Università degli Studi di Ferrara

Ph.D. course
in
Evolutionary biology and ecology

in cooperation with Università degli studi di Parma

CYCLE XXIX

DIRECTOR Prof. Guido Barbujani

Patterns of genetic and linguistic variation. A study of uniparental markers.

Scientific/Disciplinary Sector (SDS) BIO/18

Candidate
Dott. Susca Roberta Rosa

Roberta Rosa Susca
(signature)

Supervisor
Prof. Barbujani Guido

Guido Barbujani
(signature)

Years 2014/2016

Il tuo indirizzo e-mail

roberta.susca@gmail.com

Oggetto:

Dichiarazione di conformità della tesi di Dottorato

Io sottoscritto Dott. (Cognome e Nome)

Susca Roberta Rosa

Nato a:

Putignano

Provincia:

Bari

Il giorno:

28 Settembre 1987

Avendo frequentato il Dottorato di Ricerca in:

Biologia Evoluzionistica ed Ecologia

Ciclo di Dottorato

29

Titolo della tesi:

Patterns of genetic and linguistic variation. A study of uniparental markers.

Titolo della tesi (traduzione):

Pattern di variabilità genetica e linguistica. Studio di marcatori uniparentali.

Tutore: Prof. (Cognome e Nome)

Prof. Barbujani Guido

Settore Scientifico Disciplinare (S.S.D.)

BIO/18

Parole chiave della tesi (max 10):

Genetics, Linguistics, Ancient DNA, Twinning rates, mitochondrial DNA

Consapevole, dichiara

CONSAPEVOLE: (1) del fatto che in caso di dichiarazioni mendaci, oltre alle sanzioni previste dal codice penale e dalle Leggi speciali per l'ipotesi di falsità in atti ed uso di atti falsi, decade fin dall'inizio e senza necessità di alcuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni; (2) dell'obbligo per l'Università di provvedere al deposito di legge delle tesi di dottorato al fine di assicurarne la conservazione e la consultabilità da parte di terzi; (3) della procedura adottata dall'Università di Ferrara ove si richiede che la tesi sia consegnata dal dottorando in 2 copie, di cui una in formato cartaceo e una in formato pdf non modificabile su idonei supporti (CD-ROM, DVD) secondo le istruzioni pubblicate sul sito : http://www.unife.it/studenti/dottorato alla voce ESAME FINALE – disposizioni e modulistica; (4) del fatto che l'Università, sulla base dei dati forniti, archivierà e renderà consultabile in rete il testo completo della tesi di dottorato di cui alla presente dichiarazione attraverso l'Archivio istituzionale ad accesso aperto "EPRINTS.unife.it" oltre che attraverso i Cataloghi delle Biblioteche Nazionali Centrali di Roma e Firenze. DICHIARO SOTTO LA MIA RESPONSABILITA': (1) che la copia della tesi depositata presso l'Università di Ferrara in formato cartaceo è del tutto identica a quella presentata in formato elettronico (CD-ROM, DVD), a quelle da inviare ai Commissari di esame finale e alla copia che produrrà in seduta d'esame finale. Di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi; (2) di prendere atto che la tesi in formato cartaceo è l'unica alla quale farà riferimento l'Università per rilasciare, a mia richiesta, la dichiarazione di conformità di eventuali copie. PER ACCETTAZIONE DI QUANTO SOPRA RIPORTATO

Dichiarazione per embargo

12 mesi

Richiesta motivata embargo

1. Tesi in corso di pubblicazione

Liberatoria consultazione dati Eprints

Consapevole del fatto che attraverso l'Archivio istituzionale ad accesso aperto "EPRINTS.unife.it" saranno comunque accessibili i metadati relativi alla tesi (titolo, autore, abstract, ecc.)

Firma del dottorando

Ferrara, li 24/03/2017 (data) Firma del Dottorando Roberta Rosa Mosca

Firma del Tutore

Visto: Il Tutore Si approva Firma del Tutore Guido Barbujani

Thesis overview

This dissertation is divided in three sections and focuses on two of the projects I worked on during my three-years PhD, funded by a European Research Council grant whose aim was to improve our understanding about the coevolution of language and genes. Both the projects share the uniparental markers as tool used for the investigation of the human evolutionary history, but each of them addresses different scientific questions by means of a different combination of molecular and statistical methods.

Part I is a technical summary on current knowledge about uniparental markers features, and especially on the pros and cons of their usage for addressing questions stemming from the fields of linguistics and archaeology.

Part II summarizes the results of one of the ERC-founded LanGeLin works; here I describe the comparison of patterns of genetic and liguistic diversity in 36 Eurasian populations.

Part III addresses questions related with the analysis of complete mitochondrial sequences from Mesolithic times, which allowed us to address questions regarding Neolithic and pre-Neolithic peopling of Sardinia.

As a side project, I was also involved in the study of the differences in twinning frequencies among human populations of Africa (where the twinning rate is maximum), Europe and Asia (where the twinning rate is minimum). The contact point between these two projects was represented by the common bioinformatics and biostatistical tools needed for analysis of large genomic datasets in the geographical space.

Although I considered that, for the sake of consistency, this thesis will mostly focus on the ERC-funded project, I am enclosing, in the final *Manuscripts* section, both papers produced during my doctoral years.

Abstract

This dissertation is divided in three sections and focuses on two of the projects I worked on during my three-years PhD, funded by a European Research Council (ERC) grant LanGeLin.

Both the projects share the uniparental markers as tool used for the investigation of the human evolutionary history, but each of them addresses different scientific questions by means of a different combination of molecular and statistical methods.

Part I is a technical summary on current knowledge about uniparental markers features and on the pros and cons of their usage for addressing questions stemming from the fields of linguistics and archaeology.

Part II summarizes the results of one of the ERC-founded LanGeLin works; here I describe the comparison of patterns of genetic and liguistic diversity in 36 Eurasian populations. The ERC-founded LanGeLin project aims to improve our understanding about the coevolution of language and genes. R. Sokal and L.L. Cavalli-Sforza in 1988 showed that a correlation between genetic and linguistic variation within major language families is actually present, but due to imperfect methods to quantify linguistic variation, it has been difficult to compare populations belonging to distant linguistic groups. Thanks to the newly PCM linguistic method, a new way to compare languages is now available, based on stable linguistic syntax features. It is now possible to test the correlation between genetic and linguistic data in a broad geo-linguistic scale, as this thesis will do, and to interpret in evolutionary terms both the rule and the exceptions. It is also possible to study maternal and paternal lineages separately, to inquire their migrational histories. Two different migrational histories emerged, with women dispersing at a higher rate than men. When comparing genetic and linguistic features a neither obvious nor simple pattern is detectable: correlations between languages and DNA variants depend on the geographical area and the genetic markers considered.

Part III addresses questions related with the analysis of complete mitochondrial sequences from Mesolithic (Ms) times, which allowed us to address questions regarding Neolithic (Ne) and pre-Neolithic (pN) peopling of Sardinia. We investigated the role of two Ms Sardinian mtDNA sequences in the European context. Little is known about the genetic prehistory of Sardinia because of the scarcity of pN human remains. Modern Sardinians are known as genetic outliers in Europe, showing unusually high levels of internal diversity and a close relationship to early European Ne farmers. However, how far this peculiar genetic structure extends and how it originated was to date impossible to test. Here I present the first and oldest complete mtDNA sequences from Sardinia, dated back

to 10,000 yBP. These two individuals belong to rare mtDNA lineages never been found before in Ms samples and that are currently present at low frequencies also in the whole Europe. When compared with other European pN data, the Ms Sardinian sequences appeared already well differentiated, and in general more similar to pre-Last Glacial Maximum populations, than to coeval sequences. As a side project, I was also involved in the study of the differences in twinning rate (tr) among human populations of Africa (where the tr is maximum), Europe and Asia (where the tr is minimum). The contact point between these projects was represented by the common bioinformatics and biostatistical tools needed for analysis of large genomic. Although I considered that, for the sake of consistency, this thesis will mostly focus on the ERC-funded project, I am enclosing, in the final *Manuscripts* section, both papers produced during my doctoral years.

Sunto

Questa tesi, suddivisa in tre parti, riassume l'attività di ricerca da me svolta durante i tre anni di dottorato, sovvenzionato dal progetto European-Research-Council LanGeLin, il cui scopo principale è di migliorare le conoscenze sulla co-evoluzione di lingue e geni. I progetti descritti condividono l'uso di marcatori uniparentali usati per gli studi di evoluzione umana, ma differiscono per la combinazione di metodi molecolari e statistici.

La *Parte I* descrive lo stato dell'arte dei marcatori uniparentali e i pro e contro del loro utilizzo in ambito linguistico e archeologico.

La *Parte II* riassume i risultati delle ricerche condotte nell'ambito del progetto LanGeLin che descrive la diversità dei pattern genetici e linguistici in 36 popolazioni Euroasiatiche. Il progetto LanGeLin (Language and Gene Lineages), finanziato dal "European Research Council" ha lo scopo di testare l'ipotesi di Darwin presentata in "Origine delle specie". Darwin intuì che l'albero filogenetico delle diverse sottospecie umane, potesse sovrapporsi a quello ottenuto a partire dalle diverse lingue, offrendo di fatto la possibilità di studiare la genealogia delle lingue e allo stesso tempo capire come le differenze tra queste avrebbero permesso di far luce sugli aspetti elusivi della storia demografica umana. R. Sokal e L.L. Cavalli-Sforza nel 1988 hanno elucidato come la comparazione dei vocaboli rifletta la correlazione fra variabilità genetica e linguistica nelle maggiori famiglie linguistiche ma, a causa di metodi linguistici, risulta difficile comparare popolazioni derivanti da gruppi linguistici distanti. Il nuovo metodo linguistico PCM si basa sulle caratteristiche linguistiche più stabili della sintassi. È stato dunque possibile, anche in questa tesi, testare su larga scala geo-linguistica la correlazione tra dati genetici e linguistici. Lo studio delle discendenze materne e paterne è stato condotto separatamente per indagarne le relative storie migrazionali: due differenti storie migrazionali sono emerse dall'analisi del Ychr (discendenza patrilineare) e del mtDNA (discendenza matrilineare). Non ovvie considerazioni sono scaturite dalla comparazione delle caratteristiche genetiche e linguistiche, che ha portato a definire come la correlazione tra lingue e sequenza genetica sia dipendente dall'area geografica e dai marcatori genetici considerati.

La *Parte III* descrive l'analisi di sequenze di mtDNA del Mesolitico (Ms) che ci ha permesso di indagare sul popolamento della Sardegna in periodo Neolitico (Ne) e pre-Neolitico (pN). Lo studio è stato incentrato su due sequenze mitocondriali sarde Ms in relazione al contesto europeo. C'è ancora molta incertezza sulla variabilità genetica della Sardegna preistorica, a causa della scarsità di resti umani Ne. Dal punto di vista genetico, i sardi moderni possono considerarsi un gruppo a se stante rispetto al resto dell'Europa continentale, mostrando alti livelli di diversità interna e una forte vicinanza con i primi coltivatori europei del Ne. Questa tesi riporta le due prime sequenze mtDNA complete sarde, datate circa 10000 anni fa. I due individui confermano un'occupazione mesolitica dell'isola e rappresentano un aplotipo mai trovato prima in Sardegna mesolitica e con basse frequenze nell'intera Europa. Le due sequenze risultano ben differenziate se comparate con altri dati europei pN, e più simili a popolazioni dell'era pre-glaciale che a popolazioni coeve. Analisi di inferenza Bayesiana hanno mostrato come i primi abitanti dell'isola abbiano contribuito poco al popolamento attuale dell'isola, la cui diversità genetica deriva da migrazioni dal continente in tempi neolitici. Un progetto portato avanti parallelamente, ha riguardato lo studio di frequenze alleliche in gemelli dizigotici provenienti da popolazioni umane africane, europee ed asiatiche. Le tecniche bioinformatiche e biostatistiche usate per le analisi genomiche su larga scala, fanno da collante con i precedenti progetti descritti.

Contents

"I believe in intuition and inspiration.
Imagination is more important than knowledge.
For knowledge is limited, whereas imagination embraces the entire world,
stimulating progress, giving birth to evolution.
It is, strictly speaking, a real factor in scientific research."
Albert Einstein

Part I

Introduction

Chapter 1

Population genetics: a uniparental viewpoint

Human genetics is a valuable source of information regarding our past: it allows us to ask when and where we came from and how we have changed over time. Through genetic analysis we can reconstruct otherwise elusive aspects of human evolutionary history. The improvement of biomolecular technologies makes possible to obtain large-scale datasets of genomic data, using which it is possible to ask subtler questions than those that could previously be addressed based on traits of anthropological relevance.

Demographic inference from population-genetics evidence will soon be 40 years old. On September 1st, 1978, Menozzi, Piazza and Cavalli-Sforza (1978) published their principal-component analysis of genetic variation in Europe. The genetic patterns thus describe had a close match in the dates of onset of agriculture, suggesting that the two phenomena were, in fact, one and the same. Based on the parallelism between archaeological and genetic diversity, Menozzi and collaborators concluded that the Westward and Northward spread of farming from the Fertile Crescent was mainly a demic, not cultural, process. Much progress has been made ever since, both in the quality of data (polymorphisms in the DNA rather than in the proteins), in the amount of data available (now essentially covering all continents, with remarkable exceptions in North America and Australia), and in the inferential biostatistical approaches. As a matter of fact, the main problem is no longer the possibility to collect data, but rather understanding which data and which set of analyses is the best to address each specific demographic question.

Studies intended to disentangle human population history have used various types of genetic markers. The choice of which is the best genetic marker to use when it comes to the study of past events has always troubled researchers. Uniparental genomes, such as mitochondrial DNA (mtDNA) and Y chromosome, have been one of the first storytellers of human history. They reduced the problems of recombination that affects nuclear DNA (nDNA). However, they offer only a limited snapshot of genealogical information but provide valuable insights into human history and unique answers to demographic events because of the exceptional characteristics they possess.

The circular molecule of the mtDNA, has been characterized in the last century and has been widely used in evolutionary studies over the last three decades due to its small size, the technical ease and the low-cost of manipulating mitochondrial genome and the dynamics of its evolutionary rate change. Further, the ever-updating next generation sequencing (NGS) technologies increased the quality of the data and allowed the user to broaden the number of samples analyzed, optimizing the robustness of the statistical results.

1.1 - Uniparental markers

1.1.1 The mitochondrial genome and the Y chromosome features

The mitochondrial DNA (mtDNA) is a circular, double-strained DNA molecule, which is located outside the nucleus, in the energy producing mitochondria. Unlike nDNA, mtDNA is not bounded in chromosomes, but in small circular molecules, which are present in many copies in each mitochondrion. The mitochondrial genome consists of two regions (Figure 1.1) : the larger coding region, which encodes 37 genes, and a shorter control region, D-loop (displacement loop), with regulatory functions, which is the major target for evolutionary studies because it is rich in polymorphisms and so also called Hyper-Variable Region (HVR) (Stoneking 2000). MtDNA is widely used in the research of the human past, as a high number of copies of it are available; this is particularly advantageous, when it comes to the retrieval of ancient DNA (O'Rourke et al. 2000). The first sequencing of mtDNA took place in Cambridge in 1981 (Anderson et al. 1981), however it was not until 1987 that it was used for evolutionary studies by the team of Allan Wilson that conducted a worldwide survey on mtDNA to infer where and when modern humans arose (Cann et al. 1987). Since then, researchers have accumulated a considerable amount of information on the mitochondrial genome system thereby providing a better coverage of worldwide mitochondrial variation. Most of the early evidences came from studying specific populations, such as isolated tribes or isolated people. After three decades of research, we now have a relatively complete maternal phylogeny of all human populations, which enables an accurate knowledge of human origins and patterns of dispersion.

The other uniparental genetic marker is the Y chromosome, responsible for male-sex determination. Generally when we refer to the Y chromosome in evolutionary studies, we are talking about its non-recombining part (NRY), which constitutes approximately 95% of the Y chromosome (Jobling & Tyler-Smith 1995). It scarcely socialises with its neighbouring X chromosome, is poor in genes, and tends to degenerate quickly respect to the other nuclear DNA genome (Jobling & Tyler-Smith 2003). The first complete Y chromosome sequence was undertaken in 1985, but it was not until 1995 that geneticists had enough data to address worldwide variation and to reconstruct its patterns of diversity in geographical and temporal grounds.

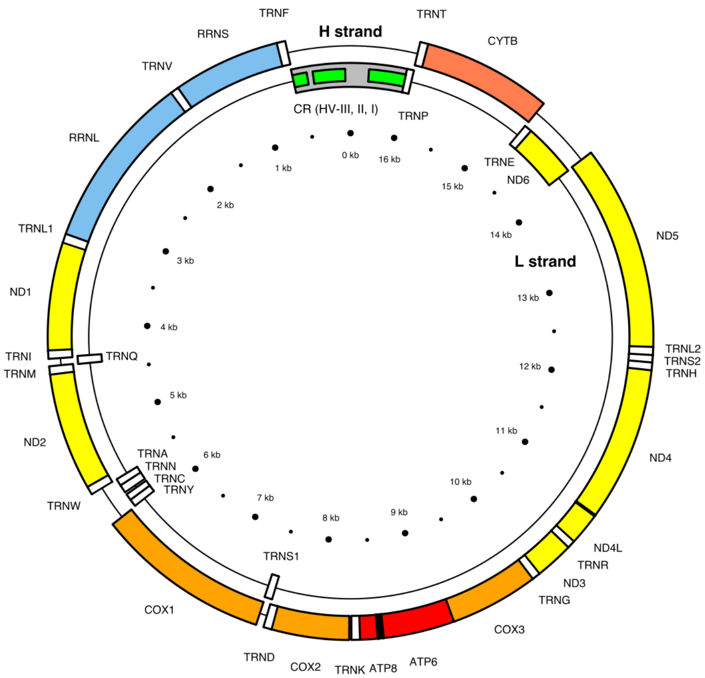

Figure 1.1: Map of the human mitochondrial DNA genome
The H (heavy, outer circle) and L (light, inner circle) strands are given with their corresponding genes. There are 22 transfer RNA (TRN) genes for the following amino acids: F, V, L1 (codon UUA/G), I, Q, M, W, A, N, C, Y, S1 (UCN), D, K, G, R, H, S2 (AGC/U), L2 (CUN), E, T and P (white boxes). There are 2 ribosomal RNA (RRN) genes: S (small subunit, or 12S) and L (large subunit, or 16S) (blue boxes). There are 13 protein-coding genes: 7 for NADH dehydrogenase subunits (ND, yellow boxes), 3 for cytochrome c oxidase subunits (COX, orange boxes), 2 for ATPase subunits (ATP, red boxes), and one for cytochrome b (CYTB, coral box). Two gene overlaps are indicated (ATP8-ATP6, and ND4L-ND4, black boxes). The control region (CR) is the longest non-coding sequence (grey box). Its three hyper-variable regions are indicated (HV, green boxes).

The first attributes that makes them interesting is the nature of their sex-specific inheritance and the haploidy. The advantage of it being inherited from only one parent is that it gives a snap-shot of only one of the parental lineages. The mtDNA sequences allow to determine the matrilineal lineages, while the Y chromosome the paternal. This feature is helpful when it comes to the reconstruction of family relationships. The sex-specific inheritance is also useful to trace the legacy of past populations. The comparison of mtDNA and Y chromosome data allows insights on population structure on the basis of sex-biased differences. Human groups have different sex apportioning due to cultural and social practices, which influence significantly differences in male and female mobility (Underhill & Kivisild 2007). For instance, data from mtDNA have underlined a South/East Asian contribution as result of a “direct train” expansion from Asia into the Polynesian islands with a low level of admixture with the Melanesian people. Conversely, Y chromosome data shows a different picture with the main genetic contribution coming from the Melanesian people (Kayser et al. 2006). These results can be interpret as a reflection of the matrilocal nature of Polynesian societies, where new families establish home near to the bride's family.

Further variables, like whom we marry (or, at least, breed with), have an impact on the patterns of diversity of human DNA. Indian society is an example that shows how sex policy influences the genetic make-up of populations. For centuries, Indian society has been ruled by a strict caste division, which regulates marriages and social interactions and separates society into endogamous groups. The caste system provides information on a restricted sex-biased admixture of social groups arising as a result of the different marriage policies applied to females and males. Caste boundaries are more permeable for females, who can more easily marry a man from a different class caste than vice versa. This has resulted in a sex-biased genetic scenario, in which Y lineages reveals distinct distribution patterns among castes with a limited gene flow between them, while mtDNA is more genetically similar across the four caste divisions (Thanseem et al. 2006).

The last advantage of the sex-specific and haploid status is that uniparental genomes are geographically clustered. mtDNA and Y chromosome are grouped in continental and regional clusters that can serve as population finger prints (Figure 12). This is due to the fact that, as they are inherited from only one parent, the dimension of the effective population size (Ne) is 1/4 in respect to the autosomal, and the number of gene copies is reduced to one. With a lower effective population size, uniparental genomes are more susceptible to genetic drift, which reduces the differences between individuals within populations and increases the divergence between groups of mtDNA and Y chromosome in different populations (Underhill & Kivisild 2007).

Another important quality of uniparental genomes is that they do not undergo recombination. Recombination occurs between two homologues autosomals during cell division, when each autosomal breaks into two pieces and then each piece re-joins with (as a rule; but there are exceptions) the same part of the homologous autosome. During the recombination process, genetic material is therefore reshuffled and combined with material coming from the other parent. Genetic information is, consequently, mixed up and this makes inference on ancestry more complicated, and sometimes even impossible, to detect. The human mtDNA escapes completely recombination and since it exists in form of a haploid DNA molecule. The Y chromosome partially recombines with the X chromosome and only genetic loci within the NRY region are considered in anthropological studies. The advantage of studying the uniparental genome is that noise in the genetic information produced by recombination is mitigated, as the DNA are not broken down; instead, genetic information is passed intact from one generation to the next, when mutation do not act.

When we analyze modern genetic variation, we see this process across generations.

The sets of genes inherited together from a single parent because physically connected along a DNA fragment are called haplotypes, and haplotypes that share a common ancestral mutation are clustered into haplogroups. The study of haplotype diversity among populations has proven to be particularly advantageous for reconstructing the events that contributed to the formation of present-day populations' genomes. Both the *male* and *female* phylogenetic trees show that the ancient branches lie in Africa, and furthermore that human groups on this continent present the highest level of heterozygosity, which progressively decreases moving away (Pakendorf & Stoneking 2005, Underhill & Kivisild 2007).

An advantageous attribute of uniparental markers is also their relatively fast rate of mutation. High mutation rate produces numerous polymorphisms and the absence of recombination facilitates possibility to trace them across generations. By counting the number of genetic mutations between two mtDNA/Y chromosome lineages it is possible to estimate the age of the common ancestor of two lineages. It is important to stress that this date (the date at which a biochemical event, a mutation in DNA, occurred) has no relationship with the time of splitting between populations (the date at which a demographic event, a separation, occurred) (Barbujani & Goldstein 2004). However, by following backwards in time the accumulation of mutations, geneticists can reconstruct the genealogy (by the coalescent theory) of the alleles or haplotypes, estimate the time to the most recent common ancestor (TMRCA), a good starting point to reconstruct processes in population history. Central to this calculation is the assumption that the mutation rate will be effectively constant in the uniparental genome and can therefore serve as a molecular clock.

1.1.2 Limitations

Uniparental genomes give many insights into the events that contributed to human history and the genetic makeup of modern groups. However, they also present two main disadvantages, which restrict the reconstruction of a complete picture of past demographic events. First, they provide only a sex-biased snap-shot of human history. They are storytellers of only the maternal and paternal lineages and do not record the entire distribution of ancestors.

Secondly, uniparental genomes have low effective population sizes making them more susceptible to random genetic drift. For instance, purifying selection removes deleterious mutations and reduces the level of diversity within the Y chromosome (Wilson Sayres et al. 2014). MtDNA is based on a single locus which makes it susceptible to the inherent stochasticity of evolutionary events. Moreover, since it is haploid and uniparentally inherited, it is effectively a quarter of the population size (Ne) of diploid nuclear DNA and this leads to low diversity within a species. On the contrary, the high genetic drift outweighs the elevated mutation rate and leads to large differences between populations.

This can create a discrepancy with the results given by nuclear DNA in terms of coalescence times and in the estimation of the molecular clock. For instance, mtDNA and nDNA gave divergent times for the separation of Eurasian and African populations (Cann et al. 1987, Green et al. 2010).

Uniparental markers are fundamental for constructing sex-based-snap shots of human population events. Researchers agree that a robust testing of demographic hypothesis requires data collected from multiple loci in order to reduce the ascertainment biases. In recent years, the era of the nuclear DNA has not replaced the information given by the uniparental genomes, but instead has contributed to further completing it.

Chapter 2

An interdisciplinary approach

The concept that the genes of living populations contain decipherable traces of their pasts is the seed from which population genetics has grown. Genetic markers, such as the uniparental mtDNA and NRY, are the core element by which are applied the methods and the theories of population and evolutionary genetics to reconstruct human population histories and demographic past events. The fields of research stretch from the process of human evolution to ancient migrations, from the biological relationships among populations to population structures. To avoid a misinterpreted events reconstruction, essential is the connection between the genetics and the research fields that broaden the genetic perspective and provide information about modern population structures in relationship to historical and cultural events. In the mid-1980es geneticists and anthropologists worked together to enlarge the field of research and to test accurate models of human origins and dispersions and the origins of modern genomes (Cann et al. 1987, Cavalli-Sforza et al. 1994). The relationship between linguistic families and ethnic groups was addressed in cooperation with one current in linguistics, which attempted to compare reconstructions of language development with genetic models (Cavalli-Sforza et al. 1988*a*). In the same years another challenging collaboration started between the teamwork of the population genetics Cavalli-Sforza and the archaeologist Ammermann, who brought together human genetics and archaeology to reconstruct the shift from hunting-gathering to agriculture in Europe; two disciplines coming from different academic areas, archaeology as division of humanities and population genetics of science, were combined as guide-example of cross-disciplinary efforts (Ammerman & Cavalli-Sforza 1984).

2.1 The Linguistics approach

Starting from Sokal's (1988) analysis of genetic distances in Europe, many anthropological and genetic studies have modeled population history and migration using language similarities looked for a large-scale correspondence between the distribution of classical genetic markers (blood groups, serum proteins, etc.) and certain long-range language classifications found in the linguistic literature (Figure 2.1). The basic idea, as Sokal (1988) put it, is that populations speaking the same language are likely to have common ancestors in the recent past, and populations speaking similar languages are likely to share ancestors in a more remote past. Therefore, by quantifying the degree of linguistic resemblance, one can make hypotheses about population history, which can then be tested against genetic evidence. Most such studies (but not all: see e.g. Mona et al. (2009)) showed that, in general, genetic change does parallel language change, hence both tend in general to reflect the same demographic processes (Barbujani & Sokal 1990, Sajantila et al. 1995, Poloni et al. 1997, Lansing et al. 2007). These works have been received with much serious criticism, though, and has remained very controversial, especially among linguists. Indeed, most linguists have denied the very possibility of a reliable global or long-range classification of languages, for clear methodological limits. The classical methods so far used (i.e. the *comparative method* and *Greenberg's mass comparison*), fail either with respect to universality or reliability because essentially based on entities ultimately characterized by lexical arbitrariness (broadly understood as to include roots and grammatical morphemes, as well as sound laws connecting them crosslinguistically). Safely identifiable similarities of words/morphemes in sound and meaning tend to dissolve within a short time span, sometimes placed around 8,000 ± 2,000 years (Nichols 1996). Linguistic evolutions over longer time periods may be impossible to reconstruct from lexical comparisons, because of deceiving affinities emerging by sheer chance and the lexicon's inability to provide exact and broad-scope taxonomic (distance) measures. Accordingly, large-scale genetic studies had to resort to very coarse classifications of languages, which are generally controversial among linguists (Belle & Barbujani 2007).

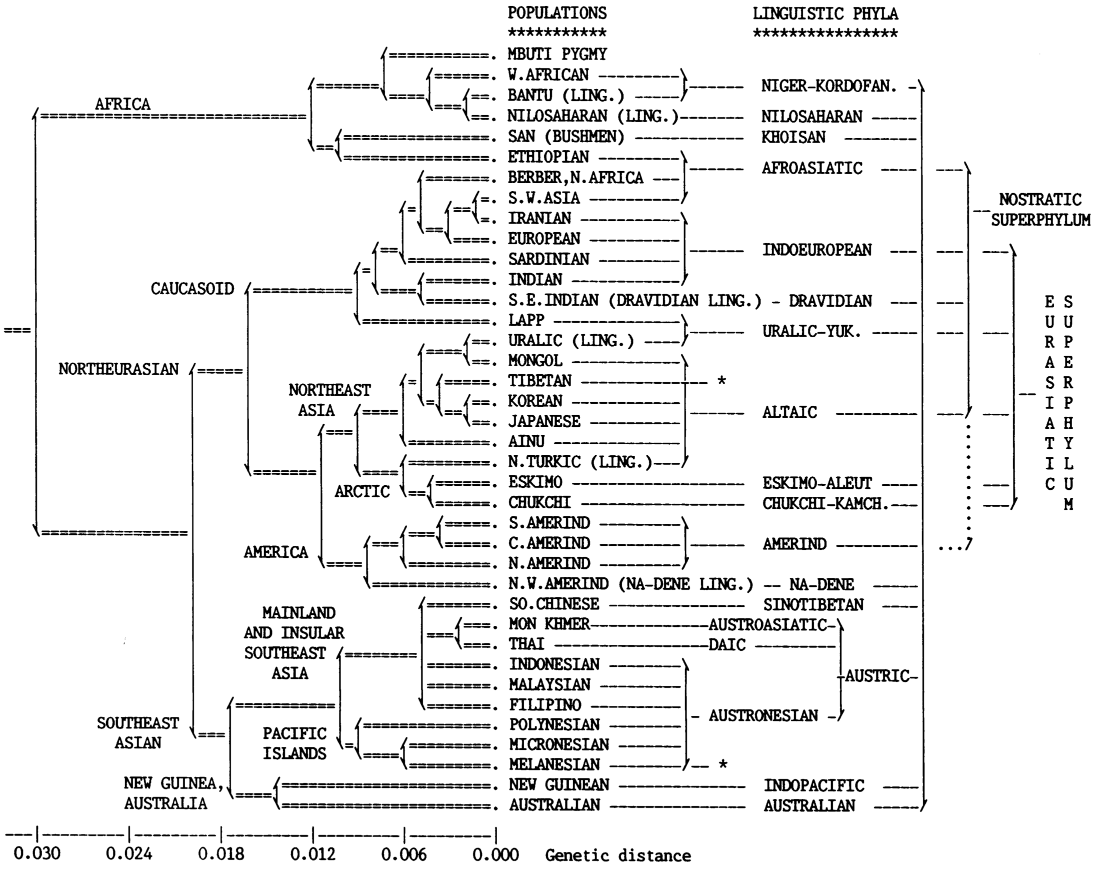

Figure 2.1: Comparison of genetic tree and linguistic phyla. Figure reproduced from Cavalli-Sforza et al., (1988)

Starting from the Eighties', comparisons between languages and genes were based on lexicon on the one hand, and on allele frequencies inferred from protein variation or blood groups on the other. Although evidence for a general correlation between patterns of linguistic and genetic variation was found in Europe (Sokal 1988, Barbujani & Sokal 1990, Sajantila et al. 1995), in the Caucasus (Barbujani et al. 1994), and even at a global scale (Cavalli-Sforza et al. 1988*b*, Barbujani & Pilastro 1993, Chen et al. 1995), these studies suffered from a radical limitation remained very controversial, especially among linguists. Indeed, most linguists have denied the very possibility of a reliable global or long-range classification of languages, for clear methodological limits. The classical methods so far used (i.e. the comparative method and Greenberg's mass comparison), fail either with respect to universality or reliability because essentially based on entities ultimately characterized by lexical arbitrariness (broadly understood as to include roots and grammatical morphemes, as well as sound laws connecting them crosslinguistically), (Bateman et al. 1990, Sims-Williams 1998). Later studies inferring genetic relationships from DNA data (Poloni et al. 1997, Gray & Atkinson 2003, Nettle & Harris 2003, Hunley & Long 2005, Ramachandran et al. 2005, Belle & Barbujani 2007, Tishkoff et al. 2009, Bouckaert et al. 2012) did not tackle the core of the problem, which was linguistic, not genetic.

However, to address and test the congruence hypothesis at a general, worldwide rather than regional, level, a broad-scope analysis of linguistic and genetic variation is inevitably needed. This has proved an impossible challenge so far, because of the mentioned weakness of long-distance language taxonomies and of the lack of dedicated datasets with a sufficient number of populations speaking different languages and, at the same time, typed for many genetic polymorphisms.

Any linguistic taxonomic method with some global ambition should be able to identify sets of correspondence characters both safe from chance (i.e. probabilistically reliable) and universally applicable. The two methods so far used (i.e. the classical comparative method and Greenberg's mass comparison), which are essentially based on entities ultimately characterized by lexical arbitrariness (broadly understood as to include roots and grammatical morphemes, as well as sound laws connecting them cross-linguistically), both fail either with respect to universality or reliability.

2.1.1 The LanGeLin project

The ERC-Founded LanGeLin project (Language Genetic Lineages) aims to combine the power of languages with the genetics to answer a hypothesis Darwin put forward almost 200 years ago in his Origin of Species (1859):

> *If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; [...] such an arrangement would, I think, be the only possible one.*

Darwin supposed that if comparing the phylogenetic tree of evolution of populations with the phylogenetic tree of the evolution of languages they should overlap. Therefore, it is possible to wonder to what extent the cultural traits can rule and limit the genetic transmission in the human species and so to ask if humankind evolution has been shaped by cultural traits as well.

