity in maintaining their social, religious (Greek Orthodox) and linguistic (Arberisht) original traits, at least until the first half of the 20th century AD (Fiorini et al., 2007). Previous bio-demographic studies (Tagarelli et al.2007; Fiorini et al. 2007) demonstrated that such cultural identity has helped Arbereshe communities to maintain a clear genetic differentiation from the surrounding Italian non-Arbereshe villages (Tagarelli et al., 2007). Importantly, it was shown that such differentiation declined drastically during the second half of the 20th century (breakdown of isolates; Fiorini et al. 2007). In accordance with bio-demographic results, our first molecular surveys (Y-chromosome, mtDNA), performed on a surname-based selected subset of Calabrian Arbereshe, confirmed this population to be genetically discontinuous compared to the Italian genetic background, while showing a shared genetic ancestry with modern Balkan populations (Boattini et al., 2010; Capocasa et al., 2014). In addition, and consistently with their peculiar condition of greater environmental isolation, surname-based studies suggested the strong influence of geography on the pattern of genetic relationships between the different Arbereshe villages. Three main clusters, exactly mirroring the areas of geographic location of the different Arbereshe communities (Pollino Massif, Pollino South-West, Crati River Valley) were particularly invoked to explain their within-ethnic-group genetic variation (Tagarelli et al., 2007).

On the other hand, the Arbereshe group of Sicily today survives in only three municipalities in the province of Palermo. They are characterized by lower geographic isolation and population size. In addition, historical evidences suggest a more troubled peopling history for these communities, that particularly involved several intermediate steps both in the Balkan and Italian Peninsulas before their final settlement in Sicily, as well as subsequent re-peopling events from Greece (Giunta, 2003). Our preliminary investigation of the uniparental genetic structures of Sicilian Arbereshe (Capocasa et al., 2014) revealed stronger links with Greek populations on the paternal side and higher similarities with Italians from the maternal point of view.

Based on these studies, here we present a new high-resolution analysis of Y-chromosome genetic variation in the Arbereshe population i) by increasing the sampling coverage for both Sicilian and Calabrian Arbereshe; ii) by deepening the resolution level of Y-chromosome analysis and combining the investigation of both slow- and fast-evolving genetic markers; and iii) by generating new data also for the Albanian comparison populations.

By using a micro-geographic approach, aimed at directly comparing linguistic isolates with both their geographic neighbors and source populations, we seek to specifically address the following issues: i) to obtain a fine-grained description of Y-chromosomal diversity within the Arbereshe ethno-linguistic minority, in order to distinguish between homogeneous or discontinuous patterns of genetic variation among the different Arbereshe groups/communities; ii) to explore the differential role of culture and geography in shaping the genetic structure of each group by assessing their degree of genetic continuity or local admixture with the Balkan-source or Italian-sink populations respectively; iii) to gain new insights into the population movements at the origin of the Arbereshe settlement in Sicily and Southern Italy by investigating their genetic ancestry.

MATERIALS AND METHODS

Ethics statement

All donors provided a written informed consent to the project and data treatment, according to the ethical standards of the institutions involved. The Ethics Committee at the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna (Italy) approved all procedures.

Population samples

The dataset primarily consist of unpublished Y-chromosome data of 150 unrelated male individuals from 13 Arbereshe villages of the province of Cosenza (Calabria, South Italy) and 2 Arbereshe communities of the province of Palermo (Sicily). From now on we will refer to the former as Calabrian Arbereshe (ARB_CAL) and to the latter as Sicilian Arbereshe (ARB_SIC). Blood samples or buccal swabs were collected from healthy males volunteers, selected according to the following criteria: i) possession of Arbereshe founding surnames; ii) self-declared affiliation to the Arbereshe ethnic-group; iii) patrilineal residence in Arbereshe village for at least three generations. Y-chromosome data (12 STRs and 30 Y-SNPs) of 40 ARB_CAL individuals were partially published in Boattini et al., 2010.

Based on previous bio-demographic results (see Introduction), and additionally considering the low sample size otherwise obtained for each single village if considered separately, the 13 Calabrian Arbershe communities were grouped in 3 ethno-geographic homogeneous clusters: i) populations located in the valley of the Crati River (referred as Crati River Valley, VAL_CRA; n=46); ii) populations located on the south-western side of the Pollino area (referred as Pollino South-West, POL_SW; n=36); iii) population located in the Pollino massif area (referred as Pollino Massif Area; POL_AREA; n=24).

The two Sicilian Arbereshe communities of Contessa Entellina (CON_ENT; n=26) and Piana degli Albanesi (PIA_ALB; n=18) were analyzed separately considering the postulated complexity and heterogeneity in their evolutionary histories (see Results).

New unpublished Y-chromosome data have been moreover generated for 223 unrelated Albanian males, collected from the two major dialect groups of Albania, respectively Ghegs from the North (N=119) and Tosks from the South (N=104). All the participants identified themselves as member of the given ethnic group, with at least two generations of unrelated paternal ancestry in their region of birth. 12 STRs and 18 Y-SNPs data for these populations were partially published in Ferri et al. 2010.

We finally included 263 samples from Sicily and Southern Italy (Boattini et al., 2013; Sarno et al., 2014) as well as 209 unpublished Greek samples from Korinth and Eubea (Anagnostou P, personal communication), reaching a total of 845 Y-chromosomes.

Genomic DNA was extracted by using a Salting Out modified protocol (Miller et al., 1988).

Y-chromosome genotyping

Haplogroups (HGs) were determined by typing 42 SNPs in the non-recombining region of the Y chromosome, following a hierarchical approach. Basal haplogroups were firstly assigned through the analysis of the seven Y-SNPs (R-M173, J-M172, I-M170, E-M35, K-M9, P-M45, F-M89) implemented in the MY1 Multiplex PCR by Onofri et al. (2006). Afterwards, 33 additional Y-SNPs (E-M78, E-V12, E-V13, E-V22, G-P15, G-P16, G-M286, G-U8, G-U13, I-M253, I-M227, I-L22, I-P215, I-M26, I-M223, J-M410, J-L27, J-M67, J-M92, J-M12, R-M17, R-M343, R-M18, R-M269, R-L51/S167, R-S127/P311, R-S21/U106, R-S116/P312, R-SRY2627/ M167, R-S28/U152, R-M126, R-M160, R-L2/S139, R-L21/S145) were typed within the major European clades by using six haplogroup-specific multiplexes (Ferri and Alù, 2012). The SNP genotyping was carried out by means of PCR Multiplex amplification, followed by Minisequencing SBE reaction performed with the SNaPshot multiplex kit (Applied Biosystem). SBE products were analyzed with capillary electrophoresis on an ABI Prism 310 Genetic Analyser. Two further SNPs (E-M81, E-M123) were finally tested with RFLP analysis as described previously (Neto et al., 2007; Gayden et al., 2008). All samples were additionally typed for the 17 Y-STRs loci implemented in the AmpFlSTR Yfiler PCR Amplification Kit (Applied Biosystems, Foster City, CA) following the manufacturer's recommendations (Mulero et al., 2006). Products were sized on an ABI Prism 310 Genetic Analyzer by using the GeneScan 3.7 software (Applied Biosystems, Foster City, CA). As the Yfiler kit amplifies DYS385a/b simultaneously, avoiding the determination of each of the two alleles, these two loci were excluded from all the analyses performed. Unless otherwise specified, statistical analyses were so performed using 15 STRs loci. The DYS389b locus was obtained by subtracting DYS389I from DYS389II (Gusmao et al., 2006).

Statistical analyses

Haplogroup frequencies were estimated by direct counting. Standard measures of intra- and inter-population genetic diversity were calculated using the Arlequin software (version, 3.5.1.2; Excoffier et al., 2007). The genetic variance within and between population groups was partitioned at different hierarchical levels through the analysis of molecular variance (AMOVA) as implemented in the Arlequin software. Fisher exact tests were performed on haplogroup frequencies in order to determine significantly over- or under-represented HGs in any of the considered populations. To reach a common level of Y-chromosome resolution, sub-lineages were concatenated in 36 HGs.

The phylogenetic relationships between haplotypes within the most frequent HGs were inferred using the Reduced Median (RM) network algorithm, whose output was subsequently exploited to calculate the Median Joining (MJ) network, through the Network software 4.5 (Bandelt et al. 1995). STRs loci were weighted according to the inverse of their variance (Meyer et al. 1999). MJ networks were visualized with Network Publisher (Fluxus Engineering, Clare, UK). Time to the Most Recent Common Ancestor (TMRCA) was calculated for population-specific clusters of STR haplotypes. These clusters were identified within the phylogenetic networks of the most important HGs following Balanosvsky et al. (2011) and selecting only those clusters which appeared to have specifically evolved within particular population groups (frequency higher than 70%). TMRCAs were estimated by means of the standard deviation (SD) estimator by Sengupta et al. (2006). The 95% confidence intervals were calculated based on the standard error (SE). This method does not estimate the population divergence time, but the relative age (amount of time) required to accumulate the observed STRs variation within the haplogroup at each population (or population-specific cluster). Because haplotype clusters are population specific, the resulting age estimations should be interpreted as lower bounds for the time that a population may have been isolated following a split (Balanosvsky et al., 2011). Being the time estimates based on variance sensitive to the presence of outliers, all Y-chromosome age estimates here obtained were corrected for the presence of outlier STR loci as described previously (Boattini et al. 2013). Since events involving Arbereshe population are relatively recent (~500 YBP), time estimates were computed based on all STR loci (minus DYS385a/b). As for mutation rates, we adopted locus-specific mutation rates as estimated by Ballantyne et al. (2010). A generation time of 28 years has been considered.

The relationships between Arbereshe ethno-linguistic minorities and both their putative Balkan source and Italian recipient populations, were investigated by using several multivariate analyses, based on both fast- and slow-evolving markers. Multi-Dimensional Scaling (MDS) based on STRs genetic distances (Reynolds et al., 1983; for the use of this distance with STRs see Laval et al., 2002) was performed by using the R software function *isoMDS* (library *MASS*; Venables and Ripley 2002). The STRs-based distance matrix was additionally exploited to perform a hierarchical cluster analysis (1,000 bootstraps) according to the Ward's algorithm implemented in the R software (function *hclust*, library *stats;* R Development Core Team, 2011). A Correspondence analysis (COA) on HGs frequencies was finally performed by using the function *dudi.coa* (R software package *ade4*, Dray and Dufour, 2007; Dray et al., 2007).

The paternal genetic relationships of Arbereshe with the other comparison populations were further addressed and compared by means of admixture-like plots based on STRs genetic distances, as described previously (Sarno et al., 2014).

In order to estimate the contribution of source and local/sink populations to present-day Arbereshe communities, we applied a hypothesis testing approach. The variation through time of DHS - a molecular distance based on the extent and type of shared haplotypes between pairs of diverging pools (Tofanelli et al. 2009) - was calculated at Y-STR and Y-SNP variants. As for STRs, we selected the 12 STRs markers with the highest duration of linearity D with time (Busby et al., 2012). Different sets of parameters (see Table S1) were modeled under a stochastic Markov chain Monte Carlo (MCMC) method as implemented in the software CUDAshes (codeplex.ashes.com). As starting pools we used three putative source samples (Tosks, Ghegs, and a 50% random blend of Greek and Tosks) and two local non-Arbereshe groups (Sicilians, Calabrians). In every simulation model we considered two populations coming into contact at time t0 and exchanging M haplotypes from the source to the sink pool. At time t1, the two populations were allowed to evolve independently for 20 generations – that is the time since the oldest evidence of the Arbereshe migration (~560 YBP) by considering 28 years per generation. For each model 500 iterations were performed and summary statistics of DHS values were calculated. To take into account the effects of demographic dynamics occurred since the event of contact, realistic prior parameters were used (Table S1). Varying parameters were the number of exchanged haplotypes or migrants (5% 10%, 20%, 30%, 50% contribution of local haplotypes to the final pool), the mutation rate (0.002485 for STRs; 0.000001 for SNPs), and the increment rate (1.00 for the stationary model, 1.07 for the population growth model, the upper estimate from the historical evidence). The starting haplotype pools of source and local populations were built upon n-time reiterations of the real data with a final effective size one-sixth of the averaged present-day census size. The distributions of simulated DHS values were compared with empirical values calculated for each pair of samples. The data were considered to fit the model when observed DHS values fell within two standard deviations by the mean of the simulated distribution.

RESULTS

We analyzed 845 Y-chromosome haplotypes from the Arbereshe ethno-linguistic minorities of Sicily and Southern Italy (Calabria), as well as from their putative source (Albania, Greece) and recipient (Calabria, Sicily) populations (Table 1). Y-STR haplotypes and Y-SNP genotyping results for the newly-typed Arbereshe (n=150) and Albanian (n=223) individuals are provided in Table S2. The frequencies of Y-chromosome haplogroups (HGs) for both the two Arbereshe groups as well as for Balkan and Southern-Italian comparison populations are detailed in Table S3. The geographic distribution of Y-chromosome main lineages in the studied populations is shown in Figure S1.

Within ethnic-group variation and population genetic diversity

Y-STRs intra-population genetic diversity of Arbereshe communities (ranging from 0.9785 to 1.000; Table S4) attains values which are similar to those reported in previous studies on Arbereshe Y-chromosome haplotype variability (Boattini et al. 2010, Capocasa et al, 2014). Higher values have been instead obtained for Y-HGs, as expected if considering the increased resolution panel of Y-SNPs here analyzed. On the whole, both haplotype and haplogroup diversity in Arbereshe communities do not significantly differ from the range of values observed among Italian and Balkan comparison populations (Mann-Whitney U test: W=10.5, p-value=0.075 for STR data, and W=12, p-value=0.113 for SNP data). Nevertheless, departures from the observed range of variability have been detected in the Sicilian Arbereshe community of Contessa Entellina as well as in the Albanian Gheg population, which showed the lowest values of within-population genetic diversity for STR (0.9785±0.0166) and SNP (0.7695±0.0251) data respectively (Table S4).

Within-ethnic group variation was explored by means of AMOVA analysis (Table S5). A significant genetic heterogeneity for both HGs (Y-SNPs) and haplotypes (Y-STRs) has been found within the Arbereshe ethno-linguistic group when all the five communities were grouped in a single cluster (3.96%, p-value<0.001 for SNPs; and 2.95%, p-value=0.0028 for STRs). When considering the variation within ARB_CAL and ARB_SIC separately, AMOVA results (Table S5) indicate the ARB_SIC as the main contributor to the heterogeneity observed within the Arbereshe ethnic-group (4.60%, p-value=0.0305 for SNPs; and 5.17%, p-value=0.0358 for STRs). On the other hand, the within-population variation for ARB_CAL appears statistically significant for SNPs markers (2.42%, p-value=0.0106) but not for STRs data (0.56%, p-value=0.2532).

As for the Italian and Balkan (excluding Arbereshe) comparison populations, both Southern Italian (including Sicily) and Greek populations appear as two quite consistent homogeneous groups (Table S5). A statistically significant intra-group variation has been instead observed for Albania

(1.53%, p-value=0.0182 for SNPs; and 1.21%, p-value=0.0258 for STRs), thus revealing a modest degree of differentiation between Tosks and Ghegs.

Populations comparison

To visualize clearly the genetic relationships of Arbereshe communities, with both their Balkan source and Italian recipient populations, a correspondence analysis (CA) was performed on HG frequencies (Figure 1). The first two components, together accounting for ~40% of the total variance, highlight both geographic and linguistic groupings. Italian and Balkan populations appear well detached along the first component (CO1, 20.6%), which differentiates the two geographic Peninsulas. In this context, the Arbereshe communities of both Sicily and Southern Italy (Calabria) appear mainly scattered within the space occupied by the Albanian and Greek populations, that way confirming the influence of ethnic and cultural features on the observed genetic structure. The distribution of samples along the second component (CO2, 19.1%) detaches all the Sicilian populations, independently from their cultural affiliation to Italian (Sicilians and Southern Italians) or Balkan (Arbereshe of Sicily) languages, from all the other populations. Consistently with CA results, a significant among-groups variance (Table S5) is observed when Arbereshe populations were grouped according to both their linguistic affiliation and geographic location, thus attesting the role of both culture and geography in shaping the observed Y-chromosome genetic composition.

Results based on STRs genetic distances (Figure 2, Figure S2), besides confirming the geographic separation between the Balkans and Southern Italy, stress even further the heterogeneity observed among the two Arbereshe ethno-linguistic groups. Coherently with CA results, Italian and Balkan populations appear generally distinguished by the two dimensions of the MDS plot (Figure S2), also clustering in two well-separated groups in the cluster analysis (respectively an Italian - bootstrap 61.2% - and a Balkan one - bootstrap 76.6%; Figure 2). Analogously, Arbereshe communities fall mostly within the Albanian-Greek cluster (the only exception being CON_ENT, Figure 2), thus supporting their shared genetic ancestry with modern Balkan populations. Nevertheless, the two Arbereshe groups appear well detached along the second dimension of the MDS plot (Figure S2), also showing different patterns of genetic similarity in the hierarchical clustering analysis (Figure 2).

Paternal contributions to the Arbereshe genetic pool

The admixture contributions of Balkan source and Italian recipient populations to the present-day Arbereshe genetic pool were estimated by performing forward DHS-based simulations under different scenarios, with the MCMC method implemented in the program CUDAshes (see Matherial and Methods for details).

By comprehensively considering both STRs- and SNPs-based results, the scenario which best fits the observed data in the case of ARB_CAL is a migration from Toskeria to Calabria with an effective size of 200-900 migrants and a local contribution to the new settlements of 5-20% (Table 2, Figure S3). Evolutionary scenarios with Ghegs as parental population fit models with a local contribution of 5-30%, corresponding to an effective size of 100-900 migrants. These results should however be considered less likely, since the observed DHS value matches the range of simulated values only in the tails of the distribution (Figure S3).

Evolutionary scenarios for ARB_SIC fit only models based on SNPs-data, with an effective size of 120 (or less) migrants from a Tosk or Tosk/Greek source population (Figure S3) and a local contribution higher than 50% (Table 2), which is unlikely unless extensive gene flows between source and local populations or complete population displacements are invoked. In the case of ARB_SIC, evolutionary scenarios with Gheghs as parental population do not fit any model (Table 2, Figure S3). In general, variation of the increment rate in the growth model did not affect appreciably the fit to the models (data not shown).

The relative contribution of local population increases when ARB_CAL were split in the three single Arbereshe communities (data not shown), as expected if considering the reduction in their relative population sizes (POL_AREA, n=24; POL_SW, n=36; VAL_CRA, n=46). Under the assumption that the lower the number the lower the probability to find shared haplotypes, DHS statistics (and consequently the distance from the source population) is expected to increase for low sizes (N<25). In a similar way, the splitting of Sicilian Arbereshe in the two single communities (CON_ENT, n=26; PIA_ALB, n=18) would implying a complete dilution of Arbereshe haplotypes in the current local genetic pools (data not shown).

Consistently with simulation results, admixture-like plots reveal ARB_CAL to be more clearly related with the Albanian populations, while ARB_SIC show a more complex pattern of genetic similarities involving partial links with Greek populations (Figure 3a). Among the two Sicilian Arbereshe communities, CON_ENT seems particularly the one mostly affected by a relevant Greek contribution (Figure 3b). It is also noteworthy the clustering of POL_SW with the Sicilian Arbereshe population of PIA_ALB, rather than with the other Calabrian Arbereshe communities of POL_AREA and VAL_CRA (Figure 3b).

Structuring of Arbereshe Y-chromosome haplogroup composition

The most frequent HGs in the whole Arbereshe dataset are E-V13 (13.3%), E-M123 (10.7%), I-P215(xM26,M223) (10.0%), I-M223 (10.0%), R-M17 (10.0%) and R-M269(xP311) (8.0%). Altogether these lineages encompass approximately the 62% of whole Y-chromosomes (Table S3). However, these average frequencies mask the real pattern of HGs distribution that appears when the two Arbereshe groups (and the single communities within them) were considered separately. In fact, despite the general accordance with the HG composition pattern described above, prevalent Y-HGs in ARB_CAL (E-V13, 16.9%; I-M223, 14.2%; Figure S1, Table S3) do not match those instead found in ARB_SIC (I-P215, 20.5%; E-M123, 18.2%; Figure S1, Table S3).

Y-chromosome haplogroup E-V13 and I-P215(xM26,M223) represent the most frequent lineages observed in ARB_CAL and ARB_SIC respectively. These haplogroups are two of the most common paternal lineages found in the Balkan Peninsula. Accordingly, E-V13 confirms to be significantly over-represented also in the Balkan populations here examined (and particularly among the Albanian Ghegs; 37.8%, p-value<0.001), reaching however noteworthy frequencies in all the three Calabrian Arbereshe communities analyzed (from 15.22% of VAL_CRA to 19.44% of POL_SW). Analogously, haplogroup I-P215, despite being frequent in the Balkans especially among Tosks and Corinthians (11.5% and 14.16% respectively), appears at significantly high frequencies in the Arbereshe populations of Sicily, and especially in the community of CON_ENT (27%, p-value=0.0304).

Some interesting patterns moreover appear when looking at HGs I-M223 for ARB_CAL and E-M123 for ARB_SIC, which represent the second most frequent lineages found in the two Arbereshe groups respectively (Table S3). Both of them are significantly over-represented in these linguistic minorities (with frequencies of 14.2%, p-value<0.001 and 18.2%, p-value=0.0015, respectively) compared to the considered Southern Italian and Balkan comparison populations. However, HG E-M123, despite confirming its high frequencies in the Arbereshe populations of Sicily (and particularly in CON_ENT; 19.2%, p-value=0.0403), appears significantly over-represented also in the Calabrian Arbereshe population of POL_SW (22.2%, p-value=0.0003). On the contrary, this lineage is completely absent in the other two Arbereshe communities of Calabria (namely VAL_CRA and POL_AREA). The opposite pattern has been observed for haplogroups I-M223. This lineage is indeed over-represented in the Calabrian Arbereshe communities of VAL_CRA and POL_AREA (17.4%, p-value=0.0074 and 25%, p-value=0.0065 respectively), being instead virtually absent in both the Sicilian Arbereshe and the remaining Calabrian Arbereshe population of POL_SW (Table S3).

The third most frequent lineage in ARB_CAL and ARB_SIC is haplogroup R-SRY10831.2, which has been found at comparatively high frequencies in both the two ethno-linguistic groups (9.43% and 11.35%).

Haplotype networks and age estimates

The five haplogroups mentioned above proved to significantly affect the genetic variability both within and between the Arbereshe groups of Sicily and Southern Italy. In order to investigate

further, we analyzed the STRs variation within these HGs by reconstructing phylogenetic networks and dating population-specific clusters of haplotypes (see Materials and Methods for more details).
Both networks of E-V13 (Figure S4) and I-P215(xM26,M223) haplotypes (Figure S5) show fairly compact star-like structures centered around Albanian (mainly Tosk) haplotypes. However, they also highlight the differential presence of Arbereshe-specific clusters of haplotype. More precisely, E-V13 exhibits a cluster (E-V13α) composed almost entirely (77.8%) of ARB_CAL haplotypes (Figure S4). On the other hand, I-P215 shows an almost exclusive (85.7%) ARB_SIC specific cluster (I-P215α, Figure S5). Both of these clusters reveal in their star-like structures signals of recent expansions. Their SD-based age estimates, dating at 469±118 YBP and 649±170 YBP respectively, are consistent with the times of Arbereshe migrations in Sicily and Southern Italy. In both E-V13α and I-P215α clusters, it is noteworthy that non-Arbereshe haplotypes were autonomously recognized as "outliers" by the jackknife-like omitting procedure of TMRCAs estimation, and consequently excluded from the age calculation (see Materials and Methods for details). These facts further attest the Arbereshe-specificity of the obtained haplotype clusters.

E-M123 and I-M223 networks reaffirm the presence of Arbereshe-specific clusters of haplotypes. In particular, HG I-M223 (Figure S6) mainly reflects its specificity for ARB_CAL (cluster I-M223α) being instead completely absent in ARB_SIC. On the contrary, HG E-M123 well emphasizes the within-ethnic group genetic variation by showing the co-existence of two distinct Arbereshe-specific clusters: respectively a 100% ARB_CAL (E-M123α) and a 100% ARB_SIC (E-M123β) (Figure S7). Age estimates are similar for both the two E-M123-identified clusters (1133±314 YBP and 883±228 YBP respectively), yet predating (as it happens also for the cluster I-M223α, 2122±548 YBP) the time of Arbereshe settlement in Sicily and Southern Italy.

Haplogroup R-SRY10831.2 proved to not significantly differ in frequency between the two Arbereshe groups. Consistently, ARB_CAL and ARB_SIC haplotypes form a shared Arbereshe-specific cluster (R-SRY10831.2α) within the corresponding network (Figure S8), which traces back at 556±146 YBP.

DISCUSSION

The Albanian-speaking Arbereshe embrace an important portion of the ethno-linguistic variability of Sicily and Southern Italy (and more generally of the whole Italy). Recently, this ethno-linguistic minority has particularly attracted the increasing attention of both bio-demographic (Tagarelli et al., 2007; Fiorini et al., 2007) and population genetic studies (Boattini et al., 2010; Capocasa et al., 2014). In fact, the documented history and traceable migration processes (especially if compared to other Italian ethno-linguistic groups), coupled with the condition of “marginality” due to cultural and geographic traits, has made it possible to ask new questions about the complex interplay between historical strata, geographic influences and cultural forces, in shaping the current distribution pattern of population genetic variability.

Previous bio-demographic and genetic studies have confirmed the important role that linguistic and cultural factors had in preserving the Balkan origins of the Arbereshe population, which indeed showed a substantial genetic discontinuity with the Italian genetic background (Tagarelli et al., 2007, Boattini et al., 2010). However, different extents in the genetic relationships between the Arbershe groups and both their source (Albanians and Greeks) and recipient (Sicialian and Calabrians) populations (Capocasa et al., 2014) have been also suggested and invoked as factors potentially affecting the within ethnic-group genetic variability.

In order to obtain a more fine-grained overview on the Arbereshe genetic diversity, in this study we compared and contrasted the paternal genetic composition of two Arbereshe groups, from Sicily and Calabria (Southern Italy) respectively, with the aim to more deeply assess the within-ethnic group population genetic structure. In addition, by directly comparing each ethno-linguistic isolate with both their geographic neighbors and source populations, we seek to gain new insights into the genetic history of each Arbereshe community and to explore the effects of cultural distinctiveness on their current pattern of genetic variability.

The analysis of within-population genetic variation (Table S4) revealed the lack of any significant reduction of gene diversity in the Arbereshe populations. With the partial exception of Contessa Entellina (which shows lower values for STRs data), all the Arbereshe communities here considered exhibite levels of gene diversity relatively high and comparable with the range of values obtained for the Balkan and Southern-Italian comparison populations. On one hand, the surname-based sampling strategy and the biodemographic-based grouping criteria (in the case of Calabrian Arbereshe) could have at least partially over-estimated the Arbereshe intra-population genetic diversity here obtained. Nevertheless, the high Y-chromosomal variability observed for all the Arbereshe surname-based communities (also at micro-geographical levels) is actually more likely to contain the ancestral genetic signature of the first migrants who settled Southern Italy from the Balkan Peninsula approximately 500 years ago (Boattini et al. 2010).

When Arbereshe were compared with Italian and Balkan populations, both SNP- (Figure 1) and STR-based analyses (Figure 2 and Figure S2) reveal the presence of geographic and cultural groupings, accounting for the genetic separation between the Balkan and Italian Peninsulas on one hand, and the shared genetic ancestry of Arbereshe groups with the Balkan populations on the other hand (see also AMOVA results, Table S5). On a broad-scale geographic perspective Balkan and

Southern Italy previously showed to constitute a quite consistent homogeneous group within the Mediterranean genetic landscape (Sarno et., al. 2014). However, from a micro-geographical point of view they actually reveal higher degrees of cultural and genetic complexity. In particular, within the specific context of the Italian-Balkan genetic link, cultural factors confirm to have played an important role for the Arbereshe communities to preserve their Balkan genetic background and to regulate their genetic exchanges with the Italian neighboring populations. Nevertheless, as preliminarily suggested in Capocasa et al., (2014) and specifically addressed in this study (Figure 3a), Calabrian and Sicilian Arbereshe show within-ethnic-group differences in their processes of ethnogenesis. In fact, the proportions of Balkan and Italian genetic admixtures - estimated by means of DHS-based simulation approach (Tofanelli et al., 2009) - vary quite substantially between the Calabrian and Sicilian ethnic-groups. In general, while the Arbereshe of Calabria confirm their Albanian origin – the most likely candidate source population being Tosks – and show relatively low levels of admixture with Italian local populations (Table 2, Figure S3**)**, on the other hand the scenario suggested for Sicilian Arbereshe implies more complex patterns of genetic inheritances, where composite founder events (possibly involving population replacements) and higher levels of admixture with local populations can be hypothesized. Although higher number of samples are needed to explore plausible scenarios also at the level of single communities, these results however suggest different genetic histories between the two different ethno-linguistic groups.

When single Arbereshe communities were directly compared on a fine-grained level by means of admixture-like plots, Contessa Entellina (CON_ENT) seems particularly the one showing greater genetic links with Greeks (Figure 3b), thus suggesting al least in part its genetic discontinuity compared to the other Arbereshe populations. This result can be explained if considering some historical documents (referable to the 1520-1521 AD) attesting that the Arbereshe municipality of Contessa Entellina (Palermo, Sicily) was repopulated with hundred families coming from the Greek island of Andros, in the Peloponnese (Giunta, 2003). On the contrary, the other Arbereshe community of Sicily (PIA_ALB), similarly to all the three Calabrian Arbereshe groups (VAL_CRA, POL_SW, POL_AREA), confirm a prevalent contribution from the Albanian Tosks (Figure 3b). This scenario is in accordance with our historical knowledge about the foundation of the Arbereshe ethno-linguistic minorities of Sicily and Southern Italy. However, it is noteworthy that POL_SW, rather than clustering with the other Calabrian Arbereshe populations, shows major links with the Sicilian Arbereshe communities (and particularly with PIA_ALB).

This picture is also reflected in the different patterns of HG composition observed both between the two Arbereshe groups, as well as among the single communities within them (Figure S1 and Table S3). Whereas some haplogroups were indeed found at comparatively high frequencies in both

Sicilian and Calabrian Arbereshe (R-SRY10831.2), other paternal lineages revealed to be differentially present in ARB_CAL (E-V13) and ARB_SIC (I-P215), or to differentially link some communities (POL_SW and ARB_SIC, E-M123) compared to other (VAL_CRA and POL_AREA, I-M223). Despite their common origin in the Balkans and the unifying cultural and linguistic traits, on a local-scale perspective the differential presence between the two Arbereshe groups (and among the communities therein) of specific lineages (at relatively high frequencies) presumptively reflects different population histories and founders events, as well as different patterns of population contacts and admixture or cultural isolation and genetic drift.

These results led us to the last aim of our research which was trying to disentangle the different genetic traces still present in the gene pool of current Arbereshe populations. At this purpose, we especially focused on those haplogroups that i) were found at significantly high frequencies in the Arbereshe genetic pool – being more likely to reflect the results of founder events - and ii) that additionally proved to distinguish the two Arbereshe groups and the single communities within them – being more likely to account for common or different genetic histories. Five haplogroups were resulted predominant in the Arbereshe genetic pool (E-V13, 13.3%; E-M123, 10.7%; I-P215, I-M223 and R-M17, 10%) and each of them showed differential frequency and distribution patterns within and between the two Arbereshe groups of Sicily and Southern Italy (Calabria).

Haplogroup E-V13 accounts for almost the 17% of Y-chromosomes in the Arbereshe of Calabria, while not reaching the 5% in the Arbereshe of Sicily. On the contrary, haplogroup I-P215 reaches frequencies of more than 20% in the Sicilian Arbereshe, being instead found at only 5% in the Calabrian ones (Table S3). Both E-V13 and I-P215 (xM26,M223) represent two of the most frequent lineages of the Balkan Peninsula. Accordingly they have been observed at significantly high frequencies in all the Albanian and Greek population groups here examined (20-34% for E-V13 and 7-10% for I-P215 respectively; Table S3). The typical Balkan origin of these lineages is moreover attested by the star-like structure of their corresponding haplotype networks (Figure S4 and Figure S5), which appear centered around Balkan haplotypes, with Arbereshe ones that instead occupy mainly peripheral positions. Although the whole history of a population cannot be identified with only single haplogroups, the values of Y-STR diversity (HD) within specific lineages can however be informative for investigating specific migration processes, under the assumption that higher diversity values should be expected in the populations where the lineage originated, and lower values might be insetad observed in the populations derived from the migration process (De Filippo et al. 2010). Interestingly, the Y-STR haplotype diversity associated with E-V13 and I-P215 lineages appears higher in the Balkan populations (0.990-0.995 for E-V13 and 0.985-0.995 for I-P215) than in Arbereshe of Calabria (0.948±0.039) and Arbereshe of Sicily (0.962±0.064) for E-

V13 and I-P215 respectively. In addition, E-V13 ARB_CAL and I-P215 ARB_SIC haplotypes form highly Arbereshe-specific clusters (E-V13α ARB_CAL cluster, Figure S4 and a I-P215α ARB_SIC cluster, Figure S5) that date back at 469±118 YBP and 649±170 YBP. These time estimates are quite consistent with our state of knowledge about the times and the diffusion patterns of Arbereshe migrations in Sicily and Southern Italy. This suggests that the expansion of both clusters might have occurred at around the same time, but probably as a result of different population movements.

Haplogroup R-SRY10831.2 has been found at comparatively high frequencies in both Calabrian and Sicilian Arbereshe (9.43% and 11.35%), representing the third most frequent Y- lineage in both of the two ethno-linguistic groups. This HG is particularly common in Eastern Europe. Accordingly it has been found at relatively high frequencies also in the Balkan populations here analyzed (especially among Greeks where it accounts for ~14% of the Y-chromosomes). Its haplotype network, despite showing a relative complex structure with many different and independent branches, actually reveals the presence of one Arbereshe-specific cluster (R-SRY10831.2α, Figure S8), which is shared between ARB_CAL and ARB_SIC haplotypes and date at 556±146 YBP. It is also noteworthy that the Y-STR haplotype diversity associated to this haplogroup in Balkan populations (0.990-1.000) appears once again higher than the one observed in the Arbereshe ethno-linguistic group (0.980±0.013). Altogether these results presumptively suggest genetic histories at least initially shared between the two Arbereshe groups, that at some point in the past might have then evolved differently.

Obviously we are not identifying the whole Arbereshe population history with single haplogroups, but haplotype clusters within specific haplogroups are actually more likely to represent or contain traces of the past population history. During the foundation and differentiation of populations, haplotype clusters within certain lineages might have emerged in response of specific demographic events, often becoming population-specific (Balanovsky et al., 2011).

Haplogroups E-M123 and I-M223 reveal a different aspect of the Arbereshe Y-chromosome genetic make-up. These HGs show quite distinct continental distributions. I-M223 reaches its highest frequencies in central Europe (especially Germany), despite being present also in Russia, Greece, Italy and around the Black Sea. E-M123 is instead widely scattered in the South-Eastern Mediterranean area, with significant peaks of frequency in populations from Horn of Africa, Southern Levant and Anatolia. Both of these haplogroups attest the differential presence of Arbereshe-specific clusters of haplotypes (Figure S6 and Figure S7). However, their time estimates predate the settlement of Arbereshe communities in Sicily and Southern Italy and possibly suggest local or more ancient introgressions in the Arbereshe genetic pool. HG E-M123 particularly was found over-represented in Sicilian Arbereshe, being present at significantly high frequency also in the Calabrian Arbereshe population of POL_SW. Interestingly, its correspondent network however shows that E-M123 haplotypes from ARB_CAL and ARB_SIC actually form two highly-specific and diverging clusters: ARB_CAL E-M123α and ARB_SIC E-M123β respectively (Figure S7).

CONCLUSIONS

In summary, this study shows how ethno-linguistic minorities may provide a powerful tool to improve our understanding of the patterns of genetic variation at finer geographical and temporal scales, and to cast new lights on the impact of cultural, historical and geographic relationships between human groups on the current population variability.

The Arbereshe communities of Sicily and Southern-Italy, despite sharing a common cultural and genetic background, actually revealed different micro-evolutionary histories. More precisely, our analyses suggest different founders events and different extents of cultural isolation and/or local admixture, both within and among the Sicilian- and Calabrian-Arbereshe ethno-linguistic groups. These factors have presumptively played a pivotal role in the shaping of the current genetic composition in the Arbereshe communities, explaining at least in part the emergence of differences in their HGs composition and accounting for the differential preservation of specific paternal lineages by the two Arbereshe groups. Future researches, by increasing the number of samples for each community and/or by exploiting the perspectives offered by complementary genetic systems (i.e. mtDNA genomes and autosomal whole-genome SNPs data), will offer the possibility to achieve an even more detailed picture of the Arbereshe genetic history and population variability, within a multi-disciplinary effort to integrate the genomic overview with the results offered by other disciplines, and in particular by linguistics.

REFERENCES

Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, Pocheshkhova E, Haber M, Platt D, Schurr T, Haak W, Kuznetsova M, Radzhabov M, Balaganskaya O, Romanov A, Zakharova T, Soria Hernanz DF, Zalloua P, Koshel S, Ruhlen M, Renfrew C, Wells RS, Tyler-Smith C, Balanovska E; Genographic Consortium (2011) Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol. 28:2905-2920. doi: 10.1093/molbev/msr126.

Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, Choi Y, van Duijn K, Vermeulen M, Brauer S, Decorte R, Poetsch M, von Wurmb-Schwark N, de Knijff P, Labuda D, Vézina H, Knoblauch H, Lessig R, Roewer L, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet. 87:341-353. doi:10.1016/j.ajhg.2010.08.006.

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

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution, 16, 37–48.

Boattini A, Luiselli D, Sazzini M, Useli A, Tagarelli G, Pettener D (2010) Linking Italy and the Balkans. A Y-chromosome perspective from the Arbereshe of Calabria. Ann Hum Biol. 38:59-68. doi: 10.3109/03014460.2010.491837.

Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, Sanz P, Yang-Yao D, Manry J, Ciani G, Luiselli D, Quintana-Murci L, Comas D, Pettener D, the Genographic Consortium (2013). Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. PLoS One. 8:e65441. doi: 10.1371/journal.pone.0065441.

Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, Bradley DG, Gusmão L, Winney B, Bodmer W, Vennemann M, Coia V, Scarnicci F, Tofanelli S, Vona G, Ploski R, Vecchiotti C, Zemunik T, Rudan I, Karachanak S, Toncheva D, Anagnostou P, Ferri G, Rapone C, Hervig T, Moen T, Wilson JF, Capelli C (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci. 279:884-892. doi: 10.1098/rspb.2011.1044.

Capocasa M, Anagnostou P, Bachis V, Battaggia C, Bertoncini S, Biondi G, Boattini B, Boschi I, Brisighelli F, Calò CM, Carta M, Coia V, Corrias L, Crivellaro F, Ferri G, Francalacci P, Franceschi ZA, Luiselli D, Morelli L, Rickards O, Robledo R, Sanna D, Sanna E, Sarno S, Tofanelli S, Vona G, Pettener D and Destro Bisol G (2014) Linguistic, geographic and genetic isolation: a collaborative study on Italian populations. J Anthropol. Sci. 92:1-32 doi:10.4436/JASS.92001.

De Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdóttir ED, Bostoen K, Nyambe T, Beyer K, Schreiber H, de Knijff P, Luiselli D, Stoneking M, Pakendorf B (2010) Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol. 28:1255-1269. doi: 10.1093/molbev/msq312.

Destro-Bisol G, Anagnostou P, Batini C, Battaggia C, Bertoncini S, Boattini A, Caciagli L, Caló MC, Capelli C, Capocasa M, Castrí L, Ciani G, Coia V, Corrias L, Crivellaro F, Ghiani ME, Luiselli D, Mela C, Melis A, Montano V, Paoli G, Sanna E, Rufo F, Sazzini M, Taglioli L, Tofanelli S, Useli A, Vona G, Pettener D (2008) Italian isolates today: geographic and linguistic factors shaping human biodiversity. J Anthropol Sci. 86:179-188.

Dray S, Dufour AB, Chessel D (2007) The ade4 package-II: Two-table and K-table methods. R News. 7:47-52.

Dray S, Dufour AB (2007) The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software. 22:1-20.

Esko T, Mezzavilla M, Nelis M, Borel C, Debniak T, Jakkula E, Julia A, Karachanak S, Khrunin A, Kisfali P, Krulisova V, Aušrelé Kučinskiené Z, Rehnström K, Traglia M, Nikitina-Zake L, Zimprich F, Antonarakis SE, Estivill X, Glavač D, Gut I, Klovins J, Krawczak M, Kučinskas V, Lathrop M, Macek M, Marsal S, Meitinger T, Melegh B, Limborska S, Lubinski J, Paolotie A, Schreiber S, Toncheva D, Toniolo D, Wichmann HE, Zimprich A, Metspalu M, Gasparini P, Metspalu A, D'Adamo P (2012) Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. Eur J Hum Genet. 21:659-665. doi:10.1038/ejhg.2012.229.

