Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
BIOLOGIA EVOLUZIONISTICA ED AMBIENTALE

CICLO XXIV

COORDINATORE Prof. Guido Barbujani

Genealogical inferences based on comparison of modern and ancient DNA

Settore Scientifico Disciplinare BIO/18

Dottorando
Dott. Ghirotto Silvia

_____

(firma)

Tutore
Prof. Barbujani Guido

_____

(firma)

Anni 2009/2011

Il tuo indirizzo e-mail
ghrslv@unife.it
Oggetto:
"Dichiarazione di conformità della tesi di Dottorato"
Io sottoscritto Dott. (Cognome e Nome)
Ghirotto Silvia
Nato a:
Rovigo
Provincia:
Rovigo
Il giorno:
12/12/1984
Avendo frequentato il Dottorato di Ricerca in:
Biologia Evoluzionistica ed Ambientale
Ciclo di Dottorato
24
Titolo della tesi (in lingua italiana):
Inferenze genealogiche basate sul confronto di DNA antico e moderno
Titolo della tesi (in lingua inglese):
Genealogical inferences based on comparison of modern and ancient DNA
Tutore: Prof. (Cognome e Nome)
Barbujani Guido
Settore Scientifico Disciplinare (S.S.D.)
BIO/18
Parole chiave della tesi (max 10):
genetica di popolazioni(population genetics), DNA antico (ancient DNA), simulazioni di coalescenza (coalescent simulations), modelli demografici (demographic models), metodi bayesiani approssimati (Approximate Bayesian Computation)
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 produrrò 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; (3) che il contenuto e l'organizzazione della tesi è opera originale da me realizzata e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura civile, amministrativa o penale e sarà da me tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi; (4) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati, non è oggetto di eventuali registrazioni di tipo brevettale o di tutela. PER ACCETTAZIONE DI QUANTO SOPRA RIPORTATO

Firma del dottorando

Ferrara, li _28/02/2012__ (data) Firma del Dottorando

_____

Firma del Tutore

Visto: Il Tutore Si approva Firma del Tutore

_____

A chi non c'è più,
ma ha dedicato la sua vita alla ricerca.

E ad Andrea,
il miglior "compagno di viaggio"
che avrei mai potuto desiderare.

Genealogical inferences based on comparison of modern and ancient DNA

ABSTRACT

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-scales of historical evidence. Demographic and evolutionary dynamics influence the distribution of the observed genetic diversity, and so one can retrospectively reconstruct episodes in population history on the basis of genetic diversity data. One way to do this is to make extensive use of simulations, considering evolution as a stochastic process in which the genetic data are modeled as random variables. The simulation of genetic data under various scenarios allows one to explore how demographic and evolutionary parameters can affect genetic variation, also making it possible to approximately estimate the historical parameters that produced the observed data. To this aim, many statistical approaches have been developed, but, when models are complex or datasets are large, they often become computationally expensive, or analytically intractable. Approximate Bayesian Computation (ABC) methods overcome these problems allowing, for the first time, to analyze large datasets and to interpret them in the light of realistic (i.e. complex) models, thus enabling the probabilistic comparison among different models of evolution, the simultaneous estimation of demographic and evolutionary parameters, and the quantitative evaluation of the results credibility. In this context, we analyzed datasets of modern and ancient genetic variation in order to understand the demographic histories of these populations, to highlight traces of past genetic variation in modern populations, and to evaluate whether, and to what extent, ancient and modern populations that have lived in the same place in different period of times can be considered genealogically related. We tried to address three anthropological questions, namely the interaction of anatomically modern humans with archaic forms (i.e. Neandertals in Europe), evidence for genealogical continuity in Sardinia since the Bronze-age, and the origins and evolution of the Etruscan population. Within the ABC framework, in each of the three studies, we explicitly compared several models, differing for the demographic processes and the genealogical relationship among population, to identify the model best accounting for the observed variation, and to estimate its demographic and evolutionary parameters. This way, it has been possible to shed light on past population history and to address questions about the nature and the extent of genealogical links between modern and ancient populations, clarifying aspects of human history that have long been controversial in population genetics and evolutionary biology.

Inferenze genealogiche basate sul confronto di DNA antico e moderno

ABSTRACT

Lo studio della variabilità genetica delle popolazioni può aiutarci a comprendere aspetti della storia demografica umana ai quali non possiamo risalire tramite evidenze storiche, o perché si tratta di eventi troppo antichi, o perché non esistono documentazioni attendibili. Le dinamiche evolutive e demografiche delle popolazioni influenzano la distribuzione della diversità genetica osservata; è quindi potenzialmente possibile, partendo dall'analisi di questa variabilità, ricostruire a posteriori quali siano stati i processi demografici ed evolutivi che possono averla generata. Un approccio ampiamente utilizzato in questo contesto riguarda l'uso di simulazioni: considerando l'evoluzione come un processo stocastico ed utilizzando un modello probabilistico adeguato, vengono simulati dati di variabilità genetica secondo diversi modelli di evoluzione delle popolazioni in esame, permettendo di testare in modo esplicito come diversi parametri evolutivi e demografici possano influenzare i livelli di variabilità genetica interna e tra le popolazioni. Confrontando la variabilità genetica che si ottiene dalle simulazioni con la variabilità genetica osservata, è possibile scegliere fra tanti quale modello evolutivo possa aver generato i livelli di variabilità osservati, e quali siano i cambiamenti demografici che hanno influenzato in misura maggiore tale variabilità. Negli ultimi anni sono stati sviluppati diversi approcci statistici allo scopo di stimare, tramite le modalità appena descritte, i parametri storici delle popolazioni. Purtroppo però, quando i dati da analizzare sono molti, o i modelli da simulare sono complessi e ricchi di parametri, il costo computazionale diventa molto elevato, tale da rendere l'analisi impraticabile. Recentemente, lo sviluppo dei metodi bayesiani approssimati (ABC) ha permesso di superare questo limite, rendendo possibile l'analisi di dataset sempre più ricchi, in linea con il recente sviluppo delle tecniche di sequenziamento su larga scala (Next Generation Sequencing), e di interpretarli alla luce di modelli sempre più complessi, e quindi realistici. Questa metodologia ha reso possibile molti confronti probabilistici tra diversi modelli di evoluzione, consentendo di stimare i valori dei parametri che meglio descrivono i dati. Abbiamo applicato questa metodologia a tre dataset di popolazioni antiche e moderne, allo scopo di determinare quale possa essere stata la loro storia demografica ed evolutiva, e al fine di evidenziare eventuali relazioni genealogiche tra popolazioni che hanno abitato le stesse località geografiche in diversi periodi temporali. Il primo studio riguarda la storia evolutiva dell'uomo moderno e la sua interazione con forme umane arcaiche preesistenti (nello specifico il Neandertal in Europa), il secondo è uno studio delle relazioni genealogiche fra popolazioni sarde antiche (le popolazioni nuragiche dell'età del Bronzo)

e moderne, e il terzo riguarda la storia della popolazione etrusca, le sue origini e le sue relazioni genetiche con i toscani moderni. Per ognuno di questi studi è stato scelto un modello genealogico più verosimile e si sono stimati i parametri demografici che si adattano meglio alla variabilità osservata. Questo ha permesso di far luce su aspetti della nostra specie prima sconosciuti, sia in termini evolutivi, sia demografici. Inoltre, è stato possibile testare per la prima volta in modo esplicito la continuità genealogica fra popolazioni antiche e moderne provenienti dalla stessa area geografica, evidenziando che anche popolazioni molto vicine geograficamente, possono avere una storia genealogica molto diversa.

Table of Contents

1. Introduction

The goal of population genetics is to understand the forces that produce and maintain genetic variation within species. These forces include mutation, recombination, gene flow (or its absence), natural selection, and the random transmission of genetic material from parents to offspring.

Even since its onset, theoretical population genetics has had strong statistical bases (Provine 1971). From a methodological perspective, the focus of this field was to develop models describing the behavior of random processes to depict the evolution of allele frequencies over time. A model can be viewed as a relatively simple mathematical formulation of the biological process producing the observed data which can incorporate parameters of interest in population genetics. Traditionally, these models (which are stochastic, since there is no predetermined outcome) have allowed researchers to predict how patterns of genetic variation would be affected by forces such as genetic drift, mutation, migration and selection (see Introduction, section 1.1).

One of the most useful stochastic models in population genetics is the *coalescent* (Kingman 1982; Wakeley 2009). In brief, the coalescent provides a theoretical description of the ancestral relationships existing in a sample of DNA sequences taken from a population, depending on the specific combination of demographic and evolutionary features of the population. A detailed description of this model is reported in Introduction, section 1.2. In simple cases, the intensity of selection, or the combination of population size (determining the impact of drift) and migration rates can be approximately inferred from the data. However, as a rule, this exercise turns out to be exceedingly complicated and to require untestable assumptions. The modern approach to understand the evolutionary and demographic forces behind the patterns of population genetic variation is then to make intensive use of simulation methods. Simulating genetic data according to the coalescent theory allows one to explore how the data can vary changing population genetics parameters such as the effective population size or the mutation rate. A limitation of the coalescent is that it can become highly computationally intensive; however, with the rapid

growth in computational power, the evolutionary models that can be simulated have grown more complex, and have therefore become more realistic.

There are two different, but related, use of the word “simulation” in this context. The first indicates the simulation of the data under a specific demographic model, thus producing datasets that are representative of the evolutionary process and that differ from each other just by chance. For example, this approach might be used to examine the degree of variability that may be found in the data that have been produced under a proposed model of evolution (Slatkin & Hudson 1991), or to test if a specific model of evolution (with a specific parameters combination) can faithfully reproduce the observed variation (Belle et al. 2009; Guimaraes et al. 2009). The second sense in which we use simulations refers to the use of simulation-based methods of statistical inference exploiting the coalescent to estimate parameters, from a particular kind of process that is described by the model. Here we start with an observed data set and we use simulations of data under a variety of parameter values, in an attempt to infer the probability of the data under a particular model, as a function of its parameters. The aim here is to find the combination of parameters value that maximize this probability, i.e. the combination of parameters able to generate datasets close to those observed. To this aim, Bayesian inference as applied to population genetics, represents a powerful tool for addressing a number of longstanding questions in evolutionary biology (see Introduction, section 1.3). Combining the intuition that is provided by complex stochastic models with the use of simulations methods for inference it is possible to address and clarify important aspects of past population history.

1.1 Processes shaping genetic variation

As just said, one of the aims of population genetics is to understand the forces that shape patterns of genetic variation. This variation has been shaped by various demographic and evolutionary factors, and hence contains information on past population changes and on the history of human adaptation to changing environment. Thus, studying how genetic variation is distributed within and between populations around the world can provide insight into (i) the place and the time of origin of our species, (ii) the degree of admixture with archaic *Homo* forms, (iii) migration of modern humans around the world, and about (iv) genealogical links between modern and ancient populations after these migratory events. Furthermore, this ability to infer past population dynamics has substantially improved with the development of methods for the typing of DNA from ancient specimens.

The genetic variation might be analyzed through two main classes of different approaches. The first one involves a **description** of the distribution of observed diversity, which allows the evaluation of the degree of variation within populations, the comparison of genetic diversity and its apportionment between populations. To this aim, relevant statistics should be calculated from the data, quantifying both the degree of internal variation (number of haplotypes, gene diversity, number of polymorphic sites), and the genetic distance between populations (Fst and allele sharing, see Methods, section 3.1). The second approach involves **testing of hypotheses** about how modern genetic diversity evolved, and this requires to develop explicit or implicit models of the evolutionary processes, allowing to make predictions about origins, movements and demography of populations, including their consequences at the DNA level (see Methods, section 3.2). Usually, studies of human genetic diversity are limited to modern populations, which severely limit our ability to investigate past processes. Prehistorical and historical processes, in this case, can only be inferred from modern diversity. However, for some years now, it has been possible to also include in the analysis samples coming from ancient specimens (ancient DNA, aDNA, see Introduction, section 1.4). The genetic information they yield is mainly from a single marker, the mitochondrial DNA (mtDNA), (Caramelli et al. 2008; Green et al. 2008), but with the development of the techniques of high throughput sequencing, it is now possible to obtain data on nuclear diversity as well (Green et al. 2006), and even sequencing entire ancient

genomes (Green et al. 2010; Reich et al. 2010). Considering the ancient genetic data allows one not only to increase the power in estimating the historical demographic processes, but also to test hypotheses about the genealogical links between modern and ancient populations living in the same place at different periods of time.

It is worth noting that there is not a single and simple way to analyze the data and to answer to complex questions of population genetics. A combination of several analytical approaches, starting from a description of the variation observed in the data, up to the use of inferential methods to estimate evolutionary and demographic parameters, might help to answer the question: "How did a particular pattern of genetic diversity arise?"

1.1.1 Hardy-Weinberg equilibrium

The first challenge of population genetics was to explain how allele frequencies in one generation could be used to calculate genotype proportions in the next generation of an infinitely large, randomly mating, population. If we consider a diploid organism, such as humans, with two allele A and a, with frequency p and q respectively, three different genotypes are possible: AA, Aa and aa. If we know the p and q values in an idealized population, we can predict the proportion of genotypes in the succeeding generation by combining gametes (containing single alleles) at random. This is known as the Hardy-Weinberg principle (Hardy 1908). The proportion of each genotype in the next generation is:

$AA = p^2$

$Aa = 2pq$

$aa = q^2$

If the genotype proportion in the succeeding generation are calculated in this manner, and any variation is found from the parental generation, the population is said to be at *Hardy-Weinberg equilibrium*. To be in Hardy-Weinberg equilibrium, the idealized population must have some additional properties other than **infinite population size**, such as no **mutation**, no **migration** and no **selection**; in other words, any factor that might change allele frequencies has to be absent. If the calculated genotype proportions are not in Hardy-

Weinberg equilibrium, we might conclude that evolution is occurring, and that one or more of the above factors are acting on the population shaping the observed variation.

1.1.2 Genetic Drift

No population is infinitely large, as assumed by the Hardy-Weinberg theorem, because each generation represents a finite sample of the previous one. This stochastic process of sampling from one generation to another determines a random variation in allele frequencies over time and is called **random genetic drift** (Wright 1931). Genetic drift may cause allelic variants to disappear completely or to be fixed (reaching frequency of 1), and therefore reduces the population genetic variation. In 1931, Wright demonstrated the extent of genetic drift in an idealized population (i.e. random mating, constant size, with nonoverlapping generations) introducing the concept of **effective population size** (*Ne*). The effective population size is the size of an idealized population that experiences the same amount of genetic drift of the population under study. It is not easy to relate this effective population size (*Ne*) to the census population size (*N*), but substantially the Ne is almost always smaller than the actual population size *N*. This concept is fundamental since it was demonstrated that the magnitude of the effects of genetic drift is correlated with the effective size of the population: the smaller the effective population size, the greater the drift effects.

The concept of effective population size allows one to calculate the probability and the rate of fixation for a new allele in a population, in the absence of mutation and selection. Fixation is a rare event, and this probability in the absence of selection is equal to the frequency of the new allele in the population, that is $^{1}/_{2N}$. From this equation it is clear that the smaller is the population, the greater chance a new mutant has of becoming fixed. Moreover, from the effective population size it is also possible to calculate the expected time (in generations) since the fixation of a new allele (i.e. equal to 4N generations). This equation demonstrates that a new allele in a smaller population will not only have a higher probability of becoming fixed, but it will also be fixed more rapidly than it would in a larger population.

The extant variation at neutral loci depends on past effective population sizes. In particular, the long-term effective population size has been shown to be approximately equal to the harmonic mean of the population sizes over time (Wright 1938; Crow & Kimura 1970), and this means that this measure is highly affected by phases in which the population size became smaller. In demographic processes involving a reduced ancestral population size, the amount of the present variation is largely determined by this smaller ancestral population size and the extent of genetic drift will be greater than expected based on current census figures. Two examples of processes reducing the effective population size are the **bottleneck** and the **founder effect**, largely documented in human populations. The first refers to the reduction in size of a single, previously larger, population, and the latter to the process of colonization and the genetic separation of a subset of the diversity present within the source population, both resulting in a loss of genetic diversity.

1.1.3 Mutation

Mutation is the sole process generating new alleles. It provides the material on which evolution can act by means of selection or other forces. In absence of these forces, an allele will decrease in frequency as new mutations arise and generate other alleles; by knowing the mutation rate for the whole gene, the initial allele frequency (p_0), and assuming no back mutation and multiple substitutions at the same site, is it possible to calculate the frequency of the same allele after *t* generations as:

$$p_t = p_0 \times e^{-\mu t}$$

This is known as mutation pressure. Mutation is a weak force (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), hence can have an appreciable impact upon genetic diversity only over long time periods.

As said above, when we consider a gene, or in general a DNA sequence, in which every mutation creates a new allele, we discount the possibility of back mutations (T->C; C->T), and recurrent mutations (same mutation at the same site in different individuals). This

model is known as the **infinite alleles model** (Kimura & Crow 1964). Another model typically used is the **infinite sites model** (Kimura 1969), which assumes that every mutation occurs at a different site in the DNA sequence and therefore, under this model, there is no need to consider multiple hits, i.e. multiple mutations at the same site. Considering that the total number of sites in each gene is so large and the mutation rate per site is so small, at first sight these models seem to be a reasonable approximation of the reality for the evolution of DNA sequences.

If we are interested in aspects of sequence evolution that require us to suppose that multiple changes might have occurred at the same site, we need more complex models of mutation. For example, these models are useful when long time scales are considered (i.e. calculating the distance between two DNA sequences separated long time ago), and not accounting for back mutation or multiple hits may result in underestimation of the real sequence divergence. In the simplest of these models, the Jukes and Cantor model (**JC69**; Jukes & Cantor 1969), all the substitutions occur at the same rate, meaning that every nucleotide in the sequence has the same probability of changing into any other nucleotide. Kimura (1980) proposed a model that accounts for transitions (A<->G; T<->C) occurring at higher rates than transversion (A,G<->T,C) (**K2P**), and Hasegawa, Kishino and Yano (1985) allow this model to account also for the differences in base frequency (**HKY**). The most complex model of nucleotide substitution is the general time reversible (**GTR**) model (Tavaré 1986) that considers six different substitution rates instead of two (i.e. transition and transversion rate). Moreover, models have been developed that can accommodate **rate variation among sites**, assuming that the mutation rate may vary along the sequence. When the rates vary, some sites (mutational hotspots) may accumulate many changes, while other sites (conserved sites) remain unchanged. One can accommodate this variation assuming that the rate of substitution for any site is a random variable drawn from a statistical distribution. The most commonly used distribution is the *gamma*, defined by the shape parameter α, that is inversely related to the extent of rate variation at sites: if $\alpha \rightarrow \infty$ the distribution degenerates into a model of a single rate for all sites; if $\alpha<1$ the distribution has a highly skewed L-shape, meaning that most sites have a very low rates of substitutions, and there are some substitution hotspots.

1.1.4 Migration

Migration is the movement of individuals from an occupied area to another, and differs from colonization since the latter regards a movement into a previously unoccupied territory. Gene flow is the outcome of the process of migration, when a migrant contributes to the next generation in the new location, and depends on the reproductive success of the migrants in the new area. Estimates of gene flow have, therefore, relied upon indirect methods linking measures of population subdivision to gene flow via a model for population structure. To describe migration processes, one has to envisage a general population subdivided in population units or demes. Alternatively, one can speak of several populations connected by gene flow into a large meta-population. From the practical standpoint, the two terminologies are equivalent; in what follows I shall use the latter.

The simplest model of gene flow is the **island model**, devised by Sewall Wright (1931), in which a meta-population is subdivided into islands of equal size *N*, exchanging genes at the same rate *m* per generation. The assumptions of this model include that all islands are equivalent, without substructure other than the division into islands; no selection is present; each population has reached an equilibrium between mutation and drift; the migrant are a random sample from the source island population; each population persists indefinitely. Under these assumptions it was demonstrated that the rate of migrants exchanged determines the level of population subdivision (as measured by Wright's Fst) by the equation:

$$Fst = \frac{1}{(1 + 4Nm)}$$

The island model does not take into account the fact that levels of migration are generally affected by the geographic distance between populations. A model considering some effect of geography is the **stepping stone** (Kimura & Weiss 1964). This model allows the exchange of genes only between adjacent discrete subpopulations. Similarly to the island model, the stepping stone assumes an equal rate of migration between populations.

A further step toward realistic modelization of migrational relationships is represented by the possibility to actually incorporate a measure of geographic distance

between potential mating partners. Migration can be modeled within a continuous population considering that mating choices are limited by distance and that these distances are typically less that the overall range of the population. This is the basis for the **isolation by distance model** (Wright 1943). Under this model, genetic similarity between neighborhoods is a function of the dispersal distance. These can be viewed either as difference between birthplaces of parent and offspring, or marital distance. Several mathematical functions have been used to relate the decline in frequency of the dispersal over geographical distance; after reaching equilibrium between genetic drift and gene flow, is it possible to predict the rate of decline of genetic similarity at increasing geographical distances.

A more realistic model of migration has been developed in 1991 by Slatkin and Voelm (Slatkin & Voelm 1991). They called this model **hierarchical island model**. The rationale behind this model is that the finite or infinite island model would not be appropriate if some of the sampled populations share some recent ancestry, if some sampled populations contribute to different migrant pools, or if there is a hierarchical population structure. In a hierarchical island model the meta-population is assumed to be made up of n neighborhoods, each of which contains d demes of effective size N. The model assume that a randomly chosen gamete after a migration event has a probability $1-m_1-m_2$ of being nonimmigrant, a probability of m_1 of being an immigrant from a different randomly chosen deme in the same neighborhood and a probability m_2 of being an immigrant from a randomly chosen deme in a different neighborhood; moreover, m_1 is assumed to be greater than m_2.

The migration models are mathematically tractable and can be generalized to many species. When the populations under study are human populations, we might have detailed information about the migratory processes, such as the migration rates, the marital distances and the migration distance. All this information can be incorporated in the migration model, which can then account for different migration rates and asymmetric migration between subpopulations.

1.1.5 Selection

Natural selection, as defined by Charles Darwin and elaborated by Ronald Fisher, is the consequence of differential ability of reproduction of genotypes through generations. Individuals exhibit differential capacities to survive and reproduce in different environments and evolution occurs by natural selection when these differences in reproductive success among organisms are correlated with their genetic differences. The individual's expected reproductive success is measured by her/his fitness, ω ($0 \leq \omega \leq 1$), and the relative fitness of an individual's genotype is obtained from a comparison of this genotype with all other genotype competing for the same resources. Usually this relative fitness is measured by a selection coefficient (s) representing the loss in fitness with respect to the fittest genotype in the population. Since the relative fitness is equal to 1- s, a selection coefficient of 0.1 represents a 10% decrease in fitness compared to the fittest genotype, which means a relative fitness of 0.9 (90%).

Natural selection can act in a population only if mutation has generated heritable polymorphism among individuals, i.e. only if any difference in fitness can be transmitted from one generation to another. That is why the genetic variance is used as a measure of the opportunity of selection in a population or species. For the purpose of this paragraph, mutation can be mainly classified into two categories: *neutral* (not having any effect on the fitness, usually located in non-coding regions), and *non-neutral*, having effect on the fitness, and which can be broadly categorized as *advantageous* (that is, adaptive) or *deleterious*. Variants that increase the fitness of an individual in its environment might increase in frequency as a result of *positive selection*, whereas moderately to severely deleterious gene variants tend to be eliminated by *purifying selection*, force that probably acts on all genes, to preserve their function.

Natural selection affects the shape of the genealogy of alleles, usually summarized in evolutionary trees whose parameters can be can be estimated through the *coalescent theory* (see Introduction, section 1.2). Positive selection, which drives an adaptive variant towards fixation, lead to an excess of low frequency variants, distorting the genealogy to create a star-like pattern (Hudson & Kaplan 1988). The genealogy of an allele under positive selection

usually presents long terminal branches connected to a common ancestor by shorter branches. This genealogy is expected to have a more recent coalescence than a genealogy of neutral alleles, since positive selection accelerates the process of allelic fixation (see Fig 1.1B).

When selection acts to remove damaging mutations, it also eliminates polymorphisms linked to the deleterious alleles, reducing the overall level of variation. The process of elimination of a deleterious mutation and the consequent reduction in variation at neutral linked polymorphisms is called *background selection*. Under the influence of background selection, an allele can be rapidly led to fixation, and, as for positive selection, this leads to an excess of polymorphisms at low frequencies. The genealogy of an allele that is driven to fixation by means of background selection has a more recent coalescent time than expected under a neutral model; this because the linked deleterious mutation caused the extinction of one lineage (the "negative selected") more quickly than would be predicted for neutral variants, hence by a simple genetic drift model (Fig 1.1C).

Natural selection does not always increase or decrease the frequency of a single allele at a locus. Sometimes, selection tends to maintain the polymorphism, preserving two or more alleles at a locus in a population. This type of selection is called *balancing selection*. We can found signatures of balancing selection, for example, in case of rare-allele advantage, which involves negative frequency-dependent selection and especially when there is generalized overdominance. In the first case, the fitness of an allele decreases as it become more common; in generalized overdominance, heterozygous individuals have a selective advantage, and this leads to an equilibrium in which two or more alleles have nonzero frequencies. This latter case is thought to be the mechanism that allows maintaining the high levels of allelic variation observed at the MHC locus (Grimsley, Mather & Ober 1998). Balancing selection tends to favor intermediate-frequency alleles, resulting in an excess of intermediate-frequency variants, and in a higher level of sequence diversity compared with neutral loci (Charlesworth, Nordborg & Charlesworth 1997; Schierup, Vekemans & Charlesworth 2000). This is reflected in genealogies with short terminal branches and longer internal branches, and having an older coalescence time respect to the genealogy expected for a neutral locus (Fig 1.1D).

Over the past few years, the interest has grown in characterizing the patterns of genetic variation in order to highlight signature of natural selection in human populations (Sabeti et al. 2006; Hernandez et al. 2011). Even so, it has been shown that most human genetic variation is neutral and that polymorphisms are fixed or eliminated in a population as a consequence of the genetic drift, reflecting the populations' historical dynamics (Balaresque, Ballereau & Jobling 2007). Demographic processes, like changes in population size or migration, are known 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).

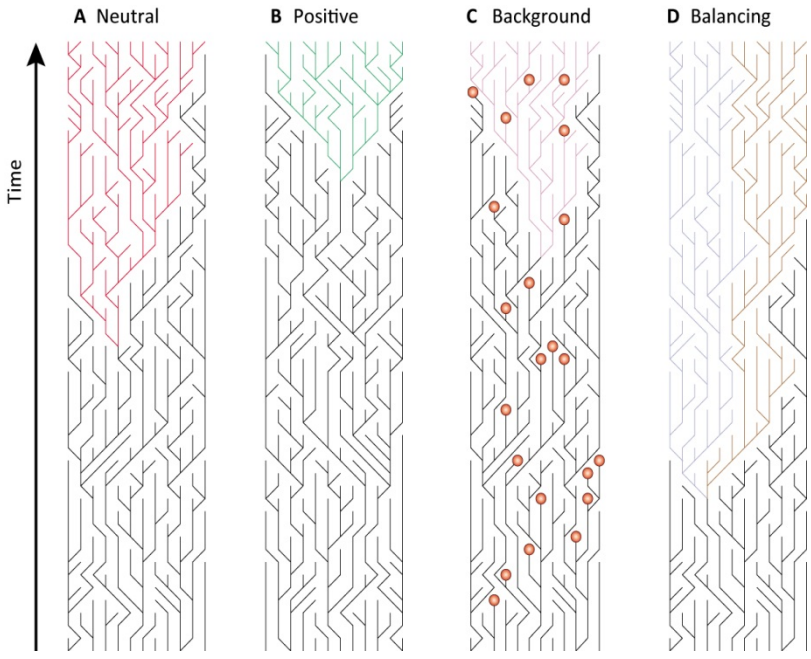


Fig 1.1. Effects of natural selection on gene genealogies and allele frequencies. Genealogies (A-D) for a population of 12 haploid individuals, considering alleles from a locus A- neutral, B- under positive selection, C- affected by background selection (each circle represents the elimination of a deleterious mutation), D-under balancing selection. Modified from Bamshad and Wooding (2003).