Past studies have shown that a parallelism between genetics and linguistic superfamilies evolution is actually present and this approach is a perfect tool to explain human expansions (Sokal 1988, Cavalli-Sforza et al. 1988*a*, Bouckaert et al. 2012). Further results were not any clearer, in fact other studies have shown the weakness of the test used to classify languages, that is the lexicon comparative method. The scientific method is based on basically two essential criteria: the method ought to be universally applicable and probabilistically reliable. This applies for linguistic taxonomy studies as well. Linguistic taxonomic methods such as the classical comparative method and the Greenberg's mass comparison, both fail respecting the two criteria mentioned. They are based on testing the lexicon comparing word roots, sounds and morphemes. For example they select a word and check its expression in different languages at different times, guessing the common ancestry simply from resemblance in sound and meaning. The problem is that these comparisons are more likely to apply well only to languages so similar that their kinship is often already obvious. Therefore, the mentioned tests use degrees of relatedness without an actual probabilistic reliability or even any possibility to quantify the values of linguistic distances, moreover they are unfit for any long-range comparison. Lexical methods have thus proved to be a poor choice for wide-range language studies and any parallelisms with biological classifications (Longobardi & Guardiano 2009, Greenhill et al. 2010, Longobardi et al. 2015).

Parametric Comparison Method: a new linguistic tool

It was only recently that Longobardi and Guardiano (2009) devised a method which, in principle, may allow quantification of language differences not involving the lexicon. This method, the Parametric Comparison method, describes abstract and general properties of languages, in grammar and syntax, thus enabling the investigator to estimate measures of linguistic diversity even among unrelated languages. Therefore the Project LanGeLin aims to reconstruct the intricate relationships between languages and genes in order to get a closer understanding of human demographic history, taking in mind the power of languages and the bread crumbs left by our genetics, which could be minutely analyzed only recently, due to improvements of genomic and statistical techniques. Moreover, LanGeLin makes use of a new comparative method to classify languages and to go beyond past results that were obtained using the lexicon comparative method. The newly PCM or Parametric Comparison Method (Longobardi & Guardiano 2009) focuses on syntactic characteristics and differences, namely language features disregarded in previous studies. This new method can provide quantitative results that are formally analogous to values obtained in genetic studies. It is suitable for broad-scope analyses such as phylogenetic studies.

The PCM implements precisely this idea of analyzing relatively large sets of identities/differences in syntax for historical purposes. The core grammar of each language is represented as a string of binary symbols, each of them encoding the value of a syntactic parameter (Chomsky 1981, Baker 2001, Biberauer 2008). Parameters are drawn from a supposedly universal list, representing the structured space of variation predefined by a species-specific human capacity, labeled by some 'universal grammar' (UG) or 'faculty of language'. Therefore, in principle, all languages, no matter how distant, could be compared by PCM, bypassing many problems emerging when lists of words are employed.

LanGeLin state of the art

The pilot LanGeLin work, published by Colonna et al. (2010), describe the preliminary analyses based on the newly proposed PCM linguistic method. They compared populations belonging to four major linguistic phyla(Ruhlen 1987, Heine & Nurse 2000), namely Afro-Asiatic (Arabic), Niger-Congo (Wolof), Uralic (Finnish and Hungarian) and Indo-European (French, Irish, Italian, English, Russian and Sindhi), and a linguistic isolate, Basques. Estimating linguistic distances among these samples from lexical comparisons would have been highly problematic, or simply unwarranted (Bateman et al. 1990). The results show that Longobardi & Guardiano (2009) index of distance based on syntax shares some useful empirical properties with other indices that are popular among geneticists. This measure shows, in fact, a broad general correlation with genetic distance, and allows one to identify outlier populations. Most genetic distances showed the well-known positive correlation with geographic (Cavalli-Sforza et al. 1988*a*, Sokal et al. 1990, Quintana-Murci et al. 2001) and linguistic distances, in agreement with many previous studies based on lexical comparisons (Sokal 1988, Poloni et al. 1997, Atkinson & Gray 2005, Libiger et al. 2009). However, these correlations are now shown to exist over several continents, because for the first time they were estimated on the basis of a robust measure of linguistic distance, suitable for such long-range comparisons.

The proof-of-concept study described (Colonna et al. 2010) of gene/language congruence in a small sample of Old-World populations has already shown how correlations can be found between a preliminary set of parametric distances and genetic ones. Further positive results were also obtained considering an European set of comparisons (Longobardi et al. 2015), where the non-independence of characters was controlled by making explicit hypotheses about implications of syntactic properties and adopting a distance calculation appropriate for them. Moreover, two tools developed for language comparison, i.e. Bouckaert et al. (2012) expanded list of Indo-European lexical cognates and Longobardi and Guardiano's (2009) Parametric Comparison Method (PCM), were compared. These linguistic resources were here used to interpret patterns of genome-wide variation in 15 European populations (from three different linguistic families), inferred from autosomal single nucleotide polymorphisms (SNPs) data. First through the quantitative approach to cognate words (Bouckaert et al. 2012), and then through the PCM syntactic method they overcome limits of previous studies (Sokal 1988). Through syntax, precise comparison and measuring is shown possible even across established linguistic families: the main families/subfamilies of Europe are discriminated by means of just 56 abstract characters suggested by formal grammatical theory, using standard methods of evolutionary biology and without resorting to unsafe long-range etymologies. Populations speaking similar languages in Europe tend to resemble each other at the genomic level, thus suggesting that cultural change and biological divergence have proceeded in parallel in Europe at least as a rule (with some exceptions (Bolnick et al. 2004)).

2.2 The Ancient DNA approach

In classical population genetics, past evolutionary processes were inferred from patterns in the data describing contemporary populations. For two decades now, ancient DNA studies, also termed molecular anthropology, have emerged as an approach to support the genetics in describing the diversity of contemporary human groups and understanding their history and genetic makeup (Cann et al. 1987). Such methods have been also applied with success to the fossil records, adding DNA from ancient individuals and extinct species to the list of proxies available to study our origins. The possibility to retrieve genetic information from ancient samples has definitely been one of the most significant achievements in the field of human genetics and has opened several new windows into the comprehension of the human past (Ermini et al. 2015). Ancient DNA (aDNA) is any sample that is retrieved from a non-living organism; such as mummified tissues, preserved plant remains, skeletal material, and so on. The first successful study was presented at the University of Berkeley in 1984, when geneticists retrieved and sequenced the DNA of an extinct zebra species, the quagga(Higuchi et al. 1984). This first success was followed shortly after by the first human sequencing, of an Egyptian mummy (Paabo 1985). At the beginning of these studies, geneticists had to deal with difficulties arising from a not yet well developed methodology, and had to rely on bacterial cloning, which made extraction and sequencing both challenging and expensive. The advent of the PCR gave a great input to the research for the analysis of a minuscule amount of highly degraded DNA through the production of an unlimited number of copies. The second major advancement was due to progress in sequencing technologies. The shotgun cloning enables the sequencing of long strands of DNA, which is of particular importance as it has resulted in the genome draft of archaic hominids. The technological advance of next-generation sequencing has helped research not only to achieve better insights into the human past, but also to overcome the problem of contamination of aDNA (Hagelberg et al. 2015). As we shall see, the damage typically carried by ancient molecules, once a major problem leading to difficulties in their amplification, is now the key feature by which one can tell endogenous from contaminating DNA (Gansauge & Meyer 2014).

Indeed, aDNA presents a number of problems in comparison to the study of DNA from present-day individuals. Ancient molecules are typically degraded, to an extent which depends on both the time elapsed since death and the conditions (humidity, temperature) of the place where it has been deposited; in most cases, aDNA is contaminated to some extent; and, lastly, it is more expensive to analyse due to the technologies adopted to overcome the first two limits. As soon as an organism dies, the process of repairing and maintaining DNA stops and the DNA is therefore subject to the attack of bacteria and microorganisms responsible for the decomposition of the body. The environmental conditions within which a specimen resides are therefore crucial for the rate of DNA decay. Heat accelerates the time of degradation, as it cleaves the bonds that link nucleotides from the DNA sugar-phosphate backbone. The result is that some traits of DNA are baseless and DNA molecules may be broken into fragments of some 100-300 nucleotides in length. In unusual, more fortunate circumstances, such as desiccation (e.g., Egyptian mummies), low temperature (ig. Ötzi and Denisova), or high salt concentration, the endonucleases process is deactivated and molecules are much better preserved.

The contamination process is due to the presence of any DNA in the sample that is derived from sources other than the specimen being analysed. One form of contamination comes from microbial and environmental organisms introduced into the fossil during and after deposition. Another major source of contamination comes from the people responsible for the retrieval of the material and its analysis. The main challenge is to identify those strands of DNA that belong to the tissue under study and to isolate them from the endogenous DNA Sampietro et al. 2006. Once the aDNA has been extracted and sequenced, it is validated as authentic DNA of the sample (Malmstrom et al. 2007). The first studies made using aDNA, did not produce convincing results due to the high levels of contamination, which had obscured the authentic aDNA signal, and results in erroneous results. For example, in the first sequencing of the Neanderthal genome the extent to which modern human activity resulted in contamination was underestimated, and the aDNA data were mixed with and compromised by DNA from present-day individuals (Paabo et al. 2004). Consequently, a second study was conducted to assess the authenticity of DNA, which resulted in the discovery of fixed differences in the nuclear genome between Neanderthal and current humans (Green et al. 2009).

The advent of aDNA analysis has extended the research possibilities for studying population history by opening several branches in the field, from the study of extinct hominids tothe direct analysis of the genetic consequences of past demographic events. First of all, aDNA has revolutionized insights into ancient hominids, Neanderthal and Denisova, leading to a better comprehension of their physical features and cognitive attributes (through the analysis of specific genes affecting the phenotype) and phylogenetic relationship with modern humans (through the analysis of both coding and noncoding genome regions).

The second major focus of research is the investigation of the dynamics of past events, especially migrations and reciprocal relationship among populations. A long standing question is whether the dispersal of modern humans happened in a single wave or multiple waves Out-of-Africa and along which routes early migrants dispersed to populate the other continents (Tassi et al. 2015).

Part II

LanGeLin uniparental project

“Biologically speaking, this hypothesis of an inheritable capability
to learn any language means that it must
somehow be encoded in the DNA of our chromosomes.
Should this hypothesis one day be verified,
then lingusitics would become a branch of biology.”
Niels K. Jerne

Aim and scope In this part of my thesis I am going to describe my contribution to the LanGeLin project.

The central goal of the ERC advanced grant project LanGeLin (LANguage-GEne LINeages) is to investigate the relationship between genetic and linguistic diversity, the latter inferred from structural language features, rather than from the vocabulary. It was only recently that Longobardi & Guardiano (2009) devised a method which, in principle, may allow comparison of languages regardless of their lexicon. This method, the Parametric comparison method, describes abstract and general properties of languages, in grammar and syntax, thus enabling the investigator to estimate measures of linguistic diversity even among unrelated languages.

The purpose of this work was to then investigate at the Eurasian scale, over an area in which many different language families are present, uniparental genetic variation together with the linguistic variation. MtDNA and Y-chr provide, in fact, complementary information and allow one to investigate the different migrational histories of males and females, and their impact over the global language-gene relationships.

I am describing first the assembling of the mtDNA genetic dataset, including 36 Eurasian populations, for which sintactical linguistic parametrization were available. The mtDNA dataset comes up beside the NRY Ychr dataset, comprehending data for the same 36 Eurasian populations, collected and analyzed by PhD Stefania Sarno at the University of Bologna. We calculated and compared phylogenetic trees and Mantel's correlations between genetic, linguistic and geographical distances starting from three matrices: dGEN based on FST (genetic distances); dSYN based on syntactic features (linguistic distances); and dGEO based on geographical distance between pairs of populations. Both similarities and differences were evident between patterns of genetic and linguistic variation, casting light on both the genealogical ties between populations, and the mechanisms of language change.

Chapter 1

Materials and Methods

The genetic data collection

Starting from the linguistic sampling indications, we searched the literature for published genetic datasets, only choosing samples from the same regions where the linguistic studies had been conducted or, in their absence, the closest possible approximation, in which the same language was spoken. Below are the cases for which some nearby areas were excluded avoiding a biased genetics sampling. The linguistics informations was always preliminarily checked against the Ethnologue database (Lewis et al. 2016).

Spanish The Spanish, or Castilian, is the only language which has official status for the whole Spain country. Various other languages have co-official or recognised status in specific territories, like Galician, Asturian, Aragonese, etc. We exclude from the sampling the regions in which the Castilian is not the main spoken language.

Italian Throughout Italy, regional variations of Standard Italian, called Regional Italian, are spoken, such as: Calabrese, Sardinian, Griko, Salentin, etc. We excluded the samples belonging to dialect speaking areas.

Basque Trying to avoid sampling people who do not speak proper Basque, we considered only the regions belonging to the orange area in Figure 1.1, corresponding to the central-Basque as considered by the linguistic group. The central-Basque includes the regions named in the map as *Gipuzcoa*, *Northern Navarre* and *Labourdin.*

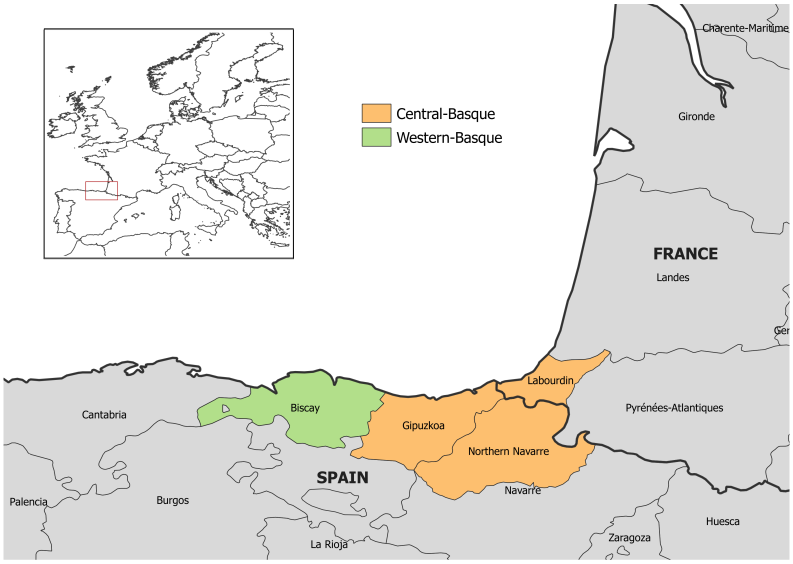

Figure 1.1: Basque geographic distribution

Hindi and Marathi The India country includes populations speaking many different languages among which the Marathi and the Hindi, the official country language, are included. The complex stratified Indian social structure, due to the presence of a caste system, made also the search of unbiased published genetic samples really difficult. The Hindi language officially is widespreads in north India regions such as: *Delhi*, *Uttar Pradesh*, *Uttarakhand*, *Rajasthan*, *Punjab*, *Madhya Pradesh*, *northern Bihar*, *Himachal Pradesh*. Whereas, the Marathi language is spoken in *Andhra Pradesh*, *Chhattisgarh*, *Goa*, *Karnataka*, and *Madhya Pradesh* states.

Eskimo Eskimo–Aleut or Eskaleut is a language family native to Alaska, the Canadian Arctic, Greenland, and the Chukchi Peninsula on the eastern tip of Siberia. Taking into account the eskimo speaker interviewed by the linguistics group in York, we searched only for published samples belonging to the Siberia region.

Chinese The two Chinese languages included in this work, namely the Mandarin and the Cantonese Chinese, are spoken across most of northern and southwestern China and the Guangdong province, respectively. We selected the published samples available in literature following the geographic indications reported in the Ethnologue database (Lewis et al. 2016) .

Japanese We excluded the samples belonging to the minor islands that surround the main Japanese land and from the regions with a large concentration of dialect speakers.

Datasets setting-up

The mtDNA HVR1 samples were retrieved in nucleotide sequence and/or haplotype (i.e. the list of polymorphic positions annotated) format. Thus, a format conversion and a sequence length constraint were performed. Firstly,, the Haplosearch tool (Fregel & Delgado 2011) provided in the Haplosite web page (http://www.haplosite.com/) was used to convert the samples collected in haplotype format into nucleotide sequence format. The rCRS mitochondrial sequence was used as reference(Andrews et al. 1999). Secondly, a sequences multi-alignment was performed using the MUSCLE software (Edgar 2004). The sequence-editing BioEdit software Hall (1999) was used to optimize the alignment and to cut all the samples to the same HVR1 sequence range shared (from position 16024 to 16383). For the male uniparental dataset, data relative to eleven Ychr-STRs (Short Tandem Repeats) were collected and analyzed by PhD Stefania Sarno at the University of Bologna.

Geographical distances

Both linguistic and genetic differences between individuals and populations reflect, in part, the effect of geography. In general, close objects tend to be more closely related than distant object, and that seems the case for both classes of variable considered in this thesis. It is therefore indispensable to keep geography into consideration, quantifying its effects upon languages and genes. It seems useful to stress that geographic distance is considered to exert its effects by limiting the possibility of contact between populations, contacts resulting in turn in a broad range of cultural similarities and genetic exchanges. Therefore, the role of geography can be meaningfully quantified if the geographic distance between populations corresponds to the migrational distance between them. Because in general no information is available about the number and entity of past migrational contacts between populations, some simplifying assumptions are unavoidable. Going from the simplest to the most complex assumptions, we might first consider that migrational distances between populations be approximated by their physical distance along the shortest path, or great circle distance. In turn, the genetic model generally associated with great circle distances is that of isolation by distance, or IBD (Wright 1931). Under isolation by distance, populations tend to diverge at the same rate because of genetic drift, but close populations maintain a greater genetic similarity than distant populations because their gene flow rates are, on average, higher.

In fact, during much of history migrating people could not know what was the shortest migrational path. Therefore, isolation by resistance, or IBR, seems a more realistic assumption. Under isolation by resistance, various features of the landscape, such as plains, arms of sea, or mountains, are associated with values representing the cost of travelling through them, so that certain paths connecting localities (e.g., those that do not require crossing seas or mountain ranges) end up being more likely than others.

The Great Circle Distance (GCD) index indicates the smallest distance between two given geographical locations taking in consideration that both are represented in a spherical surface. A Great Circle is a section of a sphere whose center coincides with center of the sphere and it gives only a unique path between any two points on the sphere surface, with exception for antipodal locations for which gives an infinite number of paths. From the path given by the Great Circle, two arcs are obtained, being the smaller one an orthodrome, also known as *Great Circle distance* (Weisstein 2015)

The Least cost path (LCP) distances measure the optimal route between two locations, defined here as the one that passes between points with the minimum accumulation of resistances or 'costs'. This effective distance may be a straightforward way to include landscape and behavioural aspects in other models which include distance as a measure for isolation (Adriaensen et al. 2003).

The conceptual basis of the IBR model lies in analogous properties of gene flow in deme networks and conductance in linear electronic circuits. All else being equal, equilibrium levels of gene flow between two demes connected by migration will increase if additional parallel movements of genes are allowed, either through increased direct movements of gametes or through indirect gene flow via intervening demes (McRae 2006). In the *Resistance distance*, the analogy between electrical and genetic connectivity is simple: as multiple or wider conductors connecting two electrical nodes allow greater current flow than would a single, narrow conductor, multiple or wider habitat swaths connecting populations allow greater gene flow. We modeled the landscape features as described in Tassi et al. (2015).

Linguistic distances

According to the formula proposed in Longobardi & Guardiano (2009) and Bortolussi et al. (2011), syntactic distances have been calculated with the normalized Jaccard distance (Jaccard 1901). The Jaccard distance is obtained by subtracting the Jaccard index from 1; it measures dissimilarity between sample sets, i.e. the number of differences between two languages divided by the sum of their identities and differences Lewandowsky & Winter (1971). The pairwise syntactic distances (dSYN) end up falling between 0 and 1.

Considering the two sets A and B, the **Jaccard index** is:

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \bigcap B|}{|A| + |B| - |A \bigcap B|}$$

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \bigcup B| - |A \bigcap B|}{|A \bigcup B|}$$

The **Jaccard distance** is then obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union

Phi-st dissimilarity index

For the mtDNA HVR1 dataset, being it constituted by nucleotide sequence information, the diversity index used to obtain the dissimilarities between populations was the Phi-st (Excoffier, Smouse, & Quattro, 1992). The diversity index used for all the other genomic datasets was the Fst index (Wright, 1951), for it calculates the diversity between populations based on the allelic information, which is on what these datasets are based. For both the calculations a 5% missing value was allowed. The distance matrices were estimated thanks to the Arlequin software (Excoffier & Lischer 2010).

Graphical data representation

Multidimensional Scaling

Distances between objects for which k measures are considered (in our case, k linguistic features, or k SNPs in the genome) can completely be represented in a k-dimensional space. As k increases, a faithful representation of the data becomes soon impossible. For that reason, many methods have been devised for reducing the dimensionality of the datasets, with minimum loss of information. Non-metric Multidimensional Scaling (hereafter Multidimensional Scaling or MDS) is one such method, designed to construct a graphic revealing the underlying structure of the input data creating thus a map of it in any n dimensions desired ($n < k$)(Wickelmaier 2005, Jaworska & Chupetlovska-Anastosova 2009). This statistical approach was first devised by Shepard (1962) and Kruskal (1964a, 1964b) and differs from the classical Multidimensional Scaling by the computation relying in the ordinal information of the data and not the distance values given. The Non-metric MDS aims to map the objects in a configuration whose distance between each object-pair is in the same rank order as in the input data. It aims to place each object in n-dimensional space such that the between-object distances are preserved as well as possible. Each object is then assigned coordinates in each of the n dimensions. Choosing n=2 optimizes the object locations for a two-dimensional scatterplot, easy to envisage and to print (Borg & Groenen 2005). With this approach based on the rank order of distances, a problem of fitness, i.e., the resemblance between the data and the final configuration, appears. In order to correct this, the programs calculates different configurations and a stress value for each, until it reaches a minimum value, thus the best configuration. The stress value thus serves to indicate the goodness-of-fit between the configuration obtained and the input data, where the smaller the value the best relation. Even though the criteria to determine the goodness of each value tends to be arbitrary and decided by each investigator, here its followed the criteria defined by Kruskal (1964a) (Table 1.1)

Stress value (%)	**Goodness-of-fit**
> 20	Poor
10	Fair
5	Good
2.5	Excellent
0	Prefect

Table 1.1: Stress and Goodness-of-fit following Kruskal (1964a) criteria

Genetics and Linguistics boundaries

The classical way to visualize genetic or linguistic variation among populations in dendrograms with multidimensional scaling is particularly suitable for identifying informative clusters or outliers. However, when the main goal of the analysis is to recognize spatial patterns related whit geography, this methods of analysis can be less suitable. Manni et al in 2004 proposed the Barrier software (Manni et al. 2004) designed for the visualization on a geographic map of the trend data contained in matrices. The idea was to implement the Monmonier's (1973) maximum difference algorithm (Monmonier 1973) in a new software in order to provide a more realistic representation of the barriers in a genetic landscape. The algorithm finds the edges associated with the highest rate of change in a given distance measure, which can be genetic, or something else. The algorithm is applied to a geometric network that connects all the populations (sampled locations) using the Delaunay triangulation derived from the Voronoï diagram(Brassel & Reif 1979) (Manni et al. 2004).

In a Voronoï diagram (Voronoi 1907) , given a plane with n points (Figure 1.2 A), the plane is partitioned into convex polygons (Figure 1.2 B) obtaining a *tessellation* in which each polygon contains exactly one generating point. Every point in a given polygon is closer to its generating point than to any other (Weisstein 2015). This *tessellation* determines which samples (populations) are neighbors, i.e. adjacent. As a consequence, two samples are adjacent if the corresponding Voronoï polygons have a common edge. Delaunay triangulation is a triangulation method to connect a set of points (localities) adjacent on a map, crossing one of the edges of the Voronoï polygons (Figure 1.2). Once a network connecting all the localities has been obtained, each edge of the network is associated with its distance value from a matrix through Monmonier's (1973) maximum difference algorithm. This algorithm is used to identify boundaries, namely, the areas where differences between pairs of populations are largest. The first boundary is traced perpendicular to the edges of the network. Starting from the edge for which the distance value is maximum and proceeding across adjacent edges, the procedure is continued until the forming boundary has reached either the limits of the triangulation (map) or closes on itself by forming a loop around a population.

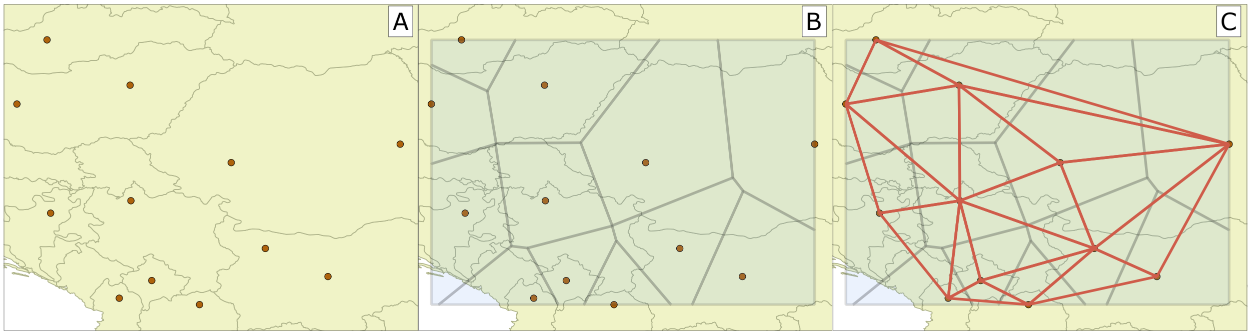

Figure 1.2: Voronoï *tessellation*
according to geographic locations (brown points) (B); Delanuay triangulation (C) scheme relative to the populations sampled in (A)

Correlation tests

Mantel test

Comparing distance matrices is not straightforward. Indeed, the values in each matrices are not independent, thus violating the assumptions of standard association tests, such as Pearson's correlation. A solution was proposed by Mantel (1967), who developed a method to estimate an empirical null distribution of a statistic, called Z, equivalent to a correlation coefficient. In practice, the corresponding values of the two matrices to be compared (say, genetic and linguistic distances, in our case) are multiplied by each other, and their sum is calculated; this is the observed Z value, which is maximum in case of a perfect positive correlation of the two distance matrices, and minimum in case of a perfect negative correlation. However, the actual value of Z also depends on the number of comparisons, i.e. on the size of the matrices. To assess its significance, Mantel proposed a randomization test, in which the entire procedure is repeated a sufficient number of times, keeping one matrix constant and randomly permuting two rows and columns of the other. This way, a null *distribution of the values of Z expected for that specific comparison* is obtained, against which the observed Z value is tested. Numerical methods then exist to convert the observed Z value so that it will fall, much like Pearson's r, in the interval between -1 and +1. Starting in the Nineties (Sokal et al. 1991, Livshits et al. 1991, Barbujani & Pilastro 1993) , this method has become very popular in evolutionary biology (see e.g. Bagley et al. (2016), Reem et al. (2016)), and has been used in the first analyses of linguistic and genetic variation based on the PCM (Longobardi et al. 2015).

Partial Mantel test

While the Mantel test only allows a comparison among two variables, a Partial Mantel test, proposed by Smouse and colleagues (Smouse et al. 1986), can be used to compare three or more variables. Essentially, the Partial Mantel test allows a comparison to be made among two variables while controlling for the third. This test is whidely used in evolutionary biology to evaluate the relationship between two variables, eliminating the effect of a third one. The method was first applied to ecological data by Legendre & Troussellier (1988). The Partial Mantel allows the identification of correlation between two variables that might be influenced by a third one. In population genetics, for instance, the first matrix may reflect genetic distances among populations, the other two matrices representing environmental factors such as linguistic distances in our case and geographic distances. Therefore, controlling for the effect of geography between populations (it means excluding the effect of demography from the observed genetic variation) we can quantify the correlation between the language and genetic distances, as if all pairs of populations were equally distant in space.

We performed the Mantel and Partial Mantel test through the `mantel` and `mantel.partial` functions available within the *vegan* R cran package (Legendre & Legendre 1998, Oksanen et al. 2016, R Core Team 2013) , empirically estimating the significance over 10 000 permutations.

Chapter 2

Results

The dataset described is a very broad compilation of mtDNA and Y-chromosome diversity. Since the main purpose of this study was to compare the two forms of genetic diversity in the same populations, we necessarily had to resort to hypervariable mtDNA region and to Y-chromosome STR markers. This way, we had available only a set of fast-mutating markers, therefore focusing on the most superficial layers of genetic diversity. In the course of the analysis, we came to realize that, for the large-scale comparisons we carried out, these data may not satisfactorily describe the differences between geographically extreme populations. On the other hand, the fact we could consider the same populations in the parallel analyses of maternally- and paternally-transmitted polymorphisms makes it possible unbiased comparisons of the results obtained. From the original dataset of 36 languages/populations, we removed Wolof (well-known outlier) and Inuit (extreme genetic outlier for mtDNA). I report below the main findings concerning the dataset at 34 populations (i.e. after the exclusion of Wolof and Inuit). That way we seek to reduce possible biases due to the presence of outlying groups and to focus on finer patterns of relationship.

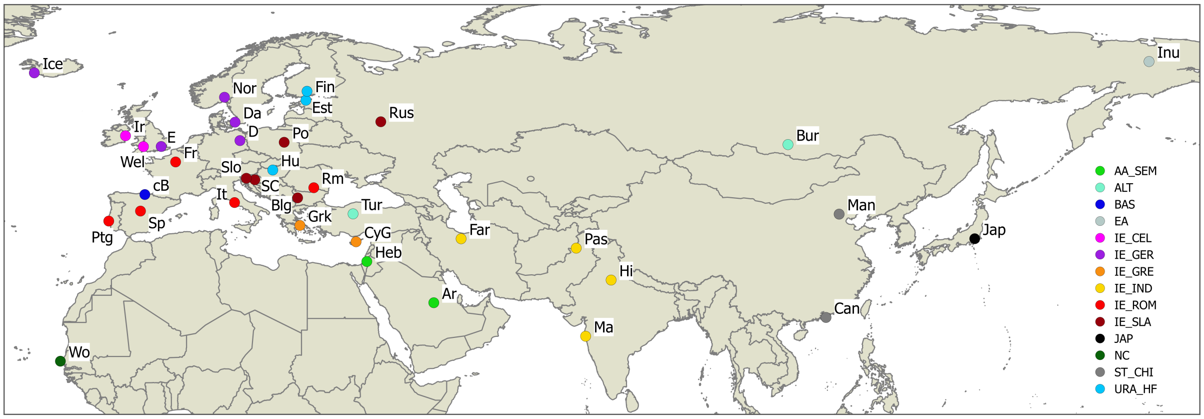

Figure 2.1: MtDNA and Ychr samples collected for 36 populations

The dataset includes 36 Old World populations, for each of which both mitochondrial and Y-chromosome information has been collected from public genomic resources. Colors in this scheme refer to one of the following language families or language isolates as listed in the legend: Afro-Asiatic_Semitic; ALT: Altaic; BAS: Basque; EA: Eskimo-Aleut; IE_CEL: IE_Celtic IE stand for Indo-European); IE_GER: IE_Germanic; IE_GRE: IE_Greek; IE_IND: IE_Indo-Iranian; IE_ROM: IE_Romance; IE_SLA: IE_Slavic; JAP: Japonic; NC: Niger-Congo; ST_CHI: Sino-Tibetan_Chinese; URA_HF: Uralic_Finno-Ugric. Minimum sample size=38, average sample size = 449 (mtDNA) and = 669 (Y-chromosome). Populations sampled detailed in Figure 2.1

population	Ychr	mtDNA		population	Ychr	mtDNA
cB	215	407		Fr	556	824
Hu	632	435		Ptg	1445	1088
Est	124	48		Rm	406	105
Fin	416	587		Grk	579	453
E	488	172		CyG	418	91
D	2063	667		Ma	162	215
Da	290	683		Hi	196	220
Ice	100	433		Far	79	465
Nor	72	343		Pas	611	230
Ir	949	319		Ar	1980	1892
Wel	118	92		Heb	38	233
Blg	96	883		Wo	74	59
SC	2058	154		Tur	60	46
Slo	102	233		Bur	215	473
Po	2200	817		Inu	287	142
Rus	944	522		Man	597	231
It	2231	1515		Can	510	205
Sp	1142	603		Jap	1643	277

Table 2.1: Populations sampled

Colours refer to the linguistic families and isolates as detailed in Figure 2.1

2.1 MDS results

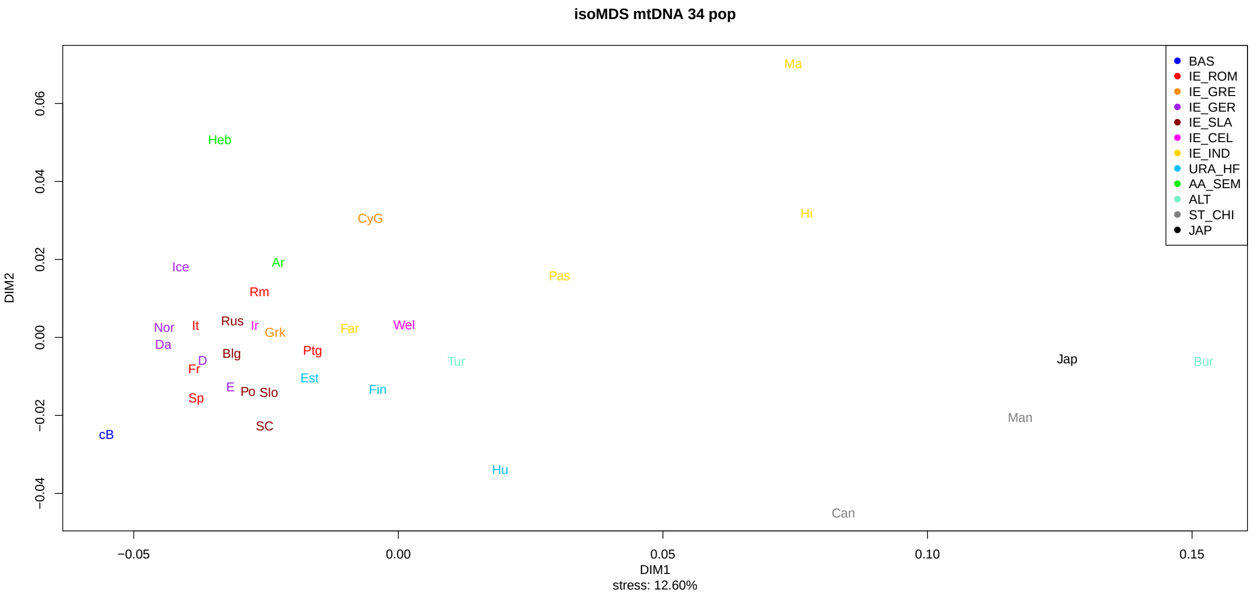

Figure 2.2: mtDNA multidimensional scaling, Inuit and Wolof removed (34 populations)

The levels of stress is not minimal, but acceptable (<20%). A continental pattern emerge, with Asian populations clustering on the right and Europe (not clearly structured) on the left. Farsi appears quite distant to their linguistic neighbors (Pashto and Hindi). Basques at the extreme left. Japan close, unlike with the Y-chromosome (see Fifure 2.3), to its geographical neighbors. Hungarians somewhat isolated from the other Europeans, towards the right. Cypriots closer to Semitic speakers than to their linguistic neighbors from Greece.
With Inuits and Wolof removed from the analyses, as clearly behaving as statistical outliers, Asian populations appear distinct from Europeans and Middle-Est Asians, so that a continental pattern emerge. In this plot, Buryats are shown as outliers, Japonic and Sino-Tibetan populations are also clustered on the left. Indo-Iranian populations on the middle are separated from European (cluster on right side) and Asians (left), with Marathi as outliers. Farsi are placed near the linguistically-related European populations rather than near their geographical neighbours, Basque are at extreme right and Cypriots are nearer to Hamito-Semitic than Greeks. The Hungarians' isolation from European populations is also evident.

