Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
BIOLOGIA EVOLUZIONISTICA ED AMBIENTALE

CICLO XXVII

COORDINATORE Prof. Guido Barbujani

Genome-based multidisciplinary approaches to the reconstruction of human demographic history

Settore Scientifico Disciplinare BIO/18

Dottorando
Dott. Tassi Francesca

Tutore
Prof. Barbujani Guido

Anni 2012/2014

UNIVERSITÀ DEGLI STUDI DI FERRARA

Sezioni

Dottorati di ricerca

Il tuo indirizzo e-mail

tssfnc@unife.it

Oggetto:

Dichiarazione di conformità della tesi di Dottorato

Io sottoscritto Dott. (Cognome e Nome)

Tassi Francesca

Nato a:

Ferrara

Provincia:

Ferrara

Il giorno:

01/04/1985

Avendo frequentato il Dottorato di Ricerca in:

Biologia Evoluzionistica ed Ambientale

Ciclo di Dottorato

27

Titolo della tesi:

Genome-based multidisciplinary approaches to the reconstruction of human demographic history

Titolo della tesi (traduzione):

Un approccio multidisciplinare alla ricostruzione della storia demografica umana mediante lo studio di polimorfismi genomici

Tutore: Prof. (Cognome e Nome)

Barbujani Guido

Settore Scientifico Disciplinare (S.S.D.)

BIO/18

Parole chiave della tesi (max 10):

Genetica delle popolazioni (Population Genetics), DNA antico (Ancient DNA), Metodi Bayesiani (Bayesian Methods), Uscita dell'uomo dall'Africa (OutOfAfrica), Lingue e geni (LanguagesAndGenes)

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

6 mesi

Richiesta motivata embargo

1. Tesi in corso di pubblicazione

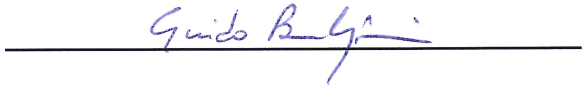
Liberatoria consultazione dati Eprints

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 16/03/2015 (data) Firma del Dottorando 

Firma del Tutore

Visto: Il Tutore Si approva Firma del Tutore 

A chi ha creduto in me

Genome-based multidisciplinary approaches to the reconstruction of human demographic history

ABSTRACT INGLESE

In my doctoral dissertation I summarize the scientific work leading to three papers in peer-reviewed journal, two submitted manuscripts. These entire studies share a common focus on human evolutionary history, but each of them address different scientific questions by means of a different combination of molecular and statistical methods.

Our cells contain a message from the past, written in their genomes; thus the study of genetic variation within and between populations can help us understand aspects of human demographic history over the past thousands of years, i.e. well beyond the time-limits of historical evidence.

Recently, extensive human genome data are becoming available, both from genome wide SNP data, and from the rapidly-increasing number of complete genome sequences, offering novel means of reconstructing human population history with a detail that was, until very few years ago, unthinkable. This abundant, and ever-growing amount of genomic data is of enormous relevant for understanding how and why human are different. Paper I (Barbujani et al., 2013) represents a review of human genetic variation and their implications for human evolutionary inference

Genetic data are indispensable to test hypothesis, generated in complementary discipline such as anthropology, linguistic and archaeology. Paper II (Tassi et al., submitted) and Paper III (Longobardi et al., submitted) provide examples of how it is possible to achieve a detailed picture of human history and evolution, taking advantage of archaeological and linguistic knowledge to interpret the genetic data.

For many years, studies of human genetic diversity have been necessarily limited to modern populations, severely limiting our ability to investigate the detail of past processes. Conversely, today, thanks to the advent of methods for reliably typing ancient DNA, it has been possible to increase our power to reconstruct historical demographic processes, and to explicitly test evolutionary hypotheses. In Paper IV (Ghirotto et al., 2013) and Paper V (Tassi et al., 2013) we analyzed ancient Etruscans sample and, within the ABC framework, we explicitly compared several models, differing for demographic and genealogical histories, to shed light on the origin and evolution of the Etruscans.

Un approccio multidisciplinare alla ricostruzione della storia demografica umana mediante lo studio di polimorfismi genomici

ABSTRACT ITALIANO

Questa tesi riassume l'attività di ricerca da me svolta durante i tre anni di dottorato che ha portato alla stesura di tre articoli pubblicati in riviste scientifiche e di due manoscritti in fase di revisione. I diversi studi sono accumunati dall'essere incentrati sullo studio della storia evolutiva umana, ma ciascuno di questi risponde a domande scientifiche diverse attraverso la combinazione di tecniche molecolari e metodologie di analisi differenti.

All'interno delle nostre cellule, racchiuso nel genoma, è contenuto un messaggio dal passato; lo studio della variabilità genetica all'interno e tra le popolazioni può così essere una valida fonte di informazione per comprendere aspetti riguardanti le ultime migliaia di anni della storia demografica umana, quindi ben oltre le testimonianze storiche.

Oggi disponiamo di una grande quantità di dati sulla variabilità genomica umana, sia grazie agli studi basati su molti marcatori a singolo nucleotide (SNP) diffusi lungo tutto il genoma, sia grazie al continuo aumento di nuove sequenze genomiche complete. Questi dati offrono nuovi mezzi per ricostruire la storia delle popolazioni umane ad un livello di accuratezza fino a poco tempo fa impensabile. In Paper I (Barbujani et al., 2013) passiamo in rassegna questi abbondanti dati genomici e analizziamo come possono essere studiati per capire come e perché gli uomini differiscono tra loro e per trarre conclusioni sulla storia evolutiva umana.

I dati genetici, inoltre, sono indispensabili per testare ipotesi proposte da discipline complementari come l'antropologia, la linguistica e l'archeologia. Paper II (Tassi et al., submitted) e Paper III (Longobardi et al., submitted) rappresentano due esempi di come sia possibile ottenere un quadro dettagliato di alcuni aspetti della storia della nostra specie e della sua evoluzione, interpretando i dati genetici alla luce delle conoscenze archeologiche e linguistiche.

Per molti anni, gli studi della variabilità genetica umana sono stati necessariamente limitati all'analisi delle popolazioni moderne, riducendo drasticamente la nostra abilità di indagare gli eventi del passato. Al contrario oggi, grazie all'avvento di nuove tecniche per ottenere in maniera affidabile il DNA da reperti antichi, è aumentato il nostro potere nel ricostruire i processi demografici del passato. In Paper IV (Ghirotto et al., 2013) e in Paper V (Tassi et al., 2013) sono stati analizzati campioni antichi di provenienza Etrusca e, grazie all'applicazione di metodi bayesiani approssimati

(ABC), sono stati confrontati in maniera esplicita diversi modelli genealogici, riuscendo a far fare luce su alcuni aspetti riguardanti l'origine e l'evoluzione del popolo Etrusco.

Table of Contents

List of Figures

Abbreviations

ABC	Approximate Bayesian computation
aDNA	ancient DNA
AMH	Anatomically Modern Human
bp	Base Pairs
DAPC	Discriminant Analysis Principal Component
GWAS	Genome-Wide Association Studies
HLA	Human Leukocyte Antigen
HVR-1	Hypervariable Region-1
IBD	identity by descent
IBS	identity by state
IE	Indo-European
ky	thousand years
LD	Linkage Disequilibrium
MD	Multiple Dispersal
mtDNA	mitochondrial DNA
my	million years
NGS	Next-Generation Sequencing
NRY	non-recombining portion of the Y chromosome
OOA	Out of Africa
PCA	Principal Component Analysis
PCM	Parametric Comparison Method
PCR	Polymerase Chain Reaction
SD	Single Dispersal
SNP	Single Nucleotide Polymorphisms
ya	years ago

Chapter 1.INTRODUCTION

Human genetic variation

The nature of the genetic data from which we can infer past processes has changed radically over the past 40 years, thanks to the development of powerful new technologies. Until a few decades ago, our knowledge about human genetic diversity was extremely limited. The first studies on human genetic variation did not directly involve DNA but rather were based on detecting and assessing variation using the so called “classical markers”. Types of classical markers range from the different variants found in the blood groups systems (starting from the ABO blood group (Landsteiner, 1900)), the different forms of particular proteins found in blood, liver and muscle such as the haemoglobins and the many different Human Leukocyte Antigen (HLA) isoforms. Through the 1970s and 80s, vast amounts of data of this kind were assembled. In 1972, Richard Lewontin analyzed allele frequencies at 15 protein loci and found that variation among major geographic regions accounts for a small percentage of total genetic variation and most of the genetic variation observed was within local populations (Lewontin, 1972). Besides, it was clear that allele frequencies for many markers were not randomly or uniformly distribute in the geographic space, rather form clines (Sokal et al., 1989a). The much higher genetic variance within, rather than between, populations, and the existence of orderly pattern of variation in space are two basic features of human diversity which have been confirmed by all following studies. In (1994) Cavalli-Sforza and coauthors attempted to synthesize data by a compendium of protein variation. This pioneering work showed that quantifying the relationship between human populations on a large scale and the synthesis with historical, archaeological and linguistic information, can provide insights into the origins and migration of history of humans. Among the others aspects, these early works already pointed out how our species is characterized by a continuous variation over the whole world with no sharp boundaries and thus, the classically defined races do not emerge from an unprejudiced biological description of human variation.

Although surveys of human polymorphism flourished in the second half of the last century, the main methodological breakthrough occurred when a host of methodological

advances enabled scientists to investigate human variation directly at the level of the DNA molecules and led to the development of "molecular markers". There are many advantages to assaying human genetic diversity through this lens. First of all, the allelic variation of the classical markers is due to amino acid level differences, therefore the genetic variation detected is limited to the DNA regions involved in transcription or translation (only the 5% of the genome(ENCODE Project Consortium, 2012). In addition, because they encode polypeptides, classical loci are likely to be under the effect of natural selection. On the other hand, DNA markers can map anywhere the genome (nuclear or mitochondrial), and because most of the genome is noncoding and thus presumably not under natural selection, hence, DNA markers occurring in these regions can be considered to be "neutral" in their effects. These kinds of markers are extremely useful for assessing the demographic history of humans, since variation in neutral region is expected to reflect mainly population level effects, such as drift, expansions, admixture and migration. The first important technical advance was the Polymerase Chain Reaction (PCR,(Mullis and Faloona, 1987)), that allow for the production of a very large quantity of a target region of a genome from even very small amounts of starting DNA. Aside from the invention of PCR itself, the other key advance in human genetic diversity studies has been the determination of the human reference genome sequence (Lander et al., 2001; Venter et al., 2001). This was made possible by the automation of Sanger sequencing (named after its inventor Frederick Sanger, and often known as chain-termination, dideoxy, or capillary sequencing (Sanger et al., 1977). The release of the reference human genome sequence provided the first foundation for studies of the genetics of the human host, but provided little insight into the extent of naturally occurring genetic variation between different individuals and populations (Kidd et al., 2010).

Later, the advent of new genomic technologies, such as DNA microarrays, has provided us with unprecedented opportunities to investigate human genetic variation at genome-wide scale. For this purpose, the International HapMap Consortium was founded in 2002 leading to a careful assessment of the common patterns of DNA sequence variation in the human genome, by characterising sequence variants, mostly single-nucleotide polymorphisms (SNP), their frequencies, and correlations between them, in DNA samples from four geographically diverse populations of Africa, Asia and Europe (The International HapMap Consortium, 2003). One of the other specific aims of the HapMap Project was to

stimulate technology to make SNP genotyping faster, more reliable, and above all cheaper, catalyzing the development of affordable SNP arrays. This technology was primarily used in the biomedical human genetics community to map disease alleles in Genome-Wide Association Studies (GWAS) (Novembre and Di Rienzo, 2009; Price et al., 2010), but then it shifted its focus. Indeed, genome-wide SNP genotyping ushered in a new phase of human genetics in which the signatures of population-genetic forces could be studied on hundreds of thousands of markers, having a big impact on our understanding of human evolution. These data have provided important insight into the finescale structure of Linkage Disequilibrium (LD; i.e., the pattern of correlation between SNPs located close together on the chromosome) in the genome (Conrad et al., 2006), the distribution and causes of recombination hotspots (Myers et al., 2005), the identity of genes that have been targeted by different forms of natural selection in the human genome (Sabeti et al., 2007; Barreiro et al., 2008), and many aspects of modern human population history, as discussed in more detail below (Novembre and Ramachandran, 2011).

The original sequencing technology was a breakthrough that helped scientists determine the human genetic code, but it would take years to sequence all of a person's DNA. However, genetics is a fast-changing field and the Sanger method, regarded as a first-generation technology, has been supplanted by a diverse set of novel technologies that have been developed more recently (beginning in 2005), collectively known as Next-Generation Sequencing (NGS). These approaches have sped up the process, taking only days to weeks to sequence a human genome, while reducing the cos. As a consequence, projects of unprecedented scales, such as the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010), are underway. The completion of 1000 Genomes Project's pilot phase has provided the location, allele frequency and local haplotype structure of ~ 15 million SNPs, 1 million indels and 20,000 structural variants, most of which being novel, creating an extensive population-scale view of human genetic variation (1000 Genomes Project Consortium et al., 2010). Since the initiation of 1000 Genomes Project the cost of sequencing an individual genome has been rapidly decreasing and will likely reach $1,000 per person within a short period of time (von Bubnoff, 2008). Many other complete genomes of individuals from different populations have been generated, leading to the discovery of a large number of previously unidentified variants, and thus suggesting that a

considerable number of human genetic variants, particularly rare variants, remain to be discovered beyond those currently known.

Thanks to this huge amount of data, we have now a very comprehensive picture of the levels and patterns of human genome diversity, from which we can draw a series of conclusions (Barbujani and Tassi, 2012).

I **Humans are genetically very close to all other ape species**. Comparing the human and chimpanzee genomes, more than 98% of the nucleotides result identical between the two species. Thirty-five million single-nucleotide changes have been identified, besides millions of chromosomal rearrangements (Chimpanzee Sequencing and Analysis Consortium, 2005). Over an estimated haploid genome length close to 3 billion nucleotides, that figure translates into a human-chimp difference equal to 1.23%. The majority of these changes, 1.06%, appear to be fixed, i.e., all members of each species have the same nucleotide. The main genetic differences between humans and other Primates do not seem to depend on point mutations, but on gain or loss of entire genes (Hahn et al., 2007), and especially on the activity of regulatory genes coordinating the expression of many other genes. These genomic regions are likely to be responsible for the key phenotypic changes in morphology, physiology, and behavioral complexity between humans and chimpanzees.

II **Humans are genetically less variable than any other ape species**. Whereas large differences are observed between pairs of orangutans, gorillas, chimpanzees and bonobos, our closest evolutionary relatives (Kaessmann et al., 2001), in humans there is polymorphism only at slightly more than 0.1% of DNA sites (Wheeler et al., 2008). Further studies will doubtless expand the list of polymorphic sites, but on average a pair of random humans is expected to share 999 out of 1000 nucleotides (Barbujani and Colonna, 2010).

III **Human populations are less genetically diverse than populations of any other ape species**. Differences among populations are often summarized by the standardized genetic variance (F_{ST}), that is, the proportion of the global genetic diversity due to

allele-frequency differences among populations (Wright, 1950). F_{ST} ranges from 0 (when allele frequencies are identical in the two populations) to 1 (when different alleles are fixed in the two populations) (for a review see (Holsinger and Weir, 2009)). Depending on the markers chosen, estimates of F_{ST} among major geographical human groups range from 0.05 to 0.13 (Lewontin, 1972; Barbujani et al., 1997). These figures mean that not only is the overall human genetic diversity the lowest in all primates but also the differences between human populations account for a smaller fraction of that diversity than in any other primate. The remaining 90% or so represents the average difference between members of the same population. Recent, extensive studies suggest that the human species' F_{ST} could even be lower (1000 Genomes Project Consortium et al., 2010), about one-third of what is observed in gorilla (F_{ST} = 0.38; (Stone et al., 2002)) and chimpanzee (F_{ST} = 0.32 (Chimpanzee Sequencing and Analysis Consortium, 2005) despite humans occupying a much broader geographic area. In short, humans show the lowest individual diversity among Primates, and are subdivided in populations more closely related than in any other Primate species. The limited degree of differentiation among human populations does not suggest a history of long-term isolation and differentiation, but rather that genome variation was mostly shaped by gene flow and admixture between populations (Hunley et al., 2009).

IV **Each human population contains a large share of the global species diversity.** One way to make sense of the above figures is to say that a random population contains on average 85% (or more) of the species' global genetic diversity. Another is to say that the expected genetic difference between unrelated individuals from distant continents exceeds by 15% (or less) the expected difference between members of the same community (1000 Genomes Project Consortium et al., 2010). A good illustration of this concept comes from the comparison of complete genomes. Among the first individuals whose genome was sequenced are James Watson and Craig Venter, two of the leading geneticists of our time, both US citizens of European ancestry. Watson's and Venter's genome sequences share more polymorphisms with a Korean subject (569,912 and 481,770 DNA sites, respectively) than with each other (461,281) (Ahn et al., 2009), so that the Korean subject is genetically intermediate

between the two persons of European ancestry (**Fig. 1.1**). This does not mean that Europeans in general are genetically closer to random Koreans than to each other, but rather that, because each population is highly variable, members of the same group, might occasionally be very different from each other, and closer to people of very distant origin. Therefore, when it comes to predicting individual DNA features, labels such as “European”, “Asian” and the like may be misleading, because they add little to the label “Human”.

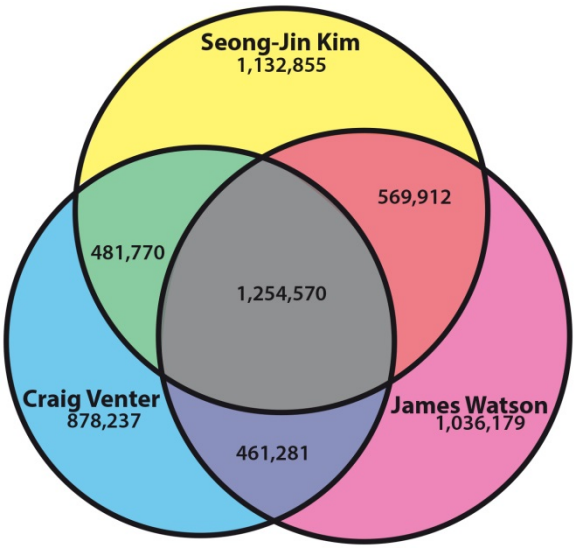


Figure 1.1 - Venn diagram of SNP alleles in Seong-Jin Kim’s, Craig Venter’s and James Watson’s genomes.
Figures within the intersections are numbers of shared alleles between individuals. Modified and redrawn from Ahn et al., 2009.

V **Africa is genetically special, and harbors the highest levels of diversity.** If we compare the main continents, we can see that African populations have the highest levels of genetic diversity at most (nearly all) loci (**Fig. 1.2**). This means that they have the largest number of unique alleles, i.e. alleles found only in one continent and not in the others (Jakobsson et al., 2008); that in many cases the alleles found out of Africa represent a subset of the African alleles; and that differences between Africans easily exceed the differences between any other pair of individuals (Schuster et al., 2010). These findings are consistent with the Recent African Origin model for the origins of modern humans (for more details see **Chapter 2**).

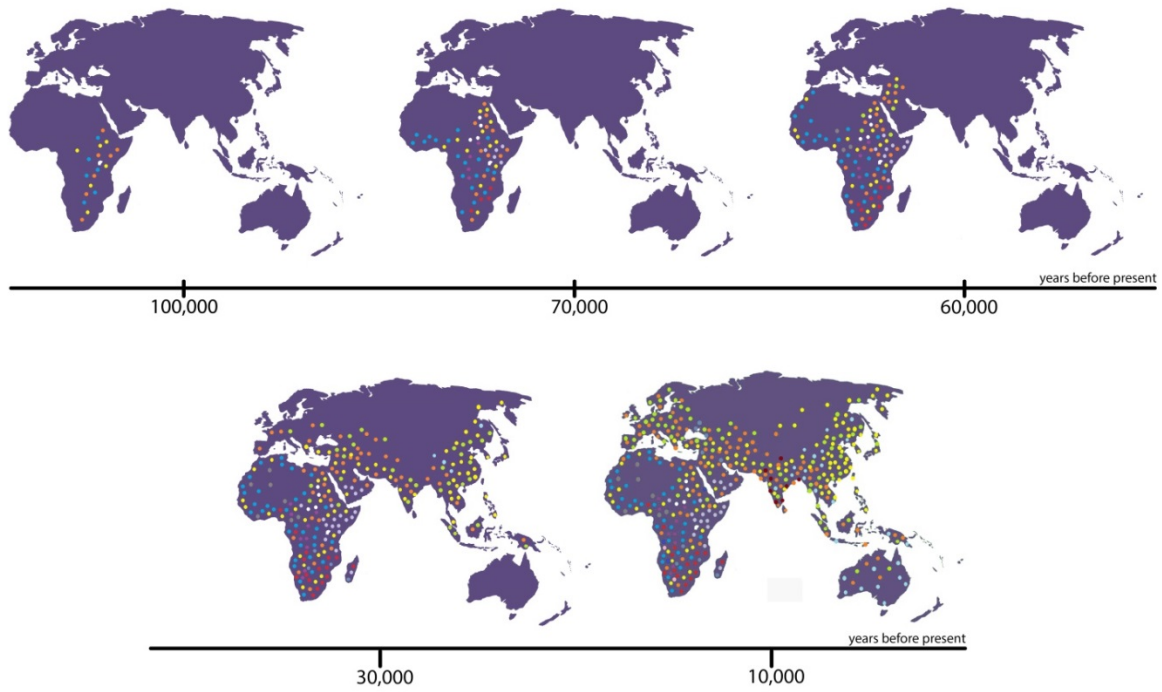


Figure 1.2 - A highly schematic view of the evolution of human biodiversity in the last 100,000 years. Dots of different colors represent different genotypes, the distribution of which roughly corresponds to archaeological evidence on human occupation of different regions. Dots of new colors appear in the maps in the course of time (e.g. red and violet in Africa at 70,000 BP, Burgundy in India at 10,000 BP), representing the effect of mutation. Because only part of the African alleles (yellow, orange and light green dots) are carried into Eurasia by dispersing Africans from 60,000 years bp, diversity in modern Eurasian populations is largely a subset of African diversity.

VI **Genetic diversity declines as a function of distance from Africa.** Several measures of genetic diversity are patterned in space, with a maximum in Africa and decreasing values, respectively, in Eurasia, the Americas, and Oceania (Prugnolle et al., 2005; Li et al., 2008). On the contrary, LD is minimal in African populations, and increases at increasing distances from there (Jakobsson et al., 2008), and the average length of haplotype blocks has a minimum in Africa around 10 kb and is close to 50 kb in Eurasia (Thomas et al., 2012). All these findings are consistent with the expected consequences of an expansion of our species outside Africa (see **Chapter 2**), and with the existence of a rather small group of founders that then rapidly populated all the world (Ramachandran et al., 2005; Liu et al., 2006). In practice, we regard these results as showing that people have lived in Africa longer than anywhere else; in this way, the African populations accumulated a higher number of mutations than any other continental group. Because only part of the African population migrated out of Africa, only part of Africa's genetic variation moved with them; and because the other continents were peopled at a relatively recent time, only few mutations are geographically restricted to these continents.

VII **There is no genetic support to the traditional idea that the human species is composed of biologically distinguishable races.** In modern biology, a race is defined as a cluster of individuals who occupy a given territory, are genetically homogeneous, and differ from other clusters of individuals. The existence of such clusters has been traditionally assumed by classical anthropologists up to the twentieth century, and many catalogs of human races were proposed, starting in the 18th century with Linnaeus. However, for these catalogs to be of any use, they must be consistent with each other, whereas in fact they are not. On the contrary, different authors' catalogs contained anything between 2 and 200 entries (Madrigal and Kelly, 2007), an incongruence that Charles Darwin had noticed, concluding that human races graduate into each other, and it is hardly possible to discover distinctive characters between them. Recent genetic studies have shown why. More than 80% of human alleles are cosmopolitan, i.e. present at different frequencies in all continents (Jakobsson et al., 2008); there are no sharp genetic discontinuities between populations or continents, and populations differ mostly for the different proportion,

in each of them, of the same alleles. In addition, the different genetic polymorphisms are differently distributed in space and not correlated over the planet, and so we can cluster people based on any set of alleles, but there is no guarantee that the same clustering will be observed when considering other alleles in the same individuals (Hammer et al., 2004; Bowden et al., 2006; Cox, 2007).

These data and their implications, are reviewed in detail **PAPER I** (Barbujani et al., 2013).

Ancient DNA

For long, past demographic changes could only be roughly inferred from patterns of current genetic diversity. Such inferential process depends heavily on assumptions on factors such as demographic growth and migration rates, for which basically no empirical information is available. In addition, the levels and the patterns of genetic variation that we observed today are strongly influenced by the particular evolutionary history of the individual sampled and it is often difficult to distinguish between competing hypotheses. Besides, people who currently live in a given territory might not represent the people who inhabited the same territory in the past and dating their presence from the coalescence of their genetic profiles poses further problems (Barbujani et al., 1998). However, with recent advances in molecular sequencing technologies and sequence data analysis, we now have an unprecedented ability to recover genetic information from archaeological and paleontological remains, which allows us to go back in time and to address directly questions about human evolution.

It is clear that, there are many practical difficulties with the analysis of ancient DNA (aDNA) in general, and of human samples in particular, caused by the nature of the studied biological material. In the cells of a living person, DNA is continually being monitored and repaired. After death, the systems that accomplish this function stop working, causing cellular degradation by endogenous nucleases and proteases with associated infiltrations of exogenous bacteria, fungi, or other organisms that further digest and non-specifically fragment the DNA (Paabo et al., 2004). The DNA survival in an ancient sample is influenced by the conditions under which it has existed since it was deposited: temperature, pH, humidity, and salt concentration affect the rates of the modifications that DNA undergoes post mortem (Smith et al., 2003). Cold, dry environments discourage the growth of microorganisms and minimize chemical damage. Remains that are quickly buried and, ideally, frozen tend to be best preserved (Hofreiter et al., 2001). The result is that the low quantities of DNA recovered from bones or other tissues of long-dead samples is severely damaged by cleavage of the sugar-phosphate backbone, resulting in short DNA fragments (usually below 70 bp (Green et al., 2008)); loss of bases; chemical modification of bases (particularly deamination that produce incorrect sequence reads, such as C to T and G to A

transitions (Fulton, 2012); and inter- or intramolecular cross-linking of sugar-phosphate backbones (Hebsgaard et al., 2005).

Aside from molecular damage to aDNA, exogenous DNA contamination of sample may and does also occur. The extracted ancient DNA is always a mixture of organismal and environmental DNA, including DNA from bacteria, fungi, and other organisms that colonize the sample during burial. Separating endogenous and contaminating DNA from microorganisms is not extremely complicated: however, the most serious problem for aDNA researchers working on humans or their close evolutionary relatives is modern human DNA contamination. PCR amplification has made it possible to analyze genetic information from such material, but amplification of the degraded and modified DNA is not very efficient and sporadically contaminating intact modern DNA molecules can be preferentially amplified. Indeed, this contamination caused erroneous results and has led to extravagant reports, including claims of DNA sequences surviving for millions of years in plants (Soltis et al., 1992) and dinosaur bones (Woodward et al., 1994). It is now believed that most or all of these results were artefacts of modern DNA contamination (with bacterial, fungal, or human DNA), and that physicochemical processes set a probable upper limit of 100 thousand years (ky) to one million years (my) on the survival of DNA (Hebsgaard et al., 2005).To deal with this issue, researchers have agreed on a series of guidelines to ensure the quality of aDNA data and the reliability of consequent conclusions that are often recapitulated as “The nine gold criteria” by Cooper and Poinar (2000). These included replicability (if an aDNA sequence is genuine, it should be possible to reproduce it) and reliability (replicates of the same target sequence should be identical).

The field of aDNA studies began thirty years ago (ya) with the extraction and sequencing of DNA material from the quagga, a South African equid (*Equus quagga quagga*) that went extinct in the 19th century (Higuchi et al., 1984) and from an Egyptian mummy (Paabo, 1985). These studies used bacterial cloning to amplify small sequences retrieved from skins of animal and human mummies, and revealed the genetic material surviving in ancient specimens was often principally microbial or fungal in origin, and that endogenous DNA was generally limited to very low concentrations of short and damaged. A few years later, with the development of PCR (Mullis and Faloona, 1987), it became possible to

routinely amplify and study surviving ancient DNA molecules even if only in a single copy, resulting in a rapid increase and diversification of ancient DNA research.

Until recently, most of aDNA studies have been restricted to short fragments, mainly from the hypervariable region-1 (HVR-1) of the mitochondrial DNA (mtDNA). This is because, first of all, mtDNA is present in several hundreds copies per cell, in contrast to the single-copy nuclear genome. Thus, integer sequences of mtDNA are more likely to be present in any single extract, and can be easily amplified, than nuclear sequences. Second, the generally higher mutation rate of vertebrate mtDNA ensures that more haplotype diversity will be seen in mtDNA than in comparable amounts of nuclear DNA. Third, because there is no recombination in mtDNA, the mutations are clonally transmitted across generations and gene genealogies of mtDNA haplotypes are readily inferred using standard phylogenetic methods. Thus, mtDNA has been successfully used to investigate the demographic history of human populations (Endicott et al., 2003; Vernesi et al., 2004; Ghirotto et al., 2010; Vai et al., 2015). In this context, I have analyzed datasets of modern and ancient genetic variation in order to understand the origins and evolution of the Etruscan population. The findings of this research are reported in this thesis (see **Chapter 4**, **PAPER IV** and **PAPER V**).

In the last few years, with the advent of new sequencing technologies, NGS (Bentley et al., 2008), the field of ancient DNA is experiencing a new era wherein what was once impossible has become possible, moving to the analysis of genome sequences, sometimes complete ones, of extinct species and population. One of the major advances introduced by high-throughput sequencing technology is the ability to sequence millions of DNA molecules in parallel, thereby increasing the amount of sequence data generated and reducing the cost of sequencing. Most importantly, NGS does not rely on targeted PCR amplification of the aDNA molecules using primers. Therefore, this technology is able to obtain useful sequence information from shorter DNA fragments (Green et al., 2010) and thus, because the number of endogenous DNA increases exponentially with decreasing fragment lengths, it permits the access to a much larger fraction of endogenous aDNA. In addition, contaminating modern DNA tends to be longer, and consequently the ratio of endogenous to contaminating DNA shifts in favour of the former when using NGS compared to PCR (Kirsanow and Burger, 2012). Another key advantage of NGS is that it allows the use of degradation patterns to

discriminate between modern DNA contaminations and ancient degraded DNA (Briggs et al., 2009).

The first paleogenomic studies using NGS produced ~13Mb of nuclear DNA from a 28,000 year old mammoth fossil (Poinar et al., 2006). After this milestone publication, many other sequencing projects of ancient DNA have been carried out based on high-throughput NGS and new perspectives to study evolution have opened up. As far as human evolution is concerned, in May 2006, the first nuclear DNA sequences from a Neandertal (*Homo Neandertalensis*) were reported, as part of the Neandertal Genome project that had started about two years earlier (Green et al., 2006). Within this project, later, a 1.3-fold coverage Neandertal genome was produced from bones from Vindija Cave in Croatia that contained only 1 to 5% endogenous DNA (Green et al., 2010). This was quickly followed by a 1.9-fold coverage genome from a morphologically uncharacterized hominin fossil from Denisova cave (Reich et al., 2010), by the genome of a 4,500 year old paleo-Eskimo at 20-fold coverage (Rasmussen et al., 2010), and an 11-fold coverage genome from an Australian aborigine (Rasmussen et al., 2011). The Denisova genome was later improved to 30-fold coverage (Meyer et al., 2012) thanks to very high (~80%) endogenous DNA content and a new, more efficient method to prepare sequencing libraries (Gansauge and Meyer, 2013). Recently, a ~50-fold coverage Neandertal paleogenome was recovered from another extremely well preserved bone with a high (~75%) endogenous content, also from a cave in the Altai Mountains of Siberia (Prufer et al., 2014). Analysis of these hominin paleogenomes revealed patterns of resemblance with modern populations suggesting potential episodes of admixture between lineages during recent evolutionary history. Neandertal DNA shares more genetic variants with present-day humans from Eurasia and Melanesia than from sub-Saharan Africa, potentially meaning that on average 2.5 % of the genome of people outside Africa derive from Neandertal ancestors. Instead, there is no evidence excess of allele sharing between Denisova and modern Europeans or East Asians, but the Denisova nuclear sequence shares around 5 to 7% of the polymorphism with modern Melanesian population, although they are far removed from the Denisova site. Although these levels of genomic similarity are doubtless there, their interpretation is not obvious, and admixture is not the only possible explanation of the data. For example, an ancient population structure in African population ancestral to humans and other hominins has been proposed as an

alternative explanation (Eriksson and Manica, 2012). The history of humans is more complex than previously supposed and many aspects have to be resolved yet (see **Chapter 2**).

More recently, the first nuclear sequences from an early modern human were determined by the capture of the chromosome 21 from an modern human male of a ~40,000 year old from Tianyuan cave near Beijing (Fu et al., 2013a). In 2014 the genome of three early modern human individuals were published: the genome of a 13,000 years old Pleistocene individual from North America (Anzick-1) (Rasmussen et al., 2014), a 24,000 year old individual from the Lake Baikal region (MA1) (Raghavan et al., 2014), and a 45,000 year old individual from Ust'-Ishim near Omsk (Ust'Ishim1) (Fu et al., 2014). The last one represents the oldest full genome of a modern human published to date. This data revealed that Ust'-Ishim individual would represent an early modern human radiation into Europe and Central Asia and that the early stages of Eurasian lineage were already complex. Anzick-1's and MA1's genomes provided indeed detailed insights into early human colonisation of the Americas, showing evidence that contemporary Native Americans and western Eurasians share ancestry through gene flow from a Siberia upper Palaeolithic population into First Americans.

Methodological strategies for maximizing the retrieval, enrichment, and sequencing of short DNA fragments are dramatically improving the quality of ancient DNA studies in the area of human evolution. Most critically, the time-depth to which ancient DNA strategies are capable of reaching has significant applications to the study of the hominin lineage. No longer is it impossible to obtain authentic DNA sequences from 100,000 year old specimens, but recently the mtDNA genome of a 400,000 year old hominin from the Sima de los Huesos in Spain has been sequenced (Meyer et al., 2014), demonstrating the possibilities of exploring DNA survival in hominin species that have yet to be sequenced, such as *Homo erectus* and *Homo heidelbergensis*. Besides the experimental challenges, the aDNA research need to address computational and analytical challenges; once useable samples are obtained and sequenced, the dataset must be processed. Then, the research field requires bioinformatics expertise, data-processing power, and data-storage solutions necessary to handle the millions or even billions of sequences that are generated. But in particular, aDNA

data require tailored bioinformatics tools for handling the short and degraded fragments (Kircher, 2012) and for evaluating the evidence of contamination (Skoglund et al., 2014).

Despite its intrinsic limitations and the necessary caution, the study of aDNA represent a fundamental tool to reveal patterns of genetic variation in past populations, to detect evidence of natural selection, or to infer past demographic events, such as migration, range expansion, and changes in population size. We expect that the coming decade will bring even more important discoveries, including a better understanding of cultural and behaviour aspects as past diet, burial practices, and also about the evolution of pathogenicity.

Evolutionary forces

The evolutionary dynamics of natural populations (be they human or not) are governed by a well-known set of evolutionary forces, causing departure from equilibrium. For long, it has only been possible to make educate guesses on the factors leading to the observed levels and patterns of within- and between-population diversity. Modern, computer-intensive methods are now permitting a much more detailed analysis of these factors and making it possible to quantitatively compare models differing for the relative weight given to mutation, selection, drift and gene flow.

Mutation is, along with recombination, the main sole source of variation in the genome generating random changes in the DNA sequence. The results of this process is a heterogeneous category of changes in DNA that come about through myriad pathways and ultimately induce changes ranging from single base pair alterations to small insertions and deletions to large-scale structural rearrangements or even the addition or deletion of whole chromosomes. It provides the raw material on which evolution can act by means of selection or other forces. Although mutations are vital to evolution, mutation rate are low (around 0.2 mutational events per million year per nucleotide for the human mitochondrial DNA (Henn et al., 2009) and around 0.001 mutational events per million year per nucleotide for a human noncoding region of autosomal DNA (Fagundes et al., 2007) and not lead, by themselves, to major changes in allele frequencies.

The second key force shaping patterns of human genetic variation is **genetic drift** (Wright, 1931), that is the stochastic process resulting from the random sampling of gametes at reproduction, and determining random variation in allele frequencies over time. Genetic drift may cause allelic variants to disappear completely or to be fixed (reaching frequency of 1), in both cases, reduces genetic variation within populations. On the other hand, because genetic drift is a random event occurring independently in different populations, the pattern of genetic drift will tend to be different on average in different population, and hence variation between populations will tend to increase under genetic drift. The magnitude of these effects depends on the size of the breeding population: the larger the population size, the smaller the change occurring from one generation to the next. Two main demographic

processes associated to genetic drift have non-negligible consequences on the genetic diversity of populations that experience them, namely bottlenecks and founder effects. The former refers to the temporary shrinking of a single, previously larger, population, and the latter to the process of range expansion and colonization of new territories, often accompanied by the sampling of a subset of the genetic diversity present within the source population, and both resulting in a loss of genetic diversity. Both those processes played an important role in human history, and their effect is still detectable in the genetic diversity pattern of modern humans.

Opposite to those of genetic drift are the effects of **migration**, that is the movement of individuals from an occupied area to another one. Migrants from other populations enter and contribute to the gene pool, changing the allele frequencies in the population, as well as introducing new genetic variation. This results in decreased the genetic differentiation between connected populations and in increased variation within a population.

All the above processes are expected to affect equally all loci in the genome (Cavalli-Sforza, 1966; Sokal et al., 1989b).The fourth force which contributes to the distribution of human genomic variation is **natural selection**, acts specifically upon single genes. During the process of evolution, some individuals with a certain trait or phenotype may tend to be more successful to reproduce than others in a certain environment. In other words, some individuals fit the environment better and have a major ability of transmit his or her genotype to the next generation. The individual's expected reproductive success is measured by her/his fitness (ω) and the relative fitness of a genotype is obtained from a comparison of this genotype with all other genotypes competing for the same resources. Natural selection can act in a population only if mutation has generated heritable polymorphisms among individuals. Selection therefore works to increase the frequency of variants that increase the fitness of an individual in its environment (*positive selection*) and to decrease the frequency of deleterious allele (*purifying selection*). In some cases, natural selection acts to maintain the polymorphism, preserving two or more alleles at a locus in a population, and tends to favour intermediate-frequency alleles (*balancing selection*). In human populations, it appears that most genetic variation is neutral and selection is a weaker force than genetic drift in shaping global pattern of genomic variation (Balaresque et

al., 2007) , but opinions differ in this area (for reviews, see (Scheinfeldt and Tishkoff, 2013) (Jeong and Di Rienzo, 2014)).

As previously mentioned, demographic processes, such as changes in population size or migration, are expected to affect the entire genome in the same way, whereas natural selection affects specific functionally important sites in the genome. However, similar patterns of genetic variation can be produced both by events in demographic history or by specific selection regimes (for example a rapid expansion in population size or positive selection can produce a similar excess of low-frequency variants: (Harpending, 1994; Braverman et al., 1995). One way to disentangle the confounding effect of population history from the effect of selection is a comparison of the pattern of variation at a candidate locus with the genome-wide pattern estimated from a set of neutral markers that have been typed in the same individual or population (Bamshad et al., 2002).

The evolution of genetic diversity under these forces is expected to behave in certain ways, defining assumptions that are used to build a set of predictions. Theoretical population genetics, using these predictions, develops mathematical models and compares genetic patterns observed in actual population with expected pattern, to elucidate how allele frequencies change in time and space.

Chapter 2.TRACING MODERN HUMAN ORIGINS

As late as 3 million ya, it is believed that all ancestors of living humans were found in the African continent. Starting ca 2.3 million ya, hominin migrations out of Africa resulted in the appearance of several population lineages in all major continents except the New World and Antarctica, succeeding in adapting to a vast diversity of environments from the frigid Siberian tundra to the lush rainforests of Southeast Asia. Over the course of 2 my, they evolved into biologically diverse groups, from the 1 meter tall Homo Floresiensis (Brown et al., 2004) to the remarkably robust Neandertals (Hublin, 2009), and the group known only from DNA information that is designated as Denisovans (Krause et al., 2010). However, there is today scientific consensus that most of these ancient human populations did not give rise to the human populations living outside of Africa today. Instead, at least 90% of the ancestry of all modern humans today can be traced to Anatomically Modern Human populations (AMH) living in Africa about 100 kya (Meyer et al., 2012).

The appearance of anatomically modern humans in Africa

Between 300 kya and 150 kya, the first evidence of what is referred to as anatomical modernity began to appear in Africa (Stringer, 2002; Tattersall, 2009). These skeletal remains are found both in eastern and southern Africa, and their absence from other parts of Africa does not necessarily mean they were not elsewhere; it could be attributed to less intense excavations or poor preservation conditions. Thus, the exact geographical point of origin of these anatomical features in Africa is not known, but they markedly predate any such evidence from outside Africa. The earliest AMHs fossils were those found in the Klasies River Mouth Caves in South Africa which date to ~130 kya and those from the Levant, at Qafzeh and Skhul which date to ~130-90 kya (Stringer, 2002). However, recent fossil evidence from Ethiopia indicates the presence of early AMHS there between ~195-154 kya (Clark et al., 2003). After the initial appearance of AMHS in the Levant, the fossil evidence suggests that they do not reappear in that region or in Europe until ~60-40 kya (Trinkaus, 2005; Mellars, 2006b). The earliest evidence of AMHS presence outside of Africa were surprisingly found in Lake Mungo in Australia and are about 45,000 years old, thus thousands of years older than fossils attributed to modern humans found in Europe and Asia (Bowler et al., 2003).

However, recent archaeological evidence from pre-Toba and post Toba (74-77 kya) artefacts from the Indian subcontinent show closer affinities to African Middle Stone Age traditions (such as Howieson‟s Poort), and may indicate modern humans might have reached the Indian sub-continent by 70 kya (Petraglia et al., 2007).

Models of modern human origins

One of the most heavily debated topics in paleoanthropology was for long the population history behind the appearance of anatomical modernity in Africa and Eurasia (Stringer, 2002). Many models have been proposed attempt to explain how AMHs became distributed throughout the globe within the last 100 kya, and how all AMHs are related to the other hominins species.

The extreme alternative scenarios are sometimes referred to as the **Out of Africa** model (OAA) (or, more precisely stated the Recent African Origin model), and the model of **Multiregional evolution** (**Fig. 2.1**). The **OOA** model posits that present-day human populations across the world trace their ancestry to Africa within the past ~200 kya, and thus that the populations with archaic morphology (such as *Homo erectus* and Neandertals) would have been replaced by these newcomers without contributing significantly to their ancestry (Stringer and Andrews, 1988; Stringer, 2002). By contrast, the **Multiregional** model proposes that anatomical modernity have emerged gradually and simultaneously from archaic forms in different continents, with natural selection acting to raise the frequency of traits associated with anatomical modernity. Although differences between geographic regions evolved over time, all human forms documented in the fossil record and modern humans would represent a single species because the archaic human groups of Africa, Asia and Europe were not reproductively isolated, but connected by gene flow occurred in the past ~ 1-2 my (Wolpoff et al., 2000).

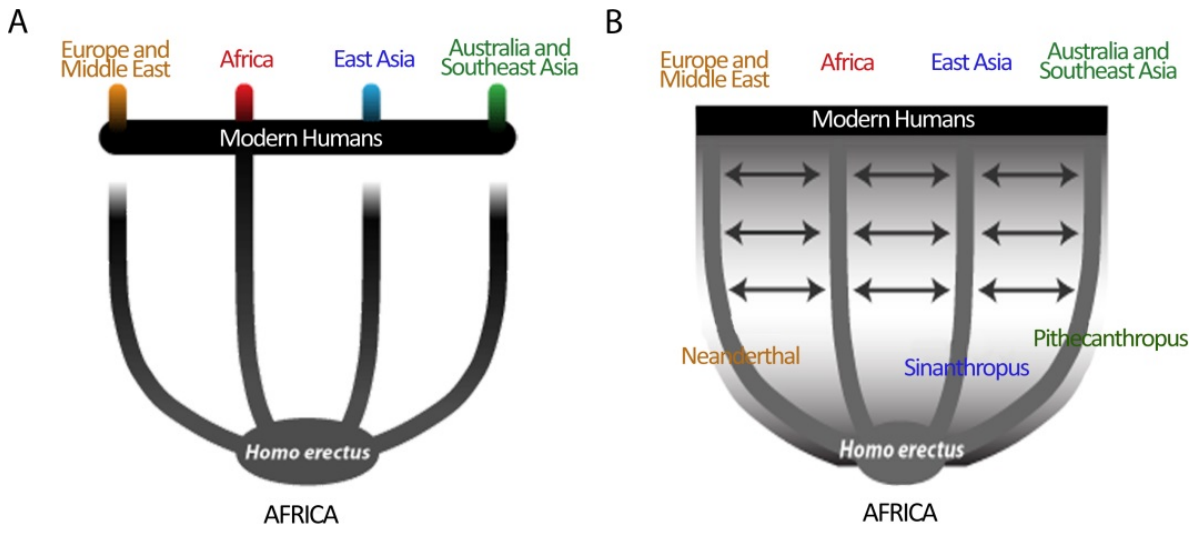


Figure 2.1 - Out of Africa (A) and Multiregional (B) model of human evolution.

For decades, the debate on the relative merits of these models, and of the many variants proposed, revolved around the interpretation of anatomical features. Now their predictions in terms of genetic variation can be tested against DNA data. At the end of the 1980s, the first available genetic evidence taken to support the **OAA** hypothesis came from the sequencing of mitochondrial DNA variation from worldwide populations (Cann et al., 1987). When a genealogical tree was reconstructed the sequences appeared to come in two clusters, one of them only including Africans, and the other containing people of different ancestries, both African and non-African. This pattern was later confirmed, and indicates that, at the mitochondrial level, non-African variation is but a subset of the variation found in Africa (Cann et al., 1987; Ingman et al., 2000). By taking mutation rate into account in the gene tree analysis, the date when the most recent common ancestor lived can be estimated, and for mtDNA analyses this date falls around 200 kya (Penny et al., 1995). These findings have often been interpreted as supporting the **OAA**, which predicts a common African ancestor at about the same time. Meanwhile, results from the non-recombining portion of the Y chromosome (NRY) were also consistent with an African origin of AMHs (Thomson et al., 2000; Underhill et al., 2000). In addition, ancient mtDNA was isolated and sequenced from a range of geographically disparate Neanderthal fossils (Krings et al., 1997; Krings et al., 2000; Ovchinnikov et al., 2000; Schmitz et al., 2002; Serre et al., 2004; Caramelli et al., 2006)

and in a paper to which I contributed was shown to be genealogically distinct from known extant mtDNA sequences (Ghirotto et al., 2011b).

Further studies of worldwide modern human variation using autosomal markers (mainly SNPs and STRs) have shown that the extant genetic pattern is remarkably consistent with a continuous decrease of genetic diversity with geographic distance from Africa (Rosenberg et al., 2002; Ramachandran et al., 2005; Li et al., 2008). These studies have shown three trends in summary statistics as a function of increasing geographic distance from Africa: a decrease in heterozygosity (Li et al., 2008), an increase in linkage disequilibrium or LD (Jakobsson et al., 2008), and a decrease in the slope of the ancestral allele frequency spectrum (indicating that derived alleles tend to be more frequent in populations at a greater distance away from Africa (Li et al., 2008). This pattern can be explained by positing a serial founder effect where populations expanding out of Africa into the rest of the world experienced the cumulative effect of genetic drift (DeGiorgio et al., 2009). On these grounds the **OOA** model has been widely adopted by the human population genetics community. However, this model was disputed by some archaeologists for whom there is evidence of the appearance of similar traits (i.e. flatness of the frontal bone and the constriction of the skull behind the orbital area) within the same geographic region over time. This evidence could be better explained by genetic continuity (according to the **Multiregional** model), rather than according to the complete replacement for which the traits would be eliminated and then, would have to appear independently (Wolpoff, 1989). On the other hand, morphological studies by Lahr and Foley (1994) found that the majority of traits analysed did not really show a specific regional continuity, and suggested that to account for them the **Multiregional** model is unnecessary.

More complex scenarios

Although paleontological and genetic data strongly suggest that Africa is the most likely geographical origin for a modern human dispersal, there is still disagreement on the extent of population replacement taking place as AMH expanded over the planet, ultimately occupying all suitable territories. In particular, open questions concern the possibility of admixture with pre-existing human forms, on the details of the dispersal process and on the exact nature of the migration events.

In the recent Neandertal genome survey (Green et al., 2010), the authors found that Neandertals are slightly but consistently closer to present-day non-Africans than to present-day Africans. Although alternative scenarios could not be ruled out on the basis of the available genomic evidence, this asymmetry was been interpreted, as evidence for hybridization between Neandertals and anatomically modern humans during the latter's exit from Africa (Green et al., 2010). Given that there is no difference between Europeans and Asians/Melanesians in their similarity to Neandertals (in fact, Asians seem, once again, slightly but consistently closer to Neandertals than Europeans, despite the latter's much longer proximity to Neandertals), it has been argued that such hybridization would have had to happen at the very beginning of the **OOA** expansion, in the Levant, before the split between Europeans and Asians/Melanesians (Green et al., 2010). The high-quality Neandertal genome recently characterized refined this estimate to ~2%. Because, as already mentioned, Neandertals appear to have contributed more DNA to modern East Asians than to modern Europeans (Meyer et al., 2012; Wall et al., 2013), simple population models where Neandertals and AMHs admixed just once when they cohabited in the Levant before the latter colonized Asia, Oceania and Europe, should probably be dismissed and more complex models preferred, where additional gene flow from Neandertals into East Asians took place after they diverged from Europeans (Vernot and Akey, 2014). Contrary to Neandertals, the genome from a newly discovered hominin from the Denisova caves in Siberia has no evidence of admixture with most present-day Eurasian populations, with the exception of unusual polymorphisms shared with Australians and Melanesians (Reich et al., 2010; Reich et al., 2011; Prufer et al., 2014), again suggesting hybridization. However, our lack of knowledge of both the geographic range of Denisovans, and of their exact taxonomic affinity to modern humans, makes it difficult to identify the exact scenario. To explain these spatially heterogeneous patterns of similarity between any ancient hominin and modern human populations, recent admixture is not the only hypothesis that has been put forward. The persistence of population substructure in early hominin populations in Africa, which has been inferred from the human paleontological record (Gunz et al., 2009; Harvati et al., 2011) and is concordant with climate fluctuations in the continent (Scholz et al., 2007; Blome et al., 2012), could produce the same pattern (Durand et al., 2011). This alternative model of population history posit that there were two or more subpopulations of hominins in Africa

with limited gene flow, with Neandertals the and ancestors of present-day non-Africans dispersing, at different times, from the same population background. Consequently, non-Africans would be slightly more genetically similar to Neanderthals than would Africans because of their more recent common ancestry. If this were the case, incomplete lineage sorting and not introgression could explain some genetic similarities between modern non-African humans and Neanderthals (Eriksson and Manica, 2012). These findings do not imply that dispersing modern people from Africa did not interbreed with other hominin populations but suggest that, at present, other hypotheses also seem to be compatible with the biological evidence.

Given the overwhelming genetic evidence for a recent origin of modern humans in Africa, an unresolved, and very relevant, question is whether there has been a single exit from Africa or more. Indeed the tempo and mode of dispersal in Eurasia and Oceania is however still controversial with different models competing. This debate has no minor implications for the issue of the possible hybridisation with Neandertals. The **single dispersal** model (**SD**) supports a unique dispersion into Eurasia about 50 kya followed (The HUGO Pan-Asian SNP Consortium, 2009) by a series of founder events and separate migrations into Asia (55-40 kya) and Europe (40-25 kya) (Liu and Zhao, 2006; Fu et al., 2013b). According to this model, a further expansion from Asia into Australia would have given rise to the ancestors of Aboriginal Australians, as early as 50-40 kya (O'Connell and Allen, 2004). In contrast to the single dispersal model, the **multiple dispersal** hypothesis (**MD**) assumes separate successive migrations from Africa to the rest of the world, an early "southern" route through the Arab peninsula and the Indian subcontinent towards Australia and Melanesia ~100-60 kya and a later "northern" route through North Africa/Middle East towards Eurasia ~70-40 kya (Cavalli-Sforza et al., 1994). According to this hypothesis, Asian populations descended from this early expansion were then largely replaced by subsequent dispersal from Africa, or underwent extensive admixture (Karafet et al., 2001; Martinez-Cruz et al., 2011), except perhaps for certain *relic* populations such as Andamanese Islanders, Malaysian and Philippine 'Negrito' groups, or aboriginal Australians and New Guineans (Mellars, 2006a) (**Fig. 2.2**).

The Southern-route hypothesis was initially proposed to account for observations of temporal and spatial patterns of cranial diversity (Lahr and Foley, 1994), and of shared phenotype (short stature, dark skin color, and tufted hair) between populations of Africa and isolated, indigenous populations of Southeast Asia (Lahr, 1996). Besides, this scenario is strengthened by mtDNA research which suggests a rapid eastward migration along the northern rim of the Indian Ocean (Macaulay et al., 2005; Thangaraj et al., 2005). The hypothesis of an early southern route also receives strong support from a different analysis of genome-wide SNPs data (Wollstein et al., 2010) which used an Approximate Bayesian Computation (ABC) framework to test various models of population history and estimate associated demographic parameters, as well as from a recent study that used both genome-wide SNP and craniometrics data to analyse the spatiotemporal predictions of various models (Reyes-Centeno et al., 2014).