1.2 The Coalescent: population genetic inference using genealogies

The genetic relationships among a sample of individuals can be described by their genealogy. Genealogies are family trees which depict ancestors and descendants of individuals. In the same way that we can construct genealogies of individuals, we can construct genealogies of genes, considering that transmission of every independent gene is a single realization of a stochastic process in which one of two alleles is passed on to the offspring. Therefore, for every independent gene, there is a potentially different genealogy. Every genealogy has exactly *n* external branches, one for each gene sampled in an individual. Proceeding backwards in time, pairs of branches have a common ancestor and the number of lineages is reduced by one. This event is called a *coalescence event*. In a genealogy there are n-1 coalescence events, until the *most recent common ancestor* (MRCA) is reached of all the gene copies in the sample (Fig 1.2). Genealogies contain information about historical demography and about the processes that have acted to shape diversity of populations. In fact, we can imagine two samples of genes, one from random people coming from a large city, and the other from random people from a small town. Intuitively, we can imagine that most pairs of people from the small town will have a common ancestor only few generations ago, whereas for two people from the big city the common ancestor would be located many generations back in the past. Moreover, this way we would realize that the number of generations separating the two individuals from the common ancestor also depends on the number of people immigrating to and emigrating from the city or the small town; migration tends to push backwards the average estimates of the time since common ancestry. Again, if we know that what is now a small town had been a metropolis for a long time, we would not be so confident that two individual from this sample have a recent common ancestor. These examples show that a number of factors determine the time of the common ancestor: the size of the population, the migration rate and the changes in population size. These examples capture the importance of reconstructing the genealogy of a sample to make inferences about historical population processes and demographies. In 1982, John Kingman described this process formally in mathematical terms and called it the ***coalescent*** (Kingman 1982). From its development, the coalescent has been the basic stochastic model in the

analysis of genetic variation, allowing, via simulation, to explore the effect that changing parameters has on the data that might be observed.

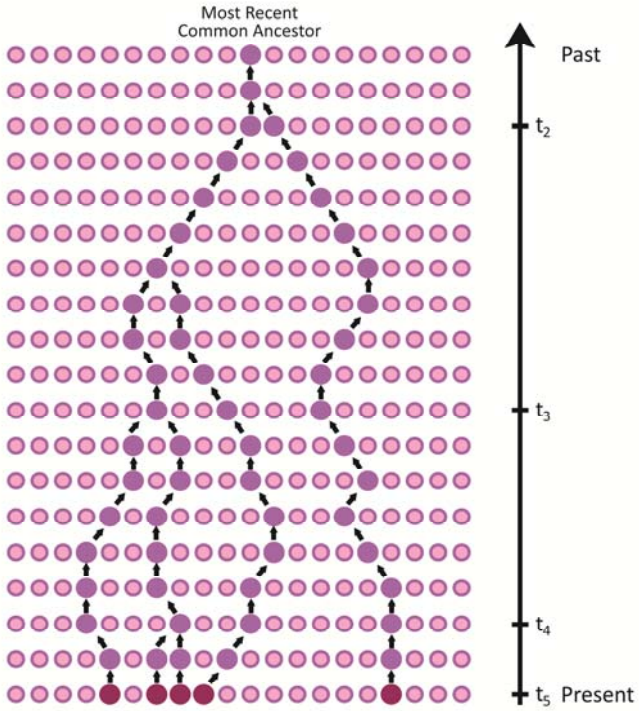


Fig 1.2. A genealogy of a sample of n individuals.

1.2.1 Kingman's Coalescent

In its simplest statement, the coalescent includes a Wright-Fisher population model (Fisher 1930; Wright 1931). In this model, a panmictic haploid population has *N* individuals, and its size remains constant over time. Generations are discrete (non-overlapping), so that at each generation only the offspring of the preceding generation survives; no selecting forces are acting on the population, and all individuals have an equal chance of producing offspring. If we sample *n* individuals from this population (with *n* larger than 2 but smaller than *N*), the history of this sample comprises *n-1* coalescence events (Fig 1.2), each event decreasing the number of lineages by one. This takes the sample from the present day when there are n lineages through a series of step in which the number of lineages decreases from *n* to *n-1*, then from *n-1* to *n-2* and so on, and finally from two to one. This last coalescent event is called the time of the most recent common ancestor, and the single lineage

remaining after this event represents the most recent common ancestor of the entire sample. At each coalescent event, two of the lineages merge into one common ancestral lineage, resulting in a bifurcating tree as shown in Fig 1.2; the time *Ti* on the figure is the time in which exactly *i* lineages remain. Because of the last assumption of the Wright-Fisher population model, individual are equally likely to reproduce, and therefore all lineages must be equally likely to coalesce; the probability that two individual will share a common ancestor in the preceding generation is $1/N$. The probability that a pair of individuals will share a common ancestor two generations ago is the probability that they will not share an ancestor in the preceding generation $(1 - \frac{1}{N})$, multiplied by the probability that their respective parents will share a common ancestor two generations ago $1/N$. We can generalize this formulation and calculate the probability that any pair of the *n* individuals will have their common ancestor *k* generations ago:

$$P(tk) = \left(\frac{1}{N}\right)(1 - \frac{1}{N})^{k-1}$$

In our sample of *n* individuals there are $n(n-1)/2$ possible pairs of individuals in the present generation that may share a common ancestor in the preceding generation. Each of these $n(n-1)/2$ possible pairs has a $1/N$ chance of having the same parent, so the probability that there will be one common ancestor in the preceding generation is:

$$P(t1) = \frac{n(n-1)}{2N}$$

and the probability that the first MRCA of any of the possible pairs in the sample will be at tK (i.e. k generations ago) is:

$$P(tk) = (\frac{n(n-1)}{2N})(1 - \frac{n(n-1)}{2N})^{k-1}$$

Kingman (1982) showed that as *N* goes to infinity, with *n* much smaller than *N*, we can move from time in discrete generation to continuous time, so that the previous equation becomes:

$$P(tk) = (\frac{n(n-1)}{2N})exp(\frac{n(n-1)}{2N}\ k)dt$$

, that is the density function of the exponential distribution, usually indicated as:

$$fTi(ti) = \binom{i}{2} e^{-\binom{i}{2}ti}$$

where *i* = 2, . . . , n, with time rescaled so that one unit of scaled time corresponds to *N* generations.

Because they are exponentially distributed, the mean and the variance of the coalescence times are:

$$E[Ti] = \frac{2}{i(i-1)}$$

$$Var[Ti] = (\frac{2}{i(i-1)})^2$$

From these equations it is clear that coalescence times are expected to increase as one proceeds backwards in time. Accordingly, the most ancient coalescence time, namely the one in which the remaining two lineages coalesce into the MRCA of the entire sample, is expected to be the longest. Especially in a large sample, many (mutually independent) coalescence events will occur over a very short period of time in the recent history of the sample. The fact that every pair of lineages is equally likely to be the pair that coalesces means that every possible genealogical tree structure is equally likely. All of the remarkable results of the standard coalescent model follow directly from these properties: the random-bifurcating nature of the coalescent trees and the independent, exponential coalescent times.

1.2.2 Demographic history

Real populations change in size over time. From the equations above it is clear that the effective size of a population correlates with the expected interval between coalescence events; changes in population sizes will result in changes to the distributions of these times.

Imagine a population that evolves according to the Wright-Fisher model, but with a different size at each generation, for example an exponentially growing population. If we sample a set of genes from this population now, hence when it has large population size (N0), we expect to find that the time to the first coalescent event will be large. After the first coalescent event, some generations in the past, the population will be smaller than N0, and the lineages will coalesce at a faster rate, proportional to the sample size at generation t (Nt). The effect of this process on the genealogy is to produce a tree with long terminal branches and shorter internal branches compared with a constant-size population tree (Fig 1.3, left panel), reflecting the fact that coalescences are more likely to have taken place when the population was small. This genealogy is said to be "star-like". Similarly, in a declining population, the effective population size at present (N0) is small relative to population sizes in the past. In this case, the first coalescence events occur rapidly, but, as one moves backwards, population sizes increase, and so on average coalescence intervals get longer (Fig 1.3, right panel).

Up to this point, the coalescent process has been described for panmictic populations. However, real populations are often spatially structured, and it is obviously important to be able to incorporate this in the model. The coalescent can be modified for a number of geographical structures, considering for example an island model (see Introduction section 1.1.4), in which the population is subdivided in demes with a certain rate of migration between them, or a stepping-stone model of migration (see Introduction section 1.1.4), where demes are arranged linearly or in a two dimensional grid, and migration can only take place between neighboring demes. In these models, the distribution of times to ancestry depends on the rate of migration between demes and on the effective population size within demes. Since two lineages can coalesce only if belonging to the same deme, if demes have small population size and low migration rate we expect that lineages within demes will coalesce relatively quickly, leaving a single ancestral lineage in each deme. Conversely, these ancestral lineages will take long time to coalesce, since this requires a rare migration event to another deme. In case of structured coalescent, the expression for the distribution of coalescence times keep in consideration the proportion of migrating individuals per generation scaled by the total number of individuals (i.e. $M = Nm$, with $0 < m < 1$).

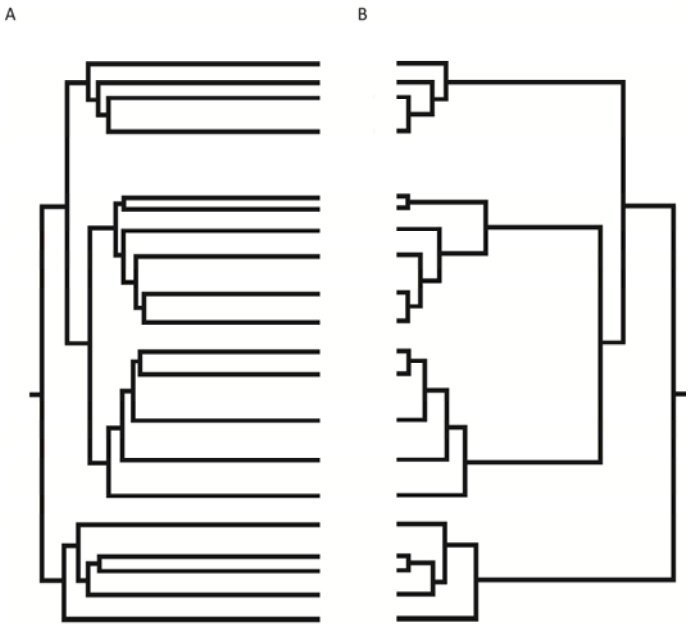


Fig 1.3. Genealogies under population expansion (left) and decline (right). Two examples of genealogies for the same number of individuals with different demographic histories. On the left, a genealogy for an expanding population, with long terminal branches and short internal branches. On the right, a genealogy for a declining population, in which lineages coalesce at a faster rate in the first part of the genealogy, but going backwards in time, coalescence intervals get longer.

1.2.3 The serial coalescent

One of the recent extensions of the coalescent involves the possibility of considering genetic samples obtained at different times. Rodrigo and Felsenstein (1999) developed the *serial coalescent*, to describe the distributions of coalescence intervals on a genealogy of samples obtained serially in time. Respect to the classical Kingman algorithm, there are two differences that arise as a consequence of sampling sequences over time. The first is the possibility to obtain a direct estimate of mutation rate simply by estimating the expected number of substitutions that accumulate over each sampling interval, and dividing by the amount of time between samples. The second difference is that in the serial coalescent, going backward in time, the number of lineages can increase. This can influence the extent to which we are able to make statements about historical population dynamics. In fact, with a standard coalescent, the number of lineages decreases steadily as one move back in time, reducing the certainty about the lengths of the coalescence intervals and so increasing the variance in the estimates of evolutionary parameters. This is particularly important when there have been changes in the population dynamics over time. On the contrary, with serial

coalescent, our ability to add sequences moving back along the genealogy means that we can increase the efficiency in estimating the time-to-ancestry. This in turns means that we have more power to detect changes in the dynamics of a population, thus rendering the analysis more informative.

1.3 The Bayesian revolution in genetics

Considerable progress in the field of population genetics has been made during the past decade, following parallel increases in computer processing speed and in the available DNA sequence data. To date, most current methods are based on the coalescent theory, the stochastic process describing how population genetic processes can shape the genealogy of the data (see Introduction, section 1.2). Coalescent-based inference methods enable population genetic parameters to be estimated directly from gene sequence data under a variety of scenarios, including variable population size (Drummond et al. 2002; Drummond et al. 2005), recombination (Kuhner, Yamato & Felsenstein 2000), and population subdivision (Beerli & Felsenstein 2001). The inference of demographic histories require a "demographic model", which is simply a mathematical function used to describe the changes in effective population size, and/or migration rate, through time. The model reflects how the data were generated, and the behavior of the model is determined by the values of a set of parameters. We use the results that have been obtained by simulation of genetic data under the tested model to estimate how populations evolved over time, i.e. to estimate population parameters defining the model under study. The traditional approach to do this is the Maximum Likelihood Estimation (MLE) method. The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data. From a statistical point of view, the method of maximum likelihood is considered to be robust and yields estimators with good statistical properties (Huelsenbeck 1995). However, although the methodology for MLE is simple, the implementation is mathematically intense and often unfeasible. The Bayesian inference is a convenient way to deal with these sorts of problems.

1.3.1 Principles of Bayesian Inference

In Bayesian and classical statistics we want to make inferences about a fixed, but unknown, parameter value; the difference is in how we approach this goal and in the interpretation of the results.

Bayesian statistics allows scientists to incorporate prior knowledge about model parameters into their data analysis, and the essence of Bayesian statistics is that there is no

logical distinction between the data and the model parameters, since they are both random variables. Being a random quantities, they have a joint probability distribution, specified by a probabilistic model in which the data are the observed variables and the parameters are unobserved variables. This joint distribution is a product of the likelihood and the prior. The likelihood measures the probability of the data given a particular set of parameter values, and is based on a model of the underlying process; the prior represents the probability distribution of the parameter values before observing the data. Together, these two functions combine all available information about the parameters. The main goal of Bayesian statistics is to manipulate this joint distribution in various ways to make inferences about the parameters; this is done by calculating the posterior distribution of the parameters, i.e. the conditional distribution of the parameters given the data. The first mathematical formulation of the Bayesian approach is attributed to Thomas Bayes, a British mathematician and Presbyterian minister. He realized that the probability of a particular value *p*, given some observed data *D,* can be calculated using the probability function:

$$P(p|D) = \frac{P(p) \times P(D|p)}{P(D)}$$

also known as **Bayes' theorem**. The function $P(p|D)$ is the posterior probability distribution, that is obtained, as said above, from the product of the prior ($P(p)$) and the likelihood ($P(D|p)$). $P(D)$ is the marginal likelihood of the data, the unconditional probability of obtaining the outcome *D* taking all passible values of *p* into account. This value is a normalizing constant, and simply ensures that the posterior probability distribution integrates to 1. These basic features of Bayesian inference are outlined in Fig 1.4.

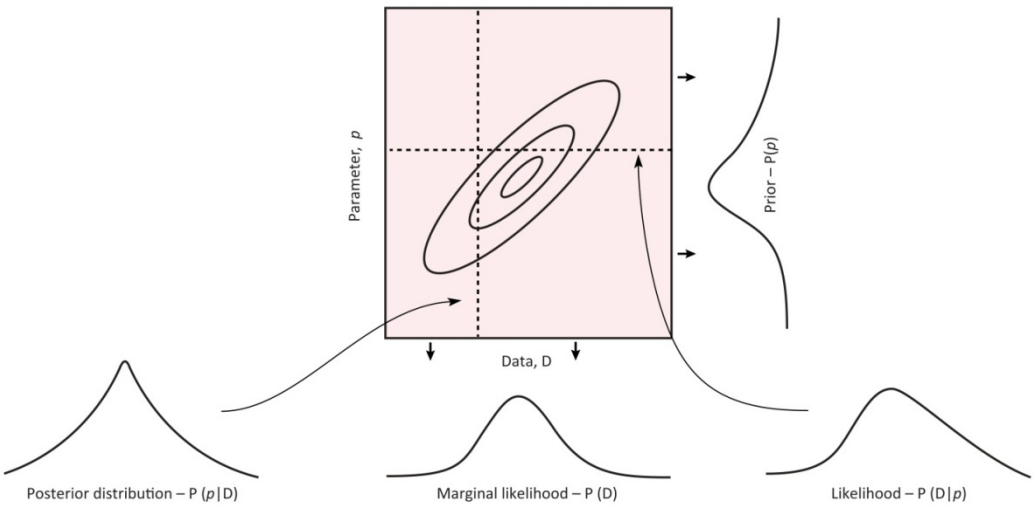


Fig 1.4. Key features of the Bayesian inference. We imagine that the data D can assume any value along the x-axis; similarly, the parameter value p can take any value along the y-axis. Bayesian inference considers the joint distribution of the parameters and the data (*P(p,D)*), represented by the contour intervals in the figure. This distribution can be obtained by the product of the prior (*P(p)*) and the likelihood (*P(D|p)*); the former is an assumed distribution of the parameters based on the background knowledge, the latter will arise from a statistical model in which it is necessary to consider how the data can be explained by the parameters. The arrows in the figure show that marginal distributions can be obtained by integrating the joint distribution over the data, recovering the prior, or over the parameter values, recovering the marginal likelihood *P(D)*. Conditional distributions are indicated by the dotted lines in the figure, and represent taking a "slice" through the joint distribution and rescaling the distribution so that the integral of possible values is equal to one. The scaling factor is given by the marginal distribution. Hence, any conditional distribution is simply the joint distribution divided by a marginal distribution. The key quantity of the Bayesian inference, the posterior distribution of the parameter given the data (*P(p|D)*), is in fact the joint distribution divided by the marginal likelihood. Modified from Beaumont and Rannala (2004).

1.3.2 Application to Phylogenetics and Population Genetics

The main purpose of phylogenetics is to make inferences about the relationships between different taxa estimating tree's parameters like topology, branch lengths and the nucleotide substitution model. By contrast population genetics is mainly interested in demographic and evolutionary parameters shaping genetic variation. For both disciplines, Bayesian methods represent an attractive development, allowing one to test complex, and so realistic, evolutionary hypotheses.

Bayesian approaches to phylogenetics generated a great deal of enthusiasm. This can be attributed to a number of factors, including that these methods enable the relatively straightforward implementation of extremely complex evolutionary models, producing both a tree estimate and a measure of uncertainty for the groups on the tree (Hughes et al. 1993; Fleming et al. 2003). Schematically, in a maximum likelihood (ML) phylogenetic analysis a hypothesis is judged by how well it predicts the observed data, and the tree that has the highest probability of producing the observed sequences is preferred; in a Bayesian phylogenetic analysis the optimal tree is the one maximizing the posterior probability, that is proportional to the likelihood multiplied by the prior probability of a phylogeny. The posterior probability of a tree can be interpreted as the probability that the tree is correct. Prior probabilities of different hypotheses (i.e. different phylogenies) convey the scientist's belief before having seen the data. In the absence of background information, a simple solution would be to use prior probability distributions largely uninformative, so that most of the differences in the posterior probability of hypotheses are attributable to differences in the likelihood. One way of doing this is to specify a uniform (or "flat") prior, in which every possible value of a parameter is given the same *a priori* probability. Thus, usually all trees are considered a priori equally probable, and the likelihood is calculated under one of a number of standard Markov models of character evolution. In principle, Bayes' rule is used to obtain the posterior probability distribution, and this probability, although easy to formulate, involves a summation over all trees and, for each tree, integration over all possible combination of branch lengths and substitution model parameter values. An important property of the Bayesian inference is that there is no sharp distinction between different types of model parameters. Once the posterior probability distribution is obtained, we can

derive any marginal distribution of interest, integrating out (marginalizing) the model parameters to which we are not interested. This is the main difference between ML and Bayesian approaches. Under the ML approach, a joint estimation of the parameters is performed, finding the highest point in the "parameter landscape". A Bayesian analysis measures the volume under a posterior probability surface rather than its maximum height. Moreover, often these parameters are nuisance parameters, not of direct interest, but must be dealt with because they are found in the likelihood equations. When complex models are used, many parameters are involved in the analysis; marginalizing becomes increasing helpful as the number of parameters increases relative to the amount of data.

In addition to phylogenetic inference, a number of Bayesian software packages have been developed for coalescent-based estimation of demographic parameters from genetic data (Rannala & Yang 2003; Kuhner 2006; Drummond & Rambaut 2007). Much like in phylogenetic analysis, they also require a gene tree in the underlying model, although, in this setting, the sequences represent different individual from the same species, rather than from different species. The development of the coalescent theory has strongly influenced many areas of population genetics, forming the basis for likelihood calculation in genealogical models (Felsenstein 1992), and allowing the use of Bayesian approaches to infer demographic history from genetic data (Atkinson, Gray & Drummond 2009; Gronau et al. 2011). In addition, Bayesian methods might be used to assign individuals to their population of origin (Pritchard, Stephens & Donnelly 2000; Tishkoff et al. 2009) and to detect selection acting on genes (Nielsen & Yang 1998; Foll & Gaggiotti 2008; Riebler, Held & Stephan 2008).

Together with progress in phylogenetic and coalescent-based population genetics, Bayesian methods have been the main factor of success in addressing many evolutionary questions. There are many practical reasons to use Bayesian inference: if a probability model includes many interdependent variables that are constrained to a particular range of values (as is often the case of genetics), maximum likelihood inference requires that a constrained multidimensional maximization be carried out to find the combined set of parameter values that maximize the likelihood function. This often entails a difficult numerical analysis problem and may require enormous computational efforts. In addition, some

approximations are required to calculate confidence intervals, approximations that are most accurate for large sample size. On the other hand, in Bayesian inference, in which the priors automatically impose the parameter constraints, inferences about parameter values on the basis of the posterior distribution require integration rather than maximization, and no other approximations are involved. Moreover, the development of numerical methods to study properties of complex probability distributions (i.e. Markov Chain Monte Carlo, see below) have greatly facilitated the evaluation of Bayesian posterior probabilities, making the calculations tractable even for complicated genetic model.

1.3.3 Markov Chain Monte Carlo Sampling

In most cases it is impossible to derive the posterior probability distribution analytically. The reason is that most of the posterior probability is likely to be concentrated in a small part of a vast parameter space. Even with a massing sampling effort, it is highly unlikely that we would obtain enough samples from the interesting region of the posterior distribution. Fortunately, a number of numerical methods allow one to approximate the posterior probability, the most useful of which is **Markov chain Monte Carlo** (MCMC) (Gilks, Richardson & Spiegelhalter 1996). MCMC has revolutionized Bayesian inference, with applications to Bayesian pylogenetic (Brown & Yang 2010) and population genetics (Choi & Hey 2011) inference. Markov chains have the property that they converge toward an equilibrium state regardless of their starting point, so we just need to set up a Markov chain that converges onto the posterior probability distribution. This can be achieved using different methods, the most flexible of which is known as the *Metropolis algorithm* (Metropolis 1953). In 1970 Hastings (Hastings 1970) introduced an important extension, and so the sampler is referred as Metropolis-Hastings method. The basic idea is to construct a Markov chain that has as its state space the parameters of the statistical model, and as stationary distribution the posterior distribution of the parameters. The MCMC algorithm involves the following steps. The chain starts at an arbitrary point in the parameters landscape (Θ). In the next generation of the chain, a new point is considered (Θ*) drawn from a proposal distribution $f(\Theta^*|\Theta)$. The ratio of the posterior probabilities at the two points is then calculated (= $P(\Theta^*|D)/P(\Theta|D)$); if the new point has higher posterior probability (the point is “uphill”), the chain moves to this state and it becomes the starting

point for the next cycle of the chain; otherwise, if the ratio is <1, the new state is accepted with a probability that is proportional to the height ratio (Fig 1.5). After this, a new state is proposed. It turns out that for a properly constructed and adequately run Markov chain, the amount of time it eventually spends sampling a particular parameter value or interval is proportional to the posterior probability of that value or interval.

The chain starts from random parameter values, and it is quite likely that the initial likelihoods are low, so low that is not really fair to consider those points as being drawn from the posterior distribution to be estimated. This early phase of the run is known as the burn-in, and the burn-in samples are often discarded because they are heavily influenced by the arbitrarily-chosen starting point. After a phase in which the posterior probabilities tend to increase, the chain reaches the stationary distribution. At this point the likelihood values tend to a plateau, and this can be confirmed from the trace plot, i. e. the plot of the likelihood values against the generation of the chain. Looking at the trace plot is important to monitor the performance of an MCMC analysis, since we are not only interested in reaching stationarity, but also in an adequate coverage of this region (which means that there has been convergence of the sample to the stationary distribution). The convergence diagnostics helps determine the quality of the sampling from the posterior distribution. Three different types of diagnostics are currently in use: examining autocorrelation times (effective sample sizes), comparing samples from successive time segments in a single chain, and comparing samples from different runs started from different space points. The speed with which the chain covers the interesting region of the posterior is known as mixing behavior. The better the mixing, the faster the chain will generate adequate sample from the posterior. To improve mixing, and thereby convergence, it might be possible to implement a Metropolis-coupled version of the algorithm (Geyer 1991) in which multiple chains are run simultaneously, with all chains but one having heated stationary distribution. This heating is achieved by raising the posterior probability to a power smaller than one. The effect is to flattened out the posterior probability surface, and if the surface is flattened, a Markov chain will move faster in the space. This is useful also if local maxima, i.e. isolated peaks of probability are present in the space, and the chain may get stuck on one of these local maxima, thus disregarding the absolute maxima of the posterior distribution. The heated chains will not individually return the correct posterior distribution but they will explore the

state space more quickly than the non-heated chain (cold chain) will. At regular intervals, there is a swap of the states between two randomly picked chains, and if the cold chain is one of them, it can jump a considerable distance in parameter space in a single step. In this way the overall mixing of the cold chain may be substantially improved.