Excoffier L, Laval G, Schneider S (2007) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 1:47-50.

Ferri G, Tofanelli S, Alù M, Taglioli L, Radheshi E, Corradini B, Paoli G, Capelli C, Beduschi G (2010) Y-STR variation in Albanian populations: implications on the match probabilities and the genetic legacy of the minority claiming an Egyptian descent. Int J Legal Med. 124:363-370. doi:10.1007/s00414-010-0432-x.

Ferri G, Alù M (2012) Development of six-Y-SNPs assay for forensic analysis in European population. DNA in Forensics 2012, 5th International EMPOP Meeting- 8th International Forensic Y-User Workshop, Innsbruck.

Fiorini S, Tagarelli G, Boattini A, Luiselli D, Piro A, Tagarelli A, Pettener D (2007) Ethnicity and evolution of the biodemographic structure of Arbereshe and Italian populations of the Pollino area, Southern Italy (1820–1984). Amer Anthropol 109:735–746.

Gayden T, Regueiro M, Martinez L, Cadenas AM, Herrera RJ (2008) Human Y-chromosome haplotyping by allele-specific polymerase chain reaction. Electrophoresis. 29:2419-2423. doi:10.1002/elps.200700702.

Giunta F (2003). Albanesi in Sicilia. In: Mandalà M. (ed.), Albanesi in Sicilia, Palermo: Mirror.

Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, Morling N, Prinz M, Roewer L, Tyler-Smith C, Schneider PM; DNA Commission of the International Society of Forensic Genetics (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci Int. 157:187-197.

Heutink P, Oostra BA (2002) Gene finding in genetically isolated populations. Hum Mol Genet. 11:2507-2515.

Kristiansson K, Naukkarinen J, Peltonen L (2008) Isolated populations and complex disease gene identification. Genome Biol. 9:109. doi: 10.1186/gb-2008-9-8-109.

Laval G, SanCristobal M, Chevalet C (2002). Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet Sel Evol 34: 481-507.

Meyer S, Weiss G, von Haeseler A (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics, 152, 1103–1110.

Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 16:1215.

Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK (2006) Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci. 51:64-75.

Neto D, Montiel R, Bettencourt C, Santos C, Prata MJ, Lima M (2007) The African contribution to the present-day population of the Azores Islands (Portugal): analysis of the Y chromosome haplogroup E. Am J Hum Biol. 19:854-860.

Onofri V, Alessandrini F, Turchi C, Pesaresi M, Buscemi L, Tagliabracci A (2006) Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. Forensic Sci Int. 57:23-35.

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Reynolds JB, Weir BS, Cockerham CC (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767-779.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet. 78:202-221.

Sarno S, Boattini A, Carta M, Ferri G, Alù M, Yang Yao D, Ciani G, Pettener D, Luiselli D (2014) An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of Sicily and Southern Italy. PLoS One (*submitted*).

Tagarelli A (2004) Studio antropologico della comunità arbëreshe della provincia di Torino. eds: Librare, pp. 47-66.

Tagarelli G, Fiorini S, Piro A, Luiselli D, Tagarelli A, Pettener D (2007) Ethnicity and biodemographic structure in the Arbëreshe of the province of Cosenza, southern Italy, in the XIX century. Coll Antropol. 31:331-338.

Tofanelli S, Bertoncini S, Castrì L, Luiselli D, Calafell F, Donati G, Paoli G (2009) On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. Mol Biol Evol. 26:2109-2124. doi: 10.1093/molbev/msp120.

Toso F (2014) The study of language islands: an interdisciplinary approach. J Anthropol. Sci. doi:10.4436/JASS.92002

Veeramah KR, Tönjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, Stumvoll M (2011) Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. Eur J Hum Genet. 19:995-1001. doi:10.1038/ejhg.2011.65.

Venables WN, Ripley, BD (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Zangari D (1941) Le colonie italo-albanesi di Calabria. Naples: Casella Editore.

FIGURES

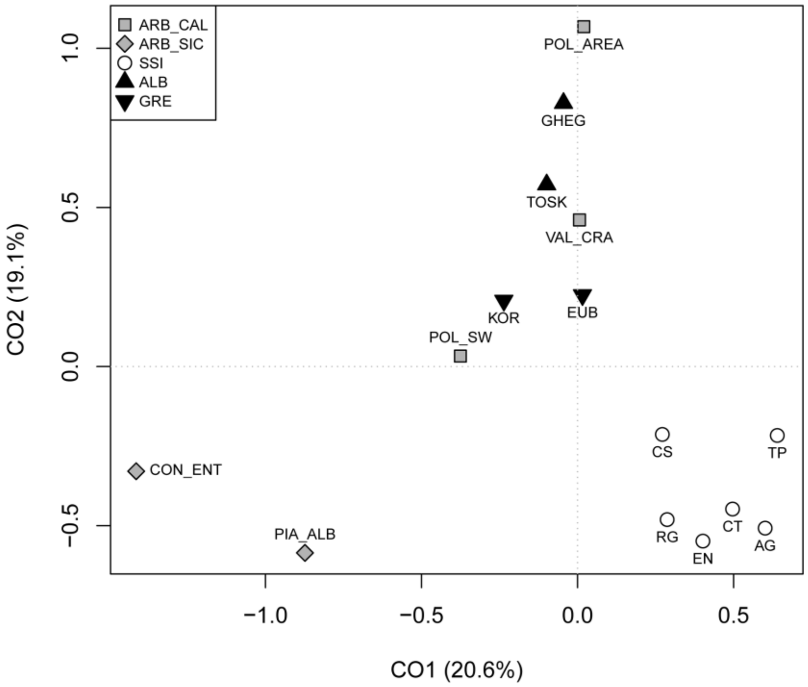


Figure 1. **Principal Component Analysis (PCA) based on Y-chromosome haplogroup frequencies**. Population codes as in Table 1. Symbols and colour codes as in the legends at the top-left. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; SSI: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks.

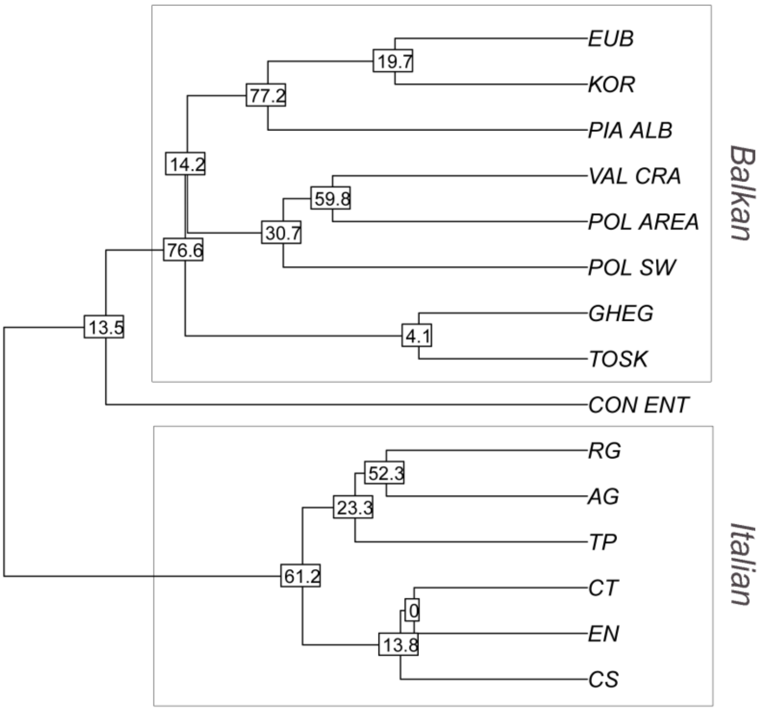


Figure 2. Hierarchical Clustering (1,000 bootstraps) computed on STR-based distances. Population codes as in Table . Balkan (plus Arbereshe) and Italian clusters are marked by boxes. Bootstrap values are reported in percentage for each nodes of the tree.

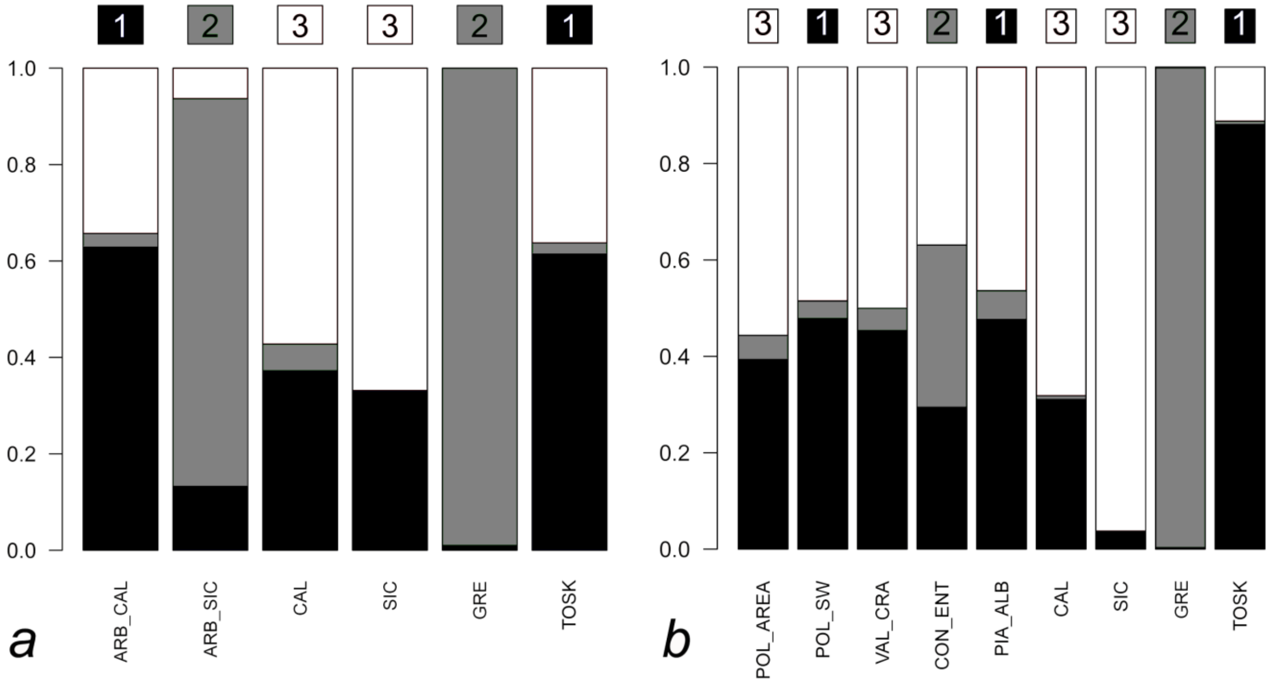


Figure 3. Admixture-like barplots for the Arbereshe groups (a) and the single arbereshe communities (b). The barplots represent DAPC-based posterior membership probabilities for each of the considered populations to belong at each of the inferred cluster (*mclust* algorithm). The affiliation of each population to a given cluster and its corresponding colour code are represented by letters (within coloured squares) on the top of each bar. Labels: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; CAL: Italians of Calabria, SIC: Italians of Sicily; GRE: Greeks; TOSK: Tosks. Codes for single Arbereshe communities as in Table 1.

TABLES

Table 1. List of the Arbereshe and comparison populations included in the present study. For each population the geographic coordinates, the region and the language affiliation are detailed. N= sample size.

Population	Pop Code	N	Longitude	Latitude	Geographic location	Language	Reference
Arbereshe of Calabria	**ARB_CAL**	**106**					
Pollino Massif Area	POL_AREA	24	16.43	39.90	Calabria (Cosenza province)	Albanian-Tosk	This study
Pollino South-West	POL_SW	36	16.12	39.75	Calabria (Cosenza province)	Albanian-Tosk	This study
Valley of Crati	VAL_CRA	46	16.43	39.58	Calabria (Cosenza province)	Albanian-Tosk	This study
Arbereshe of Sicily	**ARB_CAL**	**44**					
Piana degli Albanesi	PIA_ALB	18	13.28	38.00	Sicily (Palermo province)	Albanian-Tosk	This study
Contessa Entellina	CON_ENT	26	13.18	37.73	Sicily (Palermo province)	Albanian-Tosk	This study
Sicily and Southern Italy	**SSI**	**263**					
Trapani	TP	34	12.52	38.02	Sicily	Southern-Italian	Sarno et al. 2014
Agrigento	AG	45	13.59	37.32	Sicily	Southern-Italian	Boattini et al. 2013
Enna	EN	40	14.27	37.57	Sicily	Southern-Italian	Sarno et al. 2014
Ragusa	RG	45	15.01	37.01	Sicily	Southern-Italian	Boattini et al. 2013
Catania	CT	52	15.07	37.52	Sicily	Southern-Italian	Boattini et al. 2013
Cosenza	CS	47	16.25	39.30	Calabria	Southern-Italian	Sarno et al. 2014
Albanian	**ALB**	**223**					
Ghegh	GHEG	119	19.51	42.05	Albania (Northern)	Albanian-Ghegh	This study
Tosk	TOSK	104	19.95	40.70	Albania (Southern)	Albanian-Tosk	This study
Greek	**GRE**	**209**					
Corinth	KOR	113	22.93	37.93	Greece (Peloponnese)	Greek	Anagnostou p.c.
Eubea	EUB	96	24.00	38.50	Greece (Central)	Greek	Anagnostou p.c.

Table 2. Fitting of observed and expected values of the DHS statistics for STR and SNP genetic systems under a stationary model. A percentage contribution of local versus putative source population was accepted as the most likely model when the observed value (*Obs DHS*) fell within a two standard deviations interval (in red) of the simulated distribution at time t=20 generations.

Calabria

		% Local contribution					
Marker	***Source pop***	***5***	***10***	***20***	***30***	***50***	***Obs DHS***
SNP	Tosks	**0.0370**	0.0809	0.1957	0.3054	0.4993	**0.0234**
		0.0090	0.0291	0.0828	0.1441	0.2527	
	Ghegs	**0.0370**	0.0687	0.1459	0.2165	0.4209	**0.0112**
		0.0094	0.0255	0.0702	0.0626	0.1348	
	Tosks-Greek	**0.0492**	0.1187	0.2441	0.3664	0.5592	**0.0315**
		0.0184	0.0524	0.1239	0.1988	0.3409	
STR	Tosks	0.6751	0.7967	**0.8915**	0.9252	0.9594	**0.8605**
		0.5726	0.7483	**0.8424**	0.8796	0.9140	
	Ghegs	0.6211	0.7448	0.8187	**0.8979**	0.9379	**0.8310**
		0.5468	0.6845	0.8093	**0.8199**	0.8629	
	Tosks-Greek	0.8415	0.8973	0.9491	0.9659	0.9815	**0.9034**
		0.7988	0.8624	0.9183	0.9406	0.9619	

Sicily

		% Local contribution					
Marker	***Source pop***	***5***	***10***	***20***	***30***	***50***	***Obs DHS***
SNP	Tosks	0.0060	0.0236	0.0681	0.1079	**0.2135**	**0.1865**
		0.0000	0.0023	0.0188	0.0288	**0.0903**	
	Ghegs	0.0105	0.0267	0.0652	0.0773	0.1841	**0.2293**
		0.0010	0.0055	0.0229	0.0313	0.0421	
	Tosks-Greek	0.0140	0.0337	0.0845	0.1401	**0.2653**	**0.1968**
		0.0021	0.0116	0.0356	0.0564	**0.1362**	
STR	Tosks	0.4861	0.6105	0.7649	0.8316	0.8987	**0.9825**
		0.3988	0.5591	0.7088	0.7833	0.8504	
	Ghegs	0.4379	0.5810	0.7038	0.7760	0.8537	**1.0000**
		0.3128	0.5319	0.6475	0.7123	0.7831	
	Tosks-Greek	0.6732	0.7805	0.8810	0.9188	0.9490	**0.9770**
		0.5562	0.7428	0.8472	0.8889	0.9224	

SUPPLEMENTARY MATERIALS

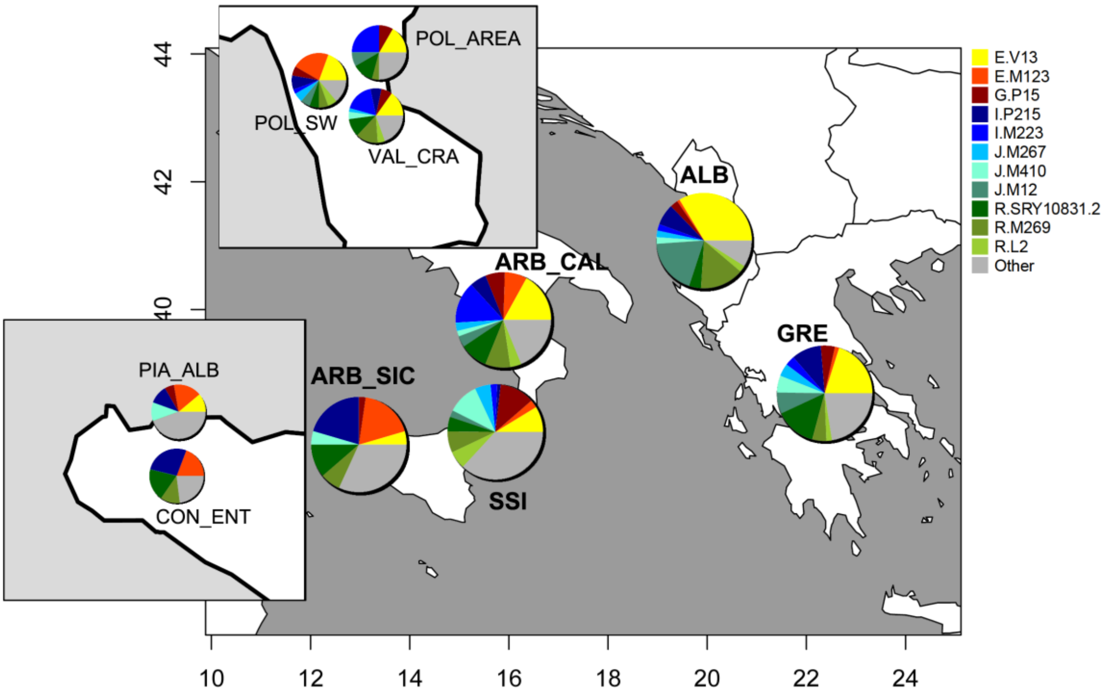


Figure S1. Geographic distribution of Y-chromosome main lineages in the studied populations. Only haplogroups reaching frequencies of at least 3% in the combined dataset were plotted. Colour code for Y-chromosome lineages as in the legend at the top right. Labels: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; SSI: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks. Haplogroup composition for each community of Calabrian and Sicilian Arbereshe are detailed in the boxes at the top and the bottom-left of the plot respectively (Arbereshe communities codes as in Table 1).

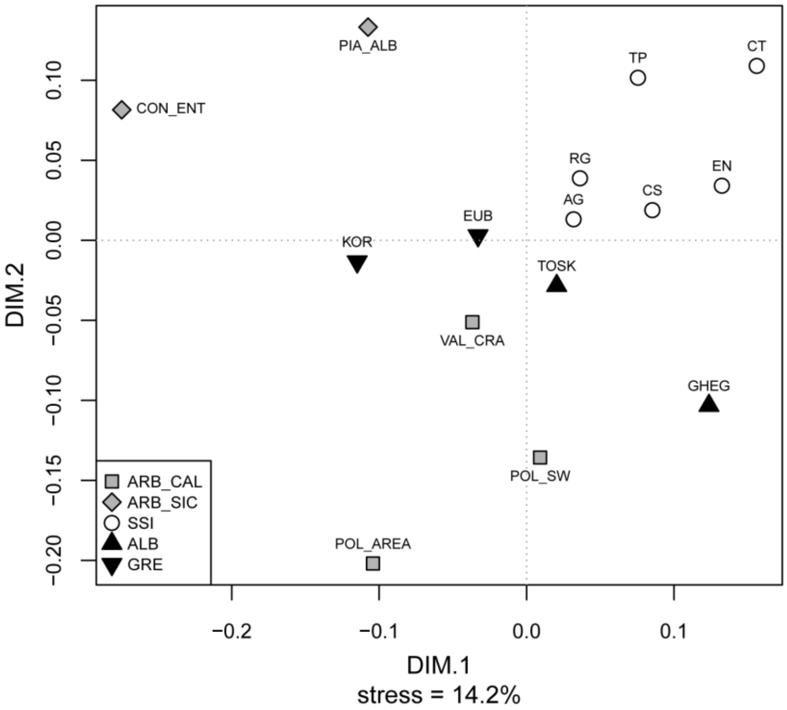


Figure S2. Non-metric MDS plot based on STRs distances. Population codes as in Table 1. Symbols and colour codes for the ethnic-group affiliations as in the legends at the bottom-left. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; SSI: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks.

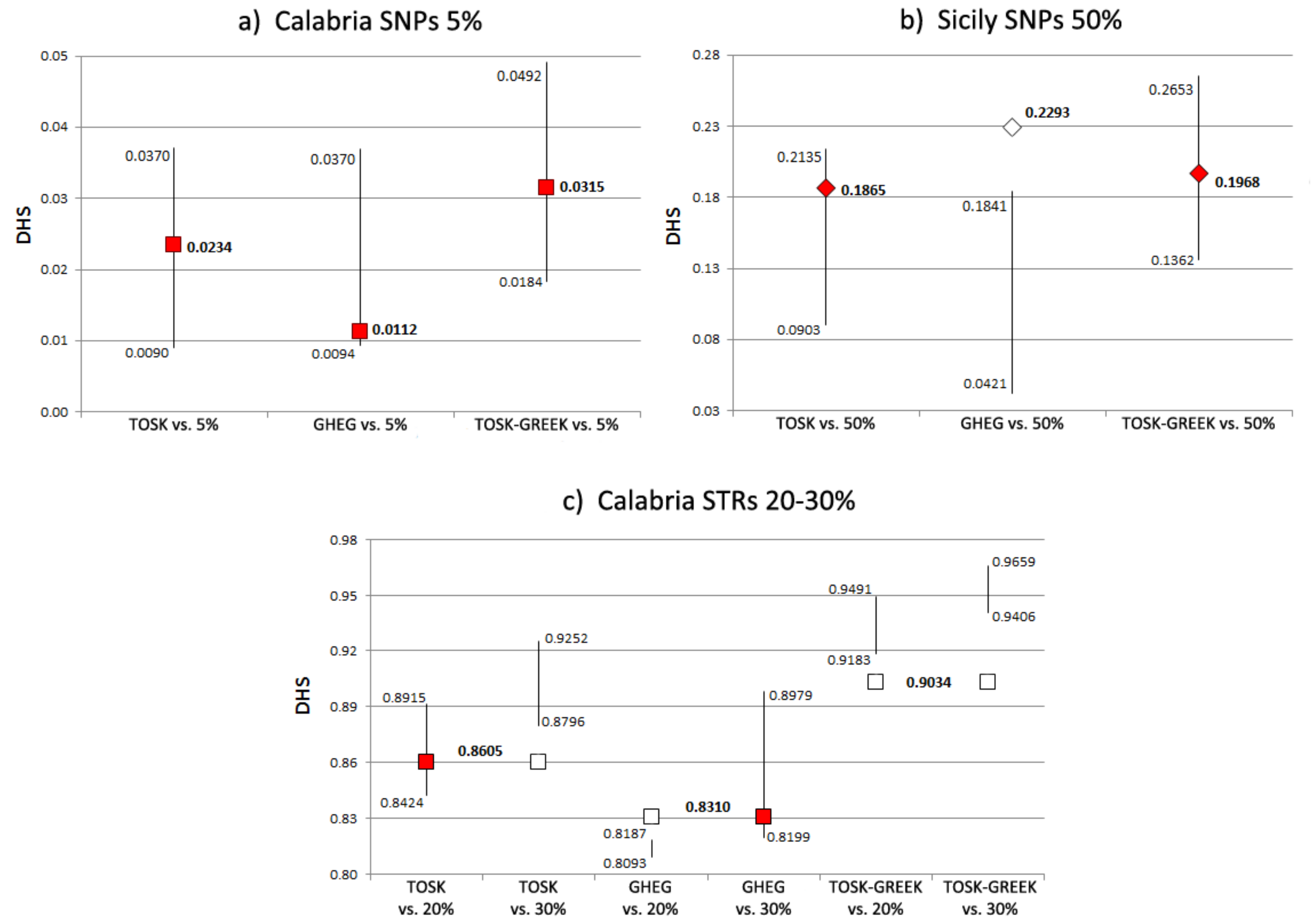


Figure S3. **Plots of observed and expected DHS values for the fitting models.** Evolutionary scenarios with the three putative source populations (Tosks, Ghegs, and a 50% random blend of Tosks and Greek) were compared for those models showing at least one fit in the simulations results (red in Table 2): a) Calabria SNPs 5%, b) Sicily SNPs 50% and c) Calabria STRs 20-30%. The standard deviations intervals of the simulated distributions are represented by vertical bars. Observed DHS values (in bold) are represented by symbols filled in red or white depending on whether or not they match the simulated intervals. Squares and rhombus stand for Calabrian and Sicilian Arbershes respectively.

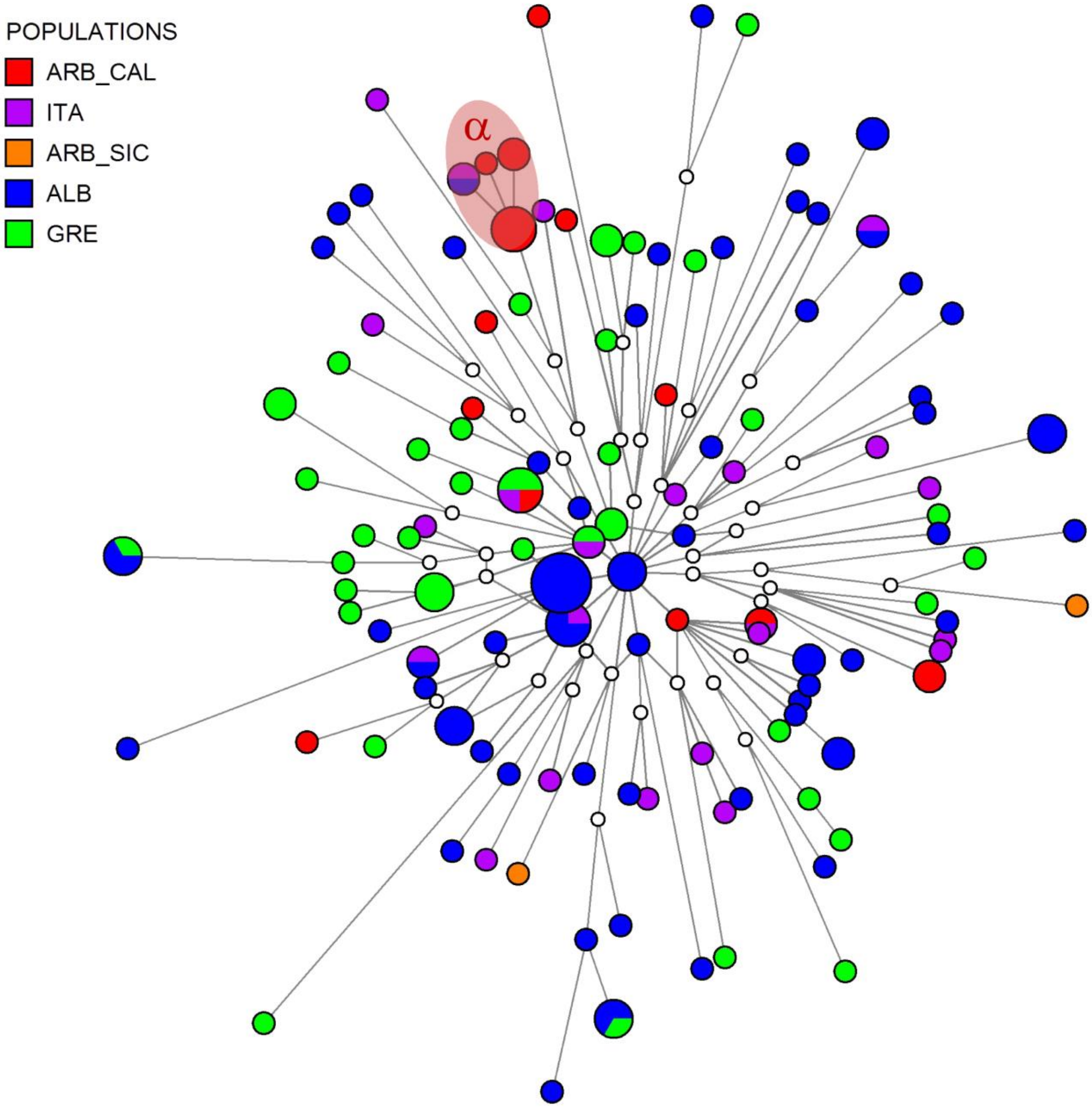


Figure S4. Phylogenetic haplotypes network of haplogroup E-V13. Each circle represents a different haplotype. Circles' areas are proportional to haplotype frequencies. White circles represent median vectors. Population colour codes as in the legends at the top-left. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; ITA: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks. Red oval designate the Calabrian Arbereshe-specific cluster (E-V13α) selected for age estimate.

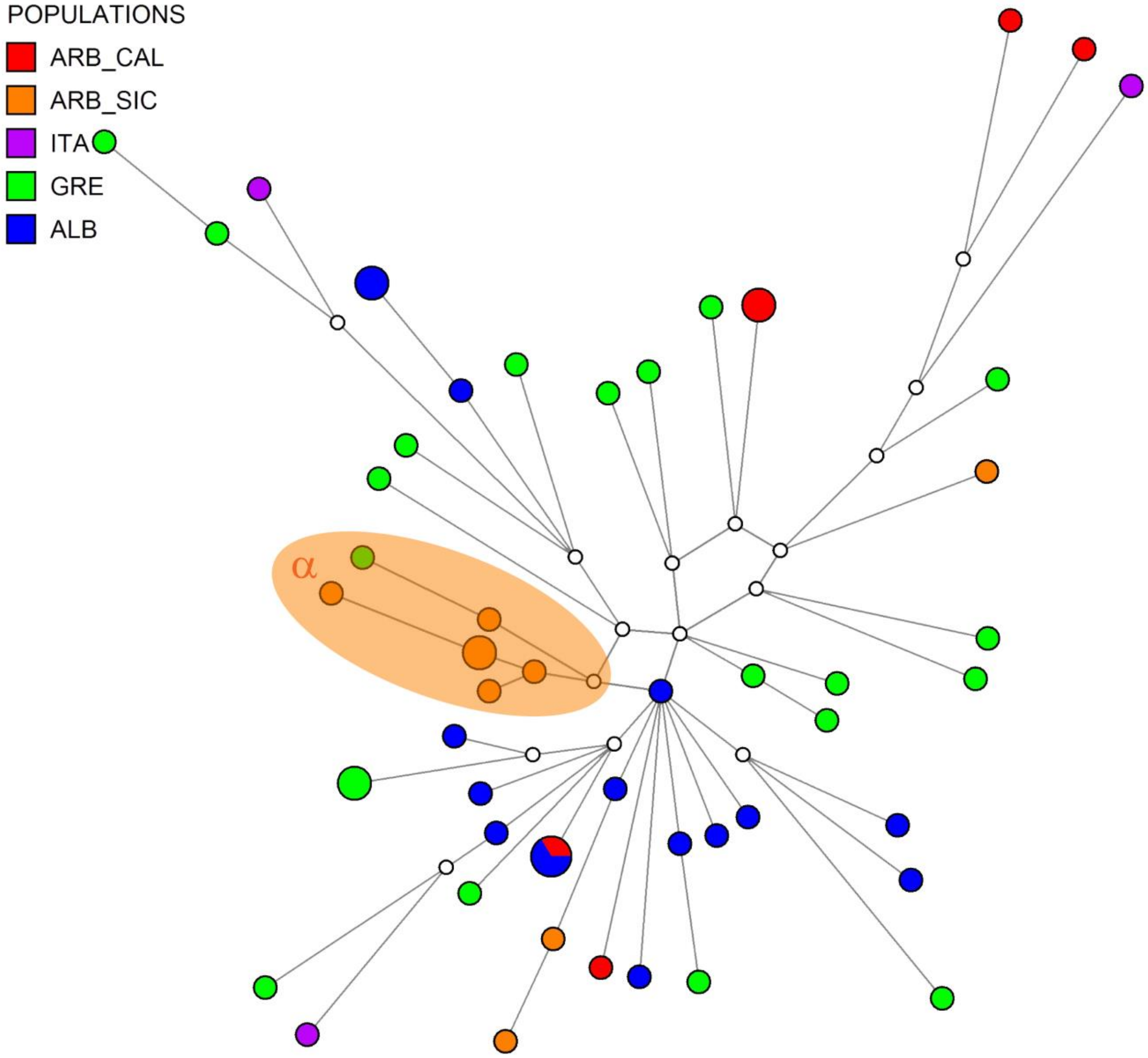


Figure S5. Phylogenetic haplotypes network of haplogroup I-P215(xM26,M223). Each circle represents a different haplotype. Circles' areas are proportional to haplotype frequencies. White circles represent median vectors. Population colour codes as in the legends at the top-left. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; ITA: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks. Orange oval designate the Sicilian Arbereshe-specific cluster (I-P215α) selected for age estimate.

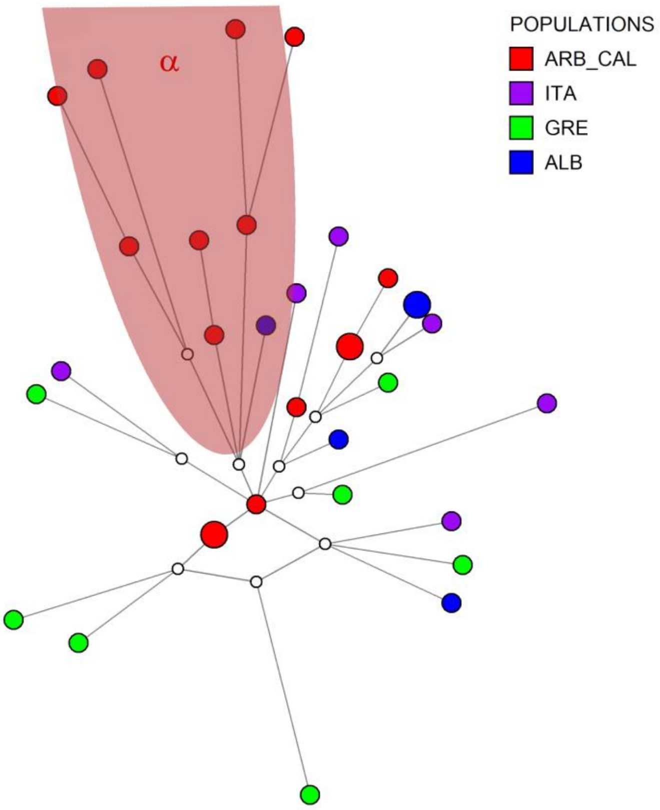


Figure S6. Phylogenetic haplotypes network of haplogroup I-M223. Each circle represents a different haplotype. Circles' areas are proportional to haplotype frequencies. White circles represent median vectors. Population colour codes as in the legends at the top-right. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; ITA: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks. Red oval designate the Calabrian Arbereshe-specific cluster (I-M223α) selected for age estimate.

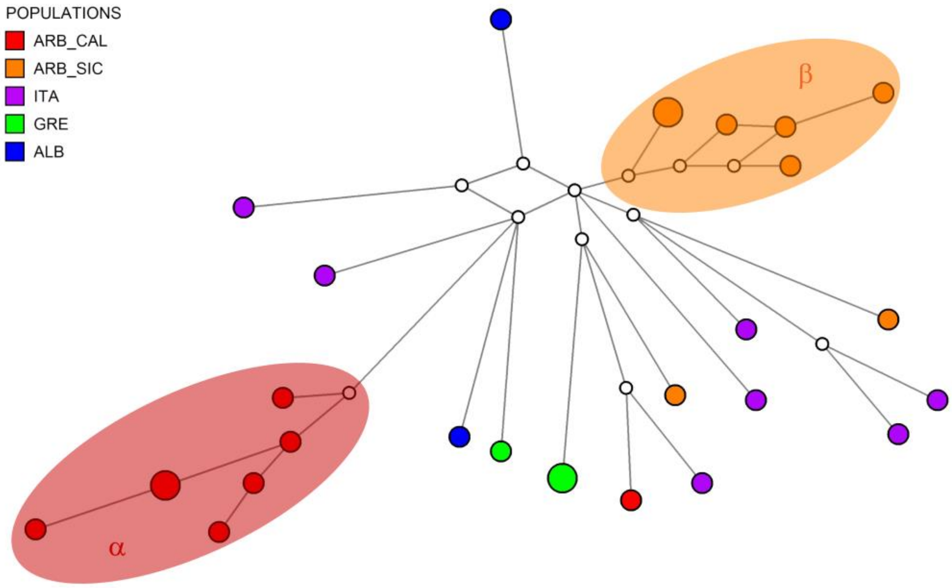


Figure S7. Phylogenetic haplotypes network of haplogroup E-M123. Each circle represents a different haplotype. Circles' areas are proportional to haplotype frequencies. White circles represent median vectors. Population colour codes as in the legends at the top-left. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; ITA: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks. Red and orange ovals designate respectively the Calabrian (E-M123α) and Sicilian (E-M123β) Arbereshe-specific clusters selected for age estimate.

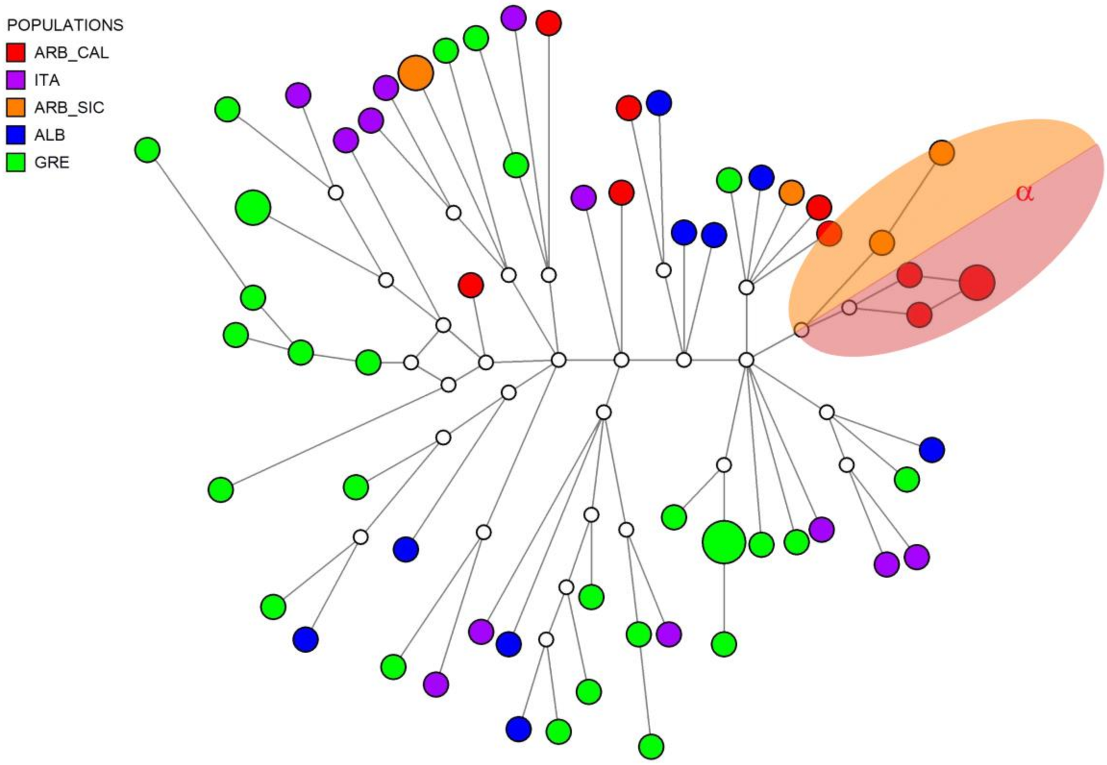


Figure S8. Phylogenetic haplotypes network of haplogroup R-SRY10831.2. Each circle represents a different haplotype. Circles’ areas are proportional to haplotype frequencies. White circles represent median vectors. Population colour codes as in the legends at the top-left. Legend abbreviations: ARB_CAL: Arbereshe of Calabria, ARB_SIC: Arbereshe of Sicily; ITA: Sicilians and Southern-Italians, ALB: Albanians, GRE: Greeks. Red/orange oval designate the shared Calabrian/Sicilian Arbereshe-specific clusters selected for age estimate.