If this really happened, in their dispersal from Africa, the ancestors of current Papuans would have missed by thousands of miles the nearest documented Neandertal population (**Fig. 2.2**), and so the similarity between Neandertals and Papuans (who have the same level of apparent Neandertal admixture as all other Eurasians tested so far (Green et al., 2010)) would call for a different explanation.

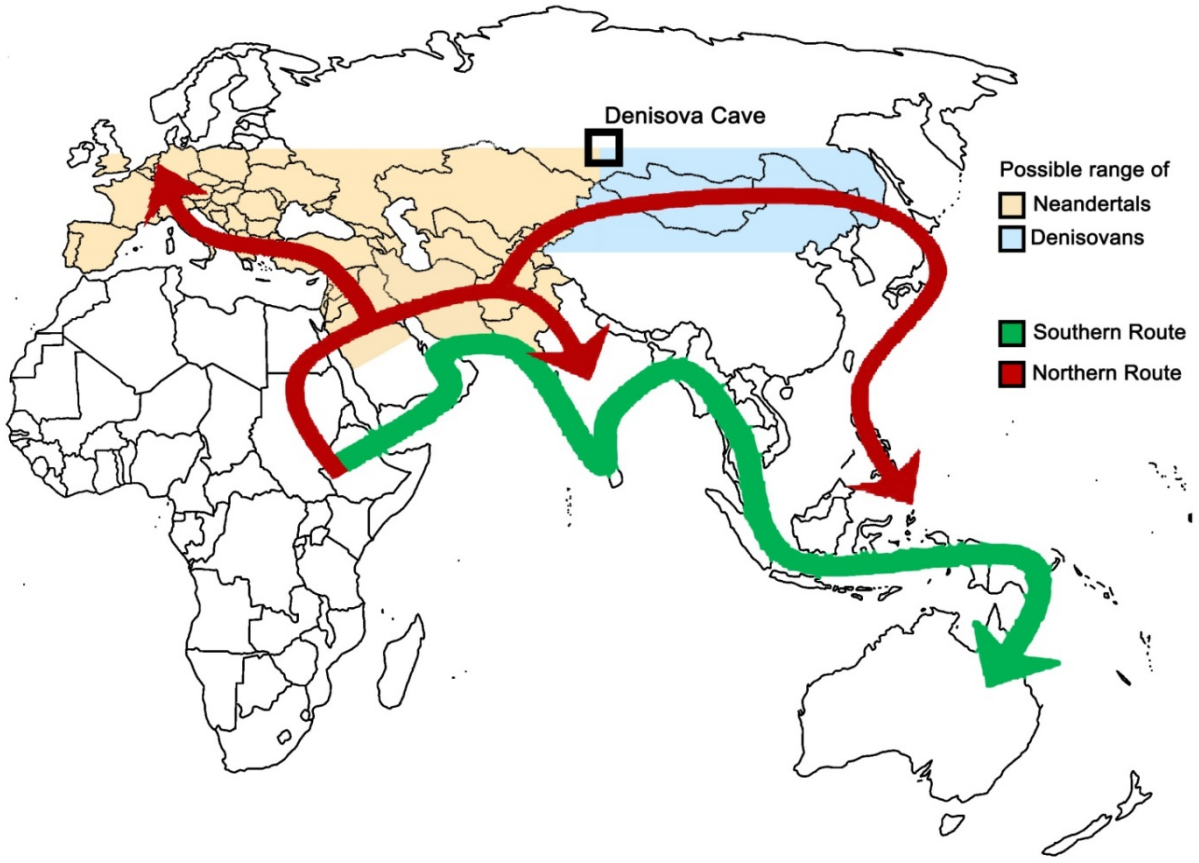


Figure 2.2 – Multiple dispersal model. According to multiple dispersal model, green arrows indicate initial colonization events along the Southern route after the origin of anatomically modern humans (AMHs) in Africa. The red arrows show the more recent expansion into Eurasian along the Northern route. The light yellow and light blue shadows, represent possible range of Neandertals and Denisovans, respectively.

Early modern human dispersal from Africa: Genomic evidence for multiple waves of migration

To obtain insight into the historical and geographical context of the emergence and dispersal of our species out of Africa, we combined different approaches, including (*i*) the analysis of population structure, (*ii*) a LD-based approach to trace the population's changes in the effective population size (N_e) through time and to estimate the respective divergence time (*T*) from Africa, and (*iii*) a comparison of genetic distances between populations with the expectations of different dispersal models.

We analysed genome-wide SNPs in 1,130 individuals belonging to 71 worldwide populations selected from several suitable public different dataset (Lopez Herraez et al., 2009; Reich et al., 2009; Xing et al., 2009; Xing et al., 2010; Reich et al., 2011; Pugach et al., 2013). We devised a careful strategy to combine the seven datasets genotyped with different platforms according to different protocols developing a pipeline built on Perl, and, after cleaning and integration, we obtained a database of more than 96,000 autosomal SNPs. We identified cryptic relatedness amongst samples computing identity-by-descent (IBD) statistic for all pairs of individuals, as unmodeled excess of genetic sharing would violate the sample independence assumptions of downstream analyses. We did not apply this screening procedure for the South-East Asia and Oceania samples, since they come from populations with extremely low effective sizes, where a certain degree of random inbreeding is inevitable (Relethford, 1985). In addition, individuals identified as population outliers were removed, evaluating their dissimilarity in terms of identity by state (IBS). The final dataset contained 1,130 individuals, and we grouped them into 24 ethnolinguistically and geographically related meta-populations.

Preliminary analyses (Principal Component Analysis (PCA), *ADMIXTURE* (Alexander et al., 2009), Discriminant Analysis Principal Component (DAPC) (Jombart et al., 2010)) allowed us to quantify the extent and the pattern of admixture and gene flow in our data, to select the appropriate populations for informative comparisons and to identify a subset of far eastern populations, which may safely be regarded as deriving from oldest expansion (under the **MD** model). Remarkably, some populations showed more than 99% contribution from the same ancestral population (e.g. West Africa, Europe, and New Guinea), whereas other

populations include several individuals with an apparently admixed genomic background, possibly resulting from successive gene flow (e.g. back migration from Europe to Northeast Africa).

We moved then to consider the patterns of LD, in order to reconstruct two key parameters of human evolution (the effective population size, N_e, and the population divergence time, T. Under neutrality, genetic differences between populations accumulate because of genetic drift, and so their extent (measured by F_{ST}) depends on two quantities; it is inversely proportional to N_e and directly proportional to T. This means that, to estimate T from F_{ST}, one needs independent estimates of N_e. Therefore, to estimate T from genetic difference between populations, independent estimate of N_e are needed, for this purpose we considered the relationship between N_e and the level of LD within each population. The levels of LD depend on both N_e and on the recombination rate between the SNPs considered (Tenesa et al., 2007). However, LD between SNPs separated by large distances along the chromosome reflects relatively recent N_e whereas LD over short recombination distances depends on relatively ancient N_e (Hayes et al., 2003). Thus, we estimated LD independently in each population using all polymorphic markers available for that population (from a minimum of ~ 90,000 to a maximum of ~ 370,000 markers), then we calculated the populations' N_es through time using the equation proposed by McEvoy et al. (2011). The obtained estimates agree well with previous studies (Yu et al., 2004; Conrad et al., 2006; McEvoy et al., 2011), suggesting that the procedure followed is accurate.

This way, from the pairwise F_{ST} values estimated over all loci as described by Weir and Cockeram (1984), and based on the independently-estimated values of N_e, we could infer the divergence times between populations as in Holsinger and Weir (2009). We found that the populations at the extremes of the geographical range considered differ substantially in the timing of their separation from the Eastern African populations, i.e. those located in the most plausible site of departure of AMH expansions (Ramachandran et al., 2005). Extreme divergence values were observed for Europe and the Caucasus on the one hand, and for Australia or New Guinea on the other, respectively at the lower and the upper tails of the distribution. Even considering the full range of uncertainty around these estimates (95% of the confidence interval) we observed no overlap, with Europe having an

upper confidence limit 77 kya and Australia having a lower confidence limit 88 kya. In addition, we showed by simulation approach based on the neutral coalescent model of the software *ms* (Hudson, 2002), that the different times of separation from East Africa estimated for Europe and Australia/New Guinea cannot be reconciled with a model assuming a single, major dispersal of all non-Africans through the classical "northern" route.

Taking into account the recent results on the genetic relationship between modern human and Denisovan (Meyer et al., 2012), we also considered the possibility that the apparent difference in African divergence times for Europe and Australia/New Guinea may somewhat reflect Denisovan admixture. Therefore, we removed from the analysis the SNPs that were identified as representing the Denisovan contribution to the latter's genome and reestimated the divergence times from Africa, finding they are still very close to those previously estimated.

Other Far Eastern populations, besides Australia and New Guineans, may have taken part in an early exit from Africa through a "southern" route; however, recent admixture events could have obscured the genomic signatures of the first migration out of Africa in these Southeast Asian populations, ultimately biasing downwards the estimates of their divergence times from Africans. To understand whether that could have actually been the case, we used a method, *TreeMix* (Pickrell and Pritchard, 2012) to estimate from genome-wide data a maximum-likelihood tree of populations, and then to infer events of gene flow after the split by identifying populations that poorly fit the tree. We selected from our dataset just the populations showing at least 30% of the ancestral genetic component to which all Australian and New Guinean genotypes could be associated in the previous *ADMIXTURE* analysis. Evidence for extensive genetic exchanges after population splits was apparent from East Asia toward populations putatively involved in the early African dispersal (i.e. Fijians, East Indonesians, Moluccans and Polynesians).

At the end, to test which model of African expansion can better explain the observed pattern of genomic variation, we compared the genomic differences between populations with alternative geographic distance matrices calculated respectively one according to (1) a **SD** model; (2) a **MD** model assuming that all Asian populations are descended from ancestors who left Africa through the Arab Peninsula and the Indian Subcontinent, all the

way to Melanesia and Australia (based on skull morphology (Lahr and Foley, 1994); (3) a **MD** model assuming that only the populations of Southeast Asia and Oceania derive from the earlier expansion, whereas Central Asian populations are attributed to the later African dispersal (Ghirotto et al., 2011a). Conversely under the **SD** scenario, anatomically modern humans left Africa through Palestine and dispersed into all of Eurasia (Stringer, 2002).

In all cases, migration routes were constrained by 5 obligatory waypoints. To obtain a realistic representation of migrational distances between populations, we did not estimate the shortest (great-circle) distances between sampling localities, but we modeled resistance to gene flow, based on the landscape features (mountain ranges, arms of sea, rivers) known to influence human dispersal.

To minimize the effects of recent gene flow unrelated with the first human dispersals, which was clearly not negligible (see previous section) we selected populations with at least 80% of a single ancestral component in the ADMIXTURE results (i.e. Australia, the Caucasus, East Africa, East Asia, Europe, New Guinea, South Africa, South India, West Africa) and we evaluated by partial Mantel tests (Mantel, 1967) the correlation between genomic (F_{ST}) and geographic distances, while holding diverge times (T) constant. This way we could control for the drift effects, due to the fact that populations separated at distinct points in time and space. The correlation between genetic and geographic distances was higher under the **MD** than under the **SD** model, but this difference was not statistically significant. This may be due, at least in part, to the fact that the three models being compared share several features, such as the same set of geographic/genetic distances for the European populations, which reduces the power of any test. However, the separation times previously estimated made us confident that the **SD** model is not inconsistent with the data, and so what was really important at this stage was the comparison between the two **MD** models. The better fit of **MD2** than **MD1** implies that the **MD** model works better if only part of the Asian genomic diversity is attributed to the earliest dispersal.

In short, analyses of genomic data based on different sets of assumptions and different methods agreed in indicating: (*i*) that a model with a single early dispersal from Africa fails to account for one crucial aspect of human genome diversity, the distribution of divergence times from Africa, and (*ii*) that within the model of multiple dispersal,

geographical patterns of genome diversity are more accurately predicted assuming that not all Asian and New Guinea/Australian populations have had the same evolutionary history. In the light of these results, we proposed that at least two major dispersal phenomena from Africa led to the peopling of Eurasia and Western Oceania. These phenomena seem clearly distinct both in their timing and in their geographical scope, with some populations of Southern Asia evolved largely independently from those of Northeast Asia retaining the signal of an early dispersal.

Conclusion

The processes of human expansion into new territories, population split, isolation, divergence and admixture are notoriously complex, and often overlap in time or place. It comes as no surprise that, despite fast progress in paleoanthropology and genetics, disentangling and identifying them has remained problematic. Our results, which look at divergence times from Africa in several worldwide populations taking into account large amount of genomic data, point to a more complex **OOA** scenario. They are unambiguous in their support of multiple dispersal into Eurasia, with Australians and New Guineans retaining the signal of "southern" route dispersal.

These results might call into question the genetic relationships between AMH and Neandertals. If dispersal through a "southern" route was substantial, most ancestors of Melanesians would have missed by 2,000 km or so the nearest documented Neandertals with whom they could have intercrossed. Thus, it may be that the 1 to 4% of apparent Neandertal contribution to non-African genomes (Green et al., 2010) reflects phenomena that did not occur after the first exit of AMH from Africa but instead date back to an earlier time. Another possibility is that an ancient structuring of populations might contribute to explaining the observed pattern of resemblance between modern humans and Neandertals. In these ways, some of contemporary humans may still be carrying in their genome traces of a closer genetic relationship with the Neandertals' ancestors, without this necessarily meaning that any admixture took place after anatomically archaic and modern human forms separated.

Continued field work, alongside rapid advances in modern and ancient genome sequencing, will allow for greater resolution in modelling the spatial and temporal dimension of modern human origin and dispersal.

The results of this research are now in the form of a manuscript submitted for publication (see **PAPER II**).

Chapter 3.TOWARD A GLOBAL TREE OF HUMAN LANGUAGES AND GENES

The evolution of human languages has probably proceeded in parallel with the evolution of human populations, although certainly not in a mechanical manner. Indeed, both are subjected to similar patterns of transmission of traits, in cultural and biological terms, and their development is influenced by the same demographic changes. Thus, studies of language phylogenies and their correlations with genetic phylogenies can enrich our understanding of human prehistory and provide insights into the processes that shaped both genetic and linguistic diversity (Pakendorf, 2014). The first intuition about this parallel can be found in the Origin of Species (1859), when Darwin suggests that biological and linguistic data could describe similar genealogies:

"*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; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one*"

Coevolution of gene and languages

The parallel study of processes of linguistic and genetic evolution was first undertaken in the late 1980s and early 1990s, when sufficient allele frequency data for a large number of human populations had been collected to make such research feasible. Some geneticists (Cavalli-Sforza et al., 1988; Sokal, 1988) advocated 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 (**Fig. 3.1**). However, their work has been received with much scepticism and has remained controversial among linguists: for virtually no professional historical linguist unconditionally subscribes to the reliability of the linguistic genealogies used as matches in such experiments. Indeed, most linguists have denied the very possibility of a reliable global or long-range classification of languages, advocating methodological reasons which brought

the interdisciplinary debate on large-scale population-language congruence close to a dead end, and which perhaps we are now in a position to overcome.

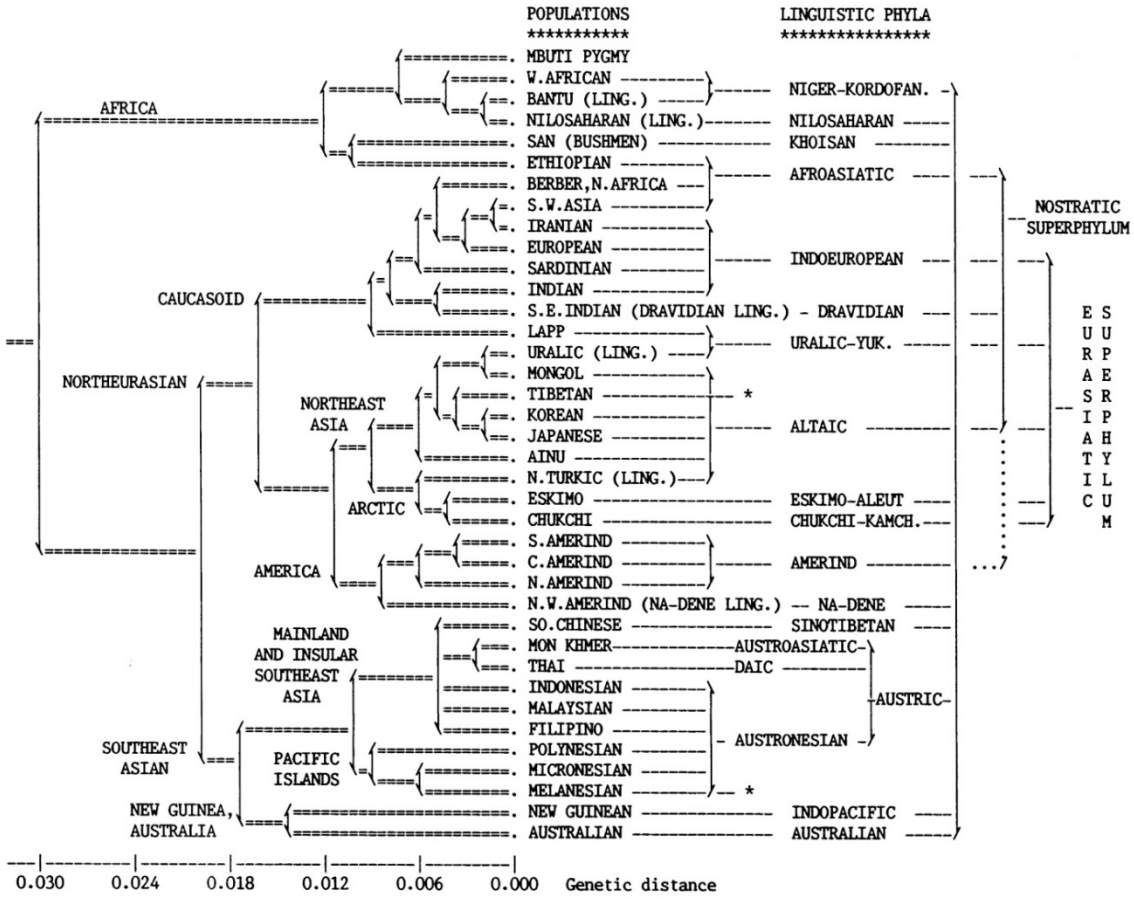


Figure 3.1 - Comparison of genetic tree and linguistic phyla. Figure reproduced from Cavalli-Sforza et al., (1988)

Indeed, recent genetic work showed that, in a large number of case studies, patterns of genetic and linguistic diversity do appear locally well correlated, implying that language can represent a barrier to gene flow (Barbujani and Sokal, 1991; Belle and Barbujani, 2007; Tishkoff et al., 2009). However, the methods adopted in these works were mostly based on the comparison of vocabulary items, hence generally lacking resolution when comparing across linguistic families, and making impossible to verify the congruence hypothesis at a general, worldwide rather than regional, scale (Nichols, 1996; Longobardi, 2003; Heggarty,

2004). A major limitation of these methods concerns the time depth that we can reach, since most linguists agree that language families cannot be traced back after an estimated age of 10 kya: beyond that time limit, there will be no detectable similarities between pairs of languages (Gray, 2005). Put in a simpler perspective, languages evolve faster than genes (Cavalli-Sforza et al., 1994), and language phylogenies coalesce to the common root (proto-language) within a narrow prehistorical depth. Besides, accidental similarities tend to emerge due to the limited constraints on possible phonological systems; accordingly, a general congruence between patterns of genetic and linguistic diversity appears to be difficult to demonstrate, as long as the latter are assessed by means of vocabulary comparisons.

New linguistic tools

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. Therefore, we chose to approach this goal in a radically different way focusing on structural linguistic features such as the order of subject, verb, and object or the presence/ absence of definite or indefinite articles. In contrast to lexical features, structural features of languages change at a slower rate, thus being more suitable for the investigation of genealogical relationships at deeper time depth (Dediu and Levinson, 2012; Sicoli and Holton, 2014). We took advantage of a new tool developed for language comparison focused on the syntactic features: that is the Parametric Comparison Method (PCM, see TableA in **Appendix** (Longobardi and Guardiano, 2009). This approach describes the core grammar of any language as a string of binary symbols, each encoding the value of a syntactic parameter (Chomsky, 1981; Baker, 2001). Since this method assumes that these parameters are part of the innate Universal Grammar, they should be found, and hence comparable, across all languages irrespective of their degree of genealogical relationship making them perfectly comparable to genetic data and avoiding the problems inherent in the use of lexical data mentioned above.

DNA diversity mirrors grammar within Europe

The study described below is part of a European Research Council project (ERC-2011-AdG_295733 grant) LanGeLin (Language and Genetic Lineages), in which the group of Population genetics of the University of Ferrara collaborates with the Linguistic group of York and the Molecular anthropology group of Bologna. Through the use of shared statistical and computational tools, LanGeLin aims to build up comparable phylogenetic trees of strategically chosen languages and populations, and therefore to test in the strongest possible way Darwin's expectation about their eventual congruity, both on local and global scales.

The first step of the LanGeLin project is a comparative analysis of genome-wide information and language structure at a cross-language families scale in Europe. For this purpose, the choice of populations was conditioned on the overlap between the languages so far analysed by the PCM method (Longobardi and Guardiano, 2009) and the publicly available genome-wide datasets. In the end, we could collect linguistic and genetic information about 15 populations belonging to three different linguistic families (i.e 12 Indo-European (hereafter: IE), two Finno-Ugric and the Basque). The final genetic dataset comprises 177,149 markers that passed the quality filters in 805 individuals.

We first checked that PCM correctly reproduces the known historical relationships of the Indo-European languages of Europe. For the our 12 languages belonging to IE, we calculated and compared distances and phylogenies both from the list of lexical cognates developed by Bouckaert et al. (2012) and through PCM. A good correlation ($r= 0.82$) was found between syntactic and lexical distances in the subset, showing that the well-established set of relationships among European IE languages can actually be reconstructed with good statistical confidence from syntactic comparisons. We then moved to analyse the complete linguistic dataset, including also the non-IE languages of Europe, with the aim to evaluate the PCM's ability to compare languages even from different families. Different standard methods of evolutionary biology (i.e UPGMA trees, PCA, DAPC) showed that the main families and subfamilies of Europe were discriminated through just 56 abstract syntactic characters, without resorting to methodologically disputable lexical comparisons.

At this point, we wanted to test whether genetic and linguistic diversity are correlated in Europe, and what is the role of geographical distances in that correlation. We

observed significant correlation between genomic and linguistic diversity (*r=0.577*), meaning that populations speaking similar languages also tend to resemble each other at the genomic level, suggesting that cultural change and biological divergence have proceeded in parallel in Europe. This correlation remained significant after removing the effects of geography by a partial Mantel test (Mantel, 1967) and after Bonferroni correction for multiple tests. Confirming what had been observed in studies in which lexical differences had been used (Sokal, 1988; Belle and Barbujani, 2007), we found that populations speaking similar languages also tend to be genetically closer than expected on the sheer basis of their geographic location. . Contrary to previous studies, however, here geographic distances appeared to be poorer predictors of genomic differences than linguistic distances in Europe (*r=0.228*).

The comparison of more detailed analysis of syntactic and genetic diversity pointed out some exceptions to the conclusion above. When the relationships were summarized by trees and by PCA, some divergences between linguistic and genetic phylogenies emerged. The main elements of disagreement were represented by the positions of Hungarians and Rumanians, which clustered genetically with speakers of Serbo-Croatian despite being highly different syntactically. Because all these populations dwell in Central Europe, these apparent violations of the biological relationships expected from linguistic history can be plausibly accounted for by the gene flow between neighbouring countries. Thus, we further investigated the evolutionary relationships between populations by a method (i.e. *TreeMix*, (Pickrell and Pritchard, 2012)) designed to identify relatively recent gene flow episodes after the main population splits. We found evidence of contacts between speakers of IE-subfamilies from Russia and Greece into Romania. These episodes of gene flow are in intriguing correspondence with the eccentric position of Rumanian in the language tree and with the observation that Rumanian forms a cluster distinct from that of the other Romance languages in a cluster membership analysis (Jombart et al., 2010). Similarly, a Southern European origin of a fraction of the Hungarians corresponds to a closer resemblance of the IE languages of our sample with Hungarian than with Finnish. Therefore, processes of relatively recent gene flow seem to nicely explain at least a fraction of the linguistic variation unaccounted for by the classical classification of languages into families. Besides, the PCM method allows one to identify some elusive aspects of population history providing insight

into processes of gene flow and cultural contacts, which would likely escape detection if only studied at the genomic level only.

Conclusion

In summary, our study proved that an effective comparison of genetic and syntactic distances at a wide scale across different linguistic families can be successfully achieved by pairing high-resolution genomic markers with the syntactic parameters underlying the PMC. In the light of our results, the PCM seems to be a powerful method to explore the relationship between distant related languages and populations, and this work represents a first step towards this direction.

The task to improve our understanding of the history and evolution of our species is hampered by the fact that human groups behave in ways that frequently make it impossible the attempt to fit the biological data to simple models of population history. Using data from several research fields, one can try to look for areas of congruence and ultimately obtain important insight about otherwise elusive past populations events (Renfrew, 2010).

So far, it is clear that the comparison of genetic and linguistic data can elucidate and complement both human population prehistory and the dynamics underlying language evolution (Cavalli-Sforza, 1997). The history of languages may, or may not, parallel the genetic history of their speakers. Thus, linguistic (as well as archaeological and historical) data are of crucial importance for generating hypotheses that can be tested at the molecular level, casting further light on the complex processes at play in the demographic history of modern humans.

The results of this study led to the publications of one paper (see **PAPER III**).

Chapter 4.GENEALOGICAL INFERENCES FROM MODERN AND ANCIENT DNA DATA

The use of genetic data can help inferring evolutionary and biological relationships of human populations. Even though genetic analysis will not answer questions about ethnic or cultural identity, it represents an essential tool to integrate archaeological data by providing information about migration dynamics, population structure, and relationships among culturally differentiated groups of individuals. The genetic data so far produced highlight the fact that the human genome is a mosaic of fragments of different origins (Henn et al., 2012), indicating a complex network of interactions between populations, a result of multiple origins, large-scale population movements and subsequent extensive gene flow (Novembre et al., 2008). For many years, studies of human genetic diversity have been necessarily limited to modern populations and the evolutionary dynamics or the genetic structure of past population were usually inferred from model-based analyses of the modern genetic diversity. However, even when inferred from large collections of data (Ralph and Coop, 2013), patterns of modern genomic variation provide ample but noisy signals that can only seldom be safely connected with specific historical events. Thus, to investigate the detail of these past processes, we necessarily need to include genetic information coming from past populations (Ramakrishnan et al., 2005) (see **Chapter 1**).

Ancient DNA data offered the possibility to test the common (and often inevitable) assumption that the unknown allelic distribution in past populations is approximated by the contemporary allelic distribution in the same area, showing that modern populations may not be in direct genetic continuity with local ancestors. Indeed, a pioneering work showed that, in Sardinia, modern populations separated by only tens of kilometers could differ sharply in their genealogical relationships with ancient populations (Ghirotto et al., 2010). In this research, carried out in our lab, for the first time an Approximate Bayesian Computation (ABC) inferential framework was applied to datasets of ancient and modern human variation, to compare several demographic models and choose the one which best accounts for the observed variation.

Once the potentiality of this approach have been demonstrated, we applied this method to address several anthropological questions, such as the interaction of anatomically modern humans with archaic forms (i.e. Neandertals in Europe) (Ghirotto et al., 2011b), the origins and evolution of the Etruscan population (**PAPER IV AND PAPER V**) and, more recently, the nature of the Longobard migrations into the Roman world (Vai et al., 2015).

Approximate Bayesian Computation

Past demographic and evolutionary dynamics influence the distribution of the genetic diversity at any given moment in time, and so, in principle one can retrospectively infer episodes in population history from genetic diversity data. In practice, though, many different combinations of evolutionary processes may lead to any observed distribution of genetic variables. One of the most powerful statistical approaches available to reconstruct populations' historical dynamics, when calculation of likelihoods is too complicated, involves the use of genealogical simulations through ABC (Beaumont et al., 2002; for a review see e.g. Bertorelle et al., 2010). The ABC machinery combines the analysis of abundant genetic data and realistic modelling, allowing the probabilistic comparison among different models of evolution, the simultaneous estimation of demographic and evolutionary parameters, and the quantitative evaluation of the results' credibility. Moreover, the Bayesian philosophy allows one to incorporate in the analysis the prior information about model parameters, such as mutation rate, effective population sizes for both modern and ancient populations, separation time (for models involving more than one population) and migration rate. This increases considerably the power to draw inference about the populations' evolutionary histories.

In short, the ABC method is based on comparison between statistics calculated on the observed dataset of genetic variation, and the same set of statistics recalculated on datasets resulting from large numbers (often millions) of simulations across a wide range of parameter values within different demographic models. Simulations producing genetic data (i.e. statistics) closest to those observed are used to identify the model best accounting for the observed data (Pritchard et al., 1999; Beaumont, 2008), as well as to estimate the posterior distributions of its parameters (Beaumont et al., 2002; Leuenberger and Wegmann, 2010). One of the recent extensions of ABC studies involves the possibility of

considering populations of different time-periods at the genomic level, thus meaning that we have more power to detect changes in the dynamics of a population, rendering the analysis more informative.

A scheme of a complete ABC analysis is outlined in **Fig. 4.1**.

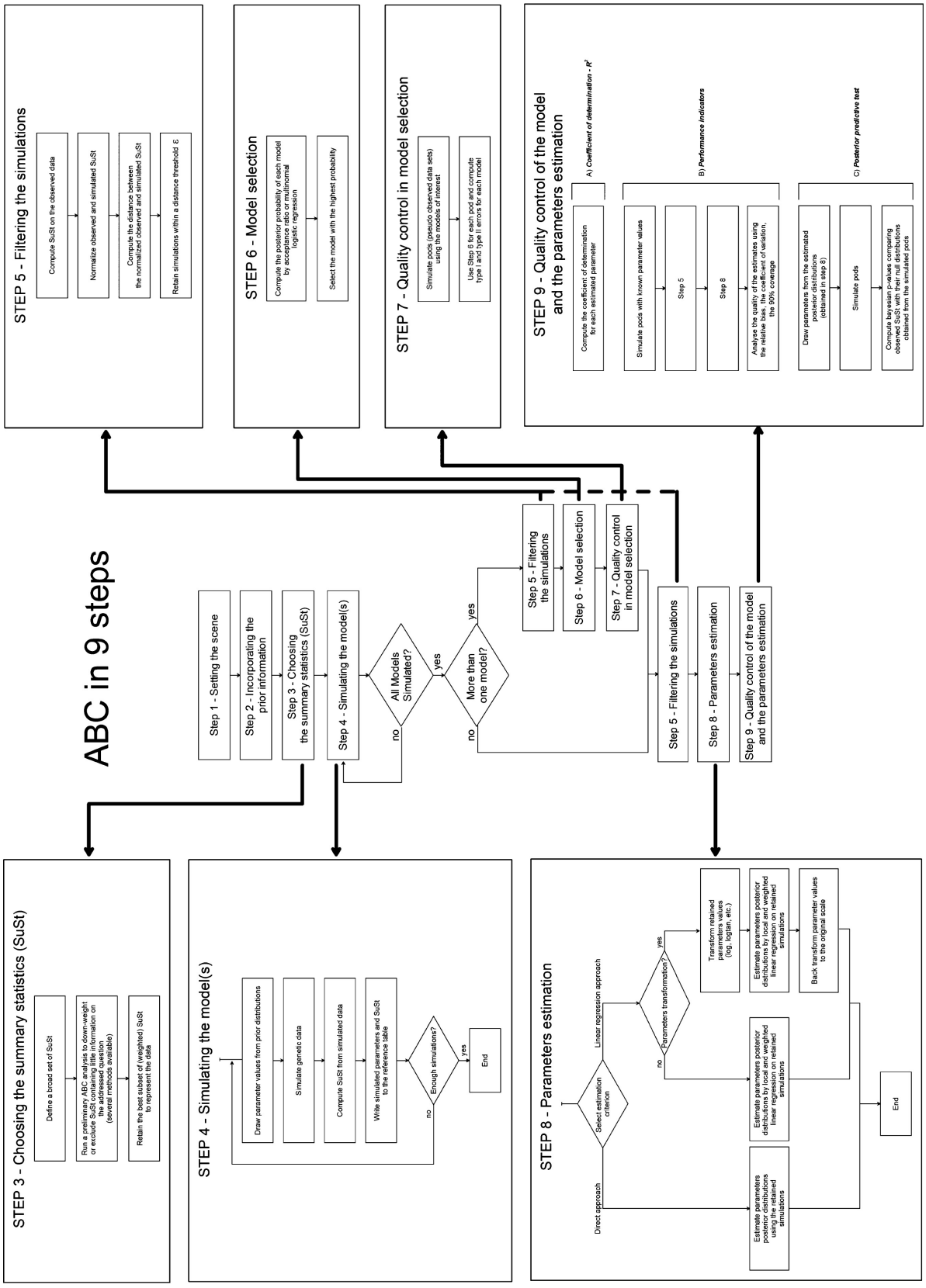


Figure 4.1 - ABC in nine steps. From Bertorelle et al., 2010.

Origin and evolution of the Etruscans' DNA

In this section of my thesis I present the results about the genealogical relationships between Etruscans and modern Tuscans. In particular, I tried to address two questions:

I whether an analysis at the small geographical scale can provide evidence of a genealogical continuity between the Etruscans and some current inhabitants of historical Etruria,

II whether the observed degree of genetic resemblance between modern inhabitants of Tuscany and Western Anatolia contains information relevant to the debated question of the Etruscans' origins.

The long-standing debate about Etruscan origin

The first urban settlements in Tuscany (Italy) date back to the Iron-Age, eighth century BC, and are associated with the onset of the Etruscan culture. Modern Tuscany broadly corresponds to the core of the Etruscan territory, or Etruria, and indeed the word 'Tuscany' itself is derived from 'Etruscan'. The Etruscan communities shared a non-Indoeuropean language, a religion and a material culture, but they never formed a political unit. According to ancient historians, the resemblances between Etruscans and other Iron-Age populations were extremely low, since they did not share language, lifestyle or customs (Barker and Rasmussen, 1998). Between the seventh and the fifth centuries, leagues of Etruscan cities exerted a crucial cultural and political role in the Mediterranean area. In the first century BC, the Etruscans obtained Roman citizenship, and their language and culture vanished from the archaeological record (Pallottino, 1975; Barker and Rasmussen, 1998). There is a long lasting controversy about the origin of the Etruscan population, whether local or Anatolian. To date, there is consensus among modern archaeologists that the Etruscan culture developed locally, with some features suggesting an Eastern influence; this hypothesis was also shared by the ancient historian Dionysius of Halicarnassus (Barker and Rasmussen, 1998). However, other ancient historians like Herodotus and Livy regarded the Etruscans as immigrants, respectively, from Lydia (modern Western Anatolia) or from North of the Alps. Modern experts definitely support the former view, but affinities between the Lydian and the Etruscan languages seem to exist (Beekes, 2002). Unfortunately, no historical documents are

available to help address this question. In fact, even if we understand reasonably well the Etruscan language, the surviving Etruscan texts are almost exclusively funerary or religious inscriptions, containing basically no historical information. However, a language or a culture can rapidly get extinct, but that is certainly not the case for the DNA of its speakers.

Genetic studies about the Etruscans without Etruscans

In the last years, in the absence of any ancient genetic information, it was generally assumed that modern Tuscans are descended from Etruscans. The Etruscans' origins were thus studied comparing Tuscans and other modern populations (Piazza et al., 1988; Achilli et al., 2007; Brisighelli et al., 2009). In the first such study, in pre-DNA times, Piazza et al. (1988) analysed 34 blood group and HLA (human leukocyte antigen) allele frequencies at 7 loci, collected in 28 Italian locations. The allele frequencies were turned into a series of synthetic variables by PCA. The main principal components were then interpolated and projected on the geographic map, obtaining a graphical representation of genetic diversity in space. A high heterogeneity was evident among Italian regions and different geographical patterns emerged, which roughly resembled the distribution of some ancient Italic cultures. The first principal component appeared distributed in a North–South gradient, which was interpreted as reflecting the Northwards dispersal of people of Greek origin. The map representing the second principal component showed a peak in an area of Central Italy not far from ancient Etruria, and a similar peak in North–Western Italy. This map was interpreted as evidence of the persistence of Etruscan genetic features in Tuscany and neighbouring regions.

Ten years later two DNA studies tried to address the relationships between Etruscans and contemporary Tuscans. Achilli et al. (2007) analysed the mtDNA of 322 samples from three areas where archaeology suggests a possible persistence of the Etruscans' biological inheritance. These were Murlo, an isolated hill village, Volterra, a former major Etruscan city, and the Casentino valley. Achilli et al. (2007) compared Tuscany sequence variation with that of 55 Eurasian populations. Eleven haplotypes were shared between Tuscans and near Eastern populations (3-fold higher than that observed in neighbouring regions) and were absent in all other European samples. The authors concluded that these haplotypes represent the Etruscan's genetic legacy and that their Eastern features support the historical validity of Herodotus' narrative. This view was further supported by Brisighelli et al.(2009),

who analysed Tuscan samples combining information from broader regions of the mtDNA chromosome. They showed that around 10% of Tuscans (26 individuals out of 258) carry haplotypes that are typically from the Near East. This observed similarity might be due to a common origin at any time in the past, but the authors viewed their data as supporting a recent historical connection with Anatolia due to migratory contacts leading to the development of the Etruscan culture. This interpretation depends strictly on the assumption that modern Tuscans are the Etruscans' descendants. However, studies of ancient DNA showed that that is not the case.

Genetic studies about the Etruscans with the Etruscans

In 2004, for the first time, Vernesi and collaborators analyzed Etruscans' mtDNA obtained from 27 different individuals from 10 necropoleis, covering Etruria in terms of both chronology and geography. The study of ancient samples highlighted the genetic similarities between the Etruscans and the current population of Turkey, but not with Italian populations other than Tuscans (even though they shared only two haplotypes). Further information on the relationships between Ironage and current inhabitants of Tuscany was sought by investigating the time window separating them. Guimaraes et al. (2009) obtained the mtDNA sequences of 27 from Medieval Tuscan individuals, clear similarities were observed between Middle-Age and contemporary Tuscans, but not with the Etruscans, thus suggesting that a substantial demographic change had taken place before AD 1,000. The claim that systematic, although unspecified, errors in the ancient DNA sequences had led to flawed genealogical inference (Bandelt and Kivisild, 2006; Achilli et al., 2007) was not supported by careful reanalysis of the Etruscan data (Mateiu and Rannala, 2008).

Inferring demographic history by Approximate Bayesian Computation analysis

Simple, eyeball comparisons of DNA data can give us a general idea of the relationships between past and present populations, but by using more complex biostatistical approaches it is possible to formally test hypotheses. My study represents the first effort to shed light on the origin and evolution of the Etruscans' DNA considering ancient DNA data and explicitly testing demographic models of evolution within the ABC framework. Besides, previous studies did not consider the potential effects of genetic divergence when populations are

structured or subdivided. If most Etruscans' descendants lived in isolated communities in the last 2,000 years, their DNAs may still persist in some localities, but will escape detection unless they are sought at the appropriate (i.e., smaller) geographical scale.

To investigate in greater geographical detail the biological relationships between contemporary and ancient populations, we explicitly tested alternative demographic models by ABC. We typed an additional set of ancient DNA sequences, and compared the levels of genetic diversity in the mtDNAs of the enlarged Etruscan sample with Medieval Tuscans (Guimaraes et al., 2009), and four modern Tuscans population; three in historical Etruria, namely Casentino, Murlo and Volterra (Achilli et al., 2007), and one from Florence (Turchi et al., 2008), representing the general Tuscan population. The results were compatible with a genealogical continuity between the Etruscans and two Tuscan isolates (Volterra and Casentino). By contrast, another population of the former Etruscan homeland, Murlo, and a forensic sample from the main city in the area, Florence, showed no special relationships with the Etruscans. These findings mean the that Etruscans cannot be regarded as the global ancestors of the people now living in what once was their territory (see **PAPER IV**), but that their genetic legacy is still present, and detectable, when modern populations are separately considered (as opposed to clumping them together).

We then asked whether genetic similarities between current Tuscans and Anatolians (Achilli et al., 2007; Brisighelli et al., 2009) provide some evidence for an Etruscan homeland in Anatolia. Because previous inhabitants of Etruria, associated with the Villanovian culture, cremated their dead, empirical genetic comparisons going further back in time are unfeasible. We exploited the algorithm of the IM methods to estimate the most probable separation time between Anatolians (from Di Benedetto et al., 2001) and the Tuscans populations showing genealogical continuity with the Etruscans. Our basic hypothesis was that if the genetic resemblance between Turks and Tuscans reflects a common origin just before the onset of the Etruscan culture, (meaning that the Etruscan population came from Anatolia as hypothesized by Herodotous) we would expect that the two ancestral populations separated around 3,000 ya. Assuming an average generation time of 25 years, a plausible mutation rate, and complete isolation after the split from the common ancestors,

the estimates of the separation time between Tuscany and Anatolia was around 7,600 ya, with a 95% credible interval between 5,000 and 10,000 (see **PAPER IV**).

We then compared the observed genetic data with the results of millions of simulations of modern and ancient mtDNAs, generated under demographic models differing for the homelands of the Etruscan people, namely, Western Anatolia or Central Italy. This way, we could test whether or not the genetic links between modern Anatolians and Tuscans may have been established through a process of gene flow occurring approximately between the tenth and eight centuries BC, and thus possibly associated with the onset of the Etruscan civilization in Italy. The results, confirming the previous analysis based on modern data only, indicated that the genetic links between Tuscany and Anatolia date back to at least 5,000 ya, suggesting that this genetic link is too old to be due to a migration occurring just before the appearance of the first archaeological evidence of the Etruscan culture. Therefore, it is safe to conclude that the Etruscan culture developed locally and not as an immediate consequence of immigration from the Eastern Mediterranean shores (see **PAPER V**).

Conclusion

For many years, studies of human genetic diversity have been necessarily limited to modern populations, severely limiting our ability to investigate the detail of past processes. With the advent of methods for reliably typing ancient DNA, it has been possible to increase the power in reconstructing historical demographic processes, and in explicitly testing evolutionary hypotheses. Combining this advance and the statistical power provided by model-based methods such as ABC, it is now possible to clarify other long-standing evolutionary questions, and to highlight aspects of human history at an unprecedented resolution.

The result of this research led to the publications of two papers (see **PAPER IV** and **PAPER V**).

BIBLIOGRAPHY

1000 Genomes Project Consortium, Durbin RM, Abecasis GR et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319):1061-1073.

Achilli A, Olivieri A, Pala M et al. 2007. Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. *Am J Hum Genet* **80**(4):759-768.

Ahn SM, Kim TH, Lee S et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**(9):1622-1629.

Alexander DH, Novembre J, and Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**(9):1655-1664.

Baker M. 2001. The atoms of languages. New York: Basic Book.

Balaresque PL, Ballereau SJ, and Jobling MA. 2007. Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet* **16 Spec No. 2**:R134-139.

Bamshad MJ, Mummidi S, Gonzalez E et al. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. *Proc Natl Acad Sci U S A* **99**(16):10539-10544.

Bandelt HJ, and Kivisild T. 2006. Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. *Annals of human genetics* **70**(3):314-326.

Barbujani G, Bertorelle G, and Chikhi L. 1998. Evidence for Paleolithic and Neolithic gene flow in Europe. *Am J Hum Genet* **62**(2):488-492.

Barbujani G, and Colonna V. 2010. Human genome diversity: frequently asked questions. *Trends Genet* **26**(7):285-295.

Barbujani G, Ghirotto S, and Tassi F. 2013. Nine things to remember about human genome diversity. *Tissue Antigens* **82**(3):155-164.

Barbujani G, Magagni A, Minch E, and Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* **94**(9):4516-4519.

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

Barbujani G, and Tassi F. 2012. Genetic Data in Forensic Science: Use, Misuse and Abuse. In: Bin R, Lorenzon S, and Lucchi N, editors. Biotech Innovations and Fundamental Rights: Springer Milan. p 243-259.

Barker G, and Rasmussen T. 1998. The Etruscans. Oxford: Blackwell.

Barreiro LB, Laval G, Quach H et al. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**(3):340-345.

Beaumont M. 2008. Joint determination of topology, divergence time and immigration in population trees. Cambridge: McDonald Institute for Archaeological Research. 135-154 p.

Beaumont MA, Zhang W, and Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162**(4):2025-2035.

Beekes R. 2002. The prehistory of the Lydians, the origin of the Etruscans, Troy and Aeneas. *Biblioteca Orientalis* **59**(3-4):206-242.

Belle EM, and 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.

Bentley DR, Balasubramanian S, Swerdlow HP et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218):53-59.

Bertorelle G, Benazzo A, and Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* **19**(13):2609-2625.

Blome MW, Cohen AS, Tryon CA et al. 2012. The environmental context for the origins of modern human diversity: a synthesis of regional variability in African climate 150,000-30,000 years ago. *J Hum Evol* **62**(5):563-592.

Bouckaert R, Lemey P, Dunn M et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* **337**(6097):957-960.

Bowden R, Sakaoka H, Ward R, and Donnelly P. 2006. Patterns of Eurasian HSV-1 molecular diversity and inferences of human migrations. *Infect Genet Evol* **6**(1):63-74.

Bowler JM, Johnston H, Olley JM et al. 2003. New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature* **421**(6925):837-840.

Braverman JM, Hudson RR, Kaplan NL et al. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**(2):783-796.

Briggs AW, Good JM, Green RE et al. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**(5938):318-321.

Brisighelli F, Capelli C, Alvarez-Iglesias V et al. 2009. The Etruscan timeline: a recent Anatolian connection. *Eur J Hum Genet* **17**(5):693-696.

Brown P, Sutikna T, Morwood MJ et al. 2004. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**(7012):1055-1061.

Cann RL, Stoneking M, and Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* **325**(6099):31-36.

Caramelli D, Lalueza-Fox C, Condemi S et al. 2006. A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr Biol* **16**(16):R630-632.

Cavalli-Sforza LL. 1966. Population structure and human evolution. *Proc R Soc Lond B Biol Sci* **164**(995):362-379.

Cavalli-Sforza LL. 1997. Genes, peoples, and languages. *Proc Natl Acad Sci U S A* **94**(15):7719-7724.

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

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

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055):69-87.

Chomsky N. 1981. Lectures on government and binding. Dodrecht: Foris.

Clark JD, Beyene Y, WoldeGabriel G et al. 2003. Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423**(6941):747-752.

Conrad DF, Jakobsson M, Coop G et al. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**(11):1251-1260.

Cooper A, and Poinar HN. 2000. Ancient DNA: do it right or not at all. *Science* **289**(5482):1139.

Cox M. 2007. Extreme patterns of variance in small populations: placing limits on human Y-chromosome diversity through time in the Vanuatu Archipelago. *Annals of human genetics* **71**(Pt 3):390-406.

Darwin C. 1859. On the Origin of Species: John Murray.

Dediu D, and Levinson SC. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS One* **7**(9):e45198.

DeGiorgio M, Jakobsson M, and Rosenberg NA. 2009. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci U S A* **106**(38):16057-16062.

Di Benedetto G, Erguven A, Stenico M et al. 2001. DNA diversity and population admixture in Anatolia. *Am J Phys Anthropol* **115**(2):144-156.

Durand EY, Patterson N, Reich D, and Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28**(8):2239-2252.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57-74.

Endicott P, Gilbert MT, Stringer C et al. 2003. The genetic origins of the Andaman Islanders. *Am J Hum Genet* **72**(1):178-184.

Eriksson A, and Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc Natl Acad Sci U S A* **109**(35):13956-13960.

Fagundes NJ, Ray N, Beaumont M et al. 2007. Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* **104**(45):17614-17619.

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, Meyer M, Gao X et al. 2013a. DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci U S A* **110**(6):2223-2227.

Fu Q, Mittnik A, Johnson PL et al. 2013b. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* **23**(7):553-559.

Fulton TL. 2012. Setting up an ancient DNA laboratory. *Methods Mol Biol* **840**:1-11.

Gansauge MT, and Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat Protoc* **8**(4):737-748.

Ghirotto S, Mona S, Benazzo A et al. 2010. Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Mol Biol Evol* **27**(4):875-886.

Ghirotto S, Penso-Dolfin L, and Barbujani G. 2011a. Genomic evidence for an African expansion of anatomically modern humans by a Southern route. *Hum Biol* **83**(4):477-489.

Ghirotto S, Tassi F, Benazzo A, and Barbujani G. 2011b. No evidence of Neandertal admixture in the mitochondrial genomes of early European modern humans and contemporary Europeans. *Am J Phys Anthropol* **146**(2):242-252.

Ghirotto S, Tassi F, Fumagalli E et al. 2013. Origins and evolution of the Etruscans' mtDNA *PLoS One*.

Gray R. 2005. Evolution. Pushing the time barrier in the quest for language roots. *Science* **309**(5743):2007-2008.

Green RE, Krause J, Briggs AW et al. 2010. A draft sequence of the Neandertal genome. *Science* **328**(5979):710-722.

Green RE, Krause J, Ptak SE et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**(7117):330-336.

Green RE, Malaspinas AS, Krause J et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**(3):416-426.

Guimaraes S, Ghirotto S, Benazzo A et al. 2009. Genealogical discontinuities among Etruscan, Medieval, and contemporary Tuscans. *Mol Biol Evol* **26**(9):2157-2166.

Gunz P, Bookstein FL, Mitteroecker P et al. 2009. Early modern human diversity suggests subdivided population structure and a complex out-of-Africa scenario. *Proc Natl Acad Sci U S A* **106**(15):6094-6098.

Hahn MW, Demuth JP, and Han SG. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* **177**(3):1941-1949.

Hammer MF, Garrigan D, Wood E et al. 2004. Heterogeneous patterns of variation among multiple human x-linked Loci: the possible role of diversity-reducing selection in non-africans. *Genetics* **167**(4):1841-1853.

Harpending HC. 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum Biol* **66**(4):591-600.

Harvati K, Stringer C, Grun R et al. 2011. The Later Stone Age calvaria from Iwo Eleru, Nigeria: morphology and chronology. *PLoS One* **6**(9):e24024.

Hayes BJ, Visscher PM, McPartlan HC, and Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* **13**(4):635-643.

Hebsgaard MB, Phillips MJ, and Willerslev E. 2005. Geologically ancient DNA: fact or artefact? *Trends Microbiol* **13**(5):212-220.

Heggarty P. 2004. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data – and to dating language? Phylogenetic methods and the prehistory of languages In: James Clackson PF, Colin Renfrew, editor. Cambridge: McDonals Institute for Archaeological Research. p 183-194.

Henn BM, Botigue LR, Gravel S et al. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**(1):e1002397.

Henn BM, Gignoux CR, Feldman MW, and Mountain JL. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Mol Biol Evol* **26**(1):217-230.

Higuchi R, Bowman B, Freiberger M et al. 1984. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**(5991):282-284.

Hofreiter M, Serre D, Poinar HN et al. 2001. Ancient DNA. *Nat Rev Genet* **2**(5):353-359.

Holsinger KE, and Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* **10**(9):639-650.

Hublin JJ. 2009. Out of Africa: modern human origins special feature: the origin of Neandertals. *Proc Natl Acad Sci U S A* **106**(38):16022-16027.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2):337-338.

Hunley KL, Healy ME, and Long JC. 2009. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. *Am J Phys Anthropol* **139**(1):35-46.

Ingman M, Kaessmann H, Paabo S, and Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**(6813):708-713.

Jakobsson M, Scholz SW, Scheet P et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**(7181):998-1003.

Jeong C, and Di Rienzo A. 2014. Adaptations to local environments in modern human populations. *Curr Opin Genet Dev* **29**:1-8.

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

Kaessmann H, Wiebe V, Weiss G, and Paabo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet* **27**(2):155-156.

Karafet T, Xu L, Du R et al. 2001. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet* **69**(3):615-628.

Kidd JM, Sampas N, Antonacci F et al. 2010. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat Methods* **7**(5):365-371.

Kircher M. 2012. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol Biol* **840**:197-228.

Kirsanow K, and Burger J. 2012. Ancient human DNA. *Ann Anat* **194**(1):121-132.

Krause J, Fu Q, Good JM et al. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**(7290):894-897.

Krings M, Capelli C, Tschentscher F et al. 2000. A view of Neandertal genetic diversity. *Nat Genet* **26**(2):144-146.

Krings M, Stone A, Schmitz RW et al. 1997. Neandertal DNA sequences and the origin of modern humans. *Cell* **90**(1):19-30.

Lahr MM. 1996. The Evolution of Modern Human Diversity: A Study of Cranial Variation. Cambridge: Cambridge Univ Press.

Lahr MM, and Foley RA. 1994. Multiple Dispersals and Modern Human Origins. *Evolutionary Anthropology* **3**:48-60.

Lander ES, Linton LM, Birren B et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860-921.

Landsteiner K. 1900. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zentralblatt Bakteriologie* **27**:357–362.

Leuenberger C, and Wegmann D. 2010. Bayesian computation and model selection without likelihoods. *Genetics* **184**(1):243-252.

Lewontin R. 1972. The apportionament of human diversity. Evolutionary biology 6. New York: Appleton-Century-Crofts. p 381-398.

Li JZ, Absher DM, Tang H et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866):1100-1104.

Liu H, Prugnolle F, Manica A, and Balloux F. 2006. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* **79**(2):230-237.

Liu N, and Zhao H. 2006. A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics* **2**(6):353-364.

Longobardi G. 2003. Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook* **3**:101-138.

Longobardi G, and Guardiano C. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua* **119**:1679-1706.