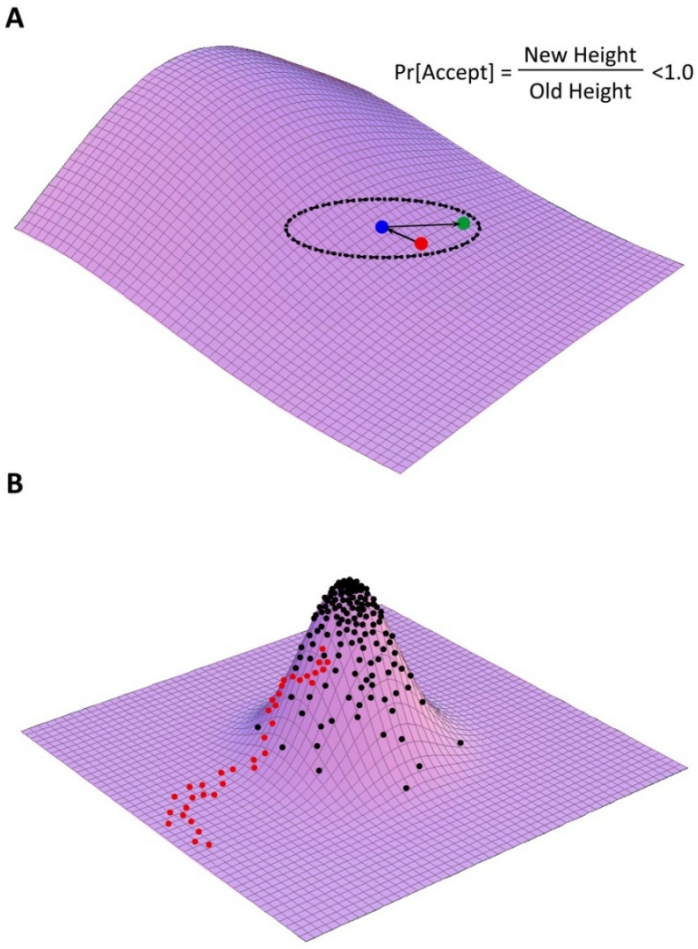


Fig 1.5. Markov chain Monte Carlo procedure. MCMC analysis is used to generate valid samples from the posterior. A: The chain is started at a random point (red), and a new state is proposed according to a proposal distribution (blue). If the new point is uphill, it will be always accepted as the new point of the chain. When another state is proposed (green) that is downhill with respect to the current state (blue), we accept it with a probability is proportional to the height ratio. B: The chain explores the parameters space until reaching stationarity. The initial running of the chain before approaching the stationary distribution is the burn-in phase (red points). After that, the chain starts to explore the posterior distribution (black points) and the amount of time it spends sampling a region of the parameters' space (proportional to the density of black circles) is proportional to its posterior probability.

1.3.4 Bayesian Model Choice

So far, in referring to the posterior distribution, we have always considered implicitly that it was conditioned on a specific model. To make it explicit, we could write Bayes' theorem as:

$$P(p|D,M) = \frac{P(p|M) \times P(D|p,M)}{P(D|M)}$$

It is now clear that the normalizing constant (*P(D|M)*), is the probability of the data given the chosen model after we have integrated out all parameters. This quantity is known as "model likelihood" and is used for Bayesian model comparison. Indeed, if we assume that we are choosing within two models, M0 and M1, the ratio of their posterior probabilities can be calculated as:

$$\frac{P(M0|D)}{P(M1|D)} = \frac{P(M0)}{P(M1)} \times \frac{P(D|M0)}{P(D|M1)}$$

The first factor is the prior odds, and the second factor is known as the Bayes Factor, which is the ratio of the model likelihoods, calculated as the harmonic mean of the likelihood values from the stationary phase of an MCMC run. When the compared models have the same prior probability, the first factor is equal to one, the Bayes Factor is the same of the posterior odds, and from it we can get information about the support of the data to model 0 with respect to model 1. The interpretation of a Bayes Factor comparison is up to the investigator, but some guidelines were suggested by Kass and Raftery (1995). An alternative of the Bayes Factor to compare models is the reversible-jump MCMC. Instead of running a full analysis on each model and then choosing among them using the estimated model likelihoods, in a reversible jump MCMC a single Bayesian analysis explore the models in a predefined model space. In this case, all parameters estimates will be based on an average across models, each model weighted according to its posterior probability.

1.3.5 Summarizing the data: the Approximate Bayesian Computation

All these methods are computationally intensive, and analyzing the data fully and accurately becomes impossible when loci are many and the models complex. In MCMC methods the difficulty lies in evaluating the likelihood, and in evaluating it in a reasonable time. In fact, even if the statistical estimation of mutation and demographic parameters have drastically improved in the last 10 years (Marjoram & Tavarè 2006), these methods are still restricted to relatively simple models for which the likelihood function can be computed, or to small dataset that can be analyzed in a reasonable amount of time. The increasing

production of genetic data and the need to simulate more realistic (which usually means complex) models, has led to the development of methods that try to approximate the likelihood. One of these methods was firstly proposed by Fu and Li (1997) and Tavaré et al (1997), and then by Weiss & Von Hassler (1998) and Pritchard et al (1999) under the name of Approximate Bayesian Computation (ABC). Few years later, Beaumont et al. (2002) formalized and generalized this approach introducing a series of improvements, so that the actual birth of the ABC coincides with his study. The ABC methods, for the first time in population genetics, combine the analysis of abundant data and realistic modeling. 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. An explanation of a complete ABC analysis, detailing the approaches used in this thesis, can be found in the Methods section (3.2.2).

In general, the idea behind the classical ABC methods is to use simulations across a wide range of parameter values within a model to find the parameter values that match most closely those in the observed data. Initially, at each iteration of the simulation step, the simulated data D' were compared with the observed data D, and if D' were identical to D, the parameters that generated that dataset were stored, and discarded otherwise (Tavaré et al. 1997). At the end of the simulation step the retained parameters were used to estimate the posterior distribution. Since this procedure is very unlikely to produce a dataset identical to the observed one, whenever the data are many and/or the models are complex, it has been proposed to replace the data with a set of summary statistics (S), and to retain a simulation only if the simulated set of summary statistics (S') are sufficiently close to the observed S (Pritchard et al. 1999). In order to account for the difference between S and S', Beaumont et al (2002) proposed to perform a local weighted linear regression to compute the posterior distribution, and this adjustment showed to substantially improve estimation. Recently, Leuenberger and Wegmann (2010) propose to reformulate the regression step using the General Linear Model (GLM) to improve the fit of the relationship between parameters and summary statistics in the retained simulations (ABC-GLM). Other improvements have recently been proposed, to increase the efficiency of the simulation step. Indeed, during an ABC analysis, all simulations are independent; this means that if a

simulated genealogy produces a data set of statistics similar to the observed one, the next simulation can be absolutely useless. Millions of simulations are needed to be sure of approaching the real values a sufficient number of times to opportunely calculate the posterior distributions. Interesting solutions, which I will not describe in detail here, were proposed by Wegmann and colleagues (2009; MCMC without likelihood (ABC-MCMC)) and Beaumont et al (2009; Population Monte Carlo (PMC)).

In short, the whole ABC machinery is based on comparisons between observed and simulated statistics, calculated respectively on observed and simulated data sets of genetic variation. The choice of the statistics is recognized as one of the most important step (Beaumont, Zhang & Balding 2002; Marjoram et al. 2003), but there is still no general rule about which and how many statistics should be used. The set of statistics has to be "sufficient", to capture the whole information contained in the data about the model parameters, but what "sufficient" means is difficult to say. Increasing the number of summary statistics calculated on the data obviously increases the amount of information considered, but other issues may arise; the larger the number of summary statistics, the larger the statistical noise included in the posterior estimation (known as "curse of dimensionality", Joyce & Marjoram 2008). In other word, by considering many variables one takes the risk to give limited or insufficient weight to the variables that would be most informative about the process of interest. Many approaches have been proposed to solve this trade-off between information and stochastic noise. Between these, Joyce and Marjoram (2008) proposed to score the different summary statistics based on their impact on the inference and Wegmann et al. (2009) proposed to transform the summary statistics via Partial Least Square to obtain a set of orthogonal linear combination of statistics best explaining the variance in the model parameter space. Alternatively, principal component analysis (PCA) can be used to select statistics most correlated with the model parameters variance (Bazin, Dawson & Beaumont 2010).

A second, critical point is the criterion to identify the model best accounting for the data. There are two main methods for the model selection in the ABC procedure, detailed in the Methods section. The first one is a "direct" method proposed by Pritchard et al. (1999) in which, after pooling all the simulations generated under different models, only those falling

within an arbitrary distance threshold from the real data are retained. The posterior probability of each model is then calculated as the fraction of retained simulations produced by each of them. This method is simple and straightforward; however it could be inaccurate if the distance threshold between observed and simulated statistics is not close to zero. To solve this problem, Beaumont (2008) proposed to improve the model selection procedure using a logistic regression approach (see Methods for details).

When the likelihood function can be evaluated, there is no advantage of using ABC as alternative. However, for many applications of population genetics, the likelihood function can be evaluated in principle, but in practice it is computationally too expensive. Moreover, the trend is analyze increasingly large datasets and to interpret them in the light of more realistic models, for which ABC methods can provide reasonable good estimates in a reasonable computational time.

1.4 Ancient DNA

For many years, inferences about the past of human populations could only come from the study of modern genetic variation. With the advent of ancient DNA techniques is it possible to add the genetic information coming from humans and pre-humans and to address directly questions such as the evolution of genes involved in human specific traits, the analysis of diversity of ancient populations and the reconstruction of their histories, the determination of past frequencies for alleles involved in phenotypes such as pigmentation, dietary adaptation linked to agriculture, and responses to particular pathogens. Unfortunately, there are lots of practical difficulties with ancient DNA analysis in general, and analysis of human samples in particular, due to the postmortem degradation of molecules of DNA and contamination with ubiquitous modern DNA.

1.4.1 Molecular damage

Within living cells, the integrity of DNA molecules is maintained by enzymatic processes (Lindahl 1993). After the death of an organism, cellular compartments that normally seize catabolic enzymes stop working, and, as a consequence, DNA is degraded by enzymes such as lysosomal nucleases. Under some rare conditions, a tissue becomes rapidly desiccated after death, or the DNA becomes adsorbed to a mineral matrix, escaping enzymatic degradation. Besides the enzymatic degradation, some other chemical processes can affect DNA in a dead cell; many of these are similar to those affecting the DNA in living cells, with the difference that in a living cell these processes are counterbalanced by cellular repair processes. After death, damages accumulate progressively until the DNA loses its integrity and decomposes, with an irreversible loss of nucleotide sequence information. With the development of polymerase chain reaction (PCR), that made it possible to produce unlimited number of copies of the same sequence of DNA from very few or even single original DNA copies, the salvage of information from rare samples in which disintegration of DNA is not yet complete is possible, although technically challenging.

Another problem of the DNA extracted from subfossil and fossil remains is its degradation to small fragments, usually between 100 and 500 base pairs in size (Hofreiter et al. 2001). This degradation is due both to enzymes and to hydrolytic cleavage of

phosphodiester bonds in the phosphate-sugar backbones (Lindahl 1993), and of glycosidic bonds between nitrous bases and the sugar backbone. The extent of degradation by these processes depends on the preservation of the specimens, and represents a limit during a PCR amplification. Moreover, the functionality of the PCR is limited by lesions blocking the elongation of DNA strands by *Taq* polymerase. These lesions are induced by free radicals, which are created by background radiation, attacking the double bounds of pyrimidines and purines (major sites of oxidative attack) leading to ring fragmentation. DNA extracted from fossil remains is susceptible to cleavage with endonuclease III, which is specific for oxidized pyrimidines (Paabo 1989); sequences with higher amounts of oxidized pyrimidines could not be amplified via PCR (Hoss et al. 1996).

In addition to fragmentation and DNA modification that prevent the extension of DNA polymerase, there are other common damages in ancient DNA. Some of these are problematic for the investigator because even if they allow the amplification of template molecules, they cause incorrect bases to be incorporated during the PCR. An example is the hydrolytic loss of amino groups from the bases adenine, cytosine, 5-methylcytosine and guanine, resulting in hypoxanthine, uracil, thymine and xanthine, respectively (Friedberg, Walker & Siede 1995). When the deamination produces uracil, thymine and xanthine are incorrectly inserted by PCR. Clearly, the risk of determination of incorrect DNA sequences due to misincorporations is great if amplification starts from a single DNA molecule and if DNA sequences are determined from a single amplification. Under such conditions, any consistent misincorporation would result in an incorrect base being determined. A way around this problem is to perform more amplifications and compare the results.

1.4.2 Contamination with exogenous DNA

Ancient samples may not contain endogenous DNA detectable with current techniques. However, if primers that amplify current human DNA are used to perform amplifications from non-human remains, they often yield DNA sequences identical to those found in contemporary humans (Serre et al. 2004). This means that, together with the ancient specimen's DNA, and sometimes in the absence of amplifiable DNA, modern DNA is present in many ancient samples. Identifying it is easy in studies of non-human species, but

not at all when the specimen's DNA does not differ by much from the contaminant's DNA (Handt et al. 1994; Handt et al. 1996; Hofreiter et al. 2001; Wall & Kim 2007). This problem might be alleviated in two ways: first, it is necessary to handle specimens, perform DNA extraction, and set up amplifications in dedicated laboratory facilities, where no post-PCR work has ever been conducted (Paabo 1990), and where all extraction work is conducted with protective clothing and the work space cleaned regularly with oxidant such as bleach and irradiated with UV lights; second, it was suggested to follow some criteria of authenticity (Paabo 1989), detailed below. The first published criteria of authenticity (Paabo 1989) were limited to three points: (a) testing of control extracts should be performed in parallel with extracts from old specimens to detect contamination introduced from reagents and solutions during the extraction procedure; (b) more than one extract should be prepared from each specimen and both should yield identical DNA sequences; (c) there should be an inverse correlation between amplification efficiency and size of the amplification product, reflecting the degradation and damage in the ancient DNA template. Later, these criteria have been expanded (Cooper & Poinar 2000; Hofreiter et al. 2001), and they can now be summarized as follows:

1. Cloning of amplification products and sequencing of multiple clones. This serves to detect heterogeneity in the amplification products, due to contamination, DNA damage, or jumping PCR (Paabo, Irwin & Wilson 1990).

2. Extraction controls and PCR controls. Each set of extractions should include at least one extraction control that does not contain any sample material but is otherwise treated identically. Similarly, for each set of PCRs, multiple negative PCR controls should be performed to differentiate between contamination that occurs during the extraction and during the preparation of the PCR.

3. Repeated amplifications from the same or several extracts. This serves two purposes. First, it allows detection of sporadic contaminants; second, it allows detection of consistent changes due to miscoding DNA lesions in extracts containing extremely low numbers of template molecules.

4. Quantitation of the number of amplifiable DNA molecules. This shows whether consistent changes are likely to occur or not. If consistent changes can be excluded (roughly for extracts containing >1000 template molecules), a single amplification is sufficient.

5. Inverse correlation between amplification efficiency and length of amplification. Because ancient DNA is fragmented, the amplification efficiency should be inversely correlated with the length of amplification. If not, there is reason to believe that the DNA extract is largely composed of exogenous molecules.

6. Biochemical assays of macromolecular preservation performed on amino acids. This method serves two purposes: first, to support the claim that a specimen is well enough preserved to allow the preservation of DNA, secondly, to perform a rapid screening to identify specimens that, according to their general state of preservation, may contain DNA. To this aim, the most widely technique used is based on the analysis of amino acids present in specimens, relating on the combination of total amount of amino acids, their composition, and their extent of racemization. Samples that contain very little amino acids, indicating that the macromolecules in the specimens have been replaced by microorganisms, or where amino acids are extensively racemized, are unlikely to contain endogenous DNA.

7. Exclusion of nuclear insertions of mtDNA. It is highly unlikely that several different primer pairs all preferentially amplify a particular nuclear insertion. Therefore, substitutions in the overlapping part of different amplification products are a warning that nuclear insertions of mtDNA may have been amplified. A lack of diversity in population studies can also be taken as an indication that nuclear insertions may have confounded the results.

8. Reproduction in a second laboratory. This serves a similar purpose as criteria 2 and 3, i.e., to detect contamination of chemicals or samples during handling in the laboratory. Note that contaminants that are already on a sample before arrival in the laboratory will be faithfully reproduced in a second laboratory.

(Paabo et al. 2004)

Even if all the criteria are followed, this hardly represents a positive proof that a DNA sequence is genuinely ancient. Indeed, if a specimen is contaminated within a certain DNA

sequence, all criteria can be verified, but the results would still be invalid. When the ancient DNA comes from animal, contamination with modern human DNA is easily retrievable; but that is not so simple when the DNA comes from ancient humans. In the last case, stricter criteria ought to be followed, such as to verify that the sequence determined from the ancient specimen is not present in all the investigators, including excavators, museum personnel, or laboratory researchers.

So far, the most common marker used in the ancient DNA study is the **mitochondrial DNA** (mtDNA). This is because mtDNA is present in several hundreds of 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 are nuclear sequences. In the last years, the development of high-throughput DNA sequencing technologies (Bentley et al. 2008) allows large-scale, genome-wide sequencing of random pieces of DNA extracted from ancient human specimens, until obtain complete ancient genomes (Green et al. 2010; Reich et al. 2010); however the degree of confidence related to these genomes is still low, and, to date, they cannot be safely used in a comparative analysis with modern humans.

2. Purpose of the Thesis

In this thesis we compare different datasets of modern and ancient human populations living in the same geographical areas in different periods of time. This has been done in order to highlight traces of past genetic variation in modern populations, and to evaluate whether, and to what extent, ancient and modern populations can be considered genealogically related.

To do this, we analyzed the data within the approximate Bayesian computation framework, that allows us to simulate complex (and hence, realistic) demographic models including the genetic information coming from ancient populations. Moreover, the Bayesian philosophy allowed us 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 evolutionary histories of the considered populations. For the first time we applied this methodology to datasets of ancient and modern human variation, studying the genealogical relationships between archaic humans (i.e. Neandertals), anatomically modern humans (i.e. Cro-Magnon) and modern Europeans (see Applications, section 4.1), between ancient (Nuragic) and modern Sardinians (see Applications, section 4.2), and between Etruscans and modern Tuscans (see Applications, section 4.3). Within this framework, for each dataset considered, we explicitly compared different demographic models estimating the most probable mechanism of evolution of the data, we estimated the combination of demographic end evolutionary parameters of the most probable model and we evaluated in several ways the quality of our estimates.

3. Methods

3.1 Measuring and summarizing genetic variation

Genetic data can be summarized by summary statistics calculated on the data (for example on DNA sequences). Even if these statistics do not encapsulate all the information present in the data, and in general are not sufficient for reliable inference about the evolutionary processes that have generated the data, the description of the data is an important starting point to have an idea about the amount of population's diversity and population's structure.

Descriptive statistics of genetic interest can be mainly grouped into two categories: statistics calculated to summarize genetic variation within populations (i.e. **intra-population statistics**, 3.1.1), and statistics calculated between populations (**inter-populations statistics**, 3.1.2) to highlight their degree of genetic differentiation. Below, I report the statistics we used to summarize ancient and modern mitochondrial DNA data; all the statistics were calculated with the software Arlequin 3.5.1 (Excoffier & Lischer 2010).

3.1.1 Genetic variation within population

We summarized genetic variation within population through the following statistics:

Haplotype number: number of different sequences in the sample.

Segregant sites: number of sites in the sample showing more than one allele per locus.

Gene diversity: the gene diversity calculated on haploid data such as mitochondrial DNA is equivalent to the expected heterozigosity for diploid data. The gene diversity at a locus is defined as the probability that two randomly chosen haplotypes are different in the sample:

$$\widehat{H} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right)$$

Where *n* is the number of gene copies in the sample, *k* is the number of different haplotypes in the sample, and p_i is the sample frequency of the i-th haplotype (Nei 1987).

Mean number of pairwise differences (π): mean number of differences between all pairs of haplotypes within the sample. It is given by:

$$\hat{\pi} = \frac{n}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j \hat{d}_{ij}$$

Where $\hat{d}_{ij}$ is a count of the number of differences between *i* and *j* (i.e. the number of mutations having occurred since the divergence of haplotype *i* and *j*),*k* is the number of haplotypes, p_i is the frequency of haplotype *i*, and *n* is the sample size (Tajima 1983). Analogous to the mean number of pairwise differences is the **nucleotide diversity**: it represents the probability that two copies of the same nucleotide drawn at random from a set of sequences will be different from one another, and is calculated as the mean number of pairwise differences divided by the total length of the sequence.

Tajima's D: this statistic compares two estimates of theta (θ), the population mutation parameter that represents the level of variation in a population under mutation-drift equilibrium. Under neutral evolution, when equilibrium is reached, the generation of new alleles by mutation is balanced by the elimination of alleles by drift; hence the expectation is that, under neutrality, different estimates of θ should be equal. Tajima's D compare two different estimates of θ, one based on the number of segregating sites (Theta S), and the other based on the nucleotide diversity (Theta π). Theta S is estimated from an infinite-site equilibrium relationship (Watterson 1975) between the number of segregating sites (S), the sample size (n) and θ for a sample of non-recombining DNA:

$$\widehat{\theta_S} = \frac{S}{a_1}$$

where:

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

(Tajima 1989)

Theta π is estimated from the infinite-site equilibrium relationship between the mean number of pairwise difference (π) e θ:

$$E(\widehat{\pi}) = \theta$$

(Tajima 1989)

The test statistic D is then defined as:

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{Var\,(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

(Tajima 1989)

Under neutrality the two estimates are expected to be equal, and so Tajima's D is expected to be zero. The significance of the D statistic should be tested by generating random samples under the hypothesis of selective neutrality and population equilibrium, using a coalescent simulation algorithm adapted from Hudson (1990). The P value of the D statistic is then obtained as the proportion of random D statistics less or equal to the observation. Significantly positive values of this statistic indicate that the differences between alleles are greater than expected from the level of variation, a phenomenon often caused by population subdivision or balancing selection. When the value of Tajima's D is significantly lower than zero, meaning that there are many alleles with respect to variation as measured by pairwise differences, this may often be due to a population expansion or positive selection.

3.1.2 Genetic distance measures

To estimate the degree of genetic differentiation between populations, we used the following statistics:

Hudson's Fst: this statistic measures the degree of variation between pairs of populations and is based on the mean number of pairwise differences within and between populations. It is calculated as:

$$Fst = 1 - (\frac{H_w}{H_b})$$

(Hudson, Slatkin & Maddison 1992)

Where H_w is the mean number of differences between different sequences sampled from the same subpopulation, and H_b is the mean number of differences between sequences sampled from the two different subpopulations sampled.

Haplotype Sharing: similar to allele sharing for genotypic data, this statistic represents the degree of genetic similarity between pairs of samples. It is calculated as the number of haplotypes that are shared between two samples (e.g. between pop1 and pop2), divided by the number of haplotypes in pop1 (haplotype shared between pop1 and pop2 respect to pop1) or in pop2 (haplotype shared between pop1 and pop2 respect to pop2).

3.2 Inference from diversity: estimating parameters from molecular data

Recent population genetics methods (i.e. coalescent based methods) can help us understand the evolutionary and demographic processes at population level. These methods are implemented in various software packages and programs, which have grown enormously in last years. In this section, I outline the two principal methodologies we used to analyze the data. The first is a likelihood method based on the Isolation with Migration model (3.2.1) (Nielsen & Wakeley 2001) that we applied to study the relationships between two modern populations in the Etruscan study (see Applications, section 4.3). Secondly, when the goal was to highlight the genealogical links between ancient and modern populations, the models became more complex (involving more populations and an elevate number of parameters) and cannot be analyzed by classical likelihood methods. To bypass this problem we referred to approximate Bayesian computation methods (3.2.2), by which the data are not fully considered but are summarized by means of statistics, allowing to simulate genetic data according to any demographic model.

3.2.1 The Isolation with Migration model

The Isolation with Migration (IM) model provides a statistical framework making it possible to discriminate between two factors leading to increased genetic similarity of populations, namely common origin and gene flow. The IM model tests for the relative weight of common ancestry, drift and gene flow in two (or more) populations. Consider a general IM model in which an ancestral population gives rise to two populations, after which there may be gene exchange between these two populations (Fig 3.1). In its original formulation the model has six main parameters, namely the size of the three populations (*NA*, *N1*, *N2*), the time of the splitting event (t), and the rates of gene flow between daughter populations (*m1*, *m2*). The IM model differs from classical population-genetics models (Wright's island model, Malécot and Morton's isolation-by-distance model) in that it does not require the (often unlikely) assumption that mutation, genetic drift and gene flow have reached an equilibrium. As such, it may be used to quantify the roles of these factors in determining the degree of genetic relatedness between populations, a classical question in

evolutionary genetics. Indeed, in principle, a certain level of similarity between two populations may reflect a recent common origin followed by isolation, or a remote common origin followed by genetic exchanges, or anything in between. By the IM method one obtains maximum-likelihood estimates of the parameters describing the effects of drift (t, NA, N1, N2) and gene flow (m1, m2) (Nielsen & Wakeley 2001; Hey & Nielsen 2004).

At first, Nielsen and Wakeley (2001) developed a Bayesian framework for fitting this six-parameters version of the IM model to data from a single, nonrecombining locus drawn from two population or closely related species. A few years later, Hey and Nielsen (Hey & Nielsen 2004) introduced an extension allowing for multilocus analysis, and wrote a computer program to implement the method (freely available at http://genfaculty.rutgers.edu/hey/software). In this formulation the IM model could not account for changes in population sizes, and for the sizes of founding populations. Later both these issues were addressed by including a seventh parameter, s, representing the proportion of members of the ancestral population giving rise to each daughter populations (respectively, s and 1-s) (Hey 2005).

Under the assumption of selective neutrality and no recombination within loci, the IM software repeatedly generates gene genealogies by Monte Carlo Markov Chain (MCMC) simulation (see Introduction, section 1.3.3). Each gene genealogy is generated choosing random values (within a predefined interval) of the six or seven parameters. Each new parameter value is accepted or rejected, according to standard criteria, until the parameter space is explored and stationarity is reached. One way to have an idea whether the program generated a good estimate of the parameter (i.e., whether there was convergence), is to run repeated analyses that differ only for the random parameter values from which the simulations start. At the end, one can see whether the same parameter distributions are obtained. Another possibility is to observe over the course of a run how accurately the parameter space is explored. In the IM software this is done by plotting recorded values over the course of the run and then by measuring how these values are autocorrelated over the length of the run. If the autocorrelation persists for a large number of steps, this means that the space is being explored slowly, and longer runs are required. To improve mixing, and

thereby the convergence, the IM software allows the Metropolis coupling of Markov chains, where multiple chains are run simultaneously (see Introduction, section 1.3.3).

We applied this method to the analysis of the Etruscan population (for details see Applications, section 4.3), in order to estimate the time of the separation between Southwestern Anatolia population (Etruscans' homeland according to Herodotous) and the Tuscan populations related to the Etruscans. Estimating the separation time between these two populations allows as understand whether the genetic resemblances between Turks and Tuscans can be referred to a common origin just before the onset of the Etruscan culture (hence not more than 3,000 years ago), placing the Etruscans' homeland in Anatolia, or, rather, the time estimated supports the autochthonous development of the Etruscan culture in Italy.