The levels of stress not minimal, but acceptable (<20%). Some geographical structuring apparent, especially among European populations. Farsi closer to their geographical neighbors (Arab- and Turkish-speakers) rather than to their linguistic neighbors. Same for

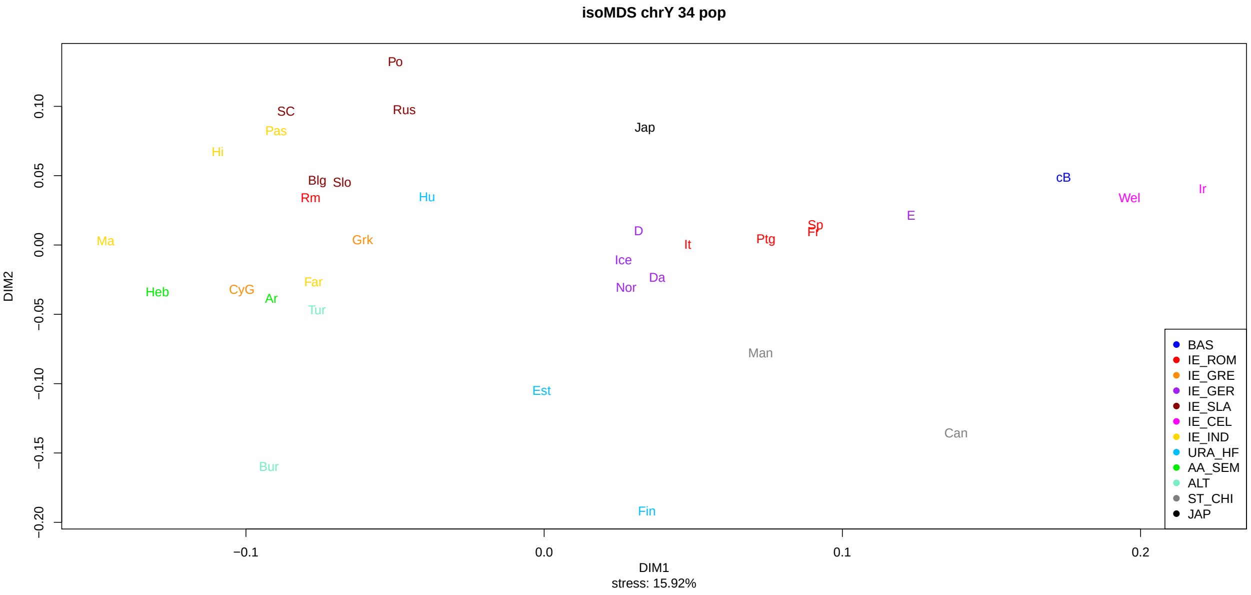

Figure 2.3: Y-chromosome multidimensional scaling, Inuit and Wolof removed (34 populations)

Hungarians and Romanians. Cypriot Greeks closer to Turkish and Semitic speakers than to their linguistic neighbors from Greece. Japan quite distant from the other East-Asian populations.
Buryats appear as outliers. Finno-Ugric populations are separated from their geographic neighbours. Finnish appear as outliers and their closest linguistic neighbors, the Estonians, are isolated from them. Once again, Hungarians are closer to their geographic neighbours but appears slightly isolated. Indo-Iranian speakers are placed to the extreme right, except for Marathi, the others are closer to European populations. Japan looks to be closer to Europe than to its geographic neighbours. Some geographic pattern is visible, especially among European populations.

Some structuring patterns appear for Y-chromosome especially among west Eurasian (and particularly European) populations: North Western Europeans (Germanic-speakers, Romance-speakers, Basques and Celtic) vs. South-Eastern Europe and Middle East (Slavic-speakers, Greek-speakers, Semitic and IE-Indian). Less evident is the differentiation of East-Asian populations (e.g. Japanese and Chinese languages). This could be probably due to the type of Y-chromosome markers considered (fast-evolving STRs; see above) which “limit” the differentiation at long-range level, while being more powerful in discriminating at narrower scales and especially within Europe. Interesting findings concern Farsi, Cypriot Greeks, Hungarians and Romanians: in all of these cases, linguistic differences do not seem to have reduced the genetic similarity with geographic neighbors speaking different languages.

2.2 Correlation test results

Geographic distances computed as GCD The two genetic uniparental markers show different levels of correlation with both language (Syn) and geography (GeoGCD) , with mtDNA (mtDNA) showing higher r correlation values (see Table 2.2). Both language-gene correlations decrease and became non-significant after removing the geographic component. Geographic-linguistic correlation became non-significant after removing the female genetic component, while remains still high and significant after removing the male genetic component.

A	mtDNA-Syn	r: 0.4598	p: 0.0001
	mtDNA-GeoGCD	r: 0.7901	p: 0.0001
	Ychr-Syn	r: 0.2791	p: 0.0017
	Ychr-GeoGCD	r: 0.3261	p: 0.0014
	Syn-GeoGCD	r: 0.5073	p: 0.0001

B	mtDNA-Syn/GeoGCD	r: 0.1117	p: 0.1261
	mtDNA-GeoGCD/Syn	r: 0.7276	p: 0.0001
	Ychr-Syn/GeoGCD	r: 0.1395	p: 0.0422
	Ychr-GeoGCD/Syn	r: 0.2230	p: 0.0080
	Syn-GeoGCD/mtDNA	r: 0.2645	p: 0.0026
	Syn-GeoGCD/Ychr	r: 0.4586	p: 0.0001

Table 2.2: A: Mantel tests (34 populations); B: Partial Mantel tests (34 populations). Geographic distances computed as GCD. Bonferroni correction (0.05/33 = 0.0015).

Geographic distances computed as LC The two genetic uniparental markers show different levels of correlation with both language (Syn) and geography (GeoGCD) , with mtDNA (mtDNA) showing higher r correlation values (see Table 2.3). Both language-gene correlations decrease and became non-significant after removing the geographic component. Geographic-linguistic correlation became non-significant after removing the female genetic component, while remains still high and significant after removing the male genetic component.

A	mtDNA-Syn	r: 0.4598	p: 0.0001
	mtDNA-GeoLC	r: 0.7646	p: 0.0001
	Ychr-Syn	r: 0.2791	p: 0.0017
	Ychr-GeoLC	r: 0.3089	p: 0.0023
	Syn-GeoLC	r: 0.4962	p: 0.0001

B	mtDNA-Syn/GeoLC	r: 0.1437	p: 0.0712
	mtDNA-GeoLC/Syn	r: 0.6958	p: 0.0001
	Ychr-Syn/GeoLC	r: 0.1523	p: 0.0296
	Ychr-GeoLC/Syn	r: 0.2044	p: 0.0177
	Syn-GeoLC/mtDNA	r: 0.2527	p: 0.0049
	Syn-GeoLC/Ychr	r: 0.4489	p: 0.0001

Table 2.3: A: Mantel tests (34 populations); B: Partial Mantel tests (34 populations). Geographic distances computed as LC. Bonferroni correction (0.05/33 = 0.0015).

Geographic distances computed as RE The two genetic uniparental markers show different levels of correlation with both language (Syn) and geography (GeoGCD) , with mtDNA (mtDNA) showing higher r correlation values (see Table 2.4). Both language-gene correlations remain significant after removing the geographic component. Geographic-linguistic correlation became non-significant after removing both female and male genetic components.

A	mtDNA-Syn	r: 0.4598	p: 0.0001
	mtDNA-GeoRE	r: 0.1730	p: 0.1291
	Ychr-Syn	r: 0.2791	p: 0.0017
	Ychr-GeoRE	r: 0.0034	p: 0.4127
	Syn-GeoLRE	r: 0.0458	p: 0.3194

B	mtDNA-Syn/GeoRE	r: 0.4593	p: 0.0001
	mtDNA-GeoRE/Syn	r: 0.1713	p: 0.1305
	Ychr-Syn/GeoRE	r: 0.2792	p: 0.0014
	Ychr-GeoRE/Syn	r: -0.0092	p: 0.4599
	Syn-GeoRE/mtDNA	r: -0.0385	p: 0.5535
	Syn-GeoRE/Ychr	r: 0.0466	p: 0.3208

Table 2.4: A: Mantel tests (34 populations); B: Partial Mantel tests (34 populations). Geographic distances computed as RE. Bonferroni correction (0.05/33 = 0.0015).

The observed significant correlations between the considered variables may have different meanings; indeed partial Mantel tests show that great part of the mtDNA-language correlation seems to be actually mediated by the geography. This is also confirmed by the still high and significant correlation between mtDNA and geography after keeping the effect of syntax constant. Despite a lower relationship with geographic, also the marginally significant correlation between syntax and Y-chromosome (once the effects of geography are removed) does not survive to Bonferroni correction for multiple testing.

2.3 Boundaries test results

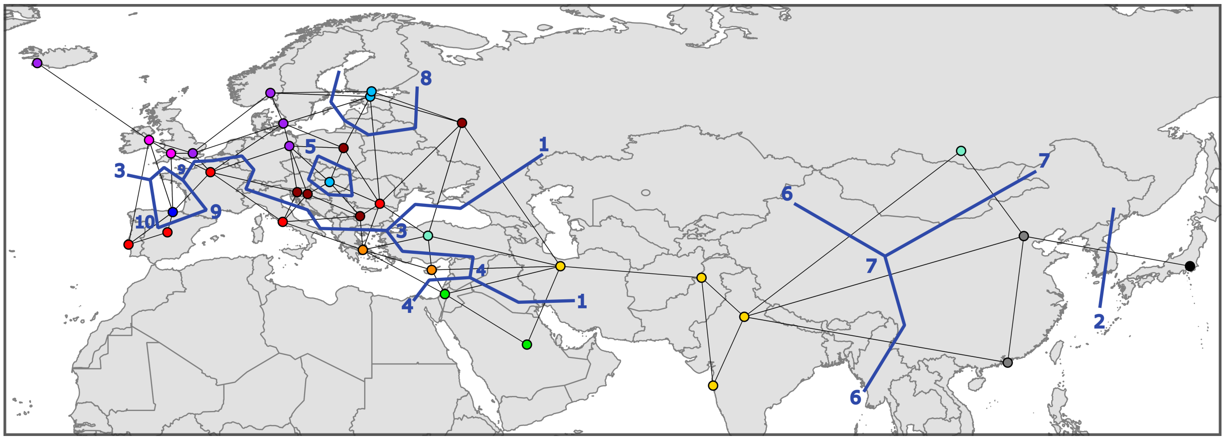

Figure 2.4: Boundaries inferred from syntactic diversity (34 populations)

Note boundary (1), separating East Asian languages, geographically non-European IE-speakers and Turkish, from geographically European IE-speakers and Semitic (the latter being then separated by boundary 4). Boundary (2) separates the Japanese, while (6) and (7) the two Chinese languages. Boundary (3) apparently separates Germanic and Slavic speakers from the bulk of Romance languages, the latter including Basques that subsequently becomes completely isolated when considering boundaries (9) and (10). Boundaries (5) and (8) distinguish the Ugro-Finnic languages, however highlighting the differentiation of Hungarians.

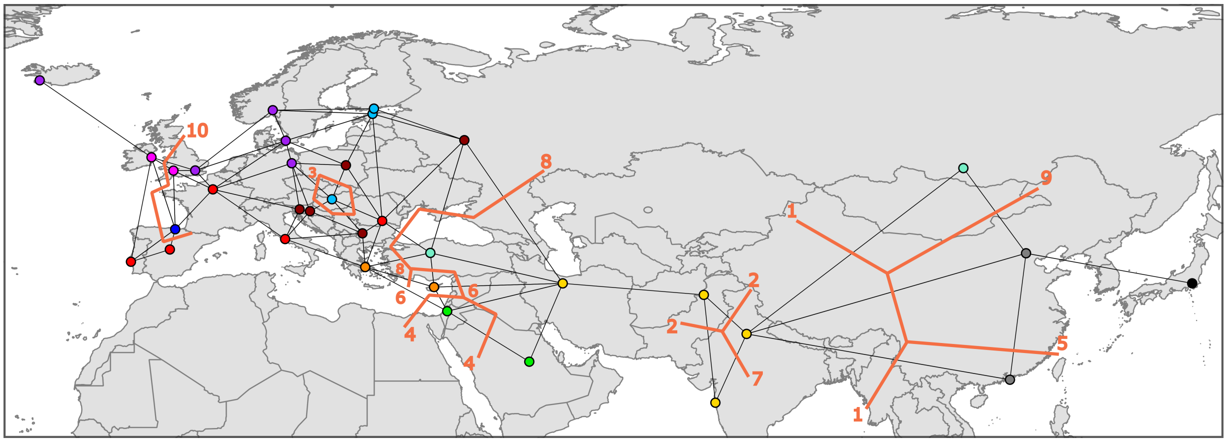

Figure 2.5: Genetic boundaries inferred from mtDNA diversity (34 populations)

Boundary (1) separates the four East Asian populations from West Eurasian ones, whereas Cantonese and Mandarin are further separated by boundary (5), and Buryat by boundary (9). Boundary (2) marks the isolation of Farsi- and Pashto-speakers from their Eastern IE-Indian neighbors (Hindi and Marathi), the latter being then separated from one another by boundary (7). Boundaries (3), (4) and (6) surround Hungarians, Hebrew and Greek Cypriots. Boundary (8) separates Turks from Europeans.

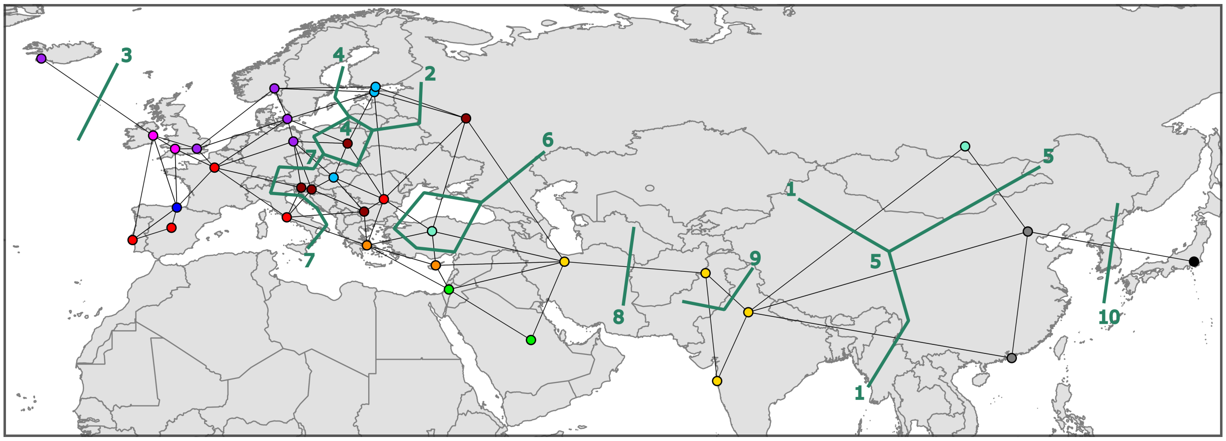

Figure 2.6: Genetic boundaries inferred from Y-chromosome diversity (34 populations)

Boundary (1) confirms what observed for mtDNA (i.e. East-Asia/West-Eurasia separation). Boundaries (2) and (4) isolates Ugro-Finnic speakers and curves to surround Poland. Boundary (3) isolates Iceland from the rest of Europe. Boundary (5) separates Buryats from Sino-Tibetan-speakers, while boundary (10) isolates Japanese. Boundary (6) completely isolates Turks. Boundary (7) crosses Europe with an East-vs-West pattern, roughly separating Germanic and Romance languages (except for Romanians) from Slavic and Hungarians. Boundary (8) and (9) separates Farsi and Pastho from the other IE-Indian languages.

Some remarkable similarities between the two uniparental genetic analyses appear. In both cases the first highest-ranking boundary (after the exclusion of Inuit and Wolof cases; see Supplementary) isolates East Asians from the bulk of the other considered populations. As for languages, the highest-ranking boundary (1) is shifted westwards, including also non-European IE-speakers and Turkish with East-Asian populations and separating them from IE-speakers from Europe and Semitic groups. Other features are different between markers. At the mtDNA level, Hungarians and Greek Cypriots are surrounded by specific boundaries, not completely confirmed for the Y-chromosome pattern, where boundaries surround the Ugro-Finnic (except Hungarians), Polish speakers, and Turks. If we add languages to the comparison, we find interesting correlations for the Finno-Ugric populations, but Hungarians are well-differentiated from all neighbors at the mtDNA level, Finns and Estonians at the Y-chromosome level. Farsi- and Pashto-speakers appear separated from their Eastern IE-Indian neighbors in both mtDNA and Y-chromosome, and even among each other in the latter paternal case, while we observe a linguistic “continuity” between all the IE-Indian languages. Some relationships between Farsi and Turkic or (partially) Semitic languages are observed also for the mtDNA and Y-chromosome patterns respectively.

Correlations between genetic and linguistic features reveal complex patterns that vary on the basis of the geographical area and the genetic markers considered. The observed complexity highlights the importance of phenomena of isolation and admixture, occurring at different time scales, sometimes influenced by the presence of language barriers, sometimes independent from them (see Hunley et al., PLoS Genetics 2008, Genetic and Linguistic Coevolution in Northern Island Melanesia).

Chapter 3

Discussion and conclusions

When comparing many variables through different set of tests, assumptions must be made with caution. Anthropological parameters have a great deal of variability; factors in our genome, in the environment, and in our cultural habits contribute to determining patterns of genome variation, along with other causes of influence, some of them probably still unknown. This analysis focused, strategically, on uniparental markers rather than the whole genomic information and considered linguistic diversity as the main cultural trait that could have, somehow, influenced genetic patterns. The idea is that linguistically-related populations are more likely to exchange mating partners than linguistically poorly-related, or unrelated populations. However, the abundance of different kind of behavioural traits in our species makes it difficult to choose just one for inferences. Even though language seems to be the most appropriate because of its greatest stability through time and has been proven correlated with genetic data (Cavalli-Sforza et al. 1988*b*, Sokal 1988, Barbujani & Sokal 1990, Barbujani & Pilastro 1993, Chen et al. 1995, Sajantila et al. 1995), there are many other cultural traits that have not been analyzed: to name just two, religion and national boundaries. That is to say that even when data show correlations with languages, it is still problematic to determine whether language itself, rather one of the many variables correlated with it, has actually been the cause of the correlation. Nevertheless the parameters used allowed us to draw some preliminary conclusions and highlight some correlations.

MDS analysis on mtDNA data shows Inuits as strong outliers causing all other populations to cluster together when jointly analyzing all 36 populations. The Inuits appears as outlier in Ychr plots aswell, although not as extreme as in the mtDNA analysis. That finding makes sense, considering that Inuits are genetically closer to American populations such as Amerind and Na-Dene, which have not been considered in this analysis, than to Asians (Bonatto & Salzano 1997). When zooming to 35 and 34 populations (also removing Wolof) to analyze the clustered populations in mtDNA MDS plot, a clearer pattern is shown, and language relationships seem to play a role in it. Farsi-speakers fall close to their geographically-distant linguistic relatives of Europe, and Basques and Hungarians stand out as genetic outliers. Japan, however, is close to its geographical neighbours. Further exceptions in the mtDNA pattern are the Cypriot-Greeks which are closer to their geographical neighbours than their linguistic family, a peculiarity also shown in Ychr plots. While Ychr plots with 36 and 35 populations are not so clear, smaller dataset show some geographic structure on the plot, highlighting geographical neighbours rather than linguistic ones, unlikely the mtDNA MDS.

Most of the pairwise and partial Mantel tests are nominally significant, showing that the three distance matrices are indeed related, as was to be expected. Correlations with geography are always significant, and those with language are stronger for mtDNA than for the Y-chromosome when 34 populations are considered. In both genetic-linguistic comparisons, however, the correlation decreases dramatically when the effects of geography are partialled out, so that the correlation of mtDNA with language is no longer greater than 0, and that of the Y-chromosome barely remains significant.

Looking at Barrier maps, some similarities are present between the two genetic analyses. In both cases the three first highest-ranking boundaries overlap, isolating the Inuits, the Wolof, and the East Asians from the other populations studied (two of these boundaries, the ones concerning Inuit and Wolof, are also high-ranking language boundaries, (1) and (3)). These boundaries can be easily compared with the MDS of the 36 populations (see Annex Figure A1) where both Inuits and Wolof appear as strong outliers, and East-Asian populations are distinguished from Western ones. Other than those, the patterns shown by mtDNA and the Y chromosome are clearly different, with Hungarians, Arabs, Cypriot Greeks and Marathi appearing strongly isolated for mtDNA but not for the Y chromosome, Turks, Poles and Finno-Ugric speakers showing the opposite pattern. Finns and Estonians are clustered both in the Ychr and in the linguistic analysis. When 30 populations have been analyzed, instead of 36 with strong outliers as reported here, boundaries appear isolating the Basque-population both in Ychr and in mtDNA, verifying the pattern of the MDS, as showed previously.

There is a relationship between genetic and linguistic diversity with geographic distances, as has been anticipated by recent studies: both markers and language distances increase at increasing geographic distances. Correlations do actually appear between parental markers and syntax diversity, meaning that languages have placed some barriers on demographic history. However, when comparing genetic and linguistic features, they reveal complex patterns that vary on the basis of the geographical area and the genetic markers considered. It seems clear that at this geographical scale, the existence of very widespread language families, geographically interspersed with one another, do not lead (and cannot possibly lead) to a simple and obvious relationship among the three sets of distances here considered. Buryat and Turkish belong to the same family, as do Icelandic and Farsi, but the several thousand km separating these populations in space appear to be a much stronger factor than language relatedness in determining their genetic similarities. Whereas at the European level the genetic patterns appeared to generally mirror patterns of linguistic variation (Longobardi et al. 2015), at the larger, Eurasian scale things are more complex, as is also the relationship between migrational movements affecting, respectively, the female and male components of the populations. The observed complexity highlights the importance of the phenomena of isolation and admixture, occurring at different time scales, sometimes influenced by the presence of language barriers, sometimes independent from them. Anyway, the complex patterns revealed show that maternal lineage diversity shows a correlation with geographical distances rather than linguistic higher than the paternal lineage. As has been proposed by several authors (Seielstad et al. 1998, Lippold et al. 2014) this might be due to a high rate of female migrations who tend to follow their partner to their native country, where their descendants would eventually learn local languages.

Part III

Sardinian project

“Are we being good ancestors?”

Jonas Salk

Aim and scope Several studies based on autosomal markers (Grimaldi et al. 2001, Battaggia et al. 2003, Falchi et al. 2004, Di Gaetano et al. 2014), mitochondrial DNA (Barbujani et al. 1995, Richards et al. 2000, Falchi et al. 2006, Caramelli et al. 2007, Ghirotto et al. 2010) and Y chromosome polymorphisms (Francalacci et al. 2003, Capelli et al. 2006, Contu et al. 2008, Francalacci et al. 2013) showed that the Sardinian population is one of the main European genetic outliers (Cavalli-Sforza & Piazza 1993, Quintana-Murci et al. 2003, Pugliatti et al. 2006, Sidore et al. 2015) and reported unusually high levels of internal diversity (Barbujani & Sokal 1991, Zei et al. 2003). Most of these studies compared variation in Sardinia and in other European populations, but there is still uncertainty about past population dynamics and demographic processes within the island, as well as about the exact nature and the extent of the genetic exchanges that occurred over millennia, actually determining the existing Sardinian genetic structure.

Controversy has also surrounded the origins and the antiquity of the colonization of Sardinia. The earliest presence of humans is still under debate. Some authors likely date it back to the end of the Middle Pleistocene, on the base of lithic artifacts typology, attributed to the Lower Paleolithic (Fenu et al. 1999, Martini & Ulzega 1989-1990, Martini 1999). Nonetheless, neither human remains nor absolute-dated contextual evidence support this hypothesis. Clues of human settlements arose only from the end of the Upper Pleistocene (Sondaar et al. 1993, 1995). The first evidence of Holocene frequentation of the island are scattered in a few rockshelters and caves, exclusively on the inside of a 20 km coast belt (Lugliè 2009).

With the advent of the agriculture, the population of the island increased in size, as demonstrated by the density of Early Neolithic (EN) sites (VI millennium BCE), and at the beginning of the IV millennium BCE there has been a rapid growth of archaeological documentation and skeletal remains (Floris 1981, 1983, Germanà 1995, Sanna et al. 1999). The fragmented anthropological and archaeological evidence of the Pre-Neolithic phase, make it difficult to properly describe a continuity towards the process of Neolithization in Sardinia.

From a genetic perspective, a recent genomic study of both ancient and modern Europeans (Sikora et al. 2014), including data from more than 400 modern-day Sardinians, revealed the existence of genetic affinities between Neolithic Europeans samples and modern Sardinians. These results suggest that Sardinians are a "modern-day snapshot of the genetic structure of the first farmers associated with the spread of agriculture in Europe".

This hypothesis has not been supported so far by evidence coming from ancient Sardinian genetic data, due to the paucity of Pre-Neolithic and actual absence of EN human remains. The only ancient data ever published were sequences of the mtDNA control region from

Bronze-Age sample (Caramelli et al. 2007) and revealed a directed genealogical continuity between Nuragic individuals and the current people of Ogliastra, but not of Gallura (Ghirotto et al. 2010).
Although the general European picture is getting clearer, many aspects of the Neolithic transition in Sardinia are still poorly understood, starting from whether, and to what extent, gene flow from mainland Europe during the time of the spread of agriculture actually contributed in shaping the genetic makeup of the island.

In this section, I am describing the methods and results that come up beside the analysis conducted and detailed in Modi et al. (2017) (see *Manuscripts* section). The analysis described allowed us to understand the Pre-Neolithic genetic relationships among the two Mesolithic Sardinians and the samples collected (detail in Matherials and Methods section below) .

Chapter 1

Sardinian samples archeo-genetics

1.1 Archaeological Site and samples description

The Su Carroppu site plays a relevant role in Sardinia, with a remarkably rich archaeological record and a series of occupational phases spanning from the Mesolithic to the historical period.

The Su Carroppu rockshelter opens onto the Paleozoic massif of the Sulcis region, around 12 km as the crow flies from the present-day South-west coast of Sardinia. Among the currently known Sardinian Early Neolithic (EN) settlements, the Su Carroppu site is one of the southernmost and farthest from the interior. This site has been the first unequivocally recognized as EN at the end of the 1960s, thus becoming for a long time representative of the early stages of the spread of the impressed ware to the Tyrrhenian basin, and eponymous of the corresponding archaeological facies. The 1978 investigation divided the site into two distinct units, called sector A and B. Sector B – covering a surface of only 6 square meters – yielded a stratigraphic sequence representative of the entire site, roughly subdivided into four levels 1 (Figure 1.1).

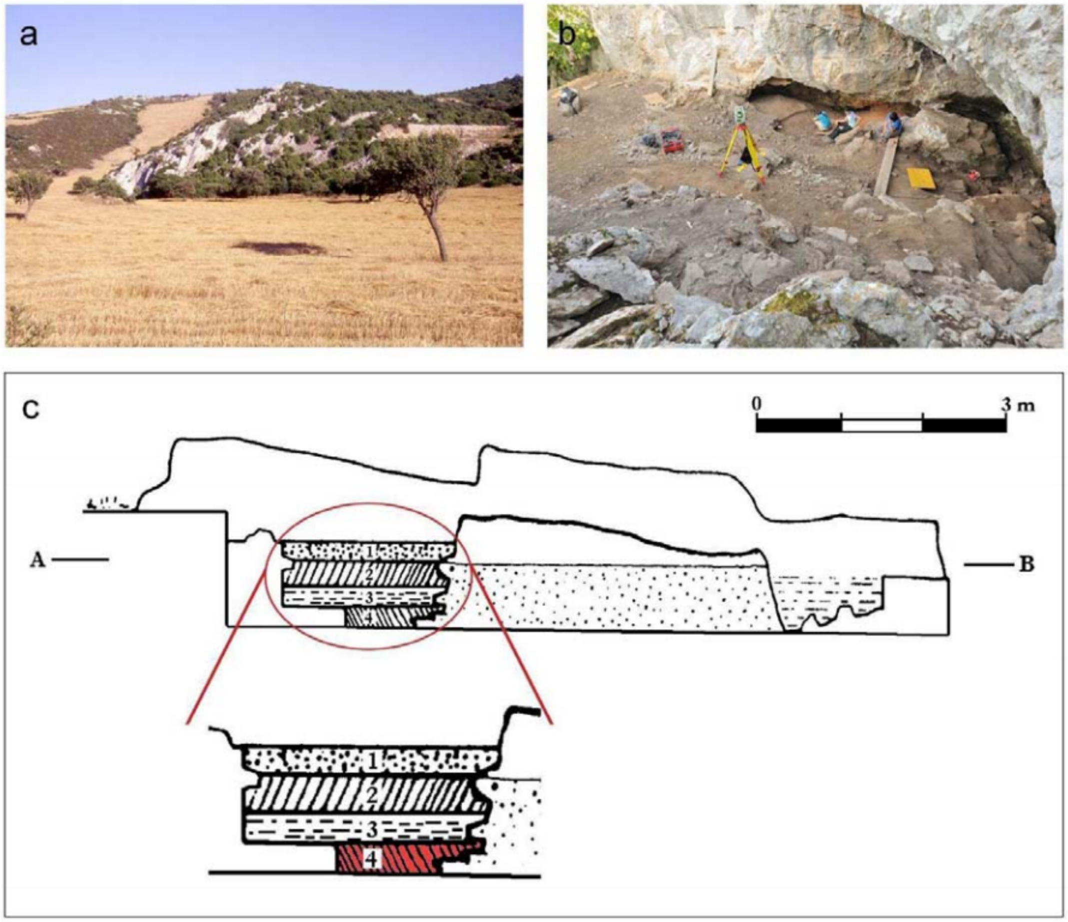

Figure 1.1: The sampling location of analyzed remains.
a: Su Carroppu site; b: archaeological site during the excavation season; c: stratigraphic profile in the sector B of the excavation. The samples analyzed in this study were retrieved in the lowermost level, horizon 4 (in red).

Levels 1 to 3, strongly perturbed by anthropic activities and burrowing animals, confirmed the presence of occupational phases spanning from EN to Bronze Age. The lowermost level, horizon 4, consisted of dark anthropogenic soil, rich in charcoal. This level, between 1.40 and 1.55 mt depth, apparently was preserved undisturbed: basing on the exclusive presence of Cardial Pottery it has been assigned to the EN and yielded large quantities of material remains and bones. The excavations carried out since 2009 and still in progress are bringing new data to the debate on the chronology and the nature of the ancient human settlement of the site, that had a long term dwelling for at least eight millennia. The revision of older excavations carried out in the Northern sector allowed to identify Mesolithic burials of the IX-VIII millennia cal BCE, suggesting an early settlement before the Neolithic age; furthermore, new analysis of archaeological and anthropological materials from level-4 confirm the presence of a rather ancient Mesolithic frequentation of the island. These new data suggest that Su Carroppu is one of the site which shows the earliest trace of human migrations and frequentations of Sardinia.

1.1.1 Analyzed samples

The skeletal remains studied here were brought to light from the base of the undisturbed level-4 during the 1978 excavation season, and were found intermingled and in concretion with bones of *Prolagus sardus* (the endemic Sardinian pika). Fragments of the ribs as well as from the left and right upper and lower limbs (ulna, radius, humerus, femur, fibula and tibia) have been identified, and some of them were covered by calcareous concretions. Three direct radiocarbon dates, performed on the human bones of ulna, tibia and humerus labeled CAR-H3, CAR-H7 and CAR-H8 respectively, placed the remains in the mid-9th millennium cal. BCE thus showing an unexpected Early Mesolithic settlement predating EN occupation (Gassin & Lugliè 2012, Luglié 2014)(see Table A2).