Lopez Herraez D, Bauchet M, Tang K et al. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* **4**(11):e7888.

Macaulay V, Hill C, Achilli A et al. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308**(5724):1034-1036.

Madrigal L, and Kelly W. 2007. Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *Am J Phys Anthropol* **132**(3):470-482.

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

Martinez-Cruz B, Vitalis R, Segurel L et al. 2011. In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur J Hum Genet* **19**(2):216-223.

Mateiu LM, and Rannala BH. 2008. Bayesian inference of errors in ancient DNA caused by postmortem degradation. *Mol Biol Evol* **25**(7):1503-1511.

McEvoy BP, Powell JE, Goddard ME, and Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* **21**(6):821-829.

Mellars P. 2006a. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313**(5788):796-800.

Mellars P. 2006b. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc Natl Acad Sci U S A* **103**(25):9381-9386.

Meyer M, Fu Q, Aximu-Petri A et al. 2014. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**(7483):403-406.

Meyer M, Kircher M, Gansauge MT et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**(6104):222-226.

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

Myers S, Bottolo L, Freeman C et al. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746):321-324.

Nichols JA. 1996. The Comparative Method Reviewed: Regularity and Irregularity in Language Change. In: Durie M, and Ross M, editors. The comparative method as heuristic. New York: Oxford University Press. p 39-71.

Novembre J, and Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet* **10**(11):745-755.

Novembre J, Johnson T, Bryc K et al. 2008. Genes mirror geography within Europe. *Nature* **456**(7218):98-101.

Novembre J, and Ramachandran S. 2011. Perspectives on human population structure at the cusp of the sequencing era. *Annual review of genomics and human genetics* **12**:245-274.

O'Connell J, and Allen J. 2004. Dating the colonization of the Sahul (Pleistocene Australia - New guinea): A review of recent research. *J Archaeol Sci USA* **31**:835-853.

Ovchinnikov IV, Gotherstrom A, Romanova GP et al. 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404**(6777):490-493.

Paabo S. 1985. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* **314**(6012):644-645.

Paabo S, Poinar H, Serre D et al. 2004. Genetic analyses from ancient DNA. *Annual review of genetics* **38**:645-679.

Pakendorf B. 2014. Coevolution of languages and genes. *Curr Opin Genet Dev* **29C**:39-44.

Pallottino M. 1975. The Etruscans. Bloomington, IN: Indiana University Press.

Penny D, Steel M, Waddell PJ, and Hendy MD. 1995. Improved analyses of human mtDNA sequences support a recent African origin for Homo sapiens. *Mol Biol Evol* **12**(5):863-882.

Petraglia M, Korisettar R, Boivin N et al. 2007. Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. *Science* **317**(5834):114-116.

Piazza A, Cappello N, Olivetti E, and Rendine S. 1988. A genetic history of Italy. *Annals of human genetics* **52**(Pt 3):203-213.

Pickrell JK, and Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**(11):e1002967.

Poinar HN, Schwarz C, Qi J et al. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**(5759):392-394.

Price AL, Zaitlen NA, Reich D, and Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**(7):459-463.

Pritchard JK, Seielstad MT, Perez-Lezaun A, and Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**(12):1791-1798.

Prufer K, Racimo F, Patterson N et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**(7481):43-49.

Prugnolle F, Manica A, and Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**(5):R159-160.

Pugach I, Delfin F, Gunnarsdottir E et al. 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc Natl Acad Sci U S A* **110**(5):1803-1808.

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

Ralph P, and Coop G. 2013. The geography of recent genetic ancestry across Europe. *PLoS Biol* **11**(5):e1001555.

Ramachandran S, Deshpande O, Roseman CC et al. 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.

Ramakrishnan U, Hadly EA, and Mountain JL. 2005. Detecting past population bottlenecks using temporal genetic data. *Mol Ecol* **14**(10):2915-2922.

Rasmussen M, Anzick SL, Waters MR et al. 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**(7487):225-229.

Rasmussen M, Guo X, Wang Y et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**(6052):94-98.

Rasmussen M, Li Y, Lindgreen S et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**(7282):757-762.

Reich D, Green RE, Kircher M et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**(7327):1053-1060.

Reich D, Patterson N, Kircher M et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* **89**(4):516-528.

Reich D, Thangaraj K, Patterson N et al. 2009. Reconstructing Indian population history. *Nature* **461**(7263):489-494.

Relethford JH. 1985. Examination of the relationship between inbreeding and population size. *J Biosoc Sci* **17**(1):97-106.

Renfrew C. 2010. Archaeogenetics--towards a 'new synthesis'? *Curr Biol* **20**(4):R162-165.

Reyes-Centeno H, Ghirotto S, Detroit F et al. 2014. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc Natl Acad Sci U S A* **111**(20):7248-7253.

Rosenberg NA, Pritchard JK, Weber JL et al. 2002. Genetic structure of human populations. *Science* **298**(5602):2381-2385.

Sabeti PC, Varilly P, Fry B et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164):913-918.

Sanger F, Nicklen S, and Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12):5463-5467.

Scheinfeldt LB, and Tishkoff SA. 2013. Recent human adaptation: genomic approaches, interpretation and insights. *Nat Rev Genet* **14**(10):692-702.

Schmitz RW, Serre D, Bonani G et al. 2002. The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc Natl Acad Sci U S A* **99**(20):13342-13347.

Scholz CA, Johnson TC, Cohen AS et al. 2007. East African megadroughts between 135 and 75 thousand years ago and bearing on early-modern human origins. *Proc Natl Acad Sci U S A* **104**(42):16416-16421.

Schuster SC, Miller W, Ratan A et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**(7283):943-947.

Serre D, Langaney A, Chech M et al. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol* **2**(3):E57.

Sicoli MA, and Holton G. 2014. Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS One* **9**(3):e91722.

Skoglund P, Northoff BH, Shunkov MV et al. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc Natl Acad Sci U S A* **111**(6):2229-2234.

Smith CI, Chamberlain AT, Riley MS et al. 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. *J Hum Evol* **45**(3):203-217.

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

Sokal RR, Harding RM, and Oden NL. 1989a. Spatial patterns of human gene frequencies in Europe. *Am J Phys Anthropol* **80**(3):267-294.

Sokal RR, Jacquez GM, and Wooten MC. 1989b. Spatial autocorrelation analysis of migration and selection. *Genetics* **121**(4):845-855.

Soltis PS, Soltis DE, and Smiley CJ. 1992. An rbcL sequence from a Miocene Taxodium (bald cypress). *Proc Natl Acad Sci U S A* **89**(1):449-451.

Stone AC, Griffiths RC, Zegura SL, and Hammer MF. 2002. High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc Natl Acad Sci U S A* **99**(1):43-48.

Stringer C. 2002. Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci* **357**(1420):563-579.

Stringer CB, and Andrews P. 1988. Genetic and fossil evidence for the origin of modern humans. *Science* **239**(4845):1263-1268.

Tassi F, Ghirotto S, Caramelli D, and Barbujani G. 2013. Genetic evidence does not support an Etruscan origin in Anatolia. *Am J Phys Anthropol* **152**(1):11-18.

Tattersall I. 2009. Out of Africa: modern human origins special feature: human origins: out of Africa. *Proc Natl Acad Sci U S A* **106**(38):16018-16021.

Tenesa A, Navarro P, Hayes BJ et al. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**(4):520-526.

Thangaraj K, Chaubey G, Kivisild T et al. 2005. Reconstructing the origin of Andaman Islanders. *Science* **308**(5724):996.

The HUGO Pan-Asian SNP Consortium. 2009. Mapping Human Genetic Diversity in Asia. *Science* **326**(5959):1541-1545.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**(6968):789-796.

Thomas E, van Zonneveld M, Loo J et al. 2012. Present spatial diversity patterns of Theobroma cacao L. in the neotropics reflect genetic differentiation in pleistocene refugia followed by human-influenced dispersal. *PLoS One* **7**(10):e47676.

Thomson R, Pritchard JK, Shen P et al. 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci U S A* **97**(13):7360-7365.

Tishkoff SA, Reed FA, Friedlaender FR et al. 2009. The genetic structure and history of Africans and African Americans. *Science* **324**(5930):1035-1044.

Trinkaus E. 2005. Early Modern Humans. *Annual Review of Anthropology* **34**(1):207-230.

Turchi C, Buscemi L, Previdere C et al. 2008. Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. *Int J Legal Med* **122**(3):199-204.

Underhill PA, Shen P, Lin AA et al. 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**(3):358-361.

Vai S, Ghirotto S, Pilli E et al. 2015. Genealogical Relationships between Early Medieval and Modern Inhabitants of Piedmont. *PLoS One* **10**(1):e0116801.

Venter JC, Adams MD, Myers EW et al. 2001. The sequence of the human genome. *Science* **291**(5507):1304-1351.

Vernesi C, Caramelli D, Dupanloup I et al. 2004. The Etruscans: a population-genetic study. *Am J Hum Genet* **74**(4):694-704.

Vernot B, and Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**(6174):1017-1021.

von Bubnoff A. 2008. Next-Generation Sequencing: The Race Is On. *Cell* **132**(5):721-723.

Wall JD, Yang MA, Jay F et al. 2013. Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**(1):199-209.

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

Wheeler DA, Srinivasan M, Egholm M et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189):872-876.

Wollstein A, Lao O, Becker C et al. 2010. Demographic history of Oceania inferred from genome-wide data. *Curr Biol* **20**(22):1983-1992.

Wolpoff M. 1989. Multiregional evolution: the fossil alternative to Eden. In: Mellars P, and Stringer C, editors. The human revolution: behavioural and biological perspectives on the origins of modern humans Edinburgh: Edinburgh University Press. p 62-108

Wolpoff MH, Hawks J, and Caspari R. 2000. Multiregional, not multiple origins. *Am J Phys Anthropol* **112**(1):129-136.

Woodward SR, Weyand NJ, and Bunnell M. 1994. DNA sequence from Cretaceous period bone fragments. *Science* **266**(5188):1229-1232.

Wright S. 1931. Evolution in Mendelian Populations. *Genetics* **16**(2):97-159.

Wright S. 1950. Genetical structure of populations. *Nature* **166**(4215):247-249.

Xing J, Watkins WS, Shlien A et al. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* **96**(4):199-210.

Xing J, Watkins WS, Witherspoon DJ et al. 2009. Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res* **19**(5):815-825.

Yu N, Jensen-Seaman MI, Chemnick L et al. 2004. Nucleotide diversity in gorillas. *Genetics* **166**(3):1375-1383.

APPENDIX

Table A

Table A: 56 nominal parameters and their settings in 15 European languages. Each parameters is identified by a progressive number (in the first column) and by a combination of three capital letters (in the third column). The alternative parameter states are encoded as '+' o '-'. The symbol '0' encodes the neutralizing effect of implicational dependencies across parameters. The conditions which must hold for each parameter to be relevant are indicated after the name of the parameter itself. They are expressed in a Boolean form, either as simple values of other parameters, or as conjunctions (written ','), disjunction ('or'), or negation ('≈') thereof. The following columns represent 26 contemporary Indo-European languages belonging to the following subfamilies:

- *Romance*: Sicilian (Sic), Northern Calabrese (Cal; data from Verbicaro, Cosenza), Italian (It), Salentino (Sal; data from Cellino S.Marco, Brindisi), Spanish (Sp), French (Fr), Portuguese (Ptg), Rumanian (Rm);

- *Greek*: Bovese Greek (BoG; data from the area of Bova, Reggio Calabria), Salentino Greek (Gri; data from Calimera, Lecce), standard Modern Greek (Grk);

- *Germanic*: English (E), German (D), Danish (Da), Icelandic (Ice), Norwegian (Nor);

- *Slavic*: Bulgarian (Blg), Serbo-Croat (SC), Slovenian (Slo), Polish (Po), Russian (Rus);

- *Celtic*: Irish (Ir), Welsh (Wel);

- Indo-Iranian: Farsi (Far), Marathi (Ma), Hindi (Hi);

	TABLE A		Sic	Cal	It	Sal	Sp	Fr	Ptg	Rm	BoG	Gri	Grk	E	D	Da	Ice	Nor	Blg	SC	Slo	Po	Rus	Ir	Wel	Far	Ma	Hi	
1	± gramm. person	FGP	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1
2	± gramm. number +FGP	FGN	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	2
3	± gramm. gender +FGN	FGG	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	3
4	± NP over D +FGP	NOD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4
5	± feature spread to N +FGN	FSN	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	5
6	± numb. on N (BNs) +FSN	FNN	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	6
7	± gramm. partial def	DGP	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	-	-	-	7
8	± gramm. def +DGP	DGR	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	+	+	0	0	0	8
9	± strong person +FGP, +DGR, -NOD	NSD	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	0	0	0	0	-	-	0	0	0	9
10	± free null partitive Q +FNN	DPQ	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10
11	± gramm. dist. art. -FSN or -FNN or +DGR	DDA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	-	-	0	0	0	11
12	± def-checking N +DGR	DCN	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	+	+	0	0	0	0	-	-	0	0	0	12
13	± def spread to N +DCN, -NSD	DSN	0	0	0	0	0	0	0	0	0	0	0	0	0	-	+	+	0	0	0	0	0	0	0	0	0	0	13
14	± def on relatives +DGR	DOR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	-	-	0	0	0	14
15	± D-controlled infl. on N +FSN	DIN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15
16	± plural spread from cardinals +FSN	CPS	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	16
17	± gramm. boundedness	CGB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	17
18	± strong article +DGR, +FNN, -CGB	CGR	+	+	+	+	+	0	+	+	+	+	+	+	+	+	-	+	-	0	0	0	0	-	-	0	0	0	18
19	± bounded.-checking N +CGB	CCN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	-	-	19
20	± null-N-licensing art -FSN or-FNN or -DCN, +NOD or +NSD	DNN	-	-	-	-	+	-	+	0	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20
21	± structured APs	AST	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	21
22	± feature spread to struct. APs +FSN, + AST	FFS	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	22
23	± feature spread to pred. APs +FGN	FSP	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-	-	-	+	+	23
24	± D-controlled infl. on A -NSD, +FFS	ADI	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	0	0	0	0	-	0	0	0	0	24
25	± DP over relatives	ADR	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	25
26	± relative extrap. -ADR	AER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	26
27	± free reduced rel +AST	ARR	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	27
28	± N-raising with obl. pied-piping +AST	NPP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	28
29	± free Gen	GFR	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	-	-	-	-	+	+	+	+	29
30	± uniform Gen +GFR	GUN	-	-	-	-	-	-	-	0	0	0	0	-	-	-	-	-	-	-	0	0	0	0	-	-	-	-	30
31	± DP over free Gen +GFR, +ADR	GPR	+	+	+	+	+	+	+	0	0	0	0	+	+	+	+	+	+	+	0	0	0	0	+	+	+	0	31
32	± GenO ¬+GUN	GFO	-	+	-	-	-	-	-	-	+	+	+	-	+	-	+	-	-	-	+	+	+	+	+	+	-	-	32
33	± Gen-feature spread to N	GFS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33
34	± D checking poss. +DGR, +NSD or ¬+CGR	PDC	-	-	-	-	+	+	?	-	-	-	-	0	0	0	+	0	-	0	0	0	0	+	+	0	0	0	34
35	± adjectival poss.	APO	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	-	-	+	+	+	+	+	-	-	-	-	35
36	± post-affix poss. +DCN	PAP	0	0	0	0	0	0	0	+	0	0	0	0	0	-	+	+	+	0	0	0	0	0	0	0	0	0	36
37	± clitic poss.	PCL	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	37
38	± N-feat. spr. to pron. poss. +FFS or -AST, +PAP or +PCL	PHS	0	0	0	0	0	0	0	+	-	-	-	0	0	0	+	+	-	0	0	0	0	0	0	0	0	0	38
39	± N-feature spread to free Gen +FFS, +GFR, ¬-PHS	GSP	-	-	-	-	-	-	-	0	0	0	0	0	-	-	-	-	-	0	0	0	0	0	-	0	0	+	39
40	± adjectival Gen +APO	AGE	-	-	-	-	-	-	-	0	0	0	0	0	-	0	0	0	+	+	+	-	+	0	0	0	0	0	40
41	± Poss°-checking N -GFS	GCN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	41
42	± Strong partial Locality ¬+CGR, -FSN or +FNN	TPL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	+	+	+	+	+	-	-	+	+	+	42
43	± Strong Locality ¬-TPL	TSL	+	+	+	+	-	+	+	-	+	+	-	+	+	+	+	+	+	+	+	+	+	0	0	+	+	+	43
44	± D Checking Dem ¬-TPL, -FSN or +DGR	TDC	+	+	+	+	+	+	+	+	?	+	-	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	44
45	± N over cardinals	NOC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45
46	± N over ordinals -NOC	NOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	46
47	± N over M1 As -NOO, -NPP	NM1	+	+	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	0	-	-	47
48	± N over M2 As -NM1	NM2	0	0	+	0	+	+	+	+	0	0	-	-	-	-	-	-	-	-	-	-	-	0	0	0	-	-	48
49	± N over As -NM2	NOA	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-	-	0	0	0	-	-	49
50	± N over GenO ¬-GFO, -NOA or -AST	NGO	0	0	0	0	0	0	0	0	0	0	+	0	+	0	+	0	0	+	+	+	+	0	0	0	0	0	50
51	± N over ext. arg. -NGO or (-GFO, -NOA or -AST)	NOE	0	0	0	0	0	0	0	0	0	0	0	-	0	-	0	-	+	0	0	0	0	0	0	0	+	+	51
52	± free MOD -NOA	AFM	0	0	0	0	0	0	0	0	0	0	+	-	-	-	-	-	-	-	-	-	-	0	0	0	-	-	52
53	± class MOD -AFM	ACM	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-	+	+	0	0	0	-	53
54	± def on APs +DGP, *+postnom. APs*	DOA	-	-	-	-	-	-	-	-	+	-	+	-	0	0	0	0	0	0	0	0	0	0	-	-	0	0	54
55	± gramm. AP marker *+postnom. APs*	AMO	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	-	-	-	-	+	0	0	55
56	± Cons. Pr. (-NM1,+A-Cpl) or (+NPP or ¬-NM2, +Cpl-A)	ACP	0	0	+	0	+	+	+	+	0	0	-	+	+	+	+	+	-	-	-	-	-	0	0	-	0	0	56
			Sic	Cal	It	Sal	Sp	Fr	Ptg	Rm	BoG	Gri	Grk	E	D	Da	Ice	Nor	Blg	SC	Slo	Po	Rus	Ir	Wel	Far	Ma	Hi	

PAPERS

PAPER I: Nine things to remember about human genome diversity.

TISSUE ANTIGENS
IMMUNE RESPONSE GENETICS

Tissue Antigens ISSN 0001-2815

REVIEW ARTICLE

Nine things to remember about human genome diversity

G. Barbujani, S. Ghirotto & F. Tassi

Department of Life Sciences and Biotechnologies, University of Ferrara, Ferrara, Italy

Key words
admixture; F_{ST}; gene flow; genetic drift; human origins; human races; pharmacogenomics; population structure

Correspondence
Guido Barbujani
Dipartimento di Scienze della Vita e Biotecnologie
Università di Ferrara
Via Borsari 46
I-44121 Ferrara
Italy
Tel: +39 0532 455312
Fax: +39 0532 249761
e-mail: g.barbujani@unife.it

doi: 10.1111/tan.12165

Abstract

Understanding how and why humans are biologically different is indispensable to get oriented in the ever-growing body of genomic data. Here we discuss the evidence based on which we can confidently state that humans are the least genetically variable primate, both when individuals and when populations are compared, and that each individual genome can be regarded as a mosaic of fragments of different origins. Each population is somewhat different from any other population, and there are geographical patterns in that variation. These patterns clearly indicate an African origin for our species, and keep a record of the main demographic changes accompanying the peopling of the whole planet. However, only a minimal fraction of alleles, and a small fraction of combinations of alleles along the chromosome, is restricted to a single geographical region (and even less so to a single population), and diversity between members of the same population is very large. The small genomic differences between populations and the extensive allele sharing across continents explain why historical attempts to identify, once and for good, major biological groups in humans have always failed. Nevertheless, racial categorization is all but gone, especially in clinical studies. We argue that racial labels may not only obscure important differences between patients but also that they have become positively useless now that cheap and reliable methods for genotyping are making it possible to pursue the development of truly personalized medicine.

Introduction

Genetics is a fast-changing field. The first sequence of a whole human genome was completed in 2003, thanks to the efforts of 2800 scientists who worked for 13 years (1). The second and the third complete sequences took 4 years and, respectively, 31 (2) and 27 scientists (3). In 2010, only 2.3 days were necessary for sequencing a whole genome (4) while the costs had dramatically dropped, from 2.7 million to a few thousand dollars. As a result, by 2012 the number of individual genomes completely typed has exceeded 1000 (4), still growing very fast. In parallel, functional elements in the genome have been systematically mapped (5) providing new, precious insights into the processes of gene regulation and gene-to-gene interaction.

Many scientists believed or hoped that the availability of this enormous mass of data would immediately improve our ability to predict phenotypes and design new therapies. Unfortunately, this has not happened yet. As a matter of fact, we still fail to understand the causes of most common diseases, and we only have a vague idea of the genetic bases of normal variation for non-pathological traits. In a sense, this is hardly surprising. Indeed, many diseases have complex causes. It is out of discussion that genetic factors play a crucial role in their onset, but there is still a substantial gap between what we can currently do, estimating the genetic predisposition to develop a disease, and what we would like to do, i.e. deciphering the complex network of interactions between genetic and non-genetic predisposing factors (exposure to chemicals, diet, lifestyle, etc.), thus coming up with accurate predictions. In the meantime, however, a much clearer picture of human diversity has emerged, only partly confirming previous ideas based on the analysis of small portions of the genome. The new genomic data have cast a different light on both normal and pathological variation, and hence understanding exactly what we know about human genome diversity seems indispensable for a rational planning of new clinical studies, for interpreting their results, and for raising public awareness of science.

In this review, we discuss nine key points about human genome variation. We present results emerging from the study of different genetic markers and complete genome sequences, emphasizing the demographic features of human evolution that can explain the observed patterns. We also stress the importance of a proper use of this information in clinical practice, with a particular focus on racial categorizations as a

poor predictor of human biological diversity and its potentially negative effects upon clinical research.

Individual genetic diversity among humans is the lowest of all primates

The comparison of genetic variation in great apes and humans is crucial to deeply investigate the origins and the evolution of our species, not to mention the fact that it can help show the molecular bases of common human diseases (6). Complete genome sequences from primates, now available (6–8), have confirmed that we are evolutionarily very close to them and have provided us with quantitative measures of that closeness. We share with the genome of our closest living relatives (chimpanzee) more than 98% of the nucleotides, over an estimated haploid genome length close to 3 billion nucleotides. Thirty-five million single-nucleotide changes (and about 5 million insertion/deletion events) have been identified, corresponding to a mean rate of single-nucleotide substitutions of 1.23% between copies of the human and chimpanzee genome. Most of these changes, 1.06% over 1.23%, appear to be fixed between species, meaning that at these sites all chimpanzees share one allele, which is different from the one shared by all humans. However, the main genetic differences between humans and other Primates do not seem to depend on point mutations, but on gain or loss of entire genes (9) that have undergone copy-number changes large enough to suggest the influence of natural selection. These genomic regions are likely to be responsible for the key phenotypic changes in morphology, physiology, and behavioral complexity between humans and chimpanzees.

What also emerged from this picture is that humans are genetically less variable than any other primate. At the beginning of 2013, 65 million nucleotide sites have been shown to vary in humans (10), and this number is steadily increasing, as more complete genomes are being sequenced. Yet, a vast majority of these polymorphisms has a very limited distribution across the species. By contrast, much larger differences are observed between pairs of orangutans, gorillas, chimpanzees, and bonobos (11). The study of the genetic relationships among three geographically close populations of common chimpanzees has shown a level of differentiation higher than that found among continental human populations (12), and the global genetic diversity of the orangutan species has been found to be roughly twice the diversity of modern humans (7), although both chimpanzees and orangutans occupy a far more restricted geographical range than we do. Further studies will doubtless expand the list of polymorphic sites, but on average a pair of random humans is expected to share 999 of 1000 nucleotides (13, 14). Quite surprisingly, as we shall see in the following section, this average similarity reflects only in part the geographic distance between the subjects being compared.

Genetic diversity between human populations is a small fraction of the species' diversity

Differences among populations are often summarized by F_{ST}, that is, the proportion of the global genetic diversity due to allele-frequency differences among populations (15). F_{ST} ranges from 0 (when allele frequencies are identical in the two populations) to 1 (when different alleles are fixed in the two populations) (for a review see Ref. 16).

Depending on the markers chosen, estimates of F_{ST} among major geographical human groups range from 0.05 to 0.13 (14). These figures mean that not only is the overall human genetic diversity the lowest in all primates but also the differences between human populations account for a smaller fraction of that diversity than in any other primate, i.e. between 5 and 13% of the species' genetic variance (17, 18). The remaining 90% or so represents the average difference between members of the same population. Different loci differ in their levels of diversity and so, for example, in 377 autosomal microsatellites (or STR, Short Tandem Repeat, markers), the differences among major groups constitute only 3–5% of the total genetic variance (19). By contrast, considering single-nucleotide polymorphisms (SNPs), the differences between continents can reach 13% (20). Recent global estimates over the whole genome from the 1000 Genomes Project suggest that the human F_{ST} could even be lower. Indeed, in analyses considering about 15 millions SNPs, 6 millions of them representing newly discovered variants, the mean values of $F_{ST} = 0.071$ between Europeans and Africans, $F_{ST} = 0.083$ between Africans and Asians, and $F_{ST} = 0.052$ between Asians and Europeans (4). This level of differentiation is less than one-third of what is observed in gorilla, $F_{ST} = 0.38$ (21) and chimpanzee, $F_{ST} = 0.32$ (6). The fact that human populations are more closely related than populations of the other primates suggests that in human evolution processes such as gene flow and admixture had a comparatively greater role than long-term isolation and differentiation.

In each individual, chromosomes are mosaics of DNA traits of different origins

When a mutation generates it, a new allele is in complete linkage disequilibrium with all the alleles that happen to lay on the same chromosome; with time, levels of linkage disequilibrium are reduced by recombination, but increase as a consequence of phenomena such as drift and admixture. The analysis of millions of SNPs over the genome has confirmed these theoretical expectations. Indeed, the combinations of alleles along the chromosome, or haplotypes, typically show blocks, namely regions of several kilobases in linkage disequilibrium, within which recombination has seldom or never occurred. The list of observed variants for every block represent the haplotype map of the human genome. Blocks vary in size across individuals and populations, depending on the relative historical weight of recombination (reducing their sizes) and drift or admixture

(increasing their size). Indeed, one of the clearest pieces of evidence supporting an African origin of humankind is the larger block size in Africans than in Europeans and Asians, a likely consequence of founder effects as small groups of Africans dispersed in the other continents (22). Information on the location and size of haplotype blocks is important for investigating the genetics of common diseases.

To understand how our genealogical history has shaped us, it is thus necessary to regard each genome as a mosaic of haplotype blocks, each with its own origin and history, brought together in the same cell by sexual reproduction. Although, as we shall see, very few genome fragments are found in a single continent (and even less so in single populations), the history of each such fragment can be inferred by comparing variation in different individuals and sometimes in different species (23). A spectacular illustration of this concept, and of how a single individual's genome may record a complex history of gene flow, is in Ref. 24.

The length of tracts assigned to distinct ancestries in an individual may be especially informative about the historical pattern of migration between populations, as well as about the time and mode of migration from one ancestral population into another. When two individuals from different parental populations mate, the first generation offspring inherits exactly one chromosome from each parental population. In subsequent generations, though, recombination events in admixed individuals generate mosaic chromosomes, essentially composed of segments having different ancestries. Intuitively, more recent admixture gives rise to longer ancestry blocks than older admixture. Thus, an excess of long blocks would indicate a recent increase in migration rate, while the opposite pattern would suggest recently reduced gene flow (25). This way, using a set of recently developed methods (26–31), it has become possible to infer with some accuracy the ancestry of many regions in individual genomes. By and large, these analyses suggest a very widespread impact of genetic admixture, a likely consequence of the absence of strong mating barriers between populations.

Allele sharing is the rule across continents

Sharing of polymorphisms across the world is extensive in humans. Jakobsson et al. (32) analyzed 525,910 SNPs and 396 CNV sites in 29 populations of five continents. They observed that 81.2% of the SNPs were cosmopolitan, i.e. occur, at different frequencies, in all continents. Less than 1% were specific to a single continent, and 0.06% were observed only in Eurasia. Combining the alleles in haplotypes, the fraction of cosmopolitan variants decreased to 12.4%, whereas 18% of the haplotypes appeared to be exclusively African. However, continent-specific haplotypes in the other four continents summed up to just 11% of the total. This small fraction of variants restricted to a single continent is in agreement with the results of a previous study of haplotype blocks. Gabriel et al. (22) sequenced 1.5 million bases of DNA in African, Asian, and European individuals: less than 2% of haplotype blocks appeared restricted to Asia, 2% appeared restricted to Europe, 25% were African specific, and the rest were shared among continents, with more than 50% occurring worldwide. Thus, with few exceptions, from the genomic standpoint, each of us can have either typically African, or generically human, features.

Several studies confirmed these results (19, 33, 34) and concurred in indicating that extensive allele and haplotype sharing across continents is the rule, not the exception, with variation within Africa exceeding that among other continents (33, 35–37). Classical population-genetics theory shows that these patterns of variation characterize species with weak or no reproductive barriers separating individuals in different groups (38). In short, it looks as though the rule for human populations is not to have independently evolved, but rather to have maintained connections through extensive gene flow. As a consequence, and as proposed by Frank Livingstone (39) on the basis of the extremely scanty data available in 1960s, genetic variation between populations tends to be continuous, without clear boundaries.

There is a clear geographic structure in human genome diversity: any population can be shown to somehow differ from any other population

Although most human variation is found within populations, the proportion that lies between continents (summarized by F_{ST}) is still significantly greater than zero. Thus, it makes sense to ask whether individuals can be assigned with good statistical confidence to their population or continent of origin on the basis of their genotype. The answer may be yes; there is indeed a relationship between patterns of genetic variation and geographical ancestry. Several recent studies have used a likelihood-based approach, implemented in the software package structure (28), to identify genetic clusters and evaluate for every individual genotype the membership to each of the inferred clusters. Rosenberg et al. (19) showed that 52 globally distributed populations can be clustered into six groups, five of which correspond to major geographic regions and one to the Kalash of Pakistan. Similar results were obtained by Li et al. (40) analyzing 650,000 common SNPs in the same populations.

However, further attempts to identify major human groups by clustering genotypes have yielded inconsistent results. Different numbers of groups and different distributions of genotypes within such groups, were observed when different datasets were analyzed (30, 41–44). The inconsistencies in these results reflect a well-known feature of human diversity, that is, different genetic polymorphism are distributed over the world in a discordant manner (44). This variation reflects in part response to different environmental pressures (Refs 45–47) and in part the different impact of demographic history

upon different genomic regions (Refs 39, 48), but in both cases leads to differences in the apparent population clusterings. It comes as no surprise, then, that if we look back at the many racial catalogs compiled since the 17th century, and at more recent genomic analyses (compare Refs 19, 32, 34, Figure 1), the only point they seem to have in common is that each of them contradicts all the others (49, 50).

But within-population diversity is very large

Above and beyond the discordant geographic patterns of population diversity, a second factor makes it difficult, or impossible, to define, once and for good, the main genetic clusters of humankind; this factor is the high level of within-population diversity. Several studies show that the

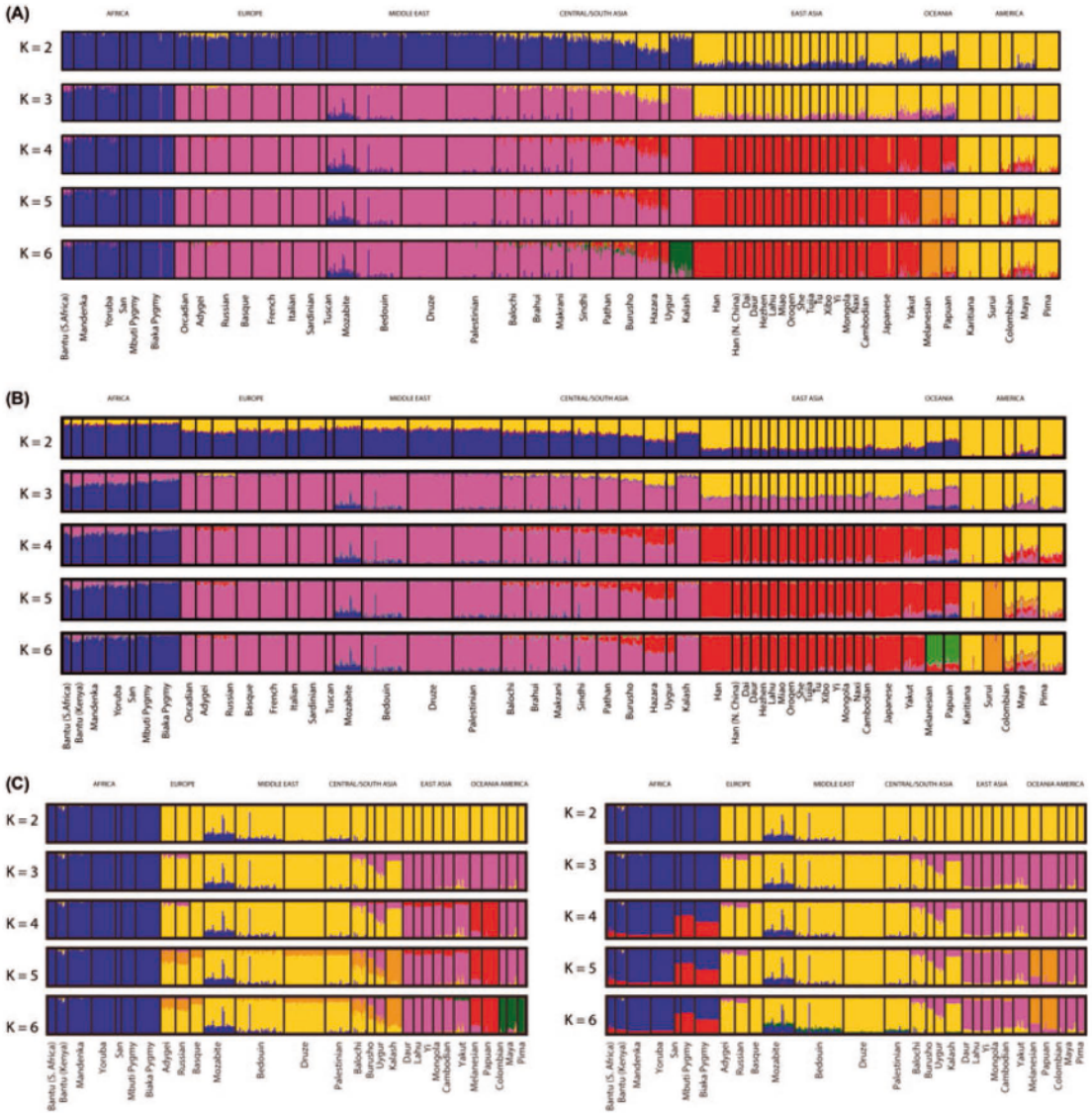


Figure 1 Comparison of four analyses of the global human population structure. Each individual genotype is represented by a thin vertical line partitioned into colored components representing inferred membership in K genetic clusters. Black lines separate individuals of different populations. Populations are labeled below each panel, with their regional affiliations above it. The analyses are based on different markers and samples: (A) 377 Short Tandem Repeat (STR) in 1056 individuals from 52 populations (19); (B) 993 STR and insertion/deletion polymorphisms in 1048 individuals from 53 populations (34); (C) 525,910 single-nucleotide polymorphisms (SNPs) (left panel) and 396 CNV sites (right panel) in 485 individuals from 29 populations (32).

largest fraction of genetic diversity in our species is due to differences between individuals of the same population, rather than to differences between populations. The pioneer analysis of this topic remains Lewontin's (18) study of protein polymorphisms from 17 loci in worldwide populations. In that study, 85.4% of total human diversity appeared due to individual variation within populations, with barely 6.3% representing the average difference between major population groups. This finding was commented with disbelief by some (51), but has been consistently replicated in protein and then DNA studies (52). The results of the analysis of 109 nuclear autosomal restriction fragment length polymorphism (RFLP) and microsatellite loci were extremely similar to Lewontin's: the variance component within population was 84.4% (17). The observation that each population harbors a large share of the global human diversity, replicated in ever-larger studies of nuclear data (19, 39, 53) means that random members of different continents tend differ, on average, just slightly more than members of the same community. On the basis of a large assemblage of microsatellite markers, Rosenberg showed that the mean proportion of alleles differing in random pairs of individuals worldwide (0.651) exceeds by 5% the mean difference for pairs from the same continent (0.618) (34).

Today, developments in DNA sequencing technology allow us to compare completely sequenced genomes. Ahn et al. (54) observed that two US scientists of European origin, namely James Watson (11) and Craig Venter (2), share fewer SNPs (461,000) than either of them shares with a Korean scientist, Seong-Jin Kim (569,000 and 481,000, respectively) (Figure 2). Of course, this does not mean that, on average, people of European origin are genetically closer to Asians than to other Europeans. However, it does show that patterns of genetic resemblance are far more complicated than any scheme of racial classification can account for. On the basis of the subjects' physical aspect, a physician would consider Venter's DNA, and not Kim's, a better approximation to Watson's DNA. Despite ideological statements to the contrary (55, 56) racial labels are positively misleading in medicine, and wherever one is to infer individual genome characteristics.

Differences between Africans are greater than between people of different continents

From a genetic standpoint, Africa is not just another continent. Paleontological data clearly indicate that anatomically modern *Homo sapiens* emerged there (57), and genetic evidence corroborates this view, showing that compared with populations from other continents, African populations have the highest level of genetic diversity at most loci (reviewed in Ref. 58). The analysis of high-quality genotypes at 525,910 SNPs in a worldwide sample of 29 populations, revealed that Africa shows the largest number of unique alleles, i.e. alleles specific to a single continent, and that in many cases the alleles found out of Africa represent a subset of the African alleles (32). In

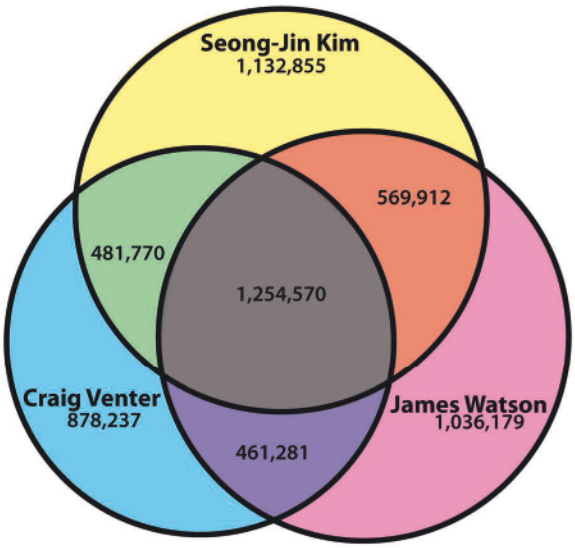


Figure 2 Venn diagram of single-nucleotide polymorphism (SNP) alleles in Seong-Jin Kim's, Craig Venter's and James Watson's genomes. Figures within the intersections are numbers of shared alleles between individuals. Modified and redrawn from Ref. 54.

the first survey of the 1000 Genomes project, populations with African ancestry contributed the largest number of variants and contained the highest fraction of novel variants roughly twice as many as in the populations of European ancestry (4). In a study of 117 megabases (Mb) of exomic sequences, the average rate of nucleotide substitutions between two hunter-gatherers from the Kalahari Desert was 1.2 per kb, compared to an average of 1.0 per kb between European and Asian individuals (35).

Gene diversity declines as a function of distance from Africa

Several measures of genetic diversity are patterned in space, with a maximum in Africa and decreasing values, respectively, in Eurasia, the Americas, and Oceania (40, 48, 59). On the contrary, linkage disequilibrium is minimal in African populations, and increases at increasing distances from there (32, 60), and the average length of haplotype blocks has a minimum in Africa around 10 kb and is close to 50 kb in Eurasia (22). All these findings are consistent with the expected consequences of an expansion of our species outside Africa, by means of dispersals of rather small groups of founders that then rapidly populated all the world (48, 61). The most likely origin of these migrational processes is East Africa (61, 62), and in fact, the geographic distance from East Africa along probable colonization routes is an excellent predictor of the genetic diversity of human populations (59). Because only a small part of the African population migrated

out of Africa, only part of Africa's genetic variation moved with them, which explains why genetic variation found in non-African populations can largely be regarded as a subsample of African variation (58, 60). Because the other continents were peopled at a relatively recent time, only few mutations are geographically restricted to these continents, i.e. those mutations that arose after the human expansion out of Africa (Figure 3).

Racial pharmacogenomics is not a step toward individual pharmacogenomics

Despite all we have seen so far, the belief that race is a reasonably good descriptor of human biological diversity is all but gone, and so is the idea that a racial categorization of patients is part of a good clinical and scientific practice. On 3 May 2013, a PubMed search using the terms 'human races' yielded 141,245 items, nearly all of them from medical journals, and this number increases at a rate of more than 20 articles per day.

The basic tenet underlying these studies is that racial categorization, although occasionally inaccurate, remains indispensable for assessing risk factors in medical and pharmaceutical research. According to Gonzalez-Burchard et al. (55): (a) reproductive barriers and endogamy have given rise to a structured human population; (b) although these barriers are mainly geographic and social, they caused genetic divergence of racial and ethnic groups; (c) as a consequence, the human population tree has major branches corresponding to five major racial groups, as defined in the US 2000 census, with secondary branches associated with indigenous populations. For all these reasons, ignoring racial background would create disadvantages to the very people that this approach means to protect (55).

Statement (a) is obviously correct; mating is not random across the whole human species, and so genetic differences exist between populations. What has proved wrong is the idea that these differences subdivide humankind in a set of recognizable genetic clusters (statement b), and it seems, at best, naïve to maintain that these clusters correspond to those in the US census classification (statement c), if only because the US census classification changes every decade. Indeed, between 2000 and 2010, races recognized in the US census have grown from 5 to infinite (15 plus 'other races'), with Hispanics or Latinos classified in a 16th group, defined as 'origin' rather than 'race' for reasons that escape us. Clearly, folk concepts change following social changes

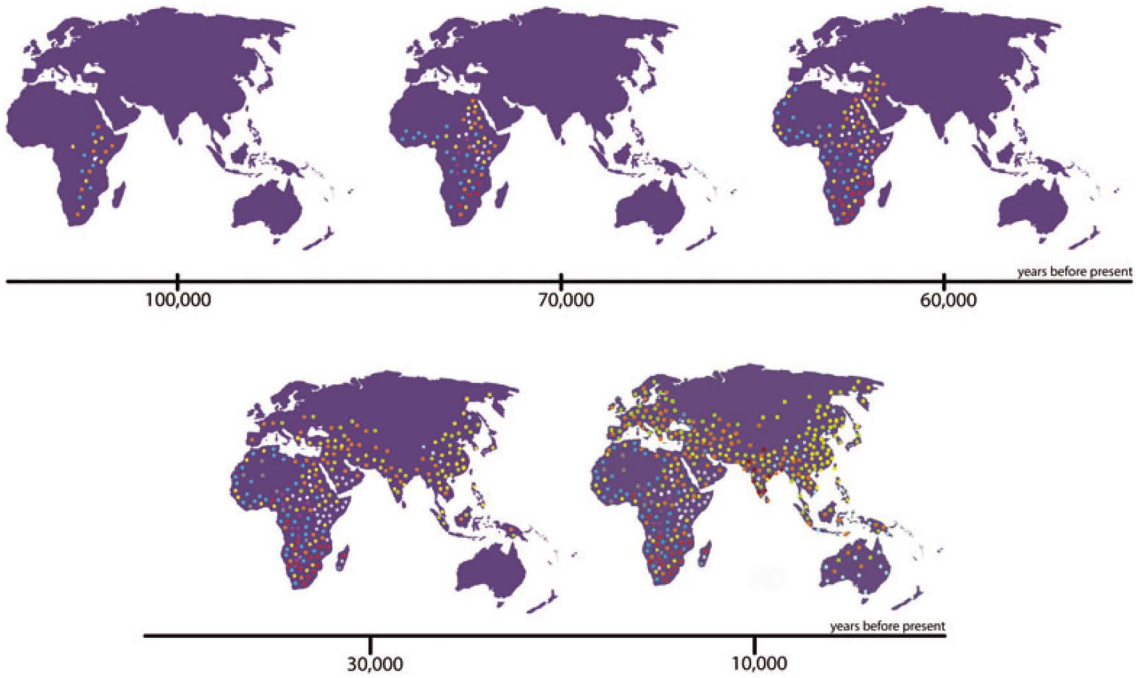


Figure 3 A highly schematic view of the evolution of human biodiversity in the last 100,000 years. Dots of different colors represent different genotypes, the distribution of which roughly corresponds to archeological evidence on human occupation of different regions. Dots of new colors appear in the maps in the course of time (e.g. red and violet in Africa at 70,000 BP, Burgundy in India at 10,000 BP), representing the effect of mutation. Because only part of the African alleles (yellow, orange and light green dots) are carried into Eurasia by dispersing Africans from 60,000 years bp (Refs 48 and 61), diversity in modern Eurasian populations is largely a subset of African diversity. Modified and redrawn from Ref. 14.

that are unrelated with biology, and hence are unsuitable for scientific purposes. One may add that differences other than those recognized in the United States may be relevant to people of different cultures (63); just as an example, in apartheid South Africa Japanese were classified as white and Chinese as colored (see Ref. 50 for more examples).

All this notwithstanding, the scientific debate on race still becomes heated on occasions. One was the patenting of the first, and so far only, drug approved by the US Food and Drug Administration for a specific racial group, BiDil. Taylor et al. (64) found that adding isosorbide dinitrate and hydralazine to standard therapy for heart failure increases survival among black patients with advanced heart failure. Critics remarked that BiDil was tested only in self-defined black patients, comparing treatments with the drug and with the placebo, but not in other groups (65); that the degree of correspondence between self-identified race and any components of the patients' genome was unknown (66, 67); and that social and economic factors probably contributing to high blood pressure were overlooked as a result of oversimplified assumptions about the existence of racial differences (65, 68). By contrast, supporters of BiDil stressed that, no matter how inaccurate was the science behind it, BiDil did save lives (69) and, more recently, that racial medicine might be a useful first step toward personalized medicine (70). Both sides accused each other to be blinded by social or political considerations that have nothing to do with science.

As a matter of fact, patenting of BiDil resulted in the resurrection of the claim that humans are naturally subdivided in biological races (71), in many cases supported by improper analyses of data available at the HapMap web site (72). Generated as part of the International HapMap Project (73), this website contains information on four populations (Nigerian Yoruba, Americans of European origin from Utah, Chinese from Beijing and Japanese from Tokyo), chosen because their well-known differences would facilitate discovery of new polymorphisms. Certainly, the HapMap samples do not provide, and are not meant to provide, a faithful description of human genome variation, but this detail, by no means secondary, was often overlooked. As a consequence, many studies based on HapMap data concluded that there are differences between Africans, Asians, and Europeans, (to nobody's surprise), but then mistook these results as evidence that indeed there are three distinct genomic clusters in the human species (see, among many examples, (74–77)). As we have seen ((32)) that is simply not true.

Still, in clinical as well as in other kinds of studies on humans, we need names to define populations and subjects. Lee et al. (78) proposed a set of guidelines on the usage of terms referring, explicitly or implicitly, to racial or ethnic categories. After stating that there is no scientific evidence supporting a biological subdivision of humans in distinct racial or ethnic groups, they urged researchers to describe how individuals were assigned category labels and to explain why samples with such labels were included in the study. They also recommended to abandon the use of race as a proxy for biological similarity, to focus on the individual rather than the group, and to avoid deterministic connections between genes and phenotype, especially when communicating to the broad non-specialist public. However, only seldom were these wise recommendations put in practice. Actually, in a large analysis of medical papers published thereafter, Ali-Khan et al. (79) found that no authors using categories such as 'race', 'ethnicity' or 'ancestry' cared to discuss the meaning of these concepts in the studied context.

Recent technical progress has dramatically reduced genotyping costs, making it possible to obtain cheap and extensive information on individual genotypes. In the future, this large amount of genetic information will likely make it possible to target drugs on specific biomarkers, so that individuals who can benefit from treatment will be identified unambiguously through their genotype, rather than through biologically inaccurate and often highly subjective racial or ethnic definitions. As for the present, Ng et al. (80) examined six drug-metabolizing genes in J. Craig Venter's and James Watson's complete genome sequences. Although both subjects identify themselves as Caucasians, they show a set of differences of clinical relevance at loci involved in drug metabolism. Venter has two fully functional **1A* alleles at the *CYP2D6* locus, and an extensive metabolizer phenotype for β-Blockers, antiarrhythmics, antipsychotics and some antidepressants; conversely, Watson is homozygous for the *CYP2D6*10* allele (common in East Asian populations, but not among Europeans), and has a decreased metabolizing activity for the same class of drugs. Doctors would not guess this and other differences by simply looking at the subjects' physical aspect. Ng et al. (80) concluded that to attain truly personalized medicine, the scientific community must leave behind simplistic race-based approaches, and look instead for the genetic and environmental factors contributing to individual drug reactions. Far from being a necessary step toward personalized medicine, racial medicine is clearly showing, on top of its long-known lack of theoretical bases, its practical irrelevance.

Conclusions and future outlooks

In clinical as well as in other kinds of studies, we need names to define populations and subjects. Therefore, the question is not whether people should or could be categorized, but how to do it. From a social standpoint, the word race is so loaded with social and political implications that avoiding it seems just reasonable. However, from the scientific standpoint, the problem is not to replace it with a more elegant synonym. Whatever term one uses to define a group of people, be it population, ethnic group, or even race, both the authors and the readers must understand that there is no deterministic connection between being part of such groups and carrying a

certain genotype or phenotype. Races are a component of our psychological and social world, and as such their importance should not be dismissed, but are scientifically ambiguous to say the least, and in scientific communication ambiguities should be kept to a minimum.

To reduce the possibility of misunderstandings, Lee et al. (78) proposed a set of guidelines on the usage of terms referring, explicitly or implicitly, to racial or ethnic categories. After stating that scientific evidence does not support a biological subdivision of humans in distinct racial or ethnic groups, they urged researchers to describe how individuals were assigned category labels and to explain why samples with such labels were included in the study. They also recommended to abandon the use of race as a proxy for biological similarity, to focus on the individual rather than the group, and to avoid deterministic connections between genes and phenotype, especially when communicating to the broad non-specialist public. However, only seldom are these wise recommendations put in practice. Actually, in a large analysis of medical papers published thereafter, Ali-Khan et al. (79) found that no authors using categories such as 'race', 'ethnicity' or 'ancestry' cared to discuss the meaning of these concepts in the studied context.

Despite all these problems, there is no doubt that recent genomic research has spectacularly improved our understanding of how humans differ, and of the demographic processes that generated human diversity. However, the attempt to convert that basic knowledge into clinical applications has been less successful. Genetics developed as a science in which data were scanty and hard to produce, and sophisticated methods had to be devised to draw inferences from the limited body of empirical evidence. Thanks to the new sequencing technologies, data have been generated on a previously unimaginable scale, but this has somewhat reversed the problem; what we seem to miss now is an intellectual framework allowing us to make complete sense of this enormous mass of information. Genome-wide association studies have shown that genetic differences account for a substantial fraction of variation among individuals, for both normal and pathological traits; we have learned that common variants predispose to, but not necessarily cause, common disease; we know less about the possible effect of rare variants, which need be better investigated, but are also difficult to recognize. However, so far only seldom has all this resulted in substantial clinical advance (81). We often conclude our papers and our talks claiming that we need more data, but it is not clear exactly what could be achieved by further expanding datasets already including thousands of cases and controls. Rather, it seems that now we need better ideas on how genetic variants and factors in the environment interact in causing the onset of disease. Only by shifting from the identification of polymorphisms associated with increased or decreased disease risk to the development of predictive models, which could then be tested against the data, genetic studies will be able to produce progress in disease treatment. For that purpose, a deep understanding of patterns of genome diversity is a necessary precondition, but just a precondition.

Acknowledgments

The research leading to this article has received funding from the European Research Council, under the European Union's 7th Framework Programme (FP7/2007-2013)/ERC Grant Agreement N 295733 (Langelin project).

Conflict of interests

The authors have declared no conflicting interests.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003: **422**: 835–47.
2. Levy S, Sutton G, Ng PC et al. The diploid genome sequence of an individual human. *PLoS Biol* 2007: **5**: e254.
3. Bentley DR, Balasubramanian S, Swerdlow HP et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008: **456**: 53–9.
4. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010: **467**: 1061–73.
5. Dunham I, Kundaje A, Aldred SF et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012: **489**: 57–74.
6. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005: **437**: 69–87.
7. Locke DP, Hillier LW, Warren WC et al. Comparative and demographic analysis of orang-utan genomes. *Nature* 2011: **469**: 529–33.
8. Scally A, Dutheil JY, Hillier LW et al. Insights into hominid evolution from the gorilla genome sequence. *Nature* 2012: **483**: 169–75.
9. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics* 2007: **177**: 1941–9.
10. Lachance J, Vernot B, Elbers CC et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 2012: **150**: 457–69.
11. Kaessmann H, Wiebe V, Weiss G, Paabo S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet* 2001: **27**: 155–6.
12. Bowden R, MacFie TS, Myers S et al. Genomic tools for evolution and conservation in the chimpanzee: pan troglodytes ellioti is a genetically distinct population. *PLoS Genet* 2012: **8**: e1002504.
13. Wheeler DA, Srinivasan M, Egholm M et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008: **452**: 872–6.
14. Barbujani G, Colonna V. Human genome diversity: frequently asked questions. *Trends Genet* 2010: **26**: 285–95.
15. Wright S. Genetical structure of populations. *Nature* 1950: **166**: 247–9.

16. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 2009: **10**: 639–50.
17. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* 1997: **94**: 4516–9.
18. Lewontin R. The apportionment of human diversity. *Evol Biol* 1972: **6**: 381–98. New York: Appleton-Century-Crofts.
19. Rosenberg NA, Pritchard JK, Weber JL et al. Genetic structure of human populations. *Science* 2002: **298**: 2381–5.
20. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* 2005: **15**: 1468–76.
21. Stone AC, Griffiths RC, Zegura SL, Hammer MF. High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc Natl Acad Sci U S A* 2002: **99**: 43–8.
22. Gabriel SB, Schaffner SF, Nguyen H et al. The structure of haplotype blocks in the human genome. *Science* 2002: **296**: 2225–9.
23. Paabo S. The mosaic that is our genome. *Nature* 2003: **421**: 409–12.
24. Henn BM, Botigue LR, Gravel S et al. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* 2012: **8**: e1002397.
25. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003: **164**: 1567–87.
26. Patterson N, Hattangadi N, Lane B et al. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 2004: **74**: 979–1000.
27. Price AL, Helgason A, Palsson S et al. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet* 2009: **5**: e1000505.
28. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000: **155**: 945–59.
29. Sankararaman S, Kimmel G, Halperin E, Jordan MI. On the inference of ancestries in admixed populations. *Genome Res* 2008: **18**: 668–75.
30. Tang H, Quertermous T, Rodriguez B et al. Genetic structure, self-identified race/ethnicity, and confounding in case–control association studies. *Am J Hum Genet* 2005: **76**: 268–75.
31. Zhu X, Zhang S, Tang H, Cooper R. A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet* 2006: **120**: 431–45.
32. Jakobsson M, Scholz SW, Scheet P et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008: **451**: 998–1003.
33. Hinds DA, Stuve LL, Nilsen GB et al. Whole-genome patterns of common DNA variation in three human populations. *Science* 2005: **307**: 1072–9.
34. Rosenberg NA. A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum Biol* 2011: **83**: 659–84.
35. Schuster SC, Miller W, Ratan A et al. Complete Khoisan and Bantu genomes from southern Africa. *Nature* 2010: **463**: 943–7.
36. Yu N, Chen FC, Ota S et al. Larger genetic differences within africans than between Africans and Eurasians. *Genetics* 2002: **161**: 269–74.
37. Zietkiewicz E, Yotova V, Gehl D et al. Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. *Am J Hum Genet* 2003: **73**: 994–1015.
38. Wright S. Isolation by distance. *Genetics* 1943: **28**: 114–38.
39. Livingstone FB. On the non-existence of human races. *Curr Anthropol* 1963: **3**: 279–81.
40. Li JZ, Absher DM, Tang H et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008: **319**: 1100–4.
41. Bamshad M, Kivisild T, Watkins WS et al. Genetic evidence on the origins of Indian caste populations. *Genome Res* 2001: **11**: 994–1004.
42. Cooper RS, Kaufman JS, Ward R. Race and genomics. *N Engl J Med* 2003: **348**: 1166–70.
43. Romualdi C, Balding D, Nasidze IS et al. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* 2002: **12**: 602–12.
44. Wilson JF, Weale ME, Smith AC et al. Population genetic structure of variable drug response. *Nat Genet* 2001: **29**: 265–9.
45. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. *Nat Genet* 2008: **40**: 340–5.
46. Hancock AM, Witonsky DB, Alkorta-Aranburu G et al. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* 2011: **7**: e1001375.
47. Hancock AM, Witonsky DB, Ehler E et al. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A* 2010: **107** (Suppl 2): 8924–30.
48. Liu H, Prugnolle F, Manica A, Balloux F. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* 2006: **79**: 230–7.
49. Barbujani G. Human races: classifying people vs understanding diversity. *Curr Genomics* 2005: **6**: 215–26.
50. Madrigal L, Barbujani G. Partitioning of Genetic Variation in Human Populations and the Concept of Race. In: Crawford MH, eds. *Anthropological Genetics: Theory, Methods and Applications*: *Cambridge University Press*, *2007*, *19–37*.
51. Edwards AW. Human genetic diversity: Lewontin's fallacy. *BioEssays* 2003: **5**: 798–801.
52. Jorde LB, Watkins WS, Bamshad MJ et al. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* 2000: **66**: 979–88.
53. Long JC, Li J, Healy ME. Human DNA sequences: more variation and less race. *Am J Phys Anthropol* 2009: **139**: 23–34.
54. Ahn SM, Kim TH, Lee S et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 2009: **19**: 1622–9.
55. Burchard EG, Ziv E, Coyle N et al. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* 2003: **348**: 1170–5.
56. Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* 2002: **3**:comment2007.

57. Rightmire GP. Out of Africa: modern human origins special feature: middle and later Pleistocene hominins in Africa and Southwest Asia. *Proc Natl Acad Sci U S A* 2009: **106**: 16046–50.
58. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 2008: **9**: 403–33.
59. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. *Curr Biol* 2005: **15**: R159–60.
60. Tishkoff SA, Goldman A, Calafell F et al. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* 1998: **62**: 1389–402.
61. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 2005: **102**: 15942–7.
62. Manica A, Amos W, Balloux F, Hanihara T. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 2007: **448**: 346–8.
63. Santos RV, Fry PH, Monteiro S et al. Color, race, and genomic ancestry in Brazil: dialogues between anthropology and genetics. *Curr Anthropol* 2009: **50**: 787–819.
64. Taylor AL, Ziesche S, Yancy C et al. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* 2004: **351**: 2049–57.
65. Brody H, Hunt LM. BiDil: assessing a race-based pharmaceutical. *Ann Fam Med* 2006: **4**: 556–60.
66. Crawley L. The paradox of race in the Bidil debate. *J Natl Med Assoc* 2007: **99**: 821–2.
67. Hoover EL. There is no scientific rationale for race-based research. *J Natl Med Assoc* 2007: **99**: 690–2.
68. Garrod JZ. A brave old world: an analysis of scientific racism and BiDil. *Mcgill J Med* 2006: **9**: 54–60.
69. Petsko GA. Color blind. *Genome Biol* 2004: **5**: 119.
70. Wolinsky H. Genomes, race and health. Racial profiling in medicine might just be a stepping stone towards personalized health care. *EMBO Rep* 2011: **12**: 107–9.
71. Kahn J. BiDil: false promises: faulty statistics and reasoning have lead to the first "racial medicine". *Genewatch* 2005: **18**: 6–9 18.
72. Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. *Genome Res* 2005: **15**: 1592–3.
73. International HapMap Consortium, Frazer KA, Ballinger DG et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007: **449**: 851–61.
74. Kishi T, Ikeda M, Kitajima T et al. No association between prostate apoptosis response 4 gene (PAWR) in schizophrenia and mood disorders in a Japanese population. *Am J Med Genet B Neuropsychiatr Genet* 2008: **147B**: 531–4.
75. O'Donnell PH, Dolan ME. Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy. *Clin Cancer Res* 2009: **15**: 4806–14.
76. Sanoff HK, Sargent DJ, Green EM, McLeod HL, Goldberg RM. Racial differences in advanced colorectal cancer outcomes and pharmacogenetics: a subgroup analysis of a large randomized clinical trial. *J Clin Oncol* 2009: **27**: 4109–15.
77. Zhang W, Ratain MJ, Dolan ME. The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform Biol Insights* 2008: **2**: 15–23.
78. Lee SS, Mountain J, Koenig B et al. The ethics of characterizing difference: guiding principles on using racial categories in human genetics. *Genome Biol* 2008: **9**: 404.
79. Ali-Khan SE, Krakowski T, Tahir R, Daar AS. The use of race, ethnicity and ancestry in human genetic research. *Hugo J* 2011: **5**: 47–63.
80. Ng PC, Levy S, Huang J et al. *Genetic variation in an individual human exome*. *PLoS Genet* 2008: **4**: e1000160.
81. Need AC, Goldstein DB. Whole genome association studies in complex diseases: where do we stand? *Dialogues Clin Neurosci* 2010: **12**: 37–46.

PAPER II: Early modern human dispersal from Africa: genomic evidence for multiple waves of migration.

Early modern human dispersal from Africa: genomic evidence for multiple waves of migration

Francesca Tassi[1*], Silvia Ghirotto[1*], Massimo Mezzavilla[2], Sibelle Torres Vilaça[1,3], Lisa De Santi[1], and Guido Barbujani[1]

1 Department of Life Sciences and Biotechnologies, University of Ferrara, Italy; 2 Institute for Maternal and Child Health- IRCCS "BurloGarofolo"- Trieste , Italy and University of Trieste, Italy; 3 Current address: Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Guido Barbujani (email: g.barbujani@unife.it).

Running title: Early modern human dispersal

Key words: Human demographic history, migration, evolutionary divergence, admixture, linkage disequilibrium, population structure

It is unclear whether early modern humans left Africa through a single, major process, dispersing simultaneously over Asia and Europe, or in two main waves, first through the Arab peninsula into Southern Asia and Oceania, and later through a Northern route crossing the Levant. Here we show that accurate genomic estimates of the divergence times between European and African populations are more recent than those between Australo-Melanesia and Africa, and incompatible with the effects of a single dispersal. This difference cannot possibly be accounted for by the effects of hybridization with archaic human forms in Australo-Melanesia. Furthermore, in several populations of Asia we found evidence for relatively recent genetic admixture events, which could have obscured the signatures of the earliest processes. We conclude that the hypothesis of a single major human dispersal from Africa appears hardly compatible with the observed historical and geographical patterns of genome diversity.

Anatomically modern humans (AMH), defined by a lightly built skeleton, a large brain, reduced face and prominent chin, first appear in the East African fossil record around 200,000 years ago (McDougall et al. 2005; Aubert et al. 2012). There is a general consensus that, while dispersing from there, they largely replaced preexisting archaic human forms (Stringer 2002). Recent DNA studies also suggest that the replacement was not complete, and there was a limited, but nonzero, interbreeding with Neandertals (Green et al. 2010), Denisovans (Reich et al. 2011), and perhaps other African forms still unidentified at the fossil level (Hammer et al. 2011; Lachance et al. 2012). As a result, modern populations might differ for the amount of archaic genes incorporated in their gene pool, which are eventually expressed and may result in phenotypic differences affecting, for example, the immune response (Abi-Rached et al. 2011), or lipid catabolism (Khrameeva et al. 2014).

Although the general picture is getting clearer, many aspects of these processes are still poorly understood, starting from the timing and the modes of AMH dispersal. The main exit from Africa, through the Levant, has been dated around 56,000 years ago (Liu and Zhao 2006; Fu et al. 2013). However, morphologic (Lahr and Foley 1994; Reyes-Centeno et al. 2014), archaeological (Field et al. 2007) and genetic (Quintana-Murci et al. 1999; Macaulay et al. 2005; Di and Sanchez-Mazas 2011; Ghirotto et al. 2011; McEvoy et al. 2011; Rasmussen et al. 2011; Reyes-Centeno et al. 2014) evidence suggest that part of the AMH population might have dispersed before that date, possibly by a Southern route into Southern Asia through the horn of Africa and the Arab peninsula.

Regardless of whether there was a single major expansion or two, several DNA studies clearly showed that genetic diversity tends to decrease (Prugnolle et al. 2005; Li et al. 2008) and linkage disequilibrium to increase (Tishkoff et al. 1998; DeGiorgio et al. 2009) at increasing distances from Africa. This probably means that, as they came to occupy their current range, AMH went through a series of founder effects (Ramachandran et al. 2005; Deshpande et al. 2009).

These results offer an excellent set of predictions which we used in the present study to test whether current genomic diversity is better accounted for by processes involving a Single major Dispersal (hereafter: SD) or Multiple major Dispersals (hereafter: MD) from Africa.

One preliminary problem, however, is how to select the appropriate populations for informative comparisons. The details of the dispersal routes, and the relationships between fossils and contemporary populations, are all but established. Whereas Europeans are consistently regarded as largely derived from the most recent African exit in all relevant studies, opinions differ as for many aspects of the peopling of Asia (Lahr and Foley 1994; Quintana-Murci et al. 1999; Macaulay et al. 2005; Field et al. 2007; Di and Sanchez-Mazas 2011; Ghirotto et al. 2011; McEvoy et al. 2011; Reyes-Centeno et al. 2014), with many populations also experiencing complex demographic histories involving admixture, as suggested by both ancient (Gonzalez-Ruiz et al. 2012) and modern (Comas et al. 2004; Martinez-Cruz et al. 2011; Nievergelt et al. 2013; Mezzavilla et al. 2014) DNA evidence. To obtain insight into the past history of Eurasian populations we analyzed genome-wide autosomal single nucleotide polymorphisms (SNPs) from 71 worldwide populations (Supplemental Fig. S1). In what follows, a number of preliminary analyses allowed us to quantify the extent and the pattern of admixture and gene flow in our data, thus making it possible to identify a subset of Far eastern populations which, under the MD model, may safely be regarded as deriving from the oldest expansion.

This way, we could address two questions, related, respectively, with the historical and geographical context of the dispersal process, namely: (1) are separation times between non-African and African populations the same (as expected under SD), or is there evidence of a longer separation between Far Eastern and Africans than between Europeans and Africans (as expected under MD)? And (2) which geographical migration routes were followed by first humans outside Africa?

Results

Genomic structure of Old World populations. We assembled genome-wide SNP data from the literature obtaining information on 71 population samples sharing, after cleaning and integration, 96,156 autosomal SNPs (see Supplemental Methods for details). By merging samples from adjacent geographical regions and with similar linguistic affiliations, we organized the data in 24 meta-populations; the final dataset comprised 1,130 individuals (Fig. 1A and Supplemental Table S1).

As a preliminary step, we visualized by Principal Component Analysis the genetic relationships between such populations, as inferred from these autosomal SNPs (Fig. 2). The first two principal components, accounting respectively for 8.4% and 4.3% of the total genetic variance, show that the populations we grouped in meta-populations do cluster together genetically. In addition, genetic relationships largely correspond to geographical distances, with Eurasian populations separated from the African ones along the axis represented by PC1, and forming an orderly longitudinal cline, all the way from Europe to East Asia and Oceania, along the PC2 axis.

Then, to further investigate the worldwide genomic structure, we applied the unsupervised ancestry-inference algorithm of the ADMIXTURE software (Alexander et al. 2009). After pruning the dataset for LD and having evaluated the best supported number of clusters using a cross-validation error procedure (Supplemental Fig. S2 and Supplemental Methods), we explored the results for *k* = 2-7 ancestral populations performing 5 iterations for each *k* value. As the number of ancestral clusters increased, we observed the emergence of several well-supported population-specific ancestry clusters (Fig. 3). At *k* = 2, the ancestry assignment differentiated between African (blue) and non-African (yellow) populations; *k*=3 further distinguishes Europeans from Asians (orange); *k*=4 identifies an Australo-Melanesian component (green) within the Asian cluster; at *k*=5

the additional component is mainly associated with the Indian subcontinent (red); the same is the case at *k*=6 for Polynesia and Fiji (pink) and at *k*=7 for many island communities of Southeast Asia and Oceania (purple). Remarkably, some populations show more than 99% contribution from the same ancestral population along different *k* values (e.g. West Africa, Europe, New Guinea), whereas other populations include several individuals with an apparently admixed genomic background, possibly resulting from successive gene flow (e.g. back migration from Europe to Northeast Africa: Henn et al. 2012). A Discriminant Analysis of Principal Components (DAPC) (Jombart et al. 2010) led to essentially the same conclusions as ADMIXTURE (Supplemental Fig. S3-S4 and Supplemental Methods).

Population divergence dates. There is a clear geographical structure in the data, which in principle allows one to test for the relative goodness of fit of the two models. The SD model implies that the separation time from Africa of all samples should be the same, whereas significantly larger times of separation are expected under the MD model for the Easternmost than for the European populations. However, for neutral genome regions, genetic differences between populations, measured by F_{ST}, are inversely proportional to the effective population sizes (N_e) and directly proportional to the time since their separation (T). Therefore, if N_e values are unknown, as is the case here, an infinite number of T values can potentially account for any observed value of F_{ST}. To circumvent this problem, we resorted to genome-wide patterns of linkage disequilibrium (*LD*). Indeed, levels of *LD* also depend on *Ne*, and on the recombination rate between the SNPs considered (Tenesa et al. 2007), with *LD* between SNPs separated by large distances along the chromosome reflecting the effects of relatively recent N_e, whereas *LD* over short recombination distances depending on relatively ancient *Ne* (Hayes et al. 2003).

To assess the robustness of this method in estimating *Ne* from *LD*, we preliminarily ran this procedure using combinations of different numbers of markers and individuals, obtaining stable

results when at least 10-15 individuals and ~100,000 markers are considered (data not shown). To measure LD levels, we considered both the r^2 statistic (Hill and Robertson 1968) and the r^2 weighted by the product of the heterozygosities at the two loci (σ^2: Ohta and Kimura 1969) as suggested by Rogers (2014). Thus, we estimated *LD* levels independently in each meta-population, using all polymorphic markers available for that sample (which means from a minimum of ~ 90,000 SNPs in Polynesia to a maximum of ~ 370,000 SNPs in North India), then calculating the populations' N_es through time using the equation in (McEvoy et al. 2011); the values obtained using the two estimators of LD gave similar results (Supplemental Fig. S5A and B, see Supplemental Methods for details). The long-term N_e for each population is simply the harmonic mean of these values (Supplemental Fig. S6A and B).

The three African populations show the largest long-term sizes (Supplemental Fig. S6A and B) and a constant declining trend through time, whereas Eurasian populations (and more markedly the Asian ones) tend to increase in size, especially in the last 10,000 years. Australians and New Guineans (represented in green in the Admixture analysis at $k \geq 4$, Supplemental Fig. S5A and B) generally maintain a constant size until present times, whereas the Negrito populations show low and declining sizes. In general, these results were not surprising, but the fact we obtained them suggests that the procedure followed is by and large accurate, and therefore that the estimated average N_es (Supplemental Fig. S6A and B) are plausible.

This way, from the pairwise F_{ST} values estimated over all loci (Supplemental Table S2), and now considering the independently-estimated values of N_e, we could infer the divergence times between populations as in Holsinger and Weir (2009) (Table 1 and Supplemental Table S3A and B). The average separation times from the East African populations, i.e. those located in the most plausible site of departure of AMH expansions (Ramachandran et al. 2005) (Table 1) are distributed along a range spanning from 60K to 100K years ago. Extreme divergence values were

observed for Europe and the Caucasus on the one hand, and for Australia and New Guinea on the other, respectively at the lower and the upper tails of the distribution. Even considering the full range of uncertainty around these estimates (95% of the confidence interval) we observed no overlap, with Europe having an upper confidence limit 77/71K years ago (depending on the LD measure used, respectively the r^2 and σ^2 statistic) and Australia having a lower confidence limit 88/80K years ago. Because we kept into consideration the effects of N_e in the estimation procedure, this difference cannot possibly be accounted for by the different impact of genetic drift upon these populations, and supports a rather complex ''Out of Africa'' scenario, suggesting at least two main phenomena of AMH dispersal from Africa. The Australo-Melanesian populations, i.e. Australians and New Guineans, with their early separation times from East Africa, may be regarded as the putative descendants of an early dispersal process, whereas the status of most Asian populations would seem, at this stage of the analysis, unclear.

Comparing the predictions of single vs multiple African exit models: Divergence times. Having shown that significantly different times of separation from Africa are estimated for Europe and Australia/New Guinea, the question arises whether it would be possible to obtain such results by chance alone, had AMH dispersed in a single wave, at the time period at which that dispersal is generally placed (in the calculations that follow, we always considered the N_e and T estimates obtained using the unweighted r^2 statistic). To answer, we needed a null distribution of T values under the SD model, which we constructed by simulation, using the software *ms* (Hudson 2002). Genetic data (1 Mb for each individual) were simulated for two populations descended from a common ancestor, which separated t years ago (Supplemental Fig. S7). During the separation, one population underwent a bottleneck (mimicking the out-of-Africa process), and then grew exponentially until the present. To account for the uncertainty in the estimates of both the timing of the process and of the effective population sizes, we considered 4 different separation times

and 6 N_e estimates, and we simulated 1,000 independent datasets for each such combination. According to Petraglia et al. (2010), Sankararaman et al. (2012) and Mellars et al. (2013), the tested separation times from Africa were 40,000ya, 50,000ya, 60,000ya, and 70,000ya, and the tested effective population sizes were 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, for a total of 24 combinations. At every iteration, we evaluated the genetic variation at 1Mb in 50 individuals (i.e. 100 chromosomes) per population, thus analyzing 200,000 Mb for each parameter combination. For each simulated dataset we estimated the effective population sizes, the F_{ST} between populations and the divergence time, the same way as we did for the observed data, following the method described in McEvoy et al (2011). Then, we plotted the (null) distribution of the 24,000 separation times derived from the simulations and we compared it with the observed *T* estimates. Whereas the value estimated in the European sample falls perfectly within the range of times predicted by the classical SD model, that is not the case for the New Guinean and the Australian values, falling in the right tail of this distribution at $P<0.05$ level (Fig. 4). This can only mean that a single exit from Africa, even considering the uncertainty in our knowledge of the relevant parameters, cannot account for the differences in the separation times from Africa observed, respectively, in Europe on the one hand, and in Australo-Melanesia on the other.

Possible effects of a Denisovan admixture in Melanesia. Recent analyses of the genetic relationships between modern humans and Denisovans suggested that a fraction possibly as high as 6-8% of the Melanesian genomes may be traced back to Denisovan ancestor (Meyer et al. 2012). To rule out the possibility that the apparent difference in African divergence times for Europe and Australo-Melanesia may somewhat reflect Denisovan admixture, we removed from the analysis the SNPs that were identified as representing the Denisovan contribution to the latter's genome (see Supplemental Methods). We recalculated the F_{ST}s from 80,621 SNPs and

reestimated the divergence times from Africa, finding they are still very close to those previously estimated (Supplemental Tables S3 and S4).

Estimates of population admixture. Other Far Eastern populations, besides Australia and New Guineans for which we estimated a remote separation from Africans, may have taken part in an early exit from Africa through a Southern route. Identifying them is not straightforward, though, because we basically have a continuous set of divergence times from East Africa, from 66K to 107K years ago (Supplemental Table S1). This result is consistent with both a continuous migration process from Africa across some 40K years (which so far has never been proposed, to the best of our knowledge) and with an early exit, followed by genetic exchanges with later-dispersing groups, which has diluted or erased altogether the genetic evidence of the earliest migration. Our previous ADMIXTURE analysis highlighted an ancestral genetic component (green, Fig. 3) to which all Australo-Melanesian genotypes could be associated. In what follows, we explored the possibility that the same component be a marker of the earliest African exit in other populations as well. To understand whether that could have actually been the case, we used a method, *TreeMix* (Pickrell and Pritchard 2012) to estimate from genome-wide data a maximum-likelihood tree of populations, and then to infer events of gene flow after the split by identifying populations that poorly fit the tree; if admixture was extensive we expect to observe extensive reticulation in the tree. We selected from our dataset just the populations showing at least 30% of the green Admixture component at *k*=5, choosing the East Asia sample as outgroup (details in Supplemental Methods, results in Supplemental Fig. S8-S9). Evidence for extensive genetic exchanges after population splits is apparent from East Asia (represented in light blue in the tree) toward populations putatively involved in the early African dispersal (represented in green in the tree). Prior to adding these migration episodes, the graph captures 87% of the global variance in the data; including the top 6 migration events inferred by *TreeMix* brought this percentage to 99%.

Therefore, these results support the hypothesis that relatively recent admixture events could have obscured the genomic signatures of the first migration out of Africa in these Southeast Asian populations, ultimately biasing downwards the estimates of their divergence times from Africans.

Comparing the predictions of single vs multiple African exit models: Geographical patterns. To conclude, we tried to better define some details of the AMH dispersal out of the African continent by evaluating which geographical migration route can better account for the current patterns of genome diversity. For that purpose, we developed explicit geographic models of demographic expansion, and looked for the model giving the closest association between genomic and geographical distances. In all cases, migration routes were constrained by 5 obligatory waypoints, identified in Ramachandran et al. (2005) and accepted by several successive studies (see e.g Reyes-Centeno et al. 2014). In addition, because of some inconsistencies in the definition of the geographic regions affected by the two waves of migration under MD (Lahr and Foley 1994; Field et al. 2007; Ghirotto et al. 2011), and of the ambiguity introduced by the previously described episodes of admixture, two different models of MD were considered.

Under Model 1, a SD model, anatomically modern humans left Africa through Palestine and dispersed into both Europe and Asia (Fig. 1B). Model 2 assumes, prior to the dispersal across Palestine, another exit through the Arab Peninsula and the Indian Subcontinent, all the way to Melanesia and Australia; according to this model, based on skull morphology (Lahr and Foley 1994) all Asians populations are derived from this earlier expansion (Fig. 1C). On the contrary, under Model 3 only the populations of Southeast Asia and Oceania are derived from the earlier expansion, whereas Central Asian populations are attributed to the later African dispersal (Ghirotto et al. 2011) (Fig. 1D).

To minimize the effects of recent gene flow unrelated with the first human dispersals, which was clearly not negligible (see previous section) we selected populations with at least 80% of a single ancestral component in the ADMIXTURE results (i.e. Australia, the Caucasus, East Africa, East Asia, Europe, New Guinea, South Africa, South India, West Africa). Geographical distances between these populations were calculated according to the different hypothetical dispersal routes and taking into account the geographical barriers (mountain ranges, arms of sea, rivers) likely to oppose, or favor, population movements (Supplemental Methods for details). We evaluated by partial Mantel tests (Mantel 1967) the correlation between genomic (F_{ST}) and geographic distances, while holding divergence times constant. This way we could control for the drift effects, due to the fact that populations separated at distinct points in time and space. All the Mantel correlations thus calculated were positive and significant (Table 2), suggesting that all tested models succeed in plausibly predicting the observed patterns of genome diversity. The highest correlation observed for Model 3 (r=0.767) supports the southern route hypothesis for populations of South-East Asia and Oceania, but the difference between Models 3 and 1 is not significant by Fisher's criterion (Fisher 1921) (Z=-1.26, P=0.08).

Discussion

Two main factors, namely the effects of population sizes and of admixture after the main population split, complicate the estimation of divergence times between populations. As for the former, large genetic differences may mean that the populations long evolved independently, or that they had small effective sizes, or that these factors interacted. In this study, we resorted to *LD* values to tell apart the effects of population size and population history. This way, we found that the populations at the extremes of the geographical range considered differ substantially in the

timing of their separation from the Eastern African populations. This difference is statistically significant, and we showed by simulation that it cannot possibly be reconciled with a model assuming a single, major dispersal of all non-Africans through the classical Northern route. The model we tested is necessarily simple and does not take into account potential admixture with archaic human forms. However, since the estimated degree of Neandertal ancestry is the same in all non-African populations (Green et al. 2010), the inclusion of this event would only affect all the *absolute* values of divergence times from Africa, and not the *ratio* between them. In addition, a very small proportion of the SNPs we considered has been identified as the Denisovan contribution to the modern genomes; therefore, as we showed, the effect of interbreeding with Denisova (Meyer et al. 2012) upon our estimates can be regarded as minimal.

As for admixture after the split from Africa, which is known to inflate estimates of the divergence time (Sved 2009), it would be unrealistic to imagine it did not occur at all. However, although we cannot quantify its impact, we argue it is unlikely to be too strong, because significant differences in time estimates were observed for populations (Europe, the Caucasus, New Guinea and Australia) showing a rather homogeneous genetic composition in the ADMIXTURE analysis, with most individual genotypes attributed to a single ancestral component (Fig. 3). This, at least, does not suggest that population admixture seriously biased the divergence times inferred.

Conversely, the method used in the present study allows us to safely rule out that fluctuations in long-term population sizes might have distorted our time estimates. Three-fold differences in very ancient (e.g. > 100,000 years ago; Supplemental Fig. S6) population sizes may appear, at a first sight, difficult to justify, because at that time all N_e values should converge to a value representing the size of the common ancestral African population. However, a similar result was also obtained in the only previous study based on the same method (McEvoy et al. 2011), and interpreted as reflecting founder effects accompanying the dispersal from Africa. In turn, these

phases of increased genetic drift may have increased *LD*, and hence caused underestimation of N_e in all non-Africans. However, the resulting distortion, if any, should have affected the *absolute* values of *T*, but not the *relative* timing of the Europeans' and Asians' separation from Africans, which is what this study is concerned with. Another possibility is that 100,000 or so years ago the ancestors of current Eurasians were already genetically distinct from the ancestors of modern Africans, as proposed by Harding and McVean (2004), and Eriksson and Manica (2012). If so, the different N_e estimates of the present study would not be a statistical artifact, but would reflect actual differences between geographically-isolated ancient populations.

Two independent analyses (by ADMIXTURE and *TreeMix*) suggest that the genotypes of most Central Asians reflect variable degrees of gene flow between populations which may have left Africa in different waves. As a result, the distribution of divergence times is essentially continuous, and hence it would make no sense to try to classify Central Asian populations as derived from either the first or the second African exit under the model of multiple dispersals.

When we modeled population dispersal in space, the correlation between genetic and geographic distances was higher under the MD than under the SD model, but this difference was statistically insignificant (Table 2). This may be due, at least in part, to the fact that the three models being compared share several features, such as the same set of geographic/genetic distances for the European populations, which reduces the power of any test. However, the separation times previously estimated made us confident that the SD model is inconsistent with the data, and so what was really important at this stage was the comparison between the two MD models. The better fit of model 3 than model 2 implies that the MD model works better if only part of the Asian genomic diversity is attributed to the earliest dispersal. A better fit of a MD than of a SD model was also observed in parallel analyses of cranial measures and of a much smaller

genomic dataset (Reyes-Centeno et al. 2014), suggesting that our findings may indeed reflect a general pattern of human diversity.

In short, analyses of genomic data based on different sets of assumptions and different methods agree in indicating: (i) that a model with a single early dispersal from Africa fails to account for one crucial aspect of human genome diversity, the distribution of divergence times from Africa, and (ii) that within the model of multiple dispersal, geographical patterns of genome diversity are more accurately predicted assuming that not all Asian and New Guinea/Australian populations have had the same evolutionary history.

The data we analyzed are probably affected, to an unknown but not negligible extent, by a bias due to the fact that most SNPs in the genotyping platforms were discovered in European populations; however the measure we used to calculate N_e and hence the separation time, *LD*, is expected to be relatively unaffected by this kind of bias (Jakobsson et al. 2008; McEvoy et al. 2011). At any rate, a likely effect of such a bias would be a spurious increase of the estimated differences between Europeans and the populations being compared with them, Africans in this case. Quite to the contrary, here the Europeans appeared significantly *closer* to Africans than Australo-Melanesians, a result which therefore cannot be explained by that kind of ascertainment bias.

Can selection account, at least in part, for these findings? In principle, we have no way to rule this out. However, in practice, even though positive selection may have extensively affected the human genome, large allele-frequency shifts at individual loci are surprisingly rare (Coop et al. 2009), so much so that so far only for very few SNPs any effects of selection have been demonstrated (Hernandez et al. 2011). If we also consider that genomic regions with large allele-frequency differences are not generally associated with high levels of linkage disequilibrium, in

contrast with what would be expected after a selective sweep (Coop et al. 2009; Pritchard et al. 2010), it seems fair to conclude that the main allele frequency shifts occurred in a rather remote past and are unlikely to reflect geographic differences in the selection regimes (Alves et al. 2012). In any case, only 8% of the SNPs we considered map within expressed loci, or in their control regions (Fig. 5); therefore, the impact of selection upon the results of this study, if any, can hardly be regarded as substantial.

In the light of these results, we propose that at least two major dispersal phenomena from Africa led to the peopling of Eurasia and Australo-Melanesia. These phenomena seem clearly distinct both in their timing and in their geographical scope.

The view whereby only part of the ancestors of current non-African populations dispersed through the Levant has some non-trivial consequences upon the possible interactions between AMH and archaic forms, traces of whose genomes have been identified in many non-African populations, including New Guineans (Green et al. 2010; Sankararaman et al. 2014).The estimated contribution of Neandertals is less in the European than in the Asian/Melanesian genomes, despite the long coexistence between Neandertals and Europeans (Higham et al. 2014). At present, the standard way to explain this finding is to assume one single, major episode of hybridization in Palestine (or perhaps further North and East: Prufer et al. 2014), 47K to 65K years ago (Sankararaman et al. 2012), followed by a split between the Europeans' ancestors on the one hand, and the Asians' and Oceanians' on the other (Stoneking and Krause 2011; Prufer et al. 2014). After that, additional contacts might have occurred, but only between Neandertals and Asians (Vernot and Akey 2014). However, if most ancestors of New Guineans dispersed through a Southern route, as this study shows, they would have missed by 2,000 km or so the nearest documented Neandertals with whom they could have intercrossed. Thus, this study raises the possibility that the current patterns of human diversity need more complex models to be fully

explained. One possibility is that admixture with Neandertals might have occurred before AMH left Africa (Sanchez-Quinto et al. 2012). Another is that common ancestry, rather than hybridization, may account for the excess similarity of Eurasians with Neandertals, in the presence of an ancient structuring of populations (Ray et al. 2005; Eriksson and Manica 2012). These hypotheses are not necessarily alternative to hybridization in Palestine, but exploring them may contribute to a better understanding of the relationships between archaic and modern human forms.

Methods

We combined genomic data from six published sources, (i.e. Lopez Herraez et al. 2009; Reich et al. 2009; Xing et al. 2009; Xing et al. 2010; Reich et al. 2011; Pugach et al. 2013) using PLINK v 1.07 (Purcell et al. 2007); after cleaning for genotyping rate, MAF and presumably related individuals showing excess allele sharing, the final dataset comprised 1,130 individuals, each typed for 96,156 shared SNPs. The Principal Component Analysis was performed using the R (R Development Core Team 2011) SNPRelate package. Individual ancestry components were inferred by the software ADMIXTURE (Alexander et al. 2009). The required LD pruning was done with the PLINK tool (Purcell et al. 2007), using a threshold r^2 of 0.3, which reduced the dataset to 54,978 markers. The 5 independent runs were combined by the software CLUMPP (Jakobsson and Rosenberg 2007) the resulting ancestry components were then plotted by the software Distruct (Rosenberg 2004). Effective population sizes and divergence times were estimated by the NeON (Mezzavilla and Ghirotto 2015) and 4P (Benazzo et al. 2014) software packages developed by the authors and available online at (www.unife.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software). To evaluate the validity of the conclusions drawn from the divergence-times analysis we used a simulation approach based on the neutral coalescent model of the software *ms*

(Hudson 2002). We simulated genetic data (multiple 1-Mb segments of DNA) according to the demographic model detailed in Supplemental Fig. S7 under the infinite sites model of mutation. Allele frequencies for the *TreeMix* (Pickrell and Pritchard 2012) analysis were estimated using PLINK tool (Purcell et al. 2007), after cleaning for LD as in the ADMIXTURE analysis. East Asia was set as an outgroup, and we used the window size of 500 (-k option). The partial Mantel (Mantel 1967) correlations were calculated using the R Vegan package, and their significance was empirically estimated over 10,000 random permutations.

Author contributions

S.G., F.T. and G.B. conceived and designed the analyses; S.G., F.T., M.M.,S.V.T. and L.D.S. analyzed the data; G.B., S.G. and F.T wrote the manuscript.

Acknowledgements

This study was supported by the European Research Council ERC-2011-AdG_295733 grant (LanGeLin), and in part by a grant of the Italian Ministry for Research and Universities (MIUR) PRIN 2010-2011. We are indebted to Mariateresa Vizzari for technical help, and to Andrea Benazzo, Cesare de Filippo and Johannes Krause for suggestions and critical discussion of this manuscript.

Table 1. Estimated divergence times from (East) Africa using the r^2 or σ^2 statistics as estimator of LD level.

TIME	r^2 (Hill and Robertson 1968)			σ^2 (Ohta and Kimura 1969)		
	5	50	95	5	50	95
Europe	63,135	69,736	76,645	57,084	65,402	70,830
Caucasian	62,363	68,143	74,016	56,531	63,196	68,797
West_Asia	60,458	66,318	73,247	55,174	61,586	67,061
Central_Asia	66,166	71,021	78,819	60,510	66,425	72,061
North_India	65,930	70,230	77,595	60,270	65,243	71,325
South_India	60,625	64,396	70,718	55,445	60,664	66,051
East_Asia	81,456	87,432	95,874	73,793	81,398	87,241
South_Asia	74,310	80,587	88,651	67,515	74,425	81,458
Malaysia	66,862	71,622	80,114	61,092	66,852	73,143
Borneo	75,517	80,056	88,234	68,644	74,579	81,799
Sumatra	75,884	82,043	90,707	69,281	76,108	82,934
East_Indonesia	66,801	71,576	79,538	61,208	67,056	73,154
Philippine	74,051	79,248	87,916	67,651	73,996	81,242
Moluccas	66,571	71,562	79,457	60,753	66,875	73,078
Australia	**87,828**	**96,599**	**108,214**	**79,827**	**89,596**	**98,794**
New_Guinea	**99,852**	**107,204**	**119,569**	**90,693**	**99,499**	**110,530**
Fiji	71,465	77,395	84,437	65,546	72,155	78,014
Polynesia	71,753	77,531	86,510	65,566	72,451	79,150
Onge	77,243	82,572	91,234	70,827	77,670	84,637
Jehai	66,414	71,521	79,922	60,820	66,859	73,613
Mamanwa	67,925	73,012	82,492	62,288	68,502	76,183

For each comparisons with East African, the three columns report the 95% lower confidence limit, the point estimate (in years, assuming a generation interval =25 years (Fenner 2005)), and the 95% upper confidence limit.

Table 2. Partial Mantel correlations between genetic and geographic distances.

	Partial Mantel Test	
	r	**p-value**
Model 1	0.67	0.0001
Model 2	0.64	0.0012
Model 3	**0.77**	0.0001

Comparisons of the genetic distance matrix (F_{ST}) with the geographic distances calculated according to the three dispersal models, while holding constant population divergence values (*T*). Values are Pearson correlation coefficients, and the *P*-values have been empirically calculated over 10,000 permutations of one matrix' rows and columns.

References

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA et al. 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 334: 89-94.

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

Alves I, Sramkova Hanulova A, Foll M, Excoffier L. 2012. Genomic data reveal a complex making of humans. *PLoS Genet* 8: e1002837.

Aubert M, Pike AW, Stringer C, Bartsiokas A, Kinsley L, Eggins S, Day M, Grun R. 2012. Confirmation of a late middle Pleistocene age for the Omo Kibish 1 cranium by direct uranium-series dating. *J Hum Evol* 63: 704-710.

Benazzo A, Panziera A, Bertorelle G. 2014. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecology and Evolution*.

Comas D, Plaza S, Wells RS, Yuldaseva N, Lao O, Calafell F, Bertranpetit J. 2004. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet* 12: 495-504.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. *PLoS Genet* 5: e1000500.

DeGiorgio M, Jakobsson M, Rosenberg NA. 2009. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci U S A* 106: 16057-16062.

Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL. 2009. A serial founder effect model for human settlement out of Africa. *Proc Biol Sci* 276: 291-300.

Di D, Sanchez-Mazas A. 2011. Challenging views on the peopling history of East Asia: the story according to HLA markers. *Am J Phys Anthropol* 145: 81-96.

Eriksson A, Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc Natl Acad Sci U S A* 109: 13956-13960.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 128: 415-423.

Field JS, Petraglia MD, Lahr MM. 2007. The southern dispersal hypothesis and the South Asian archaeological record: Examination of dispersal routes through GIS analysis. *Journal of Anthropological Archaeology* 26: 88-108.

Fisher RA. 1921. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1: 3-32.

Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J et al. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* 23: 553-559.

Ghirotto S, Penso-Dolfin L, Barbujani G. 2011. Genomic evidence for an African expansion of anatomically modern humans by a Southern route. *Hum Biol* 83: 477-489.

Gonzalez-Ruiz M, Santos C, Jordana X, Simon M, Lalueza-Fox C, Gigli E, Aluja MP, Malgosa A. 2012. Tracing the origin of the east-west population admixture in the Altai region (Central Asia). *PLoS One* 7: e48904.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH et al. 2010. A draft sequence of the Neandertal genome. *Science* 328: 710-722.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. Genetic evidence for archaic admixture in Africa. *Proc Natl Acad Sci U S A* 108: 15123-15128.

Harding RM, McVean G. 2004. A structured ancestral population for the evolution of modern humans. *Curr Opin Genet Dev* 14: 667-674.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* 13: 635-643.

Henn BM, Botigue LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J et al. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* 8: e1002397.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331: 920-924.

Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, Camps M, Arrizabalaga A, Baena J, Barroso-Ruiz C et al. 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* 512: 306-309.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet* 38: 226-231.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 10: 639-650.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337-338.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23: 1801-1806.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.

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.

Khrameeva EE, Bozek K, He L, Yan Z, Jiang X, Wei Y, Tang K, Gelfand MS, Prufer K, Kelso J et al. 2014. Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nat Commun* 5: 3584.

Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo JM, Lema G, Fu W, Nyambo TB, Rebbeck TR et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150: 457-469.

Lahr MM, Foley RA. 1994. Multiple Dispersals and Modern Human Origins. *Evolutionary Anthropology* 3: 48-60.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100-1104.

Liu N, Zhao H. 2006. A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics* 2: 353-364.

Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4: e7888.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F et al. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308: 1034-1036.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res* 27: 209-220.

Martinez-Cruz B, Vitalis R, Segurel L, Austerlitz F, Georges M, Thery S, Quintana-Murci L, Hegay T, Aldashev A, Nasyrova F et al. 2011. In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur J Hum Genet* 19: 216-223.

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

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21: 821-829.

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

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338: 222-226.

Mezzavilla M, Vozzi D, Pirastu N, Girotto G, D'Adamo P, Gasparini P, Colonna V. 2014. Genetic landscape of populations along the Silk Road: admixture and migration patterns. *BMC Genetics* 15: 131.

Mezzavilla M, Ghirotto S. 2015. Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS. *J Comput Sci Syst Biol* 037-004.

Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, Kidd JR. 2013. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet* 4: 13.

Ohta T, Kimura M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* 63: 229-238.

Petraglia MD, Haslam M, Fuller DQ, Boivin N, Clarkson C. 2010. Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim. *Annals of human biology* 37: 288-311.

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

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* 20: R208-215.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505: 43-49.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr Biol* 15: R159-160.

Pugach I, Delfin F, Gunnarsdottir E, Kayser M, Stoneking M. 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc Natl Acad Sci U S A* 110: 1803-1808.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-575.

Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* 23: 437-441.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. Vienna, Austria : the R Foundation for Statistical Computing. *ISBN: 3-900051-07-0 Available online at http://wwwR-projectorg/*.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 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: 15942-15947.

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334: 94-98.

Ray N, Currat M, Berthier P, Excoffier L. 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* 15: 1161-1167.

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89: 516-528.

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

Reyes-Centeno H, Ghirotto S, Detroit F, Grimaud-Herve D, Barbujani G, Harvati K. 2014. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc Natl Acad Sci U S A* 111: 7248-7253.

Rogers AR. 2014. How population growth affects linkage disequilibrium. *Genetics* 197: 1329-1341.

Rosenberg NA. 2004. Distruct: a program for the graphical display of population *Molecular Ecology Notes* 4: 137-138.

Sanchez-Quinto F, Botigue LR, Civit S, Arenas C, Avila-Arcos MC, Bustamante CD, Comas D, Lalueza-Fox C. 2012. North African populations carry the signature of admixture with Neandertals. *PLoS One* 7: e47765.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507: 354-357.

Sankararaman S, Patterson N, Li H, Paabo S, Reich D. 2012. The date of interbreeding between Neandertals and modern humans. *PLoS Genet* 8: e1002947.

Stoneking M, Krause J. 2011. Learning about human population history from ancient and modern genomes. *Nat Rev Genet* 12: 603-614.

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

Sved JA. 2009. Correlation measures for linkage disequilibrium within and between populations. *Genet Res (Camb)* 91: 183-192.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17: 520-526.

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK. 1998. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* 62: 1389-1402.

Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343: 1017-1021.

Xing J, Watkins WS, Shlien A, Walker E, Huff CD, Witherspoon DJ, Zhang Y, Simonson TS, Weiss RB, Schiffman JD et al. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 96: 199-210.

Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, Thara R, Mowry BJ, Bulayeva K, Weiss RB, Jorde LB. 2009. Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res* 19: 815-825.

Figure 1 | Geographic location of the 24 metapopulations analyzed (A) and geographical models of African dispersal (B, C, D). Metapopulations, each derived from the merging of genomic data from several geographically or linguistically-related populations, are South, East and West Africa, Europe, the Caucasus, South, East, West and Central Asia, North and South India, plus three Negrito (Onge, Jehai and Mamanwa) and ten Oceanian populations; the final dataset comprised 1,130 individuals. Under model 1, a SD model (B), all non-African populations are descended from ancestors who left Africa through the same, Northern route (Stringer 2002). Model 2 (C) and Model 3 (D) are MD models assuming, prior to dispersal across Palestine, another exit through the Arab Peninsula and the Indian Subcontinent; under Model 2 all Asian and Western Oceanian populations derive from this earlier expansion (Lahr and Foley 1994), whereas under Model 3 only the populations of Southeast Asia and Western Oceania derive from the earlier expansion (Ghirotto et al. 2011).

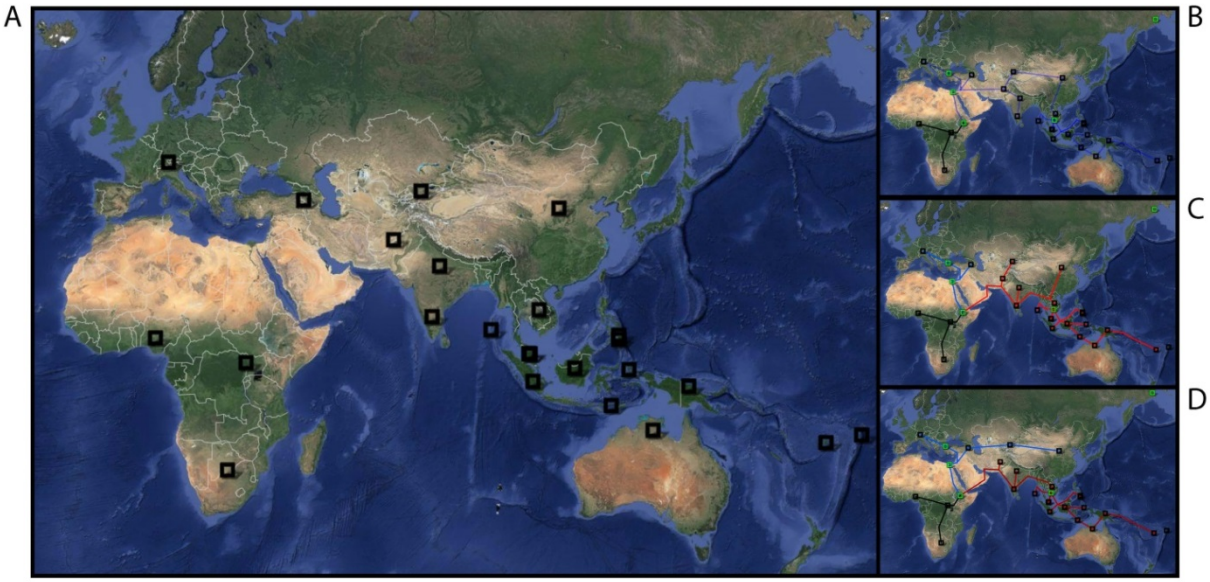


Figure 2 |Results of the Principal Component Analysis. Each symbol corresponds to an individual genotype; the first two principal components account for 12.7% of the global variation in the data. Here and in all figures, different colours represent different geographical regions.

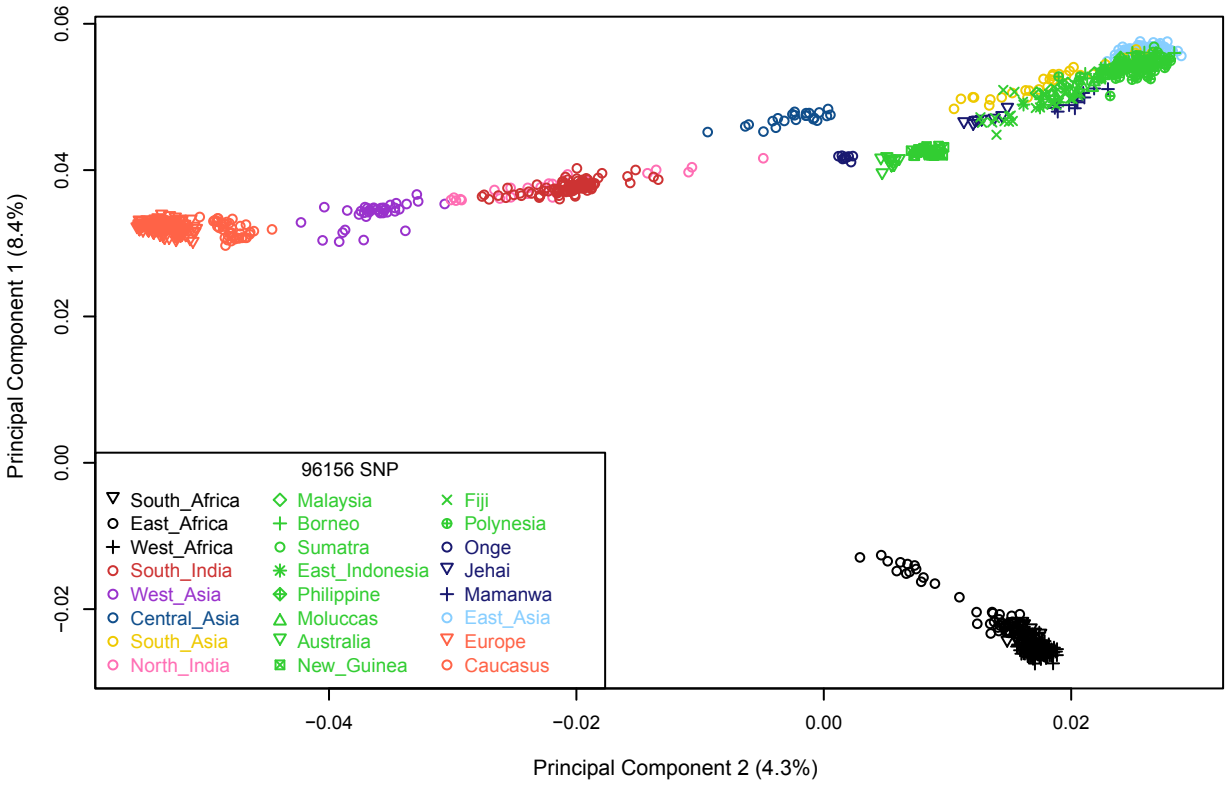


Figure 3 | Admixture analysis of 1130 individuals in 24 populations from Africa, Eurasia and Western Oceania. Each individual genotype is represented by a vertical column, the colors of which correspond to the inferred genetic contributions from k putative ancestral populations. The analysis was run for $2 \leq k \leq 7$

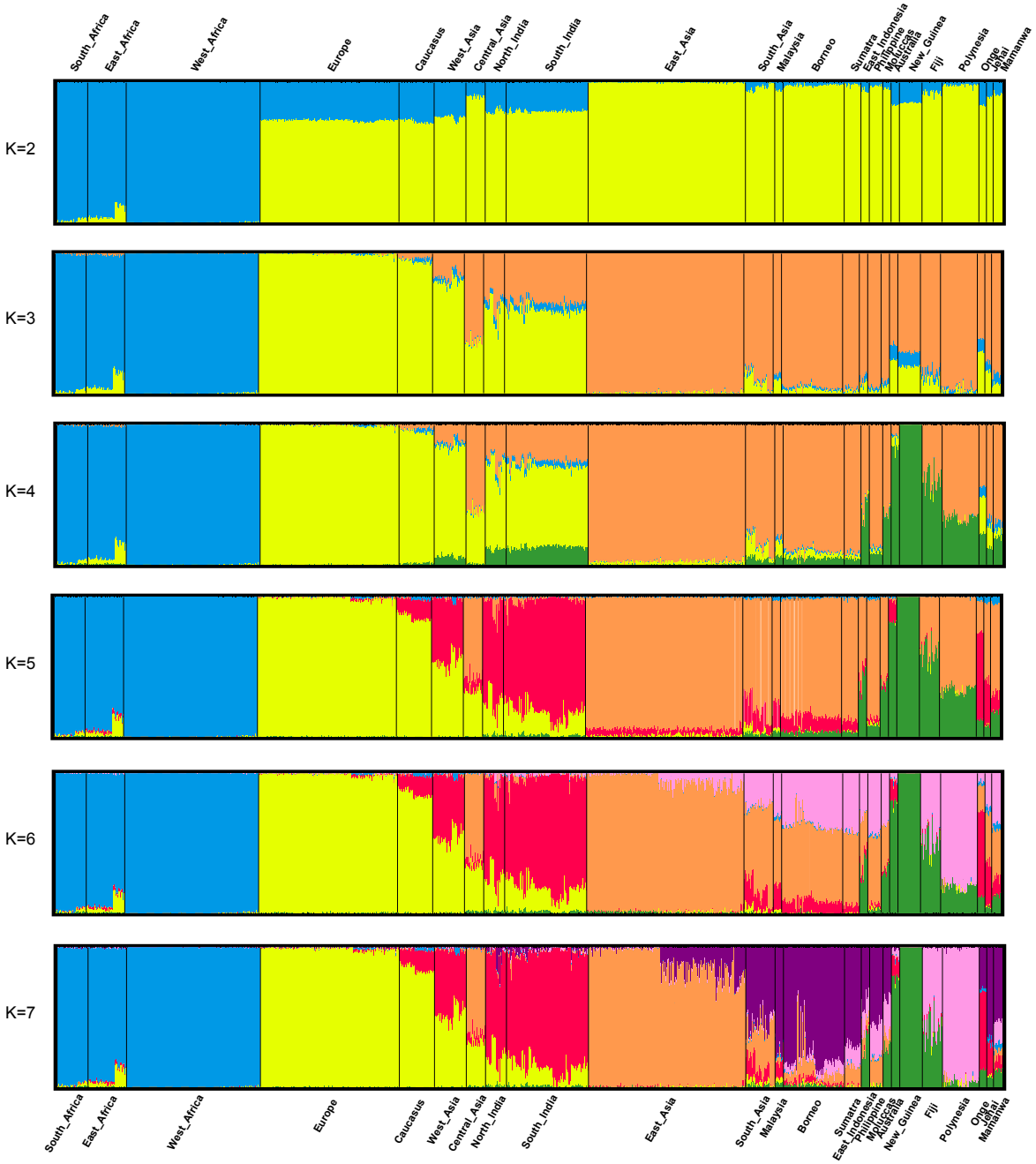


Figure 4 | Comparison of three observed divergence times with the distribution of 24,000 divergence times between African and non-African populations generated by simulation of a SD model. Data generated for 24 combinations of effective population sizes (3,000 ≤ Ne ≤ 8,000) and divergence times (40 k years ago ≤ T ≤ 70 k years ago), 1,000 independent datasets for each such combination. At every iteration, genetic variation at 1Mb was considered in 100 chromosomes per population, thus analyzing 200,000 Mb for each parameter combination (for a total of 4,800 Gb in 24,000 iterations, see Supplemental Methods for details).