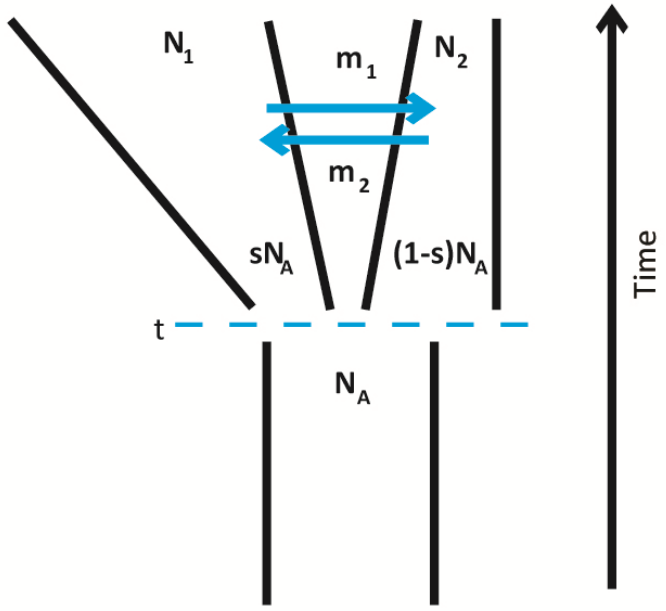


Fig 3.1. A scheme of the basic model of isolation with migration with the additional parameter s, as reported in Hey (2005). It is assumed that, *t* generations ago, an ancestral population of size N_A split into two daughter populations of sizes, respectively, N_1 and N_2, connected by gene flow. The rates of gene flow between daughter populations are expressed by m_1 and m_2, and *s* is the proportion of the members of the ancestral population giving rise to the first daughter population.

3.2.2 Likelihood free inference: the Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a flexible framework developed to choose among alternative models and to infer their parameters. Its flexibility depends on the likelihood-free inference allowing to analyze complex, and therefore realistic, demographic models (see Introduction, section 1.3; for a review see Bertorelle, Benazzo & Mona 2010). We applied this framework to test the genealogical relationships between modern and ancient populations living in the same area in different periods of time, and to estimate demographic and evolutionary parameters for the models showing the highest fit with the data. The ABC algorithm we used was firstly proposed by Beaumont in 2002 (Beaumont, Zhang & Balding 2002) as an extension of the simple rejection procedure by Pritchard et al (1999). This procedure includes the following steps:

1. First of all, one has to "set the scene", that is specify the history and the demography of the populations using a model of evolution with the specific parameters. If one is interested in testing among different hypotheses, several models can be designed and compared.

2. For each demographic model thus defined, millions gene genealogies are simulated. In our analyses, these genealogies were generated using a serial coalescent algorithm by the Bayesian version of SERIALSIMCOAL (Anderson et al. 2005; freely available on http://iod.ucsd.edu/simplex/ssc/BayeSSc.htm). Using this software it is also possible to include samples collected at different moments in time. Suppose, e.g., that one has samples of sizes n_0, n_1, $n_2 \ldots n_k$ of populations studied 0, t_1, $t_2 \ldots t_k$ generations ago. The program generates genealogies proceeding backwards in time, starting with n_0 samples in the present (t_0) and adding n_1, $n_2 \ldots n_k$ samples at the appropriate moments in the past. The genealogy is then extended backwards until it reaches the most recent common ancestor (MRCA) of the sampled lineages through a series of coalescence events (see Introduction, section 1.2). At this stage, mutations are added onto the tree according to an infinite-site model. The parameters defining the model (population sizes, mutation rates, timing of demographic processes) are considered as random variables, and their values are extracted from broad prior distributions, representing the knowledge on the parameters before the analysis; samples ages and sizes are equal to those of the observed samples.

3. Observed and simulated data are summarized using the same set of statistics; the most common statistics used in our studies were described in the previous section (Methods, section 3.1)

4. For each simulated dataset, a Euclidean distance δ between the observed and simulated summary statistic is calculated. Model selection and parameters estimation (see below) are based on the δ values thus estimated.

3.2.2.1 Model Selection

The ABC methods make it possible to compare alternative hypotheses about a process, and assign a probability to each hypothesis tested (i.e. simulated) referring to the same set of data. For our analyses, we calculated the posterior probabilities of the models in two ways.

The first criterion is based on the simple rejection procedure (AR) proposed by Pritchard et al (1999), for which model posterior probabilities are computed by counting how many simulations run under the i-th model (*ni*) are found among an arbitrarily-defined number of simulations resulting in the shortest δ between observed and simulated data (*nt*). The posterior probability for the model is then = *ni*/*nt*. Results of previous studies suggest that straightforward rejection may not be robust when considering more than a few hundred simulations (Beaumont 2008), and so, when using this approach, we considered *nt* equal to 100, i.e. we selected the 100 closest simulations.

Under the second criterion, proposed by Beaumont et al (2008), the posterior probability for each model can be computed by means of a weighted multinomial logistic regression procedure (LR). In the ABC simulations the summary statistics are the predictive variable, and the model parameters are the response variable; under the logistic regression method the model is the categorical dependent variable *Yj* ($1 \leq j \leq n$ for *n* tested models). The regression is local around the vector of observed summary statistics, and the simulations are weighted by an Epanechnikov kernel according to their distance from the observed data set. The maximum likelihood values of the β coefficients of the regression model are then estimated. The probability of the model is evaluated in the point corresponding to the observed vector of summary statistics. For this estimation procedure we considered the

50,000 simulations generating the shortest δ distance between the observed and simulated summary statistics.

For the model selection procedures performed in these studies, we used a modified version of the *calmod* function, written by M. A. Beaumont (available at http://www.rubic.rdg.ac.uk/~mab/stuff/) for the R statistical package.

3.2.2.2 Parameters Estimation

The purpose of the model selection procedure is to identify the best-fitting model, that is, the model that best explains the observed variation. After that, within the ABC framework, is it also possible to estimate the demographic and evolutionary parameters underlying this model. To do so, only a subset of simulations are retained (in general 2,000 or 5,000), i.e. the simulations producing statistics closest to the observed statistics, chosen from the total amount of simulations generated under the model. For this purpose, we implemented the approach developed by Beaumont et al. (2002) based on the computation of a local, weighted, linear regression between each parameter and the vector of the chosen summary statistics. Each retained simulation is assigned a weight (the commonly used weighting function is the Epanechnikov kernel) based on a function increasing as the distance between the observed and simulated data decreases. The regression slope is then used to adjust the parameters value from the retained simulations towards the value in correspondence of a distance zero between observed and simulated statistics. This way we obtained an estimate of the parameters' values that mimic a situation in which all simulations produce summary statistics equal to the observed values. Parameters need be transformed before the regression step (we use the logtan transformation, Hamilton, Stoneking & Excoffier 2005), to avoid adjustment outside the prior distribution.

The mode and the median value of the correspondent posterior distribution are usually used as parameter estimators; the 95% interval of the highest posterior density is also calculated, that is the interval which includes the 95% of the parameter values and within which the density is never lower than the density outside it.

For these purposes we used a modified version of the *make_pd2* script, written by M. A. Beaumont (available at http://www.rubic.rdg.ac.uk/~mab/stuff/) for the R statistical package.

3.2.2.3 Validation of the estimates

After an ABC analysis it is common to investigate the robustness of the results. To test the reliability of the model selection procedures, one can calculate the type I error, whereas to assess the quality of the parameters estimate one can calculate indices like the coefficient of determination (R^2), the bias and the root mean square error (RMSE), the coverage and the factor 2. Finally, to test whether the model we considered to best fit the observed data might actually generate patterns of genetic diversity resembling the observed ones, a posterior predictive test is commonly performed.

To assess if the models we simulated may be correctly recovered by the procedure we chose to calculate their posterior probability (that is: is there enough power in the data to allow one to distinguish the alternative models?), **Type I Error** (i.e. the probability of rejecting a true null hypothesis) is evaluated. To do this, some hundreds datasets are generated using each of the models considered in the model selection analysis; these pseudo-observed datasets are then treated as observed datasets in an ABC analysis using the previously simulated models. After that, the Type I error can be calculated as the proportion of cases in which the LR or the AR procedures were not able to recover the right model, as suggested in Fagundes et al. (2007) and Cornuet et al. (2008). If the Type I error is low, this means that the genetic data used in the analysis allow one to distinguish between the demographic models tested.

To determine whether the summary statistics we chose contain enough information to estimate model parameters, the **coefficient of determination** (R^2) can be computed. R^2 indicates the percentage of variance of the dependent variable (i.e., the parameter) explained by the predictors (i.e., the summary statistics). In the absence of an established threshold value, there is a general agreement that when $R^2 < 0.10$, the summary statistics do not convey enough information about their posterior distribution (Neuenschwander et al. 2008).

The accuracy of the median estimate of model parameters can be assessed by computing the relative **bias** and the relative **mean square error**. For these tests, *n* datasets are generated using median or mode point estimates as demographic parameters. Each of these *n* datasets is then used as a pseudo observed dataset which is analyzed with the previously described ABC methodology. Bias and RMSE depend, respectively, on the sum of differences, and on the sum of squared differences, between the *n* estimates of each parameter thus obtained, and the respective median point estimate (Neuenschwander et al. 2008). A value of 0 means that the median perfectly estimated the parameter, positive and negative values reflect, respectively, biases towards overestimation and underestimation.

To calculate the coverage and the factor 2, the same pseudo observed datasets are used. The **coverage** is defined as the proportion of times the known value (median or mode value) lies within the credible interval of the *n* estimates. For example, the 90% coverage is the proportion of instances in which the true value (i.e. the parameter value estimated during the ABC analysis) fall within the 90% credible interval of each of the *n* estimates derived from the pseudo observed dataset. The **factor 2** statistic, instead, represents the proportion of the *n* estimated median or mode values from the pseudo observed datasets lying between the 50% and the 200% of the fixed (known) value. Note that factor 2 gives information about the absolute precision of the estimator, because it is independent of the posterior distribution's variance (which, conversely, is not a property of the coverage).

Once a model has been shown to be better than any alternatives in generating data compatible with the observed one, the question is whether that model can actually generate data that faithfully reproduce the observed variation. This question can be addressed by performing a **posterior predictive test** (Gelman et al. 2004). To do this, thousands (*n*) datasets are generated under the selected model, by repeatedly drawing the parameter values from the posterior distributions estimated. These simulated data sets are summarized by summary statistics, which are then compared with the corresponding summary statistics from the observed data. This way one computes a posterior predictive P-value for each of the statistics considered, and then combines their probabilities into a global P-value, by a method that takes into account non-independence of the statistics (Voight et al. 2005). This global P-value is calculated in four steps: (1) each simulated summary statistic is compared

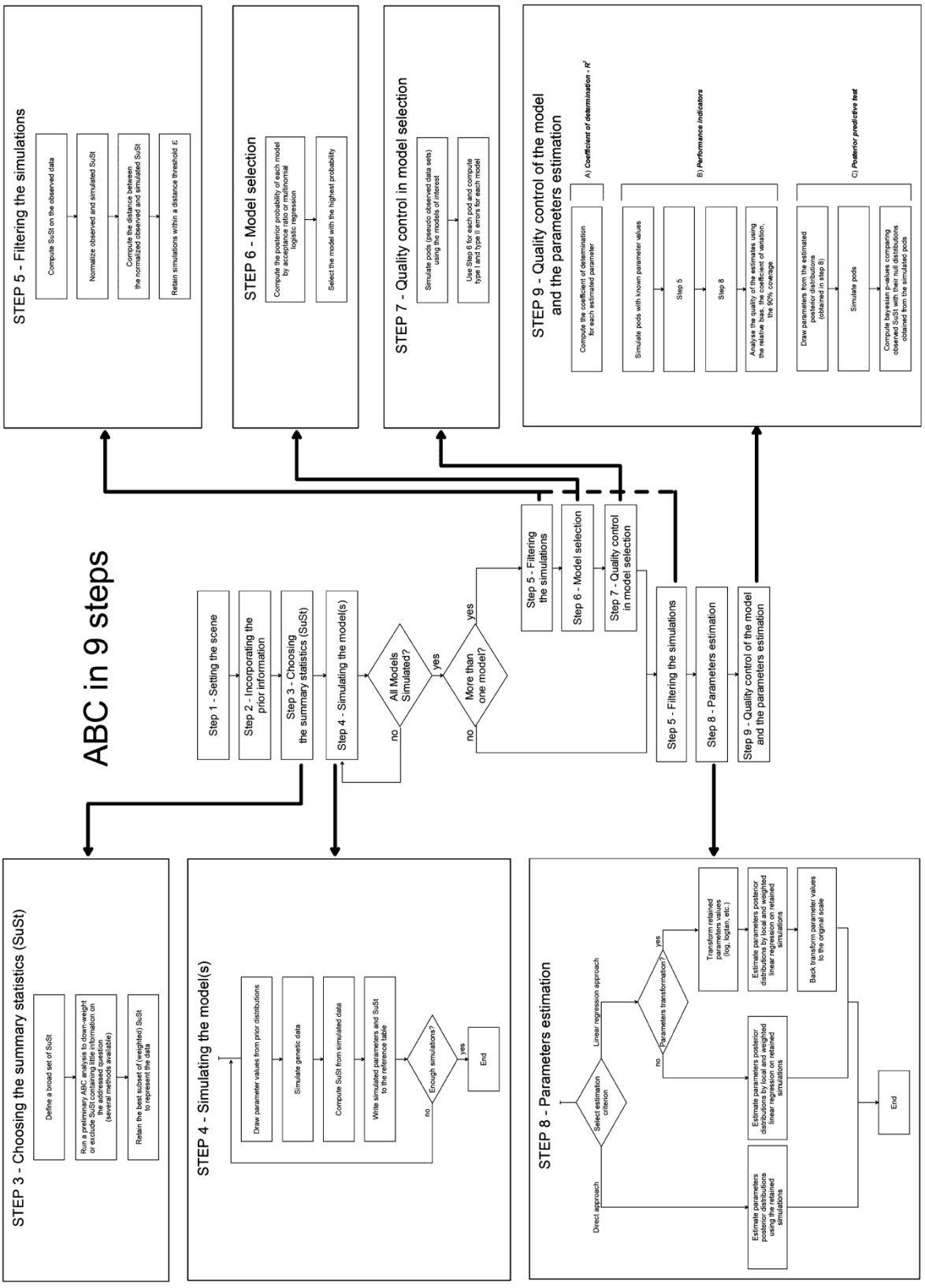
with the other *n*-1 values representing the empirical distribution of the statistic from the simulations, and thus associated with a two-tailed P-value; (2) for each simulated genealogy, a new statistic C, combining the P-values of the individual statistics (pi) is calculated as:

$$C = -2\sum \ln(pi)$$

where summation is over all P-values from each summary statistic. This step is repeated *n* times, so as to obtain a null distribution of C; (3) By repeating the same procedure over the observed statistics, we calculate an observed C value, Co; (4) by comparing Co with the C null distribution, we estimate a one tailed P-value (the Bayesian P-value) for Co.

A scheme of a complete ABC analysis is outlined in Fig 3.2.

Fig 3.2. From Bertorelle, Benazzo and Mona (2010).



4. Applications

In this chapter, I briefly present the three studies I co-authored during my PhD. The first is a study about the genealogical relationships between archaic humans (i.e. Neandertals), anatomically modern humans (i.e. Cro-Magnon) and modern Europeans (4.1); the second work regards the genealogical relationships between Bronze-age and modern population in Sardinia (4.2); the last study regards the origins and evolution of the Etruscan population (4.3). Within the ABC framework, applied here for the first time to datasets of ancient and modern human variation, we explicitly compared several models to choose the one which best accounts for the observed variation. Then we estimated the parameters of the best model and we evaluated the quality of these estimates.

A detailed description of these works is reported in the "Papers" section (7).

4.1 Neandertals, Early Modern humans and Modern Europeans

The debate on modern human origins regards the interpretation of a vast body of archaeological, fossil and genetic data, from which the relationships between ancient and modern populations and their migrational history can be approximately inferred. Many models have been proposed to account for the observed patterns of diversity and similarity, but, as a first approximation, it is fair to say the two main models are the "Out of Africa" model, and the Multiregional model (Fig 4.1). The distribution of fossils and artifacts clearly shows that, up to perhaps 2 million years ago, all human ancestors lived in Africa. Starting from that period, human forms are documented in Asia and Europe. At the end of the 1980s, the first studies of human molecular diversity suggested that our species had evolved from an African population that around 100 thousand years ago colonized the whole world, supplanting the former hominid. This replacement model is called "Out of Africa" (OOA) or "Recent African Origin" model (Fig 4.1A), and has been widely adopted by the human population genetics community. However, this model was disputed by some archaeologists for whom there is evidence of a regional continuity in the Pleistocene fossil record, which cannot be explained by a complete replacement of *Homo erectus* in Asia or Neandertal in Europe (Wolpoff 1989). They hence proposed a model of Multiregional evolution (MRE) (Fig 4.1B), where modern humans would have emerged gradually and simultaneously from

archaic forms in different continents. 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 (Wolpoff, Hawks & Caspari 2000). Further studies of worldwide modern human variation have discovered 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 (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); all these piece of evidence are in favor of the Out of Africa model. A related question is what extent of genetic exchange between archaic and modern humans is compatible with the OOA model. If the Multiregional model could only be rejected by proving that no exchange has been happened between them, the model would be impossible to falsify with scientific tools, and hence the debate would not be possible within the realm of science. The study and the comparison of DNA in ancient human forms (i.e. Neandertals), in anatomically modern humans (i.e. Cro-Magnon), and in modern populations, can be useful to address all these questions. In fact, Neandertal and Cro-Magnon coexisted in Europe for millennia, and fossil and archaeological data document a progressive withdrawal of Neandertal communities towards Western Europe as Cro-Magnoids expanded. The Neandertals anatomy and their artifacts disappeared from the record around 29,000 years ago (Mellars 1992; Mellars 2006). In analyses of mitochondrial DNA (mtDNA) Neandertal sequences fell out of the range of current European variation (Krings et al. 1997; Briggs et al. 2009), and even a small mitochondrial contribution of Neandertals to the modern human gene pool appeared unlikely (Currat & Excoffier 2004; Belle et al. 2009). However, in the first survey of the whole Neandertal nuclear genome, patterns of allele sharing with modern humans have been interpreted as suggesting 1-4% admixture between Neandertals and the ancestors of non-African people (Green et al. 2010). On the contrary, when a sample of mtDNA from Cro-Magnon was analyzed, it appeared indistinguishable from those of modern humans (Caramelli et al. 2003; Caramelli et al. 2008).

In this study, we explicitly compared models of modern human evolution using ancient and modern mtDNA sequences under the framework of Approximate Bayesian Computations (ABC, see Methods, section 3.2.2). The use of the demographic models allows

not only to compare modern and ancient variation highlighting the degree of resemblance in the sequences, but also to estimate the degree of confidence in considering Neandertals as the ancestor of modern Europeans and how much gene flow between them that can be compatible with the observed variation. To do this, we used all the ancient sequences available for Neandertals (7) and Cro-Magnon (3), and 150 modern European sequences coming from the same geographical area of the ancient samples. From the ABC analysis, the model having greater probability was the one in which the Neandertals underwent extinct around 29,000 years ago and belong to a separate genealogy respect to the Cro-Magnon and the modern Europeans. According to this model, anatomically modern humans emerged from a small population after a founder effect that followed the expansion out of Africa of the early humans. The Out of Africa model of human evolution appears to be hundreds-fold as likely as the alternative model. A direct comparison between a model without gene flow from Neandertals into Cro-Magnons and a model of gene flow during the period of the coexistence in Europe of Neandertals and Cro-magnons, showed that the best estimate of mitochondrial admixture between Neandertals and the ancestors of modern Europeans is zero. Additional tests on the reliability of the estimates confirmed the quality of the analysis, indicating that the data we analyzed contained enough information to allow one to distinguish among the models tested.

This study, albeit exploiting one of the most powerful statistical methodology of genetic inference, was limited to the mitochondrial DNA, and hence to the maternal lineage. In the recent nuclear genome survey, Neandertals appeared genetically closer to all non-Africans than to Africans. This observation was interpreted as evidence of admixture, between 1% and 4%, between Neandertals and the common ancestors of Asians and Europeans, in the Levant (Green et al. 2010). We propose a way to reconcile these findings, involving a more articulate model of genetic drift. Under such a model, the greater similarity between Neandertals and non-Africans would not necessarily require admixture between them. Indeed, if the common ancestors of Neandertals and modern humans were geographically structured, as proposed by Falush et al. (2003) and Harding & McVean (2004), all non-Africans could share with Neandertals a longer section of their genealogy, also sharing more alleles than Africans with Neandertals, including the derived alleles upon which Green et al. (2010) based their estimates. By contrast, in the mitochondrial DNA, having

lower effective population size compared with nuclear DNA, the sorting of the lineages due to genetic drift would be already complete. This view is also supported by data on the DNA of the human gastric parasite Helicobacter pilori, in which ancestral genetic clusters seem to have given rise to two distinct populations, one exclusively African, and the other cosmopolitan (Falush et al. 2003), and by the extreme levels of DNA variation still present in Africa (Schuster et al. 2010). The only additional assumption one has to make to account for the observed results is that the latter population was also ancestral to the European Neandertals typed by Green et al. (2010).

The complete published study is reported at p. 74.

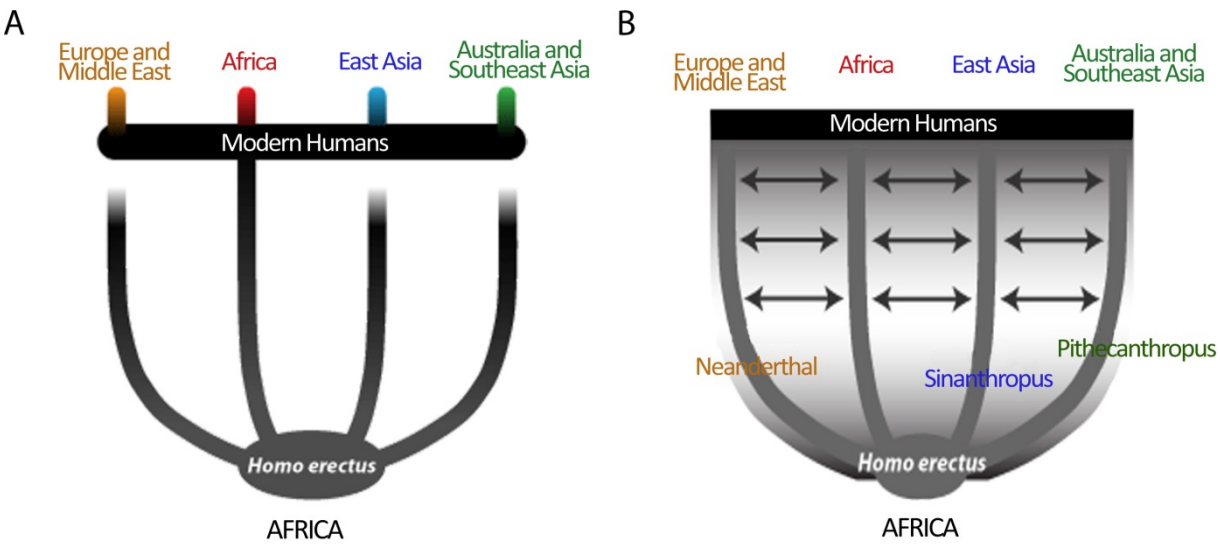


Fig 4.1. Out of Africa (A) and Multiregional (B) model of human evolution.

4.2 Modern and ancient mitochondrial variation in Sardinia

The population of Sardinia is known as one of the main genetic outliers in Europe (Cavalli-Sforza & Piazza 1993).When compared with populations from all over the world, Sardinians are clearly part of a European genetic cluster (Rosenberg et al. 2002), but they differ sharply from their European (Barbujani & Sokal 1990) and Italian (Barbujani & Sokal 1991; Barbujani et al. 1995) neighbors. Moreover, Sardinian populations show some (elsewhere rare) Y-chromosome and mitochondrial haplotypes at very high frequencies (Morelli et al. 2000; Semino et al. 2000; Quintana-Murci et al. 2003), and an unusual pattern of internal genetic diversity. Strong genetic differences are observed among Sardinian communities, both for allele frequencies (Barbujani & Sokal 1991) and polymorphism level (Fraumene et al. 2003). These peculiar features are probably due to the small effective population size combined with the reproductive isolation, caused by the fragmented habitat, that have probably enhanced the role of the genetic drift within the Sardinian communities. An ancient Sardinian sample was analyzed in a previous work (Caramelli et al. 2007), comprising 23 mitochondrial sequences from Bronze-Age Sardinia ("Nuragic" population). The authors observed very different resemblances with two modern populations of the island, separated in space by less than 120 km. One population came from Ogliastra, an isolated community in the middle-east of the island, and the other came from Gallura, an "open" region in the north-east of the island, where recent immigration is documented from mainland Italy. More than a half of the ancient haplotypes were present in the Ogliastra's sample, but only the 18% in Gallura, which is the same proportion one would observe by picking up random modern individuals from all over Europe (Caramelli et al. 2007). The existence of such sharp differences between one modern population and the ancient inhabitants of the island calls for an explanation, which lies in questions on the existence and on the strength of genealogical ties between ancient and modern people, and which can be empirically addressed by means of ABC.

In our study we defined three main models of evolution, tested both without and with migration from the mainland into Gallura (as historically documented), and differing mainly for the genealogical relationships between modern and ancient populations. In fact, in each of the three models, the ancient sample was placed respectively as ancestor of the

Ogliastra's population only, of Gallura's population only, or of both. The comparison between the observed mtDNA diversity and the patterns of variation simulated under these models clearly showed that haplotypes documented in the Bronze Age, or derived from them assuming a reasonable mutation rate, are still present and common in the isolated Ogliastra community. Conversely, the modern population of Gallura seems derived from ancestors who separated in Palaeolithic times (around 12,500 years ago) from the common ancestors of Bronze-Age and modern Ogliastra people, and have poor genealogical relationships, if any, with the ancient people of Sardinia. The only haplotype shared between Bronze-Age Sardinia and Gallura is the Cambridge Reference Sequence (CRS), which is very common all over Europe; however, the ABC analysis showed that there is no way of generating the genetic variation observed in Gallura starting from an ancient population with the same mtDNA diversity of Bronze-Age Sardinia. The most probable model estimated from the ABC analysis included also variable rates of gene flow from Latium, the mainland region nearest to Sardinia, into Gallura. Considering this migration rate we could also account for part of the excess of mtDNA variation found in Gallura with respect to Ogliastra.

This study cast new light on the nature and the extent of the genealogical ties between modern and ancient populations, a long-term source of controversy in evolutionary biology. In the case of Sardinia, we showed that, when properly analyzed, even a few tens ancient sequences can be sufficient to test hypotheses on the relationships between past and modern people and to improve the estimation of demographic and evolutionary parameters underlying their model of evolution.

The complete published study is reported at p. 97.

4.3 Origin and evolution of the Etruscans' DNA

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 language, lifestyle or customs (Barker & 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 & 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 & 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 mainly funerary or religious inscriptions. However, a language or a culture can rapidly get extinct, but that is certainly not the case for the DNA of its speakers; genetic evidence from Etruscans and other related populations may hence help one answer two questions, namely: what were the Etruscans' origins? And, what is their biological relationship with the modern inhabitants of Etruria?