Chapter 2

Materials and Methods

To study the two Mesolithic Sardinian samples in the context of the ancient genetic diversity, I have analyzed 49 already published pre-Neolithic sequences, coming from all over Europe. To highlight a potential temporal pattern of similarities the samples were classified in four chronologically-based groups: pre-LGM (ranging from 45 to 25 kya), post-LGM (ranging from 19.5 to 14.5 kya), Late Glacial (ranging from 14.5 to 11.5 kya) and Holocene (ranging from 11.5 to 7 kya). The sequences used are listed in Table A3. The sequences collected were analyzed whit the NGS pipeline described in the Modi et al. (2017) Supplementary Information and following the same criteria and filters applied to the Mesolithic Sardinian samples.

Ancient genomes processing: an *ad hoc* NGS pipeline

The technological advance of next-generation sequencing has helped research not only to achieve better insights into the human past, but also to overcome the problem of contamination of aDNA (Hagelberg et al. 2015). In order to uniform the data used and to optimize the quality of eah sample, both the two Mesolithic Sardinian and the samples collected, were analyzed by an *ad hoc* NGS pipeline detailed below.

I personally took care of the pipeline scripting phase that allowed the automatization process and the specific filtering setting steps.

Raw reads were processed with SeqPrep (John 2011) . Adaptor sequences were trimmed and paired-end reads were merged into single sequences with a minimum overlap of 10 bp in order to exclude all the sequences derived from molecules longer than 140 bp. Only reads with a minimum length of 30 bp were kept. Filtered reads were mapped to the revised Cambridge Reference Sequence (rCRS) using BWA-0.6.2 (Li & Durbin 2010). To improve the mapping efficiency in ancient molecules, we deactivated the seeding and we allowed more substitutions and up to two gaps (instead of 1) setting "-l 1000 -n 0.01 -o 2"10. PCR duplicates were removed using PicardTools-1.98 (http://picard.sourceforge.net.). The mapped reads were filtered based on mapping quality 30 using SAMtools-0.1.19 (Li et al. 2009) and to align to unique positions along the reference sequence. RealignerTargetCreator and IndelRealigner from the suite of tools Genome Analysis Tool Kit (GATK v. 3.0) (McKenna et al. 2010) were used to identify regions that contain an implausibly large number of differences to the reference genome and then to realign sequences in these regions. Consensus sequences for the mitochondrial genomes of all samples were called using mpileup and vcfutils.pl of the SAMtools-0.1.19 package (Li & Durbin 2010) . Only the reads with a minimum mapping quality of 30 were used to call confident bases for these consensus sequences. Finally, we obtained an average coverage of mtDNA of 4.83 for CAR-H3, 14.14 for CAR-H7 and 19.98 for CAR-H8 with average fragment length of 72.74 for CAR-H3, 62.59 for CAR-H7 and 53.09 for CAR-H8. Due to the low coverage obtained, sample CAR-H3 was excluded to further analysis. To avoid miscalling, all the polymorphic positions reported in the vcf output file were subsequently visually inspected. The assembly to the reference was masked taking into account all the positions covered by less than three reads. We applied the IUPAC code, when the concordance across reads was lower than 70%. The haplotype assignment was based on Haplogrep (van Oven & Kayser 2009, Kloss-Brandstatter et al. 2011).

Discriminant Analysis of Principal Component

One of the most widely applied approaches to study the genetics of populations, is the inference of population structuring with Bayesian clustering methods based on explicit population genetics models. The advantages to extracting meaningful information from genetic data come together with the computation and time cost of this kind of analyisis, especially when analyzing large datasets. Alternative to Bayesian clustering algorithms, multivariate analyses are able to identify genetic structures in very large datasets within negligible computational time, and with any assumption about the underlying population genetic model.

The Discriminant Analysis of Principal Components (DAPC) (Jombart et al. 2010) is a multivariate method, based on the Discriminant Analysis (DA) approach (Lachenbruch & Goldstein 1979), designed to identify and describe clusters of genetically related individuals. DAPC attempts to summarize the genetic differentiation between groups, while overlooking within-group variation. The method achieves the best discrimination of individuals into pre-defined groups. Further, this method allows for a probabilistic assignment of individuals to each group, as in Bayesian clustering methods. DAPC relies on data transformation using PCA (Principal Component Analysis) (Hotelling 1933) as a prior step to DA, which ensures that variables submitted to DA are perfectly uncorrelated. Along with the assignment of individuals to clusters, DAPC provides a visual assessment of between-population genetic structures, permitting to infer complex patterns such as hierarchical clustering or clines.

The analysis here described were run the with the *dapc* function within the *adegenet* R package. Ten principal component were retained and the discriminant functions were calculated for the four "a priori" groups (i.e. pre-LGM, post-LGM, Late Glacial, Holocene), corresponding to the four historical epochs in which was subdivided the Pre-Neolithic dataset. Secondly, the two Sardinian sequences were superimposed to the discriminant functions thus estimated. Then, the first two discriminant functions were plotted with the *scatter* function of the *adegenet* R package.

Pairwise distances

To highlight a potential temporal pattern of similarities, the average number of pairwise differences between each of the two Sardinian sequences and the 49 ancient samples collected from literature were estimated with the Arlequin software (Excoffier & Lischer 2010) and plotted with Qgis (http://www.qgis.org/it/site/). The pairwise differences between each of the two Sardinian sequences and all the other ancient Eurasian sequences collected, were estimated applying Kimura 2P as sequences distance method (Kimura 1980) and weighting three times the transversions over the transitions. The pairwise differences were plotted in four maps, according to the chronologically-based groups described above, with blue (CAR-H7) and orange (CAR-H8) circles located in the geographical origin of each sample and with color lightness proportional to the level of affinity between the Mesolithic Sardinians and the other ancient sequences sampled.

Chapter 3

Results

3.1 Pairwise Distance in Eurasian Ancient Samples

To better understand the genetic relationships among the two Mesolithic Sardinian and the Pre-Neolithic sequences collected (see Annex section), we calculated the average number of differences (pd hereafter) between each of the two ancient samples from Sardinia and each ancient Eurasian sample. We considered the Pre-Neolithic dataset as subdivided into four historical periods (see Methods for details). The results (shown in Figure 3.1) confirm the pattern showed by the network analysis described in Modi et al. (2017). On average, the greatest degree of resemblance is between the Mesolithics from Sardinia and the pre-LGM group (34 pairwise differences (pd) for CAR-H7 and 28 pd for CAR-H8), followed by the Holocene and the post-LGM (35 pd for CAR-H7 and 32 pd for CAR-H8) and the Late Glacial period (37 pd for CAR-H7 and 35 pd for CAR-H8). Within the pre-LGM group, the lowest level of pairwise differences (ranging between 27 and 30) includes samples from the Czech Republic (DoVe14, DoVe16 and DoVe43), from Belgium (GyQ56-16, GyQ116-1, Gy2878, GyQ53-1), an Italian (Fuma2), a Russian (Kost14) and a Siberian (UstIhm) sample. The lowest level of similarity is between Mesolithic individuals in Sardinia and the Georgian Satsurblia (Satsbl) sample (42 pd for CAR-H7 and 40 pd for CAR-H8) falling within the Late Glacial group. This pattern is confirmed by calculating the pairwise phiST distances between Sardinian Mesolithics and each historical group of sequences (Table 3.1). The Sardinians appear more similar to the pre-LGM samples than to the other historical groups (Table A3), including the Holocene, that is the group in which they would temporally belong to.

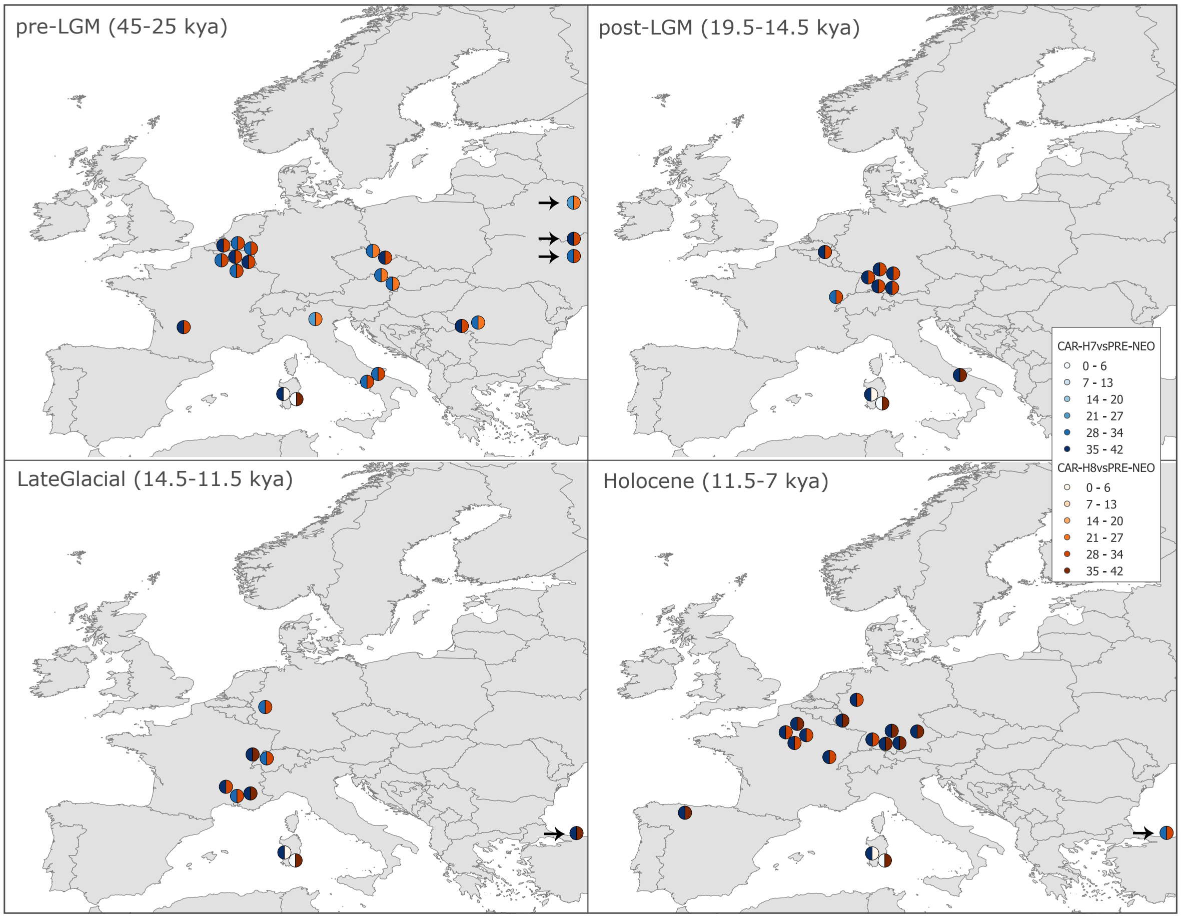

Figure 3.1: Pairwise differences between the two Mesolithic Sardinians and each PreNeolithic Eurasian sample.

The results are shown separately for the four historical epochs. Blue and orange circles correspond to CAR-H7 and CAR-H8, respectively. The colour lightness is proportional to the level of similarity between the Mesolithic Sardinians and the other ancient samples. Each map (Countries version 2.0.0) was downloaded from Natural Earth (naturalearthdata.com) and modified with QGIS (http://www.qgis.org/).

Time period	Sardinians	p-values
pre-LGM	0.362	0.0189
post-LGM	0.563	0.0237
Late Glacial	0.490	0.0271
Holocene	0.545	0.008

Table 3.1: Pairwise phiST distances between Mesolithic Sardinians and each historical group of ancient. Threshold fixed to p<0.05

3.2 Discriminant Analysis of Principal Component

To further confirm the peculiar similarity of the two Sardinian Mesolithics with the pre-LGM samples, a Discriminant Analysis of Principal Component (DAPC) was performed.

We first calculated the discriminant functions for the PreNeolithic samples grouped "a priori" in their respective time periods (pre- and post- LGM, Late Glacial and Holocene), thus obtaining three well-differentiated clusters, the pre-LGM on the left, the post-LGM on the top, and the Holocene/Late Glacial on the right. Only a few sequences do not fall within their respective clusters (Figure 3.2). We then superimposed to the discriminant functions thus estimated the two Sardinian sequences, which fell undoubtedly within the range of variation mostly driven by pre-LGM samples, thus confirming the low degree of similarity between the Mesolithics from Sardinia and the coeval Eurasian sequences considered (Figure 3.2). Within the two Sardinian sequences, CAR-H8 showed the higher affinity with the pre-LGM group.

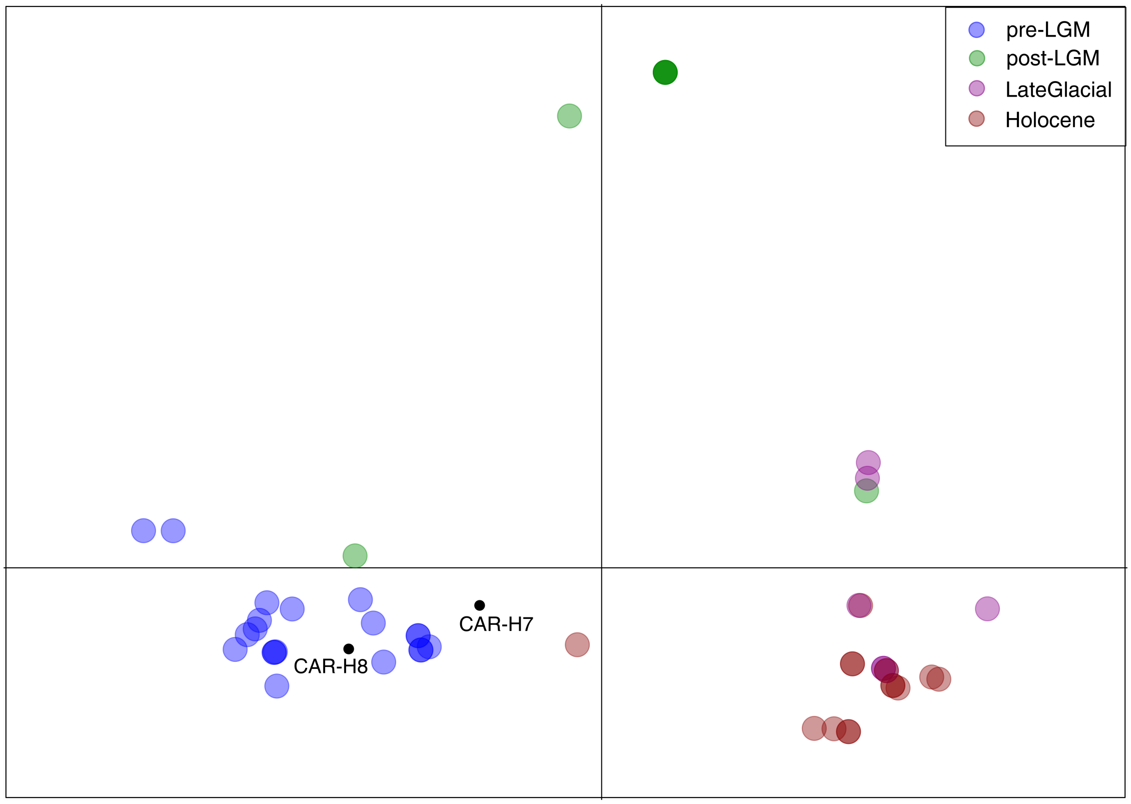

Figure 3.2: Discriminant Analysis of Principal Component
Scatterplot of the first two discriminant functions for the mitochondrial genomes in the Pre-Neolithic dataset (Supplementary Table S5). We used the four historical periods to identify prior groups. The two superimposed Mesolithic Sardinians are labelled in black. Darker spots correspond to overlapping samples.

Chapter 4

Discussion and conclusions

Archaeological evidence suggests the first human presence in Sardinia around 20,000 years ago (Spoor 1999), with sporadic and discontinuous occupations during Paleolithic and Mesolithic ages. Nowadays, Sardinians form a distinct outlier within the genetic variation of modern Europeans (Cavalli-Sforza & Piazza 1993), often interpreted as a consequence of thousands years of genetic isolation and drift, but little is known about the demographic changes that could have shaped the observed pattern of genetic variation. The, so far limited, ancient Sardinia genetic data allowed us to highlight a complete genetic continuity within a specific region of the island, Ogliastra, since the Bronze-Age (Ghirotto et al. 2010); however, cranial morphological evidence suggests that this continuity may have been established since Neolithic times, and possibly earlier (D'Amore et al. 2010).

The two Mesolithic sequences retrieved in the Su Carroppu archaeological site represents the oldest sample of DNA in Sardinia, thus providing a direct genetic evidence about the first colonizers of the island. The samples were treated following all the golden criteria before DNA extraction and sequencing to avoid any contamination. The comparable mitochondrial DNA data from European late hunter-gatherers have shown a remarkable genetic uniformity among pre-Neolithic populations, with most of the sequences (˜83 %) belonging to the haplogroup U, of which a majority carry U5 haplotypes (> 65%) (Fu et al. 2012). Neither Sardinian sequence belongs to any of the U haplotypes, documenting the presence of substantial genetic differences over the Mediterranean area. In addition, neither sequence has been observed in later, ancient or contemporary, individuals, and both belong to haplogroups and subhaplogroups now present in Europe at low (J, <16%) or very low (I, <7%) frequencies, and that are rare in modern Sardinia. Based on complete mitochondrial genomes, Posth et al. (2016) described a higher genetic diversity in pre-LGM than in post-LGM European populations and identified a major turnover around 14,000 yBP, with the subsequent expansion of haplogroup U that became widespread all around Europe until the Neolithic transition. We do not find haplogroup U in Sardinia by 11,000 yBP, which means a different impact of the LGM in the island, and probably a high isolation of the Sardinian population, not only from Neolithic times onwards (as genomes data seems to have probed), but also from former times considering the dates of our samples.

The phylogenetic network analysis of all the Pre-Neolithic complete mitochondrial sequences so far generated, actually confirmed this view (see the attached paper manuscript). The majority of the Late Pleistocene and Early Holocene sequences belong to the U lineage, and form a quite homogeneous cluster at the bottom of the network. The two Mesolithic samples from Sardinia are highly differentiated, departing from the network through long branches, so as to indicate mutations possibly arising along thousand years of geographic (and genetic) isolation. Considering the sequences' time scale, the two Mesolithic individuals fell within the variation driven by the pre-LGM haplogroups, and not together with Holocene sequences, as would have been expected. Consistent results were obtained also in the Pairwise Difference and in the DAPC analyses, with the Sardinian Mesolithic samples analyzed showing, on average, a higher degree of resemblance with preLGM sequences. This particular pattern appears to be compatible with an ancient population process, in which the first settlers of Sardinia were in rather close genealogical continuity with pre-LGM Europeans. The genome-wide data of Ice Age hunter-gatherers have shown that prehistoric Europe was characterized by recurrent populations turnover and migrations (Sanchez-Quinto et al. 2012), which resulted in a genetic homogeneity across pre-Neolithic populations. Our results support the view that these ancient populations movements did not involve genetic exchanges with Sardinians: isolation and drift may have resulted in a substantial mitochondrial differentiation between them and other Europeans.

The role and the genetic impact of migrations in Sardinia from continental Europe has been under debate for years, with particular interest on whether, and to what extent, the gene flow from the mainland during the time of the spread of agriculture in Europe contributed to shaping the present Sardinian gene pool (Ghirotto et al. 2010, Contu et al. 2008, Sikora et al. 2014). We also explicitly compared demographic models through Approximate Bayesian Computation. The question thus addressed was not whether the two Su Carroppu Mesolithic individuals are ancestral to current Sardinians along the maternal line (of course, they are not), but rather what was the posterior probability that a population of size 100-10,000 individuals (the broad interval of priors considered), and comprising the Su Carroppu individuals, may have contributed to the current gene pool. Because the alternative to genealogical continuity since Mesolithic times is immigration from the mainland, Middle and Early Neolithic sequences from Central Europe were included as a source of Neolithic migrants into the island (Haak et al. 2015). A model of genealogical continuity in Sardinia since Mesolithic times appeared very unlikely. We could not discriminate between a model assuming a certain degree of admixture and one of complete replacement by Neolithic immigrants, but if admixture occurred the contribution of Mesolithic people was apparently very limited.

We assessed the quality of these results by a number of tests. First, we evaluated exactly the probability to obtain false positives in the estimation of models' posterior probabilities, with the discontinuity model having the lowest type one error. Then, we showed in various ways (posterior predictive p-value and PCA analysis) that the discontinuity model can in fact reasonably reproduce the observed variation. Clearly, a certain degree of uncertainty necessarily affects any analysis, particularly when it is based on a single DNA region and on the necessarily small samples in which ancient DNA is usually typed. Within these unavoidable limits, we believe that we can be confident about our results.

When explicitly estimating the Neolithic admixture proportion, i.e. the amount of Neolithic genes from continental Europe that gave rise to the current Sardinian genetic pool (admixture_tot model), we obtained values of 0.8-0.9%, depending on the point estimates considered. This means that a significant proportion of modern Sardinian mitochondrial variation would came not from its first settlers, but from a subsequent migration wave from the continent. These results do not depend at any rate from the samples we chose to represent Neolithic Europe; using a set of Neolithic individuals randomly chosen from the whole dataset of Neolithic sequences available from the literature changed only marginally the posterior probabilities we estimated (data not shown). In our model comparison, we fixed the time of this second migration to 6,000 years ago, thus compatible with the archaeological evidence of Neolithic expansion in Sardinia. The spread of agriculture in Sardinia would hence been associated with demic diffusion from the continent, resulting in a large-scale population replacement. These results, for the first time supported by ancient genetic data, are also in good agreement with archaeological evidence and with what emerged from the comparison of modern Sardinian genomic data and Neolithic and Paleolithic sequences, interpreted by the authors as evidence of gene flow from mainland Europe during the time of the spread of agriculture in Europe. Sikora et al. (2014) also envisaged a genetic continuity until present times, but did not provide quantitative measures of it. Another possibility, compatible with our results, would be that Sardinian Paleo-Mesolithic males, but not females, admixed with immigrants from Neolithic Europe. This, however, would mean that in Sardinia the spread of the Neolithic culture was mainly carried out by women, in contrast with the available evidence (Rasteiro & Chikhi 2013).

Moreover, this view is also in contrast with studies of sex-biased admixture in modern communities, suggesting that the invading population tends to incorporate female residents more than males (Abe-Sandes et al. 2004, Gonzalez-Andrade et al. 2007, Goncalves et al. 2008, Stefflova et al. 2009, Quintana-Murci et al. 2010).

In conclusion, this study, albeit limited to DNA transmitted along the female lines of descent, provides the first genetic evidence on the earliest inhabitants of Sardinia, showing that they bore relationships with people who elsewhere are documented in much earlier time periods, i.e. before the Last Glacial Maximum. Formal comparison of alternative demographic models suggests that the Neolithization of the island was not a local development, but was associated with the arrival of a genetically-distinct group of immigrants from continental Europe.

The results of this study led to the publication on the paper attached in the *Manuscripts* section of this thesis.

Annex

Table A1: mtDNA HVR1 references per population and author

Language	**References**	**n samples**
Arabic	Al Balwi (unpublished)	1
	Al-Zahery, et al., 2011	315
	Badro, et al., 2013	1213
	Haber, et al., 2012	363
Bulgarian	Calafell, et al., 1996	30
	Karachanak, et al., 2012	853
Buriat	Derenko, et al., 2003	91
	Derenko, et al., 2007	295
	Gibert, et al., 2010	61
	Ingmann, et al., 2000	1
	Starikovskaya, et al., 2005	25
Cantonese	Chen, et al., 2008	106
	Kivisild, et al., 2002	69
	Yao, et al., 2002	30
central Basque	Bertranpetit, et al., 1995	45
	Cardoso, et al., 2012	34
	Cardoso, et al., 2013	210
	García, et al., 2011	115
	Prieto, et al., 2011	3
Cypriot Greek	Irwin, et al., 2008	91
Danish	Jobling et al. (personal communication)	20
	Sørensen, et al., 2010	201
	Raule, et al., 2014	429
	Richards, et al., 1996	33
English	García, et al., 2011	9
	Helgason, et al., 2001	142
	Ingmann, et al., 2000	1
	Jobling et al. (personal communication)	20
Estonian	Sajantila, et al., 1995	28
	Sajantila, et al., 1996	20

Language	**References**	**n samples**
Farsi	Metspalu, et al., 2004	435
	Schönberg, et al., 2011	30
Finish	Finnilä, et al. 2001	192
	Hedman, et al., 2007	200
	Raule, et al., 2014	146
	Sajantila, et al., 1995	49
French	Badro, et al., 2013	790
	García, et al., 2011	33
	Ingmann, et al., 2000	1
German	García, et al., 2011	11
	Hofmann, et al., 1997	67
	Jobling et al. (personal communication)	20
	Lutz, et al., 1998	200
	Richards, et al., 1996	156
	Tetzlaff, et al., 2007	213
Greek	Irwin, et al., 2008	317
	Kouvatsi, et al., 2001	54
	Jobling et al. (personal communication)	20
	Raule, et al., 2014	14
	Vernesi, et al., 2001	48
Hebrew	Behar, et al., 2008	233
Hindi	Barnabas, et al. 2005	9
	Kivisild, et al., 1999	68
	Sharma, et al., 2012	143
Hungarian	Irwin, et al., 2007	415
	Jobling et al. (personal communication)	20
Icelandic	Helgason, et al. 2000	394
	Sajantila, et al., 1995	39
Inuit	Helgason, et al., 2006	96
	Simonson (unpublished)	46
Irish	Jobling et al. (personal communication)	20
	McEvoy, et al., 2004	299
Italian	Achilli, et al., 2007	321
	Boattini, et al., 2013	600
	Brisighelli, et al., 2012	352
	Falchi, et al., 2006	61
	Ingmann, et al., 2000	1
	Jobling et al. (personal communication)	20
	Mogentale-Profizi, et al., 2001	68
	Ottoni, et al., 2009	92

Language	**References**	**n samples**
Japanese	Horai, et al., 1996	62
	Ingmann, et al., 2000	2
	Mabuchi, et al., 2007	124
	Oota, et al., 2002	89
Mandarin	Oota, et al., 2002	85
	Yao, et al., 2002	146
Marathi	Barnabas, et al., 2010	30
	Thangaraj, et al., 2010	185
Norwegian	Helgason, et al., 2001	323
	Jobling et al. (personal communication)	20
Pashto	Rakha, et al., 2011	230
Polish	Grzybowski, et al., 2007	413
	Mielnik-Sikorska, et al., 2013	404
Portuguese	González, et al., 2003	299
	Pereira, et al., 2004	549
	Prieto, et al., 2011	240
Romanian	Bosch, et al., 2006	105
Russian	Grzybowski, et al., 2007	157
	Morozova, et al., 2012	365
Serbo-Croat	Babalini, et al., 2005	96
	Fu, et al., 2012	38
	Jobling et al. (personal communication)	20
Slovenian	Malyarchuk, et al., 2003	104
	Pajnič, et al., 2004	129
Spanish	Falchi, et al., 2006	66
	García, et al., 2011	5
	Larruga, et al., 2001	196
	Jobling et al. (personal communication)	20
	Prieto, et al., 2011	316
Turkish	Di Benedetto, et al., 2001	17
	Schönberg, et al., 2011	29
Welsh	Richards, et al., 1996	92
Wolof	Ennalaa, et al., 2009	11
	Rando, et al., 1998	48

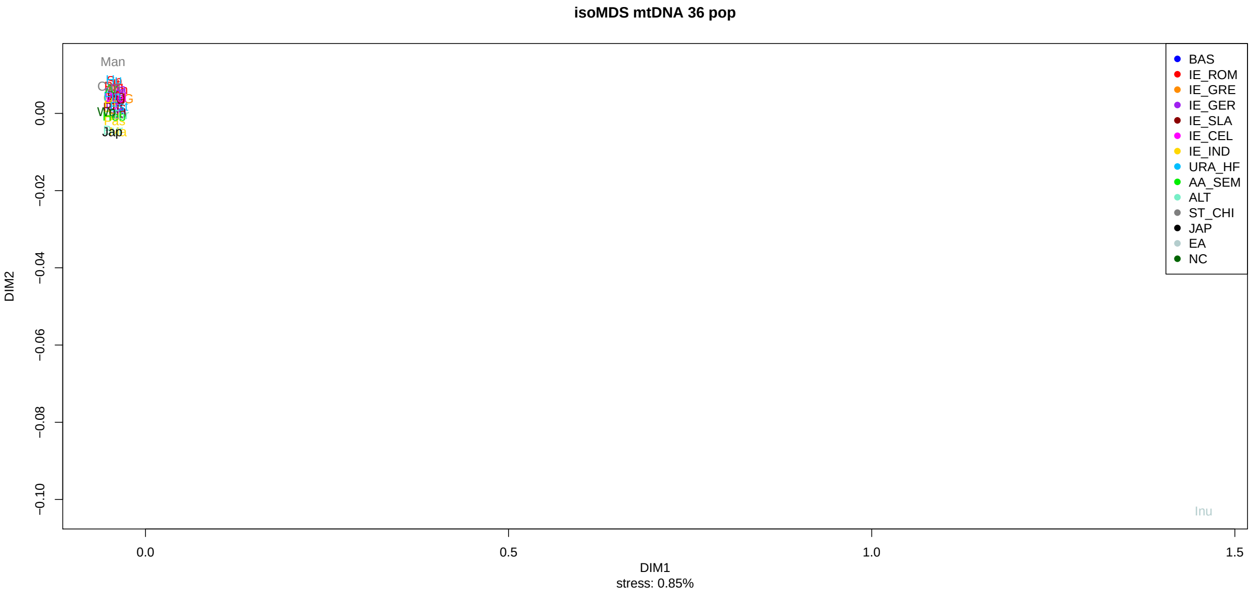

Figure A1: mtDNA multidimensional scaling (36 populations)

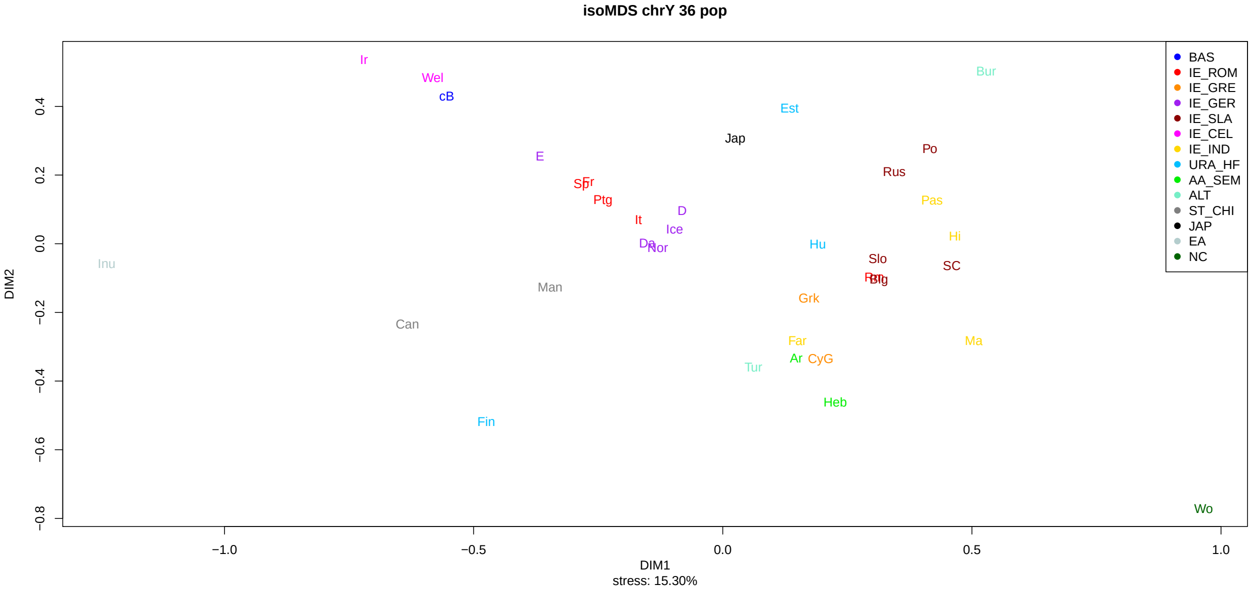

Figure A2: Ychr multidimensional scaling (36 populations)

Table A2: Sardinian samples radiocarbon dates

The table gives the archaeological context, anatomical element, uncalibrated and calibrated 14C-dates for all the analyzed samples. The service reference AA is from NSF Arizona AMS Facility, University of Arizona, Tucson AZ. Calibrated dates are computed from the OxCal computer program (v4.2).