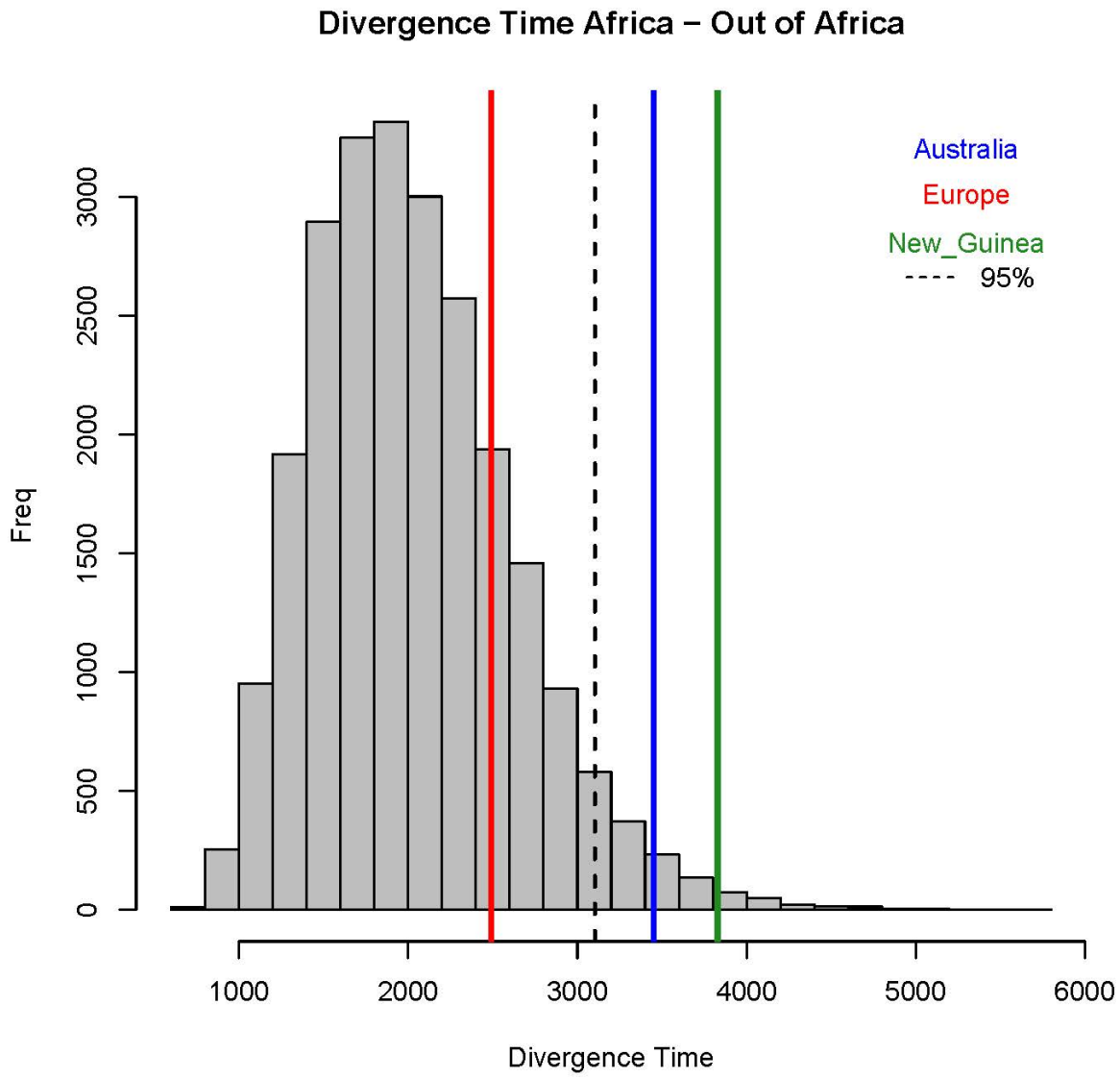
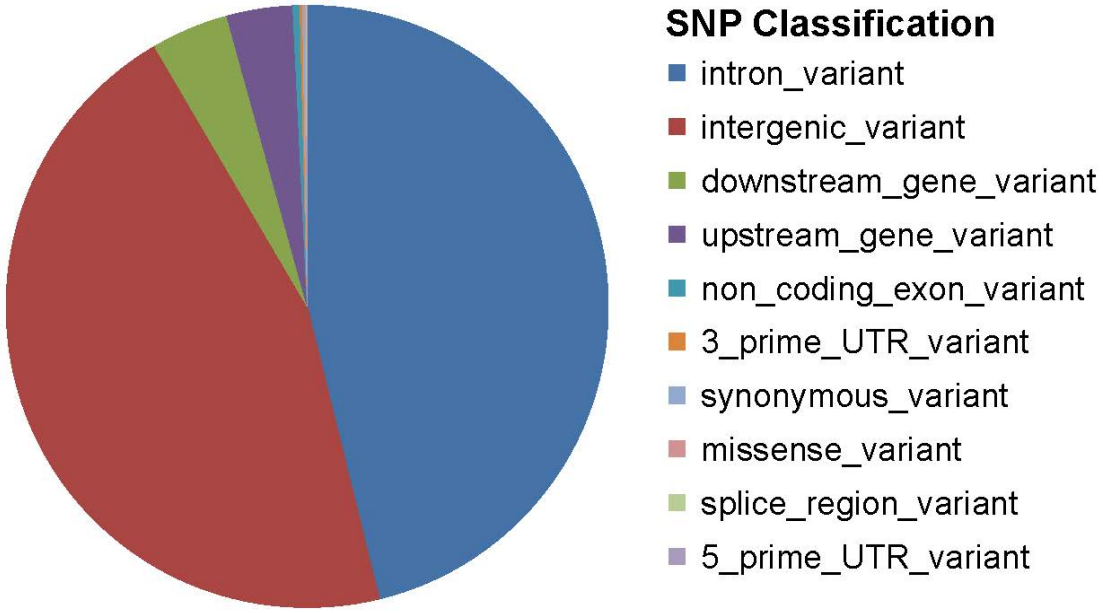


Figure 5 | Distribution of the SNPs considered in functionally different genome regions.



Materials and Methods

Subjects and markers

We combined genomic data from several published datasets: the Human Genome Diversity Cell Line Panel (Lopez Herraez et al. 2009) (n = 40 samples from 10 populations genotyped on Affymetrix GeneChip Human Mapping 500 K Array Set), Pugach et al (2013) (n= 117 samples from 12 populations genotyped on an Affymetrix 6.0 array), Reich et al (2009) (n = 56 samples from 11 populations genotyped on an Affymetrix 6.0 array), Reich et al (2011) (n = 509 samples from 13 populations genotyped on an Affymetrix 6.0 array), Xing et al (2009) (n = 243 samples from 17 populations genotyped on one array (version NspI) from the Affymetrix GeneChip Human Mapping 500K Array set), Xing et al (2010) (n = 165 samples from 8 populations genotyped on an Affymetrix 6.0 array), (Supplemental Fig.S1 and Supplemental Table S1).

We devised a careful strategy to combine the seven datasets genotyped with different platforms according to different protocols developing a pipeline built on Perl. First, for each dataset, we checked for the presence of old rs ids, if necessary changing them with the new ones (the rs merge table (RsMergeArch) can be found at the dbSNP ftp site: ftp://ftp.ncbi.nlm.nih.gov/snp/). Then, we looked for the SNPs shared among all datasets and we mapped the genome positions of these variants to the human reference genome, build hg18 (NCBI 36).

When merging data from different SNP-chip versions, strand identification can be ambiguous, possibly leading to mistakes in identifying the right alleles for A/T and G/C SNPs (as also reported in the PLINK tool documentation (Purcell et al. 2007)). Thus, to preserve as much genetic information as possible, we selected from each dataset only these ambiguous SNPs and we used the information contained in Affymetrix Annotation file to evaluate the strand polarity used to define each allele. We considered each dataset separately and we annotated the SNPs on

the plus strand, flipping only the proper SNPs. We checked the reliability of this conversion process comparing the allele frequencies for these SNPs in specific populations typed in more than one dataset (i.e Besemah, CEU, Onge), so as to verify the consistency of the frequency spectrums between the different datasets. Once these ambiguities have been resolved, with the PLINK v 1.07 software (Purcell et al. 2007) we merged progressively the datasets selecting, from each one, just the individuals from populations of our interest and flipping SNPs discordant for strand.

Using the same software, we selected only the autosomal SNPs with genotyping success rate >98% and minor allele frequency (MAF) >0.01. We identified cryptic relatedness amongst samples computing identity-by-descent (IBD) statistic for all pairs of individuals, as unmodeled excess of genetic sharing would violate sample independence assumption of downstream analyses. When pairs of individuals showed a Pi-Hat value > 0.3, we removed the individual with the lowest genotyping rate. We did not apply this screening procedure for the South-East Asia and Oceania samples, since they come from populations with extremely low effective sizes, where a certain degree of random inbreeding is inevitable (Relethford 1985). To determine whether there were genetic outliers within each population, we conducted in PLINK a “distance to the nearest neighbor analysis” (--neighbor option). Within each population, the measure of similarity in terms of identity by state (IBS) between each individual and their nearest neighbor was calculated and transformed into a Z-score. Z score distributions were examined from the first to the fifth neighbor. Outliers were identified by an extremely negative Z-score produced by *low* allele sharing with their nearest neighbor and were then dropped from the population.

After the data quality control and cleaning, the final dataset contained 1,130 individuals, each typed for 96,156 SNPs shared by all populations. We grouped these individuals into 24 ethnolinguistically and geographically related meta-populations (Fig. 1). To visualize the genetic relationships between such populations, we performed a Principal Component Analysis using the

R(R Development Core Team 2011) SNPRelate package (Fig. 2).

Population structure analysis

Individual genotypes were clustered, and admixture proportions were inferred, by the algorithm embedded in the software ADMIXTURE, based on the principle of maximum likelihood (Alexander et al. 2009). This method considers each genotype as drawn from an admixed population with contributions from *k* hypothetical ancestral populations. Because this model assumes linkage equilibrium among markers, we checked with the PLINK v1.07 tool (Purcell et al. 2007) that the set of SNPs we used did not show a level of Linkage Disequilibrium higher than r^2=0.3; this way, in the pruned dataset 54,978 markers were retained. The optimal value of *k* was evaluated through a cross-validation procedure, testing values from *k*=2 to *k*=14, thus identifying the number of ancestral populations for which the model had the best predictive accuracy (Supplemental Fig. S2). We then ran an unsupervised analysis, assuming a number of ancestral admixing populations from *k* =2 to *k* =7. The proportion of the individuals' genome belonging to each ancestral population was calculated for each k value from 5 independent runs, then combined by the software CLUMPP (Jakobsson and Rosenberg 2007) and plotted by the software *Distruct* (Rosenberg 2004) (Fig. 3).

Discriminant Analysis of Principal Components

In addition to ADMIXTURE, to identify and describe clusters of genetically related individuals we used a Discriminant Analysis of Principal Components (DAPC) (Jombart et al. 2010) implemented in the R (R Development Core Team 2011) package adegenet ver. 1.3-9.2 (Jombart and Ahmed 2011). DAPC methods allow one to assess the relationships between populations overlooking the within-group variation and summarizing the degree of between group variation.

Being a multivariate method, DAPC is suitable for analyzing large numbers of genome-wide SNPs, providing assignment of individuals to different groups and an intuitive visual description of between-population differentiation. Because it does not rely on any particular population genetics model, DAPC is free of assumptions about Hardy-Weinberg equilibrium or linkage equilibrium (Jombart et al. 2010), and so we could use the full set of 96,156 SNPs for this analysis.

By the function *find.clusters*, we determined the most likely number of genetic clusters in our dataset, using all principal components (PCs) calculated on the data. The method uses a K-means clustering of principal components (Liu and Zhao 2006) and a Bayesian Information Criterion (BIC) approach to assess the best supported number of clusters. We found K=6 to be the best supported model (Supplemental Fig. S3) and therefore used this value in the DAPC.

Then, we determined the optimal number of principal components (PCs) to retain to perform a discriminant analysis avoiding unstable (and improper) assignment of individuals to clusters. It is worth noting that, unlike K-means, DAPC can benefit from not using too many PCs: retaining too many components with respect to the number of individuals can lead to over-fitting and instability in the membership probabilities returned by the method.

Supplemental Fig. S4A shows that the main populations are distinguishable, and most individuals from the same population tend to fall in the same cluster. In the scatter plot the first two axes revealed three major clusters within the six supported by the k = 6 model (Supplemental Fig. S4B). They included (i) the three African population, (ii) most populations from Asia, and (iii) populations from Europe and Caucasus and from India and West-Asia. This clustering pattern is also observed in Admixture analysis with k = 6 (Fig. 3). Interestingly, in the Asian group the DAPC is able to distinguish three different clusters: one represented by individuals from Australia and New Guinea (in green color), one by the populations showing at least 30% of the green Admixture component at K=5 (in pink color), and one by other populations from Asia.

Population divergence dates

The divergence times between populations (*T*), was estimated from the population differentiation index (F_{ST}) and the effective population size (*Ne*). F_{ST} is the proportion of the total variance in allele frequencies that is found between groups and it was calculated between pairs of populations for each SNP individually under the random population model for diploid loci, as described by Weir and Cockeram (1984), and then averaged over all loci to obtain a single value representing pairwise variation between populations (Supplemental Table S2). Under neutrality, the differences between populations accumulate because of genetic drift, and so their extent depends on two quantities: it is inversely proportional to the effective population sizes (N_e) and directly proportional to the time passed since separation of the two populations (*T*).

Therefore, to estimate *T* from genetic difference between populations, independent estimate of N_e are needed; for this purpose we focused on the relationship between N_e and the level of linkage disequilibrium within populations. We considered that levels of Linkage Disequilibrium (*LD*) depend on both N_e and on the recombination rate between the SNPs considered (Tenesa et al. 2007). However, *LD* between SNPs separated by large distances along the chromosome mirror the effect of relatively recent N_e whereas *LD* over short recombination distances depends on relatively ancient N_e (Hayes et al. 2003). Thus, we estimated *LD* independently in each population using all polymorphic markers available for that population (MAF > 0.05), from a minimum of ~ 90,000 SNPs in Polynesia to a maximum of ~ 370,000 SNPs in North India. This way, we also reduced the impact of ascertainment bias, i.e. the bias due to the fact that most SNPs in the genotyping platforms were discovered in a single (typically European) population (Clark et al. 2005).

We assigned to each SNP a genetic map position based on HapMap2 (Release #22) recombination data, and for each pair of SNPs separated by less than 0.25 cM we quantified *LD* as

r^2_{LD} (Hill and Robertson 1968) or as σ^2_{LD} (Ohta and Kimura 1969) (hereafter: ρ). All the observed ρ values were then binned into one of 50 recombination distance classes, from 0.005 to 0.25 cM, with incremental upper boundaries of 0.005 cM. Pairs of SNPs separated by less than 0.005 cM were not considered in the analysis, since at these very short distances gene conversion may mimic the effects of recombination (Tenesa et al. 2007). We also adjusted the ρ value for the sample size using $\rho - \left(\frac{1}{n}\right)$ (Tenesa et al. 2007). Finally, we calculated the effective population size for each population in each recombination distance class as

$$Ne = \left(\frac{1}{4c}\right)\left[\frac{1}{\rho} - 2\right],$$

corresponding to the effective population size $\frac{1}{2}c$ generations ago, where c is the recombination distance between loci, in Morgans (Sved 1971; McVean 2002; Hayes et al. 2003) (Supplemental Fig. S5A and B). Finally, the long-term N_e for each population was calculated as the harmonic mean of N_e over all recombination distance classes up to 0.25 cM. The confidence intervals of these N_e values were inferred from the observed variation in the estimates across chromosomes (Supplemental Fig. S6A and B).

Based on the independently-estimated values of N_e, we could then estimate T as

$$T = \ln(1 - Fst) / \ln\left(1 - \left(\frac{1}{2Ne}\right)\right) \text{ (Holsinger and Weir 2009)}$$

where time is expressed in generations (Supplemental Table S3A and B).

All procedures were performed by in-house developed software packages, NeON (Mezzavilla and Ghirotto 2015) and 4P (Benazzo et al. 2014) available online at (www.unife.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software).

Possible effects of a Denisovan admixture in Melanesia

To rule out the possibility that the high divergence time estimated between Africans and New Guinea/Australia samples (Supplemental Table S3A and B) could reflect, largely or in part, admixture between the Denisovan archaic human population from Siberia (Meyer et al. 2012) and the direct ancestor of Melanesians, we removed from our dataset the variants could be regarded as resulting from such a process of introgression. These SNPs would carry the derived state in the archaic population and in the New Guinean/Australian samples, while being ancestral in East Africans and Europeans (i.e. those populations that did not show any signal of introgression from Denisova (Reich et al. 2011; Meyer et al. 2012)).

Using the *VCF tools* (Danecek et al. 2011) we extracted our 96,156 SNP from the high coverage Denisovan genome (http://cdna.eva.mpg.de/denisova/VCF/hg19_1000g/). We then removed from these data all transitions SNPs (C/T and G/A) because in ancient DNA these sites are known to be prone to a much higher error rate than the transversions (Reich et al. 2011) .Then, we selected the sites meeting the following set of criteria:

- the site has human-chimpanzee ancestry information;
- the human-chimpanzee ancestral allele matches one of the two alleles at heterozygous sites;
- Denisova has at least one derived allele;
- New Guineans and Australians have at least one derived allele;
- in East African and Europe individuals the ancestral allele is fixed;

When the ancestry information was missing (1,438 SNPs), to define the ancestral state, we used the East African individuals selecting the SNPs where East Africans were homozygous and considering those as ancestral.

Once we had thus identified a subset of sites putatively introgressed from Denisova, we evaluated whether the differences in divergence times between East Africans can be explained by

archaic admixture. To do so, we removed from the dataset the putatively introgressed positions, and the remaining 80,621 SNPs were used to compute the pairwise F_{ST} (Weir and Cockerham 1984) values and to infer the divergence time between the 24 meta-populations, as described above (Supplemental Table S4)

Simulations

A neutral coalescent approach was used to simulate genetic polymorphism data under the infinite sites model of mutation. We simulated data representing 1Mb chromosome segments in two populations according to the demographic scenario shown in Supplemental Fig. S7 using the coalescent simulator *ms* (Hudson 2002). We assumed an ancestral population with an initial Ne= 10,000. At t = T, the population splits into two populations. Population_2a's Ne remains constant, population_2b has a 50% reduction in Ne followed by an exponential growth, representing the genetic bottleneck experienced by populations dispersing out of Africa. In all simulations the scaled mutation rate (θ), and the scaled population recombination rate (ρ) were fixed at 400. For a sequence length of 1Mb and an effective population size of Ne= 10,000 these parameters correspond to a mutation rate of 10^{-8} and a recombination rate of 1 cM/Mb.

This model was simulated considering 4 different separation times (T) (between 40,000 and 70,000 years ago, in steps of 10,000 years) and 6 estimates of the actual effective size for population_2b (between 3,000 and 8,000, in steps of 1,000). For each of the 24 simulation conditions, 1,000 independent datasets were simulated and then analyzed according to the following procedure:

1) A sample of 50 diploid individuals was randomly selected from each population. The simulated genetic data were single nucleotide polymorphisms (SNPs) data segregating within the two populations.

2) We converted the *ms (Hudson 2002)* output file to PLINK format (Purcell et al. 2007).

3) Any SNPs with a minor allele frequency (MAF) less than 0.05 were removed from the datasets.

4) We estimated the population differentiation index, effective population size and divergence time between the two simulated populations following the same procedures used for the observed data and detailed above.

5) Estimators were calculated for each 1,000 independent replications.

Figure 4 shows the distribution of the 24,000 separation times between population_2a and population_2b and the observed divergence times between East Africans and Europeans, Australians and New Guineans.

Treemix

Using *TreeMix*, we inferred from genomic data a tree in which populations may exchange migrants after they have split from their common ancestor, thus violating the assumptions upon which simple bifurcating trees are built (Pickrell and Pritchard 2012). This method first infers a maximum-likelihood tree from genome-wide allele frequencies, and then identifies populations showing a poor fit to this tree model; migration events involving these populations are finally added. This way, each population may have multiple origins, and the contributions of each parental population provide an estimate of the fraction of alleles in the descendant population that originated in each parental population.

We applied the *TreeMix* model to the populations showing at least 30% of the green Admixture component at *k*=5 (Fig. 2) and clustering together in the third group of the DAPC scatter plot (Supplemental Fig. S4B). Allele frequencies for the *TreeMix* analysis were calculated by PLINK tool (Purcell et al. 2007), after pruning for linkage disequilibrium as we did for ADMIXTURE analysis. We modelled several scenarios allowing a number of migration events from 0 to 6, and stopping adding a migration when the following event did not increase significantly the variance explained by the model (Supplemental Fig. S8). The trees were forced to have a root in East Asia, and we used the window size of 500 (-k option).

Supplemental Fig. S9 shows the maximum-likelihood tree. Interestingly, the inferred migrations (indicated by arrows that are colored according to the intensities of the process) suggested an extensive genetic exchanges from East Asians to Southeast Asian populations.

Geographical patterns of dispersal

To obtain a realistic representation of migrational distances between populations, we did not simply estimate the shortest (great-circle) distances between sampling localities. Rather, we modeled resistance to gene flow, based on the landscape features known to influence human dispersal. We used a resistance method from the circuit theory implemented in the software Circuitscape v.3.5.2 (McRae 2006), starting from the landscape information in Oppenheimer (2012) and referring to the distribution of land masses at the last glacial maximum. Next, we added data about altitude and river presence from the Natural Earth database (http://www.naturalearthdata.com). Each area of the map was assigned a resistance value (*rv*) by the Reclassify tool in ArcGIS 10 (ESRI; Redlands, CA, USA), as follows: mountains higher than 2,000 m: *rv*=100; land or mountains below 2,000 m: *rv*=10; rivers: *rv*=5, oceans: NoData (absolute dispersal barrier); narrow arms of sea across which prehistoric migration is documented: *rv*=10. The low *rv* for rivers reflects the human tendency to follow, whenever possible, water bodies in their dispersal (see e.g. Beyin (2011)).

Under the SD model we hampered movement from Arabia to India (*rv*=100), hence preventing the dispersal along the Southern route; under the MD models, we created a buffer of low resistance value (*rv*=1) along the Southern route. For all models we then estimated least-resistance distances between the populations analyzed, when applicable going through Addis Ababa, chosen as a starting point for the African expansion (Ramachandran et al. 2005). The final output was then exported in Google Earth where geographic distances were expressed in kilometers.

Bibliography

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

Benazzo A, Panziera A, Bertorelle G. 2014. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecology and Evolution*.

Beyin A. 2011. Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate. *Int J Evol Biol* 2011: 615094.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 15: 1496-1502.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156-2158.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* 13: 635-643.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet* 38: 226-231.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 10: 639-650.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337-338.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23: 1801-1806.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27: 3070-3071.

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.

Liu N, Zhao H. 2006. A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics* 2: 353-364.

Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4: e7888.

McRae BH. 2006. Isolation by resistance. *Evolution* 60: 1551-1561.

McVean GA. 2002. A genealogical interpretation of linkage disequilibrium. *Genetics* 162: 987-991.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338: 222-226.

Mezzavilla M, Ghirotto S. 2015. Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS. *J Comput Sci Syst Biol* 037-004.

Oppenheimer S. 2012. A single southern exit of modern humans from Africa: Before or after Toba? . *Quaternary International* 258: 88-89.

Ohta T, Kimura M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* 63: 229-238.

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

Pugach I, Delfin F, Gunnarsdottir E, Kayser M, Stoneking M. 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc Natl Acad Sci U S A* 110: 1803-1808.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-575.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. Vienna, Austria : the R Foundation for Statistical Computing. *ISBN: 3-900051-07-0 Available online at http://wwwR-projectorg/*.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 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: 15942-15947.

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89: 516-528.

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

Relethford JH. 1985. Examination of the relationship between inbreeding and population size. *J Biosoc Sci* 17: 97-106.

Rosenberg NA. 2004. Distruct: a program for the graphical display of population *Molecular Ecology Notes* 4: 137-138.

Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* 2: 125-141.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17: 520-526.

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

Xing J, Watkins WS, Shlien A, Walker E, Huff CD, Witherspoon DJ, Zhang Y, Simonson TS, Weiss RB, Schiffman JD et al. 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 96: 199-210.

Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, Thara R, Mowry BJ, Bulayeva K, Weiss RB, Jorde LB. 2009. Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res* 19: 815-825.

Supplemental Table S1 General information about the 24 metapopulations analyzed.

Group	Population	Sample_Size	Dataset
South Africa(37)	Pedi	9	Xing et al. 2009
	Nguni	9	Xing et al. 2009
	Sotho/Tswana	7	Xing et al. 2009
	!Kung	12	Xing et al. 2009
East Africa(46)	Luhya	22	Xing et al. 2009
	Alur	10	Xing et al. 2009
	Hema	14	Xing et al. 2009
West Africa(160)	Yoruba	111	Reich et al. 2011
	Dogon	24	Xing et al. 2010
	Bambara	25	Xing et al. 2010
Europe(166)	CEU	111	Reich et al. 2011
	Tuscan	25	Xing et al. 2009
	French	5	Lopez Herraez et al. 2009
	Slovenian	25	Xing et al. 2010
Caucasus(42)	Urkarah	18	Xing et al. 2009
	Kurd	24	Xing et al. 2009
West Asia(38)	Pakistani	23	Xing et al. 2010
	Balochi	2	Lopez Herraez et al. 2009
	Makrani	4	Lopez Herraez et al. 2009
	Sindhi	4	Lopez Herraez et al. 2009
	Pathan	5	Lopez Herraez et al. 2009
Central Asia(23)	Kyrgyzstani	23	Xing et al. 2010
North India(25)	Bhil	7	Reich et al. 2009
	Meghawal	5	Reich et al. 2009
	Sahariya	4	Reich et al. 2009
	Satnami	4	Reich et al. 2009
	Lodi	5	Reich et al. 2009
South India(98)	Dravidian	11	Pugach et al. 2013
	Chenchu	6	Reich et al. 2009
	Kurumba	9	Reich et al. 2009
	Hallaki	6	Reich et al. 2009
	Kamsali	3	Reich et al. 2009
	Madiga	4	Reich et al. 2009
	Mala	3	Reich et al. 2009
	T.N. Dalit	13	Xing et al. 2009
	Irula	22	Xing et al. 2009
	A.P. Mala	11	Xing et al. 2009
	A.P. Madiga	10	Xing et al. 2009
East Asia(188)	Japan	86	Reich et al. 2011
	Han Chinese	88	Reich et al. 2011
	Miaozu	5	Lopez Herraez et al. 2009
	Tujia	5	Lopez Herraez et al. 2009
	Yizu	4	Lopez Herraez et al. 2009
South Asia(35)	Khmer	5	Xing et al. 2009
	Thai	20	Xing et al. 2010
	Cambodian	3	Lopez Herraez et al. 2009
	Vietnamese	7	Xing et al. 2009
Malaysia(10)	Temuan	10	Reich et al. 2011
Borneo(73)	Land Dayak	15	Pugach et al. 2013
	Barito River	23	Reich et al. 2011
	Bidayuh	10	Reich et al. 2011
	Iban	25	Xing et al. 2009
Sumatra(20)	Besemah	10	Pugach et al. 2013
	Semede	10	Pugach et al. 2013
East indonesia(10)	Flores	1	Pugach et al. 2013
	Roti	4	Pugach et al. 2013
	Timor	3	Pugach et al. 2013
	Alor	2	Pugach et al. 2013
Philippine(16)	Manobo	16	Pugach et al. 2013
Moluccas(10)	Hiri	7	Reich et al. 2011
	Ternate	3	Reich et al. 2011
Australian(10)		10	Pugach et al. 2013
New Guinea(27)	Papua New Guinea	24	Pugach et al. 2013
	Papuan	3	Lopez Herraez et al. 2009
Fiji(24)		24	Reich et al. 2011
Polinesia(44)	Polynesia	19	Reich et al. 2011
	Samoan	13	Xing et al. 2010
	Tongan	12	Xing et al. 2010
Onge(9)	Negrito	9	Reich et al. 2011
Jehai(8)	Negrito	8	Reich et al. 2011
Mamanwa(11)	Negrito	11	Pugach et al. 2013

Supplemental Table S2 Pairwise F_{ST} values estimated between populations. The matrix is symmetrical.

FST	South_Africa	East_Africa	West_Africa	Europe	Caucasus	West_Asia	Central_Asia	North_India	South_India	East_Asia	South_Asia	Malaysia	Borneo	Sumatra	East_Indonesia	Philippine	Moluccas	Australia	New_Guinea	Fiji	Polynesia	Onge	Jehai	Mamanwa
South_Africa	0.00	0.02	0.02	0.16	0.15	0.14	0.15	0.14	0.14	0.18	0.17	0.18	0.18	0.19	0.17	0.19	0.18	0.22	0.25	0.19	0.20	0.23	0.19	0.20
East_Africa	0.02	0.00	0.01	0.13	0.12	0.12	0.13	0.12	0.12	0.16	0.15	0.15	0.16	0.16	0.15	0.16	0.16	0.19	0.22	0.16	0.18	0.20	0.17	0.17
West_Africa	0.02	0.01	0.00	0.15	0.15	0.14	0.15	0.14	0.14	0.18	0.17	0.17	0.17	0.18	0.17	0.18	0.17	0.21	0.23	0.18	0.19	0.22	0.19	0.19
Europe	0.16	0.13	0.15	0.00	0.01	0.02	0.05	0.04	0.05	0.11	0.09	0.10	0.11	0.11	0.10	0.11	0.11	0.15	0.18	0.12	0.13	0.15	0.12	0.13
Caucasus	0.15	0.12	0.15	0.01	0.00	0.01	0.05	0.03	0.04	0.10	0.08	0.09	0.10	0.10	0.09	0.10	0.10	0.15	0.18	0.11	0.12	0.15	0.11	0.12
West_Asia	0.14	0.12	0.14	0.02	0.01	0.00	0.03	0.01	0.01	0.08	0.06	0.07	0.08	0.08	0.07	0.08	0.08	0.13	0.16	0.09	0.10	0.13	0.09	0.10
Central_Asia	0.15	0.13	0.15	0.05	0.05	0.03	0.00	0.03	0.04	0.02	0.02	0.04	0.03	0.04	0.04	0.04	0.04	0.12	0.16	0.07	0.07	0.12	0.07	0.07
North_India	0.14	0.12	0.14	0.04	0.03	0.01	0.03	0.00	0.00	0.07	0.05	0.06	0.06	0.07	0.06	0.07	0.06	0.11	0.15	0.08	0.09	0.11	0.08	0.09
South_India	0.14	0.12	0.14	0.05	0.04	0.01	0.04	0.00	0.00	0.07	0.05	0.06	0.06	0.07	0.06	0.07	0.06	0.11	0.14	0.08	0.09	0.11	0.08	0.08
East_Asia	0.18	0.16	0.18	0.11	0.10	0.08	0.02	0.07	0.07	0.00	0.01	0.03	0.02	0.02	0.04	0.03	0.04	0.14	0.16	0.07	0.06	0.13	0.07	0.06
South_Asia	0.17	0.15	0.17	0.09	0.08	0.06	0.02	0.05	0.05	0.01	0.00	0.02	0.01	0.01	0.03	0.02	0.03	0.13	0.16	0.06	0.05	0.12	0.05	0.05
Malaysia	0.18	0.15	0.17	0.10	0.09	0.07	0.04	0.06	0.06	0.03	0.02	0.00	0.02	0.03	0.04	0.03	0.04	0.14	0.18	0.07	0.06	0.14	0.06	0.07
Borneo	0.18	0.16	0.17	0.11	0.10	0.08	0.03	0.06	0.06	0.02	0.01	0.02	0.00	0.01	0.03	0.02	0.03	0.13	0.16	0.07	0.05	0.13	0.06	0.06
Sumatra	0.19	0.16	0.18	0.11	0.10	0.08	0.04	0.07	0.07	0.02	0.01	0.03	0.01	0.00	0.03	0.02	0.03	0.15	0.18	0.07	0.05	0.14	0.07	0.06
East_Indonesia	0.17	0.15	0.17	0.10	0.09	0.07	0.04	0.06	0.06	0.04	0.03	0.04	0.03	0.03	0.00	0.03	0.01	0.08	0.10	0.03	0.04	0.14	0.07	0.06
Philippine	0.19	0.16	0.18	0.11	0.10	0.08	0.04	0.07	0.07	0.03	0.02	0.03	0.02	0.02	0.03	0.00	0.03	0.14	0.18	0.07	0.05	0.15	0.08	0.05
Moluccas	0.18	0.16	0.17	0.11	0.10	0.08	0.04	0.06	0.06	0.04	0.03	0.04	0.03	0.03	0.01	0.03	0.00	0.10	0.11	0.03	0.04	0.14	0.08	0.06
Australia	0.22	0.19	0.21	0.15	0.15	0.13	0.12	0.11	0.11	0.14	0.13	0.14	0.13	0.15	0.08	0.14	0.10	0.00	0.08	0.07	0.13	0.20	0.16	0.15
New_Guinea	0.25	0.22	0.23	0.18	0.18	0.16	0.16	0.15	0.14	0.16	0.16	0.18	0.16	0.18	0.10	0.18	0.11	0.08	0.00	0.07	0.15	0.23	0.20	0.18
Fiji	0.19	0.16	0.18	0.12	0.11	0.09	0.07	0.08	0.08	0.07	0.06	0.07	0.07	0.07	0.03	0.07	0.03	0.07	0.07	0.00	0.03	0.15	0.10	0.09
Polynesia	0.20	0.18	0.19	0.13	0.12	0.10	0.07	0.09	0.09	0.06	0.05	0.06	0.05	0.05	0.04	0.05	0.04	0.13	0.15	0.03	0.00	0.16	0.10	0.08
Onge	0.23	0.20	0.22	0.15	0.15	0.13	0.12	0.11	0.11	0.13	0.12	0.14	0.13	0.14	0.14	0.15	0.14	0.20	0.23	0.15	0.16	0.00	0.16	0.17
Jehai	0.19	0.17	0.19	0.12	0.11	0.09	0.07	0.08	0.08	0.07	0.05	0.06	0.06	0.07	0.07	0.08	0.08	0.16	0.20	0.10	0.10	0.16	0.00	0.10
Mamanwa	0.20	0.17	0.19	0.13	0.12	0.10	0.07	0.09	0.08	0.06	0.05	0.07	0.06	0.06	0.06	0.05	0.06	0.15	0.18	0.09	0.08	0.17	0.10	0.00

Supplemental Table S3 Estimates of population divergence times, using (A) r^2 and (B) σ^2 statistics, as estimator of LD level.

A

TIME	South_Africa			East_Africa			West_Africa		
	0.05	0.50	0.95	0.05	0.50	0.95	0.05	0.50	0.95
Europe	66,865	74,595	81,894	63,135	69,736	76,645	74,138	81,354	88,626
Caucasus	68,014	74,933	81,204	62,363	68,143	74,016	74,801	81,228	87,478
West_Asia	66,495	73,546	81,188	60,458	66,318	73,247	73,105	79,715	87,337
Central_Asia	71,064	76,762	85,188	66,166	71,021	78,819	78,359	83,596	91,992
North_India	72,839	77,974	86,111	65,930	70,230	77,595	79,133	83,829	91,904
South_India	65,283	69,703	76,457	60,625	64,396	70,718	72,139	76,166	82,932
East_Asia	82,978	89,638	98,162	81,456	87,432	95,874	91,378	97,477	105,966
South_Asia	77,802	85,002	93,417	74,310	80,587	88,651	85,615	92,279	100,655
Malaysia	69,078	74,544	83,461	66,862	71,622	80,114	76,572	81,433	90,182
Borneo	77,166	82,244	90,569	75,517	80,056	88,234	85,475	90,029	98,318
Sumatra	77,742	84,720	93,621	75,884	82,043	90,707	85,877	92,222	101,013
East_Indonesia	69,028	74,492	82,788	66,801	71,576	79,538	76,719	81,629	89,830
Philippine	75,675	81,547	90,458	74,051	79,248	87,916	83,790	89,053	97,846
Moluccas	68,071	73,760	81,899	66,571	71,562	79,457	75,861	80,951	88,970
Australia	**88,134**	**97,946**	**109,839**	**87,828**	**96,599**	**108,214**	**96,568**	**105,461**	**117,025**
New_Guinea	**98,298**	**106,353**	**118,689**	**99,852**	**107,204**	**119,569**	**105,237**	**112,166**	**123,850**
Fiji	72,170	78,848	85,832	71,465	77,395	84,437	80,283	86,315	93,221
Polynesia	70,968	77,397	86,418	71,753	77,531	86,510	79,386	85,110	93,950
Onge	75,929	81,853	90,374	77,243	82,572	91,234	84,806	89,904	98,179
Jehai	66,745	72,532	81,126	66,414	71,521	79,922	74,984	80,104	88,529
Mamanwa	67,854	73,580	83,338	67,925	73,012	82,492	76,151	81,200	90,741

B

TIME	South_Africa			East_Africa			West_Africa		
	5	50	95	5	50	95	5	50	95
Europe	59,272	68,157	74,614	57,084	65,402	70,830	66,508	75,174	82,892
Caucasian	60,536	67,687	74,502	56,531	63,196	68,797	67,303	74,259	82,260
West_Asia	59,685	66,622	73,309	55,174	61,586	67,061	66,240	73,021	80,843
Central_Asia	63,899	69,991	76,739	60,510	66,425	72,061	71,134	77,072	85,081
North_India	65,580	70,732	78,142	60,270	65,243	71,325	71,857	76,857	85,375
South_India	58,715	64,075	70,532	55,445	60,664	66,051	65,495	70,736	78,358
East_Asia	73,798	81,274	87,923	73,793	81,398	87,241	82,165	89,435	97,575
South_Asia	69,438	76,417	84,626	67,515	74,425	81,458	77,213	84,003	93,562
Malaysia	61,829	67,443	74,841	61,092	66,852	73,143	69,356	74,724	83,462
Borneo	68,824	74,485	82,767	68,644	74,579	81,799	77,089	82,578	92,290
Sumatra	69,627	76,307	84,198	69,281	76,108	82,934	77,768	84,202	93,540
East_Indonesia	62,028	67,758	74,901	61,208	67,056	73,154	69,703	75,223	83,719
Philippine	67,791	73,921	82,298	67,651	73,996	81,242	75,911	81,805	91,607
Moluccas	60,818	66,794	74,031	60,753	66,875	73,078	68,619	74,353	82,968
Australia	**78,451**	**88,127**	**98,594**	**79,827**	**89,596**	**98,794**	**87,027**	**96,234**	**108,224**
New_Guinea	**87,360**	**95,573**	**107,941**	**90,693**	**99,499**	**110,530**	**94,748**	**102,325**	**116,058**
Fiji	64,875	71,258	78,000	65,546	72,155	78,014	72,991	79,118	87,323
Polynesia	63,401	69,935	77,564	65,566	72,451	79,150	71,859	78,090	87,223
Onge	67,987	74,305	82,275	70,827	77,670	84,637	76,981	82,918	92,535
Jehai	59,741	65,492	73,336	60,820	66,859	73,613	68,006	73,485	82,772
Mamanwa	60,831	66,721	75,584	62,288	68,502	76,183	69,158	74,765	85,052

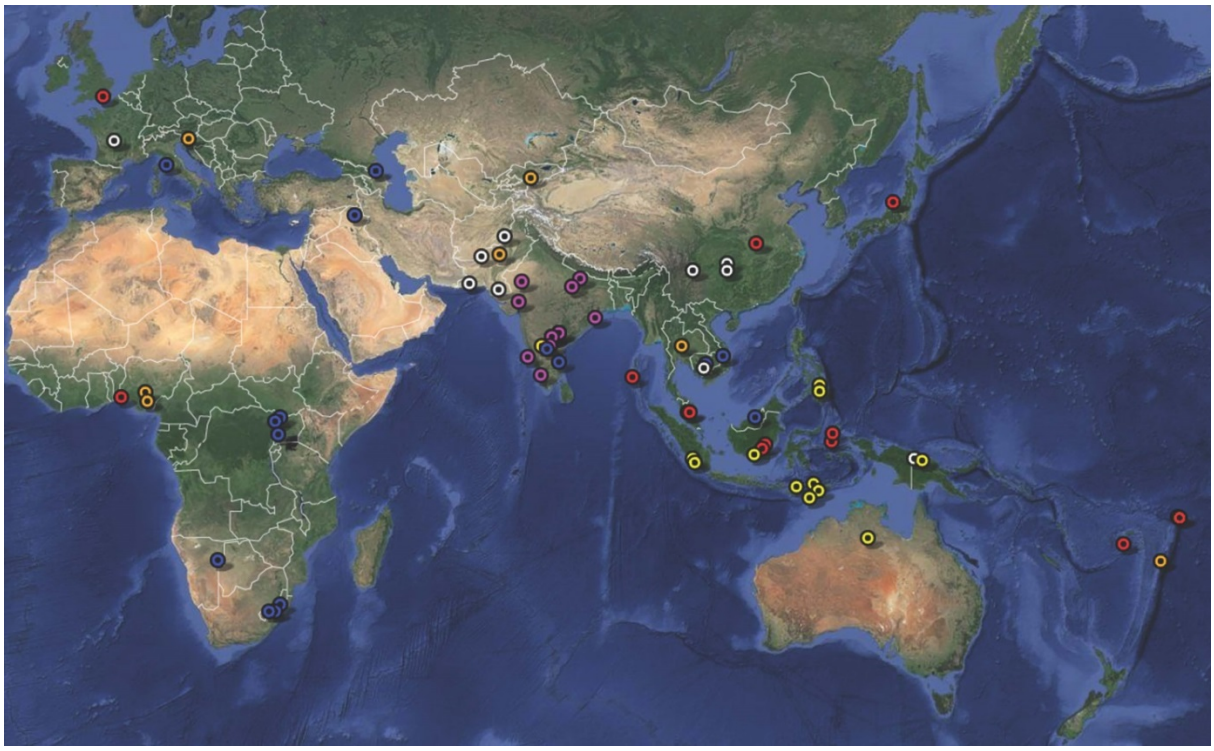
For each comparisons with one African metapopulation, the three columns report the 95% lower confidence limit, the point estimate (in years, assuming a generation interval =25 years), and the 95% upper confidence limit.

Supplemental Table S4 Estimates of population divergence times

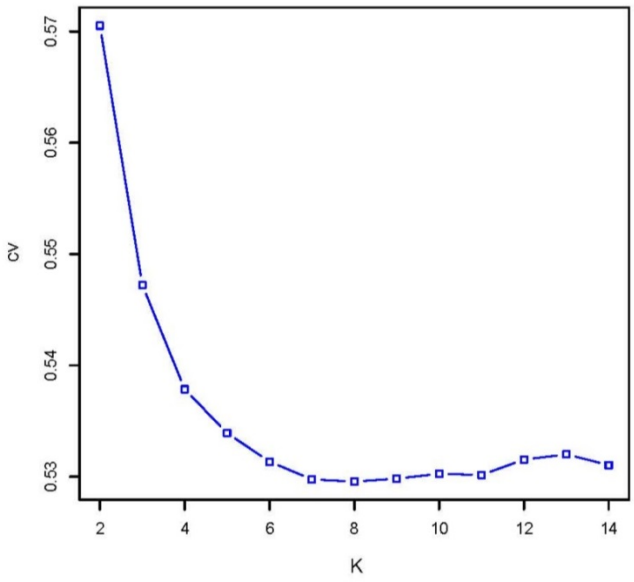
TIME	South_Africa			East_Africa			West_Africa		
	0.05	0.50	0.95	0.05	0.50	0.95	0.05	0.50	0.95
Europe	67,080	74,835	82,158	63,365	69,990	76,924	74,397	81,638	88,935
Caucasus	68,157	75,091	81,375	62,465	68,254	74,136	74,978	81,420	87,684
West_Asia	66,751	73,829	81,501	60,728	66,614	73,575	73,378	80,013	87,664
Central_Asia	71,165	76,871	85,308	66,334	71,201	79,020	78,516	83,764	92,176
North_India	73,170	78,328	86,503	66,226	70,545	77,943	79,439	84,153	92,258
South_India	65,513	69,948	76,727	60,831	64,615	70,959	72,374	76,414	83,202
East_Asia	83,335	90,024	98,585	81,812	87,813	96,292	91,761	97,885	106,410
South_Asia	78,051	85,275	93,717	74,613	80,916	89,013	85,905	92,591	100,996
Malaysia	69,262	74,742	83,683	67,069	71,844	80,363	76,771	81,646	90,418
Borneo	77,392	82,485	90,835	75,769	80,324	88,529	85,708	90,274	98,587
Sumatra	78,058	85,065	94,003	76,230	82,418	91,122	86,230	92,600	101,427
East_Indonesia	69,173	74,649	82,962	66,995	71,783	79,768	76,958	81,884	90,111
Philippine	75,957	81,850	90,794	74,435	79,658	88,371	84,170	89,457	98,290
Moluccas	68,332	74,043	82,212	66,836	71,847	79,773	76,225	81,340	89,398
Australia	**88,148**	**97,961**	**109,856**	**87,869**	**96,644**	**108,264**	**96,704**	**105,609**	**117,190**
New_Guinea	**98,404**	**106,468**	**118,818**	**99,944**	**107,302**	**119,679**	**105,518**	**112,466**	**124,180**
Fiji	72,312	79,003	86,000	71,636	77,580	84,639	80,484	86,531	93,454
Polynesia	71,180	77,628	86,677	71,972	77,768	86,774	79,621	85,362	94,228
Onge	76,202	82,147	90,699	77,532	82,881	91,576	85,091	90,206	98,508
Jehai	66,938	72,742	81,361	66,638	71,763	80,192	75,167	80,300	88,745
Mamanwa	67,928	73,660	83,429	68,034	73,129	82,624	76,254	81,310	90,864

Population divergence time estimated on a subset of SNPs chosen to exclude the effect of an archaic introgression from Denisovan.

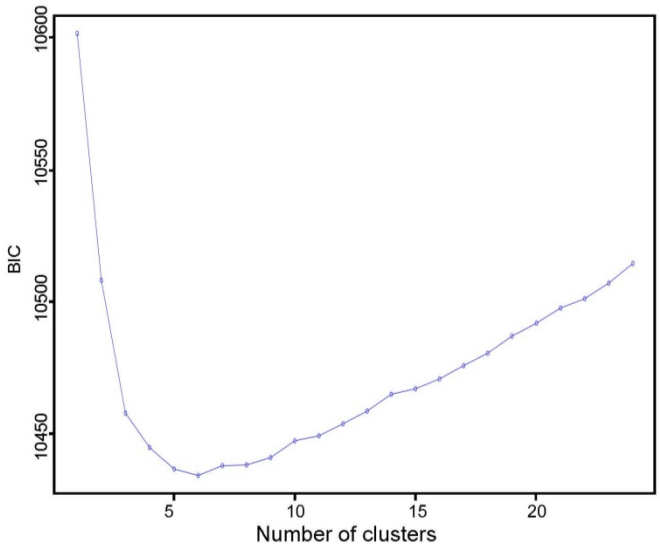
Supplemental Fig. 1 Geographic location of all the 71 populations analyzed, the different dataset we use are represented by different colors and are detailed in Supplemental Table S1.



Supplemental Fig. 2 Estimation of the most likely number of clusters in the data (X-axis) as a function of the cross-validation error observed in the attempted assignments (Y-axis).



Supplemental Fig. 3 Inference of the most likely number of clusters in the DAPC. A K value of 6 (the lowest BIC value) represents the best summary of the data.

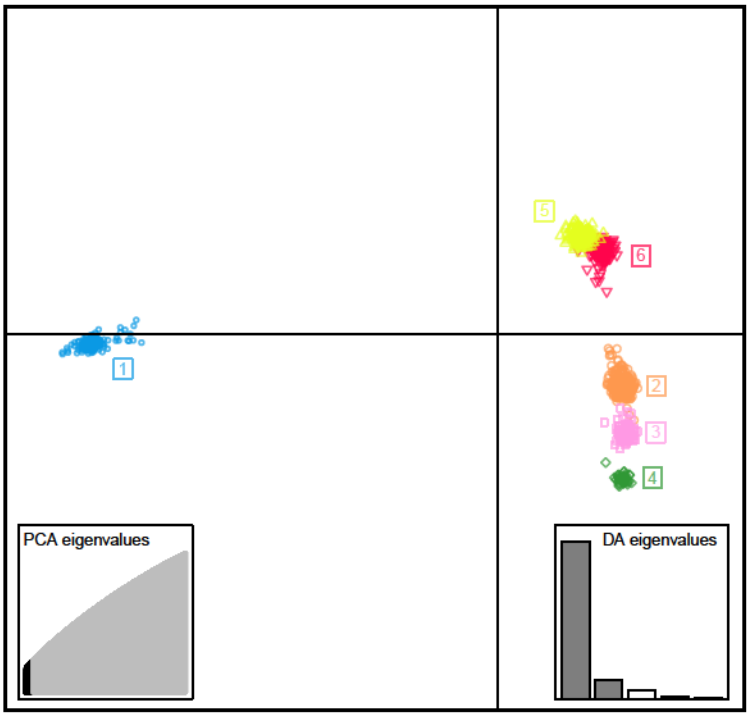


Supplemental Fig. 4 Discriminant Analysis of Principal Components, *DAPC*. (A) classification of individual genotypes; for each row (each population) the figures refer to the numbers of individuals assigned to of the *K*=6 clusters, each cluster associated with a different colour; (B) scatterplot along the first two axes; each symbol corresponds to an individual genotypes; in the insets, the fraction of Principal Components retained in the analysis (left) and the fraction of the overall variance attributed to the first five eigenvalues, with the first two columns, in grey, representing the first two Discriminant Functions (right).