Table S1. Sets of simulated parameters for the stationary model.

The extended version with the simulation parameters also for the growth model will be provided in the online manuscript once published

	Sicily						Calabria					
Marker	Incr	Migrants	% local	N X	N Y	Output file	Incr	Migrants	% local	N X	N Y	Output file
vs. TOSK												
STR	I=1	1998	5%	35 000	111	Tosk_Sic_STR_M5.csv	I=1	828	5%	35 000	46	Tosk_Cal_STR_M5.csv
		999	10%	35 000	111	Tosk_Sic_STR_M10.csv		414	10%	35 000	46	Tosk_Cal_STR_M10.csv
		444	20%	35 000	111	Tosk_Sic_STR_M20.csv		184	20%	35 000	46	Tosk_Cal_STR_M20.csv
		259	30%	35 000	111	Tosk_Sic_STR_M30.csv		107	30%	35 000	46	Tosk_Cal_STR_M30.csv
		111	50%	35 000	111	Tosk_Sic_STR_M50.csv		46	50%	35 000	46	Tosk_Cal_STR_M50.csv
vs. GREECE/TOSK												
STR	I=1	1998	5%	35 000	111	GreeceTosk_Sic_STR_M5.csv	I=1	828	5%	35 000	46	GreeceTosk_Cal_STR_M5.csv
		999	10%	35 000	111	GreeceTosk_Sic_STR_M10.csv		414	10%	35 000	46	GreeceTosk_Cal_STR_M10.csv
		444	20%	35 000	111	GreeceTosk_Sic_STR_M20.csv		184	20%	35 000	46	GreeceTosk_Cal_STR_M20.csv
		259	30%	35 000	111	GreeceTosk_Sic_STR_M30.csv		107	30%	35 000	46	GreeceTosk_Cal_STR_M30.csv
		111	50%	35 000	111	GreeceTosk_Sic_STR_M50.csv		46	50%	35 000	46	GreeceTosk_Cal_STR_M50.csv
vs. GHEG												
STR	I=1	1998	5%	35 000	111	Gheg_Sic_STR_M5.csv	I=1	828	5%	35 000	46	Gheg_Cal_STR_M5.csv
		999	10%	35 000	111	Gheg_Sic_STR_M10.csv		414	10%	35 000	46	Gheg_Cal_STR_M10.csv
		444	20%	35 000	111	Gheg_Sic_STR_M20.csv		184	20%	35 000	46	Gheg_Cal_STR_M20.csv
		259	30%	35 000	111	Gheg_Sic_STR_M30.csv		107	30%	35 000	46	Gheg_Cal_STR_M30.csv
		111	50%	35 000	111	Gheg_Sic_STR_M50.csv		46	50%	35 000	46	Gheg_Cal_STR_M50.csv

	Sicily						Calabria					
Marker	Incr	Migrants	% local	N X	N Y	Output file	Incr	Migrants	% local	N X	N Y	Output file
vs. TOSK												
SNP	I=1	2142	5%	35 000	119	Tosk_Sic_SNP_M5.csv	I=1	846	5%	35 000	47	Tosk_Cal_SNP_M5.csv
		1071	10%	35 000	119	Tosk_Sic_SNP_M10.csv		423	10%	35 000	47	Tosk_Cal_SNP_M10.csv
		476	20%	35 000	119	Tosk_Sic_SNP_M20.csv		188	20%	35 000	47	Tosk_Cal_SNP_M20.csv
		278	30%	35 000	119	Tosk_Sic_SNP_M30.csv		110	30%	35 000	47	Tosk_Cal_SNP_M30.csv
		119	50%	35 000	119	Tosk_Sic_SNP_M50.csv		47	50%	35 000	47	Tosk_Cal_SNP_M50.csv
vs. GREECE/TOSK												
SNP	I=1	2142	5%	35 000	119	GreeceTosk_Sic_SNP_M5.csv	I=1	846	5%	35 000	47	GreeceTosk_Cal_SNP_M5.csv
		1071	10%	35 000	119	GreeceTosk_Sic_SNP_M10.csv		423	10%	35 000	47	GreeceTosk_Cal_SNP_M10.csv
		476	20%	35 000	119	GreeceTosk_Sic_SNP_M20.csv		188	20%	35 000	47	GreeceTosk_Cal_SNP_M20.csv
		278	30%	35 000	119	GreeceTosk_Sic_SNP_M30.csv		110	30%	35 000	47	GreeceTosk_Cal_SNP_M30.csv
		119	50%	35 000	119	GreeceTosk_Sic_SNP_M50.csv		47	50%	35 000	47	GreeceTosk_Cal_SNP_M50.csv
vs. GHEG												
SNP	I=1	2142	5%	35 000	119	Gheg_Sic_SNP_M5.csv	I=1	846	5%	35 000	47	Gheg_Cal_SNP_M5.csv
		1071	10%	35 000	119	Gheg_Sic_SNP_M10.csv		423	10%	35 000	47	Gheg_Cal_SNP_M10.csv
		476	20%	35 000	119	Gheg_Sic_SNP_M20.csv		188	20%	35 000	47	Gheg_Cal_SNP_M20.csv
		278	30%	35 000	119	Gheg_Sic_SNP_M30.csv		110	30%	35 000	47	Gheg_Cal_SNP_M30.csv
		119	50%	35 000	119	Gheg_Sic_SNP_M50.csv		47	50%	35 000	47	Gheg_Cal_SNP_M50.csv

Table S2. Y-Chromosome STRs haplotypes and SNPs analysis results for the newly-typed samples of Arbereshe (N=150) and Albanian (N=223) populations. This table will be provided in the online version of the manuscript once published.

Table S3. Y-chromosome haplogroup frequencies in the combined dataset. For each Y-chromosome lineage the absolute number of individuals and the percentage frequency (between brackets) are reported.

Y-CHR	ARB	ARB_CAL	POL_AREA	POL_SW	VAL_CRA	ARB_SIC	PIA_ALB	CON_ENT	ALB	TOSK	GHEG	SSI[a]	GRE[b]
	150	106	24	36	46	44	18	26	223	104	119	263	209
C-M216	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.48)
E-M96	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.76)	0 (0)
E-M35	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (1.14)	0 (0)
E-M78	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.45)	0 (0)	1 (0.84)	0 (0)	1 (0.48)
E-V12	3 (2)	3 (2.83)	0 (0)	0 (0)	3 (6.52)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	8 (3.04)	1 (0.48)
E-V13	20 (13.33)	18 (16.98)	4 (16.67)	7 (19.44)	7 (15.22)	2 (4.55)	2 (11.11)	0 (0)	75 (33.63)	30 (28.85)	45 (37.82)	23 (8.75)	42 (20.1)
E-V22	1 (0.67)	1 (0.94)	0 (0)	1 (2.78)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.45)	1 (0.96)	0 (0)	5 (1.9)	2 (0.96)
E-M81	1 (0.67)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.27)	0 (0)	1 (3.85)	0 (0)	0 (0)	0 (0)	5 (1.9)	0 (0)
E-M123	16 (10.67)	8 (7.55)	0 (0)	8 (22.22)	0 (0)	8 (18.18)	3 (16.67)	5 (19.23)	2 (0.9)	1 (0.96)	1 (0.84)	7 (2.66)	3 (1.44)
F-M89	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (1.79)	3 (2.88)	1 (0.84)	0 (0)	3 (1.44)
G-M201	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.38)	2 (0.96)
G-P15	8 (5.33)	7 (6.6)	2 (8.33)	2 (5.56)	3 (6.52)	1 (2.27)	1 (5.56)	0 (0)	6 (2.69)	4 (3.85)	2 (1.68)	31 (11.79)	10 (4.78)
I-M253	8 (5.33)	6 (5.66)	4 (16.67)	0 (0)	2 (4.35)	2 (4.55)	1 (5.56)	1 (3.85)	8 (3.59)	4 (3.85)	4 (3.36)	3 (1.14)	3 (1.44)
I-P215	15 (10)	6 (5.66)	0 (0)	3 (8.33)	3 (6.52)	9 (20.45)	2 (11.11)	7 (26.92)	16 (7.17)	12 (11.54)	4 (3.36)	3 (1.14)	21 (10.05)
I-M26	3 (2)	2 (1.89)	0 (0)	0 (0)	2 (4.35)	1 (2.27)	1 (5.56)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.76)	1 (0.48)
I-M223	15 (10)	15 (14.15)	6 (25)	1 (2.78)	8 (17.39)	0 (0)	0 (0)	0 (0)	5 (2.24)	5 (4.81)	0 (0)	6 (2.28)	7 (3.35)
J-M267	3 (2)	3 (2.83)	0 (0)	2 (5.56)	1 (2.17)	0 (0)	0 (0)	0 (0)	5 (2.24)	2 (1.92)	3 (2.52)	14 (5.32)	9 (4.31)
J-M172	3 (2)	3 (2.83)	1 (4.17)	1 (2.78)	1 (2.17)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.38)	1 (0.48)
J-M410	4 (2.67)	2 (1.89)	0 (0)	0 (0)	2 (4.35)	2 (4.55)	2 (11.11)	0 (0)	5 (2.24)	4 (3.85)	1 (0.84)	28 (10.65)	12 (5.74)
J-M67	5 (3.33)	1 (0.94)	0 (0)	1 (2.78)	0 (0)	4 (9.09)	4 (22.22)	0 (0)	1 (0.45)	0 (0)	1 (0.84)	8 (3.04)	11 (5.26)
J-M92	1 (0.67)	1 (0.94)	1 (4.17)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	7 (2.66)	5 (2.39)
J-M12	4 (2.67)	4 (3.77)	2 (8.33)	2 (5.56)	0 (0)	0 (0)	0 (0)	0 (0)	42 (18.83)	12 (11.54)	30 (25.21)	6 (2.28)	15 (7.18)
K-M9	2 (1.33)	1 (0.94)	0 (0)	0 (0)	1 (2.17)	1 (2.27)	1 (5.56)	0 (0)	2 (0.9)	2 (1.92)	0 (0)	6 (2.28)	7 (3.35)
P-M45	2 (1.33)	0 (0)	0 (0)	0 (0)	0 (0)	2 (4.55)	0 (0)	2 (7.69)	0 (0)	0 (0)	0 (0)	1 (0.38)	0 (0)
R-M207	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.48)
R-M173	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.48)
R-SRY10831.2	15 (10)	10 (9.43)	3 (12.5)	2 (5.56)	5 (10.87)	5 (11.36)	0 (0)	5 (19.23)	9 (4.04)	6 (5.77)	3 (2.52)	13 (4.94)	29 (13.88)
R-M343	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (1.14)	2 (0.96)
R-M269	12 (8)	9 (8.49)	1 (4.17)	2 (5.56)	6 (13.04)	3 (6.82)	0 (0)	3 (11.54)	33 (14.8)	14 (13.46)	19 (15.97)	19 (7.22)	10 (4.78)
R-P311	2 (1.33)	0 (0)	0 (0)	0 (0)	0 (0)	2 (4.55)	1 (5.56)	1 (3.85)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
R-U106	1 (0.67)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2.27)	0 (0)	1 (3.85)	0 (0)	0 (0)	0 (0)	11 (4.18)	2 (0.96)
R-P312	1 (0.67)	1 (0.94)	0 (0)	1 (2.78)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	13 (4.94)	1 (0.48)
R-SRY2627	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.76)	1 (0.48)
R-U152	1 (0.67)	1 (0.94)	0 (0)	1 (2.78)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.9)	1 (0.96)	1 (0.84)	11 (4.18)	1 (0.48)
R-L2	4 (2.67)	4 (3.77)	0 (0)	2 (5.56)	2 (4.35)	0 (0)	0 (0)	0 (0)	5 (2.24)	2 (1.92)	3 (2.52)	16 (6.08)	4 (1.91)
R-L21	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.45)	1 (0.96)	0 (0)	5 (1.9)	0 (0)

[a] Sarno et al. 2014

[b] Anagnostou P (personal comunication)

Table S4. Y-chromosome indices for the investigated populations. Standard diversity parameters were calculated for both haplogroup frequencies (SNPs) and haplotypes data (STRs).

Y-Chromosome			STRs			SNPs
Population	N	K	STR div (*h*) ± sd	MNPD (*π*) ± sd	Nucleotyde div (π_N) ± sd	Haplogroup div (*h*) ± sd
ARB_CAL	**106**	**87**	**0.9950 +/- 0.0020**	**8.5608 +/- 3.9882**	**0.5707 +/- 0.2945**	**0.9208 +/- 0.0108**
POL_AREA	24	21	0.9891 +/- 0.0152	8.8913 +/- 4.2467	0.5557 +/- 0.2959	0.8841 +/- 0.0353
POL_SW	36	31	0.9873 +/- 0.0117	7.8762 +/- 3.7509	0.5626 +/- 0.2978	0.9079 +/- 0.0291
VAL_CRA	46	39	0.9932 +/- 0.0059	8.5672 +/- 4.0326	0.5711 +/- 0.2985	0.9159 +/- 0.0178
ARB_SIC	**44**	**37**	**0.9915 +/- 0.0070**	**8.7537 +/- 4.1173**	**0.5471 0.2857**	**0.9070 +/- 0.0229**
CON_ENT	26	20	0.9785 +/- 0.0166	8.2862 +/- 3.9672	0.5179 +/- 0.2762	0.8615 +/- 0.0374
PIA_ALB	18	18	1.0000 +/- 0.0185	7.6471 +/- 3.7409	0.5882 +/- 0.3218	0.9216 +/- 0.0391
SSI	**263**	**252**	**0.9996 0.0004**	**9.8455 +/- 4.5206**	**0.6153 +/- 0.3126**	**0.9439 +/- 0.0045**
CS	47	47	1.0000 +/- 0.0044	8.7537 +/- 4.5010	0.6029 +/- 0.3123	0.9408 +/- 0.0121
AG	45	44	0.9990 +/- 0.0050	10.1434 +/- 4.7210	0.6340 +/- 0.3276	0.9525 +/- 0.0140
TP	34	32	0.9933 +/- 0.0094	9.9529 +/- 4.6634	0.6221 +/- 0.3240	0.9234 +/- 0.0194
EN	40	40	0.9987 +/- 0.0060	9.7269 +/- 4.5504	0.6079 +/- 0.3159	0.9487 +/- 0.0182
RG.SR	45	43	0.9980 +/- 0.0052	9.3919 +/- 4.3936	0.6261 +/- 0.3252	0.9465 +/- 0.0155
CT	52	49	0.9977 +/- 0.0043	9.2232 +/- 4.3092	0.6149 +/- 0.3188	0.9367 +/- 0.0162
ALB	**223**	**177**	**0.9960 +/- 0.0012**	**8.9884 +/- 4.1554**	**0.5618 +/- 0.2874**	**0.8218 +/- 0.0168**
TOSK	104	97	0.9983 +/- 0.0017	9.2713 +/- 4.2951	0.5795 +/- 0.2973	0.8680 +/- 0.0204
GHEG	119	89	0.9899 +/- 0.0036	8.6703 +/- 4.0319	0.5419 +/- 0.2790	0.7695 +/- 0.0251
GRE	**209**	**119**	**0.9994 +/- 0.0006**	**9.7182 +/- 4.4697**	**0.6074 +/- 0.3091**	**0.9126 +/- 0.0095**
EUB	96	94	0.9996 +/- 0.0016	10.0086 +/- 4.6159	0.6255 +/- 0.3196	0.9355 +/- 0.0108
KOR	113	106	0.9986 +/- 0.0015	9.3560 +/- 4.3289	0.5848 +/- 0.2996	0.8892 +/- 0.0157

Table S5. Analyses of the molecular variance (AMOVA). Apportionment of the variance are in percentage (%) and based on both haplogroup frequencies (SNPs) and haplotype data (STRs).

Y-chromosome Grouping	N° of Groups	N° of Pop	Proportion of variation (%)					
			Among Groups		Among population Within Group		Within Population	
			Y-SNPs	Y-STRs	Y-SNPs	Y-STRs	Y-SNPs	Y-STRs
All populations	**1**	**15**			**3.70*****	**2.68*****	**96.30*****	**97.32*****
Arbereshe	**1**	**5**			**3.96*****	**2.95****	**96.04*****	**97.05****
Arbereshe Calabria	1	3			2.42*	0.56	97.58**	99.44
Arbereshe Sicily	1	2			4.60*	5.17*	95.40*	94.83*
Sicily and South Italy	**1**	**6**			**0.26**	**0.54**	**99.74**	**99.46**
Balkan	**1**	**4**			**2.58*****	**2.82*****	**97.42*****	**97.18*****
Albania	1	2			1.53*	1.21*	98.47*	98.79*
Greece	1	2			0.46	0.86	99.54	99.14
Language [a]	**4**	**15**	**2.51****	**2.00*****	**1.67*****	**1.06****	**95.82*****	**96.94*****
Geography [b]	**4**	**15**	**2.51*****	**1.68****	**1.64*****	**1.30*****	**95.85*****	**97.02*****

[a] Linguistic subdivision as follow: Albanian-Tosk, Albanian-Ghegh, Southern-Italian, Greek

[b] Geographic subdivision as follow: Sicily, Calabria, Albania, Greece

*** P-value < 0,001

** P-value < 0,01

* P-value < 0,05

Article 4

Boattini A, Sarno S, Pedrini P, Medoro C, Carta M, Tucci S, Ferri G, Alù M, Luiselli D, Pettener D. *Traces of Medieval migrations in a socially-stratified population from Northern Italy. Evidence from uniparental markers and deep-rooted pedigrees* (in preparation).

Traces of Medieval migrations in a socially-stratified population from Northern Italy. Evidence from uniparental markers and deep-rooted pedigrees.

Alessio Boattini (1)
Stefania Sarno (1)
Paola Pedrini (1)
Chiara Medoro (1)
Marilisa Carta (1)
Serena Tucci (1, 2)
Gianmarco Ferri (3)
Milena Alù (3)
Donata Luiselli (1)
Davide Pettener (1)

(1) Laboratorio di Antropologia Molecolare, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, 40126 Bologna, Italy
(2) Dipartimento di Biologia ed Evoluzione, Università di Ferrara, 44121 Ferrara
(3) Dipartimento di Medicina Diagnostica, Clinica e di Sanità Pubblica, Università degli Studi di Modena e Reggio Emilia, 41124 Modena, Italy

Corresponding Author
Alessio Boattini, Laboratorio di Antropologia Molecolare - Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Via Selmi 3, 40126, Bologna, Italy. Tel. +390512094091. E-mail: alessio.boattini2@unibo.it

Running Title
Traces of Medieval migrations in Northern Italy

Word count for main text: 6,111

ABSTRACT

Social and cultural factors had a critical role in determining the genetic structure of Europe. Therefore, socially-stratified populations may help to focus on specific episodes of the European demographic history. In this study we use uniparental markers to analyse the genetic structure of the “Partecipanza” in San Giovanni in Persiceto (Northern Italy), a peculiar institution whose origins date back to the Middle Ages and whose members form the patrilineal descent of a well-known group of founder families. If, from a maternal point of view (mtDNA), the “Partecipanza” is genetically homogeneous with the rest of the population, on the contrary we observed a significant differentiation for Y-chromosomes. In addition, by comparing 17 Y-STRs with paternal pedigrees, we estimated a Y-STR mutation rate equal to $3.90 * 10^{-3}$ and an average generation duration time of 33.38 years. When we used these values for tentative dating experiments, we estimated 1300-600 years ago for the origins of the “Partecipanza”. These results, together with a peculiar Y-chromsomal composition and historical evidences, suggest that Germanic populations (Lombards in particular) settled in the area during the Migration Period and may have had an important role in the foundation of this community.

Keywords: Y-STR mutation rates, Y-chromosome, mtDNA, recent demographic episodes, social-economical factors, Migration Period

INTRODUCTION

Recent demographic episodes and population events are rarely studied by geneticists and molecular anthropologists. In fact, if human genetic diversity can be described as a palimpsest in which different layers – reflecting different episodes – accumulate through time, dissecting a single event is a difficult task. Furthermore, recent events are rarely associated with large population substitutions; on the contrary they usually bring only limited modifications to a pre-existing genetic background, hence their genetic 'signals' are low and difficult to detect. These are some of the reasons why many studies focus on some well-known events and try to link them to the observed genetic patterns (Jobling, 2012; Larmuseau et al., 2013). Ancient DNA research is only beginning to discover what stands behind elusive labels such as 'Neolithic', 'Linear Pottery Culture', 'Etruscans', etc., but the actual genetic variability of ancient populations still remains largely unknown (Lacan et al., 2012).

It has been recently shown that an adequate sampling design may help to address some of these problems. For instance, "marginal" populations, such as ethno-linguistic minorities or geographic isolates, not only give an important contribution to the biodiversity of a region, but are also more likely to display genetic features that instead were masked by intervening demographic episodes in "open" populations (Boattini et al., 2011a,b; Capocasa et al., 2014). As a second point, surnames and/or pedigrees can be used for selecting samples within populations. In Y-chromosome studies, sampling males bearing surnames that are unequivocally associated with a certain place (either for documentary evidence or by means of inference) provides proxies for older populations (King and Jobling, 2009; Boattini et al., 2012). The same considerations hold for sample selection based on genealogies/pedigrees. Compared to surnames, genealogies yield a higher degree of temporal and geographic precision. On the other hand, extensive genealogical data are rare and their collection is very time-consuming. A small but increasing number of studies have however shown that surname- and pedigree-based samplings can reveal otherwise hidden populations genetic structures (Larmuseau et al., 2013 and references therein).

As recently observed (Boattini et al., 2013), the genetic history of Italy is particularly complicated in comparison with other European countries. In this study, we focussed on a particularly interesting population, the *Partecipanza* of San Giovanni in Persiceto, so that some of its less known aspects can be better understood. From a juridical point of view, the *Partecipanze* are an absolutely idiosyncratic way of sharing and devolving collective lands. These institutions originated in the Middle Ages and are still present in some areas of Northern Italy (in the Padana Plain). The privilege to *participate* (i.e. to share the leased assets) is inherited following a gene-like pattern, through exclusive admission granted only to the descent of a group of ‘founder’ families, usually following the paternal line (Zanarelli, 1992). In this way, the members of *Partecipanza* conserved through the centuries their social and economic identity, potentially together with some of their genetic features. Furthermore, thanks to the wide archival documentation conserved by this institution, it is possible to compare DNA samples with paternal pedigrees for the last 4-5 centuries. All these features make the *Partecipanza* an exceptional observatory on the recent genetic history of Italy.

In this study we explore the paternal (Y-chromosome) and maternal (mtDNA) genetic variability of the *Partecipanza* of S. Giovanni in Persiceto (PAR). Results are compared to those of a set of ‘control’ individuals sampled in the same place, but not sharing the affiliation to the *Partecipanza* (SGP), and interpreted considering a wide set of reference populations from Italy and Europe. Our main aims are: a) to check the genetic effects – if any – of the socio-economic separation between PAR and SGP; b) to reconstruct the time and the genetic origins of the *Partecipanza*, as well as their implications for the genetic history of Italy; and c) to estimate Y-STR mutation rates and average generation duration time by comparing paternal pedigrees and Y-chromosomal haplotypes.

MATERIALS AND METHODS

The population

The *Partecipanze* are located in the flat portions of the provinces of Bologna and Ferrara (Padana plain, Northern Italy). Their origins are related to medieval events of land reclamation, but the exact date of their foundation is unknown. Evolving from ancient collective *emphyteutic* grants, the *Partecipanze* soon became reserved to a restricted group of founder families. Currently, six *Partecipanze* are still present and active in the study area: San Giovanni in Persiceto, Nonantola, Cento, Pieve di Cento, S. Agata Bolognese and Villa Fontana. Among these, *Partecipanza* of San Giovanni in Persiceto is probably the most important. Its shared assets exceed ~2400 ha and its population size is of ~5000 individuals. This *Partecipanza* (to which we will refer exclusively from now on) is historically documented from at least 1170 and 1215 AD, when *emphyteutic* grants to the community of S. Giovanni in Persiceto (respectively from the abbot of Nonantola and the bishop of Bologna) were stipulated. The customary laws that regulate the affiliation to the *Partecipanza* were definitively coded around the year 1500 AD (*capitula*). According to these agreements, the ability to *Partecipate* (i.e. to be a member of the *Partecipanza*) was conditioned by two main rules: being a patrilineal legitimate descent from a *Partecipante* family and to maintain the place of residence within the legal boundaries of S. Giovanni in Persiceto. In addition, the procedure of *cavazione* was established. *Cavazione* is a periodic re-shuffling of shared lands among heads of the *partecipating* households. This procedure is usually repeated each nine years, leaving us census-like descriptions of the population that are conserved in the archive of the Institution. Thanks to their favoured access to lands, members of *Partecipanza* for centuries constituted the é*lite* of San Giovanni in Persiceto. Such dominance was highlighted by the fact that the leading council of *Partecipanza* was at the same time the head of the Commune. Only in 1833 the two institutions (*Partecipanza* and Commune) were definitively separated. At present, the individuals/families affiliated to the *Partecipanza* are characterised by 38 different surnames (Zanarelli, 1992).

DNA samples

Buccal swabs were collected from 149 male individuals sampled in S. Giovanni in Persiceto and its surroundings. Among them, 88 belong to the *Partecipanza* (PAR sample), while the remaining 61 form an 'open' sample (SGP), sharing the same environmental and cultural features, aside from the status of *Partecipante*. All these samples were collected according to the standard 'grandparents' criterion (i.e. at least three generations of ancestry in the SGP area) and excluding related individuals (up to second cousins). To best reconstruct the paternal genetic variability of the *Partecipanza*, we sampled at least one individual for each of the 38 Partecipanza surnames. Data from 14 individuals have been previously published in Boattini et al. (2013). The collection of biological samples was performed during various sessions from 2008 to 2012. For all subjects, a written informed consent was obtained and the Ethics Committees at the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna (Italy) approved all procedures. The confidentiality of personal information for each participant to the study was assured.

Comparison populations

Reference data include 32 populations from: North-Western Italy, South-Eastern Italy and Sicily (21, Boattini et al., 2013; Sarno et al., 2014), French Basques (3, Martinez-Cruz et al., 2012), Germany (3, Rebala et al., 2013), Poland and Slovakia (4, Rębała et al., 2013), Balkans (1, Regueiro et al., 2012). A full list is available in Table S1 and their geographic position is represented in Figure S1. In addition, 29 haplotypes for Y-chromosome hgs I1-M253 (25) and I1-P109 (4) from Brabant (Larmuseau et al., 2011; Larmuseau, personal communication) were considered.

Y-chromosome genotyping

PCR amplification of 17 Y-STR loci (DYS19,DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b, DYS437, DYS438, DYS439, DYS448, DYS456,DYS458, DYS635, and GATA H4) was carried out by using the AmpFlSTR Yfiler PCR Amplification Kit (Applied

Biosystems, Foster City, CA) following the manufacturer's recommendations in a final volume of 5μl. The PCR reaction consisted of denaturation at 95°C for 11 min, followed by 30 denaturation cycles at 94°C for 1 min, annealing at 61°C for 1 min, extension at 72°C for 1 min, and a final extension at 60°C for 80 min. Products were sized on an ABI Prism 310 Genetic Analyzer using the GeneScan 3.7 software (Applied Biosystems). As the Yfiler kit amplifies DYS385a/b simultaneously, avoiding the determination of each of the two alleles (a or b), these loci were excluded from downstream analyses except where explicitly specified. DYS389b was obtained by subtracting DYS389I from DYS389II. Basal haplogroups were assigned by typing the 7 SNPs (R-M173, J-M172 , I-M170, E-M35, K-M9, P-M45, F-M89) implemented in the MY1 Multiplex PCR by Onofri et al. (2006). Subsequently, we explored Y-chromosome variability by using 28 SNPs. 26 of them (E-M78, E-V12, E-V13, E-V22, G-P15, G-P16, G-M286, G-U8, G-U13, I-M253, I-M227, I-L22, I-P215, I-M26, J-M410, J-L27, J-M67, J-M92, J-M12, R-M17, R-M343, R-M18, R-M269, R-S21/U106, R-SRY2627/M167, R-S28/U152) were included in five haplogroup-specific multiplexes (Ferri & Alù, 2012). The SNP genotyping was carried out by means of PCR Multiplex amplification, followed by Minisequencing reaction based on dideoxy Single Base Extension (SBE), which was performed with the SNaPshot multiplex kit (Applied Biosystem). SBE products were analyzed with capillary electrophoresis on an ABI Prism 310 Genetic Analyser. Two additional SNPs (E-M81, E-M123) were finally tested with RFLP analysis, by using HpyCH4IV and DdeI enzymes respectively as in Sarno et al. (2014). Haplogroups nomenclature is in accordance with the Y-Chromosome Consortium (Karafet et al. 2008). Individual hg information and Y-STR data are provided in Table S2.

Mitochondrial DNA genotyping

Variation at the mtDNA HVS-I and HVS-II regions was investigated by sequencing a total of 750 base pairs (bp) encompassing nucleotide positions from 15975 to 155. Polymerase chain reaction (PCR) of the HVSI/II regions was carried out in a T-Gradient Thermocycler (Whatman Biometra, Gottingen, Germany) with the following amplification profile: initial denaturation 95 °C for 5 min, 35 cycles of 95 °C for 30 sec, 58 °C for 30 sec, 72 °C for 5 min and final extension at 72 °C for 15 min. PCR products were purified by ExoSap-IT1 (USB Corporation, Cleveland, OH) and sequenced on an ABI Prism 3730 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) by using a Big-Dye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions. To reduce ambiguities in sequence determination the forward and reverse primers were used to sequence both strands of HVS-I and HVS-II regions. The CHROMAS 2.33 software was used to read the obtained electropherograms. Sequences were finally aligned to both the Revised Cambridge reference sequence (rCRS: Andrews et al., 1999) and the new Reconstructed Sapiens Reference Sequence (RSRS: Behar et al. 2012) by using the DNA Alignment Software 1.3.1.1 (http://www.fluxusengineering.com/align.htm). Sequence data are provided in Table S3.

Archival Data and Pedigree Reconstruction

The Historic Archive of *Consorzio dei Partecipanti* kept a record of the enrolments (*Registri delle Iscrizioni*) since the early 17th century. The *right to participate* to the sharing of leased assets (i.e. to be a *Partecipante*) was (and still is) based on these registers. Each registration includes the name of the head of the household, his age and/or his year of birth, the composition of his family, his parish/locality of residence and all the additional information needed to qualify him as a *Partecipante* (for example, the code of the precedent enrolment). These registers were compiled in the occasion of *Cavazione*, usually each nine years.

We examined the whole series of the available *Registri delle Iscrizioni*, at present composed by registers from the years: 1606, 1612, 1631, 1634, 1640, 1643, 1650, 1653, 1659, 1662, 1668, 1671, 1677, 1680, 1686, 1689, 1695, 1698, 1707, 1713, 1716, 1722, 1725, 1731, 1740, 1743, 1752, 1761, 1764, 1770, 1779, 1788, 1797, 1806, 1815, 1824, 1833, 1842, 1851, 1860, 1869, 1878, 1887, 1896, 1905, 1914, 1923, 1932, 1941, 1950, 1959, 1968, 1977, 1986, 1995, 2004.

Information from these registers was used to draw paternal pedigrees for the individuals included in our *Partecipanza* sample. When two or more individuals were found to share a recent paternal ancestor, they were grouped into a single pedigree.

Generation duration time and Y-STR mutation rates

Average generation time was obtained starting from the ages and/or years of birth of the individuals included in the reconstructed pedigrees. For each pedigree, we calculated the number of years and the number of generations encompassed between its root and leaves. The most remote ancestors were excluded, since their year of birth is uncertain. Finally, we divided the total number of years for the total number of generations. The 95% confidence intervals were calculated by randomly sampling branches of the pedigrees with a bootstrap procedure (1000 replications).

Y-STR mutation rates were estimated comparing haplotype information within paternal pedigrees that include at least two leaves. Any individual whose hg is not compatible with the reconstructed pedigrees was excluded from calculations. ‘Outlier’ haplotypes – that are haplotypes showing an ‘outlier’ number of mutations from the common ancestor – were excluded. ‘Outlier’ haplotypes were identified by means of a Grubbs test for one outlier in a data sample (function *grubbs.test*, library *outliers*, software R; Komsta, 2011). For each pedigree, we computed the number of mutation events according to the maximum parsimony method and the number of generations separating the considered individuals/haplotypes. The average STR mutation rate was calculated by dividing the total number of mutations for the total number of generations and for the number of considered STRs. Bootstrapped confidence intervals were calculated by randomly re-sampling mutations per STR (1000 replications). To allow maximal resolution, DYS385 was included in the calculations as two separate loci; however, since mutation events were computed within pedigrees, it is highly improbable that apparently identical DYS385 configurations were actually two different haplotypes.

Within- and Between-population analyses

Standard within-population diversity parameters (Gene Diversity, Mean Number of Pairwise Differences, Nucleotide Diversity) for Y-chromosome (haplogroups, STR haplotypes) and mtDNA (HVR-I and II sequences), as well as Mismatch Distribution patterns (mtDNA sequences) were calculated with Arlequin 3.5.1.2 (Excoffier et al. 2007). The same software was used to compute pairwise Fst values between PAR and SGP based on Y-chromosome hgs and mtDNA sequences; p-values (following the null hypothesis of no differentiation) were simulated by means of a permutation procedure (1000 replications). Fisher exact tests were applied to determine if differences in Y-haplogroup frequencies between PAR and SGP were statistically significant.

In order to check the position of our populations within the Italian and European Y-chromosomal landscape, we performed a Non-Metric Multi-Dimensional Scaling (NM-MDS). Since different studies used different levels of hg resolution, the analysis was based on 15 Y-STR haplotypes, that were available for all the considered reference populations. Calculations were performed using Reynolds distance (Reynolds et al., 1983; for the use of this distance with STRs see Laval et al., 2002) and the function *isoMDS* implemented in the R software *MASS* package (R Core Team, 2013). The first and the second Dimensions were represented in a scatterplot, along with the corresponding stress value.

Within Y-hgs analysis

In order to explore genetic variability within those hgs that mostly differentiate PAR from SGP, we applied a Discriminant Analysis of Principal Components (DAPC; Jombart et al., 2010) to Y-STR haplotypes as in Boattini et al. (2013). This analysis is aimed to 1) identify well-resolved groups of haplotypes within hgs; 2) highlight possible affinities/similarities with reference haplotypes from Italy and Europe; 3) constitute a starting point for time estimates. Briefly, the procedure involves two steps. First, haplotypes are grouped using k-means, a clustering algorithm which finds a given number of clusters maximizing the variation between groups. Second, DAPC is used to describe the diversity (and the degree of separation) among such groups of observations, by maximizing the between-group variance and minimizing the within-group variance. The first 2-3 Discriminant

Functions were represented with scatterplots. All analyses were performed within the R software add-on package *adegenet* (Jombart, 2008).

Y-chromosome time estimates

Time estimates will focus on well differentiated groups of haplotypes, as described by DAPC. In order to avoid sampling biases, only clusters comprising at least ten individuals are considered. Calculations use two different sets of mutation rates: 1) our pedigree-based average mutation rate; 2) Ballantyne et al. (2010) STR-specific mutation rates. Ballantyne et al. (2010) rates, being averagely similar to our rate (see Results) but available for single STRs, were introduced for further information. Dates are computed with standard deviation (SD) estimator (Sengupta et al. 2006) as in Boattini et al. (2013). Since population events involving PAR are relatively recent, the biasing effect of STRs saturation through time is negligible, hence all STRs (minus DYS385a/b) are used for calculations. This method estimates the amount of time needed to evolve the observed STRs variation within the given clusters of haplotypes.

RESULTS

Paternal pedigrees, STR mutation rates and generation duration time.

We reconstructed paternal lineages (up to the early 17th century) for 74 (out of 88) individuals of the PAR sample. Since some individuals share a common ancestor, these paternal lineages were grouped in 41 independent paternal pedigrees. Fifteen of these pedigrees (encompassing in total 48 individuals) include two or more sampled individuals, up to a maximum of 8 (Table 1, Figure S2). All individuals included in a given pedigree share the same surname, but a given surname may be present in more than one pedigree (Table S2). This could happen because the most recent common ancestor was not reached by documented pedigrees or because the surname is polyphyletic. Y-STR mutation rates were estimated based on these 15 paternal pedigrees.

As a preliminary step, we excluded from calculations: a) four individuals displaying a different hg compared to the other members of their respective pedigrees; and b) one individual displaying five mutations from the common ancestor of his pedigree. This figure is significantly higher than all the other cases considered here (Table S2) and clearly constitutes a case of outlier (Gibbs outlier test: G = 4.1066, U = 0.5987, p-value = $8.63*10^{-5}$). Finally, we observed a 3-step mutation case at locus DYS439 within pedigree P14 (Table 1, Table S2). Since it is highly improbable that three independent mutation events involved the very same locus (in a relatively short amount of time), we consider it as a single multi-step mutational event.

On the whole, by considering 43 individuals across 15 pedigrees and 17 Y-STRs, we observed 24 mutations within 362 generations (Table 1). These values give an average mutation rate equal to 3.90 * 10^{-3} (95% CI: 2.44 * 10^{-3}, 5.68 * 10^{-3}).

Average generation duration time was calculated using all the 31 paternal pedigrees (Table 1). We observed 604 generations encompassing 20,160 years. These values give an average 33.38 years generation duration time (95% CI: 32.76, 34.00).

Diversity indices, mismatch distribution, Fst.

In general, standard within-population diversity indices show that PAR and SGP share similar levels of internal genetic variation (Table S4). The only exception is Y-SNP-based Gene Diversity, which is significantly higher in PAR. Analogously, mismatch distributions (mtDNA) are almost identical in both samples, suggesting that these populations have experienced the same recent demographic changes (Figure S3). Such changes can be interpreted as a demographic increase of the population, which is perfectly in line with what is known for the investigated area for the last two centuries.

As for Y-chromosomal hgs (Table 2), R-U152 is the most represented lineage in both populations, albeit with a significantly (Fisher Test: p = 0.0363) much higher frequency in SGP (44.26% vs. 27.27%). From the other side, PAR differentiates itself from SGP primarily due to the high frequency of hg I1-L22 (15.91%), which in turn is completely absent in SGP (Fisher Test: p-value = 0.0004). Hg J2-M67*, also, is significantly (Fisher Test: p-value = 0.0443) more frequent in PAR than in SGP (13.64% vs. 3.28%). Accordingly, we observe a slight but significant differentiation between PAR and SGP (Fst = 0.030; p-value < 0.01). In addition, we selected a PAR sub-sample based on surnames and on the reconstructed genealogies. Such sub-sample includes: a) all individuals bearing a surname that is represented a single time in our collection (even if a reconstructed paternal lineage was not available); b) one random individual for each of the 15 reconstructed pedigrees that include more than one sample. That way we selected 48 individuals simulating a putative *Partecipanza* sample of the year 1600 (PAR1600).

When comparing PAR1600 with PAR, we do not observe any significant differentiation (Fst = -0.012, p-value > 0.05). This suggests that *Partecipanza* did not undergo significant changes in its Y-chromosomal composition during the last four centuries. In other words, there is no evidence of drift effects. Instead, PAR1600 is significantly different from SGP (Fst = 0.032, p-value < 0.01).

Contrarily to Y-chromosome, mtDNA variation shows that PAR and SGP are not significantly differentiated from each other (Fst = -0.002, p-value > 0.05). Since PAR appears to be different from SGP only from the paternal point of view, all the following analyses are meant to explore such Y-chromosomal differentiation.

MDS

In order to check the position of PAR and SGP within the European Y-chromosomal genetic landscape, we performed a MDS analysis based on Y-STR haplotypes (Figure 1). Results show clearly that both populations, despite being significantly different from each other, fall within the variability spectrum of the Italian populations.

DAPC & dating

Y-chromosomal hg I1-L22 is the distinctive mark of PAR, being highly frequent here (15.91%) while completely absent in SGP. A DAPC-based exploration of I1-L22 haplotypes from PAR and from comparison European populations revealed three well-differentiated clusters (Figure 2, Table S5). Interestingly, all PAR I1-L22 haplotypes fall within the same cluster (cluster 2) together with few haplotypes from South-Eastern Italy (1), Germany (1), Poland and Slovakia (4). Other clusters (1 and 3) are composed mainly by hts from Germany and Poland, but cluster 1 includes most of the I1-L22 hts from South-Eastern Italy and the Balkans, while cluster 3 is enriched in Brabant and North-Western Italy hts. These results suggest that I1-L22 individuals from PAR share a recent common ancestor that could have been living around the age when the *Partecipanza* was established. Accordingly, we estimated: a) the age of I1-L22 PAR haplotypes; 2) the age of whole cluster 2. The first estimate can be interpreted as a lower limit to the age of foundation of *Partecipanza*, while the second one as an upper limit. For the calculations we use our own estimates of generation duration time and average STR mutation rate, as well as locus-specific rates by Ballantyne et al. (2010). Our results (Table 3) place the lower limit at around 600-700 years ago, while the upper limit is at around 1300-1400 years ago.