SampleID	**ServiceReference**	**Context**	**Matter**	^{14}C **age (BCE)**	**2 cal age (BCE)**
CAR-H3	AA-75645	testB-lev 4, bottom	Ulna	8620±80	7938-7525
CAR-H7	AA-80544	testB-lev 4, bottom	Tibia	8780±130	8227-7596
CAR-H8	AA-80545	testB-lev 4, bottom	Humerus	9200±180	9124-7851

Table A3: Dataset of the Pre-Neolithic samples

Samples	**Haplogroup**	**Epoch**	**Country**	**Reference**
BERRYAUBAC1	U5b1a	Holocene	France	Posth et al. (2016)
BICHON	U5b1h	LateGlacial	Switzerland	Jones et al. (2015)
BLA20	U5a2c3	Holocene	Germany	Bollongino et al. (2013)
BOCKSTEIN	U5b1d1	Holocene	Germany	Posth et al. (2016)
BRILLENHOHLE	U8a	post-LGM	Germany	Posth et al. (2016)
BURKHARDTSHO	U8a	post-LGM	Germany	Posth et al. (2016)
CIOCLOVINA1	U	post-LGM	Romania	Posth et al. (2016)
CUIRYLESCHAUD	U5b1b	Holocene	France	Posth et al. (2016)
DOLNIVESTONICE13	U8	pre-LGM	Czech Republic	Fu et al. (2013)
DOLNIVESTONICE14	U	pre-LGM	Czech Republic	Fu et al. (2013)
DOLNIVESTONICE16	U5	pre-LGM	Czech Republic	Posth et al. (2016)
DOLNIVESTONICE43	U5	pre-LGM	Czech Republic	Posth et al. (2016)
FALKENSTEIN	U5b2a	Holocene	Germany	Posth et al. (2016)
FELSDACH	U5a2c	Holocene	Germany	Posth et al. (2016)
FUMANE2	R	pre-LGM	Italy	Benazzi et al. (2015)
GOYET2878-21	U5	pre-LGM	Belgium	Posth et al. (2016)
GOYETQ116-1	M	pre-LGM	Belgium	Posth et al. (2016)
GOYETQ2	U8a	post-LGM	Belgium	Posth et al. (2016)
GOYETQ376-19	U2	pre-LGM	Belgium	Posth et al. (2016)
GOYETQ376-3	M	pre-LGM	Belgium	Posth et al. (2016)
GOYETQ53-1	U2	pre-LGM	Belgium	Posth et al. (2016)
GOYETQ55-2	U2	pre-LGM	Belgium	Posth et al. (2016)
GOYETQ56-16	U2	pre-LGM	Belgium	Posth et al. (2016)
HOHLEFELS10	U8a	post-LGM	Germany	Posth et al. (2016)

Table A4: Dataset of the Pre-Neolithic samples

Samples	**Haplogroup**	**Epoch**	**Country**	**Reference**
HOHLEFELS49	U8a	post-LGM	Germany	Posth et al. (2016)
HOHLEFELS79	U8a	post-LGM	Germany	Posth et al. (2016)
HOHLENSTEINSTA	U5b2c1	Holocene	Germany	Posth et al. (2016)
IBOUSSIERES25-1	U5b2a	LateGlacial	France	Posth et al. (2016)
IBOUSSIERES31-2	U5b1	LateGlacial	France	Posth et al. (2016)
IBOUSSIERES39	U5b2b	LateGlacial	France	Posth et al. (2016)
KOSTENKI14	U2	pre-LGM	Russia	Krause et al. (2010)
KOTIAS	H13c	Holocene	Georgia	Jones et al. (2015)
LABRANA	U5b2c1	Holocene	Spain	Sanchez-Quinto et al. (2012)
LAROCHETTE	M	pre-LGM	France	Posth et al. (2016)
LESCLOSEAUX3	U5a2	Holocene	France	Posth et al. (2016)
LOSCHBOUR	U5b1a	Holocene	Luxemburg	Lazaridis et al. (2014)
MA1	U	pre-LGM	Russia	Raghavan et al. (2014)
MAREUILLESMEA	U5a2	Holocene	France	Posth et al. (2016)
OASE1	N	pre-LGM	Russia	Fu et al. (2015)
OBERKASSEL998	U5b1	LateGlacial	Germany	Fu et al. (2013)
OFNET	U5b1d1	Holocene	Germany	Posth et al. (2016)
PAGLICCI108	U2'3'4'7'8'9	pre-LGM	Italy	Posth et al. (2016)
PAGLICCI133	U8c	pre-LGM	Italy	Posth et al. (2016)
PAGLICCI71	U5b2b	post-LGM	Italy	Posth et al. (2016)
RANCHOT88	U5b1	Holocene	France	Posth et al. (2016)
RIGNEY1	U2'3'4'7'8'9	post-LGM	France	Posth et al. (2016)
ROCHEDANE	U5b2b	LateGlacial	France	Posth et al. (2016)
SATSURBLIA	K3	LateGlacial	Georgia	Jones et al. (2015)
USTISHIM	R	pre-LGM	Russia	Fu et al. (2014)

Manuscripts

Complete mitochondrial sequences from Mesolithic Sardinia

Little is known about the genetic prehistory of Sardinia. The scarcity of Pre- and Early Neolithic human remains has made it difficult to properly study the first colonization of the island,as well as the Neolithic transition and the advent of the agriculture. The only ancient genetic data available so far regarded a small portion of the mitochondrial DNA of Nuragic people, and allowed us to identify a genealogical continuity between Bronze-Age and some (but not all) isolated Sardinian communities; however how far the genealogical continuity extends and how it originated was impossible to test. We present the first and oldest complete mitochondrial sequences from Sardinia, dated back to about 10,000 ya. These two sequences belong to rare mtDNA lineages and carry newly-described haplotypes, more similar to those present in European pre-LGM populations, than to those found in coeval sequences. ABC analysis allowed us to gather insight into the Paleolithic contribution to the present-day Sardinian genetic pool and to quantify the genetic impact of the Neolithic transition within the island. The most supported model suggests that the genetic diversity of present-day Sardinians derives from a massive migration from continental Europe during the time of the spread of agriculture, and that the contribution of the first colonizers of the island to the mtDNA of modern Sardinians has been negligible.

SCIENTIFIC REPORTS

OPEN

Complete mitochondrial sequences from Mesolithic Sardinia

Alessandra Modi[1,*], Francesca Tassi[2,*], Roberta Rosa Susca[2], Stefania Vai[1], Ermanno Rizzi[3,4], Gianluca De Bellis[4], Carlo Lugliè[5], Gloria Gonzalez Fortes[2], Martina Lari[1], Guido Barbujani[2], David Caramelli[1,*] & Silvia Ghirotto[2,*]

Received: 29 June 2016
Accepted: 17 January 2017
Published: 03 March 2017

Little is known about the genetic prehistory of Sardinia because of the scarcity of pre-Neolithic human remains. From a genetic perspective, modern Sardinians are known as genetic outliers in Europe, showing unusually high levels of internal diversity and a close relationship to early European Neolithic farmers. However, how far this peculiar genetic structure extends and how it originated was to date impossible to test. Here we present the first and oldest complete mitochondrial sequences from Sardinia, dated back to 10,000 yBP. These two individuals, while confirming a Mesolithic occupation of the island, belong to rare mtDNA lineages, which have never been found before in Mesolithic samples and that are currently present at low frequencies not only in Sardinia, but in the whole Europe. Preliminary Approximate Bayesian Computations, restricted by biased reference samples for Mesolithic Sardinia (the two typed samples) and Neolithic Europe (limited to central and north European sequences), suggest that the first inhabitants of the island have had a small or negligible contribution to the present-day Sardinian population, which mainly derives its genetic diversity from continental migration into the island by Neolithic times.

Due to its geographic isolation in the Mediterranean sea, the biological history of Sardinia has been the subject of extensive anthropological and population-genetics investigation. Several studies based on autosomal markers[1–4], mitochondrial DNA (mtDNA)[5–9] and Y-chromosome polymorphisms[10–13] showed that the Sardinian population is one of the main European genetic outliers[14–17] and reported unusually high levels of internal diversity[18,19]. Most of these studies compared variation in Sardinia and in other European populations, but there is still uncertainty about past population dynamics and demographic processes within the island, as well as about the exact nature and the extent of the genetic exchanges that occurred over millennia, actually determining the existing Sardinian genetic structure.

Controversy has also surrounded the origins and the antiquity of the colonization of Sardinia. The earliest presence of humans is still under debate. Some authors likely date it back to the end of the Middle Pleistocene, on the base of lithic artifacts typology, attributed to the Lower Paleolithic[20–22]. Nonetheless, neither human remains nor absolute-dated contextual evidence support this hypothesis. However, clues of human settlements arose only from the end of the Upper Pleistocene[23,24], with single human remains discovered out of context and dated back to 20,000 years ago just on the base of stratigraphic correlations[25]. The first evidence of Holocene frequentation of the island are scattered in a few rock-shelters and caves, exclusively on the inside of a 20 km coast belt[26]. After this poorly-documented phase, with around 500 years hiatus of archaeological evidence, with the advent of the agriculture, the population of the island increased in size, as demonstrated by the density of Early Neolithic (EN) sites (VI millennium BCE), and at the beginning of the IV millennium BCE, starting from the Final Neolithic culture of *Ozieri*, there has been a rapid growth of archaeological documentation and skeletal remains[27–30]. The fragmented anthropological and archaeological evidence of the Pre-Neolithic phase make it difficult to properly describe a continuity towards the process of Neolithization in Sardinia; however, the gap in the archaeological findings of the two periods suggests a lack of interaction between Mesolithic and EN groups.

From a genetic perspective, a recent genomic study of both ancient and modern Europeans, including data from more than 400 modern-day Sardinians, revealed the existence of genetic affinities between Neolithic Europeans samples and modern Sardinians. According to the authors, these results not only indicate a Neolithic

[1]Dipartimento di Biologia, Università di Firenze, 50122 Florence, Italy. [2]Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, 44121 Ferrara, Italy. [3]Fondazione Telethon, 20121 Milano, Italy. [4]Istituto di Tecnologie Biomediche, CNR, 20090 Segrate, Milano, Italy. [5]LASP, Dipartimento di Storia, Beni Culturali e Territorio, Università di Cagliari, 09124 Cagliari, Italy. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to S.G. (email: ghrslv@unife.it)

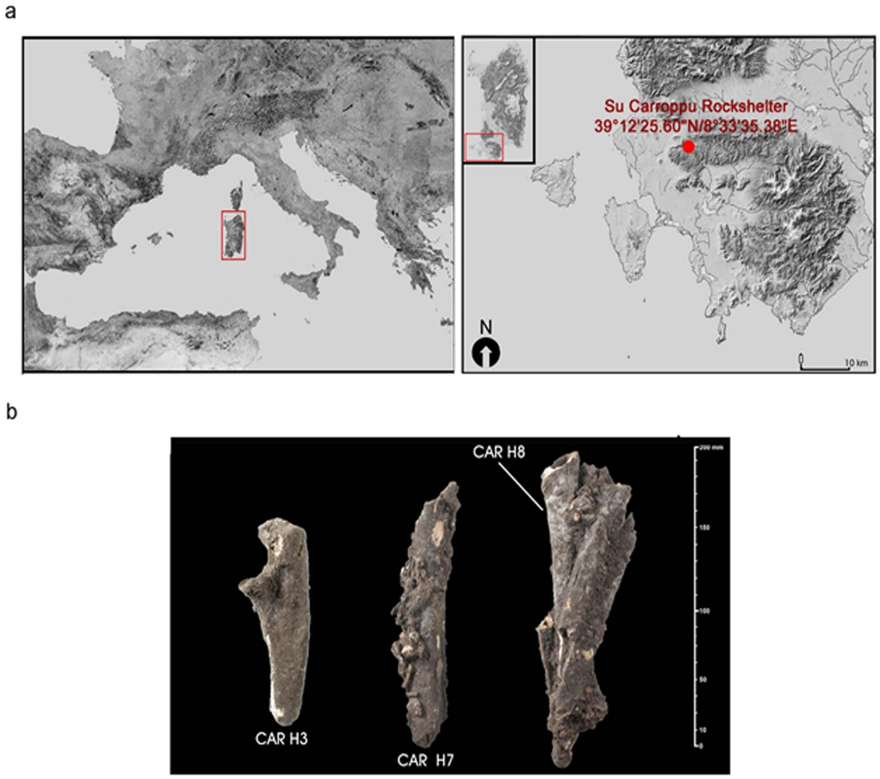

Figure 1. Su Carroppu site and samples. (**a**) the location of Su Carroppu rockshelter, Sardiania (Italy) and (**b**) pictures of the 3 samples used in this study. The map is plotted using data available on http://webgis.regione.sardegna.it/Download/raccolteCartografiche/modelliDigitaliTerreno/DTM10m/.The material is licensed under the Creative Commons attribution 4.0 International license (https://creativecommons.org/licenses/by/4.0/legalcode). The map was processed with Corel Photo-Paint 9 v9.439 (http://www.coreldraw.com/en/product/graphic-design-software/? topNav=en, version 9.439 licensed to CL) and modified with Photoshop CC (2015.5).

origin of modern Sardinians, but also suggest that Sardinians are a "modern-day 'snapshot' of the genetic structure of the first farmers associated with the spread of agriculture in Europe"[31]. Unfortunately, this hypothesis has not been supported so far by evidence coming from ancient Sardinian genetic data, due to the paucity of Pre-Neolithic and actual absence of EN human remains. The only ancient data ever published were sequences of the mtDNA control region from Bronze-Age sample[8] and revealed a directed genealogical continuity between Nuragic individuals and the current people of Ogliastra, but not of Gallura[9]. Past dispersal dynamics, genetic exchanges and replacements during the Neolithic in continental Europe have been extensively studied by means of ancient genetic data[32–42]; although the general European picture is getting clearer, many aspects of the Neolithic transition in Sardinia are still poorly understood, starting from whether, and to what extent, gene flow from mainland Europe during the time of the spread of agriculture actually contributed in shaping the genetic makeup of the island.

With this study, we present the first two complete mitochondrial genome sequences of Mesolithic human remains from Sardinia, dated back to around 10,000 yBP and associated with the earliest direct evidence of human presence in the island[43]. We analyzed these sequences along with modern and ancient genetic data in order to contextualize the Mesolithic Sardinian haplotypes into the European genetic variation, as well as to investigate the Paleolithic contribution to the current Sardinian gene pool. Preliminary model testing under an Approximate Bayesian Computation (ABC) framework is so far, given the extremely limited reference samples for Mesolithic Sardinia and Neolithic Europe supporting the hypothesis that modern-day Sardinian genetic variation is mostly derived from a massive migration from continental Europe during Neolithic times.

Results

Samples and sequencing.

We analyzed the remains of three individuals excavated from the Su Carroppu rockshelter of the Sulcis region (Fig. 1, Supplementary Fig. S1a and b). The Su Carroppu site plays a relevant role in Sardinia, with a remarkably rich archaeological record and a series of occupational phases spanning from the Mesolithic to the historical period. The 1978 archaeological excavations in the lowermost layer (level-4) (Supplementary Fig. S1c), yielded a large quantities of remains, including fragments of human bones intermingled with bones of *Prolagus sardus*. Three direct radiocarbon dates performed on the human bones placed the remains in the mid-9th millennium cal. BCE (Table 1; Supplementary Table S1) thus showing an unexpected Early Mesolithic settlement predating EN occupation[43,44].

Sample ID	^{14}C Age (BCE)	nt covered at least at 3-fold coverage (% of mtDNA)	Average fragment lenght	C to T misincorporation at 5′-end (%)	Contamination estimate (95% CI)	Hg
CAR-H3	7938–7525	13,730 (82.88%)	72.74	N/A	N/A	N/A
CAR-H7	8227–7596	16,527 (99.75%)	62.59	34.45	0.9–7.3%	J2b1
CAR-H8	9124–7851	16,446 (99.24%)	53.09	43.18	0.4–5.9%	I3

CI = credibility interval

Table 1. Samples analyzed. For each sample, radiocarbon date, the percentage of mtDNA covered at least at 3-fold coverage, average fragment length, deamination at 5′-end, contamination estimate and mitochondrial haplogroup are reported.

Here we reconstructed nearly complete mitochondrial genomes for two individuals from Su Carroppu (CAR-H7 and CAR-H8, Table 1), using hybridization capture in solution[45] coupled with high-throughput sequencing. A third individual from the same site (CAR-H3) was also captured and sequenced, but the resulted sequences did not reached the standard quality requested to guaranty the reliability of the NGS data and the sample was excluded for further analysis. The samples displayed typical features of aDNA[46]: short fragments, with average length <65 base pairs (bp), and high rate of cytosine deamination at the 5′ end of the molecules (Table 1; Supplementary Fig. S2; Supplementary Table S2). To further assess authenticity in our ancient mitochondrial genomes we evaluated the percentage of possible contaminant reads by estimating the amount of secondary bases at each haplogroup-defining positions: excluding the putative damaged bases, CAR-H7 reached the 4.04% and CAR-H8 reached the 3.38% (details in Supplementary Table S3), values that are within the range of expected contaminants considering the observed figures for published aDNA mitogenomes[33,40]. We also computed Bayesian contamination estimate[47]: the contamination ranging between 0.9–7.3% for CAR-H7 and 0.4–5.9% for CAR-H8 and the probability of authenticity was high in all the two samples, i.e. 0.95 for CAR-H7 and 0.98 for CAR-H8 (Table 1; Supplementary Fig. S3; Supplementary Table S2). The mitochondrial haplogroups were called using HaploGrep[48,49] (Table 1, Supplementary Table S4); the diagnostic variants showed a coverage ranging from 5 to 28 and were further verified by visual inspection. The CAR-H8 sample belongs to haplogroup I3, hence representing, to the best of our knowledge, the first pre-Neolithic sample carrying the haplogroup I. Studies based on complete mitogenomes have previously reported haplogroup I in ancient samples from Iran (individual I674, haplogroup I1c) and Levant (individual I1679, haplogroup I), dated to 5,105 ± 35 yBP and 8,850–8,750 yBP, respectively[39]. It was also found in two late Neolithic individuals from Germany, both belonging to haplogroup I3a and dated to around 4,000 yBP[50] but not in previous periods in Europe. Nowadays, this haplogroup is uncommon; its frequency is about 2% in modern Sardinians, 3% across Europe, and raises at maximum 6% in Northern European countries[51]. This is the first time that haplogroup I is found in a Mesolithic individual in Europe and the fact that we recovered this haplogroup in a sample of only two sequences may mean that it was present at higher frequencies in pre-Neolithic Sardinians or, in general, in the population that first settled in the island. The other sample (CAR-H7) belongs to the haplogroup J2b1. The haplogroup J has already been found in late hunter-gatherer European populations, with a frequency of about 4%[32]. The current frequency of the haplogroup J is higher than that of the haplogroup I, variable in Europe from 1.7% (Caucasus) to 15% (Wales), and representing the 13% of the total modern Sardinians mitochondrial sequences.

Network analyses. We performed a median-joining network analysis[52] to determine the phylogenetic position of the two newly-discovered sequences within the context of the genetic diversity among Pre-Neolithic complete sequences (Supplementary Table S5). Despite the network (Fig. 2a) shows a temporal pattern from left (pre-LGM) to right (Holocene), the Sardinian sequences occupy a peculiar position, not together with coeval sequences (red circles). The background shading indicates the affiliation of the lineages to the major haplogroup definition (that were determined with HaploGrep[48] based on PhyloTree Build 16). Among the non-Sardinian Pre-Neolithic samples, the most frequent major haplogroup is U, represented by 41 sequences. Just a few more haplogroups are present, namely H, K, M (three sequences each), N, R (two sequences each). The two Sardinian haplogroups (I3 and J2b1) appear well differentiated from each other and from all the other haplogroups considered in the analysis.

The Mesolithic CAR-H7 sample represents so far the oldest sequence belonging to haplogroup J2b. To better investigate the phylogeographic variation of this sequence respect to other European and Sardinian sequences belonging to the same haplogroup, we collected a dataset with 48 modern and 5 ancient J2b sequences (Supplementary Table S6) and we performed a median-joining network analysis[52]. The network confirmed that the sample CAR-H7 (green dot) falls within the variation expected for the haplogroup J2b, although carrying 5 private polymorphisms (195C, 3654T, 6053T, 9071T, 10957G) (Fig. 2b). The modern Sardinian J2b haplotypes seem to be well differentiated from the Mesolithic sequence. (Fig. 2b).

Sardinian past demographic history. We then investigated the past demographic history and the genealogical relationships through 10,000 years in Sardinia by Approximate Bayesian Computation[53,54]. We first defined three alternative models of evolution, shown in Fig. 3. The first, which we called "*continuity*", assumed modern inhabitants of Sardinia to be direct descendants of a Mesolithic Sardinian population, without any genetic exchange with continental Europe. The second, which we called "*discontinuity*", assumed a complete replacement of ancient Mesolithic Sardinia by Neolithic people from Continental Europe. Under the third model the current inhabitants of Sardinia are a genetic mixture of local Mesolithic individuals and Neolithic individuals from the

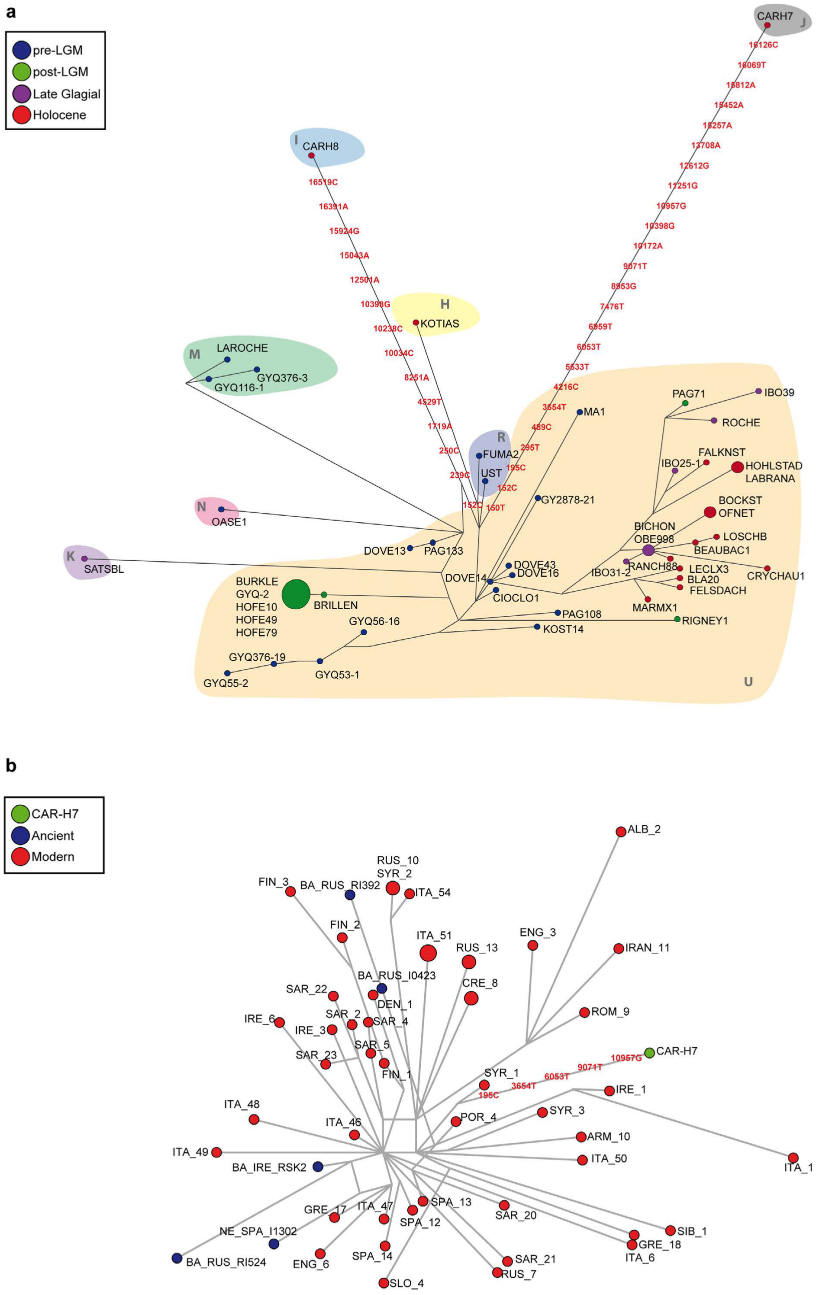

Figure 2. Median-joining network based on nucleotide variation in the whole mtDNA within (**a**) Pre-Neolithic dataset (Supplementary Table S5) (**b**) J2b dataset (Supplementary Table S6).

continent. We called this model "*admixture*". We performed 500,000 coalescent simulations under each model, with parameter values randomly chosen from prior distributions (see Supplementary Materials for details). We calculated the models' posterior probabilities by weighted multinomial logistic regression[53], evaluating different thresholds to check the stability of the results. As it is shown in Fig. 3 and in Supplementary Table S7 the *continuity* model received essentially no support, with the *discontinuity* model having the highest probability (78%). The *admixture* model received poor support (22%), with the best fit obtained when 75% of current inhabitant of Sardinia come from a continental Neolithic population (modal value, see Supplementary Fig. S4 and Supplementary Table S8), that is on the upper bound of its prior distribution.

We determined the accuracy of our model choice inference by calculating the true and the false positive rates using 1,000 random simulations from each model as pseudo-observed datasets; the results are shown in Supplementary Table S9. The true positives rate was high for all the models, ranging from 0.64 to 0.89. The false positive rate was below 0.05 for the *discontinuity* model, and relatively low, but higher, for the *continuity* and the *admixture* model (0.084 and 0.157 respectively). In general, these results mean that the model we tested can be well recognized by the model selection procedure we adopted. We also evaluated the fit of the *discontinuity* model calculating a p-value for the observed dataset under an estimated general linear model (as in Wegmann *et al.*[55]),

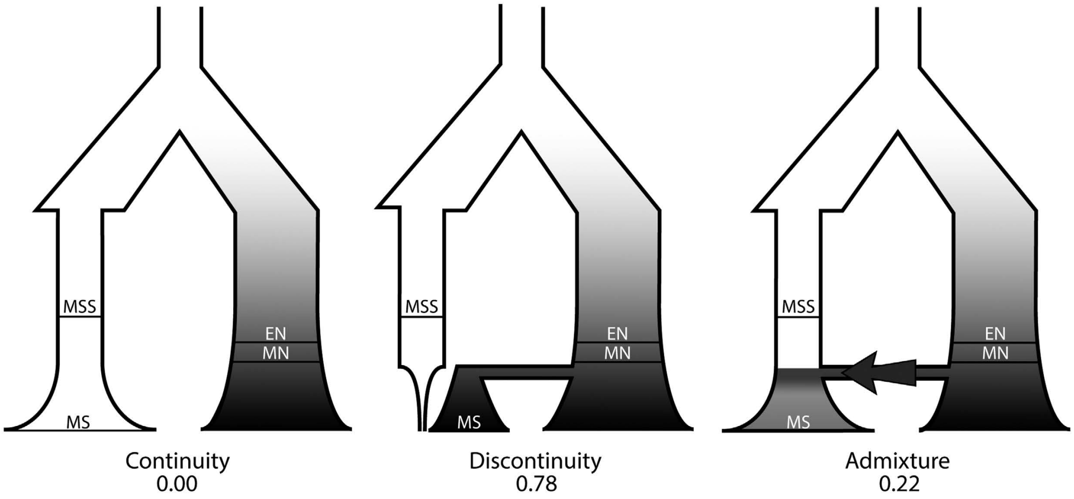

Figure 3. Alternative models of the genealogical relationships among past and present populations, and their posterior probabilities based on 50,000 best fitting simulations. MSS: Mesolithics from Sardinia; MS: Moderns from Sardinia; EN: Early Neolithics; MN: Middle Neolithics.

	Median	Mode	95% HPD-LowB	95% HPD-UppB	R Squared
P	0.872	0.966	0.518	1	0.400
rs	1.647	1	1	4.050	0.088
Nan	2,649	1,517	107	7,626	0.500
Nas	793	461	100	2,785	0.503
Ncn	38,574	14,871	1673	94,089	0.060
Ncs	21,371	9,801	1,000	84,554	0.312
mut	2.1E-08	2E-08	1.3E-08	3.1E-08	0.523

Table 2. Parameters estimation of the *admixture tot* model. *P* is the proportion of Sardinian lineages coming from Neolithic Europe, *rs* is the extent of population reduction due to the bottleneck of the first colonization of Sardinia, *Nan* is the ancient effective population size of Neolithic Europe, *Nas* is the ancient Sardinian effective population size, *Ncn* is the current European effective population size, *Ncs* is the current Sardinian effective population size and *mut* is the mutation rate per nucleotide per year.

which can also be used to judge if the observed data are in agreement with the data simulated. The so calculated p-value was not significant (0.57), meaning that the observed data are plausible under the model we selected as the best one. The Principal Component Analyses of the best 5,000 simulations coming from each model actually confirmed that the *discontinuity* model is able to generate the observed variation, and that only a poorer fit is given by the *admixture* and the *continuity* model (Supplementary Fig. S5).

To better understand the role of the Neolithic migration in shaping the current Sardinians mitogenome variation we then simulated an admixture model (that we called "*admixture_tot*") in which the proportion of lineages of current Sardinians coming from the Neolithic Europe was free to vary from 0 (complete continuity with Mesolithic in Sardinia) to 1 (complete replacement of Sardinian Mesolithics). We estimated the demographic parameters of this model (Table 2 and Supplementary Fig. S6), using summary statistic transformed via Partial Least Square[56] (see Supplementary Materials for details). All the parameters resulted to be well estimated, as it is shown by their R squared values, in some cases higher than 0.6. The median and the mode values of the proportion of Neolithic mitochondrial lineages that gave rise to current Sardinians were 0.87, and 0.96 respectively, implying that a large proportion of the current mtDNA variation in Sardinia does not come from the first inhabitants of the island. We estimated these first incomers having an effective population size of about 790 individuals, with a 95% HPD ranging between 100 and 2,700 individuals. Current Sardinian effective population size was estimated to be predictably higher, with a median value of about 21,000 individuals and a wider 95% HPD. The median value of the mutation rate was estimated to be 2.1*10–8 mutations per nucleotide per year, considering a generation time of 30 years[57], that is almost identical to the value estimated by Fu *et al.*[47]

Discussion

Archaeological evidence suggests the first human presence in Sardinia around 20,000 years ago[25], with sporadic and discontinuous occupations during Paleolithic and Mesolithic ages. Nowadays, Sardinians form a distinct outlier within the genetic variation of modern Europeans[14], often interpreted as a consequence of thousands years of genetic isolation and drift, but little is known about the demographic changes that could have shaped the observed pattern of genetic variation. The, so far limited, ancient Sardinia genetic data allowed us to highlight a complete genetic continuity within a specific region of the island, Ogliastra, since the Bronze-Age[9]; however, cranial morphological evidence suggests that this continuity may have been established since Neolithic times, and possibly earlier[58].

The two Mesolithic sequences retrieved in the Su Carroppu archaeological site represent the oldest sample of DNA in Sardinia, thus providing a direct genetic evidence about the first colonizers of the island. The samples were treated following all the golden criteria before DNA extraction and sequencing to avoid any contamination. To determine the mitochondrial haplogroups, trimmed reads were mapped against the reference sequence and only high quality calls, with a quality score of 30 or more were kept (detailed in Supplementary Materials). The comparable mitochondrial DNA data from European late hunter-gatherers have shown a remarkable genetic uniformity among pre-Neolithic populations, with most of the sequences (~83%) belonging to the haplogroup U, of which a majority carry U5 haplotypes (>65%)[32]. Neither Sardinian sequence belongs to any of the U haplotypes, documenting the presence of substantial genetic differences over the Mediterranean area. In addition, neither sequence has been observed in later, ancient or contemporary, individuals, and both belong to haplogroups and subhaplogroups now present in Europe at low (J,<16%) or very low (I,<7%) frequencies, and that are rare in modern Sardinia. Based on complete mitochondrial genomes, Posth *et al.*[41] described a higher genetic diversity in pre-LGM than in post-LGM European populations and identified a major turnover around 14,000 yBP, with the subsequent expansion of haplogroup U that became widespread all around Europe until the Neolithic transition. We do not find haplogroup U in Sardinia by 11,000 yBP, which means a different impact of the LGM in the island, and probably a high isolation of the Sardinian population, not only from Neolithic times onwards (as genomes data seems to have probed), but also from former times considering the dates of our samples.

The phylogenetic network analysis of all the Pre-Neolithic complete mitochondrial sequences so far generated, actually confirmed this view (Fig. 2a). The majority of the Late Pleistocene and Early Holocene sequences belongs to the U lineage, and form a quite homogeneous cluster at the bottom of the network. The two Mesolithic samples from Sardinia are highly differentiated, departing from the network through long branches, so as to indicate mutations possibly arising along thousand years of geographic (and genetic) isolation. The genome-wide data of Ice Age hunter-gatherers have shown that prehistoric Europe was characterized by recurrent populations turnover and migrations[42], which resulted in a genetic homogeneity across pre-Neolithic populations. So far, our two ancient Sardinian sequences seem to support the view that these ancient populations movements did not involve genetic exchanges with Sardinians: isolation and drift may have resulted in a substantial mitochondrial differentiation between them and other Europeans. A larger characterisation of ancient sequences across the Mediterranean will help to clarify this suggestion.

The role and the genetic impact of migrations in Sardinia from continental Europe has been under debate for years[9,10,31], with particular interest on whether, and to what extent, the gene flow from the mainland during the time of the spread of agriculture in Europe contributed to shaping the present Sardinian gene pool[31]. We then explicitly compared demographic models through Approximate Bayesian Computation[53,59]. The question thus addressed was not whether the two Su Carroppu Mesolithic individuals are ancestral to current Sardinians along the maternal line (of course, they are not), but rather what was the posterior probability that a population of size 100–10,000 individuals (the broad interval of priors considered), and comprising the Su Carroppu individuals, may have contributed to the current gene pool. Because the alternative to genealogical continuity since Mesolithic times is immigration from the mainland, Middle and Early Neolithic sequences from Central Europe[50] were included as a source of Neolithic migrants into the island. This is not the best reference panel but we were limited to use it given the lack of Neolithic sequences from South Europe. Results must be interpreted with caution. A model of genealogical continuity in Sardinia since Mesolithic times appeared very unlikely. We could not discriminate between a model assuming a certain degree of admixture and one of complete replacement by Neolithic immigrants, but if admixture occurred the contribution of Mesolithic people was apparently very limited (Fig. 3).

We assessed the quality of these results by a number of tests. First, we evaluated exactly the probability to obtain false positives in the estimation of models' posterior probabilities, with the discontinuity model having the lowest type one error. Then, we showed in various ways (posterior predictive p-value and PCA analysis) that the discontinuity model can in fact reasonably reproduce the observed variation. Clearly, a certain degree of uncertainty necessarily affects any analysis, particularly when it is based on a single DNA region and on the necessarily small samples in which ancient DNA is usually typed.