A

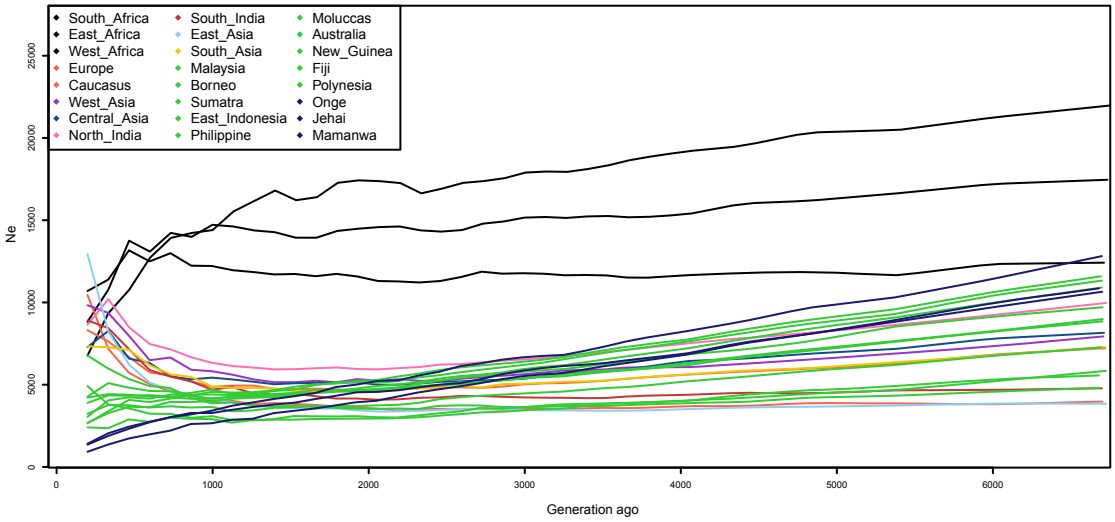
	1	2	3	4	5	6
Australia	0	0	0	10	0	0
Borneo	0	73	0	0	0	0
Caucasus	0	0	0	0	42	0
Central_Asia	0	23	0	0	0	0
East_Africa	46	0	0	0	0	0
East_Asia	0	188	0	0	0	0
East_Indonesia	0	3	7	0	0	0
Europe	0	0	0	0	166	0
Fiji	0	0	24	0	0	0
Jehai	0	8	0	0	0	0
Malaysia	0	10	0	0	0	0
Mamanwa	0	11	0	0	0	0
Moluccas	0	0	10	0	0	0
New_Guinea	0	0	0	27	0	0
North_India	0	0	0	0	0	25
Onge	0	0	0	0	0	9
Philippine	0	16	0	0	0	0
Polynesia	0	0	44	0	0	0
South_Africa	37	0	0	0	0	0
South_Asia	0	35	0	0	0	0
South_India	0	0	0	0	0	98
Sumatra	0	20	0	0	0	0
West_Africa	160	0	0	0	0	0
West_Asia	0	0	0	0	8	30

B

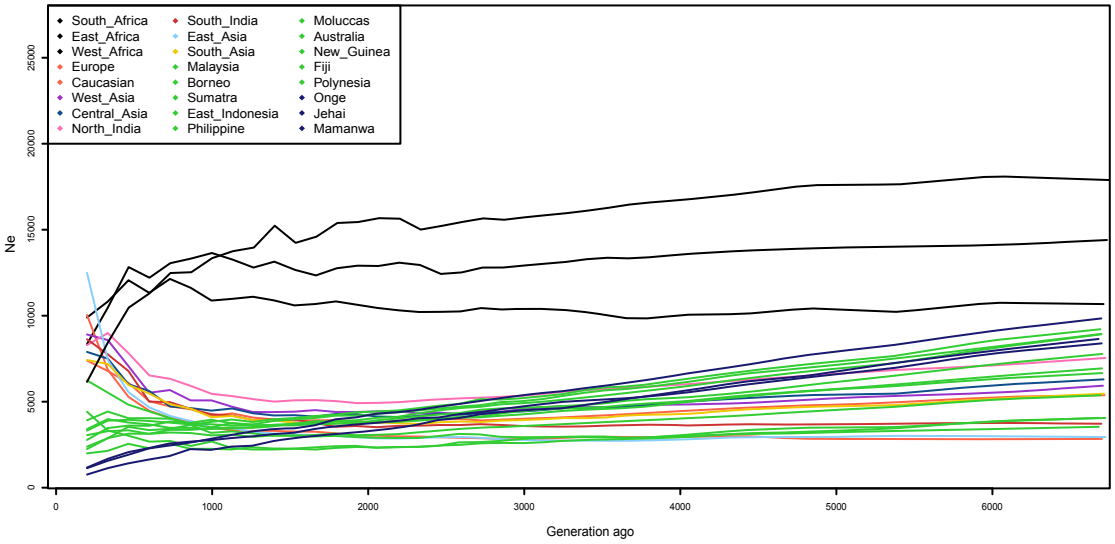


Supplemental Fig. 5 Estimates of N_e from measures of linkage disequilibrium, using the (A) r^2 and(B) σ^2 statistics as estimator of LD level. Time is on the X-axis and is expressed in generations from the present. Very recent estimates have been omitted because not reliably estimated (see McVean (2002)).

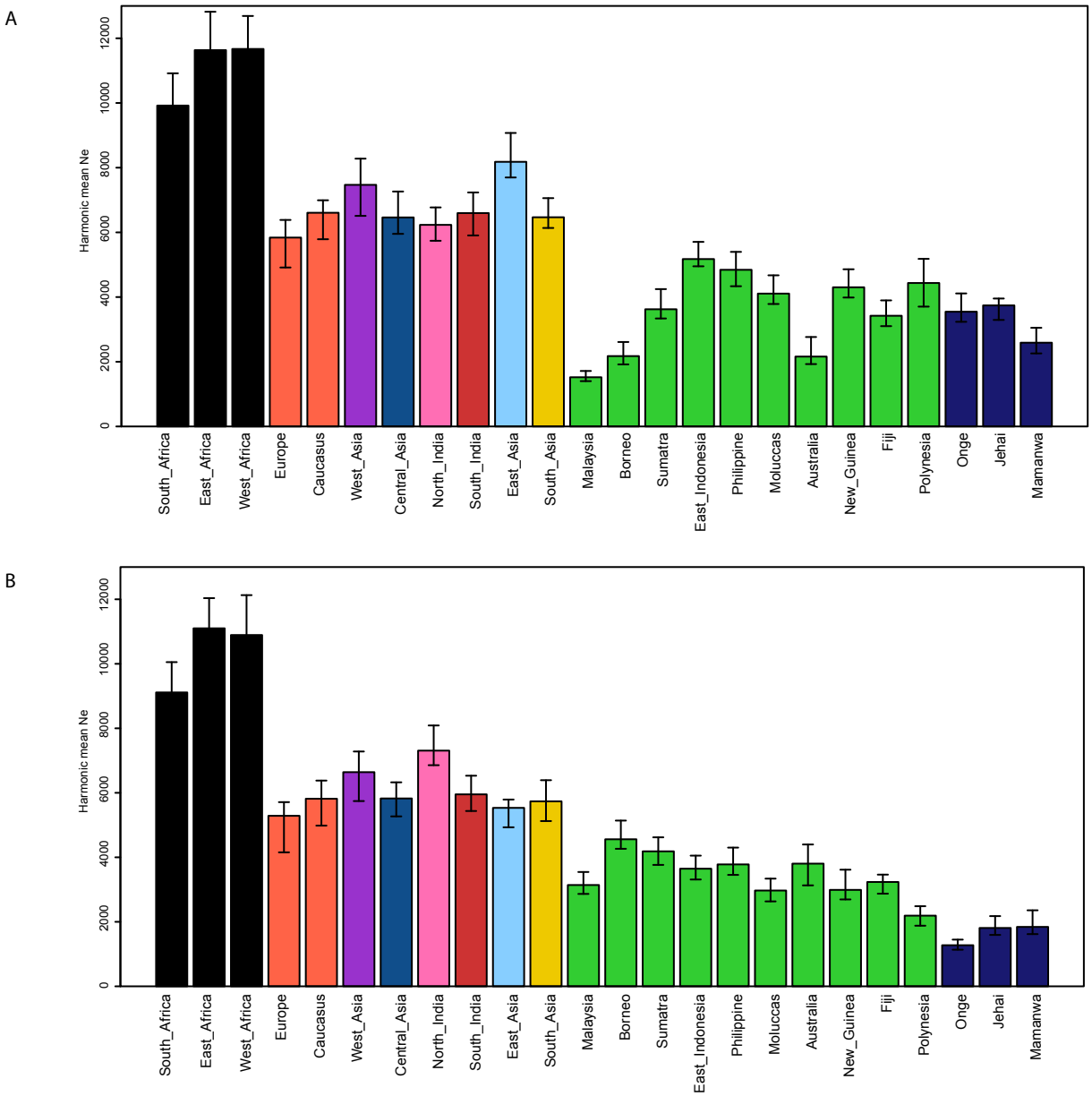
A



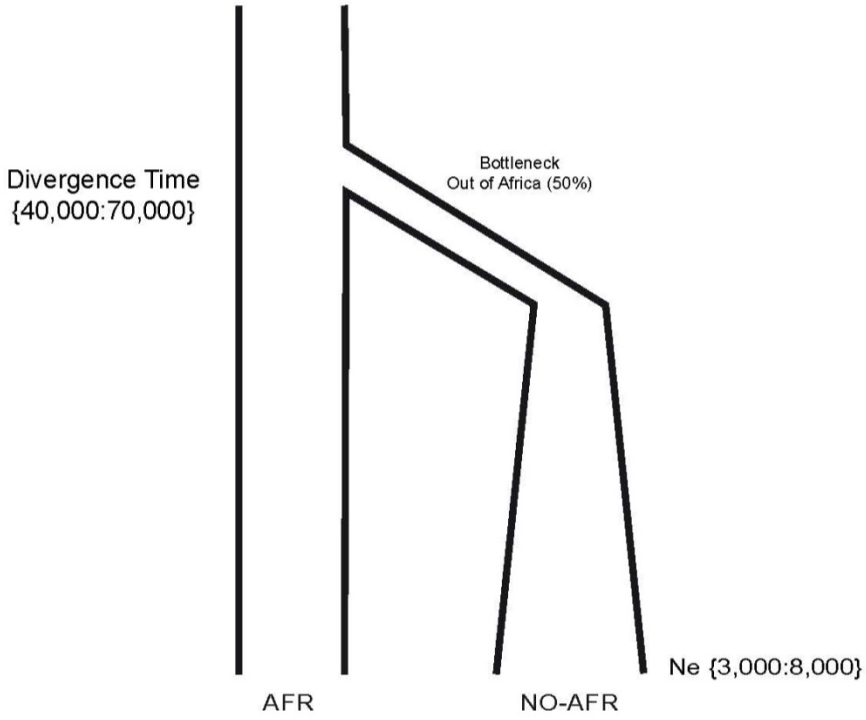
B



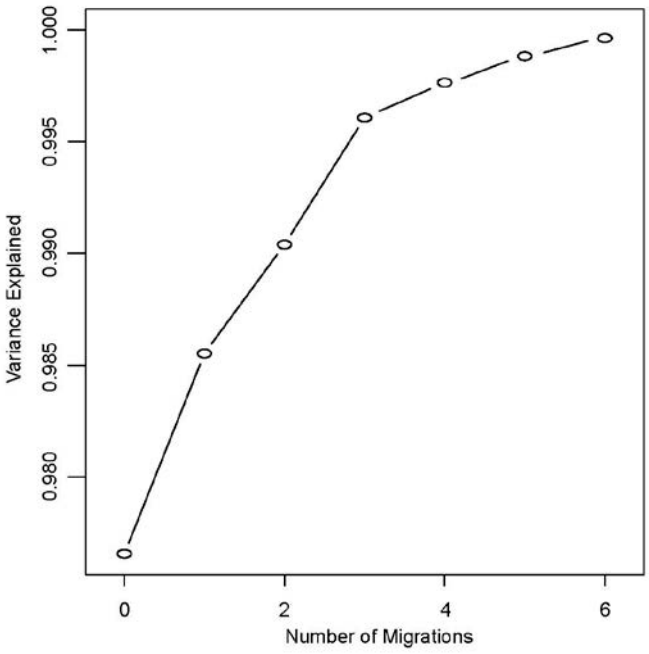
Supplemental Fig. 6 Harmonic means of the estimated population effective sizes (N_e), using the (A) r_2 and (B) σ_2 statistics as estimator of LD level. Vertical bars represent empirical 90% confidence estimates



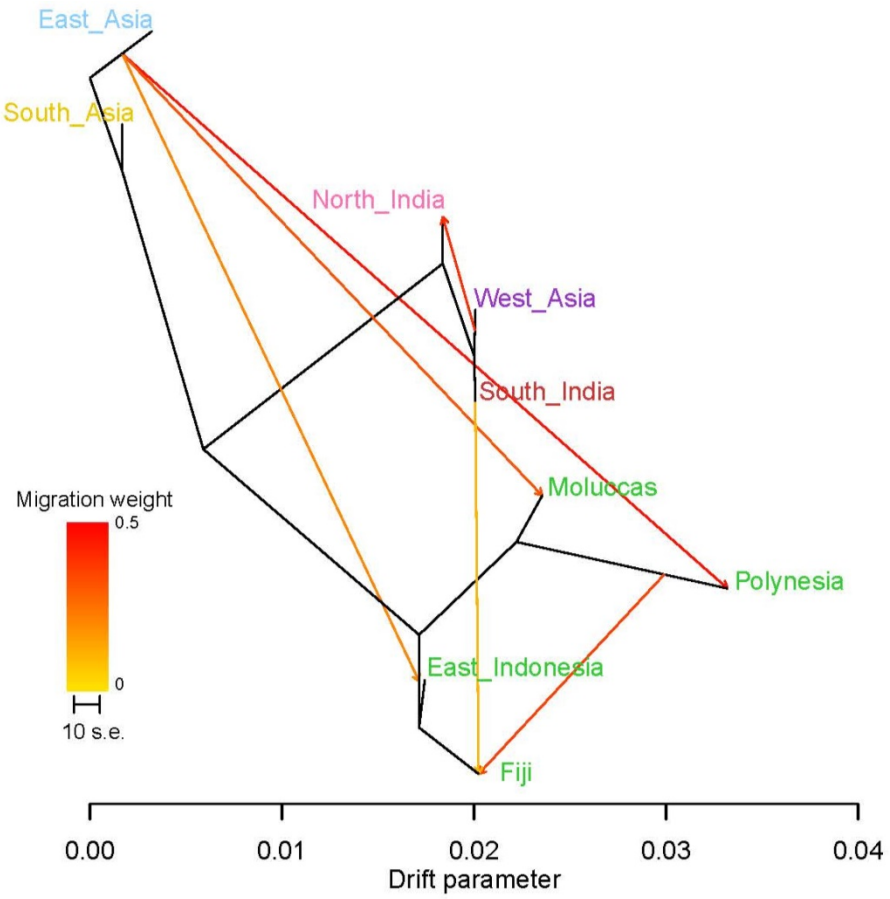
Supplemental Fig. 7 Representation of the human demographic model tested by *ms*. The past is at the top, the present is at the bottom.



Supplemental Fig. 8 Fractions of the total variance explained by the model at increasing numbers of migrations superimposed to the bifurcating tree in the *TreeMix* analysis.



Supplemental Fig. 9 Population relationships inferred by *TreeMix*. The Maximum-likelihood tree is in black; branch lengths are proportional to the impact of genetic drift, which may or may not faithfully represent separation times between populations. The inferred migration events are represented by arrows pointing from the putative source to the putative target populations, with colours of the arrows representing the relative weight of the genetic exchanges, according to the heat scale on the left.



PAPER III: Across language families: Genome diversity mirrors linguistic variation within Europe.

American Journal of Physical Anthropology

Across language families: DNA diversity mirrors grammar within Europe

Journal:	*American Journal of Physical Anthropology*
Manuscript ID:	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Longobardi, Giuseppe; University of York, Language and Linguistic Science; University of Trieste, Humanities Ghirotto, Silvia; University of Ferrara, Life Sciences and Biotechnology Guardiano, Cristina; Università di Modena e Reggio Emilia, Communication and Economics Tassi, Francesca; University of Ferrara, Life Sciences and Biotechnology Benazzo, Andrea; University of Ferrara, Life Sciences and Biotechnology Ceolin, Andrea; University of Trieste, Humanities; University of York, Language and Linguistic Science Barbujani, Guido; University of Ferrara, Life Sciences and Biotechnology
Key Words:	Parametric Comparison Method, genome-wide diversity, single-nucleotide polymorphisms, human evolutionary history
Subfield: Please select your first choice in the first field.:	Genetics [primate and human], Human biology [living humans; behavior, ecology, physiology, anatomy]

John Wiley & Sons, Inc.

7.1.2015 Research article

Across Language Families: Genome Diversity Mirrors Linguistic Variation within Europe

Giuseppe Longobardi[1,2], Silvia Ghirotto[3], Cristina Guardiano[4], Francesca Tassi[3], Andrea Benazzo[3], Andrea Ceolin[1] and Guido Barbujani[3] *

[1]*Department of Language and Linguistic Science, University of York, UK*
[2]*Department of Humanities, University of Trieste, Italy*
[3]*Department of Life Sciences and Biotechnology, University of Ferrara, Italy*
[4]*Department of Communication and Economics, University of Modena-Reggio Emilia, Italy.*

15 text pages, plus 7 pages of bibliography, 5 figures and 3 tables

Abbreviated title: Genome diversity across language families

KEYWORDS Parametric Comparison Method; genome-wide diversity; single-nucleotide polymorphisms; human evolutionary history

* Correspondence to: Guido Barbujani
Department of Life Sciences and Biotechnology, University of Ferrara
Via Borsari 46
I-44121 Ferrara, Italy
E-mail: g.barbujani@unife.it
Phone: +39 0532 455312

Grant sponsors: European Research Council ERC-2011-AdvG_295733 grant (Langelin) to GL and GB; Italian Ministry for Research and Universities (MIUR) PRIN 2010-2011 to GB; York Centre for Linguistic History and Diversity to AC.

ABSTRACT Comparing genetic and linguistic diversity may cast light on both demographic history and cultural transmission. However, classical studies were hampered, on the linguistic side, by insufficient reliance on quantitative tools and by the impossibility to compare the vocabularies of distantly-related languages. Here, we take advantage of two new tools recently proposed in comparative linguistics: first, a refined list of Indo-European cognate words for quantitative experiments, then a novel method of language comparison based on syntactic features. Since the latter method estimates linguistic diversity from a universal inventory of grammatical polymorphisms, it enables comparison even across different language families. On these grounds, by comparing a broad genome-wide SNP dataset in 15 European populations, we observed significant correlations between genomic and linguistic diversity, the latter inferred from data on both Indo-European and non-Indo-European languages. Contrary to previous observations, on the European scale, language proved a better predictor of genomic differences than geography, and inferred episodes of genetic admixture following the main population splits found convincing correlates also in the syntactic realm, supporting the relevance of a synthesis approach to cultural and biological evolution. These results pave the ground for previously unfeasible cross-disciplinary analyses at the worldwide scale, encompassing populations of different language families.

Why are humans biologically different, and how did they come to speak different languages? Taken separately, these questions have certainly been faced for millennia now, but it was Charles Darwin (1859) who explicitly put forth the idea of a parallelism between biological evolution and language diversification; Darwin foresaw that a perfect pedigree of human populations would also represent the best possible phylogenetic tree of the world's languages. Indeed, factors isolating populations from each other (such as barriers to migration, or just distance) are expected to promote both biological and cultural divergence, and factors facilitating contacts should have the opposite effect; but gene/language parallelisms might in fact be deeper than that, in many respects, and recently some scholars went as far as claiming a role even for adaptation, not only in biological evolution, but in some linguistic changes as well (Levinson and Gray, 2012).

Darwin's evolutionary framework was immediately accepted by linguists such as Schleicher (1863); however, it took more than a century for his parallelism intuition to be tried against actual data (Sokal, 1988; Cavalli-Sforza et al., 1988), and to become part of a broader research program (Renfrew, 1987, Cavalli Sforza et al. 1994). The idea is that linguistic diversification caused by demographic processes, mainly population dispersal, would generate parallel patterns of genetic and linguistic variation. That would often be the rule, but where linguistic change is not accompanied by demographic change (e.g., when a small group imposes its language upon a larger population through a process of élite dominance: Renfrew 1992), a local mismatch would arise between genetic and linguistic diversity. As a consequence, one could infer from that exception to the rule the occurrence of an important event of language replacement.

The results of the line of studies above were illuminating on the one hand, but controversial on the other. The case for analogies between linguistic and genetic variation, both in empirical fact (Barbujani and Sokal, 1990; Barbujani and Pilastro, 1993; Cavalli-Sforza et al., 1994; Sajantila et al. 1995; Poloni et al., 1997; Belle and Barbujani, 2007) and in methods (Ringe et al. 2002; Gray and Atkinson, 2003; McMahon and McMahon, 2003; Heggarty, 2004; Gray et al. 2009; Bouckaert et al., 2012; Berwick et al. 2013), has clearly emerged at a regional level; instead, at the larger geographical scale, many such results were received with skepticism, especially on the linguistic side.

One weakness of early approaches was in fact the unavailability of numerical taxonomies of languages to be matched with the biological ones: classical methods have produced impressive demonstrations of absolute relatedness for words and languages, but hardly provided quantitative measures of cognacy even for the internal articulation of well acknowledged families. An even deeper reason for skepticism was that solid linguistic relationships have so far been inferred from comparing vocabulary items (words/morphemes) and their sound structures; now, formally identifiable correspondences of such items in sound and meaning (i.e. chance-safe etymologies) are known to dissolve with time, while accidental similarities tend to emerge, due to the combination of arbitrariness of lexical variation with general constraints on possible phonological systems. Therefore, although the time depth at which these processes disrupt the potential for long-range linguistic classification is far from established (Greenhill et al., 2010, Nichols, 1996), it has been anyway impossible to reliably infer distant (across evident families) relationships from lexical comparisons. As a consequence, large-scale gene-language comparisons had to resort to – and were undermined by – ill-proven classifications of languages, often resulting from scientifically unsupported taxonomic procedures (Bolnick et al., 2004; Greenhill, 2011; Ringe, 1996; Ringe and Eska, 2013).

To overcome such problems, in this article we take advantage of two recent tools developed for language comparison: Bouckaert et al.'s (2012) list of Indo-European (hereafter: IE) lexical cognates and Longobardi and Guardiano's (2009) Parametric Comparison Method (PCM). We used these new resources to interpret the patterns of genome-wide variation in 15 European populations (belonging to three different linguistic families), inferred from autosomal single nucleotide polymorphisms (SNPs) data; the final dataset included 805 individuals, and after data cleaning and integration we had >177,000 SNPs autosomal SNPs for the analysis.

MATERIALS AND METHODS

The new linguistic approach

Our main tool for historical comparison is the PCM, that costitutes a radical departure from traditional procedures and databases. Languages are increasingly studied by theoretical linguists not merely as lists of words, but also as sets of recursive rules (technically, generative grammars: Chomsky 1955) combining words into an infinite number of sentences (Chomsky 1965). Therefore,

an alternative to comparison of vocabularies is precisely exploring the phylogenetic potential of grammatical diversity (different rules of (co-)occurrence, order, and interpretation of various classes of words, morphemes, and features: Nichols 1992; Longobardi 2003; Guardiano and Longobardi 2005).

Longobardi and Guardiano's (2009) central hypothesis is that syntactic change, though insightfully shown to be 'catastrophic' in specific 'local' instances (Lightfoot 1991), when considered as an overall phenomenon might turn out to proceed slowly enough to produce retrievable evolution (Longobardi 2003). If so, generative grammars could encode a historical signal, useful for deeper classification of languages and populations.

In the PCM approach, the core grammar of any language is represented as a string of binary symbols, each encoding the value of a syntactic parameter (Baker 2001; Biberauer 2008; Chomsky 1981; Clark and Roberts 1993; Roberts 2007). Parameters are drawn from a supposedly universal list, defining a structured variation space within the human capacity often labeled 'universal grammar' (UG) or 'faculty of language'. Therefore, through PCM, in principle, all languages, no matter how lexically distant, could now be compared, bypassing many problems arising with word collation. Case studies suggested that the chance probability of parametric resemblance can be computed and controlled for (Bortolussi et al., 2011), as well as certain amounts of homoplasy (Longobardi 2012) and admixture (Longobardi et al., 2013); finally, there is less *a priori* reason to expect external (e.g. cultural) factors to exert selective pressure on syntax than on lexical items (Guardiano & Longobardi 2005; Longobardi & Guardiano 2009; Ringe & Eska 2013). The use of such complex and explicit parameter sets has prompted a debate on language learnability issues (Boeckx & Leivada 2013), but also enabled a proof-of-concept study of a small sample of Old-World languages/populations, already showing how correlations can be found between a preliminary set of parametric distances and genetic ones (Colonna et al. 2010).

Syntactic and lexical distances

Recently, PCM has been validated on a set of 26 IE languages (Longobardi et al. 2013), producing near-perfect taxonomies within this family. Thus, for the purpose of investigating gene-language congruence in Europe, from the intersection of the languages of this syntactic dataset

with the much wider IE language sample of Bouckaert et al. (2012), we selected a further subset of 12 varieties, for whose speakers genome-wide data are publicly available. For such languages, we calculated and compared distances and phylogenies both from the new lexical list and through PCM. Then, we started to apply the results to a cross-disciplinary study, using the relationship between genomic, linguistic and geographic distances to draw inferences on the history of the corresponding populations. Finally, we expanded the analysis to include some non-IE languages.

The crucial problem for quantitative treatments of language taxonomy is the pervasive non-independence of characters, saliently emerging in grammar (Greenberg 1963; Hawkins 1983; Baker 2001; Biberauer 2008), and potentially disruptive for phylogenetic results.

The PCM has been originally designed for spelling out hypotheses on -and control for- crossparametric implications (Longobardi & Guardiano 2009; Bortolussi et al. 2011): here, the non-independence of characters is controlled by making explicit hypotheses about implication of syntactic properties and adopting a distance calculation appropriate for them (Longobardi & Guardiano 2009; Longobardi et al. 2013). In the present work, in particular, we relied on the grid of 56 nominal parameters described in the support material to Longobardi et al. (2013) (http://benjamins.com/#catalog/journals/jhl.2.2.04rat/additional), whose values have been additionally set for the three non-IE languages (Supplementary Table 1), and on the distance calculation method proposed there as well as in previous works (Longobardi and Guardiano, 2009: normalized Hamming distance or Jaccard distance): the number of differences between two languages is divided by the sum of their identities and differences. Equivalently, one can estimate the pairwise syntactic distances (d_{SYN}) by calculating the Jaccard (1901) coefficient *r*, namely the number of elements in the intersection $X \cap Y$ of two sets X and Y, here representing different languages (Lewandowsky and Winter, 1971). This is construed as the probability (between 0 and 1) that an element of at least one set is an element of both, and thus measures the overlap between the two. Then, we took the value $1-r = d_{SYN}$ as a measure of the dissimilarity of the two sets. Thus, resulting distances turn out to equally fall between 0 and 1.

The second set of linguistic distances, d_{LEX}, is based on lexical comparisons. Computational approaches to phylogenetic linguistics have led to refinements of lists of taxonomic characters

available for classifying Indo-European (IE) languages. The most recent breakthrough in such lists has been provided by Bouckaert et al.'s (2012) set of IE cognate words, which summarizes the expert cognacy judgments assigned by Dyen et al. (1992), Ringe et al. (2002), along with other sources, for a 207 Swadesh (1952) list in several IE languages. This makes available a richer device for several quantitative experiments on lexical diversification in Europe and Western Asia.

However, even lexical characters at a certain degree of sophistication encode some internal structure and redundancy of information: e.g. in Bouckaert et al. (2012), lexical roots are listed instead of meanings to take into proper account language-internal polymorphism (synonymy). This requires some calculation *caveat*, again. Thus, in order to perform our comparisons, we computed a set of pairwise distances based on the number of character differences out of the number of all lexical roots expressed at least in one of the two languages compared; again, this way all distances turn out to fall between a minimum of 0 and a maximum of 1. Yet, in our particular dataset, it turned out that almost all values of the resulting matrix were scattered around 0.9, hence scarcely informative and historically not plausible. This is likely to be a natural consequence of the criteria adopted to compute differences: indeed, Swadesh-lists require each meaning to be expressed by at least one lexical root in each of the languages; since polymorphism within the same language is expected to be a marked, albeit not uncommon, phenomenon, every lexical root displayed in a language but not in another is likely to predict a different lexical root to express the same meaning in the second language, thus doubling differences. Taking this into account and assigning differences a weight of 0.5 (rather than 1), we obtain a distance matrix (d_{LEX}) which fundamentally patterns with that previously obtained (Longobardi et al., 2013) from Dyen et al. (1992). Given that, by definition, only within the same family is it possible to compute some safe rate of common lexical etymologies, for comparison between languages from different families, which by definition share no common root, distance 1 was obviously assigned. An approximation to the distance between Hungarian and Finnish was tentatively computed from some literature references (Laasko, 2000; Peust, 2013).

The linguistic Principal Component Analysis was performed using the R *FactoMineR* program (Lê et al., 2008), with (implied) '0' values coded as 'NA'.

Genetic analyses

Genomic data on 13 populations were found in POPRES (dbGap accession phs000145.v1.p1; Nelson et al., 2008), a public resource for genetic research including 5,886 subjects genotyped at 500,568 loci using the Affymetrix 500K SNP chip. To determine the geographic location that best represents each individual's ancestry, we used a strict criterion of sample selection excluding individuals who reported mixed grandparental ancestry. A Basque (Henn et al., 2012) and a Finnish (1000 Genomes Project Consortium, 2012) sample were then added (Fig. 1). Outliers and individuals showing high levels of genetic similarity, which may point to biological relatedness, were excluded, and all data were merged using PLINK (Purcell et al., 2007). To avoid any ambiguity in strand alignment, we removed from the merged genotype datafile the alleles carrying ambiguities in strand-flipping, namely A/T and C/G polymorphisms. The final dataset comprises 177,949 markers that passed quality filters in all datasets for 805 individuals (minor allele frequency = 0.01, Genotyping Rate = 98%, Table S1). F_{ST} values between pairs of populations (Weir and Cockerham, 1984) were calculated by the 4P software (Benazzo et al., 2015).

Matrix comparisons

We started by inferring four matrices of pairwise distances between European populations: geographic (d_{GEO}: great circle distances, calculated with the R *gdistance* package), genomic (d_{GEN}: based on F_{ST}), and two types of linguistic distances, syntactic (d_{SYN}, based on the Jaccard index from syntactic parameters) and lexical (d_{LEX}, based on the criteria mentioned above).

Only 12 IE languages from Longobardi et al's (2013) original dataset of 26 find a match in the publicly available genomic data about European populations; yet, we also found genomic information for other 3 non-IE-speaking populations in Europe, which had been previously analyzed by the PCM (Longobardi & Guardiano 2009), and for which it was possible to set the same 56 universal syntactic parameters (Longobardi et al. 2013) used to classify IE, i.e. two Finno-Ugric languages - Finnish and Hungarian, and Basque. Depending on the test carried out, d_{GEN}, d_{SYN}, d_{LEX} and d_{GEO} were calculated either for the 12 IE-speaking populations, or for the whole set of 15 populations.

Correlations between pairs of these distance matrices were calculated according to the Mantel (1967) procedure, using the *mantel* function of the R *Vegan* package. The significance was empirically estimated over 10,000 permutations. To exclude the potentially confounding effect of some variable, we also ran partial Mantel tests, thus calculating the correlation between two matrices while controlling for (i.e. keeping constant) a third distance matrix. To this end, we used the *mantel.partial* function of the R *Vegan* package. Finally, to compare tree topologies (Steel and Penny, 1993), we calculated the path difference distance between trees using the *treedist* function of the R *phangorn* package, and we generated the 100,000 pairs of random trees for 12 and 15 taxa with the *rtree* function of the R *ape* package.

An improved method to describe population splits and later gene flow

Population structure depends on a number of evolutionary and demographic processes which may be difficult or impossible to summarize in the form of a simple bifurcating tree. Therefore, we also represented genomic variation by a network in which populations may exchange migrants after they have split from their common ancestors, thus violating the simplistic assumptions of most tree-building models (Pickrell and Pritchard, 2012). The first step in this exercise is the estimation of a maximum-likelihood tree. Populations poorly fitting the tree model are then identified, and migration events involving them are superimposed, so that the tree with the added migration episodes will account for a greater proportion of the overall genetic variance than the simple tree itself. This way, each population may have multiple origins, and the migrational contacts in the descendant populations are highlighted.

RESULTS

First of all, we had to make sure that the smaller subset of 12 IE languages displays a significant syntax-lexicon correlation, and retains as a plausible phylogenetic structure as that generated from the wider sample of 26 in Longobardi et al. (2013). Thus, for such 12 IE languages/populations, we compared d_{SYN} and d_{LEX} with one another. The two linguistic matrices appeared highly correlated (r=0.82).

To better understand to what extent differences in lexicon and syntax mirror each other, we represented the matrices in tree form (Fig. 2a, b), calculated the path difference distance between trees (Steel & Penny 1993), and compared this value with those obtained in 100,000 pairs of random topologies drawn, with replacement, from the total set of the possible topologies for 12 taxa. No closer match between topologies was observed (hence $P < 10^{-5}$).

The two linguistic matrices, being highly correlated with each other, expectedly showed very similar levels of correlations with genetic distances (r=0.49 and 0.51), in both cases reaching a high statistical significance, which stands Bonferroni correction for multiple tests (Table 1). Linguistic distances based on syntax also show a significantly tighter association with geography than their lexical counterparts.

Anyway, the most important result obtained is that the correlations of both lexicon and syntax with genetic distances are higher than between genes and geography (r=0.38). I.e. once precise measurements of linguistic differences are used, language emerges as a better predictor of genetic differences than geography in Europe .

Such conclusions have been reached on the already available IE databases. In order to strengthen them, we extended the analysis to the three non-IE languages of Europe mentioned above (i.e. Finnish, Hungarian and Basque). To do so, we crucially relied on PCM's ability to compare languages even from different families. Recall, indeed, that calculating lexical distances from cognates for languages from different families is an essentially vacuous procedure, since by definition such languages have no common etymologies: hence the maximum distance 1 must *a priori* be assigned, so that the result is largely uninformative. A way to overcome this shortcoming was the development of PCM, precisely because it relies on polymorphic characters which are in principle universal.

The same four matrices and six correlations as above were recalculated for the whole set of 15 populations (Table 2). The most salient result is that the correlations between genes and languages, both for syntax (0.60) and lexicon (0.54), remain much higher than between genes and geography (0.30). Indeed, on the basis of this larger evidence, the latter correlation further

decreases, while that between genes and syntax rises substantially (from 0.49 to 0.60), strongly reinforcing the conclusion that syntax, in Europe, is a better predictor of genomic variation than geography. This correlation remains significant even after removing the effects of geography by means of a partial Mantel test (d_{SYN} vs. d_{GEN} r=0.57), and after Bonferroni correction for multiple tests. The correlation of genes and lexicon appears also to rise, though only marginally, presumably because it begins to suffer from the saturation of lexical distances across different families hinted at above. For this reason, the more languages from different families will be added for comparison, the more we expect reliance on PCM to become crucial.

It is also noticeable that all the correlations with geography become lower in the 15-unit sample, probably because the 3 linguistic outliers added to the sample are not equally peripheral geographically. Thus, to better understand gene-language congruence at the cross-family European level, we focused in more detail on syntactic distances.

We drew a UPGMA tree and carried out a Principal Component Analysis (PCA) from d_{SYN}. The evolutionary tree inferred from d_{SYN} (Fig. 2C) meets all basic expectations: the deepest nodes first separate Basque, and then the pair of Finno-Ugric languages, from the cluster comprising all the *IE* varieties. The latter retained their expected articulation: Romance, Germanic and Slavic form three clusters; then Greek and Irish, as the only representatives of their subfamilies in this study, occur on separate branches, and fall close in the tree to their geographic neighbors.

In the linguistic PCA, the combination of the first two axes, jointly accounting for 34.5% of the variance, separates IE languages (but Greek) from the others: Greek, an IE language without very close relatives, falls anyway opposite to Finnish, Hungarian and Basque (Fig. 3A). This pattern is largely expected, and the position of Greek as the outlier of IE in our sample is in agreement with previous computational experiments on lexical datasets (Bouckaert et al. 2012, Gray & Atkinson 2003).

In short, through syntax, precise comparison and measuring is finally possible even across established linguistic families: the main families and subfamilies of Europe were discriminated by means of just 56 abstract characters, suggested by formal grammatical theory, through standard

methods of evolutionary biology, without resorting to methodologically disputable long-range lexical comparisons.

Then, to synthetically visualize genomic diversity in a parallel way, we also drew the corresponding tree and carried out a PCA analysis from d_{GEN}. The UPGMA tree inferred from d_{GEN} (Fig. 2D) shows that two out of the three linguistic outliers, Finns and Basques, are clearly differentiated also genomically, and connected to the other populations by long independent branches. The rest of the tree mainly reflects geographical distances, and contains all IE-speaking populations, as well as Hungarians, who appear genetically related with their geographical neighbors, Serbs and Rumanians. Once again, the path difference distance (Steel and Penny 1993) was calculated between the syntactic and the genetic tree, and the probability to obtain a closer match between random trees with 15 populations turned out P <0.004. This implies a tight relationship between the topologies of the trees inferred from syntactic and genetic distances, one that is highly unlikely to have arisen just by chance. The only salient divergence is the position of Hungarians, mostly falling within a large group of Central Europeans.

Then, we carried out a parallel PCA of >177,000 SNPs in 805 individuals from the 15 populations representative of the previously considered languages. As expected, given the well-known low levels of cross-population diversity in humans in general (Barbujani and Colonna, 2010) and in Europe in particular (Novembre et al., 2008), the proportion of the overall variance accounted for by the two main axes is much lower (less than 1%) than in the analysis of linguistic data (Fig. 3B), as previously observed. However, the two PCAs are qualitatively similar in several respects, with a main central cluster containing all *IE* speakers along with Hungarians, and with Finns and Basques appearing as outliers though both relatively close to their nearest geographical neighbors (Poles and Spaniards, respectively).

An unsupervised ancestry-inference analysis basically led to the same conclusions as the PCA and confirmed the peculiar genetic position of Hungarians. Postulating three ancestral genomic clusters for Europe, i.e. as many as the language families in the database (k=3 plot, Fig. 4), such clusters largely correspond to: (i) Basques, (ii) Finns, and (iii) all other Europeans including Hungarians; the Basque sample shows connections with the Spanish and French ones (blue

component), and Finns seem to share some ancestry with Northern Europeans (Germans and Poles, orange component). Other analyses, assuming different numbers of clusters in the genomic data, are also given for completeness of information.

We further investigated the evolutionary relationships between populations by a method designed to identify gene flow episodes after the main population splits (Fig. 5). Indeed, a tree-like representation of genomic (or linguistic, for that matter) relationships disregards the possibility of exchanges occurring after populations separated from their common ancestor. The contribution of migrants to Rumania from Russia (0.43) as well as from Greece is in agreement with the populations' geographical proximity, and their traditionally well-assessed horizontal linguistic connection: the received concept of a Balkan common linguistic area, or *Sprachbund*, has found at least some suggestive correspondence even in the parametric linguistic analysis, for in three parameters Rumanian, the outlier of the Romance branch of the language tree (Fig. 3A), shares a state with Greek in contrast to the rest of Romance, in one also with Bulgarian (Longobardi et al. 2013). The Southern European origin of a fraction of the Hungarians (0.31), instead is not apparently matched either in the linguistic PCA (Fig. 4A) or tree (Fig. 3A9, only finding a loose potential correspondence in one of the 56 syntactic characters, Parameter 7 (DGP), whose Hungarian state might in theory have been borrowed from either German or Rumanian. Relatively recent gene flows, occurring after the main population splits, seem therefore to nicely match at least a fraction of the linguistic variation not immediately representable by classifying languages into families. It is an intriguing conjecture that biological relationships unpredictable by vertical linguistic history might reflect secondary gene flow independently detected by TreeMix.

The peculiar gene-language mismatch of Hungarians was already noticed by Cavalli Sforza et al. (1994), though without any possibility of assessing it through computation of linguistic distances, now made available by PCM. Indeed, the genes-syntax correlation recalculated after removing Hungary /Table 3) further rises very significantly (0.74), while the genes-geography one remains low (0.28), confirming the status of Hungarians as an exception, in this respect. The skew is even more salient in partial Mantel tests (respectively 0.72, with geography held constant, and 0.09, with syntax held constant), the sharpest demonstration to date of a language/biology correlation for the core of Europe.

DISCUSSION

Reliable evidence for parallelism of genetic and linguistic change had previously been provided, although generally on a regional scale and without formal quantification of language distances. Based first on quantitative approaches to cognate words (Bouckaert et al. 2012), and then refined through a syntactic method (PCM, Longobardi and Guardiano 2009) designed for comparing languages even from separate families, the correlation tests in this study reach more precise conclusions on a broader continental scale: populations speaking similar languages also 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 (for exceptions, also see Bolnick et al. 2004). The partial correlation tests show that populations speaking similar languages also tend to be genetically closer than expected on the sheer basis of their geographic location, so that language, i.e. basic vocabulary and now, at an even wider scale, syntax, appears to offer a better prediction of genomic distances than geography in Europe. All these correlations provide a new type of evidence for PCM and in turn for the general biolinguistic approach it is inspired by (Berwick et al. 2013; Di Sciullo & Boeckx 2011; Lightfoot 1999).

We could thus move on to a more detailed analysis of population diversity in Europe and of the possible exceptions to the conclusions above. When population relationships were summarized by trees, the main elements of disagreement were represented by the positions of Hungarians and Rumanians, which cluster genetically with speakers of Serbo-Croatian despite being highly differentiated syntactically. These populations all dwelling in Central Europe, it is reasonable to suspect an effect of geographical proximity, enhancing gene flow between neighboring countries.

Using a method that highlights the most significant episodes of genetic exchange after population splits, a likely situation among humans (Barbujani & Colonna 2010), especially in Europe, we could precisely find evidence of the possibly relevant biological contacts among speakers of IE-subfamilies (from Slavic-speaking areas into Rumania and from Southern Europe into the Balkans) and between Ugric and IE speakers (from the Balkans into Hungary). Although these contacts must be further investigated at the appropriate geographical scale, something

beyond the purposes of the present study, it appears that where biological relationships are not those expected from vertical linguistic history, they are plausibly accounted for by relatively recent gene flow processes independently detected by Treemix.

In particular, concerning the real exception to our congruence pattern, notice that the presence in modern Hungarians of DNA markers currently common in Northern and Central Asia has been interpreted as a consequence of westward gene flow in Medieval times (Csányi et al. 2008; Biró et al. 2009; Hellenthal et al. 2014); this is obviously connected with historical migrations in the 9th century and with the fact that the current language is closely related to the Ugric-speaking communities along the Ob river. However, the current low frequency of those markers is not what one would expect to observe, had a substantial demographic replacement occurred (Hellenthal et al. 2014; Nadasi ed al 2007). Careful analyses of 10th century ancient DNA in Hungary showed a predominance of European mitochondrial haplotypes in burials attributed to the lower classes, and a high incidence of Asian haplotypes in high-status individuals of that period (Tömöry et al. 2007), which points to the Asian immigrants as representing a social elite, rather than the bulk of the population. The exception to the results of the present study is thus nicely justified in this scenario, suggesting that when a Finno-Ugric language was introduced in Hungary, the genetic buildup of the population changed only in part, thus retaining similarities with its geographic neighbors, an example of the process called élite dominance by Renfrew (1992). On the contrary, the same case cannot be easily made for Basques (Alonso et al., 2005; Rodríguez-Ezpeleta et al., 2010; Young et al., 2011; Martinez-Cruz et al., 2012) or Finns (Nelis et al., 2009), for whom, to the best of our knowledge, no available evidence suggests a similar model of limited demographic replacement associated with language replacement. Thus, the comparative linguistic/genomic analysis, attempted in the present study, seems able to single out and precisely assess these differences in the population histories of the three non-IE members of our sample.

The results obtained here confirm the fruitfulness of importing numerical and biostatistical methods into language phylogenetics (McMahon & McMahon 2005), but even more of resorting to radically new (Heggarty et al. 2005) and more structured (Longobardi 2012) levels of taxonomic characters for a thorough reconstruction of both demographic and linguistic history. In particular, we see good chances to obtain trustworthy taxonomic insights when PCM is applied to longer-

range computations that could not be safely attempted through traditional lexical methods, and we expect to find interesting and illuminating correlations between genetic and linguistic diversity across other continents, contributing to the 'New Synthesis' research line (Renfrew 1987). Sokal (1988) and Cavalli-Sforza et al. (1988) could venture into addressing Darwin's gene-language congruence issue thanks to the theoretical progress of 20th century genetics; along with the availability of broad genomic datasets, the corresponding progress of formal grammatical theory over the past 50 years may now enable us to better test the hypothesis on an ever larger and more solid basis.

Acknowledgments

We are indebted to R. Gray and M. Dunn for kindly directing us to the expanded IE database used to infer lexical distances, and all the participants in the international workshop *Advances in Phylogenetic Linguistics* (Ragusa Ibla, July 13-17, 2013).

LITERATURE CITED

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

Alonso S, Flores C, Cabrera V, Alonso A, Martín P, Albarrán C, Izagirre N, de la Rúa C, García O. 2005. The place of the Basques in the European Y-chromosome diversity landscape. Eur J Hum Genet 13:1293-302.

Baker M. 2001. The atoms of language. New York: Basic Books.

Barbujani G, Colonna V. 2010. Human genome diversity: Frequently asked questions. Trends Genet 26: 285-295

Barbujani G, Pilastro A. 1992. Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. Proc Natl Acad Sci USA 90:4670-4673.

Barbujani G, Sokal RR. 1990 Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 87:1816-1819.

Belle EM, Barbujani G. 2007. Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol 133:1137-1146.

Benazzo A, Panziera A, Bertorelle G. 2015. 4P: fast computing of basic population genetics statistics from large DNA polymorphism panels. Ecol Evol (in press).

Berwick RC, Friederici AD, Chomsky N, Bolhuis JJ. 2013. Evolution, brain, and the nature of language. Trends Cogn Sci 17:89-98.

Biberauer T. 2008. The limits of syntactic variation. Amsterdam: Benjamins.

Bíró AZ, Zalán A, Völgyi A, Pamjav H. 2009. A Y-chromosomal comparison of the Madjars (Kazakhstan) and the Magyars (Hungary). Am J Phys Anthropol 139:305-310.

Boeckx C, Leivada E. 2013. Entangled parameter hierarchies: problems for an overspecified Universal Grammar. PLoS ONE 8: e72357.

Bolnick DA, Shook BA, Campbell L, Goddard I. 2004. Problematic use of Greenberg's linguistic classification of the Americas in studies of Native American genetic variation. Am J Hum Genet 75: 519-522.

Bortolussi L, Longobardi G, Guardiano C, Sgarro A. 2011. How many possible languages are there? In: Bel-Enguix G, Jiménez-López MD, editors. Biology, Computation and Linguistics. Amsterdam: IOS Press, p 168-179.

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD. 2012. Mapping the origins and expansion of the Indo-European language family. Science 337:957-960.

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

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85:6002-6006.

Chomsky N. 1955. The logical structure of linguistic theory. Ms. (published in 1975, New York: Plenum).

Chomsky N. 1965. Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Chomsky N. 1981. Lectures on government and binding. Dodrecht: Foris.

Clark R, Roberts I. 1993. A computational model of language learnability and language change. Linguistic Inquiry 24:299-345.

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. 70:245-254.

Csányi B, Bogácsi-Szabó E, Tömöry G, Czibula A, Priskin K, Csősz A., Mende B, Langó P, Csete K, Zsolnai A, Conant EK, Downes CS, Raskó I. 2008. Y-Chromosome analysis of ancient Hungarian and two modern Hungarian-speaking populations from the Carpathian basin. Ann Hum Genet 72:519-534.

Darwin C. 1859. On the Origin of Species. London: John Murray.

Di Sciullo AM, Boeckx C. (editors). 2011. The Biolinguistic Enterprise. New Perspectives on the Evolution and Nature of the Human Language Faculty. Oxford: Oxford University Press.

Dyen I, Kruskal J, Black PJ. 1992. An Indoeuropean classification: a lexicostatistical experiment. Trans. Philos. Soc. 82, 1-132.

Gray RD, Atkinson QD. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435-439.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323: 479-483.

Greenberg J. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg J, editor, Universals of Language. Cambridge, MA: MIT Press, p 73-113.

Greenhill SJ. 2011. Levenshtein distances fail to identify language relationships accurately. Computational Linguistics 37, 689-698.

Greenhill SJ, Atkinson QD, Meade A, Gray RD. 2010. The shape and tempo of language evolution. Proc. Biol. Sci. 277:2443-2450.Guardiano C, Longobardi G. 2005. Parametric comparison and language taxonomy, in Battlori M, Picallo C, Roca F (editors). Grammaticalization and Parametric Variation, Oxford: Oxford University Press, p 149-174.

Guardiano C., Longobardi G. 2005. Parametric comparison and language taxonomy. In: in Batllori, M., Picallo, C., Roca, F. (editors) Grammaticalization and Parametric Variation. Oxford: Oxford University Press, p 149-174.

Hawkins J. 1983. Word order universals, New York: Academic Press.

Heggarty P. 2004. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data – and to dating language? In: Clarkson J, Forster P, Renfrew C, editors. Phylogenetic methods and the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research, p 183-194.

Heggarty P, McMahon A, McMahon R. 2005. From phonetic similarity to dialect classification: a principled approach. In Delbecque N, Geeraerts D, van der Auwera J. (editors) Perspectives on Variation: Sociolinguistic, Historical, Comparative, Amsterdam: Mouton de Gruyter, 43-91, http://dx.doi.org/10.1515/9783110909579.43.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343, 747-751.

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

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37:547-579.

Laasko J. 2000. "Related words" in Finnish and Hungarian, http://www.helsinki.fi/~jolaakso/f-h-ety.html

Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. J Stat Software. 25: 1-18.

Levinson SC, Gray RD. 2012. Tools from evolutionary biology shed new light on the diversification of languages. Trends Cogn Sci 16: 167-173.

Lewandowsky M, Winter D. 1971. Distance between sets. Nature 234:34-35.

Lightfoot D. 1991. How to Set Parameters. Cambridge, MA: MIT Press.

Lighfoot D. 1999. The Development of Language: acquisition, change and evolution. Cambridge, MA: MIT Press.

Longobardi G. 2003. Methods in parametric linguistics and cognitive history. Linguistic Variation Yearbook 3:101-138.

Longobardi G. 2012. Convergence in parametric phylogenies: Homoplasy or principled explanation?, in Galves C, Cyrino S, Lopes R, Sandalo F, Avelar J, editors. Parameter theory and linguistic change. Oxford: Oxford University Press, p 304-319.

Longobardi G, Guardiano C. 2009. Evidence for syntax as a signal of historical relatedness. Lingua 119:1679-1706.

Longobardi G, Guardiano C, Silvestri G, Boattini A, Ceolin A. 2013. Toward a syntactic phylogeny of modern Indo-European languages. J Histor Linguistics 3:122-152.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209-220.

Martínez-Cruz B, Harmant C, Platt DE, Haak W, Manry J, Ramos-Luis E, Soria-Hernanz DF, Bauduer F, Salaberria J, Oyharçabal B, Quintana-Murci L, Comas D; Genographic Consortium. 2012. Evidence of pre-Roman tribal genetic structure in Basques from uniparentally inherited markers. Mol Biol Evol 29:2211-2222.

McMahon A, McMahon R. 2003. Finding families: Quantitative methods in language classifying. Trans Philol Soc 101:7-55.

McMahon A, McMahon R. 2005. Language Classification by Numbers. Oxford: Oxford University Press.

Nadasi E, Gyurus P, Czakó M, Bene J, Kosztolányi S, Fazekas S, Dömösi P, Melegh B. 2007. Comparison of mtdna haplogroups in Hungarians with four other European populations: A small incidence of descents with Asian origin. Acta Biologica Hungarica 58:245-256.

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

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH. 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347-358.

Nichols JA 1992. Linguistic Diversity in Space and Time. Chicago: The University of Chicago Press.

Nichols JA. 1996. The Comparative Method reviewed: Regularity and irregularity in language change, in Durie M, Ross M. (editors) The Comparative Method as Heuristic. New York: Oxford University Press, p 39-71.

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

Peust C. 2013. Towards establishing a new basic vocabulary list (Swadesh list), http://www.peust.de/peustBasicVocabularyList.pdf.

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

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, 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 61:1015-1035.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559-575.

Renfrew C. 1987. Archaeology and language. The puzzle of Indo-European origins. London: Jonathan Cape.

Renfrew C. 1992. Archaeology,, genetics and linguistic change. Man 27:445-478

Ringe D. 1996. The mathematics of Amerind. Diachronica 13:135-154.

Ringe D, Eska J. 2013. Historical Linguistics. Cambridge: Cambridge University Press.

Ringe D, Warnow T, Taylor A. 2002. Indo-European and computational cladistics. Trans. Philol. Soc. 100:59-129.

Roberts I. 2007. Diachronic syntax. Oxford: Oxford University Press.

Rodríguez-Ezpeleta N, Alvarez-Busto J, Imaz L, Regueiro M, Azcárate MN, Bilbao R, Iriondo M, Gil A, Estonba A, Aransay AM. 2010. High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations. Hum Genet 128:113-117.

Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus ML, Aula P, Beckman L, Tranebjaerg L, Gedde-Dahl T, Issel-Tarver L, DiRienzo A, Pääbo S. 1995. Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 5:42-52.

Schleicher A. 1863. Die Darwinsche Theorie und die Sprachwissenschaft. Weimar: Böhlau.

Sokal RR. 1988. Genetic, geographic, and linguistic distances in Europe. Proc Natl Acad Sci USA 85:1722-1726.

Steel MA, Penny P. 1993. Distribution of tree comparison metrics- Some new results. Syst Biol 42: 126-141.

Swadesh M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. Proc. Philos. Soc. 96:453-463.

Tömöry G, Csányi B, Bogácsi-Szabó E, Kalmár T, Czibula A, Csosz A, Priskin K, Mende B, Langó P, Downes CS, Raskó I. 2007. Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. Am J Phys Anthropol 134:354-368.

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

Young KL, Sun G, Deka R, Crawford MH. 2011. Paternal genetic history of the Basque population of Spain. Hum Biol 83:455-475.

FIGURE LEGENDS

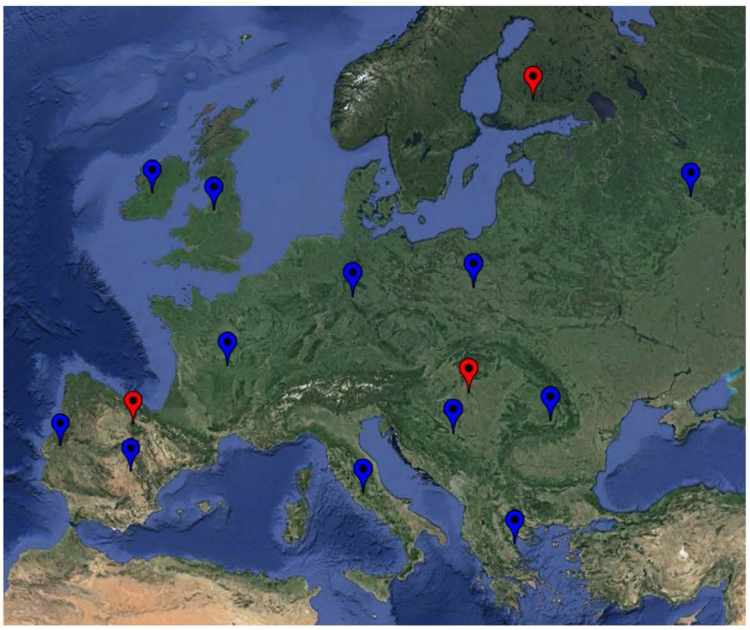
Fig. 1: Geographic distribution of the samples considered in this study. Indo-European-speaking populations in blue, populations speaking Finno-Ugric languages (Hungarian, Finnish) and the linguistic isolate (Basque) in red.