As for hg J2-M67*, by comparing PAR and SGP hts with those from European reference populations with DAPC, we obtain three different clusters (Figure S4). One of them (cluster 3) is exclusively found in PAR, where it is associated to a single documented pedigree (P05, Table 1) which includes 8 individuals. Another one (cluster 2) is found both in PAR (4 hts) and SGP (1 ht), as well as in Germany and Southern Italy. Dates for the whole of cluster 2 (Table 3) are as older as 6000-9000 years ago. Cluster 1, despite being well represented both in Germany and in Italy, has only one haplotype in SGP.

A DAPC based comparison of PAR and SGP R-U152 hts with those from European reference populations revealed five different clusters (Table S5, Figure S4). Each of these clusters is represented both in PAR and SGP. By pooling together clusters 1 and 2 – which largely overlapped in our DAPC representations – we were able to date R-U152 hts both in PAR and SGP. Results were quite similar (Table 3), pointing at a period between 3500 and 4500 years ago in both cases.

Analogously, when dating cluster 3 by considering both hts from PAR and SGP (Table 3), we obtained a time estimate at around 3000-5000 years from present.

DISCUSSION

In this study we show that it is possible to shed light on important events of the recent genetic history of a region by carefully selecting the investigated population and the individuals to be sampled therein. Our PAR and SGP samples not only help to understand poorly known aspects of the genetic history of Italy and Europe, but they offer important glimpses on issues of more general interest such as the estimation of essential parameters like Y-STR mutation rates and population average generation time.

Relevant literature provides contrasting estimates of Y-STR mutation rates, calculated on the base of different rates. Available Y-STR mutation rates can be divided into three main groups.

1) *'Evolutionary' rates*. First proposed by Zhivotovsky et al. (2004), such rate is inferred from hg variation in populations whose short-time history is known (foundation events) and is equal to 6.9 * 10^{-4} per STR per generation. It is widely used for dating purposes.

2) *'Germline' rates*. Such rates are based on direct counts of mutational events in father-son pairs. Estimates in Ballantyne et al. (2010) are based on ~2,000 pairs and, for the 17 Y-STRs set here considered, it averages to 3.2 * 10^{-3} mutations per STR per generation.

3) *'Pedigree' rates*. These rates are based on counts of mutational events within pedigrees, as in the present study. King and Jobling (2009), relying on a total of 274 generations (within 14 pedigrees) and 17 STRs, calculated a rate equal to 1.5 * 10^{-3}. Compared to King and Jobling (2009), our calculations are based on a wider set of data (368 generations within 15 pedigrees). In addition, our pedigrees do not include tight relatives. Our estimate (3.90 * 10^{-3}; 95% CI: 2.44 * 10^{-3}, 5.68 * 10^{-3}), besides being higher than King and Jobling's (2009), is very similar to Ballantyne et al.'s (2010) germline rate (which in fact is included in the CI). Finally, all these rates are an order of magnitude higher than Zhivotovsky et al.'s (2004) 'evolutionary' rate. Following another line of evidence, Balanovsky et al. (2011), observed that time estimates based on 'germline' rates have a good fit with dates obtained from linguistics and archaeology, while dates based on 'evolutionary' rates tend to be older. These facts suggest that approximately Y-STR diversity accumulates in pedigrees at the germline rate, while 'evolutionary' rate is much slower.

As for average generation duration time, our estimate (33.38 years; 95% CI: 32.76, 34.00) is comparable to the 35 years value that King and Jobling (2009) suggested as ideal for Britain. As a consequence, a 25 years value, that is often used in a vast range of analyses – from time estimates to coalescent simulations – cannot be considered as particularly realistic for paternal lineages in European populations.

As a second point, we intended to test whether the idiosyncratic social-economic structure of the *Partecipanza* could influence its genetic variability. Cultural features are one of the most discussed determinants of genetic variation in human populations. The case of *Partecipanza* shows that being part of an economically advantaged élite (i.e. to be a *Partecipante*) may actually generate significant genetic differences with the rest of the population. It is important to consider that such peculiarities are limited to Y-chromosome variation and are instead not found in mtDNA. This result is in line with the specific transmission mechanism of the status of *Partecipante*, that is strictly and exclusively male-mediated. This observation, together with the lack of any significant variability reduction in PAR compared to SGP (both for Y-chromosome and mtDNA) suggests that *Partecipanza* did not experience significant drift effects. Similarly, when a putative late 16th century Partecipanza sub-sample (PAR1600) and the complete sample (PAR) have been compared, we did not observe any significant differentiation. In other words, Y-chromosomal composition of PAR cannot be explained exclusively by isolation. Furthermore, our MDS analysis clearly showed that both PAR and SGP lie within the space defined by Italian Y-chromosomal genetic variability.

These facts lead us to our last aim, that is to reconstruct the time and the genetic origin of *Partecipanza*, as well as its implications for the genetic history of Italy and Europe. Differences between PAR and SGP are mainly determined by three Y-hgs: I1-L22, J2-M67* and R-U152. Frequency of Y-hg I1-L22 reaches 15.91% in PAR, being completely absent in SGP and rare in the whole of Italy (0.79%; Boattini et al., 2013). Furthermore, its frequency does not exceed 5% in all the considered comparison populations. As in the case of its parental clade I1-M253, I1-L22 is most frequently found in Northern Europe, around the Baltic sea, where it probably originated (Soares et al., 2010; Underhill et al., 2007).

As shown by DAPC (Figure 2), I1-L22 hts from PAR form a tight and coherent cluster, suggesting that they may share a common ancestor that lived around the period in which *Partecipanza* was founded. Further proof of this is the fact that such hg is at present found in four different surnames and nine paternal pedigrees (Table S2). Another convergent line of reasoning is provided by our time estimates. Despite the fact that Y-STR-based estimates are a controversial issue, in this study we were able to overcome at least some of the most important criticisms brought to such methods. First, our results strongly suggest that PAR I1-L22 hts are not the result of historical (or geographical) stratification; on the contrary, they are tightly related and seem to mark a single historic event. As a second, important point, we were able to estimate reliable values for important parameters, such as STR mutation rates and generation duration time, directly from our data.

Despite relatively wide confidence intervals (Table 3), our estimates confirm what is currently known about the historic origins of *Partecipanza*, pointing at a period comprised between 1300 and 600 years ago. Interestingly, the upper bound coincides with the Migration Period, and in particular with the settling of Ostrogoth and Lombard in Italy (493 AD and 568 AD, respectively). The lower bound might instead correspond to the devastating "Black Death" epidemic of the 14th century (~ 1350 AD), a strong bottleneck that may have affected our population as well. It is worth noting that the geographic distribution of hg I1-L22 and its alleged place of origin, are consistent with the alleged route followed by some German peoples – Lombards and Goths in particular – from the Baltic shores to Italy.

The area of San Giovanni in Persiceto became part of the Lombard kingdom relatively late (728 AD, under king Aistulf) and their rule lasted only half a century, after having been defeated by the Franc king Charlemagne in 776 AD (Capitani, 2009). Nevertheless, a number of historical facts seem to link this area with Lombard settlements. Among them, the most relevant involves the Abbey of Nonantola – one of the most powerful monastic centres of the area. The Abbey was founded by Lombard kings in 752 (Bottazzi, 2003). Since then, *emphyteutic* grants from Nonantola (first documented evidence in 1170 AD) had a relevant role in the formation of the assets of the *Partecipanza*. San Giovanni in Persiceto itself, according to some scholars, was the seat of a Lombard Duchy in the second half of the 8th century (Bottazzi, 2003; Santos Salazar, 2006). Eventually, it has been reported the case of a burial place – discovered in the early 1960s but originally misinterpreted as a recent mass grave – showing Germanic features and recently radiocarbon dated to ~1000 years ago (D'Adamo & Pedrini, 2013).

Putting it all together, it seems plausible to relate historical and archaeological information with our molecular results, suggesting that a Lombard component may have had a key role in the foundation of the *Partecipanza*. As it is obvious, we are not implying that *Partecipanza* is a Lombard 'living fossil', neither that hg I1-L22 coincides with ancient Lombards. Their original genetic variability is still unknown and probably varied in time and place. It is anyway reasonable to believe that I1-L22 was an important part of it.

On the contrary, hgs J2-M67* and R-U152 reveal different aspects of *Partecipanza* formative history. Contrarily to I1-L22, DAPC analysis showed that J2-M67* PAR hts form two different clusters (Table S5, Figure S4), thus suggesting that their high frequency (compared to SGP) is the result of historic/geographic stratification. One of these clusters (cluster 2) is represented both in PAR and SGP and is frequent in Italy and Germany, yielding a date as old as 6000 YBP (Table 3). Another one (cluster 3) is found only in PAR, coinciding with a single pedigree (P05: Table S5, Figure S4) whose common ancestor lived in the 16th century. A search through the Yhrd database (www.yhrd.org) showed that its modal ht has a match in Apulia (Southern Italy).

R-U152 is the most frequent hg both in PAR and SGP, as well as in the whole of North-Western Italy (Boattini et al., 2013). While its origins are still unclear (Myres et al., 2011; Busby et al., 2012), its presence in the area of S. Giovanni in Persiceto is much older than the *Partecipanza* itself. Our dating experiments yielded dates ranging between 4000 and 5000 years ago for both populations (Table 3). In addition, R-U152 shows clear evidences of stratification both in PAR and SGP. In fact, DAPC revealed five different clusters, each of them including hts from both populations (Table S5, Figure S4)

CONCLUSION

In this study we showed that a well-conceived sampling strategy (population choice, selection of individuals) facilitates the identification of otherwise hidden population events and historical stratifications. While doing this, we generated relevant information on issues that transcend the regional setting of this study. The first result is the determination of important parameters such as Y-STR mutation rates and average generation duration time. Thanks to a combination of paternal pedigrees and Y-chromosome molecular data, we estimated a Y-STR mutation rate equal to $3.90 * 10^{-3}$ (17 STRs). Such rate is one order of magnitude higher than ‘evolutionary’ rates (Zhivotovsky et al., 2004) and very similar to ‘germline’ rates (Ballantyne et al., 2010). As for generation duration time our estimate (33.38 years) is in line with what is known for other European populations (UK; King & Jobling, 2009). These results have important consequences for popular issues as Y-chromosome molecular dating.

The second aim of this study was to check if socio-economic stratification within the same population can determine specific genetic sub-structures. Our results show that this is the case for our PAR and SGP samples. In particular, such structure is related to patrilines, while mtDNA showed no difference between the two samples. This agrees with the mechanism of transmission of the shared assets within the *Partecipanza*. Finally, by comparing PAR with SGP, as well as with Italian and European populations, we concluded that most of the *Partecipanza* founders bore Y-chromosome lineages that were already present in the area by a long time. Anyway, it seems plausible that such nucleus was in some way enriched by a Lombard component that settled here during the eighth century AD and continued to live there even after the defeat of the Lombard kingdom in 776 AD. The *Partecipanza*, thanks to its particular set of rules, preserved (and possibly amplified) such a historical ‘trace’, which went instead lost in the ‘open’ population living in the same area.

Future developments of our research will further explore some of the most important results of this study. As for the *Partecipanza* ancestry, we will replicate our experiment by extending our sampling to other similar institutions. As for Y-chromosome mutation rates, comparisons between Y-haplotypes and genealogies will be extended to a higher number of Y-chromosome markers and paternal pedigrees.

ACKNOWLEDGEMENTS

We thank all the volunteers who kindly accepted to participate to this study. We are indebted with all the Personnel of the “Consorzio dei Partecipanti di San Giovanni in Persiceto”, the Municipality of San Giovanni in Persiceto and the local AVIS centre (Associazione Volontari Italiani del Sangue) for helping us to design the sampling campaign and to collect the archival data. Special thanks are due to Vittorio Cocchi (President of the “Consorzio dei Partecipanti”), Raffaella Scagliarini (Responsible of the Historical Archive of the “Consorzio dei Partecipanti”), Renato Mazzuca (Mayor of San Giovanni in Persiceto) and Sonia Camprini (Council Member of San Giovanni in Persiceto).

REFERENCES

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, Pocheshkhova E et al. (2011). Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol 28: 2905-2920.

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

Behar DM, Van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM et al. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90: 675-684.

Boattini A, Luiselli D, Sazzini M, Useli A, Tagarelli G, Pettener D (2011a). Linking Italy and the Balkans. A Y-chromosome perspective from the Arbereshe of Calabria. Ann Hum Biol 38: 59-68.

Boattini A, Griso C, Pettener D (2011b). Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci 89:161-73.

Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F (2012). General method to unravel ancient population structures through surnames, final validation on Italian data. Hum Biol 84: 235-70.

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

Bottazzi G (2003). Monteveglio e Nonantola tra bizantini e longobardi. In: Cerami D (ed) Monteveglio e Nonantola: abbazie e insediamenti lungo le vie appenniniche, Centro Studi Storici Nonantolani: Nonantola, pp. 33-66.

Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG et al. (2012). The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279: 884-92.

Capitani O (2009). Storia dell'Italia Medievale, Laterza: Roma-Bari.

Capocasa M, Anagnostou P, Bachis V, Battaggia C, Bertoncini S, Biondi G et al. (2014). Linguistic, geographic and genetic isolation: a collaborative study of Italian populations. JASs, e-pub ahead of print. doi: 10.4436/JASS.92001

D'Adamo C, Pedrini W (2013). I 34 scheletri del Poggio. Cronaca di una scoperta archeologica, Maglio Editore: S. Giovanni in Persiceto.

Excoffier L, Laval G, Schneider S (2007). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1: 47-50.

Ferri G, Alù M (2012). Development of six Y-SNPs assay for forensic analysis in European population. DNA in Forensics 2012, 5th International EMPOP Meeting - 8th International Forensic Y-User Workshop, Innsbruck.

Jobling MA (2012). The impact of recent events on human genetic diversity. Phil Trans R Soc B 367: 793-799.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403-1405.

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

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830-838.

King TE, Jobling MA (2009). Founders, Drift, and Infidelity: The Relationship between Y Chromosome Diversity and Patrilineal Surnames. Mol Biol Evol 26:1093-1102.

Komsta L (2011). outliers: Tests for outliers. R package version 0.14. http://CRAN.R-project.org/package=outliers

Lacan M, Keyser C, Crubézy E, Ludes B (2012). Ancestry of modern Europeans: contributions of ancient DNA. Cell Mol Life Sci 70: 2473-87.

Larmuseau MH, Vanderheyden N, Jacobs M, Coomans M, Larno L, Decorte R (2011). Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant. Forensic Sci Int Genet 5: 95-99.

Larmuseau MHD, Van Geystelen A, van Oven M, Decorte R (2013). Genetic Genealogy Comes of Age: Perspectives on the Use of Deep-Rooted Pedigrees in Human Population Genetics. Am J Phys Anthropol 150: 505-511.

Laval G, SanCristobal M, Chevalet C (2002). Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet Sel Evol 34: 481-507.

Martínez-Cruz B, Harmant C, Platt DE, Haak W, Manry J, Ramos-Luis E et al. (2012). Evidence of pre-Roman tribal genetic structure in Basques from uniparentally inherited markers. Mol Biol Evol 29: 2211-22.

Myres NM, Rootsi S, Lin AA, Järve M, King RJ, Kutuev I et al. (2011). A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet 19: 95-101.

Onofri V, Alessandrini F, Turchi C, Pesaresi M, Buscemi L, Tagliabracci A (2006). Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. Forensic Sci Int 57: 23-35.

R Core Team (2013). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rębała K, Martínez-Cruz B, Tönjes A, Kovacs P, Stumvoll M, Lindner I et al. (2013). Contemporary paternal genetic landscape of Polish and German populations: from early medieval Slavic expansion to post-World War II resettlements. Eur J Hum Genet 21: 415-22.

Regueiro M, Rivera L, Damnjanovic T, Lukovic L, Milasin J, Herrera RJ (2012). High levels of Paleolithic Y-chromosome lineages characterize Serbia. Gene 498: 59-67.

Reynolds JB, Weir BS, Cockerham CC (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767-779.

Santos Salazar I (2006). Castrum Persiceta: Potere e territorio in uno spazio di frontiera dal secolo VI al IX, Reti Medievali Rivista 7: 1-20.

Sarno S, Boattini A, Carta M, Ferri G, Alù M, Yang Yao D et al. (2014). An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of Sicily and Southern Italy. PLoS One submitted.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE et al. (2006). Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian Pastoralists. Am J Hum Genet 78: 202-221.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ et al. (2010). The Archaeogenetics of Europe, Curr Biol 20: R174-R183.

Underhill PA, Myres NM, Rootsi S, Chow CT, Lin AA, Otillar RP et al. (2007). New Phylogenetic Relationships for Y-chromosome Haplogroup I: Reappraising its Phylogeography and Prehistory. In: Mellars P, Boyle K, Bar-Yosef O, Stringer C (eds) Rethinking the Human Revolution, McDonald Institute for Archaeological Research: Cambridge (UK), pp. 33-42.

Zanarelli M (1992). I beni comuni e le forme di gestione attuate dalle comunità rurali: il caso di San Giovanni in Persiceto e di Medicina. In: Fregni (ed) Terre e Comunità nell'Italia Padana. Il caso delle Partecipanze Agrarie Emiliane: da beni comuni a beni collettivi, Edizioni Centro Federico Odirici: Mantova, pp. 147-173.

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

FIGURES

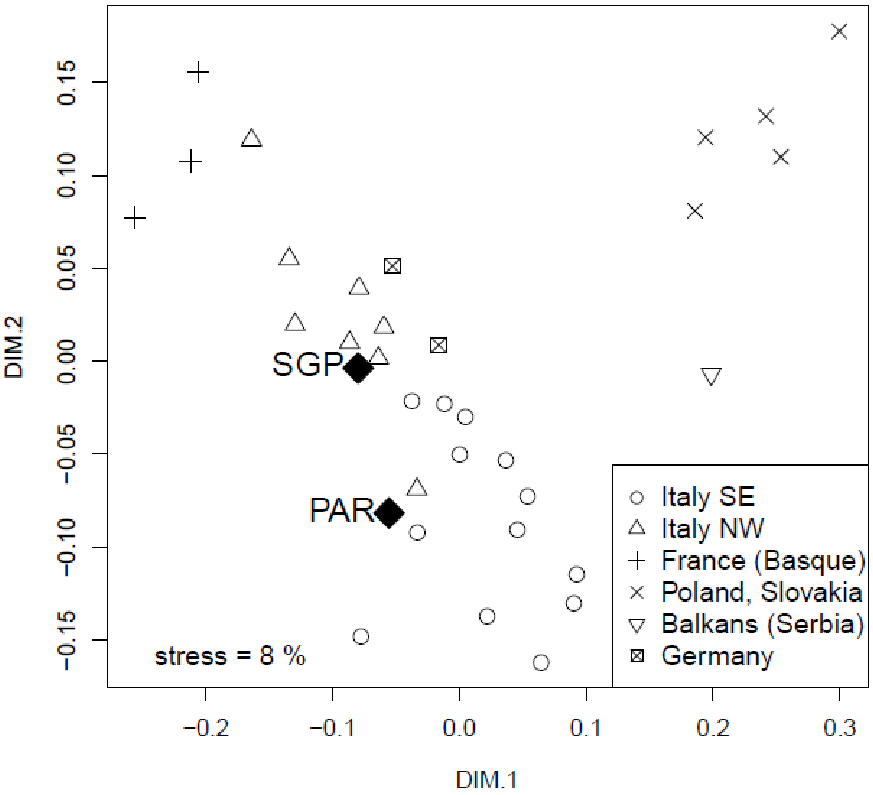


Figure 1. Non-Metric MDS representation of PAR, SGP and reference populations based on Y-STR data.

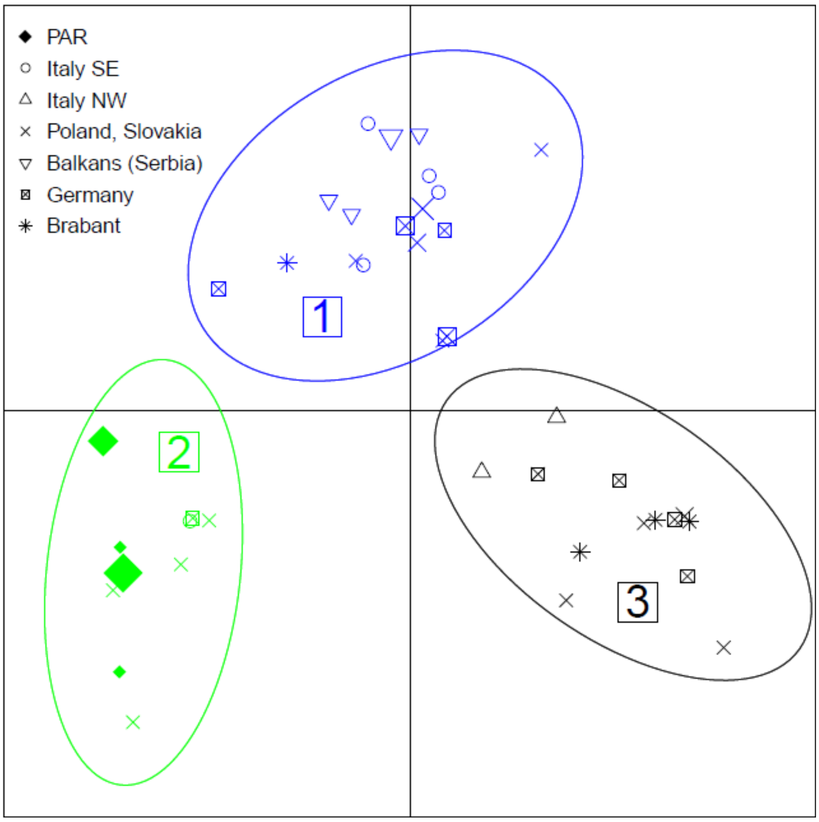


Figure 2. Discriminant Analysis of Principal Components (DAPC) of Y-STR variation in I1-L22 haplotypes from PAR and reference populations. Scatterplot of the first and the second discriminant functions. Symbol size is proportional to the frequency of Y-STR haplotypes.

TABLES

Table 1. Relationships between reconstructed paternal pedigrees and Y-STR profiles, including mutation information (NH: number of haplotypes, NM: number of observed mutations, NG: number of considered generations) and generation time (NI: number of individuals, NG: number of considered generations, YRS: total number of years).

		Mutations			Gen. Time		
Pedigree	**Surname**	**NH**	**NM**	**NG**	**NI**	**NG**	**YRS**
P01	S01	2	0	21	2	19	605
P02	S02	3	0	20	3	18	627
P03	S03	3	2	14	4	25	855
P04	S04	4	6	45	4	35	1166
P05	S05	8	3	75	8	67	2302
P06	S06	2	0	14	4	32	1066
P07	S06	3	3	34	5	37	1226
P08	S07	2	0	11	2	15	470
P09	S08	2	1	6	2	12	413
P10	S09	2	1	7	2	12	412
P11	S10	3	4	23	3	23	755
P12	S11	2	2	22	2	20	608
P13	S12	3	0	32	3	25	829
P14	S13	2	2	21	2	17	582
P15	S13	2	0	17	2	18	563
P16	S14	-	-	-	1	11	323
P17	S15	-	-	-	1	9	337
P18	S16	-	-	-	1	8	298
P19	S17	-	-	-	1	10	312
P20	S18	-	-	-	1	9	307
P21	S19	-	-	-	1	9	305
P22	S20	-	-	-	1	9	289
P23	S21	-	-	-	1	8	290
P24	S22	-	-	-	1	8	311
P25	S23	-	-	-	1	9	302
P26	S24	-	-	-	1	10	302
P27	S25	-	-	-	1	10	308
P28	S26	-	-	-	1	7	273
P29	S27	-	-	-	1	8	284
P30	S27	-	-	-	1	10	315
P31	S28	-	-	-	1	5	167
P32	S28	-	-	-	1	9	287
P33	S28	-	-	-	1	9	298
P34	S28	-	-	-	1	7	275
P35	S13	-	-	-	1	8	295
P36	S13	-	-	-	1	8	285
P37	S13	-	-	-	1	9	302
P38	S13	-	-	-	1	10	320
P39	S29	-	-	-	1	10	300
P40	S30	-	-	-	1	9	281
P41	S31	-	-	-	1	10	315
TOT		43	24	362	74	604	20160

Table 2. Y-hg frequencies in PAR, PAR1600 and SGP. Hgs whose frequencies are significantly different between PAR and SGP are in bold.

	PAR		PAR1600		SGP		PAR vs. SGP
HG	**N**	**%**	**N**	**%**	**N**	**%**	**p-val**
E-V13	1	1.14	1	2.08	3	4.92	> 0.05
E-V22	1	1.14	0	0.00	0	0.00	> 0.05
G-U8*(x U13)	0	0.00	0	0.00	3	4.92	> 0.05
G-U13	0	0.00	0	0.00	2	3.28	> 0.05
I-M253*(x L22)	4	4.55	1	2.08	3	4.92	> 0.05
I-L22	**14**	**15.91**	**9**	**18.75**	**0**	**0.00**	**0.0004**
I-M223	3	3.41	3	6.25	1	1.64	> 0.05
J-M267	1	1.14	1	2.08	2	3.28	> 0.05
J-L27*(x M67)	8	9.09	6	12.50	7	11.48	> 0.05
J-M67*(x M92)	**12**	**13.64**	**4**	**8.33**	**2**	**3.28**	**0.0443**
J-M12	5	5.68	2	4.17	1	1.64	> 0.05
R-M420	3	3.41	2	4.17	2	3.28	> 0.05
R-M343*(x M269)	0	0.00	0	0.00	1	1.64	> 0.05
R-M269*(x U152)	11	12.50	6	12.50	6	9.84	> 0.05
R-U152	**24**	**27.27**	**12**	**25.00**	**27**	**44.26**	**0.0363**
T-L445	1	1.14	1	2.08	1	1.64	> 0.05
Total	88		48		61		

Table 3. Time estimates (in years before present) based on DAPC clusters of Y-STR hts (PAR: only PAR hts; SGP: only SGP hts; ALL: all hts) using two different mutation rates (MR): our average pedigree-based rate (P) and Ballantyne et al.'s (2012) STR specific rates (B). NH = number of haplotypes, AGT = Average Generation Time.

Hg	Clusters (POP)	NH	MR	AGT	Time Estimate
I1-L22	2 (PAR)	14	B	33.38	667.98 +/- 172.47
			P	33.38	570.67 +/- 147.35
I1-L22	2 (ALL)	20	B	33.38	1410.65 +/- 364.23
			P	33.38	1318.37 +/- 340.4
J-M67	1 (ALL)	20	B	33.38	6318.2 +/- 1631.35
			P	33.38	9818.17 +/- 2535.04
R-U152	1&2 (PAR)	12	B	33.38	4275.1 +/- 1103.83
			P	33.38	3464.66 +/- 894.57
R-U152	1&2 (SGP)	17	B	33.38	3959.67 +/- 1022.38
			P	33.38	4785.77 +/- 1235.68
R-U152	1&2 (ALL)	125	B	33.38	2925.04 +/- 755.24
			P	33.38	2493.9 +/- 643.92
R-U152	3 (PAR&SGP)	11	B	33.38	5591.59 +/- 1443.74
			P	33.38	2960.48 +/- 764.39

SUPPLEMENTARY MATERIALS

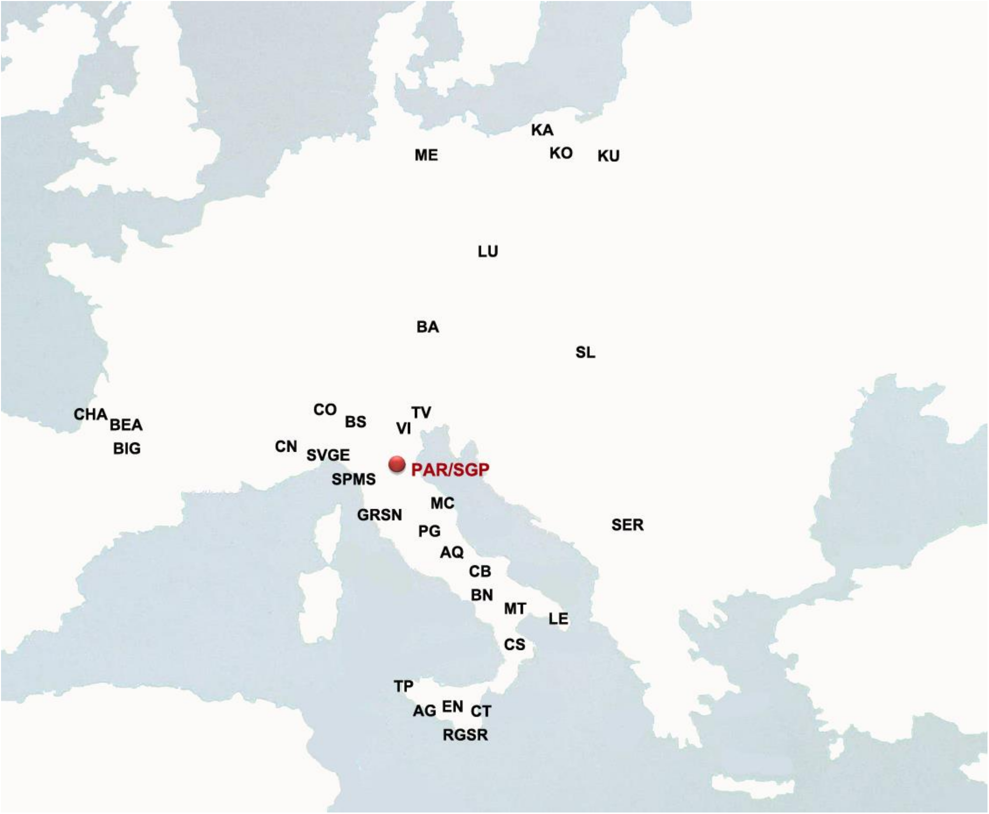


Figure S1. Geographic position of PAR, SGP and reference populations.

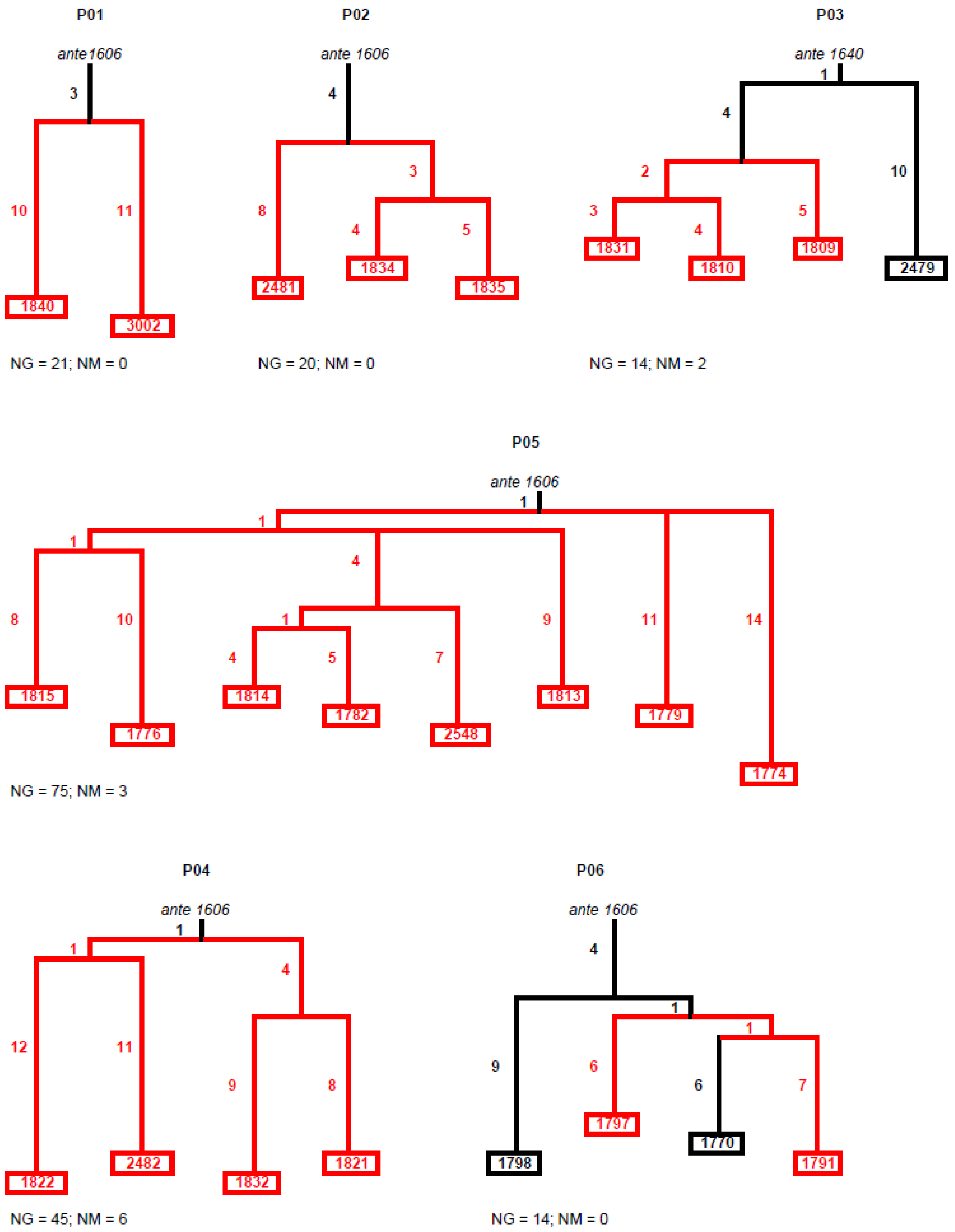


Figure S2. Paternal pedigrees encompassing at least two individuals (P01-P15). Black branches were not considered for the calculation of Y-STR mutation rate. Numbers along each branch represent the corresponding number of generations. For each pedigree is reported the number of considered generations (NG) and the number of observed mutations (NM).

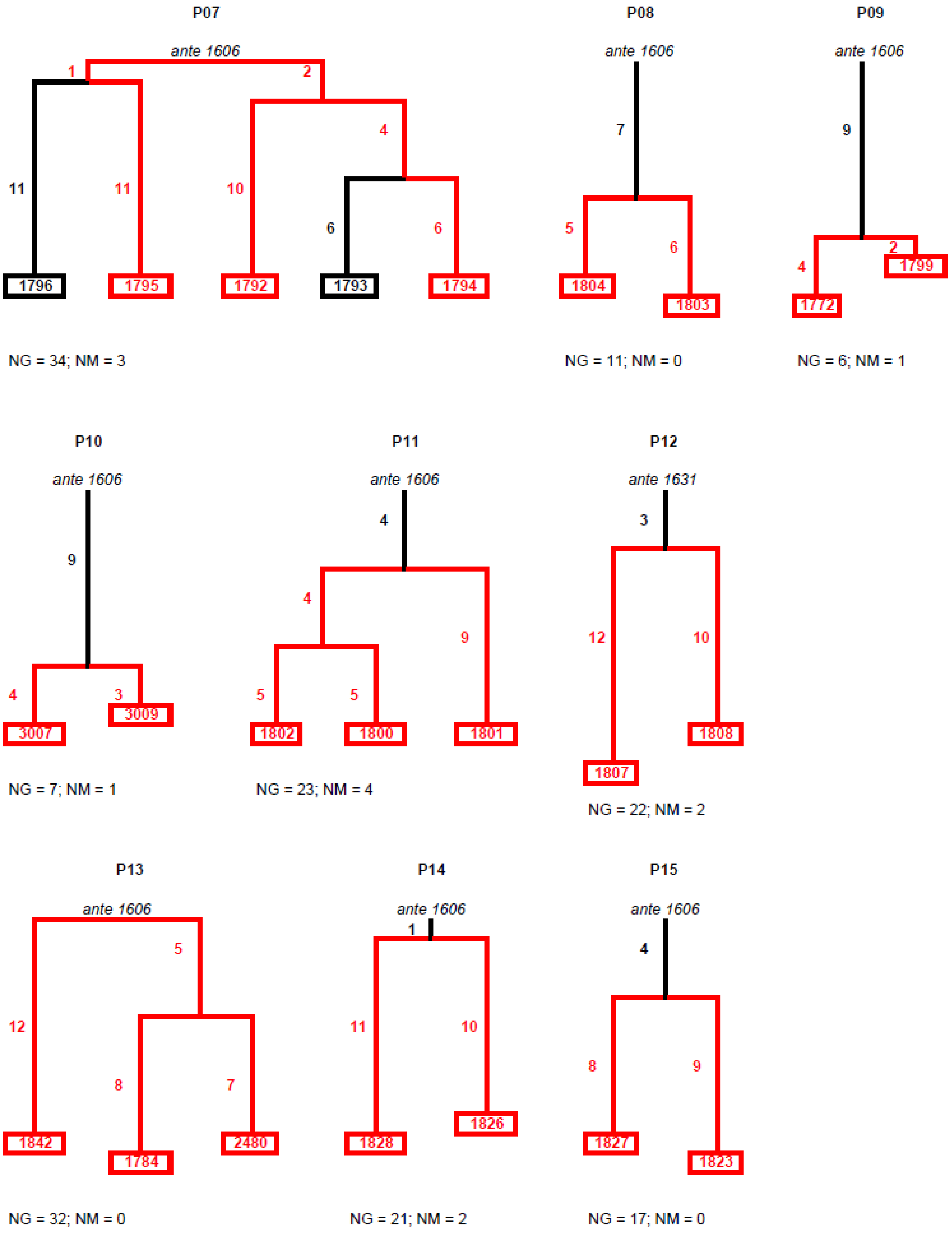


Figure S2. (*Continued*)

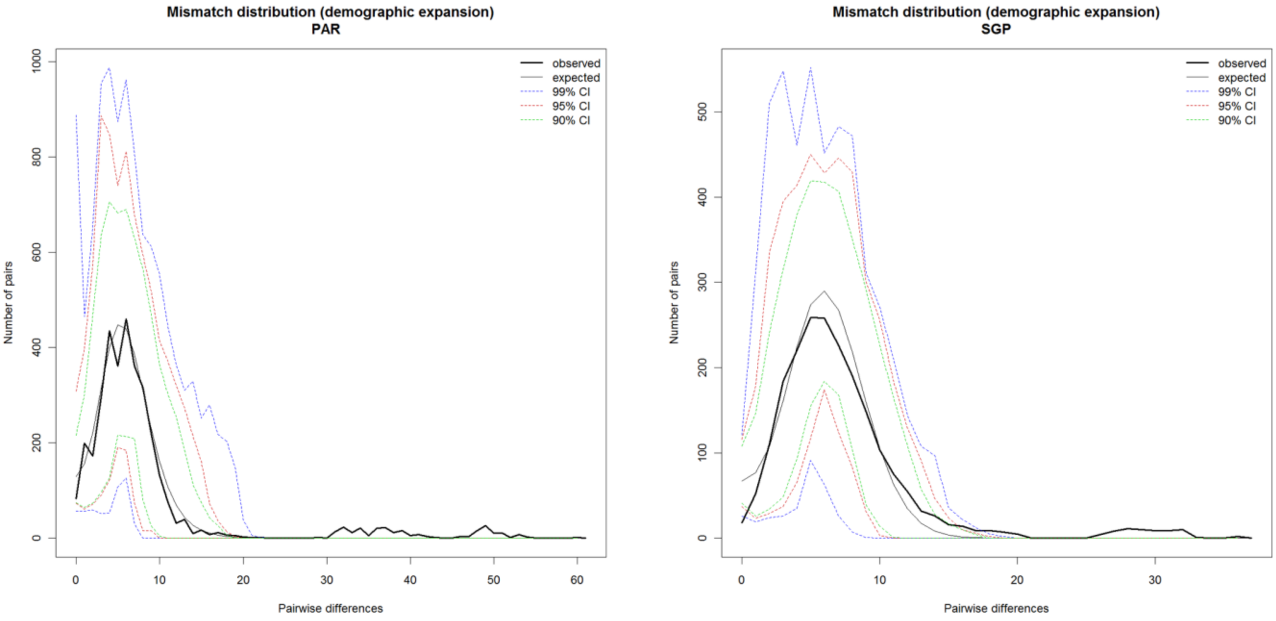


Figure S3. Mismatch distributions for PAR and SGP samples.