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). Both Achilli et al. (2007) and Brisighelli et al. (2009) observed some affinities between Tuscans and modern Anatolian people; this 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. In 2004, for the first time, Vernesi and collaborators (2004) analyzed Etruscans' mtDNA obtained from 27 different individuals, highlighting genetic similarities between the Etruscans and the current population of Turkey, but not with Italian populations other than Tuscans (even if they shared only two haplotypes). However, further studies, considering also a Medieval Tuscan sample (Guimaraes et al. 2009), do not supported a direct genealogical continuity between the Etruscans and Tuscans (or Anatolian) populations (Belle et al. 2006; Guimaraes et al. 2009). The claim that systematic errors in the ancient DNA sequences led to flawed genealogical inference (Bandelt & Kivisild 2006; Achilli et al. 2007) is not supported by careful reanalysis of the Etruscan data (Mateiu & Rannala 2008).

Previous studies did not exploit the inferential power of the ABC methods, and 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. In this study we compared an 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. In two populations, Casentino and Volterra, we found evidence of genealogical continuity from Etruscans, through Medievals, to current times. By contrast, for Murlo and Florence, the ABC analysis highlighted as most probable the model in which the modern population occupies a distinct branch of the genealogical tree with respect to Etruscans and medieval Tuscans; for these populations this model was shown to be 7 to 99 times more likely than any alternative model. 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. To answer, we exploited the algorithm of the IM methods to estimate the most probable separation time between Anatolians (from Di

Benedetto et al. 2001) and 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 years ago. 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 years ago, with a 95% credible interval between 5,000 and 10,000. Thus, there might have been a genealogical link between modern Tuscans and what Herodotus considered the Etruscans' homeland, Anatolia. However, these results do not support an oriental origin for the Etruscans, because, even under the unrealistic assumption of complete reciprocal isolation between Tuscany and Anatolia, the likely separation of the two gene pools is dated long before the onset of the Etruscan culture. To date, no available genetic evidence suggests an Etruscan origin outside Italy, and traces of genealogical links with Etruscans are still recognizable in specific localities of Tuscany. This 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 framework of approximate Bayesian computation.

This study has been submitted to Molecular Biology and Evolution; the submitted manuscript is at p. 112.

5. Future Developments

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. Until recently, the ancient genetic information derived mainly from a single marker, the mitochondrial DNA (mtDNA), thus allowing one to study the fate of maternal lineages. Many advances in this field have been made in the last years and in 2010 the first three ancient hominid nuclear genomes were published (Green et al. 2010; Rasmussen et al. 2010; Reich et al. 2010). These results were achieved thanks to the technological developments in high-throughput sequencing, making it feasible to move from single genetic locus (such as mtDNA) to (almost) complete genome sequencing of ancient populations, and offering novel means of assessing authenticity of ancient DNA, even from modern humans. Moreover, extensive human genome data are becoming available, both from genome wide SNP data (Li et al. 2008; Reich et al. 2009; Xing et al. 2009; Hatin et al. 2011; Henn et al. 2011), and from the 1000 Genome Project and other human genome and exome studies (Schuster et al. 2010; The 1000 Genomes Project Consortium 2010). In this light, we will soon have large numbers of whole genome sequences from several modern and ancient populations. Combining this advance in the availability of whole genome sequence data and the statistical power provided by model-based methods such as ABC, in the near future it will be possible to clarify other long-standing evolutionary questions, and to highlight aspects of human history at an unprecedented resolution.

6. Bibliography

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 80:759-768.

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.

Atkinson QD, Gray RD, Drummond AJ. 2009. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. Proc Biol Sci 276:367-373.

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

Bamshad M, Wooding SP. 2003. Signatures of natural selection in the human genome. Nat Rev Genet 4:99-111.

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 99:10539-10544.

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

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

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

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

Bazin E, Dawson KJ, Beaumont MA. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics 185:587-602.

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 MA, Cornuet J-M, Marin JM, Robert CP. 2009. Adaptivity for ABC algorithms: the ABC-PMC scheme. Biometrika 96:983-990.

Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. Nat Rev Genet 5:251-261.

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

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

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A 98:4563-4568.

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.

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.

Bentley DRBalasubramanian SSwerdlow HP, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

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.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783-796.

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

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.

Brown RP, Yang Z. 2010. Bayesian dating of shallow phylogenies with a relaxed clock. Syst Biol 59:119-131.

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

Caramelli D, Milani L, Vai S, et al. 2008. A 28,000 years old Cro-Magnon mtDNA sequence differs from all potentially contaminating modern sequences. PLoS One 3:e2700.

Caramelli D, Vernesi C, Sanna S, et al. 2007. Genetic variation in prehistoric Sardinia. Hum Genet 122:327-336.

Cavalli-Sforza LL, Piazza A. 1993. Human genomic diversity in Europe: a summary of recent research and prospects for the future. Eur J Hum Genet 1:3-18.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 70:155-174.

Choi SC, Hey J. 2011. Joint inference of population assignment and demographic history. Genetics 189:561-577.

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

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.

Crow JF, Kimura M. 1970. An Introduction to Population Genetics Theory. New York: Harper and Row.

Currat M, Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol 2:e421.

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.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307-1320.

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

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185-1192.

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 U S A 104:17614-17619.

Falush D, Wirth T, Linz B, et al. 2003. Traces of human migrations in Helicobacter pylori populations. Science 299:1582-1585.

Felsenstein J. 1992. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet Res 60:209-220.

Fisher R. 1930. The Genetical Theory of Natural Selection: Clarendon Press.

Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA. 2003. Understanding missense mutations in the BRCA1 gene: an evolutionary approach. Proc Natl Acad Sci U S A 100:1151-1156.

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180:977-993.

Fraumene C, Petretto E, Angius A, Pirastu M. 2003. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. Hum Genet 114:1-10.

Friedberg EC, Walker GC, Siede W. 1995. DNA Repair and Mutagenesis. Washington, DC: ASM Press.

Fu YX, Li WH. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. Mol Biol Evol 14:195-199.

Gelman A, Carlin J, Stern H, Rubin D. 2004. Bayesian Data Analysis. Boca Raton, Florida: CRC Press.

Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. In Computing Science and Statistics. Proceedings of the 23rd Symposium on the Interface. Interface Foundation of North America (Fairfax Station, VA): pp 156-163.

Gilks WR, Richardson S, Spiegelhalter DJ. 1996. Markov Chain Monte Carlo in Practice: Chapman & Hall/CRC.

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

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

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

Grimsley C, Mather KA, Ober C. 1998. HLA-H: a pseudogene with increased variation due to balancing selection at neighboring loci. Mol Biol Evol 15:1581-1588.

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

Guimaraes S, Ghirotto S, Benazzo A, et al. 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 U S A 102:7476-7480.

Handt O, Krings M, Ward RH, Paabo S. 1996. The retrieval of ancient human DNA sequences. Am J Hum Genet 59:368-376.

Handt O, Richards M, Trommsdorff M, et al. 1994. Molecular genetic analyses of the Tyrolean Ice Man. Science 264:1775-1778.

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

Hardy GH. 1908. Mendelian Proportions in a Mixed Population. Science 28:49-50.

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

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.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Hatin WI, Nur-Shafawati AR, Zahri MK, Xu S, Jin L, Tan SG, Rizman-Idid M, Zilfalil BA. 2011. Population genetic structure of peninsular Malaysia Malay sub-ethnic groups. PLoS One 6:e18312.

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.

Henn BM, Gignoux CR, Jobin M, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A 108:5154-5162.

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.

Hey J. 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol 3:e193.

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.

Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S. 2001. Ancient DNA. Nat Rev Genet 2:353-359.

Hoss M, Jaruga P, Zastawny TH, Dizdaroglu M, Paabo S. 1996. DNA damage and DNA sequence retrieval from ancient tissues. Nucleic Acids Res 24:1304-1307.

Hudson RR. 1990. Genetic Data Analysis. Methods for Discrete Population Genetic Data. Science 250:575.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. Genetics 120:831-840.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589.

Huelsenbeck JP. 1995. The performance of phylogenetic methods in simulation. Systematic Biology 44:17-48.

Hughes JM, Peters CJ, Cohen ML, Mahy BW. 1993. Hantavirus pulmonary syndrome: an emerging infectious disease. Science 262:850-851.

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

Joyce P, Marjoram P. 2008. Approximately sufficient statistics and bayesian computation. Stat Appl Genet Mol Biol 7:Article26.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. Mammalian protein metabolism. New York: Academic Press. p. 21-123.

Kass RE, Raftery AE. 1995. Bayes Factor.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893-903.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120.

Kimura M, Crow JF. 1964. The Number of Alleles That Can Be Maintained in a Finite Population. Genetics 49:725-738.

Kimura M, Weiss GH. 1964. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics 49:561-576.

Kingman JFC. 1982. The coalescent. Stochastic Processes and Their Applications. 13:235-248.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19-30.

Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768-770.

Kuhner MK, Yamato J, Felsenstein J. 2000. Maximum likelihood estimation of recombination rates from population data. Genetics 156:1393-1401.

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

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

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709-715.

Marjoram P, Molitor J, Plagnol V, Tavare S. 2003. Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci U S A 100:15324-15328.

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.

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

Mellars PA. 1992. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. Philos Trans R Soc Lond B Biol Sci 337:225-234.

Metropolis N, Rosenbluth, A.W.,Rosenbluth, M.N., Teller, A.H. and Teller, E. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1092.

Morelli L, Grosso MG, Vona G, Varesi L, Torroni A, Francalacci P. 2000. Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. Hum Biol 72:585-595.

Nei M. 1987. Molecular Evolutionary Genetics. New York, NY, USA: Columbia University Press.

Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L. 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.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885-896.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Paabo S. 1989. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci U S A 86:1939-1943.

Paabo S. 1990. Amplifying ancient DNA. PCR-Protocols and Applications. San Diego: Academic. p. 159-166.

Paabo S, Irwin DM, Wilson AC. 1990. DNA damage promotes jumping between templates during enzymatic amplification. J Biol Chem 265:4718-4721.

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

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

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

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.

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

Provine WB. 1971. The Origins of Theoretical Population Genetics. Chicago: University of Chicago Press.

Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES. 2003. Genetic structure of Mediterranean populations revealed by Y-chromosome haplotype analysis. Am J Phys Anthropol 121:157-171.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645-1656.

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

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

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

Riebler A, Held L, Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics 178:1817-1829.

Rodrigo AG, Felsenstein J. 1999. Coalescent approaches to HIV population genetics. Molecular evolution of HIV. Baltimore, Md: Johns Hopkins University Press. p. 233-272

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science 298:2381-2385.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Schierup MH, Vekemans X, Charlesworth D. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet Res 76:51-62.

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

Semino O, Passarino G, Oefner PJ, et al. 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155-1159.

Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, Paabo S. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol 2:E57.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555-562.

Slatkin M, Voelm L. 1991. FST in a hierarchical island model. Genetics 127:627-629.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437-460.

Tajima F. 1989. The effect of change in population size on DNA polymorphism. Genetics 123:597-601.

Tavaré S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences (American Mathematical Society) 17:57-86.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. Genetics 145:505-518.

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

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

Turchi C, Buscemi L, Previdere C, Grignani P, Brandstatter A, Achilli A, Parson W, Tagliabracci A. 2008. Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int J Legal Med 122:199-204.

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

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 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.

Wakeley J. 2009. Coalescent Theory: An Introduction. Greenwood Village, Colo.: Roberts & Co. Publishers.

Wall JD, Kim SK. 2007. Inconsistencies in Neanderthal genomic DNA sequences. PLoS Genet 3:1862-1866.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256-276.

Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182:1207-1218.

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. Genetics 149:1539-1546.

Wolpoff M. 1989. Multiregional evolution: the fossil alternative to Eden. In: P Mellars, C Stringer, 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, Caspari R. 2000. Multiregional, not multiple origins. Am J Phys Anthropol 112:129-136.

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.

Wright S. 1938. Size of a population and breeding structure in relation to evolution. Science 87:430-431.

Wright S. 1943. Isolation by Distance. Genetics 28:114-138.

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.

7. Papers

AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 146:242–252 (2011)

No Evidence of Neandertal Admixture in the Mitochondrial Genomes of Early European Modern Humans and Contemporary Europeans

Silvia Ghirotto, Francesca Tassi, Andrea Benazzo, and Guido Barbujani*

Department of Biology and Evolution, University of Ferrara, Italy

KEY WORDS ancient DNA; population genetics; Approximate Bayesian Computations; coalescent simulations; admixture

ABSTRACT Neandertals, the archaic human form documented in Eurasia until 29,000 years ago, share no mitochondrial haplotype with modern Europeans. Whether this means that the two groups were reproductively isolated is controversial, and indeed nuclear data have been interpreted as suggesting that they admixed. We explored the range of demographic parameters that may have generated the observed mitochondrial diversity, simulating 3.0 million genealogies under six models differing as for the relationships among contemporary Europeans, Neandertals, and Upper Palaeolithic European early modern humans (EEMH), who coexisted with Neandertals for millennia. We compared by Approximate Bayesian Computations the simulation results with mitochondrial diversity in 7 Neandertals, 3 EEMH, and 150 opportunely chosen modern Europeans. A model of genealogical continuity between EEMH and contemporary Europeans, with no Neandertal contribution, received overwhelming support from the analyses. The maximum degree of Neandertal admixture, under the model of gene flow supported by nuclear data, was estimated at 1.5%, but this model proved 20–32 times less likely than a model without any gene flow. Nuclear and mitochondrial evidence might be reconciled if smaller population sizes led to faster lineage sorting for mitochondrial DNA, and Neandertals shared a longer period of common ancestry with the non-African's than with the African's ancestors. Am J Phys Anthropol 146:242–252, 2011.

Two anatomically different human forms, the archaic Neandertals and the European early modern humans of the Upper Palaeolithic (hereafter EEMH, sometimes referred to as Cro-Magnoids), coexisted in Europe for millennia. Fossil and archaeological data document a progressive withdrawal of Neandertal communities as EEMH expanded; the Neandertal anatomy and their artifacts disappeared from the record at a moment in time which is traditionally placed around 29,000 years ago (Mellars, 1992; Mellars, 2006).

The biological relationships between these human forms are controversial. For many years, the debate focused on the relative merits of two classes of models, Regional Continuity, and Replacement. According to the former, anatomically archaic hominids of the Old World formed a subdivided population, within which a transition from archaic to modern morphology occurred; under the Replacement (or "Out of Africa") model, anatomically modern humans expanded from Africa replacing all archaic groups. More recently, Assimilation models emerged, i.e., the idea that contemporary populations are largely descended from an anatomically modern group expanding from Africa, but Neandertals contributed to the modern European gene pool to a non-negligible extent (Relethford, 2001; Trinkaus, 2007). Thus, the main controversial point became how much genetic exchange, if any, there has been between the two human forms.

Complete replacement is impossible to demonstrate, because the same genetic consequences are expected both without admixture and with extremely low levels of admixture. However, most morphological and genetic evidence seems to agree with the predictions of a model in which anatomically-modern people and archaic humans did not hybridize. With very few possible exceptions (see e.g., Zilhão, 2006), European fossils are clearly classified as either archaic or modern; the absence of intermediate morphologies suggests that levels of admixture were extremely low or nil (Tattersall and Schwartz, 1999). Neandertal mitochondrial DNA (mtDNA) sequences fall out of the range of current European variation (Krings et al., 1997; Briggs et al., 2009), so that even a small mitochondrial contribution of Neandertals to the modern human gene pool appeared unlikely (Currat and Excoffier, 2004; Belle et al., 2009). By contrast, in the first survey of the whole Neandertal nuclear genome, patterns of allele sharing with modern humans have been interpreted as suggesting gene flow from Neandertals into the ancestors of modern non-Africans, before the Eurasian populations separated (Green et al., 2010).

No clear consensus has yet emerged from studies of modern DNA diversity either; compare e.g., Labuda

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Italian Ministry of Universities (MIUR); Grant numbers: PRIN 2008.

*Correspondence to: Guido Barbujani, Department of Biology and Evolution, University of Ferrara, Via L. Borsari 46, 44121 Ferrara, Italy. E-mail: g.barbujani@unife.it

Received 23 March 2011; accepted 10 May 2011

DOI 10.1002/ajpa.21569
Published online 24 August 2011 in Wiley Online Library (wileyonlinelibrary.com).

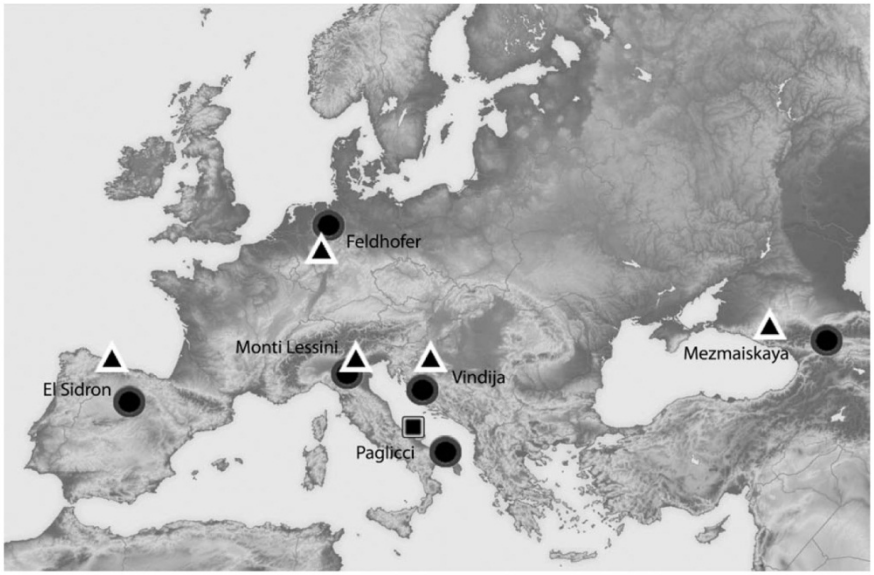


Fig. 1. Geographic localization of the samples considered. Triangles: Neandertals; Squares: Early European Modern Humans, or EEMH; Circles: Modern Europeans.

et al. (2000), Plagnol and Wall (2006) and Templeton (2007), with Excoffier (2002), Hodgson and Disotell (2008) and Jakobsson et al. (2008). In most cases, the evidence suggesting some degree of genetic exchange between archaic and modern human forms is a deep basal node in the gene tree, reflecting high levels of haplotype divergence. Such deep splits, accompanied by high linkage disequilibrium, are expected if there was admixture between groups that have long evolved in isolation (Templeton, 2005), but may also reflect genetic structuring of the common ancestors of modern humans and Neandertals (see e.g., Currat et al., 2006). Accordingly, many regard the current genetic data as insufficient to discriminate between models.

Better data may take long to accumulate, but more refined biostatistical approaches are already available. In this study, we compared for the first time ancient and modern mtDNA sequences under the framework of Approximate Bayesian Computations (ABC). The available mitochondrial data allow one to test hypotheses for which fossil evidence is inconclusive, and which cannot be currently tested at the nuclear level. We could thus evaluate posterior probabilities for a set of models differing as for the genealogical relationships of two ancient (Neandertal and EEMH) and six modern Europeans populations. We also estimated the demographic parameters of the best-fitting model, and demonstrated that the data have enough statistical power to identify the correct model. These results restrict the range of hypotheses potentially accounting for the genetic relationships between Neandertal and modern people, and show how analyses of nuclear and mitochondrial diversity can be reconciled without invoking admixture processes.

MATERIALS AND METHODS

The data

We investigated diversity in the mitochondrial hypervariable region I, spanning 360 bp, in 150 modern and 10 ancient individuals. The letter are seven Neandertal individuals: two from Feldhofer in Germany (Krings et al., 1997; Schmitz et al., 2002), two from the Vindija Cave in Croatia (Krings et al., 2000), one from Mezmaiskaya in the Russian Caucasus (Ovchinnikov et al., 2000), one from Monti Lessini in Italy (Caramelli et al., 2006), and one from El Sidrón cave, Asturias, Spain (Lalueza-Fox et al., 2006), plus 3 EEMH sequences from the Paglicci cave, Italy (Caramelli et al., 2003; Caramelli et al., 2008). No other sequences of the entire mitochondrial hypervariable region I are available for European Neandertals or EEMH (Hodgson and Disotell, 2008).

To have similar effects of geography for ancient and modern populations, we chose modern samples from Germany (Richards et al., 1996), Croatia (Babalini et al., 2005), the Caucasus (Nasidze and Stoneking, 2001), two regions of Italy (Babalini et al., 2005) and Spain (Corte-Real et al., 1996) (see Fig. 1). Sample sizes were very different; to avoid any resulting confounding effect (such as those due to an excessive weight of the largest samples upon estimates of haplotype sharing), we randomly resampled 25 sequences from each modern population. In a preliminary step, we made sure that the summary statistics calculated on the resampled datasets are consistent with those calculated on the complete datasets.

Serial coalescent simulations

Three million mitochondrial genealogies were generated by the serial coalescent algorithm implemented in the Bayesian version of SERIALSIMCOAL (Anderson et al., 2005). With this program one can generate multiple gene genealogies according to any demographic model. Suppose that one has samples of sizes n_0, n_1, $n_2 \ldots n_k$ from populations studied t_0, t_1, $t_2 \ldots t_k$ generations ago. The program generates genealogies proceeding backwards in time, starting with n_0 samples in the present (t_0) and adding n_1, $n_2 \ldots n_k$ samples at the appropriate moments in the past. The genealogies were extended

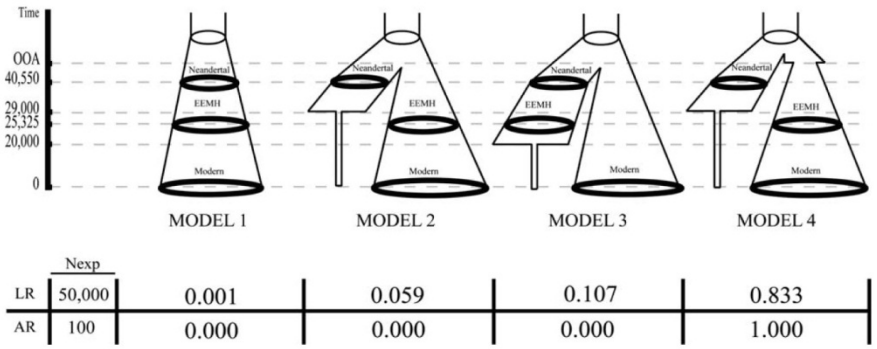


Fig. 2. Schematic presentation of the six models tested. Numbers on the Y-axis refer to years from the present. OOA (Out of Africa) is the time of the dispersal from Africa. The posterior probability is given under each model, according to two criteria: AR, or acceptance-rejection, and LR, or weighted multinomial logistic regression. *Nexp* is the number of experiments considered.

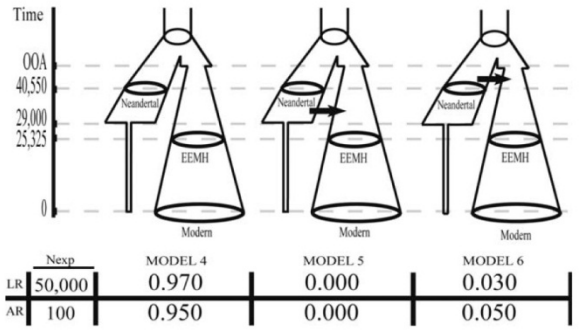


Fig. 3. Comparison between Models 4, 5 and 6. The arrows in Models 5 and 6 represent gene flow from the Neandertal to the EEMH population (Model 5) and gene flow from Neandertal to the ancestors of modern European and Asian populations (Model 6).

backwards until, through a series of coalescence events, they reached their most recent common ancestor, or MRCA. Mutations were then randomly distributed onto the tree, under a finite-site model with two potential allelic states for each site, a transition bias = 0.9375 and a rate-heterogeneity parameter = 0.26 (see Belle et al., 2009 and reference therein).

Demographic models and priors

The six models are outlined in Figures 2 and 3. Model 1 assumes genealogical continuity between Neandertal, EEMH, and modern samples; under Model 2, the Neandertal lineage separates from the lineage leading to EEMH and modern Europeans; under Model 3 the EEMH population is descended from Neandertal ancestors, whereas the modern populations are part of another lineage. Model 4 resembles Model 2, but the lineage that gave rise to EEMH and modern Europeans undergoes a founder effect associated with the dispersal from Africa. Models 5 and 6 add gene flow from the Neandertals, either (Model 5) during the maximum span of the possible coexistence of Neandertals and EEMH, 42,000 to 30,000 years ago, or (Model 6) starting 80,000 years ago, as suggested by Green et al. (2010). In all simulations, the modern samples were placed at generation 0. The EEMH and Neandertal samples were at generations 1,013 and 1,622, corresponding to the average age of the respective specimens, 25,325 and 40,550 years, assuming a generation time = 25 years (Currat and Excoffier, 2004; Fenner, 2005; Noonan et al., 2006;Fagundes et al., 2007). All population sizes increased exponentially through time, at constant rate, starting >1,622 generations ago.

Under Model 2 the Neandertal lineage got extinct 1,160 generations (29,000 years) ago (Mellars, 1992); other authors (Walker et al., 2008; Zilhão et al., 2010) proposed that this has happened earlier, around 37,000 years ago in the Iberian peninsula, but a well-dated specimen shows that Neandertals were still present, at least in the Caucasus, 29,000 years ago (Ovchinnikov et al., 2000). At any rate, choosing a late date of Neandertal disappearance increases the time interval through which there might have been contact with EEMH, thus favoring the admixture model. Under Model 3, in which EEMH and Neandertals were in the same lineage, both became extinct at an arbitrary time, 20,000 years ago. In both cases, the date of extinction has no effect on the results of the tests, in so far as it is more recent than the age of the youngest specimen.

Under Models 4–6, we added to Model 2 a founder effect in the Cro-Magnoid and Modern lineage, bringing the population size to 500 at a moment between 50,000 and 80,000 years ago (Liu et al., 2006; Fagundes et al., 2007). A complete description of the prior information considered is in Supporting Information Tables 1 and 2.

Summary statistics

Internal genetic diversity was summarized by the number of different haplotypes, the average pairwise sequence difference, and the haplotype diversity, calculated with Arlequin ver. 3.11 (Excoffier et al., 2005). We compared pairs of samples in two ways: (a) by estimating F_{ST} (Hudson et al., 1992); (b) by classifying the segregating sites into four categories, namely (1, 2) those that are polymorphic in one population and monomorphic in the other (i.e., exclusive sites for population 1 or 2); (3) polymorphic sites shared between populations 1 and 2 (shared differences); (4) fixed differences between populations 1 and 2 (Wakeley and Hey, 1997; Leman et al., 2005; Becquet and Przeworski, 2007). In this way, we

TABLE 1. Observed summary statistics describing genetic variation within and between samples: Neandertal, NE; Early European modern humans, EEMH; Modern Europeans, ME

	NE	EEMH	ME
Number of sequences	7	3	150
Number of distinct haplotypes	6	2	115
Segregating site	17	1	92
Mean pairwise difference	5.238	0.667	4.549
Haplotype diversity	0.952	0.667	0.990

Fst Hudson

NE/EEMH 0.8600	EEMH/NE 0.0966	ME/NE 0.7839

Class of Segregating Sites

SHARED SITES			FIXED SITES		
ME_EEMH	ME_NE	EEMH_NE	ME_EEMH	ME_NE	EEMH_NE
1	9	0	0	7	15

EXCLUSIVE SITES

ME_EEMH	EEMH_NE	ME_NE	NE_ME	EEMH_NE	NE_EEMH
91	0	83	8	1	17

obtained 12 statistics (four counts of segregating sites, times three pairwise comparisons) (Table 1).