When explicitly estimating the Neolithic admixture proportion, i.e. the amount of Neolithic genes from continental Europe that gave rise to the current Sardinian genetic pool (*admixture_tot* model), we obtained values of 0.8–0.9%, depending on the point estimates considered (Table 2). This means that a significant proportion of modern Sardinian mitochondrial variation would came not from its first settlers, but from a subsequent migration wave from the continent. These results need to be tested in the future when reference ancient datasets are extended for both Mesolithic Sardinia and Neolithic Mediterranean. It is well accepted in the literature that the Neolithization of Europe proceeded in two waves, one for Central and North Europe, and the other for South Europe/Mediterranean[33,35,60,61]. But currently, there is no good proxy available for the ancient Neolithic Mediterranean pool. In our model comparison, we fixed the time of this second migration to 6,000 years ago, thus compatible with the archaeological evidence of Neolithic expansion in Sardinia. The spread of agriculture in Sardinia would hence been associated with demic diffusion from the continent, resulting in a large-scale population replacement. These results, for the first time supported by ancient genetic data, are also in good agreement with archaeological evidence and with what emerged from the comparison of modern Sardinian genomic data and Neolithic and Paleolithic sequences[31], interpreted by the authors as evidence of gene flow from mainland Europe during the time of the spread of agriculture in Europe. Sikora *et al.*[31] also envisaged a genetic continuity until present times, but did not provide quantitative measures of it. Another possibility, compatible with our results, would be that Sardinian Paleo-Mesolithic males, but not females, admixed with immigrants from Neolithic Europe. This, however, would mean that in Sardinia the spread of the Neolithic culture was mainly carried out by women, in contrast with the available evidence[62]. Moreover, this view is also in contrast with studies of sex-biased admixture in modern communities, suggesting that the invading population tends to incorporate female residents more than males[63–67].

In conclusion, this study, albeit limited to DNA transmitted along the female lines of descent, provides the first genetic evidence on the earliest inhabitants of Sardinia, who bear maternal lineages distinct from current ones. Based on these two sequences, it seems that the Neolithization of the island was not a local development, but was associated with the arrival of a genetically-distinct group of immigrants from continental Europe.

Methods

DNA extraction and Sequencing. All extraction and library preparation steps before amplification were performed in the clean-room facilities of the Laboratory of Molecular Anthropology and Paleogenetics, University of Florence. Preventive measures were taken to avoid contamination during all experiments.

Sample surface was mechanically removed using a dental micro-drill with disposable tools, then the samples were UV-irradiated (254 nm) for 1 hour. Samples were ground to fine powder using the same dental micro-drill at very slow rotation (1000 rpm) and stored at −20 °C until further use. For each sample, DNA was extracted from 100 mg of bone powder following a silica-based protocol[68]. A 25 μl aliquot of each extract was used to produce double-stranded and double-indexed libraries according to a modified Illumina multiplex protocol[69]. All libraries were amplified to reach plateau and enriched for human mtDNA in a bead-capture method using long-range PCR products as bait for hybridization[45]. Negative controls were processed during each experimental step (see Supplementary Materials and Supplementary Table S10 for details).

Enriched libraries were pooled in equimolar amount with libraries from other samples and sequenced in paired-end ($2 \times 75 + 8 + 8$ cycles) on the Illumina MiSeq platform at the Institute of Biomedical Technologies, National Research Council, in Segrate (Milano).

NGS Data Processing and Authentication. Paired-end reads were merged into single reads and the adapters were trimmed using SeqPrep[70]. Filtered reads were mapped against the revised Cambridge Reference Sequence (rCRS) using BWA[71], setting "-l 1000 -n 0.01 -o 2" optimized for increased sensitivity for aDNA; reads with mapping quality below 30 were discarded and PCR duplicates were collapsed into consensus sequences. To estimate the misincorporation pattern at the end of the reads, BAM files were run on *mapDamage2.0*[72]. Then, to test for the authenticity of the consensus sequences, we used a Bayesian contamination estimate to calculate the probability that the recovered mtDNA fragments come from a single biological source[47]. A detailed description can be found in Supplementary Materials, and in Supplementary Table S11.

Haplogroup identification. Consensus sequences were called using samtools packages[73]: only high quality calls with a quality score of 30 or more were kept. The two sequences were uploaded on HaploGrep[48,49] to assign the mitochondrial genome to known haplogroups and call mtDNA SNPs, followed by manual verification of each diagnostic variant.

In order to reduce the loss of the information, the assemblies were subsequently visually inspected.

Network analysis. The phylogenetic networks based on nucleotide variation in the whole mtDNA, were constructed using the Median Joining algorithm[52] implemented in Network 5.0 program (http://www.fluxus-technology.com). The ε value was set to 0 and the transversions were weighted 3x the weight of transitions. Networks were subjected to maximum parsimony post-analysis.

Approximate Bayesian Computation. We implemented the ABC framework using the *ABCsampler* tool in the ABCToolbox package[55]. We simulated genetic data under three demographic models (*continuity*, *discontinuity* and *admixture*, see Fig. 3, detailed in Supplementary Materials) with *fastsimcoal2* (ver 2.5.2.21)[74] and running 500,000 simulations per model. The prior distributions we considered are detailed in Supplementary Table S12. The modern Sardinian sample includes 63 sequences from Ogliastra[75], the unique unbiased sample of Sardinian complete mitochondrial genomes available. As source of Neolithic variation we used 18 Middle Neolithic (6,500–5,000 BCE) and 28 Early Neolithic (7,300–6,200 BCE) sequences from Haak *et al.*[50], that are the Early and Middle Neolithic samples with the highest quality (see Supplementary Materials and Supplementary Table S13). We placed ancient samples in the corresponding branch of the demographic model, at an average sampling time. To compare models we applied the Logistic Regression procedure[59], considering different thresholds (i.e. number of retained simulations) to check the consistence oh the results. Model parameters were estimated by a locally weighted multivariate regression[53] after a *logtan* transformation[76] of the 5,000 best-fitting simulations from a specific model. To calculate the posterior probabilities for models and parameters we used R[77] scripts from http://code.google.com/p/popabc/source/browse/#svn%2Ftrunk%2Fscripts, modified by SG. We also estimated the power of our ABC procedure to correctly recognize the true model calculating for each model the proportion of true positives and false positives. We evaluated 1,000 random pseudo-observed data sets generated under each model, counting the number of times a specific model is correctly identified by the ABC procedure (true positives), and the number of times the same model is incorrectly selected as the true model (false positives). The PCA was made with the *PCA* function of the *FactoMineR* package[77,78].

References

1. Grimaldi, M. C. *et al.* West Mediterranean islands (Corsica, Balearic islands, Sardinia) and the Basque population: contribution of HLA class I molecular markers to their evolutionary history. *Tissue Antigens* **58,** 281–292 (2001).
2. Battaggia, C., Ruscitto, D., Destro-Bisol, G., Vacca, L., Calo, C. & Vona, G. Frequencies at CD4, FES, and F13A1 microsatellite loci in central-southern Sardinia (Italy). *J Forensic Sci* **48,** 442 (2003).

3. Falchi, M. *et al.* A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia. *Am J Hum Genet* **75,** 1015–1031 (2004).
4. Di Gaetano, C. *et al.* Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection. *PLoS One* **9,** e91237 (2014).
5. Barbujani, G., Bertorelle, G., Capitani, G. & Scozzari, R. Geographical structuring in the mtDNA of Italians. *Proc Natl Acad Sci USA* **92,** 9171–9175 (1995).
6. Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67,** 1251–1276 (2000).
7. Falchi, A. *et al.* Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. *J Hum Genet* **51,** 9–14 (2006).
8. Caramelli, D. *et al.* Genetic variation in prehistoric Sardinia. *Hum Genet* **122,** 327–336 (2007).
9. Ghirotto, S., Mona, S., Benazzo, A., Paparazzo, F., Caramelli, D. & Barbujani, G. Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Mol Biol Evol* **27,** 875–886 (2010).
10. Francalacci, P. *et al.* Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. *Am J Phys Anthropol* **121,** 270–279 (2003).
11. Capelli, C. *et al.* A 9-loci Y chromosome haplotype in three Italian populations. *Forensic Sci Int* **159,** 64–70 (2006).
12. Contu, D., Morelli, L., Santoni, F., Foster, J. W., Francalacci, P. & Cucca F. Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. *PLoS One* **3,** e1430 (2008).
13. Francalacci, P. *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341,** 565–569 (2013).
14. Cavalli-Sforza, L. L. & Piazza, A. Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur J Hum Genet* **1,** 3–18 (1993).
15. Quintana-Murci, L., Veitia, R., Fellous, M., Semino, O. & Poloni, E. S. Genetic structure of Mediterranean populations revealed by Y-chromosome haplotype analysis. *Am J Phys Anthropol* **121,** 157–171 (2003).
16. Pugliatti, M. *et al.* Evidence of early childhood as the susceptibility period in multiple sclerosis: space-time cluster analysis in a Sardinian population. *Am J Epidemiol* **164,** 326–333 (2006).
17. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47,** 1272–1281 (2015).
18. Barbujani, G. & Sokal, R. R. Genetic population structure of Italy. II. Physical and cultural barriers to gene flow. *Am J Hum Genet* **48,** 398–411 (1991).
19. Zei, G. *et al.* From surnames to the history of Y chromosomes: the Sardinian population as a paradigm. *Eur J Hum Genet* **11,** 802–807 (2003).
20. Fenu, P., Martini, F. & Pitzalis G. I siti paleolitici: i complessi industriali. Sa Pedrosa-Pantallinu. In: *Sardegna paleolitica. Studi sul più antico popolamento dell'isola. Museo Fiorentino di Preistoria "Paolo Graziosi"* (eds Martini, F. E.). Centro Stampa 2P (1999).
21. Martini, F. Le facies clactoniane sarde nel quadro del Paleolitico inferiore peninsulare. In: *Sardegna paleolitica. Studi sul più antico popolamento dell'isola. Museo Fiorentino di Preistoria "Paolo Grazio* (eds Martini, F. E.). Centro Stampa 2P (1999).
22. Martini, F. & Ulzega, A. L'insularità e i suoi effetti sul popolamento umano delle isole del Mediterraneo nel Pleistocene e nel primo Olocene. *Riv Sci Preist* **42,** 271–288 (1989–1990).
23. Sondaar, P. Y. *et al.* Il popolamento della Sardegna nel tardo Pleistocene: nuova acquisizione di un resto fossile umano dalla grotta Corbeddu. *Riv Sci Preist* **45,** (1993).
24. Sondaar, P. Y. *et al.* The human colonization of Sardinia: a Late-Pleistocene human fossil from Corbeddu Cave. *C R Acad Sci Paris* (*Série IIa*) **320,** 145–150 (1995).
25. Spoor, F. The human fossils from Corbeddu Cave, Sardinia: a reappraisal. in: *Elephants have a snorkel!* (eds Reumer JWFDVs J., St. John). Deinsea (1999).
26. Lugliè, C. Il Mesolitico In: Atti della XLIV Riunione Scientifica dell'IIPP La preistoria e la protostoria della Sardegna, (Cagliari-Barumini-Sassari, 23-28 novembre 2009) (eds IIPP) (2009).
27. Floris, G. Sulla variabilità dell'indice nasale dei protosardi. *Bollettino della Società Sarda di Scienze Naturali* **21,** 129–135 (1981).
28. Floris, G. La stautura nella protostoria sarda. *Arch Antrop Etnol* **113,** 263–267 (1983).
29. Germanà, F. L'uomo in Sardegna dal Paleolitico all'Età nuragica. *C. Delfino* (1995).
30. Sanna, E., Liguori, A., Fagioli, M. B. & Floris, G. Verso una revisione dell'inquadramento cronologico e morfometrico delle serie scheletriche paleo-protosarde. II: Craniometria, ulteriori aggiornamenti. *Arch Antrop Etnol* **129,** 239–250 (1999).
31. Sikora, M. *et al.* Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genet* **10,** e1004353 (2014).
32. Fu, Q., Rudan, P., Paabo, S. & Krause, J. Complete mitochondrial genomes reveal neolithic expansion into Europe. *PLoS One* **7,** e32473 (2012).
33. Skoglund, P. *et al.* Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344,** 747–750 (2014).
34. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun* **5,** 5257 (2014).
35. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
36. Gunther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc Natl Acad Sci USA* **112,** 11917–11922 (2015).
37. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326,** 137–140 (2009).
38. Malmstrom, H. *et al.* Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Curr Biol* **19,** 1758–1762 (2009).
39. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536,** 419–424 (2016).
40. Sanchez-Quinto, F. *et al.* Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Curr Biol* **22,** 1494–1499 (2012).
41. Posth, C. *et al.* Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe. *Curr Biol* **26,** 827–833 (2016).
42. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* (2016).
43. Lugliè, C. The Su Carroppu rock shelter within the process of Neolithization of Sardinia. In: *Transitions en Méditerranée, ou comment des chasseurs devinrent agriculteurs* (ed^(eds) (2014).
44. Gassin, B. & Lugliè, C. Delle frecce per far cosa? In *Atti della XLIV R.S. IIPP La preistoria e la protostoria della Sardegna*. **II,** 485–493 (2012).
45. Maricic, T., Whitten, M. & Paabo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* **5,** e14004 (2010).
46. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Paabo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* **7,** e34131 (2012).
47. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* **23,** 553–559 (2013).

48. van Oven, M. & Kayser, M. Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Hum Mutat* **30,** 386–394 (2009).
49. Kloss-Brandstatter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* **32,** 25–32 (2011).
50. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).
51. Olivieri, A. *et al.* Mitogenomes from two uncommon haplogroups mark late glacial/postglacial expansions from the near east and neolithic dispersals within Europe. *PLoS One* **8,** e70492 (2013).
52. Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human populations using median networks. *Genetics* **141,** 743–753 (1995).
53. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162,** 2025–2035 (2002).
54. Bertorelle, G., Benazzo, A. & Mona, S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* **19,** 2609–2625 (2010).
55. Wegmann, D., Leuenberger, C. & Excoffier, L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182,** 1207–1218 (2009).
56. Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11,** 116 (2010).
57. Batini, C. *et al.* Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat Commun* **6,** 7152 (2015).
58. D'Amore, G., Di Marco, S., Floris, G., Pacciani, E. & Sanna, E. Craniofacial morphometric variation and the biological history of the peopling of Sardinia. *Homo* **61,** 385–412 (2010).
59. Beaumont, M. *Joint determination of topology, divergence time and immigration in population trees*. McDonald Institute for Archaeological Research, 135–154 (2008).
60. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336,** 466–469 (2012).
61. Pinhasi, R. *et al.* The genetic history of Europeans. *Trends Genet* **28,** 496–505 (2012).
62. Rasteiro, R. & Chikhi, L. Female and male perspectives on the neolithic transition in Europe: clues from ancient and modern genetic data. *PLoS One* **8,** e60944 (2013).
63. Abe-Sandes, K., Silva, W. A. Jr. & Zago, M. A. Heterogeneity of the Y chromosome in Afro-Brazilian populations. *Hum Biol* **76,** 77–86 (2004).
64. Gonzalez-Andrade, F., Sanchez, D., Gonzalez-Solorzano, J., Gascon, S. & Martinez-Jarreta, B. Sex-specific genetic admixture of Mestizos, Amerindian Kichwas, and Afro-Ecuadorans from Ecuador. *Hum Biol* **79,** 51–77 (2007).
65. Goncalves, V. F., Carvalho, C. M., Bortolini, M. C., Bydlowski, S. P. & Pena, S. D. The phylogeography of African Brazilians. *Human heredity* **65,** 23–32 (2008).
66. Stefflova, K. *et al.* Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. *PLoS One* **4,** e7842 (2009).
67. Quintana-Murci, L. *et al.* Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. *Am J Hum Genet* **86,** 611–620 (2010).
68. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultra-short DNA fragments. *Proc Natl Acad Sci USA* **110,** 15758–63 (2013).
69. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010,** pdb prot5448 (2010).
70. John, J. St. SeqPrep.(2011).
71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
72. Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29,** 1682–1684 (2013).
73. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
74. Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet* **9,** e1003905 (2013).
75. Fraumene, C., Petretto, E., Angius, A. & Pirastu, M. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. *Hum Genet* **114,** 1–10 (2003).
76. Hamilton, G., Stoneking, M. & Excoffier, L. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc Natl Acad Sci USA* **102,** 7476–7480 (2005).
77. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (2013).
78. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25,** 1–18 (2008).

Acknowledgements

We thank Giuseppe Marasco for the help with the processing pipeline and the assembling of the datasets, Alessandro Achilli and Anna Olivieri for useful suggestions. G.B., S.G., F.T. and R.R.S. were supported by the ERC Advanced Grant Agreement No 295733, 'LanGeLin' project. C.L., A.M. and D.C. were supported by the L7/2007 CUP F71J09000060002 grant of the Regione Autonoma della Sardegna, 'Storia del primo popolamento neolitico della Sardegna (VI-V millennio a.C.): origine e processi evolutivi alla luce dei dati archeologici, linguistici ed archeogenetici' project. Research was also supported in part by PRIN 2010–2011 grant 2010EL8TXP_003 from the Italian Ministry of Education, University and Research (MIUR) assigned to D.C.

Author Contributions

D.C., S.G., A.M. and C.L. conceived the project. C.L. provided the samples. A.M., S.V., E.R. and M.L. designed the sequencing experiments. G.D.B. contributed reagents/materials for sequencing. G.G.F. provided the bioinformatic pipeline to process the data. A.M., F.T., R.R.S. and S.G. carried out data analysis. S.G., A.M., F.T. and G.B. wrote the manuscript. All authors read the manuscript and provided critical input.

Additional Information

Accession Codes: The accession numbers for the two mtDNA genome sequences reported in this paper are GenBank: KX354973-KX354974.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Modi, A. *et al.* Complete mitochondrial sequences from Mesolithic Sardinia. *Sci. Rep.* **7**, 42869; doi: 10.1038/srep42869 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Natural selection accounts for differences in dizygotic twinning rates at the worldwide scale

Patterns in the distribution of dizygotic twinning rates have long been recognized, with African and Asian populations falling, respectively, at the top and at the bottom of the range. Various genetic effects have been proposed to account for such population differences, but the evidence for association between twinning rates and specific alleles is, so far, only anecdotal. We assembled a dataset of 83 candidate twinning genes in 26 worldwide-distributed populations, and a Reference dataset of >40 000 SNPs mapping in intergenic genome regions, which we used to derive neutral expectations. We found strong evidence of positive selection for seven loci associated with twinning rates. All of these loci are involved in folliculogenesis, follicle selection, ovulation and implantation (BMP15, BMP3, GDF5, IGF2, LHCGR, IGFBPL1 and MTHFR). Population differences at these candidate genes point to slight, but not negligible, differences for the bioactivity of these enzymes, hormones and receptors, which should in turn have an influence upon the rate of twinning in the populations considered.

March 01, 2017

Natural selection accounts for differences in dizygotic twinning rates at the worldwide scale

Roberta Rosa Susca[1], Kathryn BH Clancy [2] , Lorena Madrigal[3], Guido Barbujani[*, 1]

[1] Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, Ferrara, Italy

[2] Department of Anthropology, University of Illinois at Urbana-Champaign, 109 Davenport Hall, 607 S. Mathews Ave., Urbana IL, 61801, USA

[3]Department of Anthropology, University of South Florida, 4202 E. Fowler Av. Tampa, FL, 33620, USA

[*] Correspondence: Dipartimento di Scienze della Vita e Biotecnologie, via Borsari 46, I-44121 Ferrara, Italy. Tel.: +39 0532 455745; Fax: +39 0532 249761; Email: g.barbujani@unife.it

The authors declare no competing financial interests in relation to the work described.

Running title: Twinning and fertility-related genes

Keywords: Twinning; population genetics; genome diversity; selection; BayeScan; BAYENV2

ABSTRACT: Patterns in the distribution of dizygotic twinning rates have long been recognized, with African and Asian populations falling, respectively, at the top and at the bottom of the range. Previous work has demonstrated associations between specific loci and twinning odds within homogenous populations, but no study has taken a world-wide perspective. In this paper we assembled a dataset of 83 candidate twinning genes located in all 23 pairs of chromosomes in 26 worldwide-distributed populations, and a reference dataset of >50 000 SNPs mapping in intergenic genome regions, which we used to derive neutral expectations. This way we could test, through Bayesian approaches: (a) whether candidate twinning genes show an excess of between-population diversity, traditionally associated with the effects of differential selection, and (b) whether SNP frequencies at these loci correlate with twinning rates. Combining the two approaches, we found strong evidence of positive selection for seven loci associated with twinning rates. All of these loci are involved in folliculogenesis, follicle selection, ovulation and implantation (*BMP15*, *BMP3*, *GDF5*, *IGF2*, *LHCGR*, *IGFBPL1* and *MTHFR*). Linkage disequilibrium with neighboring loci is highly unlikely to have biased these results. Population differences at these candidate genes point to slight, but not negligible, differences for the bioactivity of these enzymes, hormones and receptors, which should in turn have an influence upon the rate of twinning in the populations considered. Our results indicate that twinning is a phenotype affected by multiple loci, for which humans have genetic heritability, partly affected by geographically-variable selection.

INTRODUCTION

Human population differences in frequencies of twinning have long been recognized. Work done before and during the 1970's universally indicated that there are continent-wide differences, with Sub-Saharan African populations having the highest twinning frequencies, Asian populations the lowest, and European and Near-Eastern populations falling somewhere in between [1]. After the 1970's, these generalizations remained largely unchallenged, although they became more nuanced. It was recognized that there was variation within Sub-Saharan populations, with twinning rates ranging from 16 up to 40-50 per thousand [2], just as it was recognized that not all Asian populations had the same low rate of twinning (3-4 per thousand)[3] and that intermediate (8 per thousand) frequencies are typical of European and Middle Eastern populations, with some degree of variation. Unfortunately, in the literature on continent-wide frequencies of twinning, Native American populations have been largely ignored.

Possible explanations for differences in twinning rates among populations include several environmental and genetic factors. As for the former, the highest twinning frequencies in Sub-Saharan Africans have been attributed to their high consumption of yams, which are said to add phytoestrogens to the populations' diet[4]. Although this explanation might be partially correct, it overlooks the issue that yams are also part of the diet for many Pacific island populations, where the twinning frequency is not nearly as high as in Africa[3]. Moreover, this explanation overlooks the fact that African-derived populations in the New World also have high frequencies of twinning despite heterogeneous yam consumption, suggesting that there is a genetic component to twinning frequency differences[5].

In all of these studies DZ twinning is singled out because the frequency of monozygotic twinning (MZ) is (for all intents and purposes) a teratogenic event and therefore the same across populations[6]. That is, although there may be families in which MZ twinning is "familial" (i.e., has a genetic component), most twins which "run in families" are dizygotic, indicating that if there is a genetic component to twinning, it affects DZ, not MZ twinning.

Spontaneous dizygotic (DZ) twinning is a phenomenon related with the release and fertilization of two oocytes. A genetic component to it has long been recognized but no clear inheritance pattern has been identified, indicating that the trait is likely polygenic. Potential loci involved in DZ twinning may affect the frequency of multiple pregnancies by affecting ovulation, by being involved in the metabolic pathway, by influencing levels of circulating steroid hormones, by promoting multiple implantations or by supporting multiple pregnancies[7-12].

A reasonable starting point for investigating population-wide genetic predisposition to twinning is to consider the possibility that alleles affecting the frequency of DZ twinning may be enriched in populations with high twinning frequencies. If there is a genetic component to differences in the rate of DZ twinning across populations and continents, which genes are involved, and why? Previous papers have by necessity either been restricted to samples from particular geographic regions or ethnic groups, or have been limited to samples of mothers of twins, rather than samples representative of the entire populations. This paper takes a worldwide view of the distribution of allelic variants of genes which have been proposed as candidate genes associated with twinning in the literature (listed in Supplementary Table 1), and tests whether their patterns of variation can account for aspects of the well-known differential distribution of twinning across the world.

MATERIALS AND METHODS

Datasets

Our analyses are based on three datasets, respectively called Candidate, Reference and Twinning rates datasets. We started by considering the public data of Phase 3 of the 1000Genome project [13] representing 26 populations for a total of 2 504 unrelated individuals. The Candidate dataset refers to all SNPs of 83 candidate genes (Figure 2, Supplementary Table 1) identified by a thorough scan of the available literature[14] (the UTRs regions were included). We retrieved the SNPs annotation, based on the hg19 release, on the UCSC genome browser (https://genome.ucsc.edu/). We retrieved the HGSV ID for each SNP cited thanks to the "SNP Converter" tool available within the Mutalyzer website (https://mutalyzer.nl/)[15]. We graphically drew the 83 genes on each chromosome thanks to the NCBI GDP (Genome Decoration Page) utility (https://www.ncbi.nlm.nih.gov/genome/tools/gdp/).

From the initial Candidate dataset (168 229 SNPs) we removed loci with very low minor allele frequencies (MAF < 0.001) by the PLINK 1.07 tool (http://pngu.mgh.harvard.edu/purcell/plink/) [16].

We further pruned the dataset removing loci with levels of pairwise Linkage Disequilibrium higher than r^2=0.4 by the PLINK, so that 30 798 SNPs were in fact initially considered. Taking into account the F_{ST} distributions result (see details in the result section) we decided to apply an additional continental-based filter: considering the four continents one by one, we excluded the SNPs showing a MAF<0.005 in all continents, i.e. the subpolymorphic variation which cannot possibly account for differences in twinning rates between continents. We then combined all SNPs with MAF>0.005 in all continents. The final Candidate dataset included 3 486 SNPs.

Unlike selection, drift and gene flow (as well as, for many practical purposes, mutation) are expected to affect all loci equally [17]. The presence of outliers in the distribution of genetic variances, i.e. of loci showing very high levels of variation, is thus taken as suggestive of positive selection at those loci [18,19]. To identify such outliers, we created a Reference dataset of presumably neutral loci using the NRE software (http://nre.cb.bscb.cornell.edu/nre/) [20], according to criteria explained in Patin et al. (2009) [21]. We started from 200 086 autosomal, presumably neutral sites, sampled at random from the 1000Genomes data, in non-coding regions at least 200 kb away from any known or predicted gene. We then selected by PLINK a random subset of 20% of the loci, thus obtaining the Reference dataset (54 719 SNPs).

A third dataset (Twinning rates dataset) contained twinning rates for 22 of the 26 populations (see Supplementary table 2 and Figure 1). In 19 cases, these rates referred exactly to a population with the same ethnic specificity (even if they have migrated out of their homeland) as that of the 1000Genomes data. In three cases, we resorted to what seemed suitable proxies on anthropological grounds. The twinning rates for Afro-Caribbeans in Barbados was approximated by Afro-Costa Ricans; for Utah residents of North West European ancestry by Canadians of the same ancestry; for Puerto Ricans by Hispanic-ancestry Costa Ricans. Four populations, for which we could not find a suitable twinning rate estimate, were excluded from the analyses requiring those data. For the African samples, we purposely did not use the earlier (and highest) reported twinning rates but looked for the more recent (and less-strikingly higher) twinning rates in an effort to avoid outliers.

Linkage pedigree files (PED) and marker information files (MAP) were generated from the 1000Genomes VCF files by “VCF to PED converter” (http://browser.1000genomes.org/Homo_sapiens/UserData/Haploview) available within the 1000Genomes Online Tools. Allele frequencies were estimated by PLINK and plotted on pie chart world-maps using QGIS (http://qgis.osgeo.org).

F_{ST} distributions

Geographic patterns in SNP frequencies are shaped by the joint effects of mutation, drift, gene flow and selection. A time-honored approach to identify the effects of positive selection is based on the idea that mutation, drift and gene flow should affect all loci equally, whereas selection acts specifically on some target loci [17]. We calculated by the *pegas* R-cran package [22] F_{ST} [23] values for both a subset of the Candidate dataset (including 30 798 SNPs) and for the Reference dataset (54 719 SNPs). We then compared the two F_{ST} distributions. The null hypothesis was that candidate loci are a random subset of neutral loci in the genome, the alternative hypothesis was that some of the candidate genes show higher between-population diversity than neutral loci because of different selection regimes affecting populations.

Testing for selection 1: Comparing neutral and selection models

We then tested for positive selection at candidate loci by comparing the weight of evidence in favor of a model including selection versus a neutral model in which the observed F_{ST} values only reflect past demographic changes. For that purpose, we used a method assuming a Bayesian framework and based on a reversible jump Markov Chain Monte Carlo algorithm, BayeScan (http://cmpg.unibe.ch/software/BayeScan/; [24]). Allele-frequency data were described by a model characterized by two coefficients. One of them (ß) assumes an island model and is thus regarded as population-specific, the second one (α) is locus-specific and reflects the strength of selection acting upon that locus. For each SNP, we estimated the posterior distributions of F_{ST}, both under neutrality (α = 0) and allowing for selection (α ≠ 0), and compared their posterior odds ratio (PO). This way, if α is significantly positive or negative, the locus is considered subjected to positive and balancing selection, respectively.

We ran the BayeScan analyses on both the final Candidate dataset (3 486 SNPs) and a comparable subset of the Reference dataset (5 505 randomly chosen SNPs) under identical conditions, namely prior odds =1 000 (thus assuming the neutral model to be 1 000-fold as likely as the selection model), 20 pilot runs, 50 000 burn-in iterations followed by 50 000 output iterations with a tinning interval of 10, resulting in 5 000 iterations for posterior estimation. Note that choosing a prior probability in favor of the neutral model is a common, if somewhat arbitrary, practice, conceptually similar to the Bonferroni correction for multiple tests. By giving the neutral model a high prior probability, one reduces the risk of false positive results [25-28].

We fixed two significance thresholds. The first one was Model-based, namely the expected value of the $\log_{10}$PO (the support for the model of local adaptation relative to neutral demography) which would yield a 1% false-discovery rate based on the reference SNPs. The second threshold was a Reference-based threshold corresponding to F_{ST} values in candidate loci falling in the upper 5% of the distribution of F_{ST}s in the Reference dataset.

Testing for selection 2: Association with twinning rates

We also tested whether the covariance between allele frequencies at our candidate genes and twinning rates exceeds the expected covariance at neutral loci by the method proposed by Coop et al. [29] implemented in the program BAYENV2 (http://gcbias.org/bayenv/). For that purpose, first, we computed the neutral covariance matrix based on the randomly 10% chosen 5 505 SNPs of the Reference dataset (details above), thus summarizing the pattern of allele frequency variance among the 22 populations, for which twinning rates are available (Supplementary Table 2), according to a simple drift model. By this matrix of population differences (Ω) we could then control for the effects of demographic history when testing for covariance between twinning rates and the population-specific allele frequencies at a given SNP. For each locus BAYENV2 computes the ratio of the posterior probability (PO) of the model of adaptation vs random drift; the PO and the associated Bayes Factor (BF) represent the support for the model of local adaptation with respect to a model of random drift. Note that this parametric method assumes a linear effect of the dependent variable (here: the twinning rate), which can lead to spurious correlations in the presence of strong outliers. However, BAYENV2 also allows us to estimate the non-parametric Spearman's ρ statistic, which is less affected by extreme values. We analyzed the SNPs one by one in the Candidate dataset and determined the distribution of PO, BF and ρ separately. In order to identify candidate polymorphisms showing a strong departure from null expectations, we considered only SNPs for which the Bayes Factor for the model of adaptation vs the null model was higher than 10.We classified the biological type of each the outlier variant thanks to the VEP (Variant Effect Predictor) utility available within the Ensemble tools [30].

Possible effects of linkage disequilibrium

To test for possible effects of linked loci, we mapped levels of linkage disequilibrium upstream and downstream of a subset of 20 candidate genes, using the r^2 statistic [31], by means of the program PLINK. The pairwise LD statistics were calculated between one SNP per gene and all known SNPs falling within a range of 10 Kb up to the end of the nearest upstream gene and 10 Kb down to the start of the nearest downstream gene.

Data archiving

The datafiles used in this study have been stored in the figshare repository (https://figshare.com/). URL: https://figshare.com/s/3d359b57e7300655c934.

RESULTS

The observed F_{ST} values for the Reference dataset range from 0 to 0.60, and the 0.01 upper threshold is 0.21. By contrast, F_{ST} values in the Candidate dataset range from virtually 0 to 0.62. Within the Candidate dataset, 218 F_{ST} values are nominally significant at the P<0.01 level, and 1 149 of them exceed 0.11 (the genome-wide average F_{ST} estimated at 650 000 SNPs in 51 populations of the Human Genome Diversity Panel [32]). By applying an extra “continental-based” MAF filter (as described in Material and methods) we further reduced the data, obtaining the final Candidate dataset used for the subsequent analysis.

We looked for signals of local adaptation comparing neutral and selection models, under the assumption of an island population structure. By BayeScan we found 540 SNPs exceeding the 1% Model-based false discovery threshold, 13 of which also exceeded the Reference-based threshold. Among such outlier SNPs, 64% show $\log_{10}$PO ratios > 3 (Supplementary Table 3), a result generally considered to very strongly support selection affecting the candidate polymorphisms, according to Jeffreys [33] criteria.

Moreover, 63 out of the 83 Candidate genes showed at least one outlier SNP (Figure 4, A-J insets). Considering the locus-specific F_{ST} component (α) value at each outlier SNP, we classified the 63 genes in three classes: (i) Twenty-four of them appeared subjected to: (i) *adaptive selection*, since the outlier SNPs falling within each locus were all α-positive (Figure 4, A-D insets); (ii) Thirty-two appeared to reflect *balancing selection*, since the outlier SNPs falling within each locus were all α-negative (Figure 4, E-H insets); (iii) finally, 7 loci were classified as *ambiguous*, because some of the outlier SNPs falling within each of them wereα-positive and some α-negative (Figure 4, I and J insets). Therefore, in 38% of cases, the results were consistently in support of an *adaptive* selective pressure affecting the gene of interest (further details on these loci are in Supplementary Table 5B).