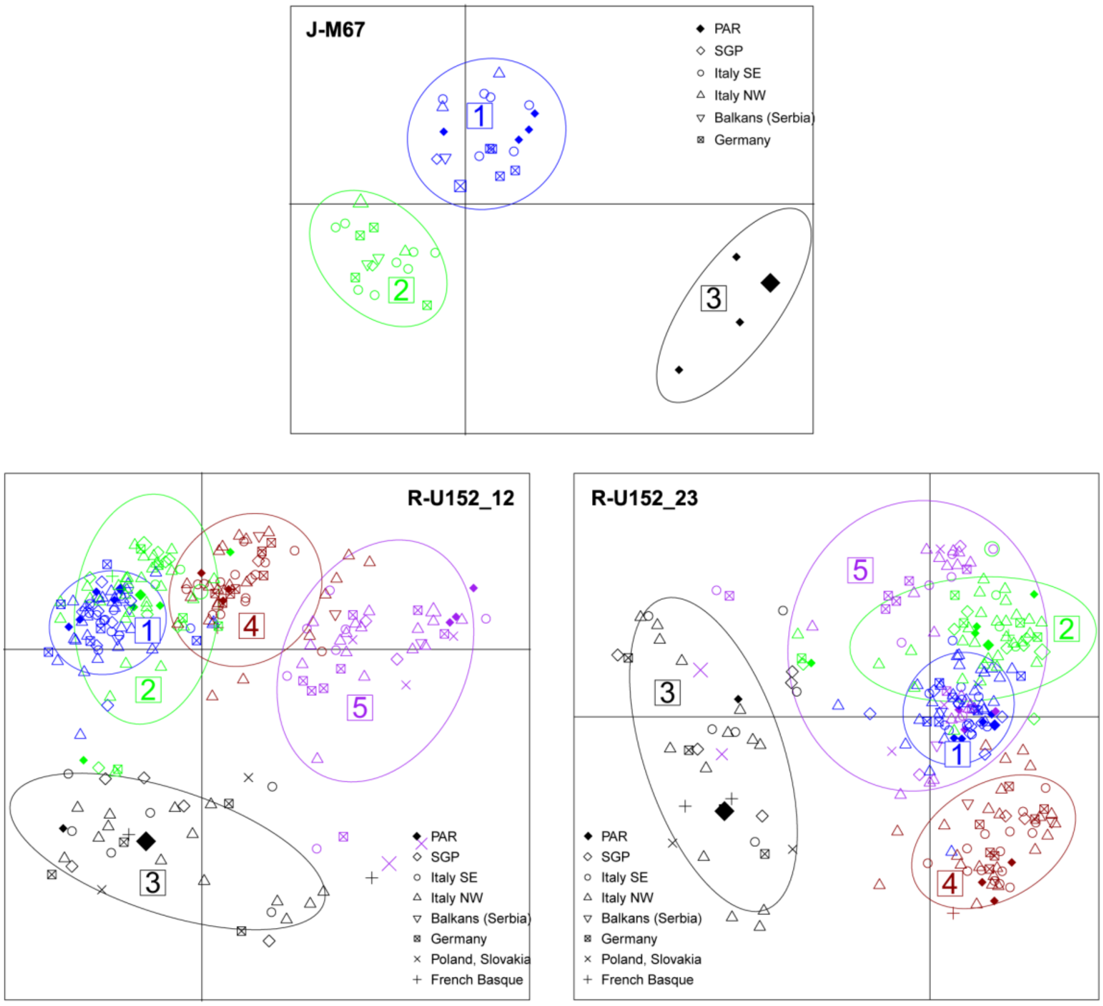


Figure S4. Discriminant Analysis of Principal Components (DAPC) of Y-STR variation in J-M67* and R-U152 haplotypes from PAR, SGP and reference populations. Scatterplots of the first and second discriminant functions (J-M67*, R-U152_12) and of the second and third discriminant functions (R-U152_23). Symbol size is proportional to the frequency of Y-STR haplotypes.

Table S1. **Reference populations for Y-STR data. N = sample size.**

COD	Description	Region	N	Reference	COD	Description	Region	N	Reference
CN	Cuneo	Continental Italy	30	Boattini et al., 2013	AG	Agrigento	Sicily	45	Boattini et al., 2013
SVGE	Savona-Genova	Continental Italy	51	Boattini et al., 2013	CT	Catania	Sicily	52	Boattini et al., 2013
CO	Como	Continental Italy	41	Boattini et al., 2013	RGSR	Ragusa-Siracusa	Sicily	45	Boattini et al., 2013
BS	Brescia	Continental Italy	39	Boattini et al., 2013	EN	Enna	Sicily	40	Sarno et al., 2014
VI	Vicenza	Continental Italy	40	Boattini et al., 2013	TP	Trapani	Sicily	34	Sarno et al., 2014
TV	Treviso	Continental Italy	33	Boattini et al., 2013	BIG	Bigorre	French Basque	44	Martinez-Cruz et al., 2012
SPMS	La Spezia-Massa	Continental Italy	24	Boattini et al., 2013	BEA	Bearn	French Basque	56	Martinez-Cruz et al., 2012
GRSN	Grosseto-Siena	Continental Italy	86	Boattini et al., 2013	CHA	Chalosse	French Basque	58	Martinez-Cruz et al., 2012
MC	Macerata	Continental Italy	40	Boattini et al., 2013	KA	Kaszuby	Poland	204	Rebala et al., 2012
PG	Perugia	Continental Italy	37	Boattini et al., 2013	KO	Kociewie	Poland	158	Rebala et al., 2012
AQ	L'Aquila	Continental Italy	30	Boattini et al., 2013	KU	Kurpie	Poland	158	Rebala et al., 2012
CB	Campobasso	Continental Italy	30	Boattini et al., 2013	LU	Lusatia	Germany	123	Rebala et al., 2012
BN	Benevento	Continental Italy	35	Boattini et al., 2013	SL	Slovakia	Slovakia	164	Rebala et al., 2012
MT	Matera	Continental Italy	25	Boattini et al., 2013	ME	Mecklenburg	Germany	131	Rebala et al., 2012
LE	Lecce	Continental Italy	40	Boattini et al., 2013	BA	Bavaria	Germany	218	Rebala et al., 2012
CS	Cosenza	Continental Italy	45	Sarno et al., 2014	SER	Serbia	Balkans	102	Regueiro et al., 2012

Table S2. **Y-STR haplotypes and haplogroups in PAR and SGP samples.** For the PAR sample, pedigree and surname codes are reported. For pedigrees encompassing at least two individuals, the number of mutations from the Most Recent Common Ancestor (NM2MRCA) is included. This table will be provided in the online version of the manuscript once published.

Table S3. **mtDNA HVSI-HVSII sequence data in PAR and SGP.** This table will be provided in the online version of the manuscript once published.

Table S4. **Standard diversity indices computed for PAR, PAR1600 and SGP.**

Y-chromosome		STR			SNP
Pop	**Sample Size**	**Gene Div. (h) ± sd**	**MNPD (π) ± sd**	**Nuc. Div. (π_N) ± sd**	**Gene div (h) ± sd**
PAR	88	0.9843 +/- 0.0055	9.5833 +/- 4.4361	0.5990 +/- 0.3072	0.8595 +/- 0.019
PAR1600	47	0.9861 +/- 0.0096	9.6022 +/- 4.4816	0.6001 +/- 0.3109	0.8677 +/- 0.0251
SGP	61	0.9995 +/- 0.0031	8.6787 +/- 4.0627	0.5786 +/- 0.3004	0.7814 +/- 0.0493
mtDNA		**SEQ**			
Pop	**Sample Size**	**Gene div (h) ± sd**	**MNPD (π) ± sd**	**Nuc. Div. (π_N) ± sd**	
PAR	88	0.9762 +/- 0.0102	5.9658 +/- 2.8739	0.0090 +/- 0.0048	
SGP	61	0.9913 +/- 0.0049	6.4629 +/- 3.0982	0.0097 +/- 0.0052	

Table S5. **Frequencies of Y-chromosome DAPC clusters in PAR, SGP and reference populations** (BALK = Balkans-Serbia, SEI = South-Eastern Italy, NWI = North-Western Italy, GER = Germany, POL/SLO = Poland/Slovakia, FRA_B = French Basque).

HG	CLUSTER	PAR	SGP	BALK	SEI	NWI	GER	BRAB	POL/SLO	FRA_B
I1-L22	1	**0**	**0**	5	4	0	6	1	8	0
	2	**14**	**0**	0	1	0	1	0	4	0
	3	**0**	**0**	0	0	2	4	3	5	0
J-M67	1	**4**	**1**	1	6	2	6	-	0	0
	2	**0**	**1**	2	8	3	4	-	0	0
	3	**8**	**0**	0	0	0	0	-	0	0
R-U152	1	**5**	**8**	2	12	26	6	-	1	0
	2	**5**	**8**	0	4	29	5	-	0	1
	3	**2**	**5**	0	6	14	4	-	2	2
	4	**3**	**2**	2	15	23	7	-	0	1
	5	**3**	**2**	1	7	17	8	-	5	0

2.4. Comments

2.4.1 Isolation, culture and genetic history

As demonstrated in the first part of this thesis, the genetic structure of the extant Italian populations retains the signatures of complex peopling and demographic processes occurred from the Palaeolithic to the more recent human history. Different migrations waves during the Neolithic and Metal Ages particularly proved to have characterised the pre- and proto-history of Italy, and were invoked to explain the genetic sub-structure (North-West Italy vs. South-East Italy) observed for Y-chromosome markers, compared to the more homogeneous maternal (mtDNA) genetic landscape (*Paper 1* and *Paper 2*). Further contributions to the genetic heterogeneity currently found in the Italian population, came from the historical gene flows and cultural exchanges that affected the population structure and admixture at more recent time frames and restricted geographical scales. Although the underlying genetic impacts of these processes remain often hidden - due to the confounding consequences of their interactions, the narrower time and/or spatial frames of their actions, as well as the lower impact and more localised extent of their effects - all these events have actually contributed to enrich the bio-anthropological and cultural variability of Italy, as attested by the large number of cultural enclaves and ethno-linguistic minorities today settled in the Italian territory.

In a recent effort to obtain a first comprehensive characterisation of the Italian micro-evolutionary genetic variability, a collaborative project – involving four research units from four Italian Universities (Bologna, Cagliari, Pisa and Roma “La Sapienza”) – have analysed and compared the patterns of genetic diversity among several geographic isolates and/or ethno-linguistic minorities – stretching from the north-eastern Italian Alps to Sicily and Southern Italy – in order to evaluate the association of linguistic and/or geographic factors with genetic isolation (Capocasa et al., 2014, *Appendix I*). Although the condition of geographic marginality demonstrated to generate by itself significant reproductive barriers (as in the case of the geographic isolate of Benetutti, in Northern Sardinia), the source of higher genetic isolation was shown to be the combination of geographic with linguistic factors (Capocasa et al., 2014). Cultural – and particularly linguistic – features acting also as genetic barriers, have been indeed invoked for geographic and cultural isolates – such as the Ligurian linguistic island of Carloforte in southern Sardinia and the German-speaking minorities of Eastern Alps (i.e. Sappada, Sauris and Timau) – thus underlying the importance of considering also the effects of socio-cultural factors when studying the human genetic variation (Capocasa et al., 2014).

Within this framework, our findings (*Paper 3* and *Paper 4*) have extended this concept to a wider perspective that transcends the specific case of “cultural traits as sources of genetic isolation”, but provides new insights concerning both the importance of sampling strategy and the role of socio-cultural elements in helping our understanding of underlying population structures and hidden genetic layers.

Besides being responsible for genetic isolation, cultural features have indeed proved to generate significant genetic sub-structures within the same population (Partecipanza, *Paper 4*) and to interact with geographic and historical factors in differentiating the micro-evolutionary histories within the same ethno-linguistic group (Arbereshe, *Paper 3*), thus getting a key role in shaping the present-day genetic composition of the so-called "marginal populations". This term identifies those human groups that experienced reduced population sizes and gene flows – compared to the neighbouring open (outbred) populations. These conditions may eventually (but not necessarily) evolve in genetic isolation. The impact of cultural features on the genetic structure of these populations explained the emerging of the peculiarities in HG composition observed for both Arbereshe communities (*Paper 3*) and the Partecipanza of San Giovanni in Persiceto (*Paper 4*). In addition, such peculiarities accounted for the differential maintenance by these marginal groups of specific ancestral paternal lineages. Due to their cultural distinctiveness - that at least partially limited the genetic exchanges with the nearby human populations - ethno-linguistic minorities (Arbereshe, *Paper 3*) and cultural enclaves (Partecipanza, *Paper 4*) have not only represented exceptional observatories to test the impact of socio-cultural traditions on population genetic structures, but have also acted as "*survival areas*" for genetic features, which can be related to migration and demographic processes occurred in continental Italy during the more recent history. The high levels of genetic diversity observed for both the two analysed populatons, besides suggesting that neither Arbereshe nor the Partecipanza have experienced significant drift effects or isolation phenomena, are indeed consistent with the hypothesis that these marginal populations have preserved much of their ancestral variability along with their cultural features. This highlights how linguistic and/or socio-economic factors may have helped maintaining a connecting link between present-day populations and specific layers of the Italian past history.

In this context, carefully designed sampling strategies demonstrated to be particularly suited to uncover such a cryptic diversity patterns and genetic layers (*Paper 3* and *Paper 4*). The widely used "*grandparents criterion*" sampling method (i.e. three generation of ancestry in the study area), if on one hand represents a quick way to exclude the more recent events of gene flows, on the other hand prevents the identification of immigration processes beyond the second generation, thus precluding the possibility to make reliable inferences about the structure of a population at deeper geographical and temporal scales (Boattini et al., 2011). At this respect, surname-based sampling strategies, coupled with the accurate selection of marginal populations, have actually proved to be a valuable solution to overcome these problems and manage to unmask otherwise hidden population genetic sub-structures (*Paper 3* and *Paper 4*). In particular, the association between patrilineal surnames and the uniparentally-transmitted male-specific Y-chromosomes, showed that surname analysis (i.e biodemography) is a particularly effective tool for sampling individuals in Y-chromosome based studies. The selection of males bearing surnames which are unequivocally associated with a certain place (founder surnames) ensures a phylogenetic picture of lineage diversity as closer as possible to the "ancestral" genetic variability, thus preceding the more recent reshuffling events of admixture within and among populations. The surname-based sampling strategy (especially if accompanied by an exhaustive representation of the "original founder surnames" and by the selection of at least one representative for each of them) allows maximising the number of different paternal lineages for a given sample size, retracing at the period of surname spread (4-5 centuries ago) and therefore excluding the more recent immigration processes by default (Boattini et al., 2011). In this context, the reconstruction of genealogies/pedigrees, despite being comparatively more difficult and time-consuming, has proved to add an even higher degree of geographic precision to the selection of samples, also offering the possibility to approach some important issues in population genetic studies which are necessary to make faithful inferences on population past history (*Paper 4*).

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

1. Integrating macro- and micro-geographic perspectives of investigation

Historical, geographic and cultural factors interact in shaping the genetic variability of current human populations at global, regional and local levels. However, the patterns of genetic variation observed at continental-wide perspectives, are not always reflected also when fine-grained geographic and temporal scales are considered (Beleza et al. 2005, Sanchez-Faddeev et al. 2013). In order to disentangle the spatial and temporal processes in the origin of present-day human genetic variability, it is therefore important to address fine-scale questions about human population history and genetic variation, by integrating continent-wide studies with the more detailed pictures that are instead achievable by means of regional and/or local case-studies.

In this thesis we approached some issues of the Italian genetic history and population variability, by particularly combining macro- and micro-geographic investigations, at different spatial and temporal levels. The main scientific questions concerned: i) exploring the uniparental genetic structures of the Italian population and tracing the times and sources of the observed diversity patterns; ii) searching for the legacy of different continental and within-continental contributions to the current genetic make-up in specific Italian local contexts, in order to clarify broad- and fine-scale aspects of the human population history; iii) characterising the putative impacts of cultural factors (i.e. language and socio-economic conditions) in preserving or reshaping the genetic variability within specific historical and geographic layers, by using key-model populations.

By combining a well-conceived sampling strategy (implying adequate sampling coverage and accurate selecting criteria), with an unprecedented in-depth resolution level for the molecular analysis (implying the joint analysis of slow- and fast-evolving markers in both uniparental systems at the same time), we were able to unveil genetic structures previously undetected in the Italian genetic landscape. Sex-biased patterns and contrasting demographic histories have been observed for males and females, highlighting different temporal and spatial phases in the construction of the Italian genetic make-up (*Paper 1*). In particular, mtDNA results suggest a more ancient and homogeneous distribution pattern of genetic variation, justified by both male-biased population incomings (Pesando et al., 2005; Lacan et al., 2011a; Lacan et al., 2011b) and higher female-mobility dynamics since the Neolithic times (Rastiero et al. 2012; Heyer et al., 2012). On the other hand the NWI-SEI genetic sub-structure observed for the paternal lineages underlines the impact of different routes of Neolithisation along the Adriatic and Tyrrhenian coasts, coupled with different post-Neolithic settlements within continental Italy (*Paper 1*). In accordance with the natural-bridge-position of the Italian Peninsula between Europe and the Mediterranean, different migration processes from the Balkans/Levant and Central Europe, mainly referable to the post-Neolithic and Metal Ages, in fact proved to have particularly affected the South-Eastern and North-Western

Italian genetic pools respectively (*Paper 1* and *Paper 2*). Consistently with these results, a recent genome-wide investigation that used long shared segments of IBD to infer the recent common ancestry across modern Europeans (Ralph and Coop, 2013), revealed that the common ancestors between Italians and the other European populations are older than 2,500 YBP, therefore suggesting that the proto-historic genetic structure of Italy was not deleted or significantly modified during the Roman and Post-Roman (Migration Period) historical invasions (Ralph and Coop, 2013).

If the Neolithic revolution could be considered one of the principal determinant of the Y-chromosome genetic structuring observed within the Italian Peninsula, on the other hand these results taken together also highlight the importance of various continental and local Post-Neolithic contributions to the current diversity patterns, suggesting that the genetic structure observed in the Italian Peninsula was not fixed in the Neolithic period, but further reinforced by subsequent demographic processes. Among them, the Greek domination during the Metal Ages confirmed to have played an important role in the cultural and genetic transitions occurred in Sicily and Southern Italy (SSI), accounting at least in part for the observed shared genetic background between SSI and the Balkan Peninsula (*Paper 2*).

Within the confounding tapestry of migration processes and cultural exchanges that affected the Italian genetic landscape, geographic outliers (Sardinians, *Paper 1*) and culturally-distinct human groups (ethno-linguistic and socio-economic marginal populations, *Paper 3* and *Paper 4*), particularly proved to have preserved specific traces of more ancient (Pre-Neolithic) or recent layers (e.g. Germanic invasions) of the Italian genetic history, also providing unique opportunities to test the effect of cultural, geographic and demographic factors on the micro-geographic genetic variability.

The demographic history and the geographically-isolated condition of the Sardinian population, confirmed to have modelled the uniqueness of its current genetic pool, indicating paralleling founder events for specific maternal and paternal lineages that today mark the genetic make-up of the people of the island relative to the rest of European populations (*Paper 1* and Pala et al., 2009). At the same time, and consistently with the suggestions of both aDNA-based studies (Keller et al., 2012) and genome-wide investigations (Di Gaetano et al., 2012), its condition of geographic impermeability has also allowed present-day Sardinians to maintain signatures of a more ancient genetic background, probably shared among all the continental European populations at least before the Post-Neolithic reshuffling migration processes (*Paper 1*).

On the other hand, the study of culturally-marginal populations allowed us to focus on the more recent layers of the Italian genetic landscape and to cast new lights into different aspects of the gene-culture interaction at finer geographical and temporal scales. Given the common Balkan origins and the shared cultural traits (Arberisht language, Greek Orthodox religion and common mytho-history), the differences in the genetic composition observed both within and among the Sicilian and Calabrian Albanian-speaking enclaves, provide an intriguing illustration of how cultural, historical and geographic relationships between human groups can intervene in shaping and differentiating the current genetic variability even among culturally and historically related populations (*Paper 3*). Similarly, the case of the *Partecipanza* of San Giovanni in Persiceto shows how social-economic stratifications may induce sex-biased genetic structuring within the same population (*Paper 4*).

The careful selection of key-model populations, coupled with the accurate selection of individuals within them - by means of founders surnames and/or deep-rooted pedigrees - have additionally offered a powerful tool to explore otherwise hidden historical strata of the Italian genetic landscape, for instance helping to retrieve genetic signals of the so-called “Migration period”. This period - spanning from the Huns invasions at the end of the 4th century to the expansion of Slavic populations throughout the 6th-10th centuries – witnessed a series of settlements by Germanic peoples (the most notably for Italy being the Ostrogoths and Lombards) that, between the 5th and the 6th centuries AD, spread into Europe ruling out most of the western Roman Empire. Despite their historical significance and widespread repercussions, these peopling events appeared to have brought negligible modifications to the Italian genetic landscape (Ralph and Coop, 2013), consequently leaving - if any - only subtle genetic traces in the gene pool of current human populations. Our results (*Paper 4*) demonstrated that the occurrence of idiosyncratic cultural conditions may have helped preserving (or even amplify) such historical “traces”. At the same time, the availability of extended genealogical data coupled with the surname-based analysis, revealed to be crucial to succeed in detecting such genetic signals. Sampling male individuals based on founder surnames can provide a set of Y-chromosome lineages that, differing from the more general patterns of genetic variation, is actually more likely to closely reflect (or contain traces of) the “ancestral” population diversity.

In addition, the association of deep-rooted pedigrees with in-depth Y-chromosome genetic profiles has allowed adding new insights into some fundamental issues in population genetic studies particularly concerning the Y-chromosome molecular dating (i.e. the estimation of the average generation duration time and the Y-STRs mutation rate per generation), which are essential to reliably infer the ages of common ancestors and thus to rigorously link specific genetic traces with particular historical events.

On the whole, the results of this research add some important pieces to the complex mosaic of the Italian genetic history, and particularly emphasise the usefulness of combining complementary scales of investigation to achieve a comprehensive evaluation of the several factors involved in the shaping of human genetic structures, both within and among extant populations. For a future perspective, extending the same study-approach to other Italian specific contexts, by including both open and marginal populations, will offer the possibility to achieve an increasingly detailed picture of the Italian genetic variability. Uniparental markers proved to be an extremely useful tool to cast new lights into specific aspects of the Italian genetic history since, due to their peculiar features, allowed to trace reliable genealogies back in time and space. However, each of these systems behaves as a single locus, thus offering only partial perspectives compared to the overall picture of population demographic history. Therefore, the complementary overview achievable with whole-genome autosomal SNPs data will surely allow deepening the perspectives offered by uniparental markers, adding further details to the "general picture" and overcoming some of the limitations of maternally and paternally inherited systems. Finally, enforcing a multi-disciplinary approach, by comparing and integrating genomic data with the results offered by other disciplines (e.g. linguistics, archaeology, aDNA-based studies), will represent a key-step to reveal and interpret some of the still unsolved questions of the Italian genetic history and population variability, at both broad and fine geographical and temporal scales.

BIBLIOGRAPHY

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogväli EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet. 75:910-918.

Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, et al. (2007) Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet 80: 759–768. doi: 10.1086/512822

Ammerman AJ, Cavalli-Sforza LL (1984) The Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton

Antonioli, F., Lambeck, K., Amorosi, A., Belluomini, G., Correggiari, A., Devoti, S., Demuro, S., Monaco, C., Marocco, R., Pagliarulo, R., Orru, P., Silenzi, S., 2004. Sea level at 8 and 22 ka cal BP along the Italian coastline. In: Antonioli, F., Vai, G.B. (Eds.), Climax Maps Italy: Explanatory Notes, Bologna, Italy, pp. 11e14.

Arcos-Burgos M, Muenke M (2002) Genetics of population isolates. Clin Genet. 61:233-47.

Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, Pocheshkhova E, Haber M, Platt D, Schurr T, Haak W, Kuznetsova M, Radzhabov M, Balaganskaya O, Romanov A, Zakharova T, Soria Hernanz DF, Zalloua P, Koshel S, Ruhlen M, Renfrew C, Wells RS, Tyler-Smith C, Balanovska E; Genographic Consortium (2011) Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol. 28:2905-2920. doi: 10.1093/molbev/msr126.

Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, Goodwin J, Moisan JP, Richard C, Millward A, Demaine AG, Barbujani G, Previderè C, Wilson IJ, Tyler-Smith C, Jobling MA (2010) A predominantly Neolithic origin for European paternal lineages. PLoS Biol 8:e1000285. doi:10.1371/journal.pbio.1000285

Banks WE, d'Errico F, Peterson AT, Vanhaeren M, Kageyama M, Sepulchre P, Ramstein G, Jost A, Lunt D (2008) Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. J Archaeol Sci, 35:481–491.

Barbujani G, Bertorelle G (2001) Genetics and the population history of Europe. Proc. Natl Acad. Sci. USA 98:22–25.

Barbujani G, Pilastro A, De Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: Neolithic demic diffusion vs. Paleolithic colonisation. Am J Phys Anthropol 95:137–154. doi:10.1002/ajpa.1330950203

Barbujani G, Bertorelle G, Capitani G, Scozzari R (1995) Geographical structuring in the mtDNA of Italians. Proc Natl Acad Sci U S A 92: 9171-9175.

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, King RJ, Rootsi S, Marjanovic D, Primorac D, Hadziselimovic R, Vidovic S, Drobnic K, Durmishi N, Torroni A, Santachiara-Benerecetti AS, Underhill PA, Semino O (2009) Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. Eur J Hum Genet 17:820–830. doi:10.1038/ejhg.2008.249

Beleza S, Gusmão L, Lopes A, Alves C, Gomes I, Giouzeli M, Calafell F, Carracedo A, Amorim A (2005) Micro-phylogeographic and demographic history of Portuguese male lineages. Ann Hum Genet 70:181-194.

Benazzi S, Douka K, Fornai C, Bauer CC, Kullmer O, Svoboda J, Pap I, Mallegni F, Bayle P, Coquerelle M, Condemi S, Ronchitelli A, Harvati K, Weber GW (2011) Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature 479:525–528. doi:10.1038/nature10617

Bertoncini S., Ferri G., Busby G., Taglioli L., Alù M., Capelli C., Paoli G. & Tofanelli S (2012) A Y Variant Which Traces the Genetic Heritage of Ligures Tribes. J. Biol. Res., 84: 143-146.

Bertorelle G, Barbujani G (1995). Analysis of DNA diversity by spatial autocorrelation. Genetics 140, 811–819.

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P et al. (1995) Human Mitochondrial DNA Variation and the Origin of Basques. Ann Hum Genet 59: 63-81

Blasi C, Filibeck G, Taglianti AV (2005) Biodiversità e biogeografia. In: Blasi C, Boitani L, La Posta S, Manes F, Marchetti M (Eds.): Stato sulla biodiversità in Italia. Contributo alla strategia nazionale per la biodiversità. Roma: Palombi Editori, pp. 40-56.

Boattini A, Luiselli D, Sazzini M, Useli A, Tagarelli G, Pettener D (2010) Linking Italy and the Balkans. A Y-chromosome perspective from the Arbereshe of Calabria. Ann Hum Biol. 38:59-68. doi: 10.3109/03014460.2010.491837.

Boattini A, Griso C, Pettener D (2011) Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci. 89:161-173. doi: 10.4436/jass.89014.

Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F (2012). General method to unravel ancient population structures through surnames, final validation on Italian data. Hum Biol 84: 235-270.

Bocquet-Appel, J.-P. & Demars, P. Y. 2000 Neanderthal contraction and modern human colonization of Europe. Antiquity 74:544–552.

Bocquet-Appela JP, Najia S, Vander Lindenb M, Kozlowskic JK (2009) Detection of diffusion and contact zones of early farming in Europe from the space-time distribution of 14C dates. J Archaeol Sci. 36:807-820

Bosch E, Calafell F, González-Neira A, Flaiz C, Mateu E, Scheil HG, Huckenbeck W, Efremovska L, Mikerezi I, Xirotiris N, Grasa C, Schmidt H, Comas D (2006) Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann Hum Genet. 70:459-487.

Bowden GR, Balaresque P, King TE, Hansen Z, Lee AC, Pergl-Wilson G et al. (2008). Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. Mol Biol Evol 25: 301–309.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, et al. (2009) Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326: 137–140.

Brisighelli F, Capelli C, Álvarez-Iglesias V, Onofri V, Paoli G, et al. (2009) The Etruscan timeline: A recent Anatolian connection. Eur J Hum Genet 17: 693–696. doi: 10.1038/ejhg.2008.224

Brisighelli F, Alvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo A, et al. (2012) Uniparental Markers of Contemporary Italian Population Reveals Details on Its Pre-Roman Heritage. PLoS ONE 7: e50794.

Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, Bradley DG, Gusmão L, Winney B, Bodmer W, Vennemann M, Coia V, Scarnicci F, Tofanelli S, Vona G, Ploski R, Vecchiotti C, Zemunik T, Rudan I, Karachanak S, Toncheva D, Anagnostou P, Ferri G, Rapone C, Hervig T, Moen T, Wilson JF, Capelli C (2011) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279:884–892. doi:10.1098/rspb.2011.1044

Calò CM, Melis A, Vona G, Piras I (2008) Sardinian population (Italy): a genetic review. International Journal of Modern Anthropology 1:39–64.

Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia' A, Vetrugno G, Tofanelli S, Onofri V, Tagliabracci A, Paoli G, Pascali VL (2007) Y-chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phylogenet Evol 44:228–239. doi:10.1016/j.ympev.2006.11.030

Capelli C, Onofri V, Brisighelli F, Boschi I, Scarnicci F, Masullo M, Ferri G, Tofanelli S, Tagliabracci A, Gusmao L, Amorim A, Gatto F, Kirin M, Merlitti D, Brion M, Verea AB, Romano V, Cali F, Pascali V (2009) Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe. Eur J Hum Genet 17:848–852. doi:10.1038/ejhg.2008.258

Capocasa M, Battaggia C, Anagnostou P, Montinaro F, Boschi I, Ferri G, Alù M, Coia V, Crivellaro F, Destro Bisol G (2013) Detecting genetic isolation in human populations: a study of European language minorities. PLoS One. 8:e56371. doi: 10.1371/journal.pone.0056371.

Capocasa M, Anagnostou P, Bachis V, Battaggia C, Bertoncini S, Biondi G, Boattini B, Boschi I, Brisighelli F, Calò CM, Carta M, Coia V, Corrias L, Crivellaro F, Ferri G, Francalacci P, Franceschi ZA, Luiselli D, Morelli L, Rickards O, Robledo R, Sanna D, Sanna E, Sarno S, Tofanelli S, Vona G, Pettener D and Destro Bisol G (2014) Linguistic, geographic and genetic isolation: a collaborative study on Italian populations. J Anthropol. Sci. doi:10.4436/JASS.92001

Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, Chiarelli B, Dupanloup I, Bertranpetit J, Barbujani G, Bertorelle G (2003) Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci U.S.A. 100: 6593–6597. doi: 10.1073/pnas.1130343100.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) History and geography of human genes. Princeton University Press, Princeton

Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci U S A 99:11008–11013.doi:10.1073/pnas.162158799

Coia V, Capocasa M, Anagnostou P, Pascali V, Scarnicci F, Boschi I, Battaggia C, Crivellaro F, Ferri G, Alù M, Brisighelli F, Busby GB, Capelli C, Maixner F, Cipollini G, Viazzo PP, Zink A, Destro Bisol G (2013) Demographic histories, isolation and social factors as determinants of the genetic structure of alpine linguistic groups. PLoS One. 8:e81704. doi:10.1371/journal.pone.0081704.

Contu D, Morelli L, Santoni F, Foster JW, Francalacci P, Cucca F (2008) Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. PLoS One. 3:e1430. doi:10.1371/journal.pone.0001430.

Cunliffe, B., 2001. The Oxford Illustrated History of PreHistoric Europe. Oxford University Press, Oxford.

Currat M, Excoffier L (2011) Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. Proc Natl Acad Sci U S A 108:15129–15134. doi:10.1073/pnas.1107450108

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglià A, Tofanelli S, Spedini G, Capelli C. (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol. 21:1673-1682.

Destro-Bisol G, Anagnostou P, Batini C, Battaggia C, Bertoncini S, Boattini A, Caciagli L, Caló MC, Capelli C, Capocasa M, Castrí L, Ciani G, Coia V, Corrias L, Crivellaro F, Ghiani ME, Luiselli D, Mela C, Melis A, Montano V, Paoli G, Sanna E, Rufo F, Sazzini M, Taglioli L, Tofanelli S, Useli A, Vona G, Pettener D (2008) Italian isolates today: geographic and linguistic factors shaping human biodiversity. J Anthropol Sci. 86:179-188.

Di Gaetano C, Cerutti N, Crobu F, Robino C, Inturri S, Gino S, Guarrera S, Underhill PA, King RJ, Romano V, Cali F, Gasparini M, Matullo G, Salerno A, Torre C, Piazza A (2009) Differential Greek and northern African migrations to

Sicily are supported by genetic evidence from the Y chromosome. Eur J Hum Genet 17:91–99. doi:10.1038/ejhg.2008.120

Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, Manzini P, Dianzani I, Betti M, Cusi D, Frau F, Barlassina C, Mirabelli D, Magnani C, Glorioso N, Bonassi S, Piazza A, Matullo G (2012) An overview of the genetic structure within the Italian population from genome-wide data. PLoS One. 7:e43759. doi: 10.1371/journal.pone.0043759.

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

Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, Brdicka R, Barbujani G, Papola F, Ciavarella G, Cucci F, Di Stasi L, Gavrila L, Kerimova MG, Kovatchev D, Kozlov AI, Loutradis A, Mandarino V, Mammi' C, Michalodimitrakis EN, Paoli G, Pappa KI, Pedicini G, Terrenato L, Tofanelli S, Malaspina P, Novelletto A (2004) Y chromosomal haplogroup J as a signature of the post-Neolithic colonization of Europe. Hum Genet 115:357–371. doi:10.1007/s00439-004-1168-9

Esko T, Mezzavilla M, Nelis M, Borel C, Debniak T, Jakkula E, Julia A, Karachanak S, Khrunin A, Kisfali P, Krulisova V, Aušrelé Kučinskiené Z, Rehnström K, Traglia M, Nikitina-Zake L, Zimprich F, Antonarakis SE, Estivill X, Glavač D, Gut I, Klovins J, Krawczak M, Kučinskas V, Lathrop M, Macek M, Marsal S, Meitinger T, Melegh B, Limborska S, Lubinski J, Paolotie A, Schreiber S, Toncheva D, Toniolo D, Wichmann HE, Zimprich A, Metspalu M, Gasparini P, Metspalu A, D'Adamo P (2012) Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. Eur J Hum Genet. 21:659-665. doi: 10.1038/ejhg.2012.229.

Fiorini S, Tagarelli G, Boattini A, Luiselli D, Piro A, Tagarelli A, Pettener D (2007) Ethnicity and evolution of the biodemographic structure of Arbereshe and Italian populations of the Pollino area, Southern Italy (1820–1984). Amer Anthropol 109:735–746.

Forsythe G (2005) A Critical History of Early Rome: From Prehistory to the First Punic War. Berkeley: University of California Press.

François O, Currat M, Ray N, Han E, Excoffier L, Novembre J (2010) Principal component analysis under population genetic models of range expansion and admixture. Mol Biol Evol. 27:1257-1268. doi: 10.1093/molbev/msq010.

Fraumene C, Belle EMS, Castrì L, Sanna S, Mancosu G et al. (2006) High resolution analysis and phylogenetic network construction using complete mtDNA sequences in Sardinian genetic isolates. Mol Biol Evol 23: 2101–2111.

Fu Q, Rudan P, Pääbo S, Krause J (2012) Complete mitochondrial genomes reveal Neolithic expansion into Europe. PLoS One 7:e32473. doi:10.1371/journal.pone.0032473

Garrigan D and Hammer MF (2006) Reconstructing human origins in the genomic era. Nat Rev Genet. 7:669-680.

Giunta F (2003). Albanesi in Sicilia. In: Mandalà M. (ed.), Albanesi in Sicilia, Palermo: Mirror.

Haak W, Forster P, Bramanti B, Matsumura S, Brandt G, et al. (2005) Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. Science 310:1016–1018

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Der Sarkissian CS, Brandt G, Schwarz C, Nicklisch N, Dresely V, Fritsch B, Balanovska E, Villems R, Meller H, Alt KW, Cooper A; Members of the Genographic Consortium (2010) Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biol. 8:e1000536. doi: 10.1371/journal.pbio.1000536.

Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD (2010) Fine-scale population structure and the era of next-generation sequencing. Hum Mol Genet. 19:R221-226. doi: 10.1093/hmg/ddq403.

Hervella M, Izagirre N, Alonso S, Fregel R, Alonso A, Cabrera VM, de la Rúa C (2012) Ancient DNA from hunter-gatherer and farmer groups from northern Spain supports a random dispersion model for the Neolithic expansion into Europe. PLoS One 7:e34417. doi:10.1371/journal.pone.0034417

Heutink P, Oostra BA (2002) Gene finding in genetically isolated populations. Hum Mol Genet. 11:2507-2515.

Heyer E, Chaix R, Pavard S, Austerlitz F (2012) Sex-specific demographic behaviours that shape human genomic variation Mol Ecol. 21:597-612. doi: 10.1111/j.1365-294X.2011.05406.x

Higham T, Compton T, Stringer C, Jacobi R, Shapiro B, Trinkaus E, Chandler B, Gröning F, Collins C, Hillson S, O'Higgins P, FitzGerald C, Fagan M (2011) The earliest evidence for anatomically modern humans in northwestern Europe. Nature 479:521–524. doi:10.1038/nature10484

Jeran N, Havas Augustin D, Grahovac B, Kapović M, Metspalu E, Villems R, Rudan P (2009) Mitochondrial DNA heritage of Cres Islanders--example of Croatian genetic outliers. Coll Antropol. 33:1323-1328.

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human evolutionary genetics: origins, peoples and disease. Garland Science, New York

Jobling MA, Hollow E, Hurles ME, Kivisild T, Tyler-Smith C (2013) Human evolutionary genetics. Second edition. Garland Science, New York.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830–838

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn

B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 3:698. doi: 10.1038/ncomms1701.

King TE, Jobling MA (2009). Founders, Drift, and Infidelity: The Relationship between Y Chromosome Diversity and Patrilineal Surnames. Mol Biol Evol 26:1093-1102.

King RJ, Di Cristofaro J, Kouvatsi A, Triantaphyllidis C, Scheidel W, Myres NM, Lin AA, Eissautier A, Mitchell M, Binder D, Semino O, Novelletto A, Underhill PA, Chiaroni J (2011) The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. BMC Evol Biol 11:69. doi:10.1186/1471-2148-11-69

Kozlowski JK (2007) The significance of blade technologies in the period 50–35 ka BP for the Middle Paleolithic – Upper Paleolithic transition in Central and Eastern Europe. In: Mellars P, Boyle K, Bar-Yosef O, Stringer C (Eds.). Rethinking the Human Revolution. Cambridge: McDonald Institute, pp.317-328.

Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Pääbo S (2010) The complete mitochondrial DNA genome of an un known hominin from southern Siberia. Nature 464:894–897. doi:10.1038/nature08976

Kristiansson K, Naukkarinen J, Peltonen L (2008) Isolated populations and complex disease gene identification. Genome Biol. 9:109. doi: 10.1186/gb-2008-9-8-109

Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, Guilaine J, Crubézy E, Ludes B (2011a) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci U S A 108:9788–9791. doi:10.1073/pnas.1100723108

Lacan M, Keyser C, Ricaut FX, Brucato N, Tarrús J, Bosch A, Guilaine J, Crubézy E, Ludes B (2011b) Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc Natl Acad Sci U S A. 108:18255-18259. doi: 10.1073/pnas.1113061108

Laland KN, Odling-Smee J, Myles S (2010) How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet. 11:137-148. doi: 10.1038/nrg2734.

Lambeck, K., Antonioli, F., Purcell, A., Silenzi, S., 2004. Sea-level change along the Italian coast for the past 10,000 yr. Quaternary Science Reviews 23, 1567e1598.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol. 18:1241-1248. doi: 10.1016/j.cub.2008.07.049.

Larmuseau MH, Vanderheyden N, Jacobs M, Coomans M, Larno L, Decorte R (2011) Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant. Forensic Sci Int Genet. 5:95-99. doi: 10.1016/j.fsigen.2010.08.020.

Larmuseau MH, Vanoverbeke J, Gielis G, Vanderheyden N, Larmuseau HF, Decorte R (2012) In the name of the migrant father--analysis of surname origins identifies genetic admixture events undetectable from genealogical records. Heredity 109:90-95. doi: 10.1038/hdy.2012.17.

Larmuseau MH, Van Geystelen A, van Oven M, Decorte R (2013) Genetic genealogy comes of age: perspectives on the use of deep-rooted pedigrees in human population genetics. Am J Phys Anthropol. 150:505-511. doi: 10.1002/ajpa.22233.