These statistics were chosen from a larger set, in order to capture the whole information contained in the data. Indeed, the larger the number of summary statistics, the larger the statistical noise included in the posterior estimation (known as "curse of dimensionality") (Joyce and Marjoram, 2008). To this aim, we assessed by principal component analysis the correlation between each descriptor of genetic diversity and the genetic diversity generated by our simulations, to choose only those statistics having a substantial impact on the inference.

Approximate Bayesian Computations

All the following procedures were developed in the R environment (R Development Core Team, 2008) using scripts available at http://www.rubic.rdg.ac.uk/~mab/stuff/. The ABC procedure included three main steps (Beaumont et al., 2002). First, for each model, 500,000 genealogies were simulated (for a total of 3,000,000 experiments), considering as random variables the demographic and evolutionary parameters of the model. Therefore, for every simulation experiment, these values were chosen at random from the corresponding prior distributions. Next, we summarized the genetic variation of the samples calculating the same set of statistics in the observed data and in each simulated dataset. Finally, we calculated for each experiment a Euclidean distance between observed and simulated statistics, thus ordering the experiments according to their distance from the observed dataset. The choice of the best model and the parameter estimation were based on the subset of simulation experiments producing the shortest Euclidean distances.

Model selection

We compared the posterior probabilities of the models in two ways, using the calmod function, also available at http://www.rubic.rdg.ac.uk/~mab/stuff/. The first criterion is a simple acceptance-rejection procedure (AR) (Pritchard et al., 1999). For each model, we initially counted the number of simulations (n_i) which were found among the N simulations with the shortest Euclidean distance. The posterior probability for the i-th model was simply n_i/N. This method is considered reliable only when based on a small set of simulations showing an excellent fit with the observed data (Beaumont, 2008); in this case, we chose to retain the 100 simulations producing statistics closest to the observed statistics. We also resorted to a second method, estimating posterior probabilities by weighted multinomial logistic regression (LR) (Beaumont, 2008). Under this procedure, each model represents a categorical dependent variable Yi (where i is again the identity number of the model), and the summary statistics are the predictive variables. The probability of the model is evaluated at the point corresponding to the observed vector of summary statistics. For this calculation we retained the 50,000 simulation experiments associated with the shortest Euclidean distances.

Parameter estimation

Within models, the parameters of the 1,000 experiments showing the shortest Euclidean distances between simulated and observed statistics were logtan transformed (Hamilton et al., 2005); we then calculated a weighted local regression, using summary statistics as predictors to adjust the parameter values towards the values expected in correspondence of the observed summary statistics (Beaumont et al., 2002). We thus obtained the posterior distributions of four classes of parameters, namely effective population sizes, separation times, mutation rates and migration rates.

Posterior predictive test and quality of the estimates

Next, we tested whether the model we chose could indeed generate patterns of genetic diversity resembling the observed ones. For that posterior predictive test (Gelman et al., 2004) we simulated 10,000 datasets according to the model with the highest probability, using the estimated posterior parameter distribution. We then estimated 9 additional descriptors of genetic diversity, namely the number of segregating sites within each population and the haplotype sharing between samples, which had not been considered during the inferential step, and compared them with the observed ones. If the model is realistic and our posterior distributions esti-

TABLE 2. Demographic parameters estimated under Model 4: prior distribution (U: uniform in all cases), median, mode, lower (0.05) and upper (0.95) limits of the 90% credible interval and coefficient of determination (R^2)

Model 4	Prior distribution	Median	Mode	0.05	0.95	R^2
Time MRCA	*	16,433	16,079	9,586	25,086	0.73
Ne Modern	{U: 1,000,000–10,000,000}	4,173,069	1,027,003	1,000,000	7,945,095	0.02
Ne Neandertal (before extinction)	{U: 5,000–100,000}	32,263	15,676	4,375	90,080	0.29
Time Out of Africa	{U: 50,000–80,000}	2,725	3,037	2,106	3,200	0.07
Separation time	{U: 80,025–900,000}	11,785	11,031	3,877	25,848	0.62
Mutation Rate	{U: 0.0002–0.008}	0.0008	0.0007	0.0006	0.0011	0.90
Ancestral Ne EEMH	{U: 5–5,000}	2,071	5	5	4,025	0.03
Ancestral Ne Neandertal	{U: 5–5,000}	766	5	5	4,105	0.10

The time to the most recent common ancestor, Time MRCA was estimated from the simulated data and not extracted from a prior distribution.

mates are plausible, summary statistics inferred from the simulated and observed datasets should not significantly differ. A posterior predictive *P*-value for each summary statistics was thus estimated, and these probabilities were combined into a global *P*-value (see Ghirotto et al., 2010), taking into account nonindependence of the individual statistics (Voight et al., 2005).

Finally, we asked whether models could actually be discriminated on the basis of the available data. To answer, we simulated 1,000 datasets from the prior distribution of each model and we treated them as observed datasets in an ABC analysis using previously simulated models. For each dataset we calculated Type I error as the number of experiments in which the simulated model was not recognized by the model selection procedure (AR), or Type I Error.

To assess the reliability of the inferred parameters we calculated the coefficient of determination (R^2, indicating the percentage of the parameter's variance explained by the summary statistics we used), the relative bias and the relative root mean square error (RMSE), that allow to evaluate biases toward overestimation and underestimation of parameters point estimate, and the 95% coverage for each estimated parameter (see Neuschwander et al., 2008 and, for an application, Ghirotto et al., 2010).

RESULTS

Model selection

Table 1 reports the statistics summarizing mtDNA sequence variation in 7 Neandertal (NE), 3 EEMH and 150 Modern Europeans (ME). Of the demographic models initially considered (see Fig. 2) Model 4 showed by far the highest ability to reproduce the observed variation, reaching probabilities >83% under the LR approach and = 100% under the AR approach. This means that almost all simulated datasets showing good agreement with the observed data had been generated assuming independent genealogies for the NE versus the EEMH and ME mtDNAs. Moreover, none of the best-fitting experiments was simulated assuming genetic continuity between NE and ME or EEMH, since Model 1 has $P = 0$ under both methods used for comparison.

We then added the possibility of migration from the NE into the EEMH between 42,000 and 30,000 years ago (Model 5), or during the early stage of the dispersal from Africa (Model 6), as suggested by Green et al. (2010). In the comparison of Models 4, 5, and 6, under both estimation procedures, Model 4 reached probabilities between 95% and 97%, appearing 20–32 times as likely as Model 6 (see Fig. 3).

Parameter estimation

Table 2 shows the posterior parameter estimates for Model 4. The age of the Most Recent Common Ancestor (MRCA), the separation time between the NE and EEMH-ME lineages and the mutation rate were well estimated, as shown by the high value of their R^2, respectively 73%, 62%, and 90% (posterior distributions in Fig. 4). The likely age of the Most Recent Common Ancestor (TMRCA) is around 411,000 years (median value), in agreement with previous estimates (Krings et al., 1999; Ovchinnikov et al., 2000; Briggs et al., 2009), but lower than the 660,000 years ago inferred from the survey of the entire mitochondrial genome (Green et al., 2008). The separation time between populations is estimated at about 295,000 years ago, close to the value inferred by Noonan et al. (2006) from 65,000 nuclear bp, and within the range estimated considering the whole genome (270,000–440,000) (Green et al., 2010). The mutation rate (0.0008 per generation for the 360-bp region, hence about 0.1 mutational events per million year per nucleotide) is lower than recently estimated (Henn et al., 2009; Soares et al., 2009) but appears reliable, considering that across our long evolutionary times multiple mutational events on the same site may occur, reducing the apparent mutation rate.

Our median estimate for the time of dispersal from Africa is about 69,000 years ago, and its 90% credible interval is between 52,650 and 80,000 years ago. These values suggest that the dispersal might have occurred earlier than inferred from studies of modern DNA diversity, i.e. 51,000 (Fagundes et al., 2007) or 56,000 (Liu et al., 2006) years ago, although the R^2 value is admittedly low (7%). We could not substantially improve the prior estimates of other parameters, including the effective sizes of the ME and EEMH populations, both associated with broad posterior distributions and low R^2. Changing the prior distribution from uniform to log-uniform had basically no effect upon the estimate and on its accuracy (Supporting Information Table 3).

Even though Models 5 and 6 appeared to be highly unlikely, we also estimated their parameters and found they have largely overlapping distributions. The gene flow rate from Neandertals to EEMH, under Model 5, has a 90% upper bound of the posterior density at 0.0057, with a mode = 0.0008 and a median = 0.001 (Supporting Information Table 4a). This means on average 3 events of migration per generation, considering the NE effective female population size estimated during the period of coexistence with the EEMH. Simulating migration right after early modern human dispersal

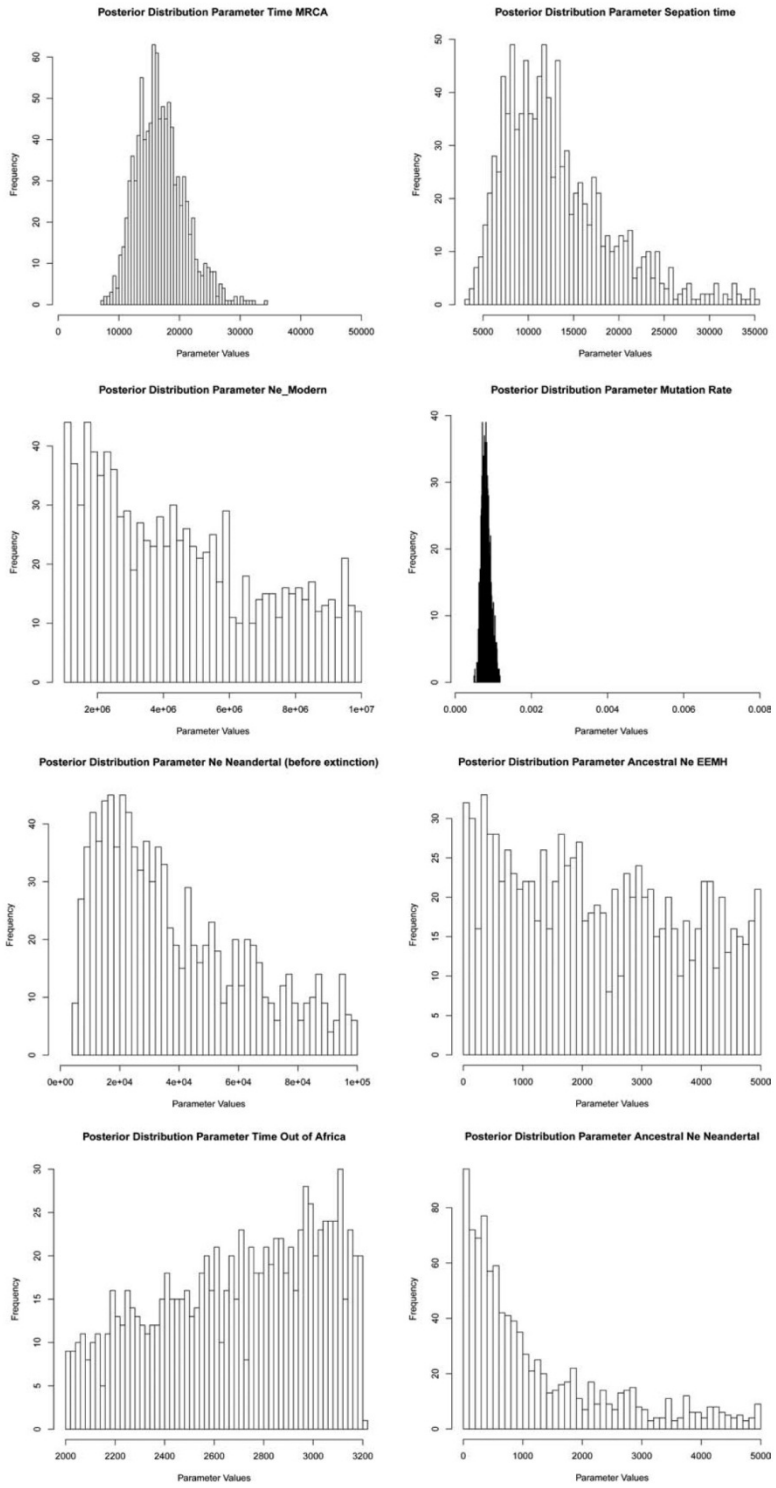


Fig. 4. Posterior distributions of parameters under Model 4, based on the best 1,000 simulations. The X axis covers the range of the (uniform) prior distributions.

TABLE 3. Power of the AR procedure to recover the true model, estimated as the proportion of cases in which data generated by the models listed on the Y-axis were attributed to the models listed on the X-axis

		% of model attribution				Type I error
		MOD1	MOD2	MOD3	MOD4	
SIMULATED MODEL	MOD1	0.992	0.005	0.003	0	0.008
	MOD2	0.031	0.947	0.009	0.012	0.052
	MOD3	0.015	0.022	0.962	0.001	0.038
	MOD4	0.002	0.031	0.002	0.963	0.035

Type I error , in the last column, is the fraction of cases in which the true model was not identified.

from Africa as suggested by Green et al. (2010) (Model 6, which proved at least 20 times less likely than a model without gene flow) the upper bound of the 90% posterior density distribution reaches 0.015, but the modal estimate is 0 (Supporting Information Table 4b). In other words, gene flow between NE and EEMH appears extremely unlikely (see Fig. 3), and if it was not zero it was nearly so. This parameter, with its narrow posterior distribution and good R^2 (45%, Supporting Information Tables 4a and 4b), seems well estimated.

Posterior predictive test and quality assessment of the estimates

Additional tests support the reliability of our estimates. For Model 4, the *P*-values representing the discrepancy between the observed data and the datasets generated drawing parameter values from the estimated posterior distributions were insignificant for all statistics (Supporting Information Table 5), and the global *P*-value was 0.478. Therefore, this model, besides having a probability close to 100% with respect to alternative models, is also capable to generate patterns of diversity fully consistent with all observed statistics.

We then asked whether the method used for the model selection (AR) is powerful enough to identify the model under which the data were generated (Table 3). Comparing Models 1 to 4, all the datasets were correctly identified, with probabilities of recovery from 94% to 99% and hence a Type I Error never >5.2%. In other words, had mtDNA diversity evolved according to one of the other simulated models, the present analysis would have shown it. As could be expected, slightly less power was shown in the comparison between Model 5 and Model 6 (Type I Error: 38%, Supporting Information Table 6). This is due to the fact that these models are very similar, differing just for the time of the gene flow events from Neandertal into EEMH, and hence generate similar patterns in the data.

Finally, we ran several tests to assess the quality of the parameters estimated under Model 4 (Supporting Information Table 7). Both relative Bias and RMSE for the modes and medians of each parameter are generally low. Only the modes of the two ancestral Ne have high values of both Bias and RMSE, suggesting, as said before, that these parameters could not be well estimated. Most of the parameters show high values of the 95% coverage, indicating that their posterior distributions are in general well estimated (Supporting Information Table 6).

DISCUSSION

Nordborg (1998) first remarked that the nonoverlap between Neandertal and modern mtDNA variation does not imply that there was no admixture, because, at the low Palaeolithic population sizes, drift could have eliminated rare, and even not-so-rare, haplotypes. The question, then, became how rare a haplotype should be, and how small the population, to produce the observed absence of Neandertal haplotypes in modern subjects, despite admixture having actually occurred.

Currat and Excoffier (2004) demonstrated by simulation that the absence of Neandertal mtDNAs in the modern gene pool is compatible with a maximum interbreeding rate = 0.1%, which translates into 1 admixture event every 100 years, during the coexistence of the two human forms in Europe. Belle et al. (2009) incorporated EEMH sequences in their analyses, but still failed to find evidence for any appreciable degree of Neandertal admixture in the European mtDNA pool. For methodological reasons, in both studies mutation rates and population sizes had to be fixed at the start of the simulation. Conversely, the ABC methods we employed in this study allowed us to explore for each model a broad and continuous range of population sizes, mutation rates and, when applicable, separation times and gene flow rates. In this way, the models were compared in a statistically rigorous way, and their final performance is independent of any specific value of the simulation parameters. We found that the best estimate by far of mitochondrial admixture between Neandertals and the ancestors of modern Europeans is zero. Even at very low population sizes and with high mutation rates, the patterns of diversity observed in ancient and modern samples appear incompatible with a Neandertal contribution to the mitochondrial genealogy of EEMH and modern Europeans.

There is reason to believe that the estimates we obtained can be trusted. The shapes of the posterior probability distributions, the posterior predictive tests, and several statistics estimated from the simulated data strongly suggest that the information available was sufficient to discriminate among models, and that most parameters are well estimated. The main area of uncertainty concerns the modern population sizes, which appear extremely large and distributed across the whole range of the prior distribution. This finding is not unusual in studies of this kind (Fagundes et al., 2007; Belle et al., 2009; Wegmann et al., 2009; Laval et al., 2010), and does not seem to reflect the choice of priors. Rather, it is probably a consequence of a simplistic, yet unavoidable, assumption, namely that populations evolved in isolation. In reality, people with different mitochondrial features must have migrated for millennia from other regions. This process resulted in an increase of genetic diversity, which the model accommodated by inflating the population size estimates. However, it is hard to imagine that the Neandertal contribution to the modern gene pool would be more likely in a (much more complicated) model also considering successive gene flow from multiple modern sources. On the other hand, such a complicated model would require the estimation of a very large number of parameters (i.e., migration rates between all possible pairs of populations), resulting in a

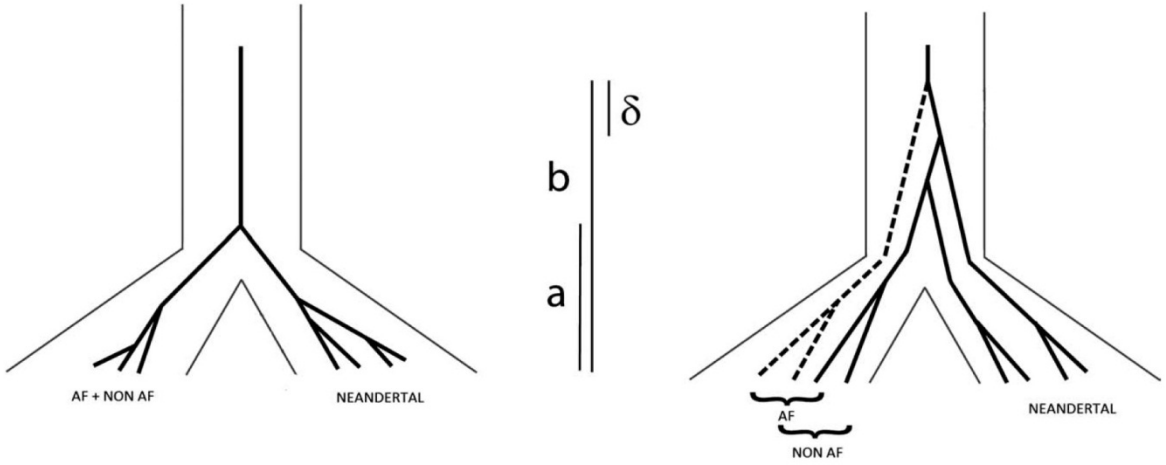


Fig. 5. Schematic view of the gene genealogies for markers transmitted by one (left) and two (right) parents, in Neandertals and modern humans. a: time since the mitochondrial most recent common ancestors; b: time since the autosomal most recent common ancestor; δ: excess evolutionary time shared with Neandertals only by the lineage leading to non-African modern people; AF, NON-AF: African and Non-African modern populations.

loss of accuracy in the estimation of the parameters that really matter, i.e., admixture rates between Neandertals and anatomically modern people.

In the recent genome survey, Neandertals appeared genetically closer to non-Africans than to Africans. This observation was interpreted as evidence of admixture between Neandertals and the common ancestors of Asians and Europeans, in the Levant, resulting in a Neandertal contribution to the modern genomes estimated between 1% and 4%. Alternative explanations are possible, but were considered less likely (Green et al., 2010). However, the poor performance of Model 6 in this study shows that the hypothesis of early admixture in the Levant has some problems too. Unless we have made serious errors in the interpretation of mitochondrial data, the model favored by the analysis of nuclear diversity seems to account very poorly, if at all, for the observed patterns of mitochondrial diversity in archaic and contemporary populations of Europe.

Only one complete Neandertal genome has been studied so far, and, given the rigid standards established to guarantee the quality of the data, sample size is not going to increase any time soon. A second problem is that the admixture model between Neandertal and anatomically modern populations proposed by Green et al. (2010) implies that the ancestors of all modern humans who left Africa had contacts with Neandertals, including those from Papua New Guinea. On the contrary, it is possible that ancestral modern humans also dispersed from Africa via a Southern route, through the Arab peninsula, the Indian subcontinent and Melanesia. This hypothesis was proposed to account for temporal and spatial patterns of cranial diversity (Lahr and Foley, 1994), has been supported by analyses of mtDNA variation (Quintana-Murci et al., 1999; Maca-Meyer et al., 2001; Macaulay et al., 2005) and, recently, by the analysis of >100,000 nuclear single-nucleotide polymorphisms (Ghirotto et al., 2011). If some modern populations of Southern Asia and Papua New Guinea are descended from people who left Africa without crossing Palestine, we see no way that their ancestors could have met, and hybridized with, Neandertals. Therefore, their genetic affinities with Neandertals must have a different origin.

It is thus necessary to find another explanation for the discrepancy between the apparent implications of the mitochondrial and nuclear analyses. In principle, two possibilities, neither simple to support empirically, would be sex-biased gene flow and hybrid selection. The former means that maybe Neandertal males, but not females, admixed with early anatomically modern Europeans. This is in contrast with studies of sex-biased admixture in modern communities, suggesting that the invading population tends to incorporate females more than males (Abe-Sandes et al., 2004; Goncalves et al., 2008; Gonzalez-Andrade et al., 2007; Stefflova et al., 2009; Quintana-Murci et al., 2010); to what extent this might also apply to prehistoric populations, nobody knows. Hybrid selection could account for the observed differences between admixture estimates if Neandertal mtDNAs had lower fitness in combination with a hybrid nuclear genome. Once again, we see no way to test empirically whether that was actually the case.

Moving on to testable hypotheses, a simple process of genetic drift after admixture is not the explanation we seek (see Fig. 3). In addition, in a simple admixture model, alleles passed from a resident to an invading population are expected to often surf to high frequencies if the invading populations also undergoes demographic growth (Currat et al., 2008). Because the incoming EEMH doubtless increased in numbers, even small Neandertal contributions should be detectable in the gene pool of their descendants, which is not the case for the European mtDNAs (Currat and Excoffier, 2004; this study).

To reconcile findings based on nuclear and mitochondrial variation we thus need a more articulate model, of which genetic drift is only a component. Many studies of modern DNA data have suggested that the common ancestors of Neandertals and modern humans might have been geographically structured (Falush et al., 2003; Harding and McVean, 2004; Lahr and Foley, 1994). A few simple calculations show that this possibility, also mentioned by Green et al. (2010), should be taken seriously (see Fig. 5). The expected time since the MRCA is $2N$ generations for mtDNA, where N is the female population size; if the sex ratio among Neandertals was 1 female : 1 male, the age of the nuclear DNA MRCA

should be 4 times as large. Briggs et al. (2009) quantified the size of the Neandertal female population around 3,500 or less. This means that the mitochondrial and nuclear MRCAs of Neandertals can be placed respectively 7,000 generations (or 175,000 years) and 28,000 generations (or 700,000 years) ago. These figures come with a large standard error, but imply that if the lineages leading to Neandertals and modern humans separated between 175,000 and 700,000 years ago, one would expect exactly what has been observed, namely independent mtDNA genealogies, and a certain degree of allele sharing at the autosomal level (see Fig. 5). On the basis of cranial measurements, anatomically archaic and modern humans separated between 311,000 and 435,000 years ago, with an upper limit of 592,000 (Weaver et al. 2008, and references therein). In this paper, we estimated that the same event occurred about 295,000 years ago (median value), with an upper 95% limit of 646,200 years. Therefore, the Replacement model with structured ancestral population is in reasonable agreement with fossil, nuclear DNA and mtDNA evidence, whereas the model of admixture fails to account for the observed relationships between ancient and modern mtDNAs.

Under a model in which the ancestral population was structured, the greater nuclear similarity between Neandertals and non-Africans would not necessarily require admixture between them. Indeed, if the non-Africans shared with Neandertals a longer section of their genealogy (represented by the interval labeled as δ in Fig. 5), they would also share more alleles than Africans and Neandertals, including the derived alleles upon which Green et al. (2010) based their estimates. This view is also supported by data on the DNA of the human gastric parasite *Helicobacter pilori*, in which ancestral genetic clusters seem to have given rise to two distinct populations, one exclusively African and the other cosmopolitan (Falush et al., 2003), and by the extreme levels of DNA variation still present in Africa (Schuster et al., 2010; Henn et al., 2011). The only additional assumption one has to make to account for the observed results is that the latter population was also ancestral to the European Neandertals typed by Green et al. (2010). Therefore, the hypothesis of genetic drift in a structured ancestral population, in which Neandertals shared a longer period of common ancestry with the non-African's than with the African's ancestors, seems to reconcile most findings about DNA diversity in Neandertal and modern people. This hypothesis predicts that the nuclear alleles preferentially shared by Neandertals and non-African will have MRCAs falling in the upper part of the genealogy (δ interval in Fig. 5), and we are preparing to test this hypothesis.

ACKNOWLEDGMENTS

The authors thank Giorgio Bertorelle and Arnaud Estoup for their useful suggestions.

LITERATURE CITED

Abe-Sandes K, Silva WA Jr, Zago MA. 2004. Heterogeneity of the Y chromosome in Afro-Brazilian populations. Hum Biol 76:77–86.

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.

Babalini C, Martinez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, Tarsi T, Contini I, Barac L, Janicijevic B, Martinovic Klaric I, et al. 2005. The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. Eur J Hum Genet 13:902–912.

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 MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res 17:1505–1519.

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.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, et al. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318–321.

Caramelli D, Lalueza-Fox C, Condemi S, Longo L, Milani L, Manfredini A, de Saint Pierre M, Adoni F, Lari M, Giunti P, et al. 2006. A highly divergent mtDNA sequence in a Neandertal individual from Italy. Curr Biol 16:R630–R632.

Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, Chiarelli B, Dupanloup I, Bertranpetit J, Barbujani G, et al. 2003. Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci USA 100:6593–6597.

Caramelli D, Milani L, Vai S, Modi A, Pecchioli E, Girardi M, Pilli E, Lari M, Lippi B, Ronchitelli A, et al. 2008. A 28,000 years old Cro-Magnon mtDNA sequence differs from all potentially contaminating modern sequences. PloS one 3:e2700.

Corte-Real HB, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, Sykes BC. 1996. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60:331–350.

Currat M, Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol 2:e421.

Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S. 2006. Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science 313:172; author reply 172.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. Evolution 62:1908–1920.

Excoffier L. 2002. Human demographic history: refining the recent African origin model. Curr Opin Genet Dev 12:675–682.

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

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.

Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, et al. 2003. Traces of human migrations in *Helicobacter pylori* populations. Science 299:1582–1585.

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.