We then proceeded to test whether SNP frequencies show correlation with twinning frequencies. BAYENV2 identified 114 SNPs in 30 candidate genes with a BF>10 and absolute ρ-value higher than 0.25, both results indicating a strong signal of correlation with the twinning rates (Supplementary Table 4). The gene *BMPR1B* is the most represented (with 19 SNP under the fixed threshold) and the one with the highest BF and ρ values. All the 30 BAYENV2 outlier genes (Figure 4, all insets but A, H and J) are included within the 63 BayeScan outlier genes, and 11 of them are *adaptive* genes (Figure 4, B, C, D insets and Supplementary Table 5A, 5B and 5C).

We searched for the variants with the strongest signal of local adaptation by merging the two sets of outlier SNPs that emerged as significant from the two selection tests described. The 49 common SNPs retrieved are α-positive and fall within 21 genes (Figure 4, C, D, E, F insets). Roughly half of the SNPs fall in intronic regions, 15 in regulatory regions, five close to genes (annotated as upstream and downstream variants); in addition, three variants, two falling in

BMP1 and one in *BMPR1A* are “NMD_transcript_variant” (Non-sense mediated decay target sites).

Visual inspection of these 49 SNP frequency distributions suggests that not all of them are necessarily related with twinning. Indeed, roughly 26% did not show the expected continental pattern of allele frequencies, that is, at each SNP the allele with highest frequencies in African populations has the lowest frequencies in Asian populations, with Europeans and Americans with intermediate values. (see Figure 3, Supplementary Figure 1, Supplementary Figure 2 and Supplementary Table 5D). Conversely, 17 out of the 21 genes (Figure 4, D and E insets) show the expected continental pattern for at least half of their outlier SNPs.

Therefore, we proceeded to screen these genes intersecting the information about the gene type (whether subjected to *adaptive* selection, *ambiguous,* or subjected to *balancing* selection) and about the number of SNPs (at each gene) with the expected continental pattern. Moreover, we picked out only the genes subjected to *adaptive* selection and showing that pattern in at least half of the outlier SNPs (see an example in Figure 3 and details in Supplementary Table 5D and Supplementary Figure 2). Following these criteria, we identified seven genes, namely *BMP15, BMP3, GDF5, IGF2, LHCGR, IGFBPL1, MTHFR*, represented by a total of 13 SNPs (Figure 4, D inset; see also Table 1). The process of selection of the subset of seven loci showing consistent evidence of positive selection under all the criteria we followed is recapituleted in Figure 4. The allele frequencies for the 13 SNPs falling in these 7 genes (Figure 4, D inset), and 36 additional SNPs (Figure 4, C E and F insets) are, respectively, in Supplementary Figures 1 and 2.

Finally, we checked on a subset of 20 candidate genes whether there was reason to believe we could be actually identifying the consequences of positive selection affecting nearby genes. Although we cannot rule it out completely, the results do not suggest that that was the case. Indeed, 3 out of 20 tested genes showed LD at $r^2 > 0.7$ with SNPs in the flanking genes, either upstream or downstream gene (see Figure 5 for examples). , MDM4 shows high levels of LD with the upstream PIK3C2B gene, a member of the PI3K kinase family and a likely candidate for selection, being involved in cell proliferation, oncogenic transformation, cell survival, cell migration, and intracellular protein trafficking; MTHFR is in LD with the C1orf167 uncharacterized protein coding gene; finally, FSHB is flanked by ARL14EP, a gene that controls the export of major histocompatibility class II molecules (annotation from GeneCards Human Gene Database [34]. For the remaining 17 genes there was no evidence that LD may have generated the patterns we observed.

DISCUSSION

DZ twinning is a multifactorial process, whose inheritance pattern has been defined as ‘more complex than previously thought’ in a linkage study identifying regions of potential interest on chromosomes 2, 7 and 18[11]. Mbarek et al. (2016) calculated a significant polygenic risk score which reflected the polygenic contribution to the susceptibility to DZ twinning[9]. Therefore, by considering several genes, we are acknowledging that DZ twinning is best understood as a trait with a polygenic inheritable component.

In this paper, we attempted to answer two related questions: 1. Are population-specific differences in gene frequencies due to allele differences? 2. Are these allele differences the result of positive natural selection? Our work builds upon recent work on the genetic bases of twinning in mothers of twins [8-11] which by necessity had to focus on mothers of twins from specific samples hardly reflective of all of human variation. Our purpose, instead, was to look at the variation in allele frequencies around the world for understanding the causes of differences in DZ twinning rates among populations at a worldwide scale.

With this approach we disregarded balancing selection, for a selection regime maintaining two or more alleles at intermediate frequencies is unlikely to result in the observed cline of twinning rates Across continents. An effect of negative or purifying selection can also be conceived. However, in many cases negative selection will just rapidly eliminate new detrimental alleles generated by mutation. In other cases, negative selection might be the same across all continents, and so it would not result in different twinning rates among populations. Thus, continental differences will emerge only if there is purifying selection against twinning alleles in some, but not all, populations. In this case, however, positive and negative selection are two sides of the same coin, and the alternative allele would be identified as subjected to positive selection.

Results are not always fully consistent, depending on the approach chosen, but there seems to be little doubt that, by and large, variation at the candidate loci departs from expectations based on presumably neutral genome regions. Most candidate loci showed a highly significant excess of between-population variance when compared with an empirical distribution inferred from variation at neutral genomic sites; support for a model including selection at candidate loci is much stronger than that for a neutral model; and 30 loci showed SNP frequencies significantly correlated with twinning rates (see Supplementary Table 5A and 5C).

When we tried to fit by BayeScan under a neutral island model of demography, 63 out of 83 candidate genes, a remarkably high proportion, showed a significant signal of local adaptation. In particular, 38% of such genes showed consistent signals of *adaptive* selective pressure for all the SNPs considered (Figure 4 and Supplementary Table 5A).

Only 20 genes did not show any signal of local adaptation and correlation with twinning rates. Among these genes is the *FSHB* locus, coding for the follicle stimulating hormone (*FSH*) beta subunit, involved in the specific interaction with the *FSHR* receptor and recognized as a highly-conserved vertebrate gene [35]. Actually, one SNP upstream of *FSHB* emerged as strongly associated with twinning in a broad GWAS, considering almost 2 000 mothers of DZ twins and 13 000 controls[9]. The two studies are hardly comparable; here we analyzed 26 worldwide-distributed populations, whereas the GWAS was based on four populations which may largely share ancestors in Northern Europe. Above and beyond that fact, however, there is no reason to expect that the loci accounting for the observed differences among populations (investigated in this study) be also associated with a higher risk of DZ twinning (investigated in Ref. 9).

Linkage disequilibrium seems to have played a negligible or even no role in determining our results. Indeed, only three of the candidate genes tested showed any LD with flanking genes, and only *MTHFR* emerged as significantly correlated with twinning rates, and likely subjected to selection. Therefore, if LD affected our results, it did so only to a very limited extent.

As a whole, these results achieve an extremely high level of statistical significance; the patterns we identified cannot be accounted for by processes occurring by chance through demographic history. Variation in allelic frequencies of these genes may point to slight differences for the bioactivity of these enzymes, hormones and receptors and in turn to their influence in the rate of twinning in the populations here considered. However, we focus our discussion on seven loci, namely *BMP15*, *BMP3*, *GD5*, *IGF2*, *LHCGR*, *IGFBPL1*, and *MTHFR* which: (a) showed strong signals of local adaptation; (b) resulted as genes subjected to *adaptive* selective pressures; (c) showed parallel patterns of twinning rates and allele frequencies (Figure 4 and Supplementary Table 5D).

There are two pathways by which multiple ovulation may occur: the promotion of a larger cohort of follicles, or fewer inhibitions of follicle selection and atresia [36]. That is, either an individual with DZ twins can make more follicles to start, or she discards fewer of them. Then, there appears to be variation in the extent to which the endometrium can support multiple implantation. Inhibins, activins, the *p53* pathway, and *LHCG* receptors are all implicated in processes of endometrial receptivity and implantation [37]. Therefore, variations in their bioactivity could lead to greater or lesser support of multiple implantation. It is not surprising that these systems were found to be implicated in this study.

Out of our seven genes, two code for bone morphogenetic proteins (*BMP15* and *BMP3*), two for growth-differential factors (*GDF5* and *IGF2*), one codes for the receptor of luteinizing hormone and choriogonadotropin (*LHCGR)*, one is an insulin-like growth-factor binding protein (*IGFBPL1*), and another one is essential in folate metabolism (*MTHFR*). Bone morphogenetic protein 15 (*BMP15*) is necessary for normal ovarian development and folliculogenesis. This gene is expressed in the oocytes within the ovary, with expression increasing as the oocytes mature [38]. A previous study of BMP15 showed association between DZ twinning and a common intronic variant, but this result did not reach significance after correction for multiple testing[39]. Bone morphogenetic protein 3 (*BMP3*) has not been previously associated with folliculogenesis. However, Monestier et al. propose that two bone morphogenetic protein and two growth-differential factors share a common evolutionary origin, namely *BMP3*/*GDF10* and *BMP15*/*GDF9*. They note that it is likely that the four genes help explain ovulation variability in poly and uni-ovulatory mammals [40].

Growth differentiation factor 5 (*GDF5*) is a regulator of cell growth and differentiation in embryonic tissue, particularly skeletal and joint development [41]. Its contribution to twinning probably stems from its critical role in endometrial decidualization, which is important to embryo implantation [42]. Insulin-like growth-factor 2 receptor (*IGF2*) has been shown to be directly involved in inducing steroidogenesis in bovine granulosa cells [43].

LHCGR produces a protein which acts as a receptor for both luteinizing hormone (LH) and human chorionic gonadotropin (*hCG*). *LHCGR* may be involved in DZ twinning because LH triggers the release of eggs from the ovaries. Moreover, *hCG* is produced by fetal tissue during pregnancy and is necessary for the pregnancy to continue [44]. Thus, *LHCGR* may be involved in DZ twinning by triggering multiple ovulations or by maintaining multiple pregnancies.

The insulin and insulin-like factor binding protein family includes at least six proteins (*IGFBP1-6*) which are expressed in human follicles at varying stages of development as well as in the endometrium [45,46]. *IGFBP*s regulate IGF in the ovary [47,48], and the stromal cells of the endometrium during implantation [46]. In the ovary, *IGFBP* is important to dominant follicle selection. *IGFBP*s multiply the expression of IGF-*II* from the dominant follicle, and sequester IGF in subordinate follicles [47]. In the endometrium, IGFBPs regulate decidualization of the endometrium and thus the timing of endometrial receptivity [49,50]. Thus, variation in IGFBP could influence the number of follicles selected or the receptivity of the endometrium to multiple embryos.

Finally, the *MTHFR* enzyme is essential for folate metabolism. Women who are homozygotes of a common variant of the *MTHFR* enzyme, the C677T mutation, have lower serum estradiol concentrations at ovulation during infertility treatments, produce fewer oocytes for retrieval, and are known to have a lower risk of multiple pregnancies. The *MTHFR* gene may be affecting twinning frequencies by affecting ovarian function, implantation or even embryogenesis since folate metabolism is involved at all stages [51]. This mutation is almost absent in African populations, which helps explain why this gene shows such clear geographical distribution. Previous work in Australasian and Dutch families failed to find evidence of association between MTHFR genotypes and twinning in mothers of twins[52] .

The next question, then, is whether twinning is advantageous from a Darwinian fitness perspective in humans. The existing literature is mixed. It is well known that twin pregnancies are associated with preterm and premature labor, higher risk of pre-eclampsia and pregnancy-induced hypertension and gestational diabetes, and other maternal and fetal/neo-natal complications [53]. However, mothers of twins under two different socio-cultural and ecological conditions have been reported to achieve higher selective fitness than mothers of singletons, suggesting that at least in some environments, the genetic propensity for twinning may be favored [54,55].

This study is an attempt to get insight into the genetic factors accounting for the higher DZ twinning rates in Africa than in Asia. It is reasonable to think the seven loci identified as associated with these continental differences, and most likely reflecting differences in selection regimes, are involved in the causation of DZ twinning. However, the reverse is not true. Other loci, perhaps many, are likely to contribute to the phenomenon of DZ twinning, but cannot emerge in our analysis if their allele frequencies are the same across continents

The results of our study raise profound questions about variability of the life history patterns of our species. While it has been well known that non-industrialized human populations will control their fertility with various cultural means such as post-partum taboos and effective breast-feeding patterns, a demonstration that human populations differ significantly in their fertility-related genetic make-up is new. Here we have shown that human populations differ in gene frequencies of several genes potentially related with twinning and that these frequencies are significantly correlated with DZ twinning. A recent study has shown that in Guinea Bissau the perinatal death rate is higher for DZ twins [56]. In the long run, this should lead to a decrease in twinning frequencies, which has not actually been observed[56]. A logical implication of our and that study is that, at least in some African groups, natural selection has been acting to increase the frequency of twinning because it, or a trait co-selected with it, confers some degree of selective advantage.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Silvia Ghirotto for her help and critical comments on this work. This paper was partly supported by the European Research Council ERC-2011-AdG_295733 grant (LanGeLin) to GB.

1. Bulmer MG. *The Biology Of Twinning In Man.* Clarendon, Oxford, 1970.
2. Pollard R. Ethnic variation of twinning rates in Malawi. *Acta Genet Med Gemel* 1996; **45:** 361-365.
3. Pollard R. Twinning rates in Fiji. *Ann Hum Genet* 1985; **49:** 65-73.
4. Nylander PPS. The factors that influence twinning rates. *Acta Genet Med Gemel* 1981; **30:** 189-202.
5. Madrigal L, Saenz G, Chavez M and Dykes D. Frequency of twinning in two Costa Rican ethnic groups: An update. *Am J Hum Biol* 2001; **13:** 220-226.
6. Hoekstra C, Zhao ZZ, Lambalk CB, *et al.* Dizygotic twinning. *Hum Reprod Update* 2008; **14:** 37-47.
7. Duffy DL, Montgomery GW, Hall J, *et al.* Human Twinning is not linked to the region of chromosome 4 syntenic with the sheep twinning gene FecB. *Am J Med Genet* 2001; **100:** 182-186.
8. Huang H, Clancy KBH, Burhance C, Zhu Y and Madrigal L. Women who deliver twins are more likely to smoke and have high frequencies of specific SNPs: Results from a sample of African-American women who delivered preterm, low birth weight babies. *Am J Hum Biol* 2015; **27:** 605-612.
9. Mbarek H, Steinberg S, Nyholt DR, *et al.* Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. *Am J Hum Genet* 2016; **98:** 898-908.
10. Painter JN,Willemsen G, Nyholt D, *et al.* A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins. *Hum Reprod* 2010; **25:** 1569-1580.

11. Derom C, Jwaheer D, Chen WV, *et al.* Genome-wide linkage scan for spontaneous DZ twinning. *Eu J Hum Gent* 2006; **14:** 117-122.

12. Sirugo G, Edwards DRV, Ryckman KK, *et al.* PTX3 Genetic Variation and Dizygotic Twinning in The Gambia: Could Pleiotropy with Innate Immunity Explain Common Dizygotic Twinning in Africa? *Ann Hum Genet* 2012; **76:** 454-463.

13. Auton A, Brooks LD, Drubin RM, *et al.* A global reference for human genetic variation. *Nature* 2015; **526:** 68-74.

14. Harris RA, Tardiff SD, Vinar T, *et al.* Evolutionary genetics and implications of small size and twinning in callitrichine primates. *P Natl Acad Sci USA* 2014; **111:** 1467-1472.

15. Wildeman M, van Ophuizen E, den Dunnen JT and Taschner PEM. Improving sequence variant descriptions in mutation Databases and literature using the mutalyzer sequence variation nomenclature checker. *Human Mutation* 2008; **29:** 6-13.

16. Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.

17. Cavalli-Sforza LL. Population structure and human evolution. *P R Soc B* 1966; **164:** 362-379.

18. Coop G,Pickrell JK, Novembre J, *et al.* The Role of Geography in Human Adaptation. *Plos Genet* 2009; **5**: Article Number: e1000500.

19. Pickrell JK, Coop G, Novembre J, *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Gen Res* 2009; **19:** 826-837.

20. Arbiza L**,** Zhong E and Keinan A. NRE: a tool for exploring neutral loci in the human genome. *Bmc Bioinformatics* 2012; **13:** Article Number: 301.

21. Patin E, Laval G, Barreiro L, *et al.* Inferring the Demographic History of African Farmers and Pygmy Hunter-Gatherers Using a Multilocus Resequencing Data Set. *Plos Genet* 2009; **5:** Article Number: e1000448.

22. Paradis E. Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 2010; **26:** 419-420.

23. Weir BS and Cockerham CC. Estimating F-statistics for the analysis of population-structure. *Evolution* 1984; **38:** 1358-1370.

24. Foll M and Gaggiotti O. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* 2008; **180:** 977-993.

25. Dall'Ara I, Girotto S, Ingusci S, *et al.* Demographic history and adaptation account for clock gene diversity in humans. *Heredity* 2016; **117:** 165-172.

26. Pais A, Whetten R and Xiang Q. Ecological genomics of local adaptation in Cornus florida L. by genotyping by sequencing. *Ecol. Evol* 2016; **20:** 441-465.

27. Schweizer RM, Vonholdt BM, Harrigan R, *et al.* Genetic subdivision and candidate genes under selection in North American grey wolves. *Mol Ecol* 2016; **25:** 380-402.

28. Shim HJ, Laurent S, Matuszewski S, Foll M and Jensen JD. Detecting and Quantifying Changing Selection Intensities from Time-Sampled Polymorphism Data. *G3* 2016; **6:** 893-904.

29. Coop G, Witonsky D, Di Rienzo A and Pritchard JK. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics* 2010; **185:** 1411-1423.

30. McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. *Genom Biol* 2016; **17:** Article Number: 122.

31. Devlin B and Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29:** 311-322.

32. Li JZ, Absher DM, Tang H, *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319:** 1100-1104.

33. Jeffreys H. *Theory of probability*. Oxford University Press, Oxford. 1961

34. Rebhan M, Chalifa-Caspi V, Prilusky J and Lancet D. GeneCards: Integrating information about genes, proteins and diseases. *Trends Genet* 1997; **13:** 163-163.

35. Nagirnaja L, Rull K, Uuskuela L, *et al.* Genomics and genetics of gonadotropin beta-subunit genes: Unique FSHB and duplicated LHB/CGB loci. *Mol Cell Endo* 2010; **329:** 4-16.

36. Lambalk CB, Boomsma DJ, De Boer L, *et al.* Increased levels and pulsatility of follicle-stimulating hormone in mothers of hereditary dizygotic twins. *J Clin Endocr Metab* 1998; **83:** 481-486.

37. Florio P, Luisi, S, Ciarmela P, *et al.* Inhibins and activins in pregnancy. *Mol Cell Endo* 2004; **225:** 93-100.

38. de Castro FC, Cruz MHC and Leal CLV. Role of Growth Differentiation Factor 9 and Bone Morphogenetic Protein 15 in Ovarian Function and Their Importance in Mammalian Female Fertility - A Review. *Asian-Austral J Anim Sc* 2016; **29:** 1065-1074.

39. Zhao ZZ, Painter JN, Palmer JS, *et al.* Variation in bone morphogenetic protein 15 is not associated with spontaneous human dizygotic twinning. *Hum Reprod* 2008; **23:**. 2372-2379.

40. Monestier O, Servin B, Auclair S, *et al.* Evolutionary Origin of Bone Morphogenetic Protein 15 and Growth and Differentiation Factor 9 and Differential Selective Pressure

Between Mono- and Polyovulating Species. *Biology of Reproduction* 2014; **91:** Article Number: 83.

41. Peng J, Wigglesworth K. Rangarajan A, *et al.* Amino Acid 72 of Mouse and Human GDF9 Mature Domain Is Responsible for Altered Homodimer Bioactivities but Has Subtle Effects on GDF9:BMP15 Heterodimer Activities. *Biol Reprod* 2014; **91:** Article Number: 142.

42. Kanamarlapudi V, Gordon UD and Bernal AL. Luteinizing hormone/chorionic gonadotrophin receptor overexpressed in granulosa cells from polycystic ovary syndrome ovaries is functionally active. *Reproduct Biomed Online* 2016; **32:** 635-641.

43. Pyun JA, Kim S, Cha DH and Kwack K. Epistasis between IGF2R and ADAMTS19 polymorphisms associates with premature ovarian failure. *Hum Reprod* 2013; **28:** 3146-3154

44. Choi J and Smitz J. Luteinizing hormone and human chorionic gonadotropin: Origins of difference. *Mol Cell Endo* 2014; **383:** 203-213.

45. Brogan RS, Mix S, Puttabyatappa M, VandeVoort CA and Chaffin CL. Expression of the insulin-like growth factor and insulin systems in the luteinizing macaque ovarian follicle. *Fertil Steril* 2010; **93:** 1421-1429.

46. Lathi RB, Hess AP, Tulac S, *et al.* Dose-dependent insulin regulation of insulin-like growth factor binding protein-1 in human endometrial stromal cells is mediated by distinct signaling pathways. *J Clin Endo Metabol* 2005; **90:** 1599-1606.

47. Baerwald AR, Adams GP and Pierson RA. Ovarian antral folliculogenesis during the human menstrual cycle: a review. *Hum Reprod Update* 2012; **18:** 73-91.

48. Silva JRV, Figueiredo JR and van den Hurk R. Involvement of growth hormone (GH) and insulin-like growth factor (IGF) system in ovarian folliculogenesis. *Theriogenology* 2009; **71:** 1193-1208.

49. Ganeff C. Chatel G, Munaut C, *et al*. The IGF system in in-vitro human decidualization. *Mol Hum Reprod* 2009; **15:** 27-38.

50. Kutsukake M, Ishihara R, Yoshie M, Kogo . and Tamura K. Involvement of insulin-like growth factor-binding protein-related protein 1 in decidualization of human endometrial stromal cells. *Mol Hum Reprod* 2007; **13:** 737-743 .

51. Thaler CJ. Folate Metabolism and Human Reproduction. *Geburtshi Frauenheilk* 2014; **74:** 845-851.

52. Montgomery GW, Zhao ZZ, Morley KL, *et al*. Dizygotic twinning is not associated with methylenetetrahydrofolate reductase haplotypes. *Hum Reprod* 2003; **18:** 2460-2464.

53. Wennerholm UB. The risks associated with multiple pregnancies. In Gerris J, Adamson G, Sutter PD and Racowsky C (eds.) *Single Embryo Transfer* 3-16. Cambridge University Press. Cambridge, UK. 2009.

54. Madrigal L. Differential fertility of mothers of twins and mothers of singletons - study in Limon, Costa Rica. *Hum Biol* 1995; **67:** 779-787.

55. Sear R, Shanley D, McGregor I. and Mace R. The fitness of twin mothers: evidence from rural Gambia. *J Evol Biol* 2001; **14:** 433-443.

56. Bjerregaard-Andersen M, Lund N, Jepsen, FS, *et al.* A prospective study of twinning and perinatal mortality in urban Guinea-Bissau. *Bmc Preg Childb* 2012; **12:** Article Number: 140.

Legends to figures and Table caption

Figure 1 Map of 1000G sampled populations together with twinning rate information. The black squares indicate populations for which information about twinning rates was not available.

Figure 2 Chromosomes ideogram scheme showing the 83 Candidate gene mapping sites.

Figure 3 Frequency distributions for one SNP of four candidate loci. The maps at the top show two out of the 13 outlier SNPs in Table 1 (see also Supplementary Figure 1). The maps at the bottom show two out of the 36 outlier SNPs in Supplementary Figure 2 (see also Supplementary Table 5D).

Figure 4 Venn diagram summarizing the procedure to identify seven loci (D inset), for which there is evidence of adaptive selection. A: Loci identified by BayeScan (BS); B: Loci also identified by BAYENV2 (BE); C: Loci whose SNPs have been identified both by BS and BE; D: Loci whose SNPs have been identified both by BS and BE and with at least half SNPs showing a continental pattern associated with rates of DZ twinning.

Figure 5 LD plots for four candidate loci tested. A: MDM4, B: FSHB, C: TP53, D: ACVR1.

Table 1 List of the 13 SNPs in the 7 genes for which consistent evidence of positive selection was found in all analyses.

Supplementary Figure 1 Frequency distributions for 13 alleles at the 7 outlier loci described in Table 1.

Supplementary Figure 2 Frequency distributions for 36 alleles at 21 additional loci, identified as subjected to selection by both BayeScan and BAYENV2,

Supplementary Table 1 Candidate twinning genes considered. Genes without source reference were retrieved on the UCSC genome browser (h19 release) and added to the analysis in order to complete the gene family panel.

Supplementary Table 2 Populations in the dataset and twinning rates source

Supplementary Table 3 BayeScan result for significant SNPs (qval<0.01)

Supplementary Table 4 BAYENV2 result for significant SNPs (BF>10)

Supplementary Table 5A Summary results

Supplementary Table 5B BayeScan 63 outlier genes

Supplementary Table 5C BAYENV2 30 outliers genes

Supplementary Table 5D 49 common SNPs summary results

Figure 1. Map of 1000G sampled populations together with twinning rate information. The black squares indicate populations for which information about twinning rates was not available.

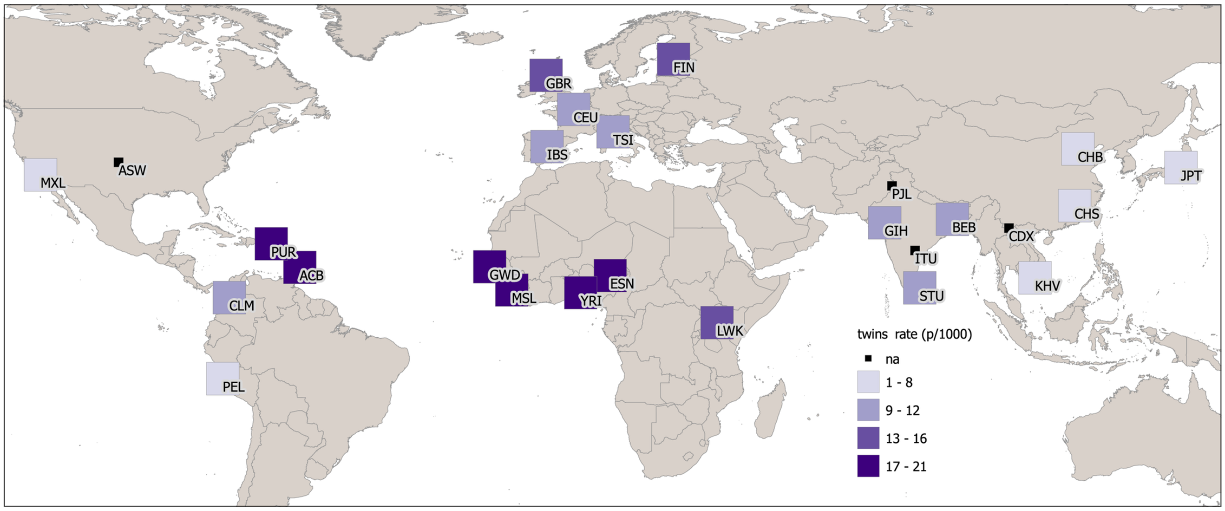

Figure 2. Chromosomes ideogram scheme showing the 83 Candidate gene mapping sites.

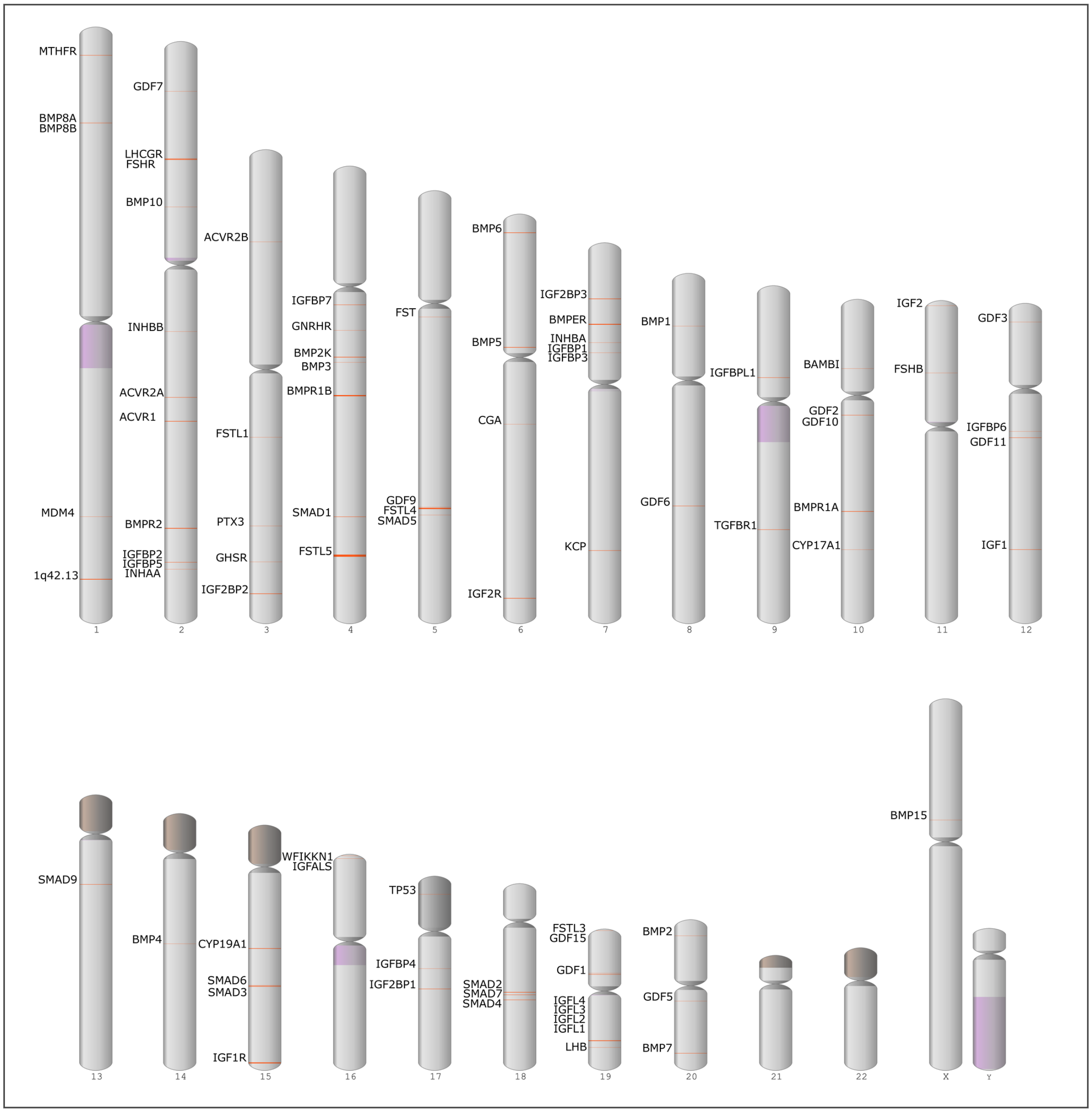

Figure 3. Frequency distributions for one SNP of four candidate loci. The maps at the top show two out of the 13 outlier SNPs in Table 1

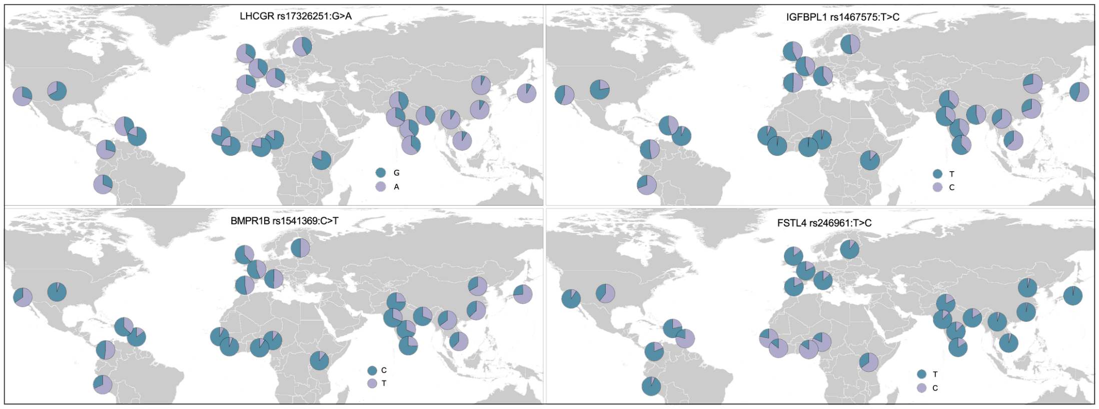

Figure 4. Venn diagram summarizing the procedure to identify seven loci (D inset), for which there is evidence of adaptive selection. A: Loci identified by BayeScan (BS); B: Loci also identified by BAYENV2 (BE); C: Loci whose SNPs have been identified both by BS and BE; D: Loci whose SNPs have been identified both by BS and BE and with at least half SNPs showing a continental pattern associated with rates of DZ twinning.

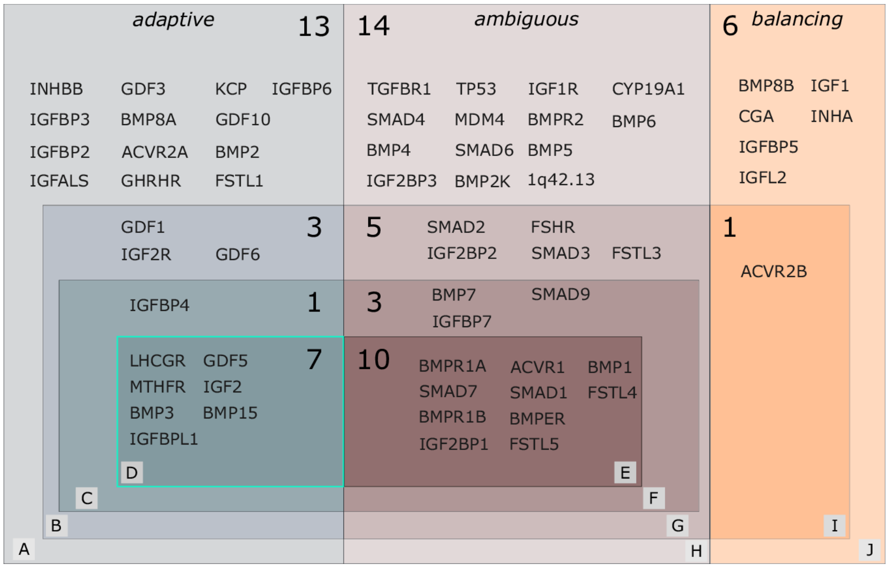

Figure 5. LD plots for four candidate loci tested. A: MDM4, B: FSHB, C: TP53, D: ACVR1.