López-Parra AM, Gusmão L, Tavares L, Baeza C, Amorim A, Mesa MS, Prata MJ, Arroyo-Pardo E (2009) In search of the pre- and post-neolithic genetic substrates in Iberia: evidence from Y-chromosome in Pyrenean populations. Ann Hum Genet. 7:42-53. doi:10.1111/j.1469-1809.2008.00478.x.

Malmstrom H, Gilbert MT, Thomas MG, Brandstrom M, Stora J, et al. (2009) Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. Curr Biol 19: 1758–1762.

Malyarchuk, B, Grzybowski T, Derenko M, Perkova M, Vanecek T, Lazur J, Gomolcak P, and Tsybovsky I (2008). Mitochondrial DNA phylogeny in Eastern and Western Slavs. Mol. Biol. Evol. 25:1651–1658.

Malyarchuk B, Derenko M, Grzybowski T, Perkova M, Rogalla U, et al. (2010) The Peopling of Europe from the Mitochondrial Haplogroup U5 Perspective. PLoS ONE 5(4): e10285. doi:10.1371/journal.pone.0010285

Manni F, Toupance B, Sabbagh A, Heyer E (2005). New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. Am J Phys Anthropol 126: 214–228.

Mannino MA, Di Salvo R, Schimmenti V, Di Patti C, Incarbona A, Sineo L, Richards MP (2012) Upper Palaeolithic hunter-gatherer subsistence in Mediterranean coastal environments: an isotopic study of the diets of the oldest directly-dated humans from Sicily. J Archaeol Sci 38: 3094–3100. doi: 10.1016/j.jas.2011.07.009.

Marchani EE, Watkins WS, Bulayeva K, Harpending HC, Jorde LB (2008) Culture creates genetic structure in the Caucasus: Autosomal, mitochondrial, and Y-chromosomal variation in Daghestan. BMC Genetics, doi:10.1186/1471-2156-9-47.

Marjanovic D, Fornarino S, Montagna S, Primorac D, Hadziselimovic R, Vidovic S, Pojskic N, Battaglia V, Achilli A, Drobnic K, Andjelinovic S, Torroni A, Santachiara-Benerecetti AS, Semino O (2005). The peopling of modern Bosnia-Herzegovina: Y-chromosome haplogroups. Ann Hum Genet. 69:757-763.

Mellars P (2011) Palaeoanthropology: the earliest modern humans in Europe. Nature 479:483–485. doi:10.1038/479483a

Mithen S (2006) After the ice: a global human history, 20.000 – 5.000 BC. Cambridge: Harvard University Press.

Myres NM, Rootsi S, Lin AA, Järve M, King RJ, Kutuev I, Cabrera VM, Khusnutdinova EK, Pshenichnov A, Yunusbayev B, Balanovsky O, Balanovska E, Rudan P, Baldovic M, Herrera RJ, Chiaroni J, Di Cristofaro J, Villems R, Kivisild T, Underhill PA (2011) A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet 19:95–101. doi:10.1038/ejhg.2010.146

Nasidze I, Stoneking M. (2001) Mitochondrial DNA variation and language replacements in the Caucasus. Proc R Soc Lond B 268:1197-1206

Nasidze I, Quinque D, Dupanloup I, Cordaux R, Kokshunova L, Stoneking M (2005) Genetic evidence for the Mongolian ancestry of Kalmyks. Am J Physical Anthropol. 128:846–854

Nasidze I, Quinque D, Udina I, Kunizheva S, Stoneking M (2007) The Gagauz, a linguistic enclave, are not a genetic isolate. Ann Hum Genet. 71:379-389.

Neel J (1992) Minority populations as genetic isolates: the interpretation of inbreeding results. In: Bittles AH, Roberts DF, editors. Minority Populations: Genetics Demography and Health. London: The MacMillan Press.

Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskácková T, Balascák I, Peltonen L, Jakkula E, Rehnström K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgár N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julià A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A (2009) Genetic structure of Europeans: a view from the North-East. PLoS One. 4:e5472. doi:10.1371/journal.pone.0005472. Epub 2009 May 8.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40, 646–649

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature. 456:98-101. doi: 10.1038/nature07331.

Novembre J, Ramachandran S (2011) Perspectives on human population structure at the cusp of the sequencing era. Annu Rev Genomics Hum Genet 12: 245-274..

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet. 29:20-21.

Onofri V, Alessandrini F, Turchi C, Fraternale B, Buscemi L, et al. (2007) Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. Int J Legal Med 121: 234–237. doi: 10.1007/s00414-007-0153-y

Ostrer H, Skorecki K (2013) The population genetics of the Jewish people. Hum Genet 132: 119-127.

Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, et al. (2009) Human mitochondrial DNA variation in Southern Italy. Ann Hum Biol 36: 785-811.

Pala M, Achilli A, Olivieri A, Kashani BH, Perego UA, Sanna D, Metspalu E, Tambets K, Tamm E, Accetturo M, Carossa V, Lancioni H, Panara F, Zimmermann B, Huber G, Al-Zahery N, Brisighelli F, Woodward SR, Francalacci P, Parson W, Salas A, Behar DM, Villems R, Semino O, Bandelt HJ, Torroni A (2009) Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 84:814–821. doi:10.1016/j.ajhg.2009.05.004

Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A (2009) Genetic markers and population history: Finland revisited. Eur J Hum Genet 17:1336-1346.

Pesando F (2005) L'Italia antica. Culture e forme del popolamento nel I millennio a. C. Roma: Carocci editore. 326 p

Pereira L, Richards M, Goios A, Alonso A, Albarrán C, Garcia O, Behar DM, Gölge M, Hatina J, Al-Gazali L, Bradley DG, Macaulay V, Amorim A (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. 15:19-24

Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, Costa MD, Martins H, Soares P, Behar DM, Richards MB, Macaulay V (2010) Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol. 10:390. doi:10.1186/1471-2148-10-390.

Pessina A, Tinè V (2008) Archeologia del Neolitico. L'Italia tra il Vi e il IV millennio a.C. Roma: Carrocci editore. 375 p

Piazza A, Cappello N, Olivetti E, Rendine S (1988) A genetic history of Italy. Ann. Hum. Genet. 52, 203–213

Pinhasi R, Fort J, Ammerman AJ (2005) Tracing the origin and spread of agriculture in Europe. PLoS Biol 3:e410. doi:10.1371/journal.pbio.0030410

Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J (2012) The genetic history of Europeans. Trends Genet. 28:496-505. doi: 10.1016/j.tig.2012.06.006.

Prat S, Péan SC, Crépin L, Drucker DG, Puaud SJ, Valladas H, Lázničková-Galetová M, van der Plicht J, Yanevich A (2011) The oldest anatomically modern humans from far southeast Europe: direct dating, culture and behavior. PLoS One 6:e20834. doi:10.1371/journal.pone.0020834

Price TD (2000) Europe's First Farmers. Cambridge: Cambridge University Press.

Ralph P, Coop G (2013) The geography of recent genetic ancestry across Europe. PLoS Biol. 11:e1001555. doi: 10.1371/journal.pbio.1001555.

Ramachandran S, Tang H, Gutenkunst R and Bustamante CD (2010) Human Population Genetics and Genomics In: Speicher M, Antonarakis SE, Motulsky AG (Eds.). Vogel and Motulsky's Human Genetics: Problems and Approaches, 4th ed. Berlin: Springer, pp. 589-615. doi:10.1007/978-3-540-37654-5_20

Rasteiro R, Bouttier PA, Sousa VC, Chikhi L (2012) Investigating sex-biased migration during the Neolithic transition in Europe, using an explicit spatial simulation framework. Proc Biol Sci. 279:2409-2416. doi: 10.1098/rspb.2011.2323.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251–1276. doi:10.1016/S0002-9297(07)62954-1

Rickards O, Martinez-Labarga C, Scano G, De Stefano GF, Biondi G, Pacaci M, Walter H. (1998) Genetic history of the population of Sicily. Hum Biol 70:699–714.

Romano V, Cali F, Ragalmuto A, D'Anna RP, Flugy A, De Leo G, Giambalvo O, Lisa A, Fiorani O, Di Gaetano C, Salerno A, Tamouza R, Charron D, Zei G (2003). Autosomal microsatellite and mtDNA genetic analysis in Sicily (Italy). Ann Hum Genet 67:42–53.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barać L, Pericić M, Balanovsky O, Pshenichnov A, Dion D, Grobei M, Zhivotovsky LA, Battaglia V, Achilli A, Al-Zahery N, Parik J, King R, Cinnioğlu C, Khusnutdinova E, Rudan P, Balanovska E, Scheffrahn W, Simonescu M, Brehm A, Goncalves R, Rosa A, Moisan JP, Chaventre A, Ferak V, Füredi S, Oefner PJ, Shen P, Beckman L, Mikerezi I, Terzić R, Primorac D, Cambon-Thomsen A, Krumina A, Torroni A, Underhill PA, Santachiara-Benerecetti AS, Villems R, Semino O (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet 75:128–137.doi:10.1086/422196

Rootsi S, Myres NM, Lin AA, Järve M, King RJ, et al. (2012) Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. Eur J Hum Genet 20: 1275-1282.

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, et al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet 67:1526-1543.

Sampietro ML, Lao O, Caramelli D, Lari M, Pou R, Martí M, Bertranpetit J, Lalueza-Fox C (2007) Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe. Proc Biol Sci 274:2161–2167. doi:10.1098/rspb.2007.0465

Sanchez-Faddeev H, Pijpe J, van der Hulle T, Meij HJ, van der Gaag KJ, Slagboom PE, Westendorp RG, de Knijff P (2013) The influence of clan structure on the genetic variation in a single Ghanaian village. Eur J Hum Genet 21: 1134-1139.

Sazzini M, Sarno S, Luiselli D (2014) The Mediterranean human population: an Anthropological Genetics perspective. In: Goffredo S, Baader H, Dubinsky Z (Eds.). The Mediterranean Sea: Its History and Present Challenges. Berlin: Springer, pp. 529-551. doi:10.1007/978-94-007-6704-1_31

Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Arbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., et al. (2000). The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290, 1155–1159.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB (2010) The archaeogenetics of Europe. Curr Biol 20:174–183. doi:10.1016/j.cub.2009.11.054

Soule M (1973) The epistasis cycle: a theory of marginal populations. Annual Review of Ecology and Systematics, 4:165–187.

Stringer C, Davies W (2001) Archaeology. Those elusive Neanderthals. Nature 413:791–792. doi:10.1038/35101688

Tagarelli A (2004) Studio antropologico della comunità arbëreshe della provincia di Torino. eds: Librare, pp. 47-66.

Tagarelli G, Fiorini S, Piro A, Luiselli D, Tagarelli A, Pettener D (2007) Ethnicity and biodemographic structure in the Arbereshe of the province of Cosenza, southern Italy, in the XIX century. Coll Antropol 31:331–338.

Tassi F, Ghirotto S, Caramelli D, Barbujani G (2013) Genetic evidence does not support an Etruscan origin in Anatolia. Am J Phys Anthropol. 152:11-8. doi: 10.1002/ajpa.22319.

Thomas MG, Weale ME, Jones AL, Richards M, Smith A et al. (2002) Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. Am J Hum Genet 70: 1411-1420.

Toso F (2014) The study of language islands: an interdisciplinary approach. J Anthropol. Sci. doi: 10.4436/JASS.92002.

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumäki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet. 69:844-852.

Tresset A and Vigne JD (2011) Last hunter-gatherers and first farmers of Europe. C R Biol. 334:182-189. doi: 10.1016/j.crvi.2010.12.010.

Underhill PA, Kivisild T (2007) Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 41:539-564.

Van Oven M, Hämmerle JM, van Schoor M, Kushnick G, Pennekamp P, Zega I, Lao O, Brown L, Kennerknecht I, Kayser M (2011) Unexpected island effects at an extreme: reduced Y chromosome and mitochondrial DNA diversity in Nias. Mol Biol Evol. 28:134913-61. doi:10.1093/molbev/msq300.

Veeramah KR, Tönjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, Stumvoll M (2011) Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. Eur J Hum Genet. 19:995-1001. doi:10.1038/ejhg.2011.65.

Varzari A, Kharkov V, Stephan W, Dergachev V, Puzyrev V, Weiss EH, Stepanov V (2009) Searching for the origin of Gagauzes: inferences from Y-chromosome analysis. Am J Hum Biol. 21:326-336. doi: 10.1002/ajhb.20863.

Ward AM, Heichelheim FM, Yeo CA (2009). Roman History: Its Geographic and Human Foundations. In: History of the Roman People. Pearson Education (US).

Whittle A (1996) Europe in the Neolithic. The creation of new worlds. Cambridge: Cambridge University Press.

Wilkins JF (2006) Unraveling male and female histories from human genetic data. Curr Opin Genet Dev. 16:611-617.

Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I et al. (2012). People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. Eur J Hum Genet 20: 203–210.

Zalloua PA, Platt DE, El Sibai M, Khalife J, Makhoul N, Haber M, Xue Y, Izaabel H, Bosch E, Adams SM, Arroyo E, López-Parra AM, Aler M, Picornell A, Ramon M, Jobling MA, Comas D, Bertranpetit J, Wells RS, Tyler-Smith C, Genographic Consortium (2008) Identifying genetic traces of historical expansions: Phoenician footprints in the Mediterranean. J Hum Genet 83:633–642. doi:10.1016/j.ajhg.2008.10.012.

doi 10.4436/JASS.92001
e-pub ahead of print

JASs Reports
Journal of Anthropological Sciences
Vol. 92 (2014), pp. xx-xx

Linguistic, geographic and genetic isolation: a collaborative study of Italian populations

Marco Capocasa[2,3], Paolo Anagnostou[1,2], Valeria Bachis[4], Cinzia Battaggia[1], Stefania Bertoncini[5], Gianfranco Biondi[6], Alessio Boattini[7], Ilaria Boschi[8], Francesca Brisighelli[1,9§], Carla Maria Calò[4], Marilisa Carta[7], Valentina Coia[10,11§], Laura Corrias[4], Federica Crivellaro[12], Sara De Fanti[7], Valentina Dominici[1,2], Gianmarco Ferri[13], Paolo Francalacci[14], Zelda Alice Franceschi[15], Donata Luiselli[7], Laura Morelli[14], Giorgio Paoli[5], Olga Rickards[16], Renato Robledo[17], Daria Sanna[14], Emanuele Sanna[4], Stefania Sarno[7], Luca Sineo[18], Luca Taglioli[5], Giuseppe Tagarelli[19], Sergio Tofanelli[5], Giuseppe Vona[4], Davide Pettener[7] & Giovanni Destro Bisol[1,2]

1) Sapienza Università di Roma, Dipartimento di Biologia Ambientale, Roma, Italy

2) Istituto Italiano di Antropologia, Roma, Italy

3) Sapienza Università di Roma, Dipartimento di Biologia e Biotecnologie "Charles Darwin", Roma, Italy

4) Università di Cagliari, Dipartimento di Scienze della Vita e dell'Ambiente, Cagliari, Italy

5) Università di Pisa, Dipartimento di Biologia, Pisa, Italy

6) Università dell'Aquila, Dipartimento di Scienze Ambientali, L'Aquila, Italy

7) Università di Bologna, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Bologna, Italy

8) Università Cattolica del Sacro Cuore, Istituto di Medicina Legale, Roma, Italy

9) Universidade de Santiago de Compostela, Instituto de Ciencias Forenses, Santiago de Compostela, Spain

10) Università di Trento, Dipartimento di Filosofia, Storia e Beni Culturali, Trento, Italy

11) Accademia Europea di Bolzano, Istituto per le Mummie e l'Iceman, Bolzano, Italy

12) Sezione di Antropologia, Museo Nazionale Preistorico Etnografico "Luigi Pigorini", Roma, Italy

13) Università di Modena e Reggio Emilia, Dipartimento di Medicina Diagnostica, Clinica e di Sanità Pubblica, Modena, Italy

14) Università di Sassari, Dipartimento di Scienze della Natura e del Territorio, Sassari, Italy

15) Università di Bologna, Dipartimento di Discipline Storiche, Antropologiche e Geografiche, Bologna, Italy

16) Università degli Studi di Roma "Tor Vergata", Centro di Antropologia Molecolare per lo studio del DNA Antico, Dipartimento di Biologia, Roma, Italy

17) Università di Cagliari, Dipartimento di Scienze Biomediche, Cagliari, Italy

18) Università di Palermo, Dipartimento di Biologia Ambientale e Biodiversità, Palermo, Italy

19) Istituto di Scienze Neurologiche, CNR, Mangone, Cosenza, Italy

§ Present affiliation

e-mail: marco.capocasa@uniroma1.it; davide.pettener@unibo.it; destrobisol@uniroma1.it

the **JASs** is published by the ***Istituto Italiano di Antropologia*** www.isita-org.com

2 Exploring biodiversity of Italian populations

Summary - *The animal and plant biodiversity of the Italian territory is known to be one of the richest in the Mediterranean basin and Europe as a whole, but does the genetic diversity of extant human populations show a comparable pattern? According to a number of studies, the genetic structure of Italian populations retains the signatures of complex peopling processes which took place from the Paleolithic to modern era. Although the observed patterns highlight a remarkable degree of genetic heterogeneity, they do not, however, take into account an important source of variation. In fact, Italy is home to numerous ethno-linguistic minorities which have yet to be studied systematically. Due to their difference in geographical origin and demographic history, such groups not only signal the cultural and social diversity of our country, but they are also potential contributors to its bio-anthropological heterogeneity. To fill this gap, research groups from four Italian Universities (*Bologna, Cagliari, Pisa *and* Roma Sapienza*) started a collaborative study in 2007, which was funded by the Italian Ministry of Education, University and Research and received partial support by the* Istituto Italiano di Antropologia*. In this paper, we present an account of the results obtained in the course of this initiative. Four case-studies relative to linguistic minorities from the Eastern Alps, Sardinia, Apennines and Southern Italy are first described and discussed, focusing on their micro-evolutionary and anthropological implications. Thereafter, we present the results of a systematic analysis of the relations between linguistic, geographic and genetic isolation. Integrating the data obtained in the course of the long-term study with literature and unpublished results on Italian populations, we show that a combination of linguistic and geographic factors is probably responsible for the presence of the most robust signatures of genetic isolation. Finally, we evaluate the magnitude of the diversity of Italian populations in the European context. The human genetic diversity of our country was found to be greater than observed throughout the continent at short (0-200 km) and intermediate (700-800km) distances, and accounted for most of the highest values of genetic distances observed at all geographic ranges. Interestingly, an important contribution to this pattern comes from the "linguistic islands" (e.g. German speaking groups of* Sappada *and* Luserna *from the Eastern Italian Alps), further proof of the importance of considering social and cultural factors when studying human genetic variation.*

Keywords – *Genetic structure, Linguistic diversity, Minority languages, Linguistic islands.*

Introduction

The plant and animal biodiversity found on Italian territory is among the richest in the Mediterranean basin and in Europe as a whole. This depends on the presence of different biomes, with Alpine tundra and the Mediterranean arid zones at the extremes, whose variety may be seen as a consequence of the natural role of Italy as a bridge connecting central European and North African environments (Blasi *et al.*, 2005). As has happened throughout the continent, biodiversity has been influenced by human activities, such as cattle breeding and the agricultural exploitation of the lands since the Neolithic (Goudie, 2013). However, the remarkable ethno-cultural diversity of human populations which have settled on Italian territory since prehistory has led to different ways of managing natural resources (Padovani *et al.*, 2009). This has probably reduced the loss of biodiversity produced by man-driven modifications of natural ecosystems to some extent. The latter points introduce us to another important aspect, which is more directly concerned with the anthropological dimension, i.e. biodiversity of human populations.

Looking at biodiversity from a holistic perspective, it is worth raising the question: does the great variety observed in Italy for plants and animals hold for human populations? Since prehistory, numerous peopling events have occurred on Italian territory. During the Last

Glacial Maximum, most of the country provided a refugium not only for animal and plant species (e.g. see Taberlet *et al.*, 1998; Petit *et al.*, 2003; Grassi *et al.*, 2009), but also for human groups (Banks *et al.*, 2008). Migrations during the Late Paleolithic, the Neolithic and, even more so, the Metal Ages characterized the complex peopling process of pre-historic Italy, leaving more (e.g. Greeks; see Scheidel, 2003) or less (e.g. Etruscans; see Barker & Rasmussen, 1998; Beekes, 2003) clear signatures. Under Roman rule, there were heavy demographic reshuffles caused by warfare and slavery and this had a significant impact on population composition (Hin, 2013). In more recent times, further contribution to the population heterogeneity came from invasions of the Barbarians (Lombards and Normans among others; see Jörg, 2002; Donald, 2008) and Arabs (Aghlabids; see Metcalfe, 2009). All these events explain at least part of the genetic structure of extant Italian populations, structured in two main geographic clusters (North-West and South-East) for Y-chromosome markers but more homogeneous for mtDNA polymorphisms (Boattini *et al.*, 2013). However, the investigations carried out so far have yet to provide an exhaustive picture of the genetic diversity of Italian populations, since they could not take into account another important component of the human biodiversity of our country. We are referring to the historic ethno-linguistic minorities which total about 5% of the population today. They differ according to area of origin (central and southeastern Europe) and demographic history, and the majority of them settled in their present locations in relatively recent times (from the Middle Ages to the 18th century). The linguistic isolation from neighboring populations and, in many cases, the settlement as small communities in secluded areas make such groups very important when evaluating the entire spectrum of biodiversity of Italian populations. At present, twelve ethno-linguistic minorities are officially safeguarded by Italian legislation: Albanian, Catalan, Croatian, French, Franco-provençal, Friulian, German, Greek, Ladin, Occitan, Slovene and Sardinian (Toso, 2008). Furthermore, other minor linguistic groups are recognized at regional level (e.g. Tabarkian in Sardinia) and ancient linguistic substrates can be still recognized in marginal areas.

The interest in ethnic minorities and local communities is considerable in the present-day international debate concerning the preservation of cultural diversity and its influence on the biodiversity in the globalized world. Since 1999, the United Nations Environment Programme (UNEP) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) have turned their attention to the “indigenous” peoples for the assessment of global biodiversity. UNEP focused on the relationships between biodiversity and cultural diversity with the publication of the report *Cultural and Spiritual Values of Biodiversity* (Addison Posey, 1999). After having defined *Cultural Diversity* as the “common heritage of humanity” (UNESCO, 2001), UNESCO has recently pointed out to the risk of extinction of 2500 languages, among which most of the minority languages spoken in Italy (Moseley, 2010). The message emerging from these initiatives is that paying attention to ethno-linguistic minorities is important not only for their contribution to the overall human biodiversity, but also for their role in maintaining cultural traditions which have an impact on biodiversity. This seems to expand the view, shared by many evolutionary Anthropologists, that biological and cultural diversity should be studied together to understand how the past has shaped the present of human species: “*The subjects are (on the one hand) data, and (on the other) the cultural history surrounding the collection and interpretation of those data*” (Marks, 1995).

In order to achieve a more complete characterization of the genetic structure of Italian populations, four research units of four Italian Universities (Bologna, Cagliari, Pisa and Roma Sapienza) initiated a collaborative research on linguistic minorities in 2007. During the realization of the project, the working group started further collaborations with researchers from other Italian Universities. A first part of work was done in the framework of the project “Isolating the isolates:

geographic and cultural factors of human genetic variation" supported by the Italian Ministry of Education, University and Research (PRIN project 2007 and 2009). In this initiative, our aim was to understand the relationships between cultural and geographical factors and the genetic structure of the Italian isolates (Destro Bisol *et al.*, 2008). The research work continued under the umbrella of the project "Human biodiversity in Italy: micro-evolutionary patterns" (PRIN project 2009-2011). In this second step, we also focused on the role of social structures in shaping the human biodiversity of isolated groups[1]. Further support to the whole initiative was provided by the *Istituto Italiano di Antropologia* (project "Bio-cultural Atlas of Italian populations").

As genetic tools, we used mtDNA and Y chromosome polymorphisms, a choice based on their power to detect microevolutionary events on a wider timescale than bi-parentally transmitted polymorphisms, their relatively low costs, abundance of comparative data both at micro- and macro-geographic scale and the possibility to estimate gender-biased processes. Notably, the outputs of the study are not limited to scientific papers (e.g. Destro Bisol *et al.,* 2008; Robledo *et al.*, 2009; Boattini *et al.*, 2010; Bertoncini *et al.*, 2012; Capocasa *et al.*, 2013b; see Supplementary File S1 for a complete list of papers and communications to congress), but also include the return of scientific results to the investigated communities (e.g. Capocasa *et al.*, 2012, 2013a) and a dedicated genetic online database (http://www.isita-org.com/Anthro-Digit/italianisolates/index.html).

In this paper, we present an account of the results obtained by the research units (RUs) involved in the collaborative study. To this aim, four case studies spanning from the eastern Alps to southern Italy are described and discussed, with a focus on their micro-evolutionary and anthropological implications. Thereafter, we present the results of a systematic analysis of the relations between linguistic, geographic and genetic isolation combining the data obtained in the course of the long-term study with literature and unpublished results. More in particular, we: (i) analyze the genetic structure of numerous Italian populations in order to evaluate the association of linguistic and geographic factors with genetic isolation; (ii) compare the extent of genetic diversity observed in Italy and Europe taking into account both open populations and geographically and/or linguistically isolated groups.

Materials and methods

The dataset

Our original dataset includes data from 873 individuals for mtDNA and 795 individuals for Y chromosome from 19 populations (see Table 1) studied in the course of the PRIN projects. It was integrated with other published and unpublished results, reaching a total of 2875 and 1811 entries for mtDNA and Y chromosome polymorphisms, respectively (see Appendix 1; Congiu *et al.*, 2012).

Buccal swabs were collected in apparently healthy donors (see Tab. 1). A detailed description of laboratory analyses for DNA extraction, amplification, purification and genotyping is reported in Supplementary File S2.

Statistical analyses

Haplotype diversity (HD) and its standard errors were calculated according to Nei 1987. Pairwise differences among all the populations of the datasets were calculated using the genetic distance measure Fst (Reynolds *et al.*, 1983; Slatkin, 1995). Analyses of molecular variance (AMOVA) were performed in order to test genetic differences among Sardinian populations. Demographic descriptive indices (Tajima's D, Fu's Fs and Harpending's raggedness) were calculated to test for the presence of signs of demographic expansion (Tajima, 1989; Harpending, 1994; Fu, 1997). An analysis of haplotype sharing was performed in order to evaluate the specific contribution of private haplotypes to the observed patterns of genetic differentiation among populations.

[1] As a rule, by the term "isolated populations" we mean linguistic and/or geographic isolates.

Tab. 1. Populations analyzed in the course of the collaborative study. Abbreviations: ALT, altitude; FS, founder surnames; G, geographic isolate; GL, geographic/linguistic isolate; GP, grandparents; L, linguistic isolate; O, open population.

						MTDNA		Y CHROMOSOME	
POPULATION	ABB.	STATUS	SAMPLING STRATEGY	CENSUS SIZE*	ALT.°	N	REFERENCE	N	REFERENCE
Arbereshe (Calabria)	ARC	GL	FS	28034	535	87	This study	87	Boattini *et al.*, 2011; Pettener, *pers. com.*
Arbereshe (Sicily)	ARS	L	FS	7875	645	40	This study	42	Pettener, *pers. com.*
Benetutti	BEN	G	FS	2010	406	50	This study	45	Calò *et al.*, 2013
Cadore	CAD	O	GP	10797	984	32	Caglià, *pers. com.*	52	Caglià, *pers. com.*
Carloforte	CFT	GL	FS/GP	6420	10	51	Calò *et al.*, 2012	41	Robledo *et al.*, 2012
Cimbrians (Lessinia)	LES	GL	GP	13455	758	40	Capocasa *et al.*, 2013	29	Coia *et al.*, 2013
Cimbrians (Luserna)	LUS	GL	GP	286	1333	21	Coia *et al.*, 2012	25	Coia *et al.*, 2013
Circello	CIR	G	FS	2501	700	27	This study	34	Tofanelli, *pers. com.*
Cosenza	COS	O	GP	69131	238	42	Pettener, *pers. com.*	28	Pettener, *pers. com.*
Fiemme Valley	FIE	G	GP	18990	1033	41	Coia *et al.*, 2012	41	Coia *et al.*, 2013
Ladins (Fassa Valley)	LVF	GL	GP	9894	1345	47	Coia *et al.*, 2012	47	Coia *et al.*, 2013
Lucca plain	PIL	O	GP	154928	81	50	This study	50	This study
North Sardinia	NSA	O	GP	67253	440	40	This study	47	Calò *et al.*, 2013
Sappada	SAP	GL	GP	1307	1217	59	Capocasa *et al.*, 2013b	36	Coia *et al.*, 2013
Sauris	SAU	GL	GP	429	1212	48	Capocasa *et al.*, 2013b	29	Coia *et al.*, 2013
Sulcis-Iglesiente	SGL	O	GP	128614	96	50	Robledo *et al.*, 2012	46	Robledo *et al.*, 2012
Timau	TIM	GL	GP	500	830	46	Capocasa *et al.*, 2013b	22	Coia *et al.*, 2013
Trapani-Enna	TEN	O	GP	96834	-	80	Pettener, *pers. com.*	71	Pettener, *pers. com.*
Vagli	VAG	G	FS	995	575	22	This study	23	Tofanelli *pers. com.*

* Source: ISTAT (2011) http://demo.istat.it.

° meters above sea level.

All parameters of intra- and inter-population genetic diversity were calculated using the Arlequin 3.5 software (Excoffier & Lischer, 2010). Multidimensional scaling (MDS) was applied to genetic distance matrices to visualize genetic differentiation among populations using the SPSS software (release 16.0.1 for Windows, S.P.S.S. Inc.). In order to evaluate the effect of census size on the intra- and inter-population diversity, we performed a stepwise multiple regression analysis using the current census size and altitude as covariates. Census size was log-transformed in order to linearize its relation to HD and average Fst. Unless otherwise stated, analyses were performed using mtDNA HVR-1 region (from 16033 to 16365 n.p.) and 15 Y chromosome STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and GATA H4.1).

Genetic relationships among Arbereshe, putative source populations (Albania, Greece) and sink populations (Calabria, Sicily) were analysed as follows: populations were first grouped in clusters using a non-hierarchical EM algorithm (mclust R package; see Fraley & Raftery, 2002, 2006), then posterior membership probabilities (for each population and for each cluster) were calculated using Discriminant Analysis of Principal Components (DAPC, adegenet R package; see Jombart *et al.*, 2010) and graphically represented with a bar-plot.

To assess which interval of stepwise mutational differences between the male specific region of Y chromosome (MSY) haplotypes is the most suitable to represent the *Ligures* contribution to Samnium, we used the Equation 31 described by Walsh (2001) implemented in the software ASHEs 1.1 (ashes.codeplex.com, Tofanelli *et al.*, 2011). The Bayesian posterior distribution of expected pairwise mutation differences was calculated for haplotypes separated by 78 generations (2,190 years ago using 28 years per generation), assuming a strict stepwise mutational model, a mutation rate of $3.09\text{x}10^{-3}$ per locus per generation (averaged values from observed single-locus germ-line mutations following Ballantyne *et al.*, 2010) and a lambda value of 0.0002.

Genes, geography and language: a tale of four stories

The collaborative project covered a substantial part of the linguistic and/or geographic isolates settled on Italian territory. What follows is a selection of the results obtained by each of the four RUs, which illustrates the variety of factors to be taken into account when studying human isolated populations and the diversity of approaches needed to carry out these investigations.

The Arbereshe: between Italy and the Balkans

The Arbereshe are one of the largest linguistic minorities settled in Italy (~100,000 individuals; see Fig. 1). They originate from migratory waves from Albania between the end of the 15th and beginning of the 16th century in response to the invasion of the Balkans by the Ottoman Empire. Historians agree that most of the immigrants came from the south of Albania (Toskeria), often passing along the Peloponnese peninsula (Zangari, 1941). This hypothesis is supported by evidence that Arberisht, the language spoken by the Arbereshe, is part of the Tosk dialect group originally spoken in Toskeria. At present, there are 50 Arbereshe communities in Italy. They are located in Southern Italy, where they are separated from each other and interspersed among Italian villages. As such, they constitute an interesting case study allowing us to explore the effects of cultural isolation on genetic variability, as well as their relationships with the source (Southern Balkans) and sink populations (Southern Italy).

The Arbereshe of Calabria are probably the most important and interesting group, due to the high number of villages (30, with a total of ~ 60,000 individuals) and an exceptional preservation of their cultural identity. The latter feature is evident not only in the language (Arberisht), but also by the Greek Orthodox religion and a common mytho-history (Fiorini *et al.*, 2007). The investigated populations are scattered around the Pollino massif, an area that was historically characterised by strong geographic isolation and an economy based on sheep breeding and crop cultivation. During

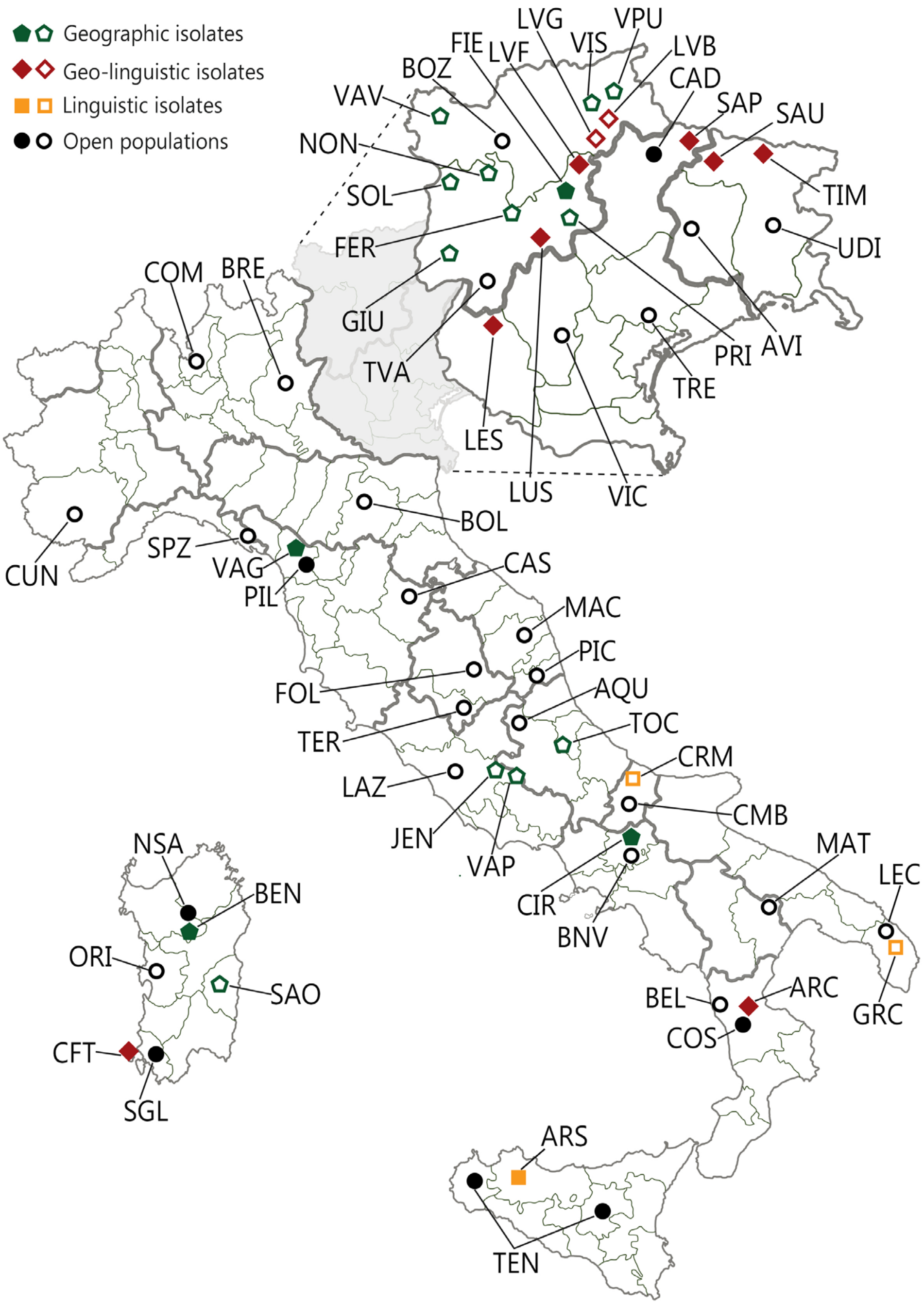


Fig. 1 - Geographic location of the populations analysed in this study. Filled and empty symbols indicate populations studied in the PRIN projects and literature data, respectively.

the second half of the 20th century, the whole area was heavily struck by emigration which led to a partial loss of original cultural traits (breakdown of isolates) (Fiorini *et al.*, 2007).

In Sicily, the Arbereshe are currently found in only three municipalities of the Palermo province (Contessa Entellina, Piana degli Albanesi and Santa Cristina Gela). Characterized by a relatively low population size (~15,000 individuals), they are settled in less geographically isolated areas than Calabrians. Their migration history is particularly complex, involving various intermediate steps (both in the Balkans and in Italy) before their definitive settlement. Furthermore, there is documented evidence of partial re-peopling events from Greece, the most well-known case being that of approximately 100 families coming from the island of Andros in 1520 AD (Giunta, 2003).

Calabrian Arbereshe were sampled in 13 different villages (Acquaformosa, Cerzeto, Civita, Firmo, Frascineto, Lungro, Plataci, S. Basile, S. Cosmo Albanese, S. Demetrio Corone, S. Giorgio Albanese, S. Sofia d'Epiro and Spezzano Albanese). Sicilian Arbereshe were sampled in two villages (Contessa Entellina, Piana degli Albanesi). In both cases, founder surnames were used as inclusion criterion (Boattini *et al.*, 2011) and "control" samples were collected from open neighbouring populations. All the sampled individuals have been unrelated for at least three generations. Data for Albanian and Greek populations were obtained from literature (Bosch *et al.*, 2006).

We first detected the genetic signatures of ethno-linguistic affiliation of the Arbereshe from Calabria by means of biodemographic analyses (Fiorini *et al.*, 2007; Tagarelli *et al.*, 2007). We showed that Arbereshe populations are clearly distinguishable from their Italian neighbours, and observed a correlation between geographic- and surname-based distances between the Arbereshe groups. In addition, Arbereshe groups from the Pollino area showed considerably higher marital isonymy levels than in Italians (0.080 and 0.050, respectively). All these lines of evidence suggest that Arbereshe of Calabria are a group of populations that have undergone cultural and geographical isolation. Importantly, diachronic analyses revealed that such isolation features declined progressively during the last two centuries, and more rapidly in the last 50 years. A Y-chromosome study based on a first sampling of Calabrian Arbereshe (a total of 40 individuals) confirmed and extended these findings (Boattini *et al.*, 2011). Arbereshe were found to provide a signal of discontinuity in the Italian genetic landscape, while showing a strong affinity with some modern Balkan populations (in particular Albanians and Kosovars).

Further insights are provided by taking into account unpublished results for Y-chomosomal and mitochondrial polymorphisms from a larger sampling of Calabrian Arbereshe (104 and 101 individuals, respectively) together with new data on Sicilian Arbereshe (44 and 57 individuals). Haplotype diversity values observed for male and female lineages in Calabrian Arbereshe (0.976 and 0.995, respectively) and Sicilian Arbereshe (0.979 and 0.992) are very close to those of Italian neighbour populations of Cosenza (0.994 and 1.000) and Trapani-Enna (0.985 and 0.999). These results are in line with our previous conclusions that Arbereshe have "conserved much of [their] ancestral genetic diversity along with [their] founder surnames and cultural features" (Fiorini *et al.*, 2007; Boattini *et al.*, 2011) and that their Y-chromosomal gene pool may mirror "the genetic structure of the migrants that came to Italy from the Southern Balkans (Albania) five centuries ago" (Boattini *et al.*, 2011). As shown in Fig. 2, Calabrian Arbereshe cluster (with high membership probabilities) with Albanians both for male and female lineages, as predicted by the isolation pattern inferred from surname analysis and their similarity with the source population. Intriguingly, our results suggest a different and more complex scenario for the Arbereshe from Sicily. Their Y-chromosome clustering pattern supports a significant admixture with Greeks, which agrees with historical information. On the other hand, mtDNA results highlight a substantial similarity to populations from Southern Italy, a likely consequence of admixture processes (and/or sexually imbalanced migration patterns).