Fig. 2: UPGMA trees summarizing population relationships. Distances inferred from: (A) lexical and (B) syntactic comparisons among 12 Indo-European-speaking European populations; (C) syntactic comparisons among 15 European languages, and (B) F_{ST} distances among 15 populations sharing 177,949 SNPs. Lexical distances were estimated from list of cognate words, amounting to over 6,000 roots (http://ielex.mpi.nl/); syntactic distances were measured over 56 parameters of nominal phrases (http://dx.doi.org/10.1075/jhl.3.1.07lon.additional). In (D), numbers indicate the support of the branching after 100 bootstrap replicates.

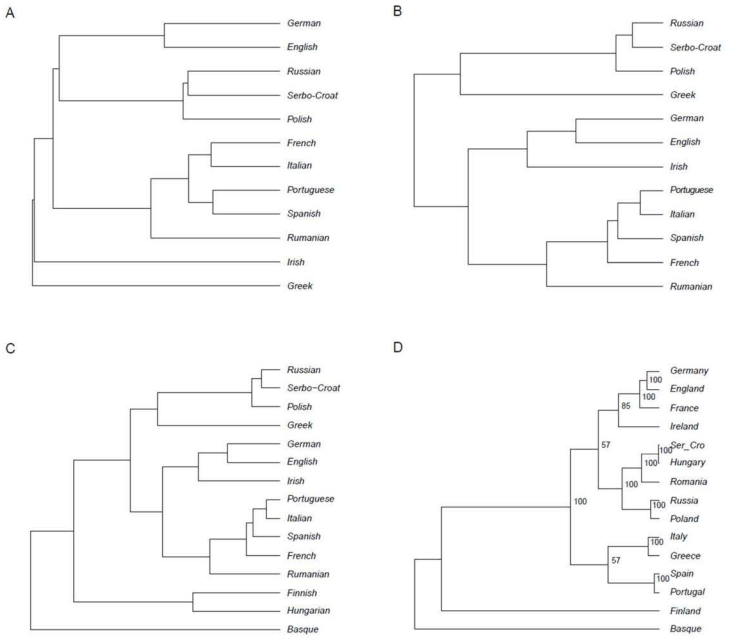
Fig. 3: Projection on two dimensions of the main components (PCA) of linguistic (A) and individual genomic (B) variation. The linguistic PCA was performed using the *R FactoMineR* program, with (implied) '0' values coded as 'NA', whereas the genomic PCA was calculated with the *R SNPRelate* package. Note that the linguistic scatter diagram accounts for a fraction of the total variance that is >25 fold as large as that accounted for by the genomic scatter diagram.

Fig. 4: Unsupervised ancestry-inference analysis based on the software ADMIXTURE. Each individual genotype is represented by a column in the area representing the appropriate population, and colors correspond to the fraction of the genotype that can be attributed to each of the K groups ($2 \leq K \leq 5$) assumed to have contributed to the populations' ancestry.

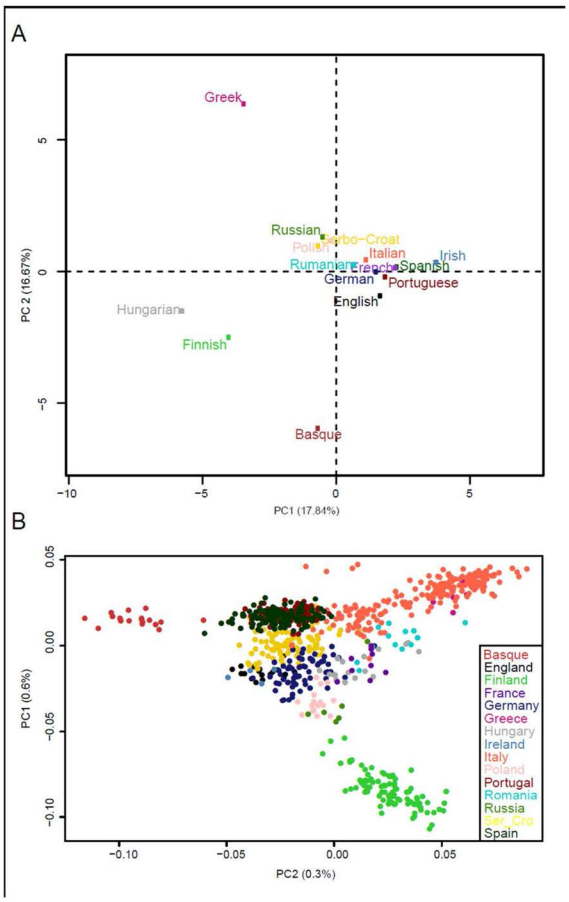
Fig. 5: Maximum-likelihood population trees. The algorithm chosen, TreeMix (*28*), estimates phylogenetic relationships with (A) three, (B) one, and (C) two superimposed migration events after the main population splits.



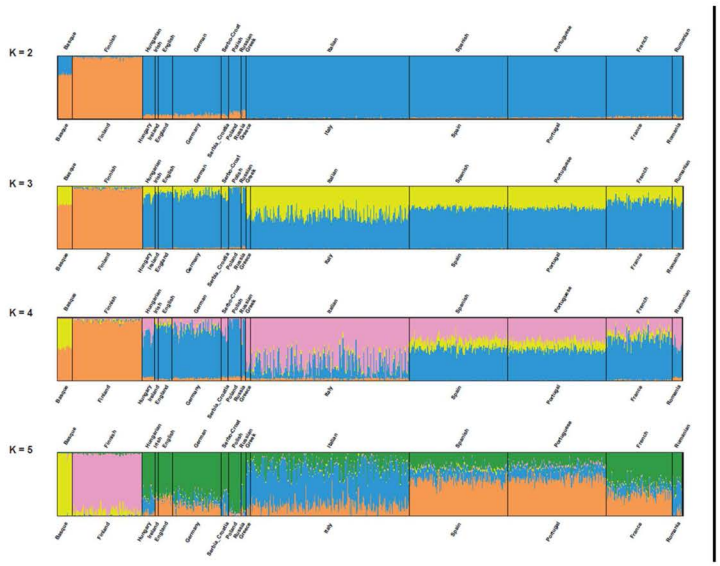
Geographic distribution of the samples considered in this study. Indo-European-speaking populations in blue, populations speaking Finno-Ugric languages (Hungarian, Finnish) and the linguistic isolate (Basque) in red.
104x87mm (300 x 300 DPI)



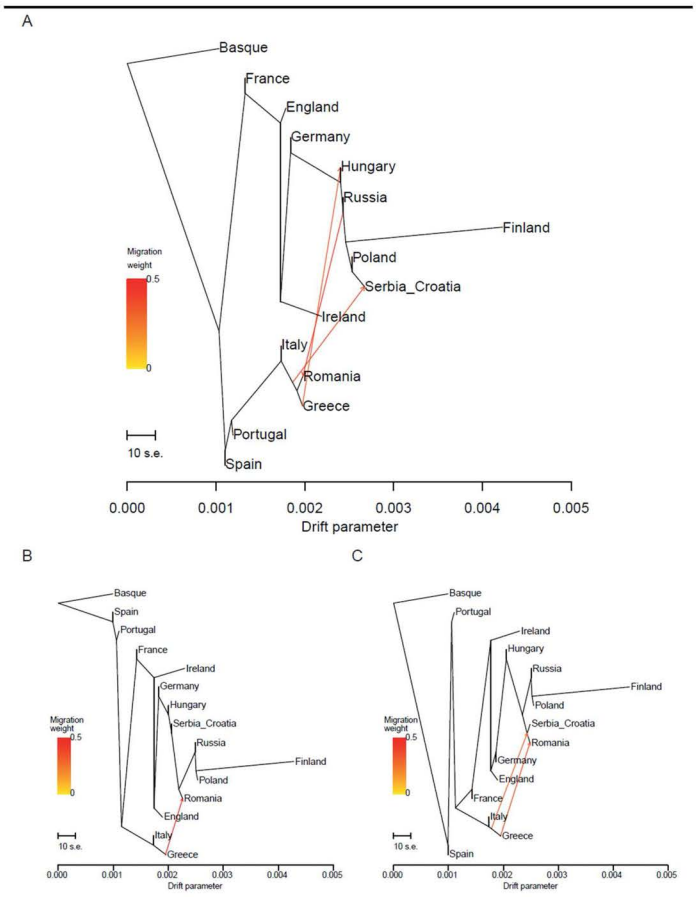
UPGMA trees summarizing population relationships. Distances inferred from: (A) lexical and (B) syntactic comparisons among 12 Indo-European-speaking European populations; (C) syntactic comparisons among 15 European languages, and (B) FST distances among 15 populations sharing 177,949 SNPs. Lexical distances were estimated from list of cognate words, amounting to over 6,000 roots (http://ielex.mpi.nl/); syntactic distances were measured over 56 parameters of nominal phrases (http://dx.doi.org/10.1075/jhl.3.1.07lon.additional). In (D), numbers indicate the support of the branching after 100 bootstrap replicates.
323x296mm (96 x 96 DPI)



Projection on two dimensions of the main components (PCA) of linguistic (A) and individual genomic (B) variation. The linguistic PCA was performed using the R FactoMineR program, with (implied) '0' values coded as 'NA', whereas the genomic PCA was calculated with the R SNPRelate package. Note that the linguistic scatter diagram accounts for a fraction of the total variance that is >25 fold as large as that accounted for by the genomic scatter diagram.
237x376mm (96 x 96 DPI)



Unsupervised ancestry-inference analysis based on the software ADMIXTURE. Each individual genotype is represented by a column in the area representing the appropriate population, and colors correspond to the fraction of the genotype that can be attributed to each of the K groups ($2 \leq K \leq 5$) assumed to have contributed to the populations' ancestry.
317x238mm (96 x 96 DPI)



Maximum-likelihood population trees. The algorithm chosen, TreeMix (28), estimates phylogenetic relationships with (A) three, (B) one, and (C) two superimposed migration events after the main population splits.
282x365mm (96 x 96 DPI)

Table 1. Mantel correlations between genetic, geographic and two kinds of linguistic distances in Indo-European-speaking populations of Europe. After Bonferroni correction for multiple tests, these results are significant at the P=0.0006 level.

Distance matrices	*r*	*P*
d_{LEX} d_{GEO} Linguistic (lexical)-Geographic	0.206	0.077
d_{LEX} d_{GEN} Linguistic (lexical)-Genetic	0.514	0.0001
d_{SYN} d_{GEO} Linguistic (syntactic)-Geographic	0.385	0.008
d_{SYN} d_{GEN} Linguistic (syntactic)-Genetic	0.491	0.0004
d_{LEX} d_{SYN} Linguistic (lexical)-Linguistic (syntactic)	0.822	0.0001
d_{GEN} d_{GEO} Genetic-Geographic	0.390	0.011

Table 2. Mantel correlations and partial Mantel correlations between matrices of Syntactic, Lexical, Geographic and Genetic distance in 15 populations in Europe. After Bonferroni correction for multiple tests, these results are significant at the *P*=0.012 level.

Distance matrices	*r*	*P*
d_{GEN} d_{GEO} Genetic - Geographic	0.299	0.030
$d_{SYN}d_{LEX}$ Syntactic - Lexical	0.850	0.001
d_{SYN} d_{GEO} Syntactic - Geographic	0.240	0.039
d_{LEX} d_{GEO} Lexical - Geographic	0.084	0.264
d_{SYN} d_{GEN} Syntactic - Genetic	0.599	0.001
d_{LEX} d_{GEN} Lexical - Genetic	0.537	0.001
d_{GEN} d_{GEO} (d_{SYN}) Genetic - Geographic (Syntax held constant)	0.200	0.114
d_{GEN} d_{GEO} (d_{LEX}) Genetic - Geographic (Lexicon held constant)	0.302	0.035
d_{SYN} d_{GEO} (d_{GEN}) Syntactic - Geographic (Genetics held constant)	0.079	0.264
d_{LEX} d_{GEO} (d_{GEN}) Lexical - Geographic (Genetics held constant)	-0.095	0.736
d_{SYN} d_{GEN} (d_{GEO}) Syntactic - Genetic (Geography held constant)	0.570	0.002
d_{LEX} d_{GEN} (d_{GEO}) Lexical - Genetic (Geography held constant)	0.538	0.001

Table 3. Mantel correlations and partial Mantel correlations between matrices of Syntactic, Lexical, Geographic and Genetic distance for 14 populations in Europe (after removing Hungary). After Bonferroni correction for multiple tests, these results are significant at the P=0.012 level.

Distance matrices	*r*	*P*
d_{GEN} d_{GEO} Genetic - Geographic	0.275	0.048
$d_{SYN}d_{LEX}$ Syntactic - Lexical	0.850	0.001
d_{SYN} d_{GEO} Syntactic - Geographic	0.291	0.026
d_{LEX} d_{GEO} Lexical - Geographic	0.152	0.144
d_{SYN} d_{GEN} Syntactic - Genetic	0.740	0.001
d_{LEX} d_{GEN} Lexical - Genetic	0.687	0.001
d_{GEN} d_{GEO} (d_{SYN}) Genetic - Geographic (Syntax held constant)	0.093	0.254
d_{GEN} d_{GEO} (d_{LEX}) Genetic - Geographic (Lexicon held constant)	0.238	0.083
d_{SYN} d_{GEO} (d_{GEN}) Syntactic - Geographic (Genetics held constant)	0.135	0.178
d_{LEX} d_{GEO} (d_{GEN}) Lexical - Geographic (Genetics held constant)	-0.053	0.615
d_{SYN} d_{GEN} (d_{GEO}) Syntactic - Genetic (Geography held constant)	0.717	0.001
d_{LEX} d_{GEN} (d_{GEO}) Lexical - Genetic (Geography held constant)	0.679	0.001

PAPER IV: Origins and Evolution of the Etruscans' mtDNA.

OPEN ACCESS Freely available online

PLOS ONE

Origins and Evolution of the Etruscans' mtDNA

Silvia Ghirotto[1], Francesca Tassi[1], Erica Fumagalli[1,2¤a], Vincenza Colonna[1,3], Anna Sandionigi[4], Martina Lari[4], Stefania Vai[4], Emmanuele Petiti[4], Giorgio Corti[5¤b], Ermanno Rizzi[5], Gianluca De Bellis[5], David Caramelli[4], Guido Barbujani[1]*

1 Department of Biology and Evolution, University of Ferrara, Ferrara, Italy, **2** Department of Biotechnologies and Biosciences, University of Milano-Bicocca, Milan, Italy, **3** Institute of Genetics e Biophysics "Adriano Buzzati-Traverso", National Research Council, Naples, Italy, **4** Department of Evolutionary Biology, University of Florence, Florence, Italy, **5** Institute for Biomedical Technologies, National Research Council, Segrate, Milan, Italy

Abstract

The Etruscan culture is documented in Etruria, Central Italy, from the 8th to the 1st century BC. For more than 2,000 years there has been disagreement on the Etruscans' biological origins, whether local or in Anatolia. Genetic affinities with both Tuscan and Anatolian populations have been reported, but so far all attempts have failed to fit the Etruscans' and modern populations in the same genealogy. We extracted and typed the hypervariable region of mitochondrial DNA of 14 individuals buried in two Etruscan necropoleis, analyzing them along with other Etruscan and Medieval samples, and 4,910 contemporary individuals from the Mediterranean basin. Comparing ancient (30 Etruscans, 27 Medieval individuals) and modern DNA sequences (370 Tuscans), with the results of millions of computer simulations, we show that the Etruscans can be considered ancestral, with a high degree of confidence, to the current inhabitants of Casentino and Volterra, but not to the general contemporary population of the former Etruscan homeland. By further considering two Anatolian samples (35 and 123 individuals) we could estimate that the genetic links between Tuscany and Anatolia date back to at least 5,000 years ago, strongly suggesting that the Etruscan culture developed locally, and not as an immediate consequence of immigration from the Eastern Mediterranean shores.

Citation: Ghirotto S, Tassi F, Fumagalli E, Colonna V, Sandionigi A, et al. (2013) Origins and Evolution of the Etruscans' mtDNA. PLoS ONE 8(2): e55519. doi:10.1371/journal.pone.0055519

Editor: John Hawks, University of Wisconsin, United States of America

Received July 20, 2012; **Accepted** December 24, 2012; **Published** February 6, 2013

Funding: Study supported by the Italian Ministry for Universities Funds PRIN 2008 to GB and DC and FIRB 2008 (RBFR08U07M) to ER, DC and GB, by the "Futuro in ricerca" grant RBFR08U07M to ML, ER, GC, GD and DC, by the Fondazione Cassa di Risparmio di Ferrara and by Associazione Archeologica Odysseus Casale di Pari. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: g.barbujani@unife.it

¤a Current address: Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
¤b Current address: Institute for Cancer Research and Treatment, Candiolo (Turin), Italy

Introduction

The Etruscan culture is documented in Central Italy (current Tuscany and Northern Latium, formerly known as Etruria) between the 8th and the 1st century BC. Questions about the Etruscans' origins and fate have been around for millennia. Herodotus and Livy regarded them as immigrants, respectively from Lydia, i.e. Western Anatolia, or from North of the Alps, whereas for Dionysius of Halicarnassus they were an autochthonous population [1]. Previous DNA studies, far from settling the issue, have raised further questions. The Etruscans' mitochondrial DNAs (mtDNAs) appear similar, but seldom identical, to those currently observed in Tuscany [2,3]. Assuming reasonable effects of genetic drift and mutation, these levels of resemblance proved incompatible with the notion that modern Tuscans are descended from Etruscan ancestors [4,5]. Explanations for this result include the (extreme) possibility that the Etruscans became extinct, but also that their modern descendants are few and geographically dispersed, or that the ancient sample studied represents a small social elite rather than the entire population [4]. As for the Etruscans' origins, ancient DNA is of little use, because pre-Etruscan dwellers of Central Italy, of the Villanovan culture, cremated their dead [1], and hence their genetic features are unknown. DNAs from modern humans and cattle in Tuscany show affinities with Near Eastern DNAs, which was interpreted as supporting Herodotus' narrative [2,6], but in these studies modern Tuscans were assumed to be descended from Etruscan ancestors, in contrast with ancient DNA evidence [5]. The claim that systematic errors in the Etruscan DNA sequences led to flawed genealogical inference [2,7] is not supported by careful reanalysis of the data [8].

What previous studies overlooked is the potential genetic effect of population subdivision. If most Etruscans' descendants lived in isolated communities in the last 2,000 years, their DNAs may still persist in some localities, but will escape detection unless they are sought at the appropriate (i.e., smaller) geographical scale. Indeed, previous work in another area of Italy [9] showed that modern populations separated by only tens of kilometers can differ sharply in their genealogical relationships with ancient populations. To investigate in greater geographical detail the biological relationships between contemporary and ancient populations, we thus sampled multiple burials in classical Etruria. MtDNA was extracted from bones, amplified and sequenced by a combination of classical methods and Next Generation Sequencing. After adding these sequences to the other Etruscan sequences produced in our lab [3] we compared them through methods of

Approximate Bayesian Computation with those of relevant ancient and modern human populations. These include Medieval Tuscans (n = 27) [5], contemporary Tuscans from three sites in historical Etruria (Casentino, n = 122; Murlo, n = 86; Volterra, n = 114) [2] and from Florence [10] (n = 48) (Figure 1). The sample from Florence here represents a control, since no special relationships is expected between the DNAs of the Etruscans and those of the inhabitants of a large city, after millennia of immigration.

We thus tried to address two questions, namely (1) whether an analysis at the small geographical scale can provide evidence of a genealogical continuity between the Etruscans and some current inhabitants of historical Etruria, and (2) whether the observed degree of genetic resemblance between modern inhabitants of Tuscany and Western Anatolia has anything to do with the Etruscans' origins. To answer, for each modern population we designed and compared three demographic models differing for the genealogical relationships with the ancient samples (see Material and Methods for details). We identified the model best fitting each set of the observed data, and then we moved to estimating, under an isolation-with-migration (IM) framework, the separation time between Tuscan and Anatolian populations [11], evaluating whether the estimated time can be reconciled with an Etruscan origin in Anatolia and a subsequent migration in Italy around the 8th century BC.

Results

Ancient DNA Sequences

We could obtain amplifiable DNA from 14 Etruscan specimens. Four of them, from Tarquinia, were analyzed in 2004 but were still unpublished. Ten samples come from 18 initial bone samples (each represented by two fragments of the right tibia) from a 3rd century BC multiple burial in Casenovole, Southern Tuscany. The bones were freshly excavated and collected according to the most stringent ancient DNA criteria (see Materials and Methods) by one of us (EP); they can safely be regarded as belonging to different individuals. After a first round of DNA extraction, the 18 Casenovole samples were subjected to multiple PCRs, cloning and cycle sequencing. In ten of them we could determine the sequence of the complete mtDNAhypervariable region I (hereafter: HVR-I), whereas the remaining eight gave no results (Figure S1). Their final consensus sequences (Table S1) were determined by comparing results obtained using the standard procedures (575 clones overall) and Next Generation Sequencing (127,837 reads) (Figure S2). We added to these the sequences of four individuals from Tarquinia, (GenBank accession numbers: bankit1285669 GU186064; bankit1285680 GU186065; bankit1285699 GU186066; bankit1285702 GU186067).

The Etruscans in the Context of Modern and Ancient Genetic Diversity

We analyzed four non-overlapping datasets (Table 1). The ETR dataset comprises the 14 newly produced DNA sequences, along with 16 already available sequences from necropoleis in historic Etruria [3]; individuals from geographically distant Etruscan populations, Adria and Capua, were excluded. The TUS dataset comprises four modern Tuscan populations, i.e. Casentino, Murlo, Volterra and Florence; the last mentioned is a forensic sample, representing random members of a large city, to the exclusion of recent immigrants (Figure 1). In addition, this dataset includes a sample of Medieval Tuscans from Guimaraes et al. [5]. Finally, the ANC dataset and the EUR dataset include, respectively, data on ancient and modern populations from Europe and from the Near East.

In Table 2 we show several statistics summarizing genetic variation in the ETR and TUS datasets. Estimates of the internal genetic diversity of the Etruscans, as expressed by their mean pairwise difference (2.966±1.560) and by haplotype diversity (0.943±0.032), appear close to those obtained in Vernesi et al. [3] using a partly different dataset. We also calculated two measures of genetic distance between the Etruscans (ETR) and modern populations (EUR), namely Wright's pairwise Fst and allele sharing, the latter measured as the fraction of modern sequences

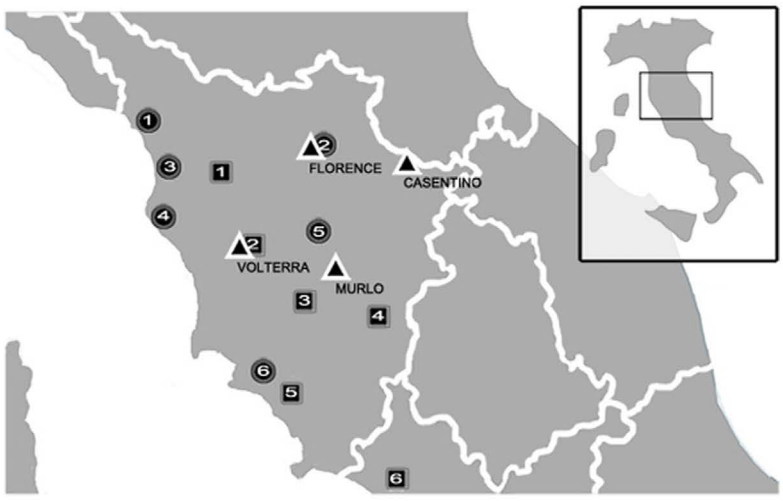


Figure 1. Geographic location of the samples considered in the ABC analysis. Triangles, Contemporary Tuscans (n = 370); Circles, Medieval Tuscans: 1. Massa Carrara (n = 3); 2. Florence, (n = 10); 3. Pisa, (n = 6); 4. Livorno, (n = 3); 5. Siena, (n = 4); 6. Grosseto (n = 1); Squares, Etruscans: 1. Castelfranco di Sotto (n = 1); 2. Volterra (n = 3); 3. Casenovole (n = 10); 4. Castelluccio di Pienza (n = 1); 5. Magliano/Marsiliana (n = 6); 6. Tarquinia (n = 9).
doi:10.1371/journal.pone.0055519.g001

Table 1. A synopsis of the datasets analyzed.

Dataset	N populations	N individuals	Notes
ETR	1	30	Etruscan sequences from the present paper and from Vernesi et al. (2004)
TUS	5	397	Medieval and modern sequences from Tuscany
EUR	52	4,910	Modern European sequences
ANC	9	190	Ancient European sequences

doi:10.1371/journal.pone.0055519.t001

also observed in the Etruscan sample (Figure S3). A general decline of genetic resemblance with geographic distance is evident (Figure 2).

Among the 30 Etruscan individuals (ETR dataset) we observed 21 different sequences with 24 variable sites (Table 2); the network describing the relationship among the Etruscans' haplotypes is reported in Figure 3. Comparisons with 52 modern populations in the TUS and EUR datasets (listed in Table S2) show that 11 of these sequences are shared with at least one of 4,910 individuals from Western Eurasia and the Southern Mediterranean shore (Table S1). The Etruscan sample falls within the range of contemporary genetic variation (EUR dataset, Figure S4A, S4B). In the comparison with the samples of the ANC dataset, the Etruscans appear to fall very close to a Neolithic population from Central Europe and to other Tuscan populations; geographically distant Bronze and Iron-age samples, from Iberia and Sardinia, appear genetically differentiated from the Etruscans (Figure S4C).

Genealogical Relationships between the Etruscans and Contemporary Populations

We investigated the genealogical relationships between ancient and contemporary samples by Approximate Bayesian Computation (ABC), a set of methods to fit complex evolutionary models to genetic data. We proceeded in 5 steps, namely: (i) we defined 3 alternative models of the genealogical relationships between ancient and current inhabitants of Tuscany (TUS dataset) (Figure 4A); (ii) we generated by serial coalescent simulation millions of gene genealogies for each model; (iii) we summarized genetic diversity in the observed and simulated data by the same set of statistics (Table 2); (iv) by comparing these statistics in the observed and simulated data, we selected a set of simulations best reproducing variation in the data (the number of simulations retained depends on the criterion chosen for the model selection: 100 for the simple rejection procedure and 50,000 for the weighted multinomial logistic regression); and (v) we estimated the models' posterior probabilities (*PP*) by counting how many of the selected simulations were generated under each model (normalizing so that the sum of *PP*s for all models is equal to 1). Demographic (population sizes) and evolutionary (mutation rates) parameters were explored in the simulations within a broad range of possible values defined by priors, and finally estimated from the simulated data.

In total, 24 million simulations were run (1 million for each of 3 models, 4 modern populations in the TUS dataset, and 2 demographic scenarios, respectively including or not including a bottleneck at the time of the Medieval plague epidemics [12]).

We found evidence for genealogical continuity all the way from Etruscan to current times in two contemporary populations (Figure 4A); the posterior probability (*PP*) of Model 1 was between 0.65 and 0.76 for Volterra and 0.95 and 0.99 for Casentino, and this result did not change considering different numbers of best-fitting simulations (say, 500 instead of 100, or 100,000 instead of 50,000). Similar results were obtained incorporating in the model a recent population bottleneck (Figure S5), although an explicit comparison between models with and without plague favoured the latter (Figure 4B). At any rate, the relative success of the models does not depend on the presence of a bottleneck in the late Middle Age. Therefore, this event was not considered in subsequent analyses.

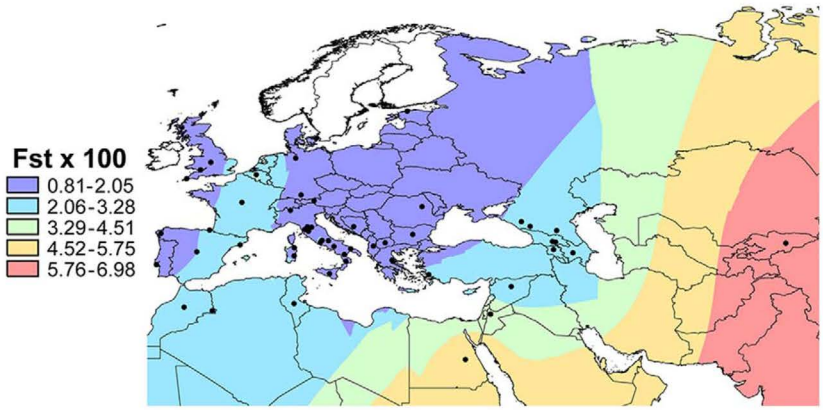


Figure 2. Genetic distances (percent F_{ST} values) between the Etruscan and modern population samples. Different colors represent different levels of genetic differentiation from the Etruscans.
doi:10.1371/journal.pone.0055519.g002

Table 2. Statistics summarizing intra-(A) and inter- (B) population genetic diversity.

A						
	Etruscans	**Medieval**	**Casentino**	**Murlo**	**Volterra**	**Florence**
Number of sequences	30	27	122	86	114	48
Number of distinct Haplotypes	21	14	72	59	57	40
Mean pairwise difference	2.966	1.972	4.105	4.278	3.850	4.152
Haplotype diversity	0.943	0.860	0.976	0.975	0.955	0.980
Segregating sites	24	14	62	64	58	48
B						
Fst						
	Etruscans	**Medieval**	**Casentino**	**Murlo**	**Volterra**	**Florence**
Etruscans	0.000	0.015	0.020	0.010	0.012	0.014
Medieval	0.015	0.000	0.020	0.015	0.013	0.022
Allele sharing						
Etruscans	1.000	0.238	0.333	0.143	0.238	0.095
Medievals	0.357	1.000	0.500	0.214	0.429	0.143

These values were used in the ABC analysis. Allele sharing was calculated as the number of alleles shared between pairs of populations, over the total number of alleles in the ancient sample.

doi:10.1371/journal.pone.0055519.t002

By contrast, for Murlo and Florence, Model 2, with the modern DNAs occupying a distinct branch of the genealogical tree with respect to Etruscans and Medieval Tuscans, was shown to be 7 to 99 times more likely than any alternative model (*PP* between 0.86 and 0.99) (Figure 4A); Model 3 received essentially no support. Choosing different sets of statistics to summarize the data did not change the essence of the results.

We then asked whether there is enough power in the data for these models to be discriminated. To answer, we generated by simulation (separately for Casentino, Murlo, Volterra and Florence) 1,000 pseudo-observed datasets according to each model analyzed (Models 1–3), with parameter values randomly chosen from the correspondent prior distribution. We analyzed these pseudo-observed data with the standard ABC procedure, and counted the fraction of cases in which the model used to generate the data was not recognized, or Type I error. We found that Type I error was always $\leq$0.08 and that the model emerging from the analysis of the observed data (Model 1 for Casentino and Volterra, Model 2 for Murlo and Florence) was correctly identified in at least 95% of cases (Table 3).

Under Model 1, archaic population sizes appear small in both Tuscan populations, with an exponential growth starting around 10,000 years ago for Casentino and 16,500 years ago for Volterra (Figure S6). The estimated mutation rate (around 0.3 mutational events per million years per nucleotide) is in agreement with previous independent reports [9,13]. In general, all the parameters appear well estimated; indeed, their R^2value are always higher than 0.1, an empirical figure generally accepted to be the value beyond which an estimate may be considered reliable [14]. We note that the posterior distribution of the modern effective population sizes drives to the upper limit of the priors (Figure S6). This has also been observed in previous comparable studies [15–17] and reflects the fact that the estimated population size is basically a function of the existing genetic diversity. Clearly, immigration processes have introduced new haplotypes in populations that we had to model as genetically isolated; the resulting excess of diversity is reflected in an increase of the estimated population size. However, in simulations based on the parameters estimated for Model 1 (posterior predictive tests) we succeeded in generating patterns of variation fully compatible with the observed variation; the model's *P*-values (0.332 for Casentino, 0.380 for Volterra) show that the statistics estimated from the observed and simulated data do not differ significantly, and imply that problems related with the estimation of modern population sizes did not undermine the general validity of our approach.

An Etruscan Origin in Anatolia?

Going back to the issue of the Etruscans' origins, if the genetic resemblance between Turks and Tuscans reflects a common origin just before the onset of the Etruscan culture, as hypothesized by Herodotus and as considered in some recent studies [2,6,18], we would expect that the two populations separated about 3,000 years ago. To discriminate between the potentially similar effects of remote common origin and recent gene flow, we ran four independent analyses based on the IM method [19,20]. In the model we tested, the two populations originate from a common ancestor, and may or may not exchange migrants after the split (Figure S7A). Assuming an average generation time of 25 years [16,21] and no migration after the split from the common ancestors, the most likely separation time between Tuscany and Western Anatolia falls around 7,600 years ago, with a 95% credible interval between 5,000 and 10,000 (Figure 5). These results are robust to changes in the proportion of members of the initial population being ancestral to the two modern populations (Figure S7B). We also considered an expanded Anatolian sample (total sample size = 123 [11,22]) coming from all over Turkey, to test whether a founder effect might have enhanced the role of the genetic drift in the previous analysis, inflating the divergence time estimates; the resulting distributions of separation times completely overlapped with those previously estimated, with a lower bound of

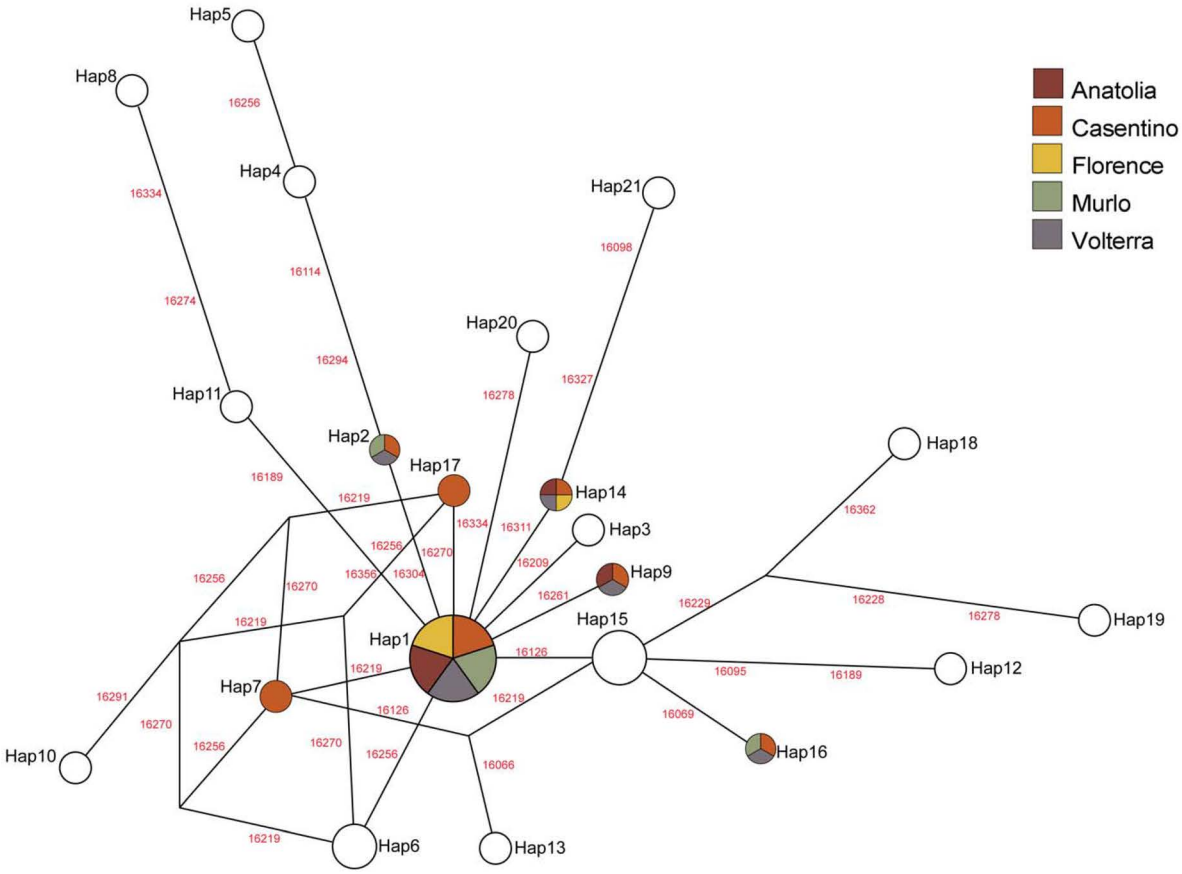


Figure 3. Median-joining network of the Etruscans' haplotypes. The width of the circles is proportional to the frequency of that haplotype in the Etruscan sample; the labels on the edges of the network indicate the position of the nucleotide substitution in the mtDNA reference sequence. The colour of each haplotype represents whether that sequence is also present in five modern populations from Tuscany and Anatolia.
doi:10.1371/journal.pone.0055519.g003

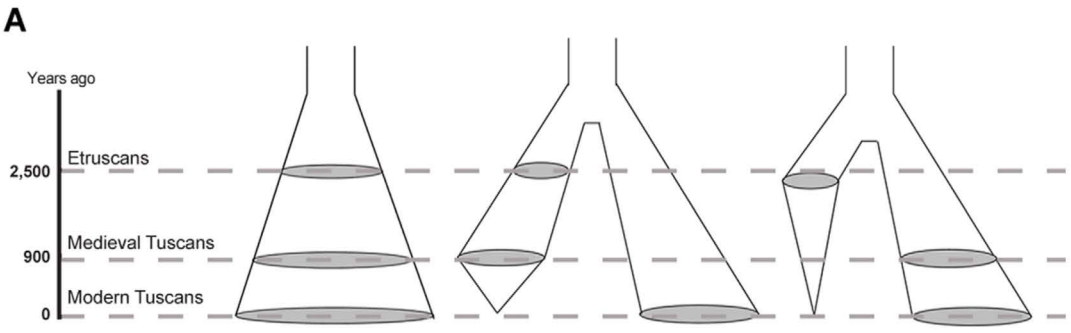
the 95% credible interval never smaller than 5,300 years ago (Figure 5).

For these tests we chose the mutation rate (μ) estimated from the data in the previous ABC analyses (very close to the figure accounting for the time-dependency of the mitochondrial molecular clock [13], $\mu = 0.003$). Tests were also run using the value incorporating a correction for the effects of purifying selection [23] ($\mu = 0.0014$), always finding that it results in a further increase of the estimated separation times (Figure S7B). Only assuming very high mutation rates, at least twice as large as estimated in Henn et al. [13], was it possible to obtain separation times <5,000 years (Figure S7B). With both Anatolian samples, any degree of gene flow after separation between the ancestors of Tuscans and Anatolians resulted in more remote separation times.

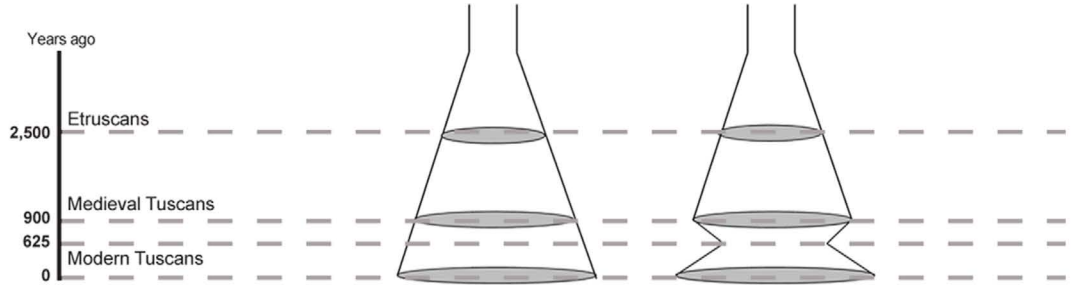
Discussion

MtDNA data give much stronger support to a model of genetic continuity between the Etruscans and some Tuscans than to any other model tested, characterized by plausible population sizes and mutation rates. However, this clear picture emerges only when modern Tuscan communities are separately considered, highlighting the importance of population structure even at the small geographical scale. In a previous analysis of smaller samples we found no evidence of genealogical continuity since Etruscan times [5]. In this study, the larger sample sizes allowed us to separately investigate the relationships of each modern population with the Etruscans. A model of genealogical continuity across 2,500 years thus proved to best fit the observed data for Volterra, and especially Casentino, but not for another community dwelling in an area also rich with Etruscan archaeological remains (Murlo), nor (as expected) for the bulk of the current Tuscan population, here represented by a forensic sample of the inhabitants of Florence. Therefore, the present analysis indicates that the Etruscan genetic heritage is still present, but only in some isolates, whereas current Tuscans are not generally descended from Etruscan ancestors along the female lines. It also shows that there is no necessary correlation between the presence of archaeological remains and the biological roots of the inhabitants of the areas where these remains occur. Because Medieval Tuscans appears directly descended from Etruscan ancestors, one can reasonably speculate that the genetic build-up of the Murlo and Florence populations was modified by immigration in the last five centuries.

As for the second question, the IM analysis shows that indeed there might have been a genealogical link between modern Tuscans and the inhabitants of what Herodotus considered the Etruscans' homeland, Western Anatolia. However, even under the



	Model 1	Model 2	Model 3
Casentino			
LR	0.986	0.012	0.002
AR	0.950	0.040	0.010
Murlo			
LR	0.012	0.988	0.000
AR	0.060	0.940	0.000
Volterra			
LR	0.757	0.227	0.016
AR	0.650	0.330	0.020
Florence			
LR	0.020	0.980	0.000
AR	0.140	0.860	0.000



	Model 1, No plague	Model 1, plague
Casentino		
LR	0.916	0.084
AR	0.840	0.160
Volterra		
LR	0.799	0.201
AR	0.810	0.190

Figure 4. Alternative models of the genealogical relationships among past and present populations, and their posterior probabilities. Shaded areas represent the modern population (at 0 years ago on the Y axis), the Medieval population (900 years ago) and the Etruscans (at 2,500 years ago). Model 1 assumes genealogical continuity between ancient and modern samples, Model 2 assumes continuity only between Etruscan and Medieval individuals, and in Model 3 the Etruscan lineage separates from the lineage leading to Medieval and Modern Tuscans. Under each model is the proportion of the best-fitting simulations supporting it, for the four modern populations considered, using the acceptance rejection (AR) and logistic regression (LR) methods [43]. (A) Comparison among Models 1–3 for four modern Tuscan populations. (B) Comparison of the fit of Model 1, with and without a bottleneck corresponding to the Plague epidemics at 625 BP [12].
doi:10.1371/journal.pone.0055519.g004

unrealistic assumption of complete reciprocal isolation for millennia, the likely separation of the Tuscan and Anatolian gene pools must be placed long before the onset of the Etruscan culture, at least in Neolithic times; if isolation was incomplete, the estimated separation must be placed further back in time. Consistent with this view is the observation that Etruscan and

Table 3. Type I errors for the 3 Models in the 4 Tuscan samples.

Simulated Model					
CASENTINO					
		MOD 1	**MOD 2**	**MOD 3**	**Type I error**
	MOD 1	**0.98**	0.00	0.02	0.02
	MOD 2	0.01	**0.99**	0.00	0.01
	MOD 3	0.02	0.00	**0.98**	0.02
MURLO					
		MOD 1	**MOD 2**	**MOD 3**	**Type I error**
	MOD 1	**0.95**	0.01	0.04	0.05
	MOD 2	0.02	**0.98**	0.00	0.02
	MOD 3	0.07	0.00	**0.93**	0.07
VOLTERRA					
		MOD 1	**MOD 2**	**MOD 3**	**Type I error**
	MOD 1	**1.00**	0.00	0.00	0.00
	MOD 2	0.07	**0.93**	0.00	0.07
	MOD 3	0.05	0.00	**0.95**	0.05
FLORENCE					
		MOD 1	**MOD 2**	**MOD 3**	**Type I error**
	MOD 1	**0.92**	0.03	0.05	0.08
	MOD 2	0.04	**0.95**	0.01	0.05
	MOD 3	0.05	0.01	**0.94**	0.06

For each of the modern populations listed on the Y axis, data were simulated according to three models and attributed by the LR procedure to one of the models on the X-axis. The power of the procedure in recovering the correct model is represented by the rates of correct attribution (along the main diagonal; shaded cells); the last column (Type I error) represents the fraction of cases in which the correct model was not identified.

doi:10.1371/journal.pone.0055519.t003

Neolithic mtDNAs are close to each other in the two-dimensional plot of Figure S4C; however, a formal test would be necessary to draw firm conclusions from the simple observation of a genetic similarity. Separation times were very close when estimated both using a sample from Western Anatolia, and an expanded sample including individuals from much of Anatolia, and so the choice of the Anatolian population does not seem to affect the results of this analysis.

A general problem in ancient human DNA studies is the quality of the data; errors resulting from contamination, or from poor preservation of DNA in the specimens, are common. However, there are several reasons to be confident that the Etruscan sequences obtained in this study are authentic: (i) bones were recovered from burials according to the most stringent existing procedures and sent directly to the ancient DNA laboratory without manipulations; (ii) the mtDNA HVR-I motifs of the people who came in contact with the bones at any stage of the analysis do not match those obtained from the ancient samples (Table S1); (iii) the ancient samples were typed following the most stringent standard criteria for ancient DNA authentication; (iv) we used two different sequence determination procedures (classical methodology and high throughput methodology) and the results obtained from different extractions and different sequencing methodologies are concordant except in the regions of homopolymeric strings $\geq$5 bp that are problematic for the 454 pyrosequencing technology; in these cases, consensus sequences were determined considering only the results of the standard sequencing procedure; (v) sequences make phylogenetic sense, i.e. do not appear to be combinations of different sequences, possibly suggesting contamination by exogenous DNA.

Using such ancient DNA data for testing complex evolutionary models has become possible with the development of ABC and other recent Bayesian inference methods [24,25]. These models, albeit more articulate than those that can be tested otherwise, are still a necessarily schematic representation of the processes affecting populations in the course of millennia. Many phenomena that could not be incorporated in the models, such as immigration from other sources or additional demographic fluctuations, most likely occurred and left a mark in the patterns of genetic diversity. In addition, specific phenomena may have involved mostly or exclusively males, resulting in genetic changes that are not recorded in mtDNA variation. Still, if we rule out the unlikely hypothesis that the Etruscans' and their descendants' population history was radically different for males and females, the picture emerging from this study is rather clear. The additional tests we ran (Type I error, Table 3) show that, at these sample sizes, we had a high probability to identify the correct evolutionary model.

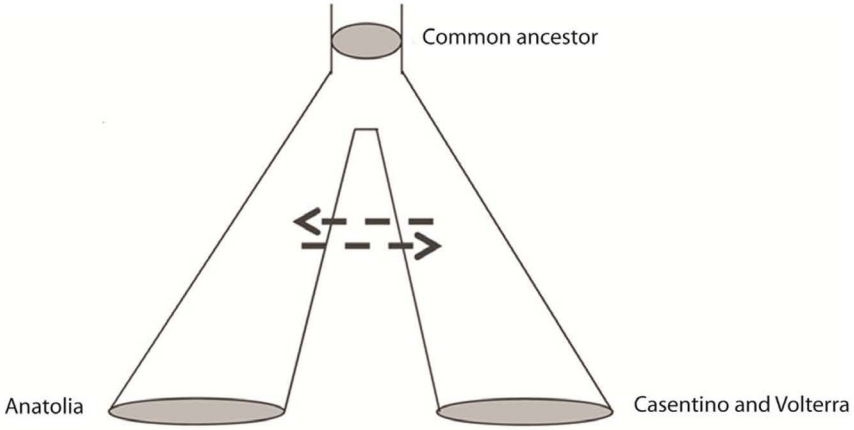
As also suggested by the analysis of skull diversity [26], contacts between people from the Eastern Mediterranean shores and Central Italy likely date back to a remote stage of prehistory, possibly to the spread of farmers from the Near East during the Neolithic period [27,28], but not necessarily so (we only estimated a minimum separation time between gene pools). At any rate, these contacts occurred much earlier than, and hence appear unrelated with, the onset of the Etruscan culture (Figure 5). We conclude that no available genetic evidence suggests an Etruscan origin outside Italy. While their culture disappeared from the records, the Etruscans' mtDNAs did not; traces of this heritage are still recognizable. However, most current inhabitants of the ancient Etruscan homeland appear descended from different ancestors along the female lines, as clearly shown by the analysis of the urban (Florence) sample. Genetic continuity since the Etruscan's time is still evident only in relatively isolated localities, such as Casentino and Volterra.

Materials and Methods

DNA Extraction and Characterization of the Etruscan Samples

We obtained 18 bone samples (each represented by two fragments of the right tibia) from a multiple burial from Casenovole, Southern Tuscany, near Grosseto. Their approximate age, based on archaeological evidence, is the 3rd century BC. The permit to genetically characterize these fossil samples came from Soprintendenza Archeologica per la Toscana (Archaeological Authority for Tuscany), Siena. The bone fragments were freshly excavated and collected according to the most stringent ancient DNA criteria [29] by one of us (EP) and can safely be regarded as belonging to different individuals (Minimum number of individuals estimated in the burial = 21). These fragments were processed in the ancient DNA facilities at the University of Florence using standard ancient DNA procedures [30]. After a first round of DNA extraction, the samples were subjected to multiple PCRs, cloning and cycle sequencing.

In a successive step, DNA was independently reextracted from the samples that had given positive results in the previous analysis. In this case, after multiple PCRs, the amplicons were not cloned but ligated to the appropriate adaptor sequences and directly sequenced with 454/Roche technology. Low Molecular Weight DNA (LMW DNA) 454/Roche protocol was applied and a final procedure modification was added to increase the recovery of



Western Anatolian sample	Run 1	Run 2	Run3	Run 4
Mean	7,775	7,980	7,663	6,692
95% LowB CI	5,572	5,945	5,665	6,002
95% UppB CI	9,958	10,556	10,444	9,861

Expanded Anatolian sample	Run 1	Run 2	Run3	Run 4
Mean	6,680	6,711	7,065	7,140
95% LowB CI	5,535	5,329	5,758	5,815
95% UppB CI	8,671	9,343	9,642	8,558

Figure 5. Separation time estimated by the IM model. Estimation of the separation time between the gene pools of Anatolians (whether only Western Anatolians, or the expanded sample) and contemporary Tuscans (Casentino and Volterra). Means, upper bound and lower bound of the 95% credible intervals in four independent runs, obtained fixing the migration rate (indicated by dashed arrows) at 0, with mutation rate = 0.003 and assuming that the proportion of the ancestral population is equal in each descendant population (i.e. s = 0.5). Each analysis consisted of five coupled Markov chains, and 10,000,000 steps. Any degree of gene flow between the ancestors of Anatolians and Tuscans results in an increase of the estimate of the time since the population separation.
doi:10.1371/journal.pone.0055519.g005

a single stranded library [31]. Libraries were quantitated using a quantification Real Time PCR (qPCR) by KAPA Library Quant Kits (KAPA Biosystems, MA, USA). Samples libraries were independently amplified on beads by emulsion PCR (emPCR), then enriched and counted beads were loaded onto 454/Roche PicoTiterPlate (PTP) divided in 16 regions. Sequencing was performed as in 454/Roche protocol and the obtained reads were filtered and mapped using the Cambridge reference sequence [32]. For each sample and amplicon, a masking procedure allowed to remove primer sequences from the reads and obtain a multi-alignment using the 454/Roche Amplicon Variant Analysis (AVA) software. A consensus was generated by custom scripting and then mapped on the mitochondrial DNA reference sequence (GenBank accession number: J01415). Complete mtDNA HVR-I sequences could be retrieved in all samples. At each site the most frequent nucleotide was observed in a range of 97.7–98.8% of the reads in the different samples. Unmapped reads were then analyzed in order to characterize them and we found that they are mostly primer dimers. Final consensus sequences of the 10 samples were determined by comparing results obtained from both standard procedures (575 Clones) and Next Generation Sequencing (127,837 reads).

Four additional samples from Tarquinia, sequenced in 2004, but never published so far, brought to 14 the total of Etruscan samples typed for this study.

Ancient and Modern mtDNA Diversity

In all statistic analyses, we replaced the nucleotides occupying position 16180–16188 and 16190–16193 with the nucleotides in the CRS, because they contain two stretches of Adenines and Citosines known to result in apparent length polymorphism of the mtDNA sequence [33,34]. Summary statistics were estimated by Arlequin ver. 3.5.1 [35]. The Fst values between the populations in the EUR dataset and the Etruscans were interpolated in a map representing using the Spatial Analyst extension in ArcGIS 10 (ESRI; Redlands, CA, USA) using the Kriging procedure. Genetic distances between the Etruscans and each population in the ANC, TUS and EUR datasets were visualized by Multidimensional Scaling (MDS), using the *cmdscale* function in the R environment [36].