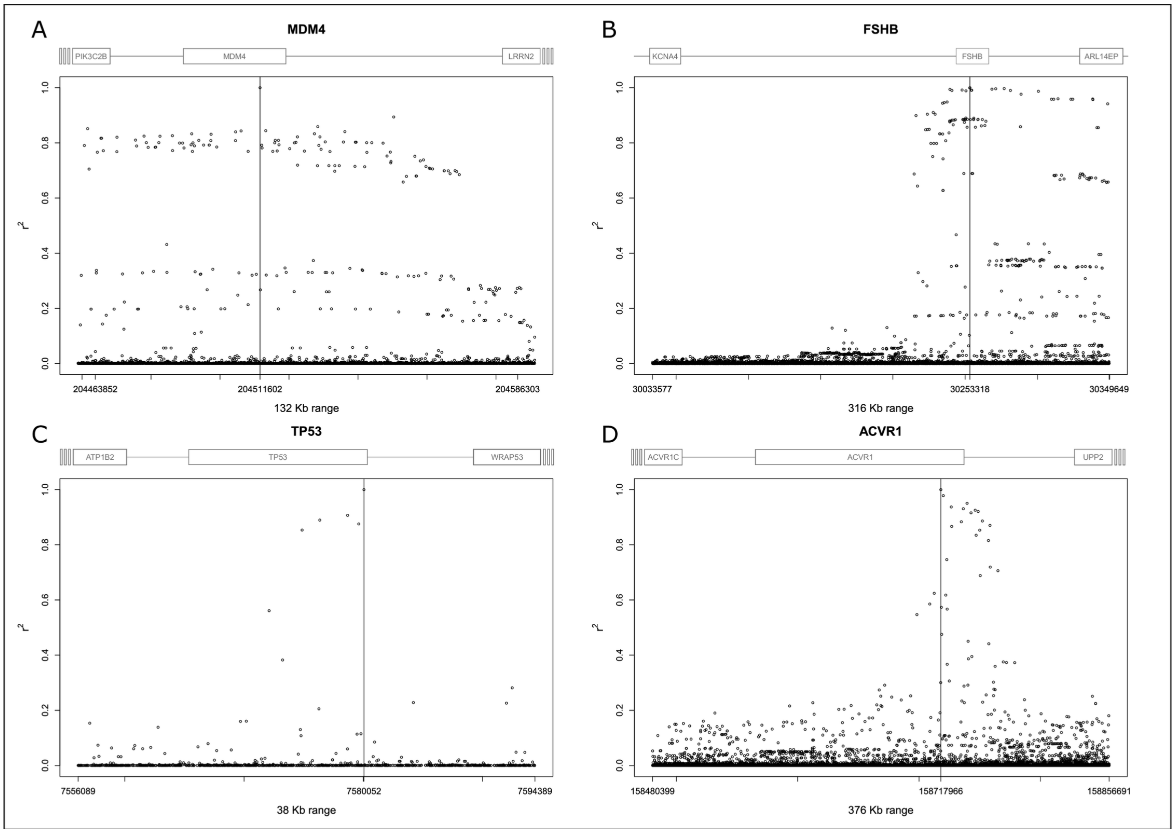

Table 1. List of the 13 SNPs in the 7 genes for which consistent evidence of positive selection was found in all analyses

SNP	hg19_HGVS	gene	Variant type	MAJOR ALLELE in AFRICAN POPULATIONS	MAJOR ALLELE in ASIAN POPUPATIONS
rs3897937	chrX:g.50655016A>G	BMP15	intronic	G	A
rs12642476	chr4:g.81966095T>C	BMP3	intronic	T	C
rs5022942	chr4:g.81959966A>G	BMP3	intronic	A	G
rs224333	chr20:g.34023962G>A	GDF5	intronic	A	G
rs3741212	chr11:g.2161858A>G	IGF2	upstream gene	A	G
rs1467575	chr9:g.38422045T>C	IGFBPL1	regulatory region	T	C
rs17037566	chr2:g.48940073A>G	LHCGR	intronic	G	A
rs17326251	chr2:g.48916822A>G	LHCGR	intronic	A	G
rs4637174	chr2:g.48960317A>G	LHCGR	intronic	G	A
rs6733079	chr2:g.48935849T>C	LHCGR	intronic	C	G
rs3737966	chr1:g.11847759C>T	MTHFR	3' UTR; regulatory region	C	T
rs4846052	chr1:g.11857951T>C	MTHFR	regulatory region	T	C
rs865907	chr1:g.230644086G>A	MTHFR	regulatory region	A	G

Abbreviations

ABC Approximate Bayesian computation

aDNA ancient DNA

bp Base Pairs

DAPC Discriminant Analysis Principal Component

HVR-1 Hypervariable Region-1

IE Indo-European

ky thousand years

LGM Last Glacial Maximum

mtDNA mitochondrial DNA

NGS Next-Generation Sequencing

NRY non-recombining portion of the Y chromosome

PCA Principal Component Analysis

PCM Parametric Comparison Method

PCR Polymerase Chain Reaction

SNP Single Nucleotide Polymorphisms

ya years ago

List of Figures

List of Tables

Bibliography

Abe-Sandes, K., Silva, W. A. & Zago, M. A. (2004), 'Heterogeneity of the Y chromosome in Afro-Brazilian populations', *Hum. Biol.* **76**(1), 77–86.

Adriaensen, F., Chardon, J., De Blust, G., Swinnen, E., Villalba, S., Gulinck, H. & Matthysen, E. (2003), 'The application of 'least-cost' modelling as a functional landscape model', *Landscape and Urban Planning* **64**(4), 233–347–567.

Ammerman, A. & Cavalli-Sforza, L. (1984), 'The Neolithic transition and the genetics of populations in Europe', *Princeton: Princeton University Press* .

Anderson, S., Bankier, A., Barrell, B. et al. (1981), 'Sequence and organization of the human mitochondrial genome', *Nature* **290**(5860), 457–465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999), 'Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA', *Nat. Genet.* **23**(2), 147.

Atkinson, Q. D. & Gray, R. D. (2005), 'Curious parallels and curious connections–phylogenetic thinking in biology and historical linguistics', *Syst. Biol.* **54**(4), 513–526.

Bagley, R. K., Sousa, V. C., Niemiller, M. L. & Linnen, C. R. (2016), 'History, geography, and host use shape genome-wide patterns of genetic variation in the redheaded pine sawfly (Neodiprion lecontei)', *Mol. Ecol.* .

Baker, M. (2001), *The Atoms of Language.*

Barbujani, G., Bertorelle, G., Capitani, G. & Scozzari, R. (1995), 'Geographical structuring in the mtDNA of Italians', *Proc. Natl. Acad. Sci. U.S.A.* **92**(20), 9171–9175.

Barbujani, G. & Goldstein, D. B. (2004), 'Africans and Asians abroad: genetic diversity in Europe', *Annu Rev Genomics Hum Genet* **5**, 119–150.

Barbujani, G., Nasidze, I. S. & Whitehead, G. N. (1994), 'Genetic diversity in the Caucasus', *Hum. Biol.* **66**(4), 639–668.

Barbujani, G. & Pilastro, A. (1993), 'Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily', *Proc. Natl. Acad. Sci. U.S.A.* **90**(10), 4670–4673.

Barbujani, G. & Sokal, R. R. (1990), 'Zones of sharp genetic change in Europe are also linguistic boundaries', *Proc. Natl. Acad. Sci. U.S.A.* **87**(5), 1816–1819.

Barbujani, G. & Sokal, R. R. (1991), 'Genetic population structure of Italy. II. Physical and cultural barriers to gene flow', *Am. J. Hum. Genet.* **48**(2), 398–411.

Bateman, R., Goddard, I., O'Grady, R., Funk, V., Mooi, R., Kress, V. & Cannel, P. (1990), 'Speaking of Forked Tongues. The Feasibility of Reconciling Human Phylogeny and the History of Language', *Curr Anthropol* **31**(1).

Battaggia, C., Ruscitto, D., Destro-Bisol, G., Vacca, L., Calo, C. & Vona, G. (2003), 'Frequencies at CD4, FES, and F13A1 microsatellite loci in central-southern Sardinia (Italy)', *J. Forensic Sci.* **48**(2), 442.

Belle, E. M. & Barbujani, G. (2007), 'Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity', *Am. J. Phys. Anthropol.* **133**(4), 1137–1146.

Benazzi, S., Slon, V., Talamo, S. et al. (2015), 'Archaeology. The makers of the Protoaurignacian and implications for Neandertal extinction', *Science* **348**(6236), 793–796.

Biberauer, T. (2008), 'The Limits of Syntactic Variation', *Amsterdam/Philadelphia, Jhon Benjamins* .

Bollongino, R., Nehlich, O., Richards, M. P., Orschiedt, J., Thomas, M. G., Sell, C., Fajkosova, Z., Powell, A. & Burger, J. (2013), '2000 years of parallel societies in Stone Age Central Europe', *Science* **342**(6157), 479–481.

Bolnick, D., Shook, B., Campbell, L. & Goddard, I. (2004), 'Problematic use of Greenberg's linguistic classification of the Americas in studies of native [A]merican genetic variation', *Am J Hum Genet* **75**, 519–522.

Bonatto, S. L. & Salzano, F. M. (1997), 'A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data', *Proc. Natl. Acad. Sci. U.S.A.* **94**(5), 1866–1871.

Borg, I. & Groenen, P. (2005), *Modern Multidimensional Scaling: theory and applications*, Vol. 2nd Edition.

Bortolussi, L., Longobardi, G., Guardiano, C. & Sgarro, A. (2011), 'How many possible languages are there? In: Bel-Enguix G, Jimenez-Lopez MD, editors.', *Biology, computation and linguistics. Amsterdam: IOS Press* pp. 168–179.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. & Atkinson, Q. D. (2012), 'Mapping the origins and expansion of the Indo-European language family', *Science* **337**(6097), 957–960.

Brassel, K. & Reif, D. (1979), 'A procedure to generate Thiessen polygons', *Geogr. Anal.* **325**, 31–36.

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987), 'Mitochondrial DNA and human evolution', *Nature* **325**(6099), 31–36.

Capelli, C., Arredi, B., Baldassari, L. et al. (2006), 'A 9-loci Y chromosome haplotype in three Italian populations', *Forensic Sci. Int.* **159**(1), 64–70.

Caramelli, D., Vernesi, C., Sanna, S. et al. (2007), 'Genetic variation in prehistoric Sardinia', *Hum. Genet.* **122**(3-4), 327–336.

Cavalli-Sforza, L. L. & Piazza, A. (1993), 'Human genomic diversity in Europe: a summary of recent research and prospects for the future', *Eur. J. Hum. Genet.* **1**(1), 3–18.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988*a*), 'Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data', *Proc. Natl. Acad. Sci. U.S.A.* **85**(16), 6002–6006.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988*b*), 'Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data', *Proc. Natl. Acad. Sci. U.S.A.* **85**(16), 6002–6006.

Cavalli-Sforza, L., Menozzi, P. & Piazza, A. (1994), 'The history and geography of human genes', *Princeton University Press* .

Chen, J., Sokal, R. R. & Ruhlen, M. (1995), 'Worldwide analysis of genetic and linguistic relationships of human populations', *Hum. Biol.* **67**(4), 595–612.

Chomsky, N. (1981), 'Lectures on Government and Binding', *Dordrecht, Foris* .

Colonna, V., Boattini, A., Guardiano, C., Dall'ara, I., Pettener, D., Longobardi, G. & Barbujani, G. (2010), 'Long-range comparison between genes and languages based on syntactic distances', *Hum. Hered.* **70**(4), 245–254.

Contu, D., Morelli, L., Santoni, F., Foster, J. W., 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**(1), e1430.

D'Amore, G., Di Marco, S., Floris, G., Pacciani, E. & Sanna, E. (2010), 'Craniofacial morphometric variation and the biological history of the peopling of Sardinia', *Homo* **61**(6), 385–412.

Di Gaetano, C., Fiorito, G., Ortu, M. F. et al. (2014), 'Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection', *PLoS ONE* **9**(3), e91237.

Edgar, R. C. (2004), 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Res.* **32**(5), 1792–1797.

Ermini, L., Der Sarkissian, C., Willerslev, E. & Orlando, L. (2015), 'Major transitions in human evolution revisited: a tribute to ancient DNA', *J. Hum. Evol.* **79**, 4–20.

Excoffier, L. & Lischer, H. E. (2010), 'Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows', *Mol Ecol Resour* **10**(3), 564–567.

Falchi, A., Giovannoni, L., Calo, C. M., Piras, I. S., Moral, P., Paoli, G., Vona, G. & Varesi, L. (2006), 'Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms', *J. Hum. Genet.* **51**(1), 9–14.

Falchi, M., Forabosco, P., Mocci, E. et al. (2004), 'A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia', *Am. J. Hum. Genet.* **75**(6), 1015–1031.

Fenu, P., Martini, F. & G., P. (1999), *I siti paleolitici: i complessi industriali. Sa Pedrosa-Pantallinu. In: Sardegna paleolitica. Studi sul più antico popolamento dell'isola*, Vol. Museo Fiorentino di Preistoria "Paolo Graziosi" (eds Martini FE).

Floris, G. (1981), *Sulla variabilità dell'indice nasale dei protosardi*, Vol. 21.

Floris, G. (1983), 'La stautura nella protostoria sarda', *Arch. Antrop. Etnol.* **113**, 263–267.

Francalacci, P., Morelli, L., Angius, A. et al. (2013), 'Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny', *Science* **341**(6145), 565–569.

Francalacci, P., Morelli, L., Underhill, P. A. et al. (2003), 'Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability', *Am. J. Phys. Anthropol.* **121**(3), 270–279.

Fregel, R. & Delgado, S. (2011), 'HaploSearch: a tool for haplotype-sequence two-way transformation', *Mitochondrion* **11**(2), 366–367.

Fu, Q., Hajdinjak, M., Moldovan, O. T. et al. (2015), 'An early modern human from Romania with a recent Neanderthal ancestor', *Nature* **524**(7564), 216–219.

Fu, Q., Li, H., Moorjani, P. et al. (2014), 'Genome sequence of a 45,000-year-old modern human from western Siberia', *Nature* **514**(7523), 445–449.

Fu, Q., Mittnik, A., Johnson, P. L. et al. (2013), 'A revised timescale for human evolution based on ancient mitochondrial genomes', *Curr. Biol.* **23**(7), 553–559.

Fu, Q., Rudan, P., Paabo, S. & Krause, J. (2012), 'Complete mitochondrial genomes reveal neolithic expansion into Europe', *PLoS ONE* **7**(3), e32473.

Gansauge, M. T. & Meyer, M. (2014), 'Selective enrichment of damaged DNA molecules for ancient genome sequencing', *Genome Res.* **24**(9), 1543–1549.

Gassin, B. & Lugliè, C. (2012), *La preistoria e la protostoria della Sardegna. Atti della XLIV R.S. IIPP II.*

Germanà, F. (1995), *L'uomo in Sardegna dal Paleolitico all'Età nuragica.*

Ghirotto, S., Mona, S., Benazzo, A., Paparazzo, F., Caramelli, D. & Barbujani, G. (2010), 'Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia', *Mol. Biol. Evol.* **27**(4), 875–886.

Goncalves, V. F., Carvalho, C. M., Bortolini, M. C., Bydlowski, S. P. & Pena, S. D. (2008), 'The phylogeography of African Brazilians', *Hum. Hered.* **65**(1), 23–32.

Gonzalez-Andrade, F., Sanchez, D., Gonzalez-Solorzano, J., Gascon, S. & Martinez-Jarreta, B. (2007), 'Sex-specific genetic admixture of Mestizos, Amerindian Kichwas, and Afro-Ecuadorans from Ecuador', *Hum. Biol.* **79**(1), 51–77.

Gray, R. D. & Atkinson, Q. D. (2003), 'Language-tree divergence times support the Anatolian theory of Indo-European origin', *Nature* **426**(6965), 435–439.

Green, R., Briggs, A., Krause, J. et al. (2009), 'The Neandertal genome and ancient DNA authenticity', *EMBO Journal* **28**(17), 2494–2502.

Green, R., Krause, J., Briggs, A. et al. (2010), 'A draft sequence of the Neandertal genome', *Science* **5979**(328), 710–722.

Greenhill, S. J., Atkinson, Q. D., Meade, A. & Gray, R. D. (2010), 'The shape and tempo of language evolution', *Proc. Biol. Sci.* **277**(1693), 2443–2450.

Grimaldi, M. C., Crouau-Roy, B., Amoros, J. P., Cambon-Thomsen, A., Carcassi, C., Orru, S., Viader, C. & Contu, L. (2001), 'West Mediterranean islands (Corsica, Balearic islands, Sardinia) and the Basque population: contribution of HLA class I molecular markers to their evolutionary history', *Tissue Antigens* **58**(5), 281–292.

Haak, W., Lazaridis, I., Patterson, N. et al. (2015), 'Massive migration from the steppe was a source for Indo-European languages in Europe', *Nature* **522**(7555), 207–211.

Hagelberg, E., Hofreiter, M. & Keyser, C. (2015), 'Introduction. Ancient DNA: the first three decades', *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **370**(1660), 20130371.

Hall, T. (1999), 'BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT', *Nucl. Acids. Symp. Ser.* **41**, 95–98.

Heine, B. & Nurse, D. (2000), 'African Languages. An Introduction', *Cambridge, Cambridge University Press* .

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A. & Wilson, A. C. (1984), 'DNA sequences from the quagga, an extinct member of the horse family', *Nature* **312**(5991), 282–284.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *The Journal of Educational Psychology* **24**, 417–441.

Hunley, K. & Long, J. C. (2005), 'Gene flow across linguistic boundaries in Native North American populations', *Proc. Natl. Acad. Sci. U.S.A.* **102**(5), 1312–1317.

Jaccard, P. (1901), 'Etude comparative de la distribution ?oraledans une portion des alpes et des jura', *Bull De La SocVaudoise Des Sci Nat* (37), 547–579.

Jaworska, N. & Chupetlovska-Anastosova, A. (2009), 'A review of Multidimensional Scaling (MDS) and its utility in various psychological domains', *Tutorials in Quantitative Methods for Psychology* pp. 1–10.

Jobling, M. A. & Tyler-Smith, C. (1995), 'Fathers and sons: the Y chromosome and human evolution', *Trends Genet.* **11**(11), 449–456.

Jobling, M. A. & Tyler-Smith, C. (2003), 'The human Y chromosome: an evolutionary marker comes of age', *Nat. Rev. Genet.* **4**(8), 598–612.

John, J. S. (2011), 'SeqPrep. eds'.

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.

Jones, E. R., Gonzalez-Fortes, G., Connell, S. et al. (2015), 'Upper Palaeolithic genomes reveal deep roots of modern Eurasians', *Nat Commun* **6**, 8912.

Kayser, M., Brauer, S., Cordaux, R. et al. (2006), 'Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific', *Mol. Biol. Evol.* **23**(11), 2234–2244.

Kimura, M. (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *J. Mol. Evol.* **16**(2), 111–120.

Kloss-Brandstatter, A., Pacher, D., Schonherr, S., Weissensteiner, H., Binna, R., Specht, G. & Kronenberg, F. (2011), 'HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups', *Hum. Mutat.* **32**(1), 25–32.

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P. & Paabo, S. (2010), 'The complete mitochondrial DNA genome of an unknown hominin from southern Siberia', *Nature* **464**(7290), 894–897.

Kruskal, J. (1964a), 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* pp. 1–27.

Kruskal, J. (1964b), 'Nonmetric multidimensional scaling: a numerical method', *Psychometrika* pp. 115–129.

Lachenbruch, P. & Goldstein, M. (1979), 'Discriminant analysis', *Biometrics* **35**, 69–85.

Lansing, J. S., Cox, M. P., Downey, S. S. et al. (2007), 'Coevolution of languages and genes on the island of Sumba, eastern Indonesia', *Proc. Natl. Acad. Sci. U.S.A.* **104**(41), 16022–16026.

Lazaridis, I., Patterson, N., Mittnik, A. et al. (2014), 'Ancient human genomes suggest three ancestral populations for present-day Europeans', *Nature* **513**(7518), 409–413.

Legendre, P. & Legendre, L. (1998), *Numerical Ecology*, Vol. 2nd English Edition.

Legendre, P. & Troussellier, M. (1988), 'Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation', *Limnol. Oceanogr.* **33**, 1055–1067.

Lewandowsky, M. & Winter, D. (1971), 'Distance between sets', *Nature* (234), 34–35.

Lewis, M. P., Simons, G. F. & Fennig, C. D. (2016), 'Ethnologue: Languages of the World, Nineteenth edition. Dallas, Texas: SIL International'.
URL: *http://www.ethnologue.com*

Li, H. & Durbin, R. (2010), 'Fast and accurate long-read alignment with Burrows-Wheeler transform', *Bioinformatics* **26**(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* **25**(16), 2078–2079.

Libiger, O., Nievergelt, C. M. & Schork, N. J. (2009), 'Comparison of genetic distance measures using human SNP genotype data', *Hum. Biol.* **81**(4), 389–406.

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., Schroder, R. & Stoneking, M. (2014), 'Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences', *Investig Genet* **5**, 13.

Livshits, G., Sokal, R. R. & Kobyliansky, E. (1991), 'Genetic affinities of Jewish populations', *Am. J. Hum. Genet.* **49**(1), 131–146.

Longobardi, G., Ghirotto, S., Guardiano, C., Tassi, F., Benazzo, A., Ceolin, A. & Barbujani, G. (2015), 'Across language families: Genome diversity mirrors linguistic variation within Europe', *Am. J. Phys. Anthropol.* **157**(4), 630–640.

Longobardi, G. & Guardiano, C. (2009), 'Evidence for syntax as a signal of historical relatedness', *Lingua* (119), 1679–1706.

Lugliè, C. (2009), *Il Mesolitico. In: Atti della XLIV Riunione Scientifica dell'IIPP La preistoria e la protostoria della Sardegna*, Vol. Cagliari-Barumini-Sassari.

Lugliè, C. (2014), *The Su Carroppu rock shelter within the process of Neolithization of Sardinia. In: Transitions en Méditerranée, ou comment des chasseurs devinrent agriculteurs.*

Malmstrom, H., Svensson, E. M., Gilbert, M. T., Willerslev, E., Gotherstrom, A. & Holmlund, G. (2007), 'More on contamination: the use of asymmetric molecular behavior to identify authentic ancient human DNA', *Mol. Biol. Evol.* **24**(4), 998–1004.

Manni, F., Guerard, E. & Heyer, E. (2004), 'Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm', *Hum. Biol.* **76**(2), 173–190.

Mantel, N. (1967), 'The detection of disease clustering and a generalized regression approach', *Cancer Res.* **27**(2), 209–220.

Martini, F. (1999), *Le facies clactoniane sarde nel quadro del Paleolitico inferiore peninsulare. In: Sardegna paleolitica. Studi sul più antico popolamento dell'isola*, Vol. Museo Fiorentino di Preistoria "Paolo Graziosi" (eds Martini FE).

Martini, F. & Ulzega, A. (1989-1990), 'L'insularità e i suoi effetti sul popolamento umano delle isole del Mediterraneo nel Pleistocene e nel primo Olocene', *Riv. Sci. Preist.* **42**, 271–288.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. (2010), 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Res.* **20**(9), 1297–1303.

McRae, B. H. (2006), 'Isolation by resistance', *Evolution* **60**(8), 1551–1561.

Modi, A., Tassi, F., Susca, R. R., Vai, S., Rizzi, E., Bellis, G., Luglie, C., Gonzalez Fortes, G., Lari, M., Barbujani, G., Caramelli, D. & Ghirotto, S. (2017), 'Complete mitochondrial sequences from Mesolithic Sardinia', *Sci Rep* **7**, 42869.

Mona, S., Grunz, K. E., Brauer, S., Pakendorf, B., Castri, L., Sudoyo, H., Marzuki, S., Barnes, R. H., Schmidtke, J., Stoneking, M. & Kayser, M. (2009), 'Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis', *Mol. Biol. Evol.* **26**(8), 1865–1877.

Monmonier, M. (1973), ‘Maximum-difference barriers: An alternative numerical regionalization method’, *Geogr. Anal.* (3), 245—261.

Nettle, D. & Harris, L. (2003), ‘Genetic and linguistic affinities between human populations in Eurasia and West Africa’, *Hum. Biol.* **75**(3), 331–344.

Nichols, J. (1996), ‘The comparative method as heuristic; in Durie M, Ross M (eds): The Comparative Method Reviewed: Regularity and Irregularity in Language Change ’, *New York, Oxford University Press* pp. 39–71.

Oksanen, P., Guillaume Blanchet, F., Kindt, R. et al. (2016), *vegan: Community Ecology Package.* R package version 2.3-4.
URL: *http://CRAN.R-project.org/package=vegan*

O’Rourke, D., Hayes, M. & S., C. (2000), ‘Ancient DNA studies in physical anthropology’, *Annu. Rev. Anthropol.* **29**(1), 217–242.

Paabo, S. (1985), ‘Molecular cloning of Ancient Egyptian mummy DNA’, *Nature* **314**(6012), 644–645.

Paabo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L. & Hofreiter, M. (2004), ‘Genetic analyses from ancient DNA’, *Annu. Rev. Genet.* **38**, 645–679.

Pakendorf, B. & Stoneking, M. (2005), ‘Mitochondrial DNA and human evolution’, *Annu Rev Genomics Hum Genet* **6**, 165–183.

Poloni, E. S., Semino, O., Passarino, G., Santachiara-Benerecetti, A. S., Dupanloup, I., Langaney, A. & Excoffier, L. (1997), ‘Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics’, *Am. J. Hum. Genet.* **61**(5), 1015–1035.

Posth, C., Renaud, G., Mittnik, A. et al. (2016), ‘Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe’, *Curr. Biol.* **26**(6), 827–833.

Pugliatti, M., Riise, T., Sotgiu, M. A., Satta, W. M., Sotgiu, S., Pirastru, M. I. & Rosati, G. (2006), ‘Evidence of early childhood as the susceptibility period in multiple sclerosis: space-time cluster analysis in a Sardinian population’, *Am. J. Epidemiol.* **164**(4), 326–333.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P. D., Hoal, E. G. & Behar, D. M. (2010), 'Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture', *Am. J. Hum. Genet.* **86**(4), 611–620.

Quintana-Murci, L., Krausz, C., Zerjal, T., Sayar, S. H., Hammer, M. F., Mehdi, S. Q., Ayub, Q., Qamar, R., Mohyuddin, A., Radhakrishna, U., Jobling, M. A., Tyler-Smith, C. & McElreavey, K. (2001), 'Y-chromosome lineages trace diffusion of people and languages in southwestern Asia', *Am. J. Hum. Genet.* **68**(2), 537–542.

Quintana-Murci, L., Veitia, R., Fellous, M., Semino, O. & Poloni, E. S. (2003), 'Genetic structure of Mediterranean populations revealed by Y-chromosome haplotype analysis', *Am. J. Phys. Anthropol.* **121**(2), 157–171.

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/*

Raghavan, M., Skoglund, P., Graf, K. E. et al. (2014), 'Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans', *Nature* **505**(7481), 87–91.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. (2005), 'Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa', *Proc. Natl. Acad. Sci. U.S.A.* **102**(44), 15942–15947.

Rasteiro, R. & Chikhi, L. (2013), 'Female and male perspectives on the neolithic transition in Europe: clues from ancient and modern genetic data', *PLoS ONE* **8**(4), e60944.

Reem, E., Douek, J., Paz, G., Katzir, G. & Rinkevich, B. (2016), 'Phylogenetics, biogeography and population genetics of the ascidian Botryllus schlosseri in the Mediterranean Sea and beyond', *Mol. Phylogenet. Evol.* **107**, 221–231.

Richards, M., Macaulay, V., Hickey, E. et al. (2000), 'Tracing European founder lineages in the Near Eastern mtDNA pool', *Am. J. Hum. Genet.* **67**, 1251–1276.

Ruhlen, M. (1987), 'Guide to the World's Languages.', *Edward Arnold* .

Sajantila, A., Lahermo, P., Anttinen, T. et al. (1995), 'Genes and languages in Europe: an analysis of mitochondrial lineages', *Genome Res.* **5**(1), 42–52.

Sampietro, M. L., Gilbert, M. T., Lao, O., Caramelli, D., Lari, M., Bertranpetit, J. & Lalueza-Fox, C. (2006), 'Tracking down human contamination in ancient human teeth', *Mol. Biol. Evol.* **23**(9), 1801–1807.

Sanchez-Quinto, F., Schroeder, H., Ramirez, O. et al. (2012), 'Genomic affinities of two 7,000-year-old Iberian hunter-gatherers', *Curr. Biol.* **22**(16), 1494–1499.

Sanna, E., Liguori, A., Fagioli, M. & Floris, G. (1999), 'Verso una revisione dell'inquadramento cronologico e morfometrico delle serie scheletriche paleo-protosarde. II: Craniometria, ulteriori aggiornamenti', *Arch. Antrop. Etnol.* **129**, 239–250.

Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. (1998), 'Genetic evidence for a higher female migration rate in humans', *Nat. Genet.* **20**(3), 278–280.

Shepard, R. (1962), 'The analysis of proximities: multidimensional scaling with an unknown distance function', *Psychometrika* pp. 219–246.

Sidore, C., Busonero, F., Maschio, A. et al. (2015), 'Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers', *Nat. Genet.* **47**(11), 1272–1281.

Sikora, M., Carpenter, M. L., Moreno-Estrada, A. et al. (2014), 'Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe', *PLoS Genet.* **10**(5), e1004353.

Sims-Williams, P. (1998), 'Genetics, linguistics, and prehistory: thinking big and thinking straight', *Antiquity* **72**(277), 505–527.

Smouse, P., Long, J. & Sokal, R. (1986), 'Multiple regression and correlation extensions of the Mantel test of matrix corrispondence', *Systematic Zoology* **35**, 627–632.

Sokal, R., Oden, N., Legendre, P., Fortin, M., Kim, J., Thomson, B., Vaudor, A., Harding, R. & Barbujani, G. (1990), 'Genetics and language in European populations', *American Naturalist* **135**, 157–175.

Sokal, R. R. (1988), 'Genetic, geographic, and linguistic distances in Europe', *Proc. Natl. Acad. Sci. U.S.A.* **85**(5), 1722–1726.

Sokal, R. R., Oden, N. L. & Wilson, C. (1991), 'Genetic evidence for the spread of agriculture in Europe by demic diffusion', *Nature* **351**(6322), 143–145.

Sondaar, P. et al. (1993), ‘Il popolamento della Sardegna nel tardo Pleistocene: nuova acquisizione di un resto fossile umano dalla grotta Corbeddu’, *Riv. Sci. Preist.* **45**.

Sondaar, P. et al. (1995), ‘The human colonization of Sardinia: a Late-Pleistocene human fossil from Corbeddu Cave’, *C. R. Acad. Sci. Paris* **320**((Série IIa)), 145–150.

Spoor, F. (1999), *The human fossils from Corbeddu Cave, Sardinia: a reappraisal. In: Elephants have a snorkel!*, Vol. eds Reumer JWFDVs, J. St. John.

Stefflova, K., Dulik, M. C., Pai, A. A., Walker, A. H., Zeigler-Johnson, C. M., Gueye, S. M., Schurr, T. G. & Rebbeck, T. R. (2009), ‘Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas’, *PLoS ONE* **4**(11), e7842.

Stoneking, M. (2000), ‘Hypervariable sites in the mtDNA control region are mutational hotspots’, *Am. J. Hum. Genet.* **67**(4), 1029–1032.

Tassi, F., Ghirotto, S., Mezzavilla, M., Vilaca, S. T., De Santi, L. & Barbujani, G. (2015), ‘Early modern human dispersal from Africa: genomic evidence for multiple waves of migration’, *Investig Genet* **6**, 13.

Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V. K., Bhaskar, L. V., Reddy, B. M., Reddy, A. G. & Singh, L. (2006), ‘Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA’, *BMC Genet.* **7**, 42.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R. et al. (2009), ‘The genetic structure and history of Africans and African Americans’, *Science* **324**(5930), 1035–1044.

Underhill, P. A. & 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. & Kayser, M. (2009), ‘Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation’, *Hum Mutat* **30**(9), 386–394.

Voronoi, G. (1907), ‘Nouvelles applications des paramètres continus à la théorie des formes quadratiques’, *J. reine angew. Math.* **133**, 97–178.

Weisstein, E. W. (2015), ‘”Great Circle” from MathWorld-A Wolfram Web Resource’.
URL: *http://mathworld.wolfram.com/GreatCircle.html*

Wickelmaier, F. (2005), *An introduction to MDS.*

Wilson Sayres, M. A., Lohmueller, K. E. & Nielsen, R. (2014), ‘Natural selection reduced diversity on human y chromosomes’, *PLoS Genet.* **10**(1), e1004064.

Wright, S. (1931), ‘Evolution in mendelian populations’, *Genetics* **16**, 97–159.

Zei, G., Lisa, A., Fiorani, O., Magri, C., Quintana-Murci, L., Semino, O. & Santachiara-Benerecetti, A. S. (2003), ‘From surnames to the history of Y chromosomes: the Sardinian population as a paradigm’, *Eur. J. Hum. Genet.* **11**(10), 802–807.