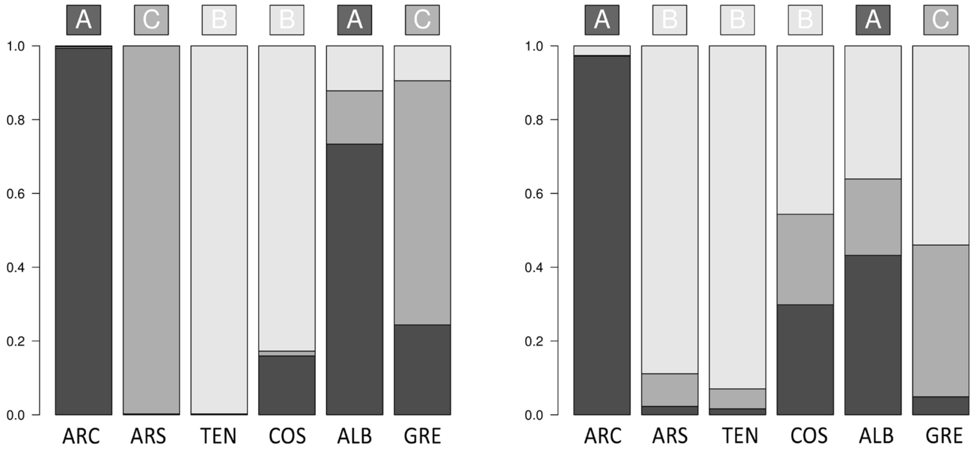


Fig. 2 - Admixture-like barplots for Y-chromosome (left) and mtDNA (right). The barplots represent DAPC-based posterior membership probabilities for each of the considered populations and for each inferred cluster (mclust algorithm). The affiliation of each population to a given cluster and its corresponding colour code are represented by letters (within coloured squares) on the top of each bar. Labels: Arbereshe from Calabria (ARC), Arbereshe from Sicily (ARS), Trapani-Enna (TEN), Cosenza (COS), Albania (ALB), Greece (GRE).

These findings suggest a noticeable difference in the ethnogenesis of Calabrian and Sicilian Arbereshe, with a more evident continuity with source populations for the former and lower impact of isolation for the latter. The study of Albanian-speaking minorities of Southern Italy illustrates, therefore, how the genetic structure of populations sharing the same ethno-linguistic label might have been shaped by diverse microevolutionary histories.

An analysis of isolated populations in Sardinia

Populations from Sardinia provide a paradigmatic example of genetic differentiation in an insular context. The combined effect of small population size, endogamy, and the associated consanguinity, resulting from geographical and cultural barriers, has increased among-group diversity, mainly via long-term genetic drift. A further source of genetic heterogeneity was provided by the different incidence of malaria morbidity. Thus, the unique genetic structure and high level of isolation of some Sardinian villages create opportunities to map genes involved in multifactorial diseases (Zavattari *et al.*, 2000, Angius *et al.*, 2001). However, the interplay between microevolutionary, demographic and cultural factors makes Sardinian populations of great interest also for anthropological studies.

During the collaborative project, the RU of the University of Cagliari focused on the isolates of Carloforte (linguistic) and Benetutti (geographic), in order to assess the degree of isolation and the relative contribution of cultural and geographical barriers (see Fig. 1). Carloforte (6420 inhabitants), located in the small island of San Pietro, off the Southwestern coast of Sardinia, is an alloglot community founded in 1738 AD by Ligurian migrants coming from Tabarka, a Tunisian island (Vallebona, 1974). As a historical imprint, the present-day Carloforte inhabitants speak the ancient dialect of Pegli (Liguria), also referred to as “tabarkino” (a minority language officially recognized by Sardinian regional legislation). Benetutti, a small village of 2010 inhabitants, is located in an area of Northern Sardinia, Goceano, which is characterized by strong geographic isolation. The first historical-demographic documents regarding Benetutti may be found in the “*Quinque Libri*”, dating back to 1618 AD.

By studying Y chromosome and mtDNA polymorphisms, we were able to detect signatures of genetic isolation and of a constant demographic state in both populations (Robledo *et al.*, 2012; Calò *et al.*, 2013; present study). In fact, haplotype diversity observed in Carloforte (0.975 and 0.921 for Y chromosome and mtDNA, respectively) and Benetutti (0.975 and 0.918 for Y chromosome and mtDNA, respectively) are lower than the ones observed in open Sardinian populations (Robledo *et al.*, 2012). In fact, using both founder surnames and grandparent rule as a selection criteria, we were able to select individuals descending from village founders and capture a large extent of within-population genetic diversity at the same time. A similar conclusion was reached in a different study on Y chromosome lineages in isolates located in the Italian Western Alps (Boattini *et al.*, 2010). In any case, the lack of a substantial drop in within-population diversity is consistent with historical records which do not point to any dramatic change in population size. This is also supported by demographic inferences based on mtDNA analysis. In fact, the negative but statistically insignificant values of Tajima's D and Fu's index together with values of Harpending's raggedness suggest that these populations have kept their sizes small and constant over time.

The analysis of inter-population variation proved useful to shed light on other aspects of the genetic history of the two populations under study. Effects of genetic isolation were detected with both uniparental markers, since genetic distances of Benetutti and Carloforte from neighbouring open populations were found to be statistically significant (p<0.05). In the multidimensional scaling (MDS) plot based on Y chromosome data (Fig. 3a), Carloforte lies in a rather isolated position, although not very distant from other Italian populations. Genetic differentiation between the sampled Ligurian population and the founders of Carloforte may account for the former evidence. Benetutti is rather separated from other groups, including Sardinians. The mtDNA based plot (Fig. 3b) confirms both the peculiar behaviour of the Carloforte, which is the only Sardinian population located in the x-positive values of the plot, and the genetic differentiation of Benetutti from other Sardinians. However, differently from what we observed using paternal lineages, we found that Carloforte lies close to African populations, particularly Tunisia. Although no support comes from historical sources, a hypothesis of admixture of Carloforte males with Tunisian females is worth considering due to the occurrence at appreciable frequency (8%) of the mtDNA haplogroup M, characteristic of Northern African and Western Asian populations but so far undetected in Sardinia and Italy (Gonzalez *et al.*, 2007; Fraumene *et al.*, 2003, 2006; Morelli *et al.*, 2000).

An analysis of molecular variance among Sardinian populations was carried out to evaluate the apportionment of genetic diversity. Consistent with previous research work in Sardinia, only a small portion of variation could be attributed to differences between populations (3.78% for mtDNA and 3.31% for Y chromosome; p<0.05). Interestingly, data on preferential male mobility in the Carloforte (Calò & Vona, 1994) and, more in general, the historically documented practice of matrilocal behaviour of some areas of Sardinia (Murru Corriga, 1995) might explain the slightly larger variance observed for maternal lineages.

Genetic and cultural isolation in Northern Apennines

Recent studies suggest an ethno-linguistic continuity between the human groups currently inhabiting the inner valleys of North-West Tuscany (Lunigiana, Garfagnana and orographically linked areas) and populations settled in this area in pre-Roman times. The genetic evidence of this connection was formerly reported by Piazza *et al.* (1988), who associated one of the main principal components obtained from blood protein alleles to the inheritance of those ancient tribes named *Ligures* according to the Latin etymology. This possible link is supported by recent studies of variation in the male-specific portion of the Y chromosome, which have identified North Western Tuscany as an area where lineages

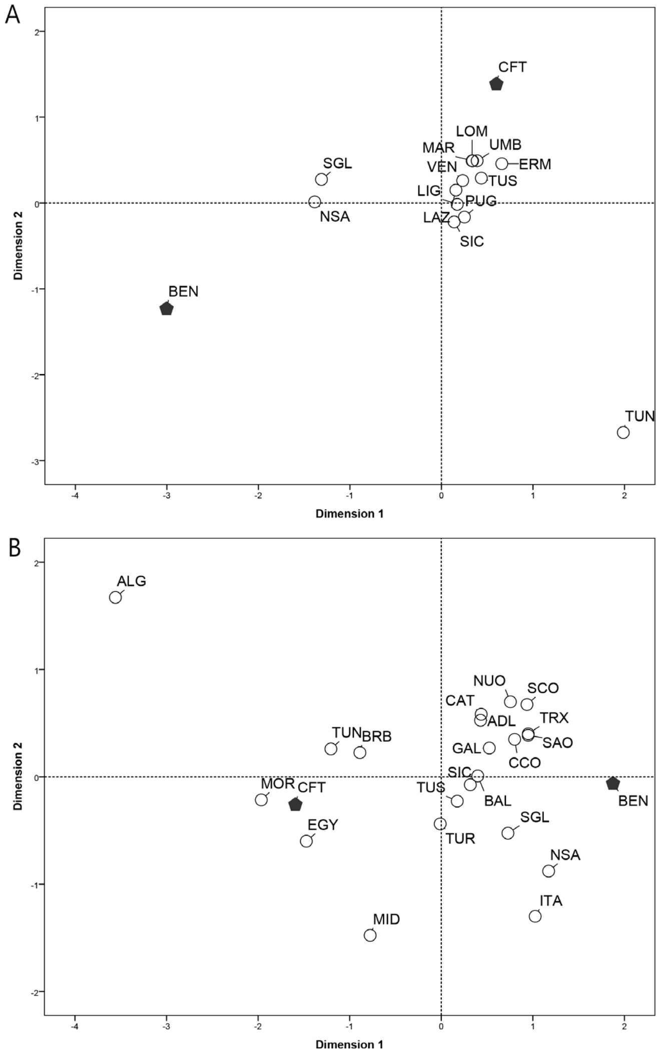


Fig. 3 - Multi-dimensional scaling plots of Fst genetic distances among Sardinian and Mediterranean populations based on (A) 7 Y chromosome STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393; stress value=0.046) and (B) mtDNA HVR-1 sequences (stress value=0.130). Abbreviations and references: (a) Benetutti, BEN; North Sardinia, NSA (Calò et al., 2013). Carloforte, CFT; Sulcis-Iglesiente, SGL (Robledo et al., 2012). Emilia Romagna, ERM (Ferri et al., 2009). Lazio, LAZ; Liguria, LIG; Puglia, PUG; Toscana, TUS; Umbria, UMB (Presciuttini et al., 2001). Lombardia, LOM (Cerri et al., 2005). Marche, MAR (Onofri et al., 2007). Sicilia, SIC (Ghiani et al., 2004); Tunisia, TUN (Frigi et al., 2006). Veneto, VEN (Turrina et al., 2006). (b) Algeria, ALG; Catalogna, CAT (Corte-Real et al., 1996). Andalusia, ADL; Balearic Islands, BAL; Gallura, GAL; Morocco, MOR; Nuorese, NUO; South Corsica, SCO; Trexenta, TRX (Falchi et al., 2006). Benetutti, BEN, North Sardinia, NSA (this study). Carloforte, CFT (Calò et al., 2012). Central Corsica, CCO (Varesi et al., 2000). Egypt, EGY (Krings et al., 1999). Middle East, MID (Di Rienzo & Wilson, 1991). North Italy, ITA (Stenico et al., 1996). Ogliastra, SAO (Fraumene et al., 2006). Sicilia, SIC (Vona et al., 2001). Sulcis Iglesiente, SGL (Robledo et al., 2012). Toscana, TUS (Francalacci et al., 1996). Tunisia, TUN (Plaza et al., 2003). Tunisian Berbers, BRB (Fadhlaoui-Zid et al., 2004). Turchia, TUR (Comas et al., 1996).

with a putative post-Neolithic date of origin persist at relatively high frequencies (Busby *et al.,* 2012, Boattini *et al.,* 2013).

Modern historical sources consider *Ligures Apuani* to be not entirely eradicated by the Roman military campaigns of the II century BC and the following mass deportations to *Ager Taurasinum* (Samnium) of at least 47,000 people (*Titus Livius*, In: Storia Romana XXXVIII). Rather, they depict the fate of natives in the form of confederated tribes surviving in hilly or mountainous areas, until they were eventually absorbed in the demo-territorial institutes later imposed by the Roman authorities (i.e. *res publicae*, *arcifinalis, excepta*), often maintaining a certain degree of autonomy and ownership (Dilke, 1971; Gagliardi, 2006; Marcuccetti, 2012).

From a linguistic point of view, many hamlets remained in the Apuan Alps and the Northern Apennines show similar forms of resistance to the "*tuscanization"* of the spoken language, maintaining a sort of bilingualism exerted through the use of lexemes and phonemes which are traceable to a pre-Latin matrix (Ambrosi, 1956; Giacomelli, 1975; ALT-Web). Moreover, the distribution of many toponyms with pre-Latin roots, in particular those referring to general orographic names such as rivers, mountain reliefs, open places (i.e. *nava, asc, var, dur, mat, penn, aus, kar*), is limited over Europe to conservative areas and have a peak of density in the territory once inhabited by *Ligures* tribes (Marcuccetti, 2008).

The hypothesis of a genetic legacy between *Ligures Apuani* and present Apuan and Samnite isolates (Ligures Legacy Model, LLM) was tested by the RU of the University of Pisa through the genetic characterization of uni-linear markers in a sample of unrelated donors from the communities of Vagli (Province of Lucca) and Circello (Province of Benevento) selected according to founder surnames. Vagli is located in the core of the area which has many archeological records linked to *Ligures Apuani.* Its elder inhabitants still speak a language characterized by a number of linguistic relicts (Ambrosi, 1956; Guazzelli, 2001). Circello lays in close proximity to the remains of Macchia, the town in the Samnium around which the deported *Ligures* (*L. Baebiani*) were forced to settle in 180 BC (Patterson, 2009).

A slight reduction of HD at both mtDNA and Y Chromosome has been observed for Vagli (0.948 and 0.984) and Circello (0.960 and 0.975) relative to the neighbouring populations of Piana di Lucca (0.983 and 0.999) and Benevento (0.989 and 1.000). As expected for communities of Indo-European ancestry, usually practicing prevalence of female *vs* male mobility (patrilocal), genetic distances based on mtDNA are weaker discriminators than distances based on MSY haplotypes (see Supplementary Tab. S1). Preliminary comparative assessments of MSY profiles suggest that the diversity of Apuans might be due to an excess of R-U152 haplotypes, whose diffusion in Italy is thought to coincide with the diffusion of *Ligures* cultural features in the Middle-Late Bronze Age (Bertoncini *et al.,* 2012).

As a whole, the two communities under study (Vagli and Circello) showed a genetic pattern which is compatible with a long history of isolation but also with quite diverging micro-evolutionary histories after contacts implied by the LLM. As a more direct test of a genetic continuum with *Ligures* tribes (Fig. 4), we assessed whether an enrichment of matches compatible with the *Titus Livius* deportation hypothesis is detectable when comparing any MSY haplotype of the local population (Vagli) with haplotypes of both, the putative displaced (Circello) and the open Samnite population (Benevento). The enrichment of LLM-compatible matches in the Vagli-Circello curve totaled about 80% and was extremely statistically significant (Fisher exact test, p<0.0001).

The hypothesis of direct descent of the resident males in the Apuan and in the Circello area from members of *Ligures* tribes, who escaped deportation, cannot be ruled out. Further data from wider samples and haplogroup diagnostic markers, as well as more extensive simulation analyses will help achieve more robust inferences. Nonetheless, our case study shows that even mild geographical and cultural isolation may lead to the preservation of a long genetic thread connecting present populations to ancient layers of

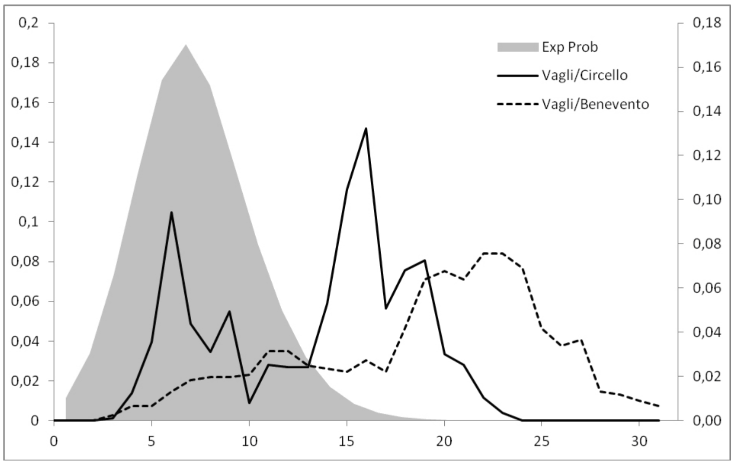


Fig. 4 - Distribution of 17-locus Y-STR haplotype pairwise mutational differences between Vagli and Circello (solid line) and between Vagli and Benevento (dotted line). The shaded area is the Bayesian posterior distribution of pairwise mutation differences compatible with the LLM hypothesis.

pre-Roman Italy. As a corollary, it suggests that many isolated Italian communities other than Northern Apennine ones may escape simplistic classification schemes (i.e. linguistic *vs* geographic isolate), owing to the gradual fringing and recent oblivion of a common ancient cultural identity. Finally, our study highlights the usefulness of accurate non random genetic samplings to uncover genetic layers obscured by recent reshufflings within and among human populations.

Genetic and cultural isolation in the Eastern Italian Alps

Despite the presence of several linguistic islands, the Eastern Alps had not been thoroughly investigated from a bio-anthropological perspective before the start of the collaborative project. The RU from the University of Rome “La Sapienza” focused on six populations: Ladins from the Fassa Valley, Cimbrians from Lessinia and Luserna and the ethno-linguistic “islands” of Sappada, Sauris and Timau (see Fig. 1). Dolomitic Ladins probably originated from a founder group of pre-Indoeuropean speakers that underwent a process of “Latinization” under the pressure of the Roman Empire. At the beginning of the Middle Ages, they fragmented as a result of the arrival of German people. The Ladin community of the Fassa Valley consists of approximately 6,000 individuals settled in seven villages (Toso, 2008). The other populations under study are German-speaking groups thought to be in continuity with peoples migrated from Carinthia, Bavaria and Tyrol in the high and late Middle Ages (Denison, 1971; Rizzi, 1993). After their initial settlement, these ethno-linguistic isolates maintained close social and cultural links (De Concini, 1997; Navarra, 2002). At present, their common background is evident due to their sharing of cultural aspects, such as language and traditions, which persist despite a certain degree of cultural exchanges with the surrounding neo-latin groups. The two Cimbrian settlements have different social and environmental characteristics. Luserna is a small town located over 1,300 meters a.s.l. in the Trentino region, and inhabited by approximately 300 people. Conversely, Lessinia is populated by over 13,000 inhabitants who are distributed in a wider and less “hostile” mountainous territory in the province of Verona in the Veneto region. Sappada is located at an altitude of 1,245 m.a.s.l. in the Veneto region (province of Belluno; 1,307 inhabitants). Sauris and Timau are two small villages (429 and 500 inhabitants, respectively) located in the Carnic Alps. The former is placed at 1,212 m.a.s.l. within the province of Udine, in the Friuli Venezia Giulia region. Timau is situated at 830 m.a.s.l. in the But valley.

The analysis of genetic variation suggests a certain degree of genetic isolation from surrounding populations for Luserna, Sappada, Sauris and Timau. In fact, in accordance with the James Neel's statement that genetic isolates are "derived from a relatively small population sample, which then slowly expand, with very little recruitment from outside the group" (Neel, 1992), we observed low haplotype diversity values and lack of any signal of demographic expansion (Fu's Fs and Harpending's raggedness) (Coia *et al.*, 2012, 2013; Capocasa *et al.*, 2013b). Analyses of genetic distances highlight the differentiation among the four communities and their diversity from other European populations. Furthermore, genetic differentiation among Alpine populations was detected even at individual level using a Bayesian method to cluster multilocus genotypes based on 15 autosomal microsatellites (Montinaro *et al.*, 2012). A different pattern was detected for Ladins from Fassa and Cimbrians from Lessinia, who showed neither an evident reduction of mtDNA and Y chromosome intrapopulation diversity nor a significant departure from the European genetic background.

Although patterns of genetic variation and the lack of signature of demographic expansion observed in Sappada, Sauris and Timau are compatible with what is to be expected in human isolates, it is possible that these results may simply reflect their small initial effective size. As a more direct test of genetic isolation, we performed an analysis of gene flow using a Bayesian approach, applying an Isolation with Migration model (Hey & Nielsen, 2007). Through this approach, we inferred a reduced incoming gene flow for Sappada, Sauris and Timau, both from a wide Central Western European and a neighbouring open Italian-speaking population (Fig. 5). Therefore, our results support the hypothesis that the peculiar linguistic and geographical factors acting on the communities under study might have determined a substantial degree of genetic isolation, so shaping their genetic structure (Capocasa *et al.*, 2013b).

Another aspect which we noticed was the extreme genetic differentiation among Sappada, Sauris and Timau, an unexpected finding considering their relative geographic proximity and closeness in terms of language and traditions (De Concini, 1997; Navarra, 2002). In order to evaluate the magnitude of the genetic differentiation among the three linguistic islands, we first compared them with other European language minorities (Cimbrians and Ladins from Eastern Alps, Aromuns from Albania and Macedonia) whose members are settled in geographically close but distinct locations. Interestingly, Sappada, Sauris and Timau showed inter-population distance values which were two (Ychromosome) or three (mtDNA) times higher than in other groups (Capocasa *et al.*, 2013b). We then wondered whether cultural factors might help explain this pattern. This seems to be worth testing since members of the Eastern Alps linguistic islands do not self-identify as belonging to the same community despite their cultural homogeneity (Steinicke *et al.*, 2011). Such behaviour seems to be in sharp contrast with the sense of belonging of Cimbrians, Ladins and Aromuns and their respective ethnic groups. Therefore, we went on by hypothesizing that this "local ethnicity" may have played a role in marriage strategies, decreasing the genetic exchange among the three linguistic islands. Carrying out coalescent simulations, we inferred that the combined effect of a small initial effective size, as indicated by historical sources (Petris, 1980; Brunettin, 1998; Peratoner, 2002), and a substantial reduction of gene flow, as implied by a local ethnicity model, can explain the pattern observed for Sappada, Sauris and Timau communities (Capocasa *et al.*, 2013b). More in general, our case study shows that complementing classical measures of genetic diversity with Bayesian estimates of gene flow and simulations of micro-evolutionary models may help us better understand patterns of genetic isolation and its relations with demographic and cultural factors in human populations.

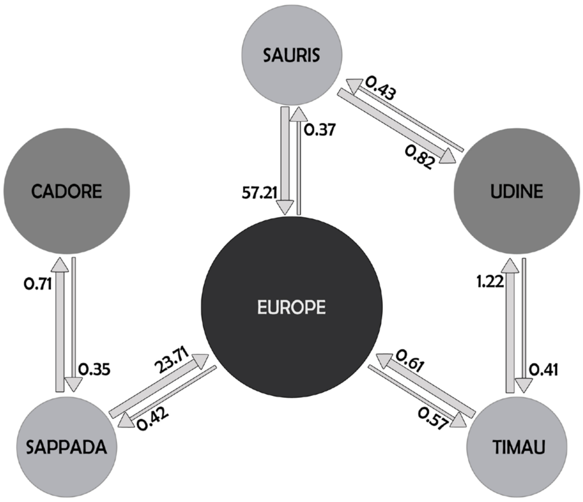


Fig. 5 - Estimates of gene flow among German-speaking Eastern Alps linguistic islands, a neighbouring population and an European metapopulation based on mtDNA HVR-1 region (values reported in effective number of migrating haplotypes).

Linguistic, geographic and genetic isolation in Italian populations

Studies of cultural and geographical factors of genetic isolation have been so far mainly concerned with single or a limited number of groups, focusing on specific populational contexts (e.g. Jeran *et al.*, 2009; van Oven *et al.*, 2011; Veeramah *et al.*, 2011). Combining data relative to numerous isolated groups, produced in the course of this collaborative project, with literature and unpublished results gives us the opportunity to draw inferences on the relations between culture, geography and genetics on a larger geographic scale. To this aim, we built large mtDNA and Y Chromosome databases (see Tab. 1 and Appendix 1), including 57 populations (29 isolates and 28 open groups) and 46 populations (22 isolates and 24 open), respectively. In order to reach the maximum coverage of the Italian territory/populations, we decided to include geographic and/or linguistic isolates studied with different sampling schemes.

As a preliminary step, we compared the genetic structure of isolated and open populations, without considering any further subdivision. We decided to focus on HD and Fst values, since these parameters are little influenced by differences in sample size among those commonly used to detect genetic isolation (Anagnostou *et al.*, in preparation). The HD of isolated populations was found to be significantly lower (Mann-Whitney test p<0.001) than those of open groups, both for mtDNA (from 0.778 to 0.990 vs 0.903-0.995) and Y Chromosome (from 0.590 to 0.998 vs 0.993-1.000) (see Supplementary Tab. S2 and S3; Supplementary Fig. S1). Most of the highest average genetic distances from open populations were observed in geographic and geographic/linguistic isolates, which is another expected effect of genetic isolation (Fig. 6; Supplementary Fig. S2). The difference between open and isolated population groups was slightly more evident for Y chromosome (0.011-0.322 for isolates vs 0.004-0.046 for open populations) than mtDNA (0.000-0.085 vs 0.000-0.013), but highly statistically significant for both polymorphisms (Mann-Whitney test p<0.001). Private haplotypes do not contribute in any significant way to these patterns (Supplementary Fig. S3), a finding which is consistent with the relatively recent age of most isolated populations under study.

On the whole, these results suggest that populations subject to geographic and linguistic isolation are more likely to show lower HD and higher average Fst than open groups. However, they do not clarify whether these differences may be actually regarded as an effect of genetic isolation and the relative weight of geographic and linguistic barriers. To explore these issues, we carried out further analyses in three steps.

First, we used the departure from the pattern of open populations as a criterion to distinguish between simple fluctuations of values and signatures of genetic isolation. To do this, we used the inter-quartile method to identify weak and strong outliers for HD and average Fst distributions. To be considered as bearing a signature of genetic isolation, a given population should display a departure falling in the "weak outlier" interval for both parameters. Furthermore, they were labelled as strong only when both reduction in HD and increase in average Fst exceeded the threshold delimiting the strong outlier area. Using these criteria, we identified 11 populations (out of 29 geographic and/or linguistic isolates) bearing signatures of genetic isolation (10 strong and 1 weak) for mtDNA. Results for Y chromosome do not markedly differ for the number of groups showing signatures of genetic isolation (8 out of 22), but, rather, for the proportion between strong (3) and weak (5) signals (see Fig. 7).

Second, we took into account the possible confounding effect of census sizes in order to validate the obtained inferences. This parameter was preferred to population density, for which no statistically significant correlation with either HD or average Fst was observed (data not shown). To this purpose, we carried out a stepwise multiple regression analysis on all our populations subject to isolation factors using census as a dependent variable. In these calculations, altitude was used as a control variable. This choice was based on two reasons: (i) it may be regarded as one of the ways in which geography may determine isolation between populations, rather than a confounding factor; (ii) we observed a significant correlation between altitude and census size in our dataset (Pearson's ρ=-0.495, p=9.201*10^{-5}; Spearman's ρ=-0.546, p=1.566*10^{-5}), as already reported in previous studies (Cavalli Sforza & Bodmer, 1971; for the Italian context see Franceschi & Paoli, 1994; Morelli *et al.*, 2002; Capocasa *et al.*, in press).

We found a significant correlation between census size and average Fst for both polymorphisms. However, they account only for a minor portion of the total variance (10.5% for mtDNA and 31.8% for Y chromosome). Census size and HD were found to be significantly correlated only for Y chromosome, but again with a weak effect (26.5% of variance explained) (see Supplementary Tab. S4). Therefore, the effect of census size, although detectable, does not seem so marked to undermine our inferences. Altitude was found to have a low and significant, but yet negligible effect only on Y chromosome Fst distribution (10.5% of total variance), implying that other factors are responsible for the observed signals of genetic isolation.

Third and finally, we tried to understand whether the conditions of geographic isolation alone, or combined with linguistic isolation, may have different outcomes on the genetic structure of populations under study. Unfortunately, the presence of only 3 linguistic isolates do not make it possible to perform any test of the effect of linguistic barriers alone. In the mtDNA plot (Fig. 7), signatures of genetic isolation were observed more frequently in geographic/linguistic (6 out of 10; 60%) than in geographic isolates (5 out of 16: 31%). A comparable disproportion was shown by Y chromosome polymorphisms, with 2 geographic (out of 9; 22%) and 5 geographic/linguistic isolates (out of 10; 50%) showing the signal. Limiting the comparison to strong signatures (defined as above), the ratio of positives between geographic/linguistic and geographic isolates does not change substantially either for mtDNA (5 out of 10 *vs* 5 out of 16, respectively) or Y chromosome (3 out of 10 *vs* 0 out of 9). A further indication of the effect of linguistic isolation on the genetic structure of populations is provided by the demographic inferences based on mtDNA data.

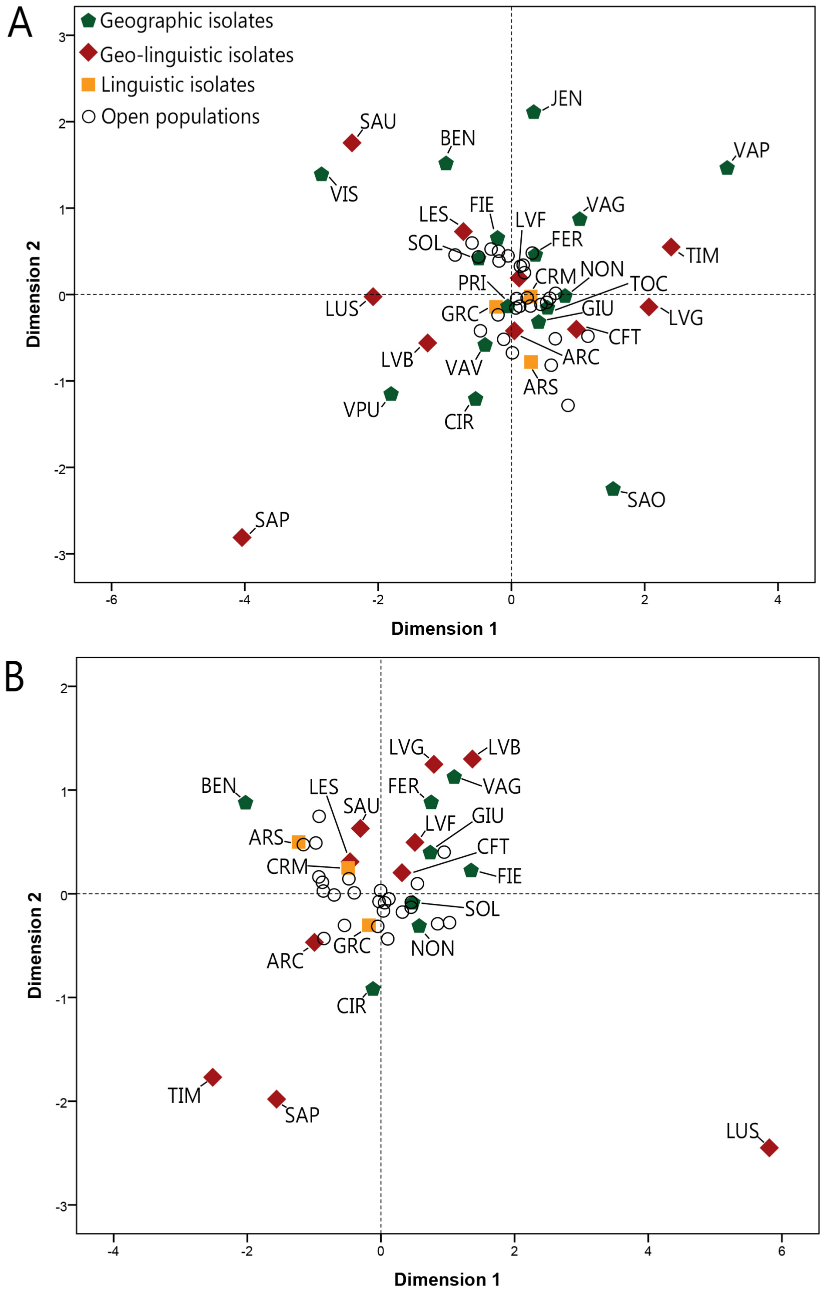


Fig. 6 - Multi-Dimensional Scaling plot of Fst genetic distances among Italian populations based on (A) mtDNA HVR-1 sequences (stress value=0.236) and (B) 15 Y Chromosome STRs (stress value=0.156). Population abbreviations as in Table 1.

In fact, four of the 6 groups which showed a lack of signals of demographic expansion (non significant Fu's and high raggedness values) belong to the geographic /linguistic isolates (Cimbrians from Luserna, Sappada, Sauris, Timau), whereas only 2 are geographic isolates (Vallepietra and Val Pusteria) (Supplementary Tab. S2).

What do all these findings tell us? We admit that no simple cause-effect relationship between linguistic barriers and genetic structure can be inferred when studying populations with different ethnogenesis and demographic histories. Nonetheless, it seems clearly evident that the combination of linguistic and geographic isolation factors is frequently associated with the presence of robust signatures of genetic isolation and a condition of slow, if any, demographic expansion, fitting the classical definition of genetic isolate given by James Neel in 1992. This is not trivial since, to date, evidence of a joint effect of geographic and linguistic isolation factors in Italian populations has been provided by biodemographic investigations only (e.g. Fiorini *et al.*, 2007; Boattini *et al.*, 2011). It is also worth noting that the extreme reduction of genetic heterogeneity observed in some geographic isolates in which language has acted as an additional barrier to gene flow makes such groups of particular usefulness for genetic studies of complex diseases (see Heutink & Oostra, 2002).

Re-assessing genetic diversity of Italian populations

Another opportunity offered by our collaborative study was to produce a more complete picture of the diversity of Italian populations, combining data from large populations, which are presumably exempt from isolation factors, with the results obtained for ethno-linguistic minorities and geographical isolates. A contextualization into the European background seemed to us to represent a convenient approach in order to evaluate the magnitude of genetic heterogeneity, a necessary step to understand whether human genetic variation parallels plant and animal biodiversity (see introduction). Also in this case, we started by assembling *ad hoc* mtDNA and Y chromosome datasets, in which we included populations settled throughout the continent (Supplementary Tab. S5 and S6).

Both Italian and European open populations are found in the middle of the MDS plots of genetic distances, with a more tight clustering observed for mtDNA (Fig. 8). Although strong signals of genetic differentiation for both maternal and paternal lineages were detected for some European isolates (e.g. Aromuns from Stip and Basques), the most evident outliers are represented by some Italian isolates (Sappada and Luserna). Looking at mtDNA only, the genetic distinctiveness of the Isarco Valley, Ogliastra and Vallepietra is also noticeable. To obtain a quantitative assessment of the diversity among Italian and European populations, we carried out an AMOVA using an extended dataset (see appendices 2 and 3). Including geographically and/or isolated groups in the analyses, we observed increased estimates of diversity among Italians (mtDNA from 0.38% to 1.89%; Y chromosome from 3.19 to 6.89%) which were higher than for Europeans (mtDNA from 0.33% to 1.52%; Y chromosome 8.95% from to 10.60%). It seems, therefore, that inclusion of isolates in the dataset makes the extent of diversity of Italian populations slightly higher than Europeans for mtDNA and reduces the negative difference observed for Y chromosome. However, given the substantial difference in the ratio between open and isolated groups and spatial dispersion between Italy and Europe, we thought it useful to compare genetic and geographic distances using box plots of quartile distributions (Fig. 9). In the case of mtDNA, genetic distances of Italian isolates between each other and with Italian open populations are characterized by wider distributions and lower median values than Europeans for 7 and 8 ranges of geographic distances (out of 10), respectively. However, a high level of differentiation among Italian populations is signalled by the outlying genetic distances observed at practically all geographic ranges. Although this finding may be again related to the availability of a larger number of data on Italian compared to European isolates, it should be noted that large genetic distances

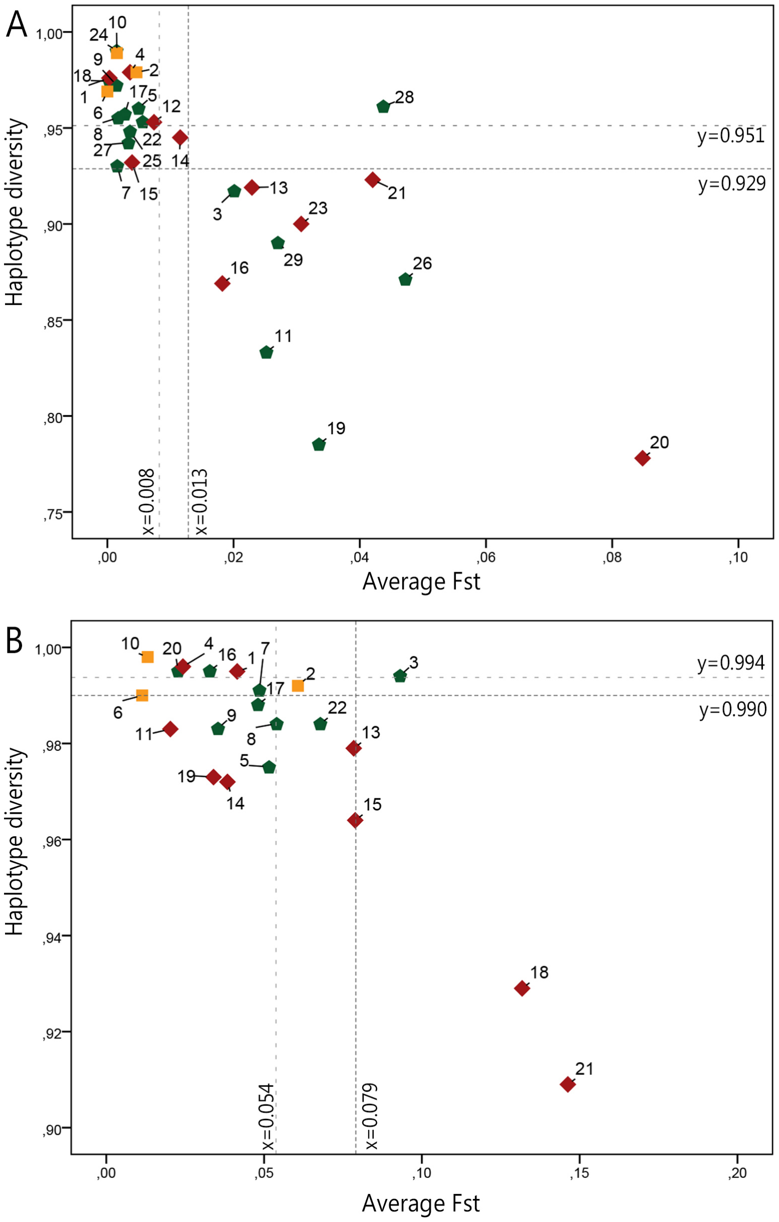


Fig. 7 - Scatter plot of haplotype diversity and average Fst of isolated populations for (A) mtDNA and (B) Y chromosome (LUS not shown HD=0.590; average Fst= 0.322). Dotted lines represent the boundaries defined by the analysis of outliers (interquartile method). Symbols and colors as in Fig. 6. MtDNA labels: 1=ARC, 2=ARS, 3=BEN, 4=CFT, 5=CIR, 6=CRM, 7=FER, 8=FIE, 9=GIU, 10=GRC, 11=JEN, 12=LES, 13=LUS, 14=LVB, 15=LVF, 16=LVG, 17=NON, 18=PRI, 19=SAO, 20=SAP, 21=SAU, 22=SOL, 23=TIM, 24=TOC, 25=VAG, 26=VAP, 27=VAV, 28=VIS, 29=VPU. Y chromosome labels: 1=ARC, 2=ARS, 3=BEN, 4=CFT, 5=CIR, 6=CRM, 7=FER, 8=FIE, 9=GIU, 10=GRC, 11=LES, 12=LUS, 13=LVB, 14=LVF, 15=LVG, 16=NON, 17=PRI, 18=SAP, 19=SAU, 20=SOL, 21=TIM, 22=VAG.