Approximate Bayesian Computation

Inferring demographic and evolutionary processes from genetic data requires the testing of models which are often too complex for their likelihoods to be derived. Approximate Bayesian Computation (ABC) [37] offers a valid alternative. Summary statistics estimated from the data are compared with those generated by simulation, and posterior distributions of the models' parameters can be approximated by simulating large numbers of gene genealogies. We generated gene genealogies in which individuals are sampled at different moments in time using the Bayesian version of SERIALSIMCOAL [38]. At every iteration, the

parameters of the model (population sizes, mutation rates, timing of demographic processes) were considered as random variables, and their values were extracted from broad prior distributions; ages and sizes of the samples were equal to those of the observed samples. We then calculated a Euclidean distance between observed and simulated statistics, and we ordered the simulations according to this distance. In total, 24 million simulations were run (1 million for each of 3 models, 4 modern populations in the TUS dataset and two demographic scenarios, respectively including or not including a recent bottleneck). All the procedures were developed in the R environment [36] using scripts from [39]. We selected the summary statistics via PCA, keeping for the ABC analysis those statistics which have shown to be more correlated with the parameters' variance (Table S2).

Demographic Models and Priors

The three demographic models tested differ for the relationships between modern and ancient samples (Figure 4); under each model, each population in the TUS dataset was independently compared with the Etruscan and Medieval populations. All prior distributions were uniform and wide. The effective modern population size ranged between 100 and 200,000; for the time of the onset of the expansion (under Model 1) and the separation time (under Models 2 and 3) the priors ranged from 101 (one generation before the Etruscans) to 1,500 generations ago. Priors for the mutation rate encompassed the low value estimated from phylogenies [40], and the high value estimated from pedigrees [41], from 0.0003 to 0.0075 mutations per generation for HVR-I. The Medieval and the Etruscan effective population sizes were extracted from a prior distribution spanning from 100 to 50,000, as suggested in Guimaraes et al. [5]. Ancestral population sizes varied from 5 to 6,000 individuals. The entire procedure was repeated under a demographic scenario including a population bottleneck corresponding to the 14th century plague epidemics, in which an estimated one-third of the population was lost [42].

Model Selection and Parameter Estimation

The posterior probabilities of the 24 combinations of models (3), modern populations (4) and demographic scenarios (2), were calculated either: (i) by a simple rejection procedure (AR) [43] for which we retained the 100 simulations associated with the shortest distance between observed and simulated statistics [44]; or (ii) by a weighted multinomial logistic regression (LR) [44] for which we retained the 50,000 simulations generating the shortest distance between the observed and simulated statistics. In both cases, we normalized the PPs so that their sum for all models being compared is 1. The parameters of the best-fitting model were estimated from the 2,000 simulations closest to the observed dataset, after a *logtan* transformation of the parameters [45] and according to Beaumont [37].

Additional Tests: Type I Error and Posterior Predictive Tests

We estimated the probability that the true null hypothesis be rejected by evaluating the Type I Error, i.e. the proportion of cases in which 1,000 pseudo-datasets generated under each model are not correctly identified by the ABC analysis. In addition, to test whether the data can be actually reproduced under a specific demographic model, we carried out a posterior predictive test [9,25]. For that purpose, we simulated 10,000 datasets according to the model with the highest probability using the estimated posterior parameter distribution, and we calculated a posterior predictive P-value for each statistic; these probabilities were then combined into a global P-value, taking into account their non-independence [46].

The Isolation with Migration (IM) Model

We estimated the likely separation time between the Tuscan and Anatolian gene pools by Isolation with Migration (IM), a method generating posterior probabilities for complex models in which populations need not be at equilibrium [19]. Seven parameters were estimated from the data, namely the size of the ancestral and daughter populations (N_A, N_1, N_2), the rates of gene flow between daughter populations (m_1, m_2), the time since the split (t), and the proportion of the members of the ancestral population giving rise to the first daughter population (s) [47]. Because any degree of genetic exchange increases the t estimate, after some preliminary tests we set to 0 the values of m_1 and m_2. Most tests were run fixing the mutation rate at the value estimated in the ABC analysis (0.003 mutational events per locus per generation), but we repeated the whole IM analysis with both lower and higher values (respectively, 0.0014 and 0.0060 mutational events per locus per generation; [13,23]) under a Hasegawa-Kishino-Yano (HKY; [48]) mutational model with inheritance scalar 0.25, as recommended for mtDNA data. For each mutation rate tested we ran several analyses starting from different random seeds, in order to assess the consistency of the results; moreover, to improve the exploration of the parameters' space, and thereby the convergence, we coupled the Markov chains, running simultaneously 5 chains per run.

Supporting Information

Figure S1 Amplicons of the 10 sequences from Casenovole. DNA sequences from the575 clones analysed for the 10 Casenovole Etruscan samples. The sequences of the external primers are not reported in the figure. The Cambridge reference sequence with the numbering of the nucleotide positions is at the top. Nucleotides identical to the Cambridge reference sequence are indicated by dots. The clones are identified by a code (from S1 to S17, indicating the individual), the first number is the extraction, the second number is the PCR.
(PDF)

Figure S2 Results of the mapping step for the 10 Etruscan samples analyzed. (A) The number of sequences that map to the reference and those that do not map is plotted as a histogram. Some samples had a large amount of unmapped reads that were afterwards characterized as primers' dimers. (B) Frequency distribution (% on the Y axis) of the frequency of the most frequent nucleotide for the 10 Etruscan samples analyzed (the upper limits of the % intervals are reported in the legend). For example, in sample S1 at around 84% of the positions the frequency of the most frequent allele among reads is between 99% and 100%.
(PDF)

Figure S3 Measures of genetic distance. Allele sharing (A) and Fst (×100) (B) in 52 modern populations of Western Eurasia and the Mediterranean basin. Population labels and sample sizes are provided in Table S2. Allele sharing estimated as the number of sequences shared between Etruscans and every modern population, divided by the sample size of the modern sample.
(TIF)

Figure S4 Multi Dimensional Scaling. Multi Dimensional Scaling summarizing genetic affinities between the Etruscans and (A) 52 modern populations of Western Eurasia and the Mediterranean basin; (B) Medieval and modern Italian popula-

tions; (C) 9 ancient populations of Europe. Population labels and sample sizes are provided in Table S2.
(PDF)

Figure S5 Results of model selection. Results of model selection with or without a bottleneck representing the plague epidemics at 625 BP, in Casentino, Murlo and Volterra. Dashed lines represent the presence of plague epidemic that killed one third of the population. For each sample we report the posterior probabilities calculated comparing Models 1–3, either considering or disregarding this demographic event.
(PDF)

Figure S6 Parameter estimates and posterior distributions under Model 1, for Casentino (A) and Volterra (B). Upper panels: Prior distributions (all the priors were uniform), median and mode estimates, the 95% of the highest posterior density (lower and upper bound), and coefficient of determination R^2. The time is expressed in years, the mutation rate in number of mutational events per generation per locus. Lower panels: histograms and smoothed distributions of the parameters estimated.
(PDF)

Figure S7 IM model (A) and estimates (B) for the separation time between Anatolians and Tuscans. N_1 and N_2: modern population size; N_A: ancestral population size; m_1 and m_2: migration rates; s: proportion of the ancestral population that founds descendent population 1; t: separation time. Different mutation rates and proportions of the ancestral population founding the descendant populations were considered.
(PDF)

Table S1 Consensus HVR-I Etruscans mtDNA and sequences of all the investigators. Upper panel: Consensus HVR-I mtDNA sequences in 30 individuals from historical Etruria. Tarq represents individuals from Tarquinia, Cas from Casenovole, Vol from Volterra, Pie from Castelluccio di Pienza, Sot from Castelfranco di Sotto and MM from Magliano and Marsiliana. CRS is the Cambridge reference sequence [32]. The HVR-I motif is the position (−16,000) where substitution were observed, with respect to the CRS; the observed transversions are indicated with a capital letter. The haplotypes shared with EUR dataset are in bold type. For the Casenovole sample, the labels of the individuals used in Figure S1 are between parentheses. Lower panel: Sequences of all the investigators who had direct contact with the ancient specimens.
(DOCX)

Table S2 Detailed description of the samples in the EUR and ANC datasets.
(DOC)

Acknowledgments

Computational support for the data analysis has been provided by CINECA (Bologna) and CASPUR (Roma) HPC facilities. We thank Carlo Previderé for sharing with us unpublished data, Sibelle Vilaça for her help with the graphics, Alessandro Achilli, Andrea Benazzo, Mathias Currat, Martin Richards and especially Stefano Mona for discussion and suggestions.

Author Contributions

Conceived and designed the experiments: SG DC GB. Performed the experiments: SG FT EF AS ML SV EP GC ER GDB. Analyzed the data: SG FT EF VC. Wrote the paper: SG DC GB.

References

1. Barker G, Rasmussen T (1998) The Etruscans. Oxford: Blackwell.
2. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, et al. (2007) Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet 80: 759–768.
3. Vernesi C, Caramelli D, Dupanloup I, Bertorelle G, Lari M, et al. (2004) The Etruscans: a population-genetic study. Am J Hum Genet 74: 694–704.
4. Belle EM, Ramakrishnan U, Mountain JL, Barbujani G (2006) Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. Proc Natl Acad Sci U S A 103: 8012–8017.
5. Guimaraes S, Ghirotto S, Benazzo A, Milani L, Lari M, et al. (2009) Genealogical discontinuities among Etruscan, Medieval, and contemporary Tuscans. Mol Biol Evol 26: 2157–2166.
6. Pellecchia M, Negrini R, Colli L, Patrini M, Milanesi E, et al. (2007) The mystery of Etruscan origins: novel clues from Bos taurus mitochondrial DNA. Proc Biol Sci 274: 1175–1179.
7. Bandelt HJ (2004) Etruscan artifacts. Am J Hum Genet 75: 919–920; author reply 923–917.
8. Mateiu LM, Rannala BH (2008) Bayesian inference of errors in ancient DNA caused by postmortem degradation. Mol Biol Evol 25: 1503–1511.
9. Ghirotto S, Mona S, Benazzo A, Paparazzo F, Caramelli D, et al. (2010) Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. Mol Biol Evol 27: 875–886.
10. Turchi C, Buscemi L, Previdere C, Grignani P, Brandstatter A, et al. (2008) Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int J Legal Med 122: 199–204.
11. Di Benedetto G, Erguven A, Stenico M, Castri L, Bertorelle G, et al. (2001) DNA diversity and population admixture in Anatolia. Am J Phys Anthropol 115: 144–156.
12. Livi-Bacci M (2007) A concise history of world population. Oxford: Blackwell.
13. Henn BM, Gignoux CR, Feldman MW, Mountain JL (2009) Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. Mol Biol Evol 26: 217–230.
14. Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, et al. (2008) Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. Mol Ecol 17: 757–772.
15. Belle EM, Benazzo A, Ghirotto S, Colonna V, Barbujani G (2009) Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations. Heredity 102: 218–225.
16. Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, et al. (2007) Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A 104: 17614–17619.
17. Laval G, Patin E, Barreiro LB, Quintana-Murci L (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS One 5: e10284.
18. Brisighelli F, Capelli C, Alvarez-Iglesias V, Onofri V, Paoli G, et al. (2009) The Etruscan timeline: a recent Anatolian connection. Eur J Hum Genet 17: 693–696.
19. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167: 747–760.
20. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158: 885–896.
21. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128: 415–423.
22. Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, et al. (2004) Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corrido. Am J Hum Genet 74: 827–845.
23. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740–759.
24. Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19: 2609–2625.
25. Gelman A, Carlin J, Stern H, Rubin D (2004) Bayesian Data Analysis. Boca Raton, Florida: CRC Press.
26. Claassen H, Wree A (2004) The Etruscan skulls of the Rostock anatomical collection–how do they compare with the skeletal findings of the first thousand years B. C.? Ann Anat 186: 157–163.
27. Barker G (2006) The Agricultural revolution in prehistory: Why did foragers become farmers?. Oxford: Oxford University Press.
28. Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, et al. (2011) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci U S A 108: 9788–9791.
29. Caramelli D, Lalueza-Fox C, Condemi S, Longo L, Milani L, et al. (2006) A highly divergent mtDNA sequence in a Neandertal individual from Italy. Curr Biol 16: R630–632.

30. Caramelli D, Milani L, Vai S, Modi A, Pecchioli E, et al. (2008) A 28,000 years old Cro-Magnon mtDNA sequence differs from all potentially contaminating modern sequences. PLoS One 3: e2700.
31. Maricic T, Paabo S (2009) Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. Biotechniques 46: 51–52, 54–57.
32. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.
33. Bandelt HJ, Kivisild T (2006) Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. Ann Hum Genet 70: 314–326.
34. Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57: 248–256.
35. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10: 564–567.
36. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. http://www.R-project.org. Vienna, Austria: Foundation for Statistical Computing. Accessed 2013 January 3.
37. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.
38. Anderson CN, Ramakrishnan U, Chan YL, Hadly EA (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. Bioinformatics 21: 1733–1734.
39. PopABC website. Available at: http://code.google.com/p/popabc/source/browse/#svn%2Ftrunk%2Fscripts. Accessed 2013 January 3.
40. Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6: 165–183.
41. Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, et al. (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72: 659–670.
42. Biraben J-N (1979) Essai sur l'evolution du nombre des hommes. Population (French ed) 34: 13–25.
43. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16: 1791–1798.
44. Beaumont M (2008) Joint determination of topology, divergence time and immigration in population trees. Simulations, genetics and human prehistory. Cambridge: McDonald Institute for Archaeological Research. 135–154.
45. Hamilton G, Stoneking M, Excoffier L (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc Natl Acad Sci U S A 102: 7476–7480.
46. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, et al. (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A 102: 18508–18513.
47. Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol 3: e193.
48. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.

PAPER V: Genetic Evidence Does Not Support an Etruscan Origin in Anatolia

AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 152:11–18 (2013)

Genetic Evidence Does Not Support an Etruscan Origin in Anatolia

Francesca Tassi,[1] Silvia Ghirotto,[1] David Caramelli,[2] and Guido Barbujani [1]*

[1]*Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy*
[2]*Department of Evolutionary Biology, University of Firenze, Ferrara, Italy*

KEY WORDS ancient DNA; mitochondrial DNA; coalescent simulations; approximate Bayesian computation

ABSTRACT The debate on the origins of Etruscans, documented in central Italy between the eighth century BC and the first century AD, dates back to antiquity. Herodotus described them as a group of immigrants from Lydia, in Western Anatolia, whereas for Dionysius of Halicarnassus they were an indigenous population. Dionysius' view is shared by most modern archeologists, but the observation of similarities between the (modern) mitochondrial DNAs (mtDNAs) of Turks and Tuscans was interpreted as supporting an Anatolian origin of the Etruscans. However, ancient DNA evidence shows that only some isolates, and not the bulk of the modern Tuscan population, are genetically related to the Etruscans. In this study, we tested alternative models of Etruscan origins by Approximate Bayesian Computation methods, comparing levels of genetic diversity in the mtDNAs of modern and ancient populations with those obtained by millions of computer simulations. The results show that the observed genetic similarities between modern Tuscans and Anatolians cannot be attributed to an immigration wave from the East leading to the onset of the Etruscan culture in Italy. Genetic links between Tuscany and Anatolia do exist, but date back to a remote stage of prehistory, possibly but not necessarily to the spread of farmers during the Neolithic period. Am J Phys Anthropol 152:11-18, 2013.

The Etruscan civilization is documented in Etruria, roughly corresponding to current Tuscany, starting from the eighth century BC, and is defined by a material culture, by a non-Indo-European language and by an alphabet derived from the Greek alphabet. Availability of copper and iron and ability in seafaring were the main factors leading to an Etruscan expansion over much of Central Italy in the sixty and fifth centuries BC. Later, military defeats and the Roman expansion caused a decline of the Etruscans' political influence. From the first century BC, the Etruscan language disappeared from the archeological record (Barker and Rasmussen, 1998).

Questions about the Etruscans' origins date back to antiquity and are still open. In the fifth century BC, Herodotus described them as a group emigrating from Lydia, in Western Anatolia; by contrast, Dionysius of Halicarnassus regarded them as an indigenous Italic population, and this view is shared by most modern archeologists. The first genetic studies assumed that current inhabitants of Tuscany are the direct mitochondrial descendants of the Etruscans, and their results suggested an evolutionary link with Anatolia that would support Herodotus' view (Achilli et al., 2007; Brisighelli et al., 2009). However, analyses of mitochondrial DNA (mtDNA) sequences from bones excavated in Etruscan necropolis (Vernesi et al., 2004) raised questions about the significance of the similarities observed between modern populations. Indeed, based on ancient DNA data, the Etruscans appeared to represent a single biological population, connected by genetic links across its territory, but showed a limited genetic resemblance with modern people of the same area (Vernesi et al., 2004). Explicit tests comparing mtDNAs of ancient (i.e., Etruscan) and modern inhabitants of Tuscany ruled out the hypothesis that the former might be the latter's direct ancestors (Belle et al., 2006; Guimaraes et al., 2009). Among the possible explanations for these results are the presence of massive errors in the Etruscan sequences (as suggested by Bandelt and Kivisild, 2006), a complete population extinction, and a persistence of the Etruscans' genetic heritage only in isolated localities, whereas most contemporary Tuscans would be descended from different mitochondrial ancestors.

Two recent studies contributed to clarifying this complex picture. First, Bayesian analysis of patterns of mutation showed no evidence of systematic errors in the ancient Etruscan sequences, which might explain the observed differences between modern and ancient inhabitants of Etruria (Mateiu and Rannala, 2008). Errors are always possible and sometimes hard to identify in ancient DNA analyses, but as far as one can test, Vernesi et al.'s (2004) Etruscan sequences comply with the highest quality standards. Second, in a study including a new set of ancient DNA sequences, in which alternative demographic models were explicitly tested by Approximate Bayesian Computation (ABC), the results were compatible with a genealogical continuity between

Grant sponsor: University of Ferrara; Italian Ministry for University and Research (MIUR), PRIN 2012 funds.

*Correspondence to: G. Barbujani, Department of Life Sciences and Biotechnology, University of Ferrara, via Borsari 46, 44121 Ferrara, Italy. E-mail: g.barbujani@unife.it

Received 11 January 2013; accepted 20 May 2013

DOI: 10.1002/ajpa.22319
Published online 30 July 2013 in Wiley Online Library (wileyonlinelibrary.com).

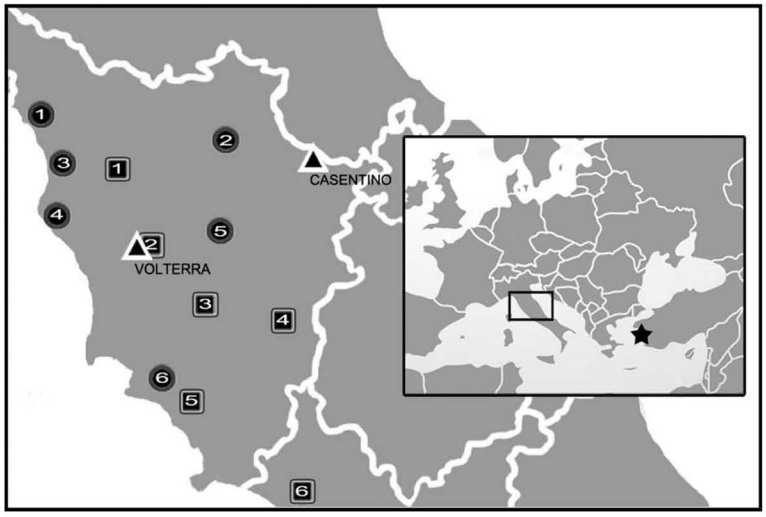


Fig. 1. Geographic location of the samples considered in the ABC analysis. Triangles, contemporary Tuscans (n = 236); Circles, Medieval Tuscans: 1. Massa Carrara (n = 3), 2. Florence (n = 10), 3. Pisa (n = 6), 4. Livorno (n = 3), 5. Siena (n = 4); 6. Grosseto (n = 1); Squares, Etruscans: 1. Castelfranco di Sotto (n = 1); 2. Volterra (n = 3); 3. Casenovole (n = 10); 4. Castelluccio di Pienza (n = 1); 5. Magliano/Marsiliana (n = 6); 6. Tarquinia (n = 9); Star, Turks (n = 35).

the Etruscans and two Tuscan isolates (Volterra and Casentino) that had not been previously compared with them. By contrast, another population of the former Etruscan homeland, Murlo, and a forensic sample from the main city in the area, Florence, showed no special relationships with the Etruscans (Ghirotto et al., 2013). The former finding means that populations separated by short distances may differ in their genetic relationships with ancient populations (as already seen in Sardinia; Ghirotto et al., 2010), and the latter confirms that the Etruscans cannot be regarded as the global ancestors of the people now living in what once was their territory.

In this article, we used all the available ancient mtDNA samples from classical Etruria, the inferential power given by the ABC methods, and the information on the genealogical relationships between the Etruscans and the communities of Volterra and Casentino (Ghirotto et al., 2013) to further investigate the Etruscans' biological origins. Because previous inhabitants of Etruria, associated with the Villanovian culture, cremated their dead, empirical genetic comparisons going further back in time are unfeasible. We then compared the observed genetic data with the results of millions of simulations of modern and ancient mtDNAs, generated under demographic models differing for the homelands of Etruscan people, namely, Western Anatolia or Central Italy. This way, we could test whether or not the genetic links between modern Anatolians and Tuscans may have been established through a process of gene flow occurring approximately between the tenth and eight centuries BC, and thus possibly associated with the onset of the Etruscan civilization in Italy.

MATERIALS AND METHODS

Genetic data

Historical Etruria comprises much of current Tuscany and the northernmost region of current Latium (Fig. 1). Therefore, in this study, we excluded previously published specimens coming from the regions of Etruscan expansion to the North (Adria) and to the South (Capua), because (a) we had no idea of the levels of admixture with non-Etruscan people in these localities and (b) there was no reason to assume that these populations could have contributed to the ancestry of modern Tuscans.

We analyzed 360 bp [positions 16024–16383 of the Cambridge reference sequence (CRS) (Andrews et al., 1999)] within the HVRI (hypervariable I) region of mtDNA. In all statistical analyses, we replaced the nucleotides occupying position 16180–16188 and 16190–16193 with the nucleotides in the CRS, to avoid the stretch of adenines and cytosines known to result in apparent length polymorphism of the mtDNA sequence (Bendall and Sykes, 1995; Bandelt and Kivisild, 2006).

We considered samples of three main historical periods: the Etruscans (with specimens dated around 2,500 years ago, on average), the medieval Tuscans (dated around 900 years ago), and modern subjects. The Etruscan sample is composed of 30 sequences from different necropolis (Vernesi et al., 2004; Ghirotto et al., 2013). The Medieval sample comprises 27 sequences collected in various Tuscan localities (Guimaraes et al., 2009). The modern sample comprises the following: (a) two Tuscan populations [Casentino, 122 sequences and Volterra, 114 sequences (Achilli et al., 2007)] for which we previously demonstrated a high level of genealogical continuity since Etruscan times (Ghirotto et al., 2013) and (b) a population from Western Anatolia [35 sequences (Di Benedetto et al., 2001)], representing the putative Etruscans' homeland according to Herodotus.

Summary statistics

In this study, statistics summarizing genetic diversity were calculated by using Arlequin ver. 3.5.1. (Excoffier and Lischer, 2010). The within-population diversity of each sample is described by (i) sample size, (ii) number

TABLE 1. Statistics summarizing (A) intra- and (B) interpopulation genetic diversity

A					
	Etruscans	Medieval	Casentino	Volterra	Turks
No. of sequences	30	27	122	114	35
No. of distinct Haplotypes	21	14	72	57	29
Mean pairwise difference	2.966	1.972	4.105	3.850	4.689
Haplotype diversity	0.943	0.860	0.976	0.955	0.965
Segregating sites	24	14	62	58	43
B					
Fst					
Etruscans	0.000	0.015	0.020	0.012	0.033
Medieval	0.015	0.000	0.020	0.013	0.045
Allele sharing					
Etruscans	1.000	0.238	0.333	0.238	0.143
Medieval	0.357	1.000	0.500	0.429	0.286

These values were used in the ABC analysis.

of different haplotypes, (iii) mean pairwise difference, (iv) haplotype diversity, and (v) number of segregating sites (Table 1a). In addition, we quantified the relationships between Tuscans and Anatolians calculating, for each comparison, Hudson's F_{ST} (Hudson et al., 1992) an index of genetic diversity particularly suitable for haploid sections of the genome; and a measure of allele sharing, defined as the number of haplotypes of the Anatolian sample also present in each Tuscan sample, scaled by the total number of haplotypes in the latter. In this way, we divided the number of shared haplotypes, respectively, by 21, 14, 72, and 57, namely, the number of haplotypes in the Etruscan, Medieval, Casentino, and Volterra's samples (Table 1b).

Tested demographic scenarios and priors

We designed two models, both assuming a common origin of populations of the Eastern and Northern Mediterranean shores during the Paleolithic dispersal of anatomically modern humans from Africa (see, e.g., Otte, 2000), but differing in the timing of a migration event. We did that by simulating an ancestral population split 1,000 generations ago, roughly equivalent to 25,000 years ago if one assumes, according to Fenner (2005), an average generation time of 25 years. In time, the first resulting lineage gave rise to the Etruscans (100 generations or 2,500 years ago) who, in turn, are the ancestors of the medieval Tuscans (36 generations or 900 years ago) and of current inhabitants of Casentino and Volterra (placed 0 generations ago); the second lineage gave rise to the Western Anatolian population. Both lineages grew exponentially in size after the split. Under Model A, migration from the East, followed by admixture, took place in a relatively remote past [between 6,000 and 10,000 years ago, based on estimates in Ghirotto et al., (2013)], whereas under Model B this event happened just before (and was crucial for) the onset of the Etruscan culture. For each model, we tested various admixture rates.

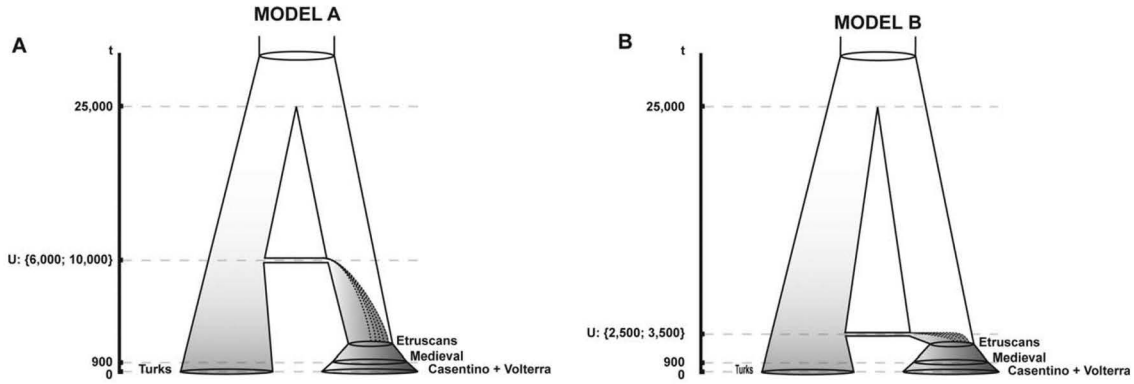
The models are characterized by demographic and evolutionary parameters whose values are independently drawn in each simulation experiment from uniform and wide prior distributions. The ancestral population sizes ranged from 5 to 6,000 individuals and the effective population sizes for modern Tuscans and Anatolians were independently sampled from a prior distribution ranging between 100 and 400,000. The prior for mutation rate was between 0.0003 and 0.0075, in agreement with recent papers based on ABC methods (Sanchez-Quinto et al., 2012; Ghirotto et al., 2013). A schematic outline of the models is in Figure 2 and a complete description of the prior information considered is in Table 2.

Approximate Bayesian computations: Model choice

The two models were compared, and their parameters were estimated, under an ABC framework (Beaumont et al., 2002). The ABC methods combine the analysis of abundant data and realistic models. They allow the probabilistic comparison of different models of evolution accounting for the observed variation, the simultaneous estimation of demographic and evolutionary parameters, and the quantitative evaluation of the results' credibility (Beaumont et al., 2010). ABC method is intuitively very easy: in principle, to test hypotheses on the genealogical relationships between samples, millions of genealogies are generated under different models and assuming different parameter values. The simulations that produce genetic variation patterns close to the observed data are retained and analyzed in detail. Indeed, parameter values and model features in the retained simulations are of course interesting because they are able to generate datasets with some properties found in the observed data. This approach could meet with difficulties because of the large number of parameters needed to fully describe the genealogy underlying the observed data, but the flexibility of ABC makes it possible to evaluate the likelihood also for complex demographic models (Marjoram and Tavarè, 2006). Indeed, the ABC approach allows one to approximate the likelihoods by comparing summary statistics extracted from the data, rather than the DNA sequences themselves, thus reducing the amount of information to account.

The various steps of the ABC procedure and their rationale are described in detail in Bertorelle et al. (2010), and summarized below:

1. For both models, we ran a large number of coalescent-based simulations using the program BayeSSC (Anderson et al., 2005; see http://iod.ucsd.edu/simplex/ssc/BayeSSc.htm); in particular, we



	Admixture Rate				
	50	60	70	80	90
AR	0.127	0.168	0.176	0.240	**0.289**
LR	0.115	0.158	0.181	0.216	**0.330**

	Admixture Rate				
	50	60	70	80	90
AR	**0.292**	0.204	0.191	0.155	0.158
LR	**0.334**	0.272	0.188	0.118	0.088

C

	Model A_{90}	Model B_{50}
AR	**0.665**	0.335
LR	**0.698**	0.302

Fig. 2. (**A**) and (**B**) Schematic presentation of the two models tested and ABC results among the admixture rates tested for each model. (**C**) ABC results of the comparison between submodels A_{90} and B_{50}.

generated 2,000,000 simulated datasets for each model (two) and each admixture rate (nine), for a total of 36,000,000 simulations. The values of the other parameters (i.e., population sizes, timing of the demographic events, mutation rates) defining the demographic processes described by the model were drawn from the prior distributions for each simulated dataset.

2. Summary statistics were estimated from both the observed data and from each simulated dataset. Then, after normalization of all statistics, for each simulated dataset, a Euclidean distance between the observed and simulated summary statistics was calculated.
3. Models were then compared for the goodness of their fit to the data. For this procedure, we followed two methods, both based on the *Calmod* function written by M. A. Beaumont (available at PopABC website: https://code.google.com/p/popabc/source/browse/trunk/scripts/calmod.r) for the *R* statistical package. Under the first method, the AR, or acceptance–rejection procedure (Pritchard et al., 1999), the posterior probability of a model is obtained by simply counting the proportion of the simulations in which either model generates statistics arbitrarily close to the observed statistics. This method is considered reliable only when applied to a few simulations showing an excellent fit with the observed data (Beaumont, 2008); therefore, we retained the 750 simulations resulting in the shortest Euclidean distances. The second criterion is a weighted multinomial logistic regression (LR) between the summary statistics and a categorical variable indicating either demographic model (Beaumont, 2008). For this calculation, we retained the 150,000 simulation experiments associated with the shortest Euclidean distances.

Parameter estimation

For the best model, we retained the 5,000 simulations generated under that model showing the shortest associated Euclidean distances, from a total of 5,000,000 simulations. Then, parameters were estimated by a locally weighted multivariate regression (Beaumont et al., 2002), after a *logtan* transformation to prevent the estimate from exceeding the bounds of the prior distribution (Hamilton et al., 2005).

Quality of the estimation: Type I error and posterior predictive test

To check whether the power of the LR and AR procedures is sufficient to actually identify the best model based on the available data, we followed an approach suggested by Fagundes et al. (2007) and Cornuet et al. (2008). First, we simulated 1,000 datasets from the prior distribution under the submodel emerging as the most likely, and we analyzed them as if they were observed datasets in an ABC analysis. We then assigned each of the 1,000 simulated datasets to the model showing the highest posterior probability. Finally, we calculated Type I error as the number of experiments in which the simulated model was not recognized by the model selection procedures.

TABLE 2. Priors, estimates, and R^2 of the parameters estimated under Submodel A_{90}. LowB and UppB are, respectively, the lower and upper bound of the 95% credible interval of the posterior probability distributions

	Priors	Median	Mode	95% HPD LowB	95% HPD UppB	R^2
Time MRCA	[a]	44,424	40,712	9,864	183,807	0.51
Mutation rate	(0.0003–0.0075)	0.0016	0.0016	0.0007	0.0032	0.77
Time admixture	(6,000–10,000)	8,396	10,000	6,007	10,000	0.02
Ne ancestral Tuscans	(5–6,000)	140	36	5	4,729	0.16
Ne ancestral Turks	(5–6,000)	676	229	5	5,814	0.31
Ne Tuscans	(100–400,000)	180,421	128,721	18,257	400,000	0.53
Ne Turks	(100–400,000)	302,391	400,000	6,259	400,000	0.23

HPD, highest posterior density; MRCA, most common recent ancestor.
[a] The time to the most recent common ancestor, Time MRCA, was estimated from the simulated data and not extracted from a prior distribution.

The above-described procedures are suitable to identify the model better reproducing the observed statistics, but do not test whether either model is realistic at all. To that end, we eventually evaluated by a posterior predictive test whether the model we chose has the ability to reproduce the observed data (Gelman et al., 2004). Therefore, we simulated 1,000 datasets according to the model with the highest probability using the estimated posterior parameter distributions. Then, we calculated 20 summary statistics that had not been considered during the previous inferential step, namely, nucleotide diversity and Tajima's *D* within each sample, and five Hudson's F_{ST} and five allele-sharing measures that describe the genetic distance between the Tuscans' samples. We compared these values with the same observed statistics and estimated a posterior predictive *P*-value for each summary statistic. Finally, these probabilities were combined into a global *P*-value, following a procedure described in Ghirotto et al. (2010).

RESULTS

Table 1 shows the statistics summarizing genetic variation in the five samples considered. The posterior probability of the alternative models being compared essentially measures the model's ability to generate data closely resembling the observed data.

Under the ABC framework we actually started by comparing two sets of models. We assumed either (Model A) that the genetic resemblance between Central Italy and Anatolia is because of a relatively ancient gene flow between these geographical regions, or (Model B) that migration from Anatolia brought into Central Italy the immediate ancestors of the Etruscans. The first test we ran was a comparison of the probabilities of various admixture rates (0.50, 0.60, 0.70, 0.80, and 0.90) within each demographic model (as a consequence, each Model was divided into five submodels, namely, A_{50}, A_{60}, A_{70}, A_{80}, A_{90} and B_{50}, B_{60}, B_{70}, B_{80},B_{90}).

We found that the highest posterior probabilities corresponded to an admixture rate of 0.90 for Model A (supported by 29% of the experiments under the LR approach and 33% under the AR approach; Fig. 2A) and 0.50 for Model B (supported by 29% of the experiments under the LR approach and 33% under the AR approach; Fig. 2B). We then proceeded to compare the models keeping constant the admixture rates thus estimated (submodel A_{90} vs. submodel B_{50}), so as to use for both models the most likely admixture level.

Submodel A_{90} proved about twice as likely as the alternative model, regardless of the criterion used for model selection (posterior probabilities were 66% under the LR approach and 70% under the AR approach; Fig. 2C). This result also held when different numbers of simulations were considered to compare models. In addition, the Model A results provided the most probable scenario even when compared with Model B, considering lower admixture rate (data not shown).

Once submodel A_{90} proved able to generate statistics in better agreement with the observed data than the alternative submodel B_{50}, we calculated its parameters' posterior probabilities, here reported in Table 2 and Figure 3, along with the priors. Narrow posterior distributions of the estimates mean that independent simulation experiments suggest similar values, and hence that these estimates are reliable. That seems the case for the ancient populations' sizes, the time to the Most Recent Common Ancestor (MRCA), and the mutation rate; the median for this statistic is 0.17 mutational events per million years per nucleotide, close to the values estimated in previous comparable studies (Hill et al., 2007; Ghirotto et al., 2010). The median time of admixture, 8,396 years ago, is probably an underestimate because the modal value corresponds to the upper limit of the priors we imposed. In other words, had we chosen a broader distribution of priors, we would have likely inferred an older gene flow event. Our purpose, however, was not to estimate a specific date for an event that may well have occurred across many years, or even centuries, but to see if that event has any chance to have closely preceded the appearance of Etruscan artifacts in the archeological record. The answer is that the probability of such a recent event is less than one-third, which means that the alternative is more than twice as likely. The estimates for the archaic population size of Tuscans and Turks are low, not an unexpected finding for the Mediterranean basin in Paleolithic times. On the contrary, the sizes of both modern populations show broad distributions of posterior probabilities. Such a finding is common in studies comparing populations across time (Fagundes et al., 2007; Belle et al., 2009; Laval et al., 2010), and probably reflects the effect of immigration, resulting in incorporation of novel mtDNA variants from external sources that are not easily incorporated into the models. This input of external DNAs increases the internal diversity of populations, and hence the (correlated) estimate of population size.

To be reliable, these results must be supported by evidence that, at the sample sizes we considered, our methods for model selection (AR and LR) were powerful enough to identify the correct model. To answer this question we calculated, for Model A_{90} and Model B_{50},

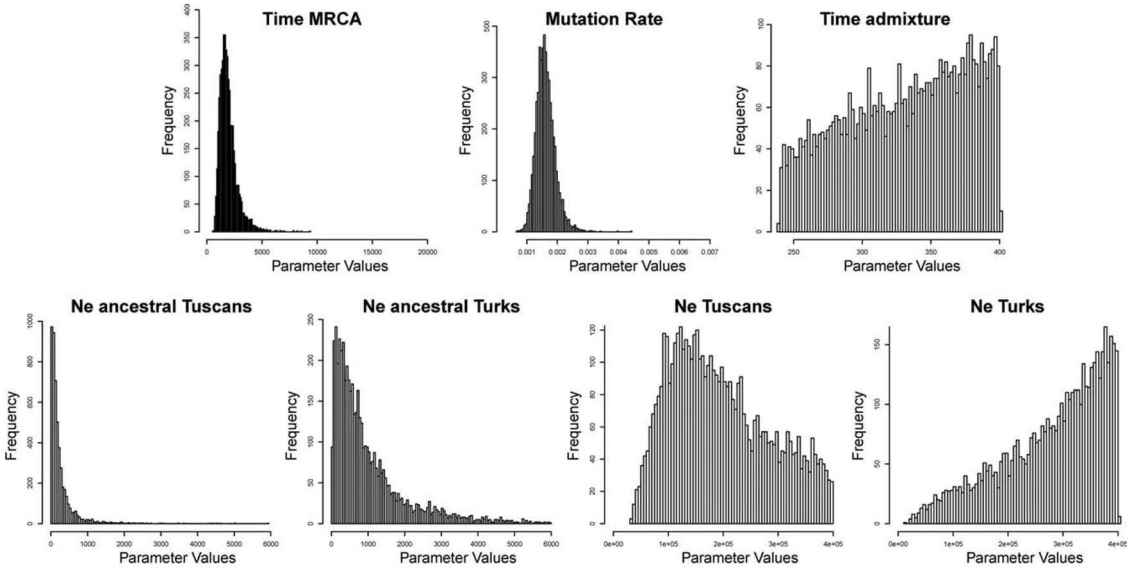


Fig. 3. Posterior distribution of parameters under submodel A_{90}. On the *x*-axis are the parameter values and the width of the axis expresses the range of the (uniform) prior distribution of each parameter, on the *y*-axis their frequencies in the 5,000 best-fitting simulation experiments (out of 5,000,000 performed).

TABLE 3. Type I errors for the two best models emerging from the ABC analysis

	Model A_{90}	Model B_{50}	Type I error
AR			
Model A_{90}	**0.81**	0.19	0.19
Model B_{50}	0.28	**0.72**	0.28
LR			
Model A_{90}	**0.85**	0.15	0.15
Model B_{50}	0.27	**0.73**	0.27

AR, acceptance–rejection criterion; LR, logistic regression criterion. The numbers in bold indicate the percentage of experiments in which the simulated model was successfully recognized.

the Type I error, generating by simulation 1,000 pseudo-observed datasets according to each model, with samples having the same size and age as the observed samples. Analyzing by ABC method these datasets generated under submodels A_{90} and B_{50} as if they were observed datasets, we found that both were in general correctly identified with a probability of recovery ranging from 72% to 84%, so that Type I error was, respectively, 28–16%. As could be expected, the statistical power in this comparison is not very high, because these models are rather similar, differing just for the timing and extent of an admixture event (Table 3).

A related, but different, question would be whether submodel A_{90} can indeed generate patterns of variation compatible with variation in the observed data. The *P*-values calculated by the posterior predictive test led to a global *P*-value for the whole model of 0.48. This probability means that the statistics generated by the model we chose as the best one broadly overlap with, and do not significantly depart from, those in the observed data.

DISCUSSION

The genetic patterns observed at the mtDNA level in the past and present Tuscany have a higher probability of resulting from an ancient migration process from Anatolia than by a migration occurring just before, and associated with, the origins of the Etruscan culture.

This finding is supported by explicit tests of hypotheses against ancient and modern DNA data. It confirms in part previous results based on modern data only, showing that the main separation between the Anatolian and Tuscan mitochondrial pools most likely occurred at least 6,500 years ago, or earlier if the two populations kept exchanging migrants after separation (Ghirotto et al., 2013). Similarly, we did not incorporate into our models the possibility of genetic exchanges between Anatolia and Tuscany after the main admixture event; we see no reason to exclude that these exchanges might have occurred, but, had we considered them, the admixture event would have been placed in a more remote past. Therefore, no genetic evidence, either based on ancient or modern DNA variation, suggests an input of people emigrating from Anatolia into Tuscany as a likely causal factor in the origin of the Etruscan civilization.

The analysis of modern mtDNAs (Ghirotto et al., 2013) and the comparison of ancient and modern DNAs (this study) have another result in common. Despite being based on different methods, and on largely (even if not completely) independent datasets, both dated the contact between the ancestors of current Anatolians and Tuscans at a moment in which gene flow was extensively occurring in Europe, namely, the Neolithic period. Indeed, studies of mitochondrial (Simoni et al., 2000) and nuclear (Chikhi et al., 1998) DNAs in modern Europeans, and comparisons of mitochondrial haplogroups between modern and ancient populations (Bramanti et al., 2009; Fu et al., 2012; Sanchez-Quinto et al., 2012)

show that the Neolithic spread of farming technologies in Europe was accompanied by significant demographic changes. The actual impact of Neolithic processes on European genetic diversity is still debated (e.g., Soares et al., 2010; Arenas et al., 2013) but there is little doubt that a Westward gene flow from the Near East or Anatolia into Europe took place in the Neolithic period (Barbujani and Goldstein, 2004; Barbujani, 2012).

The difference in the probabilities of the two models compared, approximately twofold, is not large. However, one has to consider that (a) genetic differences between populations are minimal in Europe, with the main geographical gradients accounting for some 0.45% of the global diversity (Novembre et al., 2008); (b) because only mtDNA has been typed on a sufficiently large scale to allow for diachronic comparisons, only one locus, albeit a highly variable one, could be considered; (c) a large number of population movements is documented in the ethnohistorical record of Europe (Sokal et al., 1993; Sokal et al., 1996), each of them potentially confounding the genetic pattern left by a more remote event, such as the one we were analyzing; and (d) the two models we were comparing differed only as for the timing of a migration event, and so could not possibly be expected to differ by much in their consequences. In light of these factors, it would have been unrealistic to expect greater levels of statistical significance in this study, and it seems remarkable that we could demonstrate a substantial difference in the models' posterior probabilities. Availability of nuclear DNA sequences from ancient specimens will radically improve our inferential power, but at present only a handful of ancient individuals have been studied at the nuclear level (Sanchez-Quinto et al., 2012).

We are fully aware that the processes that occurred in the Mediterranean area over the long time span considered in this study are far more complex than the one actual model. That is also the reason why we chose to focus mainly on admixture rates equal to or greater than 50%. Although lower values were also taken into consideration, comparing a large number of hypotheses differing for only one parameter, the admixture rate in this case is notoriously complicated (e.g., Konečný et al., 2013). At any rate, smaller Anatolian contributions to the Etruscan gene pool would hardly have been compatible with the notion that the Etruscans are an immigrant Eastern population. In addition, it would have been extremely difficult to tell apart smaller admixture rates from the effects of some of the migration processes that occurred in later prehistoric and historical times. Therefore, we do not give any special importance to the point estimate we obtained for the date of a likely contact between the ancestors of Anatolians and Tuscans, which was admittedly calculated in a rather rough manner. What matters, and has historical relevance, is that this date is clearly earlier than 2,500 or 3,000 years ago, and that a similar date was also inferred from the analysis of only modern DNAs (Ghirotto et al., 2013). More to the point, in this study the posterior probability of Model B increased for older dates of the migration episode, thus suggesting that, if anything, our date might underestimate, certainly not overestimate, the age of the contact.

Therefore, this study shows that inference based on DNA diversity in modern populations is well complemented by ancient DNA studies, and that considering both kinds of data is important if one is to identify the genealogical links of populations. Future studies also considering nuclear DNA diversity in ancient samples will add further details to the general picture, and may possibly lead us to reconsider some of the conclusions of this study. However, the analysis of ancient nuclear genes is still in its infancy and it will take time to accumulate sufficient sample sizes to explicitly test models on the genealogical links between the past and current populations. For the time being, it seems safe to say that, based on the best available data as analyzed by the most advanced biostatistical methods, ancient and modern DNA evidence converges in not suggesting a biological origin of the Etruscans outside Italy. The existing similarities between the Anatolian and Tuscan gene pools (Achilli et al., 2007) can simply be accounted for by the effects of older, or much older, prehistoric contacts, unrelated to the later development of the Etruscan culture.

LITERATURE CITED

Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, Battaglia V, Accetturo M, Kutuev I, Khusnutdinova E, Pennarun E, Cerutti N, Di Gaetano C, Crobu F, Palli D, Matullo G, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Semino O, Villems R, Bandelt HJ, Piazza A, Torroni A. 2007. Mitochondrial DNA variation of modern Tuscans supports the Near Eastern origin of Etruscans. Am J Hum Genet 80:759–768.

Anderson CN, Ramakrishnan U, Chan YL, Hadly EA. 2005. Serial SimCoal: a population genetics model for data from multiple populations and points in time. Bioinformatics 21: 1733–1734.

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

Arenas M, Francois O, Currat M, Ray N, Excoffier L. 2013. Influence of admixture and paleolithic range contractions on current European diversity gradients. Mol Biol Evol 30:57–61.

Bandelt HJ, Kivisild T. 2006. Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. Ann Hum Genet 70:314–326.

Barbujani G. 2012. Human genetics: message from the Mesolithic. Curr Biol 22:R631–R633.

Barbujani G, Goldstein DE. 2004. Africans and Asians abroad: genetic diversity in Europe. Annu Rev Genomics Hum Genet 5:119–150.

Barker G, Rasmussen T. 1998. The etruscans. Oxford: Blackwell.

Beaumont M. 2008. Joint determination of topology, divergence time and immigration in population trees. Simulations, genetics and human prehistory. Cambridge: McDonald Institute for Archaeological Research. p 135–154.

Beaumont M. 2010. Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst 41:379–406.

Beaumont M, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Belle EM, Ramakrishnan U, Mountain JL, Barbujani G. 2006. Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. Proc Natl Acad Sci USA 103:8012–8017.

Belle EM, Benazzo A, Ghirotto S, Colonna V, Barbujani G. 2009. Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations. Heredity 102:218–225.

Bendall KE, Sykes BC. 1995. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57:248–256.

Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19:2609–2625.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ, Lueth F, Terberger T, Hiller J, Matsumura S, Forster P, Burger J. 2009. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326:137–140.

Brisighelli F, Capelli C, Alvarez-Iglesias V, Onofri V, Paoli G, Tofanelli S, Carracedo A, Pascali VL, Salas A. 2009. The Etruscan timeline: a recent Anatolian connection. Eur J Hum Genet 17:693–696.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. 1998. Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. Proc Natl Acad Sci USA 95:9053–9058.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24:2713–2719.

Di Benedetto G, Erguven A, Stenico M, Castri L, Bertorelle G, Togan I, Barbujani G. 2001. DNA diversity and population admixture in Anatolia. Am J Phys Anthropol 115:144–156.

Excoffier L, Lischer HE. 2010. Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567.

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA 104:17614–17619.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415–423.

Fu Q, Rudan P, Paabo S, Krause J. 2012. Complete mitochondrial genomes reveal Neolithic expansion into Europe. PLoS One 7:e32473.

Gelman A, Carlin J, Stern H, Rubin D. 2004. Bayesian data analysis. Boca Raton, FL: CRC Press.

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 27:875–886.

Ghirotto S, Tassi F, Fumagalli E, Colonna V, Sandionigi A, Lari M, Vai S, Petiti E, Corti G, Rizzi E, De Bellis G, Caramelli D, Barbujani G. 2013. Origins and evolution of the Etruscans' mtDNA. PLoS One 8:e55519.

Guimaraes S, Ghirotto S, Benazzo A, Milani L, Lari M, Pilli E, Pecchioli E, Mallegni F, Lippi B, Bertoldi F, Gelichi S, Casoli A, Belle EM, Caramelli D, Barbujani G. 2009. Genealogical discontinuities among Etruscan, Medieval, and contemporary Tuscans. Mol Biol Evol 26:2157–2166.

Hamilton G, Stoneking M, Excoffier L. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc Natl Acad Sci USA 102:7476–7480.

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, Richards M. 2007. A mitochondrial stratigraphy for island Southeast Asia. Am J Hum Genet 80:29–43.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589.

Konečný A, Estoup A, Duplantier JM, Bryja J, Bâ K, Galan M, Tatard C, Cosson JF. 2013. Invasion genetics of the introduced black rat (Rattus rattus) in Senegal, West Africa. Mol Ecol 22:286–300.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010 Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS One 5: e10284.

Marjoram P, Tavarè S. 2006. Modern computational approaches for analysing molecular genetic variation data. Nature 7:759–770.

Mateiu LM, Rannala BH. 2008. Bayesian inference of errors in ancient DNA caused by postmortem degradation. Mol Biol Evol 25:1503–1511.

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

Otte M. 2000. The history of European populations as seen by archaeology. In: Renfrew C, Boyle K, editors. Archaeogenetics: DNA and the population prehistory of Europe. Cambridge: McDonald Institute for Archaeological Research. p 41–44.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16:1791–1798.

Sanchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez AM, Marcos ME, Encinas JM, Bertranpetit J, Orlando L, Gilbert MT, Lalueza-Fox C. 2012. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. Curr Biol 22:1494–1499.

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G. 2000. Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262–278.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB. 2010. The archaeogenetics of Europe. Curr Biol 20:R174–R183.

Sokal RR, Jacquez GM, Oden NL, DiGiovanni D, Falsetti AB, McGee E, Thomson BA. 1993. Genetic relationships of European populations reflect their ethnohistorical affinities. Am J Phys Anthropol 91:55–70.

Sokal RR, Oden NL, Walker J, Di Giovanni D, Thomson BA. 1996. Historical population movements in Europe influence genetic relationships in modern samples. Hum Biol 68:873–898.

Vernesi C, Caramelli D, Dupanloup I, Bertorelle G, Lari M, Cappellini E, Moggi-Cecchi J, Chiarelli B, Castri L, Casoli A, Mallegni F, Lalueza-Fox C, Barbujani G. 2004. The Etruscans: a population-genetic study. Am J Hum Genet 74:694–704.

RINGRAZIAMENTI

Iniziare a scrivere questi ringraziamenti mi è tanto difficile. Non perché non mi piaccia ringraziare, anzi, ma perché con essi si chiude un altro ciclo della mia vita personale e accademica. I sentimenti che mi attraversano in questo momento sono tanti e in contraddizione tra loro, è un misto di allegria e tristezza, entusiasmo e stanchezza, soddisfazione e frustrazione.

Questi tre anni sono stati i più duri della mia vita in cui ho dovuto affrontare tante difficoltà, tanti impegni e tanti cambiamenti. Penso che l'attività di ricerca, e soprattutto la mia passione verso questa, siano state l'appiglio a cui mi sono aggrappata saldamente nei momenti più difficili. L'università e in particolare la "stanza 25" sono state per me un rifugio. Non potrei quindi consegnare questa tesi senza ringraziare tutte le persone che in questi anni hanno trascorso un po' di tempo con me a fare ricerca o anche solo per due chiacchiere.

Primo fra tutti voglio ringraziare Guido Barbujani che, oltre avermi dato la possibilità di entrare nel mondo della ricerca e avermi guidato in questo percorso di crescita personale e lavorativa, si è dimostrato molto comprensivo di fronte alle avversità che ho dovuto affrontare in questi anni.

Non troverò mai tutte le parole necessarie per ringraziare a sufficienza Silvietta, da cinque anni oltre a collaborare ai vari progetti di ricerca, è l'amica a cui mi affido ogni giorno. Grazie alla componente Y della stanza: Andrea e Alex, per il loro supporto scientifico ma soprattutto per l'allegria che mi hanno regalato in questi anni.

Grazie a Silvia e Giorgio che ho sempre sentito vicini e disponibili a darmi un consiglio.

Grazie a tutto il gruppo di genetica delle popolazioni di Ferrara del presente e del passato, grazie ai membri della stanza 10: Leti, Roby, Serena e Andrea, grazie ai dottorandi che mi hanno preceduto, ai tesisti che si sono succeduti e a chi oggi è da qualche parte nel mondo e quindi grazie a Angy, Sibelle, Marco, Solange, Massimo, Pierpaolo, Sean, Luca Cornetti, Michela, Dolfin, Francesca, Lisa, Emanuele, Esteban, Teresa, Luca, Denise e come non dimenticare Marcella e Aurelio.

Un grazie particolare alla piccola Susi che, fin dai primi giorni dal suo concepimento, ha dovuto sopportare la mia voce durante le lunghe e appassionanti discussioni con la sua mamma, e che mi regala sempre dei super sorrisi quando mi vede.

Dopo aver ringraziato la mia seconda famiglia, è giunto il momento di ringraziare chi nelle poche ore fuori dall'università deve sopportare i miei malumori: grazie al mio papà, alla mia mamma e a mia sorella. E infine grazie ad Ale che è arrivato così dal nulla in questa bufera e ha saputo fin dall'inizio adattarsi ai miei ritmi e ai miei colpi di testa.