Gelman A, Carlin J, Stern H, Rubin D. 2004. Bayesian data analysis. Boca Raton, Florida: 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, Penso-Dolfin L, Barbujani G. 2011. Genomic evidence for an African expansion of anatomically-modern humans by a southern route. Hum Biol 00:00–00.
Goncalves VF, Carvalho CM, Bortolini MC, Bydlowski SP, Pena SD. 2008. The phylogeography of African Brazilians. Hum Hered 65:23–32.
Gonzalez-Andrade F, Sanchez D, Gonzalez-Solorzano J, Gascon S, Martinez-Jarreta B. 2007. Sex-specific genetic admixture of Mestizos. Amerindian Kichwas, and Afro-Ecuadorans from Ecuador. Hum Biol 79:51–77.
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.
Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416–426.
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.
Harding RM, McVean G. 2004. A structured ancestral population for the evolution of modern humans. Curr Opin Genet Dev 14:667–674.
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.
Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodríguez-Botigué L, Ramachandran S, Hon L, Brisbin A, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA 108:5154–5162.
Hodgson JA, Disotell TR. 2008. No evidence of a Neanderthal contribution to modern human diversity. Genome Biol 9:206.
Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.
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.
Joyce P, Marjoram P. 2008. Approximately sufficient statistics and Bayesian computation. Stat Appl Genet Mol Biol 7:Article26.
Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, von Haeseler A, Grossschmidt K, Possnert G, Paunovic M, Paabo S. 2000. A view of Neandertal genetic diversity. Nat Genet 26:144–146.
Krings M, Geisert H, Schmitz RW, Krainitzki H, Paabo S. 1999. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. Proc Natl Acad Sci USA 96:5581–5585.
Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30.
Labuda D, Zietkiewicz E, Yotova V. 2000. Archaic lineages in the history of modern humans. Genetics 156:799–808.
Lahr MM, Foley RA. 1994. Multiple dispersals and modern human origins. Evol Anthropol 3:48–60.
Lalueza-Fox C, Krause J, Caramelli D, Catalano G, Milani L, Sampietro ML, Calafell F, Martinez-Maza C, Bastir M, Garcia-Tabernero A, et al. 2006. Mitochondrial DNA of an Iberian Neandertal suggests a population affinity with other European Neandertals. Curr Biol 16:R629–R630.
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.
Leman SC, Chen Y, Stajich JE, Noor MA, Uyenoyama MK. 2005. Likelihoods from summary statistics: recent divergence between species. Genetics 171:1419–1436.
Liu H, Prugnolle F, Manica A, Balloux F. 2006. A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79:230–237.
Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM. 2001. Major genomic mitochondrial lineages delineate early human expansions. BMC Genet 2:13.
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.
Mellars P. 2006. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proc Natl Acad Sci USA 103:9381–9386.
Mellars PA. 1992. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. Philos Trans R Soc Lond B Biol Sci 337:225–234.
Nasidze I, Stoneking M. 2001. Mitochondrial DNA variation and language replacements in the Caucasus. Proc Biol Sci 268: 1197–1206.
Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L. 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.
Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Paabo S, Pritchard JK, et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113–1118.
Nordborg M. 1998. On the probability of Neanderthal ancestry. Am J Hum Genet 63:1237–1240.
Ovchinnikov IV, Gotherstrom A, Romanova GP, Kharitonov VM, Liden K, Goodwin W. 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature 404:490–493.
Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. PLoS Genet 2:e105.
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.
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.
Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, van Helden PD, Hoal EG, Behar DM. 2010. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. Am J Hum Genet 86:611–620.
R Development CoreTeam. 2008. R: A Language and Environment for Statistical Computing. Available at:http://www.R-project.org. Vienna, Austria: Foundation for Statistical Computing.
Relethford JH. 2001. Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. Am J Phys Anthropol 115:95–98.
Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, Sykes B. 1996. Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59:185–203.
Schmitz RW, Serre D, Bonani G, Feine S, Hillgruber F, Krainitzki H, Paabo S, Smith FH. 2002. The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley. Germany. Proc Natl Acad Sci USA 99:13342–13347.
Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947.
Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740–759.
Stefflova K, Dulik MC, Pai AA, Walker AH, Zeigler-Johnson CM, Gueye SM, Schurr TG, Rebbeck TR. 2009. Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. PloS one 4:e7842.

Tattersall I, Schwartz JH. 1999. Hominids and hybrids: the place of Neanderthals in human evolution. Proc Natl Acad Sci USA 96:7117–7119.

Templeton AR. 2005. Haplotype trees and modern human origins. Am J Phys Anthropol Suppl 41:33–59.

Templeton AR. 2007. Genetics and recent human evolution. Evolution 61:1507–1519.

Trinkaus E. 2007. European early modern humans and the fate of the Neandertals. Proc Natl Acad Sci USA 18:7367–7372.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102:18508–18513.

Wakeley J, Hey J. 1997. Estimating ancestral population parameters. Genetics 145:847–855.

Walker MJ, Gibert J, Lòpez MV, Vincent Lombardi A, Pérez A, Zapata J, Ortega J, Higham T, Pike A, Schwenninger JL, et al. 2008. Late Neandertals in Southeastern Iberia: Sima de las Palomas del Cabezo Gordo. Murcia, Spain. Proc Natl Acad Sci USA 105:20631–20636.

Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182:1207–1218.

Zilhão J. 2006. Neandertals and moderns mixed, and it matters. Evol Anthropol 15:183–195.

Zilhão J, Davis SJM, Duarte C, Soares AMM, Steier P, Wild E. 2010. Pego do Diabo (Loures. Portugal): dating the emergence of anatomical modernity in westernmost Eurasia. PLoS ONE 5:e8880.

Supplementary Material

Supplemental Table 1. Information upon which the prior distributions are based. For missing parameters (i.e. ancestral Ne and time of the onset of expansion) we chose a wide, uniform, prior distribution.

MRCA (most recent common ancestor)		
Years ago	**Note**	**References**
550,000 - 690,000	based on mtDNA	(Krings et al., 1997)
465,000 (CI: 317,000 - 741,000)	based on mtDNA	(Krings et al., 1999)
500,000	based on mtDNA	(Paabo, 1999)
365,000 - 853,000	based on mtDNA	(Ovchinnikov et al., 2000)
516,000 (CI: 465,000 - 569,000)	DNA	(Green et al., 2006)
461,000 - 825,000	mtDNA	(Green et al., 2006)
706,000 (CI: 468,000 - 1,015,000)	DNA	(Noonan et al., 2006)
560,000 (CI: 509,000 - 615,000)	based on Green et al. (2006)	(Wall and Kim, 2007)
706,000 (CI: 466,000 -1,028,000)	based on Noonan et al. (2006)	(Wall and Kim, 2007)
244,200 ± 2,200	based on mtDNA	(Belle et al., 2009)
511,200 (CI: 388,900 - 641,300)	based on mtDNA	(Briggs et al., 2009)
439,000 (CI: 321,900 - 553,800)	based on mtDNA	(Briggs et al., 2009)
441,000 - 684,000	based on nuclear genome	(Green et al., 2010)
Separation time		
Years ago	**Note**	**References**
250,000 - 400,000	archaeological evidence	(Foley and Lahr, 1997; Rightmire, 2001)
106,000 - 246,000	based on mtDNA	(Ovchinnikov et al., 2000)
440,000 (CI: 170,000 - 620,000)	based on the European data	(Noonan et al., 2006)
325,000 (CI: 135,000 - 557,000)	based on Noonan et al. (2006)	(Wall & Kim, 2007)
150,000	based on mtDNA	(Belle et al., 2009)
182,000 - 592,000	based on cranial measurements	(Weaver et al., 2008)
270,000 - 440,000	based on nuclear genome	(Green et al., 2010)
Neandertal Persistence in Europe:		
Years ago		**References**
29,000 years ago		(Mellars, 1992)
Admixture Rate Neandertal_EEMH		
Lower than		**References**
25%		(Nordborg, 2001)
45%		(Cooper et al., 2004)
0,1%		(Currat and Excoffier, 2004)
25%		(Serre et al., 2004)
0%		(Noonan et al., 2006)
5%		(Blum and Rosenberg, 2007)

0%		(Wakeley and Hey, 1997)
0,001%		(Belle et al., 2009)
Neandertal Population Size		
	Note	**References**
250,000		(Biraben, 2003)
5,000 - 9,000	female population	(Lalueza-Fox et al., 2005)
3,000	female population	(Green et al., 2006)
3,500	female population	(Briggs et al., 2009)
Modern Population Size		
	Geographic Area	**References**
4,079,702	Apulia, Italy	http://demo.istat.it/pop2009/index.html
4,333,979	Emilia Romagna, Italy	http://demo.istat.it/pop2009/index.html
4,435,056	Croatia	http://www.eustat.es/idioma_i/indice.html
3,758,878	Southern Spain	http://www.ine.es/censo/es/consulta.jsp
463,000	Cherkessian	http://www.kcr.narod.ru
14,926,820	Northern Germany	http://www.destatis.de
Dating		
Years ago	**Sample**	**References**
Neandertal		
43,000	ElSidron1252	(Lalueza-Fox et al., 2006)
40,000	Feldhofer1	(Schmitz et al., 2002)
40,000	Feldhofer2	(Schmitz et al., 2002)
29,000	Mezmaiskaya	(Ovchinnikov et al., 2000)
50,000	Monti Lessini	(Caramelli et al., 2006)
42,000	Vindija75	(Krings et al., 2000)
38,000	Vindija80	(Serre et al., 2004)
EEMH		
25,000	Paglicci 12	(Caramelli et al., 2003)
28,000	Paglicci 23	(Caramelli et al., 2008)
23,000	Paglicci 25	(Caramelli et al., 2003)

Supplemental Table 2. Prior distribution of all the parameters for each model

Model 1	**Prior distribution[a]**
Ne Modern	{U: 1,000,000 - 10,000,000}
Time at the start of exponential growth	{U: 40,575 - 900,000}
Mutation Rate	{U: 0.0002 - 0.008}
Ancestral Ne	{U: 5 - 5,000}

Model 2	**Prior distribution[a]**
Ne Modern	{U: 1,000,000 - 10,000,000}
Ne Neandertal (before extinction)	{U: 5,000 - 100,000}
Separation Time	{U: 40,575 - 900,000}
Mutation Rate	{U: 0.0002 - 0.008}
Ancestral Ne EEMH	{U: 5 - 5,000}
Ancestral Ne Neandertal	{U: 5 - 5,000}

Model 3	**Prior distribution[a]**
Ne Modern	{U: 1,000,000 - 10,000,000}
Ne EEMH (before extinction)	{U: 5,000 - 100,000}
Separation time	{U: 40,575 - 900,000}
Mutation Rate	{U: 0.0002 - 0.008}
Ancestral Ne Modern	{U: 5 - 5,000}
Ancestral Ne Neandertal	{U: 5 - 5,000}

Model 4	**Prior distribution[a]**
Ne Modern	{U: 1,000,000 - 10,000,000}
Ne Neandertal (before extinction)	{U: 5,000 - 100,000}
Time Out of Africa	{U: 50,000 - 80,000}
Separation Time	{U: 80,025 - 900,000}
Mutation Rate	{U: 0.0002 - 0.008}
Ancestral Ne EEMH	{U: 5 - 5,000}
Ancestral Ne Neandertal	{U: 5 - 5,000}
Ne before Out Of Africa	{U: 5,000 - 100,000}

Model 5	**Prior distribution[a]**
Ne Modern	{U: 1,000,000 - 10,000,000}
Ne Neandertal (before extinction)	{U: 5,000 - 100,000}
Time Out of Africa	{U: 50,000 - 80,000}
Separation time	{U: 80,025 - 900,000}

Mutation Rate	{U: 0.0002 - 0.008}
Ancestral Ne EEMH	{U: 5 - 5,000}
Ancestral Ne Neandertal	{U: 5 - 5,000}
Ne before Out Of Africa	{U: 5,000 - 100,000}
Gene flow rate from Neandertal to EEMH	{U: 0.000 - 0.02}

Model 6	**Prior distribution[a]**
Ne Modern	{U: 1,000,000 – 10,000,000}
Ne Neandertal (before extinction)	{U: 5,000 - 100,000}
Time Out of Africa	{U: 50,025 – 80,000}
Separation time	{U: 80,025 - 900,000}
Mutation Rate	{U: 0.0002 - 0.008}
Ancestral Ne EEMH	{U: 5 - 5,000}
Ancestral Ne Neandertal	{U: 5 - 5,000}
Ne before Out Of Africa	{U: 5,000 - 100,000}
Gene flow rate from Neandertal to EEMH	{U: 0.000 - 0.02}

Time is expressed in years
Ne= Effective female population size
a: U= Uniform probability, in the range between the two values

Supplemental Table 3. Comparison of Ne estimates obtained under Model 4 using different distributions of priors

Prior distribution	Median	Mode	0.05	0.95	R^2
Uniform {U: 1,000,000 - 10,000,000}	4,173,069	1,027,003	1,000,000	7,945,095	0.02
Log uniform {U: 1,000,000 - 10,000,000}	2,358,466	1,000,000	1,000,000	8,069,306	0.08

Supplemental Table 4a. Demographic parameters estimated under Model 5: prior distribution, median, mode, 90% of the posterior distributions credible interval and coefficient of determination .

Model 5	Prior distribution	Median	Mode	0.05	0.95	R^2
Time MRCA	*	18,858	19,034	9,684	30,329	0.41
Ne Modern	{U: 1,000,000 - 10,000,000}	5,493,974	1,000,000	1,000,000	3,264,626	0.02
Ne Neandertal (before extinction)	{U: 5,000 - 100,000}	30,269	16,246	5,235	82,059	0.16
Time Out of Africa	{U: 50,000 – 80,000}	2,601	2,609	2,000	2,314	0.03
Separation time	{U: 80,025 - 900,000}	13,618	10,171	4,047	32,238	0.25
Mutation Rate	{U: 0.0002 - 0.008}	0.0005	0.0002	0.0002	0.0015	0.82
Ancestral Ne EEMH	{U: 5 - 5,000}	2,610	3,725	5	4,678	0.01
Ancestral Ne Neandertal	{U: 5 - 5,000}	979	274	5	4,213	0.09
Gene flow rate from Neandertal to EEMH	{U: 0.000 - 0.01}	0.0011	0.0008	0.0000	0.0057	0.26

Time is expressed in years, and time of MRCA is calculated at any simulation by the program

[a] U = uniform probability, in the range between the two values

[b] Upper and lower limits of the 90% credible interval

[c] Coefficient of determination

[d] Effective female population size

Supplemental Table 4b. Demographic parameters estimated under Model 6: prior distribution, median, mode, 90% of the posterior distributions credible interval and coefficient of determination.

Model 6	Prior distribution	Median	Mode	0.05	0.95	R^2
Time MRCA	*	13,755	12,825	8,116	21,636	0.54
Ne Modern	{U: 1,000,000 – 10,000,000}	4,620,743	1,000,000	1,000,000	9,211,521	0.02
Ne Neandertal (before extinction)	{U: 5,000 - 100,000}	35,822	20,506	5,955	92,080	0.19
Time Out of Africa	{U: 50,025 - 80,000}	2,087	2,000	2,000	2,757	0.38
Separation time	{U: 80,025 - 900,000}	7,855	6,498	3,201	20,547	0.26
Mutation Rate	{U: 0.0002 - 0.008}	0.0010	0.0010	0.0006	0.0013	0.85
Ancestral Ne EEMH	{U: 5 - 5,000}	1,987	5	5	3,835	0.01
Ancestral Ne Neandertal	{U: 5 - 5,000}	1,173	224	5	4,309	0.08
Gene flow rate from Neandertal to EEMH	{U: 0.000 - 0.01}	0.0018	0.0000	0.0000	0.0156	0.13

Time is expressed in years, and time of MRCA is calculated at any simulation by the program

[a] U = uniform probability, in the range between the two values

[b] Upper and lower limits of the 90% credible interval

[c] Coefficient of determination

[d] Effective female population size

Supplemental Table 5. Posterior predictive test for Model 4.

	P_value
Global for Model 4	0.478
Seg_site_Mod	0.130
Seg_Site_EEMH	0.122
Seg_Site_N	0.262
AS_Mod_EEMH(Mod)	0.186
AS_Mod_EEMH(EEMH)	0.271
HS_EEMH_N(EEMH)	0.483
HS_EEMH_N(N)	0.484
HS_Mod_N(Mod)	0.472
HS_Mod_N(N)	0.474

Mod: Modern
EEMH: Early European modern humans
Nea: Neandertal
Seg_site: Number of segregating Sites
AS: Allele sharing
HS: Fraction of haplotypes in common

Supplemental Table 6. Power of the AR procedure to recover the true model, comparing Models 4, 5 and 6.

		% of model attribution [a]			
		MOD4	**MOD5**	**MOD6**	**Type I error**
SIMULATED MODEL	**MOD4**	0.93	0.02	0.05	0.07
	MOD5	0.03	0.89	0.08	0.11
	MOD6	0.1	0.37	0.53	0.38

[a] These figures represent the proportion of simulations generated under the Models on the Y-axis, which were assigned to the models on the X-axis by the AR procedure. Shaded cells represent the proportion of correct assignments.

Supplemental Table 7. Accuracy of the estimated parameters Model 4: Bias, RMSE, 95%Coverage.

MEDIAN	**Bias**	**RMSE**	**95% Coverage**
Ne Modern	0.296	0.370	1
Ne Neandertal	0.788	1.026	1
Time Out Of Africa	-0.060	0.071	1
Separation Time	0.014	0.353	0.983
Mutation Rate	0.067	0.138	0.972
Ancestral Ne Modern	0.150	0.287	1
Ancestral Ne Neandertal	0.220	0.386	1

MODE	**Bias**	**RMSE**	**95% Coverage**
Ne Modern	0.900	1.716	1
Ne Neandertal	-0.021	0.319	1
Time Out Of Africa	-0.023	0.028	1
Separation Time	-0.370	0.398	0.981
Mutation Rate	-0.052	0.121	0.989
Ancestral Ne Modern	160.933	338.820	0
Ancestral Ne Neandertal	49.543	64.098	0

Supplementary References

Belle EM, Benazzo A, Ghirotto S, Colonna V, and Barbujani G. 2009. Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations. Heredity 102(3):218-225.

Biraben J. 2003. L'évolution du nombre des hommes. Popul & Sociétés 394:1-4.

Blum MG, and Rosenberg NA. 2007. Estimating the number of ancestral lineages using a maximum-likelihood method based on rejection sampling. Genetics 176(3):1741-1757.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z et al. . 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science (New York, NY 325(5938):318-321.

Caramelli D, Lalueza-Fox C, Condemi S, Longo L, Milani L, Manfredini A, de Saint Pierre M, Adoni F, Lari M, Giunti P et al. . 2006. A highly divergent mtDNA sequence in a Neandertal individual from Italy. Curr Biol 16(16):R630-632.

Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, Chiarelli B, Dupanloup I, Bertranpetit J, Barbujani G et al. . 2003. Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proceedings of the National Academy of Sciences of the United States of America 100(11):6593-6597.

Caramelli D, Milani L, Vai S, Modi A, Pecchioli E, Girardi M, Pilli E, Lari M, Lippi B, Ronchitelli A et al. . 2008. A 28,000 years old Cro-Magnon mtDNA sequence differs from all potentially contaminating modern sequences. PloS one 3(7):e2700.

Cooper A, Drummond AJ, and Willerslev E. 2004. Ancient DNA: would the real Neandertal please stand up? Curr Biol 14(11):R431-433.

Currat M, and Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS biology 2(12):e421.

Foley R, and Lahr M. 1997. Mode 3 technologies and the evolution of modern humans. Cambridge Archaeological Journal 7:3-36.

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 (New York, NY 328(5979):710-722.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M et al. . 2006. Analysis of one million base pairs of Neanderthal DNA. Nature 444(7117):330-336.

Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, von Haeseler A, Grossschmidt K, Possnert G, Paunovic M, and Paabo S. 2000. A view of Neandertal genetic diversity. Nature genetics 26(2):144-146.

Krings M, Geisert H, Schmitz RW, Krainitzki H, and Paabo S. 1999. DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proceedings of the National Academy of Sciences of the United States of America 96(10):5581-5585.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, and Paabo S. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90(1):19-30.

Lalueza-Fox C, Krause J, Caramelli D, Catalano G, Milani L, Sampietro ML, Calafell F, Martinez-Maza C, Bastir M, Garcia-Tabernero A et al. . 2006. Mitochondrial DNA of an Iberian Neandertal suggests a population affinity with other European Neandertals. Curr Biol 16(16):R629-630.

Lalueza-Fox C, Sampietro ML, Caramelli D, Puder Y, Lari M, Calafell F, Martinez-Maza C, Bastir M, Fortea J, de la Rasilla M et al. . 2005. Neandertal evolutionary genetics: mitochondrial DNA data from the iberian peninsula. Molecular biology and evolution 22(4):1077-1081.

Mellars PA. 1992. Archaeology and the population-dispersal hypothesis of modern human origins in Europe. Philosophical transactions of the Royal Society of London 337(1280):225-234.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Paabo S, Pritchard JK et al. . 2006. Sequencing and analysis of Neanderthal genomic DNA. Science (New York, NY 314(5802):1113-1118.

Nordborg M. 2001. On detecting ancient admixture. Genes, fossils and behaviour: An integrated approach to human evolution

Amsterdam: Ios Press. p 123-136.

Ovchinnikov IV, Gotherstrom A, Romanova GP, Kharitonov VM, Liden K, and Goodwin W. 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature 404(6777):490-493.

Paabo S. 1999. Human evolution. Trends in cell biology 9(12):M13-16.

Rightmire G. 2001. Patterns of hominid evolution and dispersal in the Middle Pleistocene. . Quaternary International 75:77-84.

Schmitz RW, Serre D, Bonani G, Feine S, Hillgruber F, Krainitzki H, Paabo S, and Smith FH. 2002. The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. Proceedings of the National Academy of Sciences of the United States of America 99(20):13342-13347.

Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, and Paabo S. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. PLoS biology 2(3):E57.

Wakeley J, and Hey J. 1997. Estimating ancestral population parameters. Genetics 145(3):847-855.

Wall JD, and Kim SK. 2007. Inconsistencies in Neanderthal genomic DNA sequences. PLoS genetics 3(10):1862-1866.

Weaver TD, Roseman CC, and Stringer CB. 2008. Close correspondence between quantitative- and molecular-genetic divergence times for Neandertals and modern humans. Proceedings of the National Academy of Sciences of the United States of America 105(12):4645-4649.

Inferring Genealogical Processes from Patterns of Bronze-Age and Modern DNA Variation in Sardinia

Silvia Ghirotto,[1] Stefano Mona,†[,1] Andrea Benazzo,[1] Francesco Paparazzo,‡[,1,2] David Caramelli,[3] and Guido Barbujani*[,1]

[1]Dipartimento di Biologia ed Evoluzione, Università di Ferrara, Ferrara, Italy

[2]Dipartimento di Biologia, Università di Milano, Milano, Italy

[3]Laboratorio di Antropologia, Dipartimento di Biologia Evoluzionistica, Università di Firenze, Firenze, Italy

†Present address: Computational and Molecular Population Genetics, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

‡Present address: Section of Evolutionary Biology, Ludwig-Maximilians-University BioCenter, Planegg-Martinsried, Germany

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

Associate editor: Jody Hey

Abstract

The ancient inhabitants of a region are often regarded as ancestral, and hence genetically related, to the modern dwellers (for instance, in studies of admixture), but so far, this assumption has not been tested empirically using ancient DNA data. We studied mitochondrial DNA (mtDNA) variation in Sardinia, across a time span of 2,500 years, comparing 23 Bronze-Age (nuragic) mtDNA sequences with those of 254 modern individuals from two regions, Ogliastra (a likely genetic isolate) and Gallura, and considering the possible impact of gene flow from mainland Italy. To understand the genealogical relationships between past and present populations, we developed seven explicit demographic models; we tested whether these models can account for the levels and patterns of genetic diversity in the data and which one does it best. Extensive simulation based on a serial coalescent algorithm allowed us to compare the posterior probability of each model and estimate the relevant evolutionary (mutation and migration rates) and demographic (effective population sizes, times since population splits) parameters, by approximate Bayesian computations. We then validated the analyses by investigating how well parameters estimated from the simulated data can reproduce the observed data set. We show that a direct genealogical continuity between Bronze-Age Sardinians and the current people of Ogliastra, but not Gallura, has a much higher probability than any alternative scenarios and that genetic diversity in Gallura evolved largely independently, owing in part to gene flow from the mainland.

Key words: ancient DNA, mitochondrial DNA, coalescent simulations, approximate Bayesian computation.

Introduction

For several decades now, important aspects of human evolutionary history have been reconstructed by studying patterns of genetic variation in the geographical space. Traditionally, these studies start from a description of genetic diversity in samples of contemporary people, from which inferences are drawn on the relative weight of natural selection, mutation, drift, long-range migration, and short-range gene flow in the population's history (see e.g., Menozzi et al. 1978; Sokal et al. 1991; Nielsen 2005; Nielsen and Beaumont 2009).

Two recent developments have substantially improved the power of such studies. One is the availability of ancient DNA data. Although information on genetic diversity in the past remains essentially limited to mitochondrial DNA (mtDNA; see e.g., Bramanti et al. 2009) because of the well-known risk of undetected modern contamination (Pääbo et al. 2004), in principle, questions on the existence and strength of genealogical ties between ancient and modern people can now be empirically addressed. The second development is a new set of statistical methods, designed to compare alternative evolutionary models and to estimate the relevant parameters, referred to as approximate Bayesian computations (ABC; Beaumont et al. 2002). These methods proved powerful in addressing several biological questions, ranging from the introduction of corn worm in Europe (Miller et al. 2005), the evolution of intrahost HIV genetic diversity (Shriner et al. 2006), and the spread of tuberculosis (Tanaka et al. 2006) to the origin of early modern humans (Fagundes et al. 2007), Polynesians (Kayser et al. 2008), and pygmies (Verdu et al. 2009). So far, ABC methods have never been used to compare ancient and modern human DNA data and test alternative models of their genealogical relationships.

This study stemmed from the observation that a sample of mtDNA sequences from Bronze-Age Sardinia, known in archeology as the "nuragic" people, shows very different relationships with two modern populations of the island, separated in space by less than 120 km. More than half of the mitochondrial haplotypes of the nuragic sample are present in one region, Ogliastra, but only 18% in the other region, Gallura, which is the same proportion one would observe by picking up random modern individuals from all over Europe (Caramelli et al. 2007). Sardinia is known as one of the

main genetic outliers in Europe (Cavalli-Sforza and Piazza 1993; Quintana-Murci et al. 2003; Pugliatti et al. 2006) and shows unusually high levels of internal diversity (Barbujani and Sokal 1991; Zei et al. 2003), but the existence of such sharp differences between one modern population and the ancient inhabitants of the island calls for an explanation.

To find such an explanation, we generated by serial coalescent simulation (Anderson et al. 2005) a total of 10.5 million mtDNA genealogies, considering alternative models of the genetic relationships among populations and a wide range of parameter values within models. Under the ABC framework, we then compared the posterior probabilities of the models and we estimated the most likely parameter values. Finally, we showed that using the parameter values estimated under the most likely model, we could generate patterns of genetic diversity that closely resemble the observed ones.