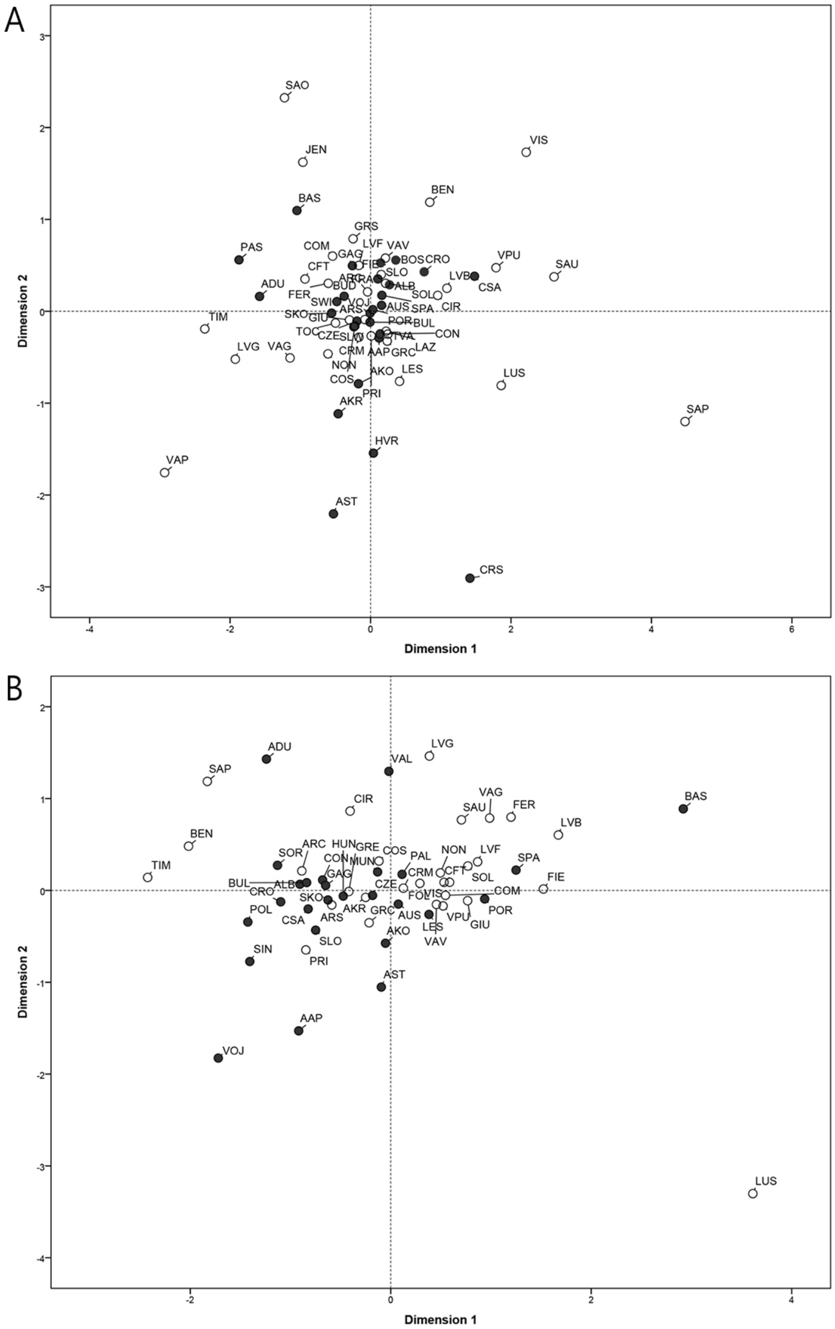


Fig. 8 - Multi-Dimensional Scaling plot of Fst genetic distances among Italian and European populations based on (A) mtDNA HVR-1 sequences (stress value=0.226) and (B) 5 Y chromosome STRs (stress value=0.183). Filled and empty dots represent European and Italian populations, respectively. Population abbreviations as in Supplementary Table S3 and S4.

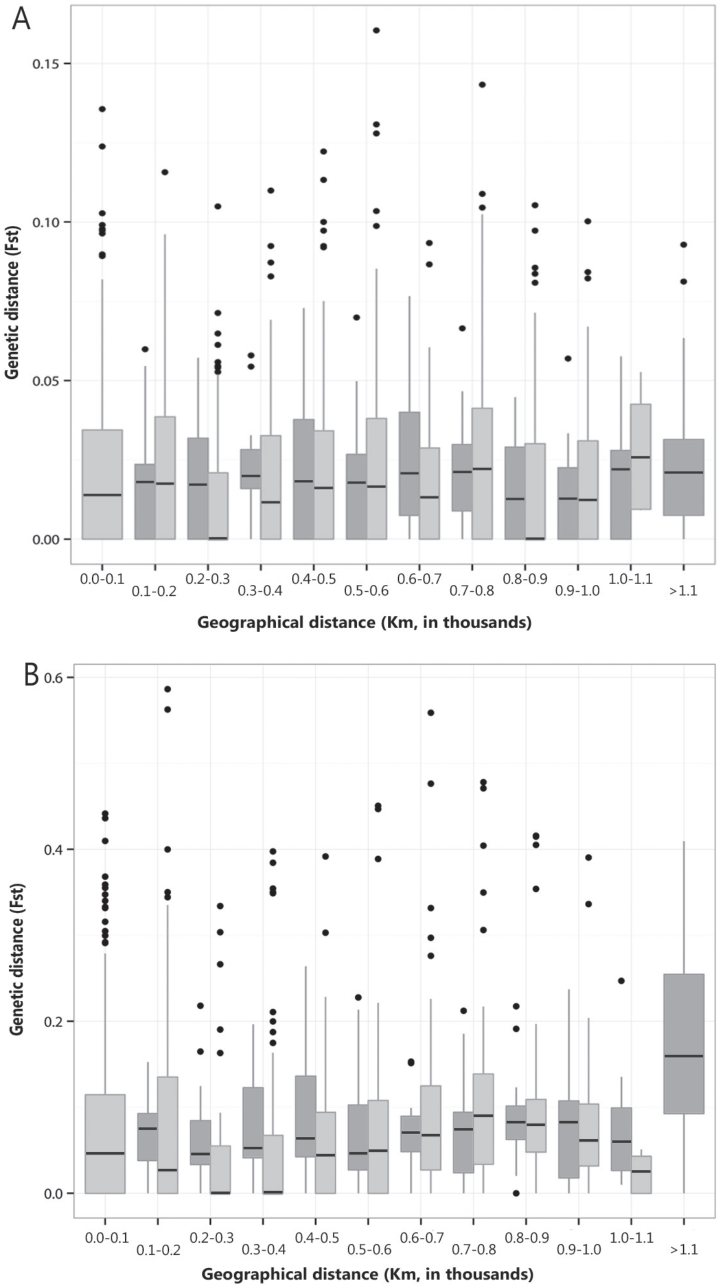


Fig. 9 - Box plot analyses performed on Fst values within different ranges of geographic distances (expressed as thousands of kilometers) for mtDNA (A) and Y chromosome (B). The light gray boxes describe the distribution of all pairwise genetic distances between Italian isolates and those between Italian isolates and open Italian populations; the dark gray boxes indicate the distribution of pairwise genetic distances between European isolates and those between European isolates and open European populations.

between Italian groups are observed even at short geographic distances (<200km). As predicted by previous analyses (Fig. 7), an important contribution to this finding comes from ethno-linguistic minorities of the Eastern Alps. Another range of geographic distances in which the signal of heterogeneity of Italian populations is more marked is between 700 and 800 km. Here, a substantial effect comes from the genetic distances of Sardinian isolates (Benetutti, Carloforte and Ogliastra). Results obtained using Y chromosome data point to a closer correlation between genetic and geographical distances (r^2 value of 0.226, $p<0.001$, for European populations), which is in line with previous observations at a wider geographic scale (Rosser *et al.*, 2000). As observed for mtDNA, a greater differentiation is shown by Italian populations in the ranges between 0-200 and 700-800 km, with a higher frequency of extreme values for all intervals of geographic distances.

Concluding remarks

Our study moves from the assumption that human biological variation is an integral part of what we mean by the term biodiversity. This concept may seem self-evident. However, it is contradicted by many examples in which biodiversity is thought to be necessarily linked to conservation, rather than a value *per se*, while human factors are considered only in relation to anthropic effects or educational aspects (e.g. Johns, 2009; Marton-Lefèvre, 2010). Not less importantly, the study of human biodiversity is often perceived as a difficult, if not risky, task, given the long tradition of misuse of this concept which still survives in bio-medical literature (e.g. Burchard *et al.*, 2003).

Once shortcomings and prejudices have been removed, the question of whether human genetic diversity observed in a certain territory shows similarities with other layers of biodiversity becomes worth answering. Certainly, comparing observations on a large number of species studied at various ecological scales using different approaches (e.g. morphological and genetic) with results from populations of a single species could appear to be simplistic. Obviously, our survey is intended to be just a first step towards further studies driven by explicit hypotheses regarding the impact of environmental and demographic factors on different layers of biodiversity. We also maintain that studying human populations in the framework of biodiversity has an intrinsic value, since it may help us gain insights into the impact of socio-cultural factors on human genetic variation and on biological diversity as a whole.

The results of this collaborative study may be summarized in two main points. We could shed light on the genetic structure of most of the Italian ethno-linguistic minorities, showing that a combination of linguistic and geographic factors is probably responsible for the presence of the most robust signatures of genetic isolation. Furthermore, drawing a picture of Italian populations which also includes linguistic and/or geographic isolates allowed us to better appreciate the noticeable extent of the genetic diversity of our country. It turned out to be slightly greater than that of Europeans for maternal lineages on the whole, and at specific ranges of geographic distances for both genetic markers. Therefore, our survey provides a first affirmative response to the question of whether the genetic diversity of human populations shows a pattern which is comparable to what has been observed in our country for plant and (the rest of) animal biodiversity. The longitudinal extension and the nature of the natural bridge between Central Europe and the Mediterranean seems to provide an explanation also to the high level of inter-population variation. In fact, some of the isolates which determine the noticeable diversity we observed in our study are located along the boundaries of Italian territory. Furthermore, the small demographic size of some of these groups has probably increased their departure from the surrounding genetic background. From an anthropological point of view, the finding that an important contribution to the genetic diversity of Italy comes from the so called "linguistic islands" (e.g. German speaking groups of Sappada and Luserna from the Eastern Italian Alps) is of

Info on the web

http://www.isita-org.com/Anthro-Digit/Italian_isolates_data/index.html
An online repository of genetic data regarding linguistic and/or geographic isolates from the Italian territory.

http://www.isita-org.com/Anthro-Digit/Italian_isolates_pictures/index.html
A photogallery of places and peoples.

http://ec.europa.eu/languages/languages-of-europe/regional-and-minority-languages_en.htm
Information about projects regarding European minority languages supported by the European Commission.

http://it.wikipedia.org/wiki/Minoranze_linguistiche_d'Italia
A useful starting point for a web search on Italian minority languages (in Italian only).

http://www.isolelinguistiche.it/home.page
Web site of the unitary committee of the historical German linguistic islands (in Italian and German, with a good link page).

particular significance. In fact, this is a further proof of how taking into account social and cultural factors may help us understand the structure of human genetic variation more in depth.

Obviously, our study should be seen as a first step towards a more complete assessment. Extending the sampling to the few linguistic isolates that have yet to be included in the survey (e.g. Walser, Occitan and Franco-Provençal linguistic minorities from the Western Alps) could offer an even more detailed picture. Not less importantly, some limitations of unilinearly transmitted polymorphisms (e.g. their nature of single loci and the possible confounding effect of small sample sizes) might be overcome by increasing their resolution or, even better, exploiting the power of large panels of SNPs. Our research network has already planned to deepen the analysis of the genetic structure of isolated populations using a genome-wide approach implemented in the new genotyping array "GenoChip" (Elhaik *et al.*, 2013). Enforcing the interdisciplinary collaborations already initiated in the course of this study (e.g. Fiorini *et al.*, 2007; Robledo *et al.* 2012; Coia *et al.*, 2013) will be a key-step to really take advantage of these new genomic data.

Author Contributions

Designed the research: MCP CMC VC GDB DP GP GV ST.
Collected the samples: PA AB CB GB SB CMC LC MCP VC FC GDB ZAF DL LM GP OR RR ES LS SS GT ST GV.
Performed the experiments: MA CB FB IB SB VB VC MCP MCR SDF GF PF SS.
Analyzed the data: AB PA CMC MCP VD FM LT ST.
Wrote the paper: PA AB CMC MCP GDB DP ST GV.
Read and approved the manuscript: All authors.

Acknowledgements

This paper is dedicated to our friend and colleague Laura Morelli. We would like to thank: the medical and nursing staff of the Azienda Provinciale per i Servizi Sanitari, Trento; the volunteers of the Associazione Volontari Italiani Sangue del Trentino (AVIS), Biblioteca Intercomunale di Primiero, Istitut Cultural Ladin and Kulturinstitut Lusern; all the people who kindly contributed to the collection of *samples in Trentino (*http://laboratoriobagolini.it*); Marcella Benedetti (Municipality of Sappada), Nino Pacilè and Lucia Protto (Municipality of Sauris), Vito Massalongo (Giazza), Ottaviano Matiz and Velia Plozner (Timau) for their warm assistance in sample collection and hospitality; all the people from Benetutti, Circello, Carloforte and Vagli who agreed to participate to this study; Silvano Zaccone (Consorzio "Il Cigno"), Sauro Mattei, Alfonso Tatavitto and Massimo Del Grosso for their fundamental help in collecting samples in Samnite and Apuan areas; Lorenzo Marcuccetti for his commentary to Ligures' history and linguistics; Michele Musacchia (Azienda Sanitaria Locale 4, Corleone) and Giorgio Ales (MD, Piana degli Albanesi) for their help in collecting Sicilian Arbereshe.*

*This research was made possible by grants from the Italian Ministry of Education, University and Research (PRIN projects 2007 and 2009, Principal Investigator DP). Research work in the Eastern Alps was further supported by the Istituto Italiano di Antropologia (project "*Atlante bioculturale Italiano*"; funds to GDB), the University of Rome "La Sapienza" (project "L'isolamento genetico in popolazioni europee", funds to GDB) and the Provincia Autonoma di Trento (Project "Biostre"; funds to VC).*

References

Addison Posey D. 1999. *Cultural and Spiritual Values of Biodiversity*. Intermediate Technology Publications, London.

Ambrosi A.C. 1956. Gli attuali limiti dell'area fonetica cacuminale nelle Alpi Apuane. *Giornale Storico della Lunigiana e del territorio lucense*, 7: 5-25.

Angius A., Melis P.M., Morelli L., Petretto E., Casu G., Maestrale G.B., Fraumene C., Bebbere D., Forabosco P. & Pirastu M. 2001. Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. *Hum. Genet.*, 109: 198-209.

Ballantyne K.N., Goedbloed M., Fang R., Schaap O., Lao O., Wollstein A., Choi Y., van Duijn K., Vermeulen M., Brauer S., Decorte R., Poetsch M., von Wurmb-Schwark N., de Knijff P., Labuda D., Vézina H., Knoblauch H., Lessig R., Roewer L., Ploski R., Dobosz T., Henke L., Henke J., Furtado M.R. & Kayser M. 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.*, 87: 341-353.

Banks W.E., d'Errico F., Peterson A.T., Vanhaeren M., Kageyama M., Sepulchre P., Ramstein G., Jost A. & Lunt D. 2008. Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. *J. Archaeol. Sci.*, 35: 481–491.

Barker G. & Rasmussen T. 1998. *The Etruscans*. Blackwell, Oxford.

Beekes R.S.P. 2003. *The origins of Etruscans*. KNAW, Amsterdam.

Bertoncini S., Ferri G., Busby G., Taglioli L., Alù M., Capelli C., Paoli G. & Tofanelli S. 2012. A Y Variant Which Traces the Genetic Heritage of Ligures Tribes. *J. Biol. Res.*, 84: 143-146.

Blasi C., Filibeck G. & Taglianti A.V. 2005. Biodiversità e biogeografia. In C. Blasi, L. Boitani, S. La Posta, F. Manes & M. Marchetti (eds.): *Stato sulla biodiversità in* Italia. Contributo alla strategia nazionale per la biodiversità, pp. 40-56. Palombi Editori, Roma.

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

Boattini A., Luiselli D., Sazzini M., Useli A., Tagarelli G & Pettener D. 2011. Linking Italy and the Balkans. A Y-chromosome perspective from the Arbereshe of Calabria. *Ann. Hum. Biol.*, 38: 59-68.

Boattini A., Pedrosi M.E., Luiselli D. & Pettener D. 2010. Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). *Ann. Hum. Biol.*, 37: 604–609.

Bosch E., Calafell F., Gonzalez-Neira A., Flaiz C., Mateu E., Scheil H.G., Huckenbeck W., Efremovska L., Mikerezi I., Xirotiris N., Grasa C., Schmidt H. & Comas D. 2006. Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann. Hum. Genet.*, 70: 459–487.

Brunettin G. 1998. L'insediamento di Sauris tra storiografia e rappresentazione di un'origine. In D. Cozzi, D. Isabella, E. Navarra (eds.): *Sauris, Zahre, una comunita' delle Alpi carniche*, pp. 43-61. Forum Editrice Universitaria Udinese, Udine.

Burchard E.G., Ziv E., Coyle N., Gomez S.L., Tang H., Karter A.J., Mountain J.L., Pérez-Stable E.J., Sheppard D. & Risch N. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.*, 348: 1170–1175.

Busby G.B., Brisighelli F., Sánchez-Diz P., Ramos-Luis E., Martinez-Cadenas C., Thomas M.G., Bradley D.G., Gusmão L., Winney B., Bodmer W., Vennemann M., Coia V., Scarnicci F., Tofanelli S., Vona G., Ploski R., Vecchiotti C., Zemunik T., Rudan I., Karachanak S., Toncheva D., Anagnostou P., Ferri G., Rapone C., Hervig T., Moen T., Wilson J.F. & Capelli C. 2012. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc. Roy. Soc. B. - Biol. Sci.*, 279: 884-892.

Calò C.M., Corrias L., Vona G., Bachis V. & Robledo R. 2012. Sampling strategies in a linguistic isolate: results from mtDNA analysis. *Am. J. Hum. Biol.*, 24: 192-194.

Calò C.M., Corrias L., Bachis V., Vona G., Brandas A., Scudiero C.M., Di Fede C., Mameli A. & Robledo R. 2013. Analisi di due isolati della Sardegna (Italia) attraverso lo studio dei polimorfismi del cromosoma Y. *Antropo*, 29: 1-7.

Calò C.M., Vona G. 1994. Carloforte: evoluzione della struttura matrimoniale. *Rendiconti Seminario Facoltà Università Cagliari*, 64: 73-87.

Capocasa M., Anagnostou P., Battaggia C. & Destro Bisol G. 2012. Il patrimonio genetico dei Cimbri della Lessinia, uno studio interdisciplinare. *Cimbri/Tzimbar*, 47: 99-107.

Capocasa M., Anagnostou P., Battaggia C. & Destro Bisol G. 2013a. Omogeneità culturale e peculiarità genetiche della comunità alpina di Timau. *Asou Geats*, 72: 6-7.

Capocasa M., Battaggia C., Anagnostou P., Montinaro F., Boschi I., Ferri G., Alu M., Coia V., Crivellaro F. & Destro Bisol G. 2013b. Detecting genetic isolation in human populations: a study of European language minorities. *PLoS One*, 8: e56371.

Capocasa M., Taglioli L., Anagnostou P., Paoli G. & Danubio M.E. Determinants of marital behaviour in five Apennine communities of Central Italy inferred by surname analysis, repeated pairs and kinship estimates. *Homo* (in press), doi: 10.1016/j.jchb.2013.08.001.

Cavalli Sforza L.L. & Bodmer W.F. 1971. *The Genetics of Human Populations*. Freeman, San Francisco.

Cerri N., Verzeletti A., Bandera B. & De Ferrari F. 2005. Population data for 12 Y-chromosome STRs in a sample from Brescia (northern Italy). *Forensic. Sci. Int.*, 152: 83-87.

Coia V., Boschi I., Trombetta F., Cavulli F., Montinaro F., Destro Bisol G., Grimaldi S. & Pedrotti A. 2012. Evidence of high genetic variation among linguistically diverse populations on a micro-geographic scale: a case study of the Italian Alps. *J. Hum. Genet.*, 57: 254-260.

Coia V., Capocasa M., Anagnostou P., Pascali V., Scarnicci F., Boschi I., Battaggia C., Crivellaro F., Ferri G., Alù M., Brisighelli F., Busby G.B.J., Capelli C., Maixner F., Cipollini G., Viazzo P.P., Zink A. & Destro Bisol G. 2013. Demographic histories, isolation and social factors as determinants of the genetic structure of Alpine linguistic groups. *PLoS One*, 8: e81704.

Comas D., Calafell F., Mateu E., Perez-Lezaun A. & Bertranpetit J. 1996. Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol. Biol. Evol.*, 13: 1067-1077.

Congiu A., Anagnostou P., Milia N., Capocasa M., Montinaro F. & Destro Bisol G. 2012. Online databases for mtDNA and Y chromosome polymorphisms in human populations. *J. Anthropol. Sci.*, 90: 201-215.

Côrte-Real H.B., Macaulay V.A., Richards M.B., Hariti G., Issad M.S., Cambon-Thomsen A., Papiha S., Bertranpetit J. & Sykes B.C. 1996. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.*, 60: 331-350.

De Concini W. 1997. *Gli altri delle Alpi. Minoranze linguistiche dell'arco alpino italiano*. Comune di Pergine Valsugana, Trento.

Denison N. 1971. Some observations on language variety and plurilingualism. In E. Ardener (ed.): *Social anthropology and language*, pp. 157-183. Tavistock, London.

Destro Bisol G., Anagnostou P., Batini C., Battaggia C., Bertoncini S., Bottini A., Caciagli L., Caló C.M., Capelli C., Capocasa M., Castrì L., Ciani G., Coia V., Corrias L., Crivellaro F., Ghiani M.E., Luiselli D., Mela C., Melis A., Montano V., Paoli G., Sanna E., Rufo F., Sazzini M., Taglioli L., Tofanelli S., Useli A., Vona G. & Pettener D. 2008. Italian isolates today: geographic and linguistic factors shaping human biodiversity. *J. Anthropol. Sci.*, 86: 179-188.

Di Rienzo A. & Wilson A.C. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA*, 88: 1597-1601.

Dilke O.A.W. 1971. *The Roman Land Survivors: an introduction to the agrimensores*. David & Charles, Newton Abbott.

Donald M. 2008. *I Normanni in Italia*. Laterza, Roma-Bari.

Elhaik E., Greenspan E., Staats S., Krahn T., Tyler-Smith C., Xue Y., Tofanelli S., Francalacci P., Cucca F., Pagani L., Jin L., Li H., Schurr T.G., Greenspan B., Spencer Wells R. & Genographic Consortium. The GenoChip: a new tool for genetic anthropology. *Genome Biol. Evol.*, 5: 1021-31.

Excoffier L. & Lischer H.E.L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, 10: 564-567.

Fadhlaoui-Zid K., Plaza S., Calafell F., Ben Amor M., Comas D. & Bennamar Elgaaied A. 2004. Mitochondrial DNA heterogeneity in Tunisian Berbers. *Ann. Hum. Genet.*, 68: 222-233.

Falchi A., Giovannoni L., Calò C.M., Piras S.I., Moral P., Paoli G., Vona G. & Varesi L. 2006. Genetic history of some western Mediterranean human isolates through mtDNA HVRI polymorphisms. *J. Hum. Genet.*, 51: 9-14.

Ferri G., Alù M., Corradini B., Radheshi E. & Beduschi G. 2009. Slow and fast evolving markers typing in Modena males (North Italy). *Forensic Sci. Int. Genet.*, 3: e31-e33.

Fiorini S., Tagarelli G., Boattini A., Luiselli D., Piro A., Tagarelli A. & Pettener D. 2007. Ethnicity and evolution of the biodemographic structure of Arbereshe and Italian populations of the Pollino area, Southern Italy (1820–1984). *Amer. Anthropol.*, 109: 735–746.

Fraley C. & Raftery A.E. 2002. Model-based Clustering, Discriminant Analysis and Density Estimation. *J. Amer. Statist. Assoc.*, 97: 611-631.

Fraley C. & Raftery A.E. 2006. MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical Report No. 504, Department of Statistics, University of Washington (revised 2009).

Francalacci P., Bertranpetit J., Calafell F. & Underhill P.A. 1996. Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am. J. Phys. Anthropol.*, 100: 443-460.

Franceschi, M.G. & Paoli, G. 1994. Isolation factors and kinship by isonymy in a group of parishes in northern Tuscany (Italy): influence of within-parish similarity level on between-parish similarity pattern. *Hum. Biol.*, 66: 905–916.

Fraumene C., Belle E.M., Castrì L., Sanna S., Mancosu G., Cosso M., Marras F., Barbujani G., Pirastu M. & Angius A. 2006. High resolution analysis and phylogenetic network construction using complete mtDNA sequences in sardinian genetic isolates. *Mol. Biol. Evol.*, 23: 2101–2111.

Fraumene C., Petretto E., Angius A. & Pirastu M. 2003. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. *Hum. Genet.*, 114: 1–10.

Frigi S., Pereira P., Pereira L., Yacoubi B., Gusmao L., Alves C., Khodjet el Khil H.L., Cherni A., Amorim A. & El Gaaied A. 2006. Data for Y-chromosome haplotypes defined by 17 STRs

(AmpFLSTR1 YfilerTM) in two Tunisian Berber communities. *Forensic Sci. Int.*, 160: 80–83.

Fu Y.X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. *Genetics*, 147: 915-925.

Gagliardi L. 2006. *Mobilità e integrazione delle persone nei centri cittadini, Aspetti giuridici. I. La classificazione degli incolae*. Giuffrè Editore, Milano.

Ghiani M.E., Piras I.S., Mitchell R.J. & Vona G. 2004. Y-chromosome 10 locus short tandem reapeat haplotypes in a population sample from Sicily Italy. *Leg. Med. (Tokyo)*, 6: 89-96.

Giacomelli G. 1975. Aree lessicali toscane. *La ricerca dialettale*, 1: 115-152.

Giunta F. 2003. Albanesi in Sicilia. In: M. Mandalà (ed.): *Albanesi in Sicilia*, p. 25. Mirror, Palermo.

Gonzalez A.M., Larruga J.M., Abu-Amero K.K., Shi Y., Pestano J. & Cabrera V.M. 2007. Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics*, 8: 223.

Goudie A.S. 2013. *The human impact on the natural environment: past, present, and future*. Wiley-Blackwell, Oxford.

Grassi F., De Mattia F., Zecca G., Sala F. & Labra M. 2008. Historical isolation and Quaternary range expansion of divergent lineages in wild grapevine. *Biol. J. Linn. Soc.*, 95: 611–619.

Guazzelli F. 2001. *Suddivisione dialettale della Garfagnana*. Verlag, Köln.

Harpending R.C. 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum. Biol.*, 66: 591-600.

Heutink P. & Oostra B.A. 2002. Gene finding in genetically isolated populations. *Hum. Mol. Genet.*, 11: 2507-2515.

Hey J. & Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *P. Natl Acad. Sci. USA*, 104: 2785-2790.

Hin S. 2013. *The Demography of Roman Italy: Population Dynamics in an Ancient Conquest Society 201 BCE-14 CE*. Cambridge University Press, Cambridge.

Jeran N., Havaš Auguštin D., Grahovac B., Kapović M., Metspalu E., Villems R. & Rudan P. 2009. Mitochondrial DNA heritage of Cres islanders-Example of Croatian genetic outliers. *Coll. Antropol.*, 33: 1323-1328.

Johns D. 2009. The International Year of Biodiversity–From Talk to Action. *Conserv. Biol.*, 24: 338–340.

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

Jörg J. 2002. *Storia dei longobardi*. Einaudi, Milano.

Krings M., Salem A.E., Bauer K., Geisert H., Malek A.K., Chaix L., Simon C., Welsby D., Di Rienzo A., Utermann G., Sajantila A., Paabo S. & Stoneking M. 1999. mtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration?. *Am. J. Hum. Genet.*, 64: 1166-1176.

Livius Titus. *Storia Romana XXXVIII*.

Marcuccetti L. 2008. *La lingua dimenticata*. Luna Editore, La Spezia.

Marcuccetti L. 2012. *La fondazione di Luca e Luna*. Centro Stampa Photos, La Spezia.

Marks J. 1995. *Human biodiversity. Genes, race, and history*. Transaction Publishers, New Brunswick.

Marton-Lefèvre J. 2010. Biodiversity is our life. *Science*, 313: 1179.

Metcalfe A. 2009. *The Muslims of medieval Italy*. Edinburgh University Press, Edinburgh.

Montinaro F., Boschi I., Trombetta F., Merigioli S., Anagnostou P., Battaggia C., Capocasa M., Crivellaro F., Destro-Bisol G. & Coia V. 2012. Using forensic microsatellites to decipher the genetic structure of linguistic and geographic isolates: a survey in the eastern Italian Alps. *Forensic Sci. Int. Genet.*, 6: 827-833.

Morelli L., Grosso M.G., Vona G., Varesi L., Torroni A. & Francalacci P. 2000. Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. *Hum. Biol.*, 72: 585-595.

Morelli L., Paoli G. & Francalacci P. 2002. Surname analysis of the Corsican population reveals an agreement with geographical and linguistic structure. *J. Biosoc. Sci.*, 34: 289–301.

Moseley C. 2010. *Atlas of the World's Languages in Danger*. UNESCO, Paris.

Murru Corriga G. 1995. Il cognome della madre. Tendenze matrilineari nella parentela in Sardegna (XVI-XVIII sec.). *Antropologia Contemporanea*, 18: 47-66.

Navarra E. 2002. Comportamenti demografici e organizzazione socio-economica in due comunità germanofone delle Alpi orientali: Sappada e Sauris (sec. XVIII e XIX). In A. Fornasin & A. Zannini (eds.): *Uomini e comunità delle montagne*, pp. 113-132. Forum Editrice Universitaria Udinese, Udine.

Neel J. 1992. Minority populations as genetic isolates: the interpretation of inbreeding results. In A.H. Bittles & D.F. Roberts (eds.): *Minority Populations: Genetics, Demography and Health*, pp. 1-13. The MacMillan Press, London.

Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Onofri V., Alessandrini F., Turchi C., Fraternale B., Buscemi L., Pesaresi M. & Tagliabracci A. 2007. Y-chromosome genetic structure in sub-Apennine populations ofCentral Italy by SNP and STR analysis. *Int. J. Legal Med.*, 121: 234–237.

Padovani L.M., Carrabba P., Di Giovanni B. & Mauro F. 2009. *Biodiversità. Risorse per lo sviluppo*. ENEA, Roma.

Patterson J. 2009. *Samnites, Ligurians and Romans*. Edizioni Comune di Circello, Benevento.

Peratoner A. 2002. *Sappada/Plodn. Storia, etnografia e ambiente naturale*. Tiziano Editore, Pieve di Cadore.

Petit R.J., Aguinagalde I., de Beaulieu J.L., Bittkau C., Brewer S, Cheddadi R., Ennos R., Fineschi S., Grivet D., Lascoux M., Mohanty A., Müller-Starck G., Demesure-Musch B., Palmé A., Martin J.P., Rendell S. & Vendramin G.G. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, 300: 1563–1565.

Petris B. 1980. *Tischlbong Tamau Timau*. Del Bianco, Udine.

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

Plaza S., Calafell F., Helal A., Bouzerna N., Lefranc G., Bertranpetit J. & Comas D. 2003. Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann. Hum. Genet.*, 67: 312-328.

Presciuttini S., Caglià A., Alù M., Asmundo A., Buscemi L., Caenazzo L., Carnevali E., Carrà E., De Battisti Z., De Stefano F., Dominici R., Piccinini A., Resta N., Ricci U. & Pascali V.L. 2001. Y-chromosome haplotypes in Italy: the GEFI collaborative database. *Forensic Sci. Int.*, 122: 184-188.

Reynolds J., Weir B.S. & Cockerham C.C. 1983. Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105: 67-779.

Rizzi E. 1993. *Storia dei Walser*. Fondazione E. Monti, Anzola d'Ossola.

Robledo R., Corrias L., Bachis V., Puddu N., Mameli A., Vona G. & Calò C.M. 2012. Analysis of a Genetic Isolate: The Case of Carloforte (Italy). *Hum. Biol.*, 86: 735-754.

Robledo R., Piras I., Beggs W. & Calò C. 2009. Analysis of 31 STR loci in the genetic isolate of Carloforte (Sardinia, Italy). *Genet. Mol. Biol.*, 32: 462–465.

Rosser Z.H., Zerjal T., Hurles M.E., Adojaan M., Alavantic D., Amorim A., Amos W., Armenteros M., Arroyo E., Barbujani G., Beckman G., Beckman L., Bertranpetit J., Bosch E., Bradley D.G., Brede G., Cooper G., Côrte-Real H.B., de Knijff P., Decorte R., Dubrova Y.E., Evgrafov O., Gilissen A., Glisic S., Gölge M., Hill E.W., Jeziorowska A., Kalaydjieva L., Kayser M., Kivisild T., Kravchenko S.A., Krumina A., Kucinskas V., Lavinha J., Livshits L.A., Malaspina P., Maria S., McElreavey K., Meitinger T.A., Mikelsaar A.V., Mitchell R.J., Nafa K., Nicholson J., Nørby S., Pandya A., Parik J., Patsalis P.C., Pereira L., Peterlin B., Pielberg G., Prata M.J., Previderé C., Roewer L., Rootsi S., Rubinsztein D.C., Saillard J., Santos F.R., Stefanescu G., Sykes B.C., Tolun

A., Villems R., Tyler-Smith C. & Jobling M.A. 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.*, 67: 1526-1543.

Scheidel M. 2003. The Greek Demographic Expansion: Models and Comparisons. *J. Hellenic Stud.*, 123: 120-140.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139: 457-462.

Steinicke E., Walder J., Löffler R. & Beismann M. 2011. Autochtonous linguistic minorities in the Italian Alps: new legislation – new identifications – new demographic processes. *Journal of Alpine Research*, 99: 2.

Stenico M., Nigro L., Bertorelle G., Calafell F., Capitanio M., Corrain C. & Barbujani G. 1996. High mitochondrial sequence diversity in linguistic isolates of the Alps. *Am. J. Hum. Genet.*, 59: 1363-1375.

Taberlet P., Fumagalli L., Wust-Saucy A.G. & Cosson J.F. 1998. Comparative phylogeography and postglacial colonization routes in Europe. *Mol. Ecol.*, 7: 453–464.

Tagarelli G., Fiorini S., Piro A., Luiselli D., Tagarelli A. & Pettener D. 2007. Ethnicity and biodemographic structure in the Arbereshe of the province of Cosenza, southern Italy, in the XIX century. *Coll. Antropol.*, 31: 331–338.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123: 585-595.

Tofanelli S., Taglioli L., Merlitti D. & Paoli G. 2011. Tools which simulate the evolution of uni-parentally transmitted elements of the human genome. J. Anthropol. Sci., 89: 201-219.

Toso F. 2008. Le minoranze linguistiche in Italia. Il Mulino, Bologna.

Turrina S., Atzei R. & De Leo D. 2006. Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay. *Int. J. Legal Med.*, 120: 56–59.

UNESCO 2001. *Universal Declaration on Cultural Diversity*. UNESCO, Paris.

Vallebona G. 1974. *Storia di una Colonizzazione*. Edizioni Della Torre, Cagliari.

van Oven M., Hämmerle J.M., van Schoor M., Kushnick G., Pennekamp P., Zega I., Lao O., Brown L., Kennerknecht I. & Kayser M. 2011. Unexpected island effects at an extreme: reduced Y chromosome and mitochondrial DNA diversity in Nias. *Mol. Biol. Evol.*, 28: 1349-1361.

Varesi L., Memmi M., Cristofari M.C., Mameli G.E., Calò C.M. & Vona G. 2000. Mitochondrial control region sequence variation in Corsican population (France). *Am. J. Hum. Biol.*, 12: 339-351.

Veeramah K.R., Tönjes A., Kovacs P., Gross A., Wegmann D., Geary P., Gasperikova D., Klimes I., Scholz M., Novembre J. & Stumvoll M. 2011. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur. J. Hum. Genet.*, 19: 995-1001.

Vona G., Ghiani M.E., Calò C.M., Memmi M. & Varesi L. 2001. Mitochondrial DNA sequence analysis in Sicily. *Am. J. Hum Biol.*, 13: 576-589.

Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics*, 158: 897-912.

Zangari D. 1941. *Le colonie italo-albanesi di Calabria*. Editore Casella, Napoli.

Zavattari P., Deidda E., Whalen M., Lampis R., Mulargia A., Loddo M., Eaves I., Mastio G., Todd J.A. & Cucca F. 2000. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum. Mol. Genet.*, 9: 2947-2957.

Associate Editor, Rita Vargiu

Appendix 1 - Other unpublished and literature data used in this study. Abbreviations: Alt, altitude; G, geographic isolate; GL, geo-linguistic isolate; L, linguistic isolate; O, open population.

					mtDNA		Y CHROMOSOME	
POPULATION	**ABB.**	**STATUS**	**CENSUS***	**ALT**	**N**	**REFERENCE**	**N**	**REFERENCE**
Adige Valley	TVA	O	166394	199	56	Coia *et al.*, 2012	56	Coia *et al.*, 2013
Ascoli Piceno	PIC	O	49892	154	53	Brisighelli *et al.*, 2012	38	Brisighelli *et al.*, 2012
Aviano	AVI	O	9270	159	29	Boattini *et al.*, 2013	-	-
Belvedere	BEL	O	9388	150	-	-	27	Brisighelli *et al.*, 2012
Benevento	BNV	O	61700	207	36	Boattini *et al.*, 2013	33	Boattini *et al.*, 2013
Bologna	BOL	O	373020	54	100	Bini *et al.*, 2003	-	-
Bolzano	BOZ	O	105774	262	59	Thomas *et al.*, 2008	-	-
Brescia	BRE	O	188602	149	40	Boattini *et al.*, 2013	35	Boattini *et al.*, 2013
Campobasso	CMB	O	48479	701	37	Boattini *et al.*, 2013	29	Boattini *et al.*, 2013
Casentino	CAS	O	46039	465	122	Achilli *et al.*, 2007	-	-
Como	COM	O	83132	201	39	Boattini *et al.*, 2013	41	Boattini *et al.*, 2013
Croatian Molise	CRM	L	1884	493	41	Babalini *et al.*, 2005	15	This study
Cuneo	CUN	O	55627	534	40	Boattini *et al.*, 2013	30	Boattini *et al.*, 2013
Fersina Valley	FER	G	51478	1066	25	Coia *et al.*, 2012	26	Coia *et al.*, 2013
Foligno	FOL	O	58363	234	-	-	29	Boattini *et al.*, 2013
Giudicarie Valley	GIU	G	36282	680	52	Coia *et al.*, 2012	51	Coia *et al.*, 2013
Grecanici Salento	GRC	L	10000	80	47	Brisighelli *et al.*, 2012	46	Brisighelli *et al.*, 2012
Isarco Valley	VIS	G	44500	801	34	Pichler *et al.*, 2006	-	-
Jenne	JEN	G	407	834	103	Messina *et al.*, 2010	-	-
La Spezia	SPZ	O	95378	3	50	Brisighelli *et al.*, 2012	45	Brisighelli *et al.*, 2012
Ladins (Badia Valley)	LVB	GL	10644	1341	56	Thomas *et al.*, 2008	44	Coia *et al.*, 2013
Ladins (Gardena Valley)	LVG	GL	10198	1320	46	Thomas *et al.*, 2008	51	Coia *et al.*, 2013
L'Aquila	AQU	O	110268	714	25	Boattini *et al.*, 2013	27	Boattini *et al.*, 2013
Lazio	LAZ	O	5536292	-	52	Babalini *et al.*, 2005	-	-
Lecce	LEC	O	89839	49	39	Boattini *et al.*, 2013	35	Boattini *et al.*, 2013

Appendix 1 - Continued.

					mtDNA		Y CHROMOSOME	
POPULATION	**ABB.**	**STATUS**	**CENSUS***	**ALT**	**N**	**REFERENCE**	**N**	**REFERENCE**
Macerata	MAC	O	43019	315	39	Boattini *et al.*, 2013	34	Boattini *et al.*, 2013
Matera	MAT	O	59973	401	36	Boattini *et al.*, 2013	24	Boattini *et al.*, 2013
Non Valley	NON	G	37832	816	48	Coia *et al.*, 2012	48	Coia *et al.*, 2013
Ogliastra	SAO	G	57959	392	175	Fraumene *et al.*, 2003	-	-
Oristano	ORI	O	32156	5	39	Boattini *et al.*, 2013	40	Boattini *et al.*, 2013
Primiero Valley	PRI	G	9959	763	40	Coia *et al.*, 2012	41	Coia *et al.*, 2013
Pusteria Valley	VPU	G	73000	997	37	Pichler *et al.*, 2006	-	-
Sole Valley	SOL	G	15235	884	63	Coia *et al.*, 2012	65	Coia *et al.*, 2013
Terni	TER	O	109482	130	31	Boattini *et al.*, 2013	-	-
Tocco da Casauria	TOC	G	2782	356	50	Verginelli *et al.*, 2003	-	-
Treviso	TRE	O	82535	15	39	Boattini *et al.*, 2013	28	Boattini *et al.*, 2013
Udine	UDI	O	100514	113	51	Brisighelli *et al.*, 2012	45	Brisighelli *et al.*, 2012
Vallepietra	VAP	G	308	825	21	Messina *et al.*, 2010	-	-
Venosta Valley	VAV	G	34307	1024	112	Pichler *et al.*, 2006	-	-
Vicenza	VIC	O	113639	39	40	Boattini *et al.*, 2013	33	Boattini *et al.*, 2013

* Source: ISTAT (2011) (http://demo.istat.it)