Materials and Methods

Genetic Data

The data analyzed are sequences of the first hypervariable region of mtDNA (HVR1) spanning 360 bp. The ancient Sardinian data set is represented by 23 Bronze-Age, or nuragic, sequences (Caramelli et al. 2007), and the modern Sardinian data set includes two samples, respectively, from Ogliastra ($n = 175$), generally considered a genetic isolate (Fraumene et al. 2003), and Gallura ($n = 27$), an area in which immigration is documented in historical times (Morelli et al. 2000). Modern samples from mainland Italy, namely Latium ($n = 52$; Babalini et al. 2005) and Tuscany ($n = 197$; Achilli et al. 2007), were used as proxies for DNA diversity in recent immigrants. In fact, only 52 random Tuscan sequences were considered, so as to have the same sample sizes for both modern Italian populations.

Summary Statistics

We estimated in each sample 1) the number of different haplotypes, 2) the number of segregating sites, 3) the average number of pairwise differences, 4) haplotype diversity, and 5) Tajima's *D*, as measures of internal genetic diversity. In addition, we quantified the relationships between samples by (6–8) three measures of haplotype sharing (estimated as the number of shared haplotypes between two populations scaled by the total number of haplotypes in the ancient sample or, for the comparison between modern populations, in the Ogliastra sample), and (i) Hudson's F_{st} (Hudson et al. 1992). We preliminarily tested different sets of summary statistics, always obtaining comparable results. In particular, because the two modern Sardinian samples have very different sizes (175 vs. 27), we resampled 1,000 times 27 sequences from the larger sample, Ogliastra, and calculated from them the haplotype sharing. This procedure had the purpose to determine whether the haplotype sharing values somewhat reflected the different sample sizes; however, in the ABC procedure, we always considered values estimated from the whole Ogliastra sample. In the more complex simulations taking into account modern samples from the mainland (Models 4–7), we summarized variation within samples only by three statistics (haplotype number, segregating site, and pairwise differences). Summary statistics in the observed data were calculated by Arlequin version 3.1 (Excoffier et al. 2005).

The Simulations

Mitochondrial genealogies of samples collected at different moments in time were simulated using a serial coalescent algorithm, according to specific demographic models. Suppose that one has samples of size n_0, n_1, $n_2 \ldots n_k$, of individuals studied t_0, t_1, $t_2 \ldots t_k$ generations ago. The serial coalescent algorithm (Anderson et al. 2005) generates genealogies proceeding backward in time, starting with n_0 samples in the present (t_0) and adding $n_1, n_2 \ldots n_k$ samples at the appropriate moments in the past. The genealogy was extended backward in time until it reached the most recent common ancestor of the sampled lineages through a series of coalescence events. Then, mutations were added onto the tree according to an infinite-site model. Each of the demographic models tested was characterized by a series of parameters, detailed below. The Bayesian version of the SERIALSIMCOAL program (Anderson et al. 2005) freely available on http://iod.ucsd.edu/simplex/ssc/BayeSSc.htm was used to generate simulated genealogies and to estimate summary statistics from the simulated data.

Demographic Models

We considered seven demographic models, differing for the relationships between ancient and modern samples and for the presence of immigration from the mainland. Under Model 1, the ancient sample is ancestral to the Ogliastra but not to the Gallura population; under Model 2, to the Gallura but not to the Ogliastra population, and under Model 3, to both. Models 4 through 6 are analogous, with additional gene flow from the mainland into Gallura in the time period separating ancient and modern samples. We fixed the separation time between the populations of mainland Italy and Sardinia at 721 generations ago (=18,000 years ago), corresponding to the first likely human presence in Sardinia (Vona 1997). The last model, Model 7, is equivalent to Model 4, but migration rate from Latium to Gallura is fixed to 0, so as to essentially replicate the features of Model 1, making it comparable with models with immigration.

In all simulations, the modern samples were placed at generation 0, and the nuragic samples at generation 126, corresponding to the average age of the ancient specimens, 3,146 years, thus assuming that a generation lasts on average 25 years (Fenner 2005; see also Currat and Excoffier 2004; Noonan et al. 2006; Fagundes et al. 2007). The ancestors of the Ogliastra and Gallura populations could separate from their common ancestor at a time >126 generations under Models 1, 2, 4, 5 and 7 or <126 generations under Models 3 and 6 because only in this way could the nuragic people be regarded as ancestral to the appropriate modern samples.

Bronze-Age and Modern DNA Variation · doi:10.1093/molbev/msp292 MBE

Approximate Bayesian Computations

Models were compared, and parameters were estimated, by ABC. Approaches based on ABC algorithms include the following steps: 1) a large number of simulations are performed under the chosen model, with demographic parameters extracted from prior distributions, representing the prior knowledge on the possible parameter values; 2) a vector of summary statistics is computed in each simulation; 3) the euclidean distance is computed between each simulated vector of summary statistics and the vector of observed statistics; 4) the parameter values associated with an arbitrary number d (or "threshold") of simulations, that is, the d simulations closest to the observed data, are retained; 5) after a transformation of the parameters (see Hamilton et al. 2005), a weighted local regression is performed to adjust the values of the retained parameters using summary statistics as predictors. Parameters were estimated by retaining, for each model tested, the 2,000 simulations associated with the shortest euclidean distances, chosen from a total of 1.5 millions simulations per model. This was done in the R environment (R Development Core Team 2008) using a modified version of the makepd4 script, freely available at http://www.rubic.rdg.ac.uk/~mab/stuff.

Priors

For all models, all priors were taken from uniform distributions, in the range described below: Modern N_e, Gallura and Ogliastra, between 100 and 200,000; ancestral N_e, one generation after the split, between 5 and 6,000; and separation time between Gallura and Ogliastra, between 127 and 720 generations (or between 0 and 125 generations for Model 3 and 6). HVR1 mutation rate between 0.0003 and 0.006, corresponding to between 0.06 and 1.3 mutations per million years per site (commonly accepted estimates range from 0.05 to 0.5; Pakendorf and Stoneking 2005). Models were also tested with a fixed mutation rate of 0.0027 substitutions per generation for the 360 bp of the mitochondrial HVR1, which was shown to be compatible with the time window under investigation (Henn et al. 2009).

Under Models 4–7, modern Latium N_e was 400,000, that is, one-twelfth of the 2001 census population size. In a panmictic population, the individuals who actively reproduce are around one-third of the census size (see e.g., Tishkoff and Gonder 2007; Cela-Conde and Ayala 2007). Because females are one-half of the reproductively active individuals, a rough estimate of the N_e for mtDNA would be around one-sixth of the census size. We further divided this value by 2 to take into account the fact that the current population increased dramatically in recent times because of massive immigration into Rome and the increased effects of drift in subdivided populations. The time since separation of the Sardinia and mainland populations was fixed at 721 generations (Vona 1997); migration rate from Latium into Gallura was between 0 and 0.01 per generation. The same set of priors were also used when we simulated immigration from Tuscany, rather than from Latium.

Model Selection

Models were compared by estimating their posterior probabilities in two ways. The posterior probability can easily be estimated by acceptance–rejection sampling (Pritchard et al. 1999), comparing the distribution of normalized distances between observed and simulated summary statistics (acceptance-rejection [AR] method). If all models have the same prior probability, the posterior probability of the i-th model is simply obtained by ranking simulations according to their associated distances. One then counts how many simulations run under the i-th model (n_i) are found among an arbitrary number, d, of the simulations resulting in the shortest distances between observed and simulated data. The posterior probability for the model is then equal to n_i/d.

Results of previous studies suggest that straightforward rejection may not be robust when d is greater than a few hundred simulations (Beaumont 2008). The alternative approach (logistic regression [LR] method) estimates the models' posterior probabilities by multinomial logistic regression, which is known to perform better than the AR method particularly when investigating the population tree topology (Beaumont 2008). Under the LR method, a logistic regression is fitted where the model is the categorical dependent variable Y_j ($1 \leq j \leq 3$ when comparing Models 1–3 and 4–6; $1 \leq j \leq 4$ in the comparisons of Models 4–7) in the ABC simulations and the summary statistics are the predictive variables (Fagundes et al. 2007; Beaumont 2008). The regression is local around the vector of observed summary statistics in the same way as in the parameter estimation procedure. The probability of the model is finally evaluated in the point corresponding to the observed vector of summary statistics.

The β coefficients of the regression model were estimated by maximum likelihood; the standard error of the estimates was taken as a measure of the accuracy of the posterior probabilities. For both AR and LR, we used the "calmod" function, written by M. A. Beaumont (available at http://www.rubic.rdg.ac.uk/~mab/stuff/) for the R statistical package. Model selection within each set of scenarios was based on 1,500,000 simulations for each model. Different numbers of simulations (i.e., different thresholds) were considered for both approaches. Finally, we analyzed the power of the LR procedure to correctly recover the true model as suggested by Fagundes et al. (2007) and Cornuet et al. (2008). Specifically, we first simulated 1,000 data sets from the prior distribution under each model considered (for a total of 7,000 simulated data sets) and analyzed them using the same simulations and setting as in the observed data. We thus assigned each of the 1,000 simulated data sets to the model showing the highest posterior probability and counted how many times the true model was correctly identified. Type I error is the fraction of cases in which the true model was not recovered.

Quality of the Estimation

To determine whether the summary statistics we chose contain enough information to estimate model parameters, during the regression step, we computed the

coefficient of determination (R^2). R^2 indicates the percentage of variance of the dependent variable (i.e., the parameter) explained by the predictors (i.e., the summary statistics). In the absence of an established threshold value, there is a general agreement that when $R^2 < 0.10$, the summary statistics do not convey enough information about their posterior distribution (Neuenschwander et al. 2008).

The accuracy of the median estimate of model parameters was assessed computing relative bias and relative mean square error. For these tests, we generated 1,000 data sets using our median point estimates as demographic parameters. Each of these 1,000 data sets was used as a pseudo-observed data set, which was analyzed with the 1,500,000 simulations previously performed for ABC estimation in the observed data. Bias and root mean square error (RMSE) depend, respectively, on the sum of differences and on the sum of squared differences, between the 1,000 estimates of each parameter thus obtained and the respective median point estimate (Neuenschwander et al. 2008). A value of 0 means that the median perfectly estimated the parameter, positive and negative values reflect, respectively, biases toward overestimation and underestimation.

We also calculated the factor 2 statistic, representing the proportion of the 1,000 estimated median values lying between the 50% and the 200% of the fixed (known) value, and the 50% coverage, defined as the proportion of times that the known value lies within the 50% credible interval of the 1,000 estimates. Note that factor 2 gives information about the absolute precision of the estimator because it is independent of the posterior distribution's variance (which, conversely, is not a property of the coverage).

Posterior Predictive Tests for the Models

Finally, we evaluated by a posterior predictive test whether, under any specific model, we were able to reproduce the observed data (Gelman et al. 2004). This test is the Bayesian analogue of the parametric bootstrap under the frequentist framework. Its rationale is, if our posterior distribution estimates are plausible, they should be able to generate data sets similar to the observed data. The discrepancy between the model and the data is measured by a test quantity yielding a final Bayesian *P* value that can be interpreted as the probability of accepting the null hypothesis that our data have been generated by that model (Gelman et al. 2004). For each demographic model, we first computed a posterior predictive *P* value for each of the statistics considered and then combined the probabilities of the single statistics into a global *P* value, by a method that takes into account nonindependence of the statistics (Voight et al. 2005). Briefly, for each model of interest, random draws from the posterior probability of the demographic parameters inferred by ABC were used to generate by coalescent simulations 10,000 data sets with the sample size of the sample considered. Summary statistics were then computed in these data sets to obtain their null distribution (under the model of interest), against which we tested the summary statistics computed in our observed data obtaining a Bayesian *P* value for each statistic. The global *P* value was calculated in four additional steps: 1) Each simulated summary statistic was compared with the other 9,999 values representing the empirical distribution of the statistic from simulation and thus associated with a two-tailed *P* value; 2) For each simulated genealogy, a new statistic *C*, combining the *P* values of the individual statistics (p_i), was calculated as:

$$C = -2\sum \ln(p_i),$$

where summation is over all *P* values from each summary statistic. This step was repeated 10,000 times, so as to obtain a null distribution of *C*; 3) By repeating the same procedure with the observed statistics, we calculated an observed *C* value, C_o; 4) By comparing C_o with the *C* null distribution, we estimate a one-tailed *P* value (the Bayesian *P* value) for C_o.

Results

A median-joining network (Bandelt et al. 1999) summarizing the relationships among the DNA sequences of modern and ancient populations is in figure 1, and a list of the nucleotide substitutions observed in the ancient specimens is in supplementary table S1 (Supplementary Material online).

Choosing the Best Model

Summary statistics computed from the observed data (table 1), namely mtDNA sequences in nuragic Sardinians ($n = 23$) and modern people from Ogliastra ($n = 175$), Gallura ($n = 27$), and Latium ($n = 52$) (fig. 2), were compared with the statistics calculated from the simulated data. In addition, we also ran the same simulations and analyses using Tuscany ($n = 52$) instead of Latium. The results were absolutely consistent when immigrants came from either mainland population, and so, unless otherwise specified, our comments will refer to the simulations in which immigrants were taken from the Latium data set. Because the two modern Sardinian samples have different sizes, as a preliminary test of the effects of sample size, we resampled 1,000 times 27 sequences from the Ogliastra data set and calculated from them the summary statistics. We found that sample size had but a minimal effect on the estimates, and so, we could conclude that the higher fraction of Bronze-Age haplotypes shared by Ogliastra than by Gallura is not simply an artifact and is informative for the inference of genealogical relationships.

We started from six demographic models, differing from each other as for the genealogical relationships between the nuragic and the modern samples (fig. 3). Sardinia is genetically isolated under Models 1–3, whereas Models 4–6 incorporate variable rates of immigration from mainland Italy.

Model 1 was favored among the models without immigration (fig. 4A), showing a posterior probability up to 0.97 and in any case never less than 0.70, depending on the criterion chosen to compare the results across models. Alternative models received only scanty, if any, support. When immigration was added to the previously simulated

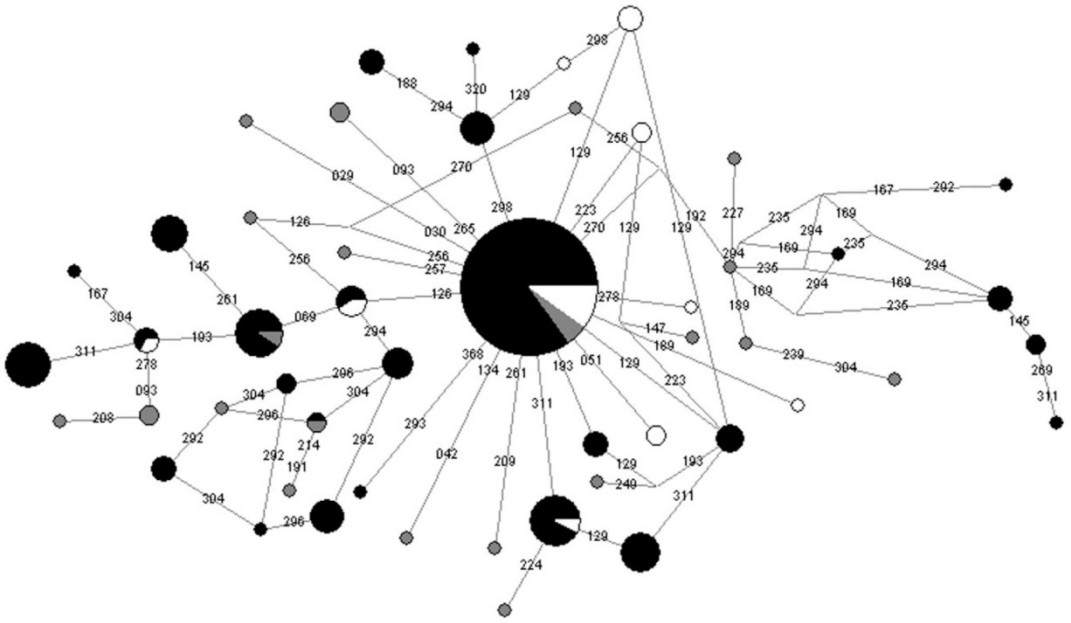


FIG. 1. Median-joining network of the DNA sequences considered. Ancient samples are represented by white areas in the pies, Gallura by gray areas and Ogliastra by black areas. Figures on the edges of the network indicate the position of the nucleotide substitution in the mtDNA reference sequence minus 16,000.

scenarios, we fixed the separation time between the populations of Latium and Sardinia at 721 generations ago (=18,000 years ago), corresponding to the first likely human presence in Sardinia (Vona 1997), so that that separation would necessarily precede the split between the ancestors of current people from Gallura and Ogliastra. Equal levels of genetic diversity can be obtained through many generations of gene flow at low rates, a few generations of intense gene flow, or any combinations of factors in between (Hey 2006). Therefore, fixing the separation time was expected to simplify the estimation of migration rates, and it did. In the comparison of Models 4–6 (fig. 4*B*), Model 4, analogous to Model 1 in the assumed genealogical links, with the addition of gene flow from Latium into Gallura, showed the highest posterior probability (between 0.71 and 0.79, depending on the criterion chosen). Little changed if the Tuscany, and not the Latium, data set was used as a source of migrants into Sardinia (fig. 4*C*).

Models without immigration and models with immigration from Latium could not be directly compared because of the different data sets analyzed. However, in both cases, the models of genealogical continuity with Ogliastra (1 and 4) proved better than the others. The question to address, at that point, was only whether migration adds to the ability of the model to account for the data. To answer, we developed a seventh model, identical to Model 4 but with *m* set to 0. In this way, we obtained a way to test on the same data set (including the Latium data set in addition to the ancient and modern Sardinians) whether considering gene flow from the mainland improves the resemblance between observed and simulated statistics. In fact, little changed when the models including immigration were compared with Model 7, which shows a posterior probability between 0.15 and 0.30, versus 0.10 or less for Models 5 and 6 (data not given). When the comparison was restricted to the two best models, 4 and 7, both considering

Table 1. Observed Summary Statistics Describing Genetic Variation in the Samples.

	Bronze Age	Ogliastra	Gallura	Latium
Haplotype number	10	26	21	36
No. of segregating sites	10	22	31	45
Mean pairwise difference	1.39	2.49	4.42	4.07
Haplotype diversity	0.83	0.79	0.97	0.95
Tajima's *D*	−1.64	−0.97	−1.66	−2.02
F_{st} (Ogliastra-Gallura)		0.0218		
Haplotype sharing	Ogliastra/Bronze Age = 0.400 Gallura/Bronze Age = 0.100 Gallura/Ogliastra = 0.095			

FIG. 2. A map of Sardinia (left) and its geographical relationship with mainland Italy. Gallura and Ogliastra are shaded in gray; solid circles represent the archeological sources of the ancient specimens considered in this study.

the nuragic people as ancestral to the Ogliastra populations, nonzero immigration from the mainland into Gallura (Model 4) resulted in a roughly 2-fold greater posterior probability when compared with no immigration (Model 7), with values ranging from 0.64 to 0.71 (fig. 4*D*).

These results show altogether that what really made the difference among models was to represent the Ogliastra people as direct descendants of local nuragic ancestors (Models 1, 4, and 7), to the exclusion of the Gallura people. Considering immigration from the mainland into Gallura did increase the resemblance between simulated and observed statistics, although up to one-third of the simulations favored Model 7, both when compared with all alternative models and when compared with Model 4 only.

In all these tests, the mitochondrial mutation rate was estimated from the data. We also repeated the experiments assuming a fixed molecular clock at a rate of 0.3 substitutions per nucleotide per million year (0.0027 per generation for the 360 bp of the HVR1 assuming a generation time of 25 years). This value was taken from a recent study (Henn et al. 2009) and seems plausible for the time window we are considering. Results with the fixed rate were essentially the same as above (data not given).

Estimating Population Parameters

Table 2 shows the posterior distribution of the parameters estimated under Models 1 and 4, respectively, along with the priors. The mutation rate (median values of 0.0020 and 0.0014 per generation for the 360-bp hypervariable mtDNA region for Models 1 and 4, respectively) is close to the values accepted in most studies of mtDNA diversity (Vigilant et al. 1991; Forster et al. 1996) and barely lower than the

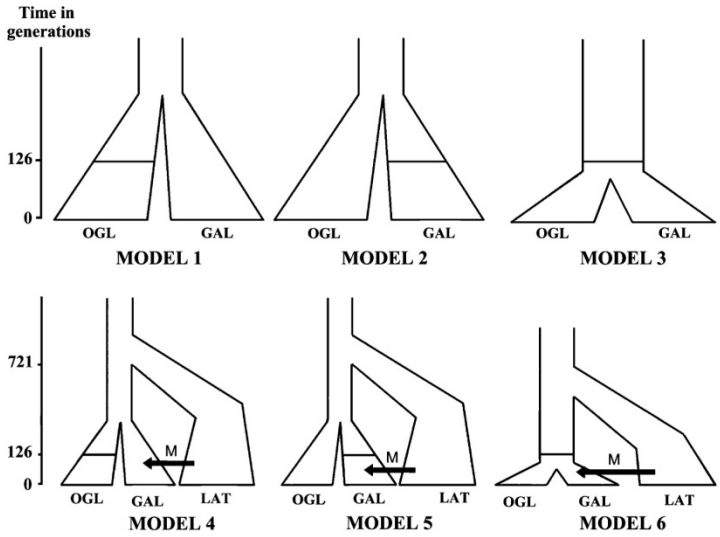


FIG. 3. A schematic summary of the six models tested. Numbers on the *y* axis are generations from the present. The horizontal line at generation 126 represents the nuragic population; the arrows represent gene flow at a rate *M*, which was estimated from the data, from the mainland into Gallura. OGL, Ogliastra; GAL, Gallura; LAT, Latium.

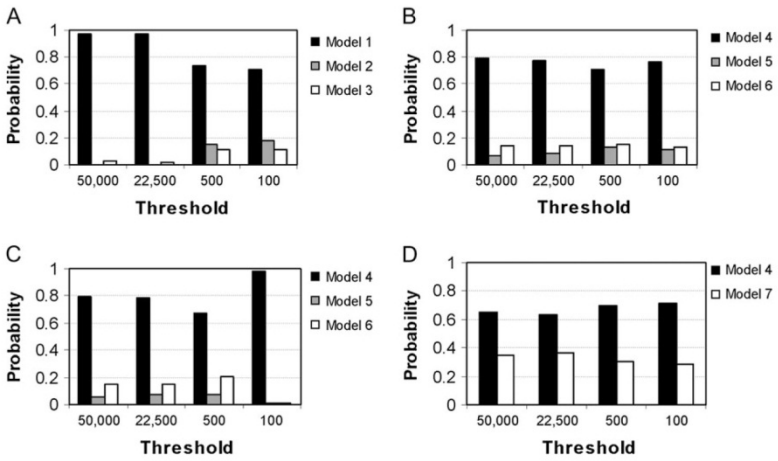


FIG. 4. Posterior probabilities of the models evaluated over different thresholds. Thresholds on the *x* axis indicate the number of simulations retained under the AR (acceptance–rejection sampling: 500 and 100) and the LR (logistic regression: 50,000 and 22,500) approaches. (*A*) Models without immigration; (*B*) models with immigration from Latium; (*C*) models with immigration from Tuscany; and (*D*) a comparison of the two best models, with immigration from Latium (Model 4) and without it (Model 7).

value (0.0027) estimated by Henn et al. (2009), under a model that differed from ours in that it did not include migration. Credible intervals of mutation rate include the values estimated in recent studies (Henn et al. 2009; Soares et al. 2009), and in one case (Model 4), the median value overlaps exactly with the estimate of Soares et al. (2009). Therefore, our estimates for this parameter appear reasonable, and robust, as shown by the high values of the coefficient of determination, R^2.

Under both models, the ancient effective population sizes appear to be around one or a few thousand individuals, in agreement with Y-chromosome–based estimates for the European Palaeolithic (Contu et al. 2008) and with the finding that N_e is systematically larger for females than males in humans (Dupanloup et al. 2003; Wilder et al. 2004). Estimated modern population sizes can be compared with the data of the 2008 census, that is, 58,389 for Ogliastra and 153,339 for Gallura (see http://demo.istat.it/). Because, as a rule of thumb, approximately one-third of the population is considered to be reproductively active in humans and because half of the reproductively active individuals are females, one should expect N_e values around one-sixth of the census values, that is, 10,000 and 26,000, respectively (or less, if the population is subdivided). In fact, under Model 4, we found N_e of 11,290 for Ogliastra, which seems an excellent approximation, especially considering that our study is necessarily based on a single locus. On the other hand, we estimated N_e at 104,000 in Gallura,

Table 2. Demographic Parameters Estimated under Models 1 (Upper Panel) and 4 (Lower Panel).

	Median	0.025[a]	0.975[a]	R^{2b}	Prior[c]
Model 1					
N_e[d] Ogliastra	8,947	2,645	65,724	0.550	U: 100–200,000
N_e[d] Gallura	128,534	21,010	196,314	0.171	U: 100–200,000
Separation time	551	230	714	0.286	U: 127–2,5000
Mutation rate	0.0020	0.0009	0.0044	0.688	U: 0.0003–0.006
Ancestral N_e[d] Ogliastra	346	65	2,061	0.446	U: 5–6,000
Ancestral N_e[d] Gallura	811	121	4,655	0.351	U: 5–6,000
Model 4					
N_e[d] Ogliastra	11,290	2,646	70,762	0.540	U: 100–200,000
N_e[d] Gallura	104,183	10,450	195,062	0.129	U: 100–200,000
Migration rate from Latium	0.00497	0.00031	0.00972	0.081	U: 0–0.001
Separation time (Sardinia)	513	185	709	0.291	U: 127–720
Mutation rate	0.0014	0.0008	0.0023	0.746	U: 0.0003–0.006
Ancestral N_e[d] Ogliastra	824	158	4,177	0.455	U: 5–6,000
Ancestral N_e[d] Gallura	683	37	4,564	0.149	U: 5–6,000
Ancestral N_e[d] Latium	1,137	124	5,291	0.337	U: 5–6,000

[a] Upper and lower limits of the 95% credible interval about the estimated median.
[b] Coefficient of determination.
[c] U, uniform probability, in the range between the two values.
[d] Effective female population size.

Table 3. Power of the LR Procedure to Recover the True Model.

		% of Model Attribution[a]				
Without immigration		**MOD1**	**MOD2**	**MOD3**	**Total**	
Simulated model	**MOD1**	92.9	1.2	5.9	100.0	
	MOD2	1.6	89.8	8.6	100.0	
	MOD3	4.8	6.1	89.1	100.0	
With immigration		**MOD4**	**MOD5**	**MOD6**	**Total**	
Simulated model	**MOD4**	94.8	0.4	4.8	100.0	
	MOD5	3.8	91.9	4.3	100.0	
	MOD6	6.0	3.4	90.6	100.0	
With immigration, plus Model 7		**MOD4**	**MOD5**	**MOD6**	**MOD7**	**Total**
Simulated model	**MOD4**	71.2	0.5	4.1	24.2	100.0
	MOD5	2.9	90.9	5.1	1.1	100.0
	MOD6	3.7	2.9	87.0	6.4	100.0
	MOD7	26.6	1.9	5.6	65.9	100.0

[a] Proportion of cases in which the analysis correctly recovered the true model. One-thousand replicates were generated for each model using random values drawn from the prior distributions. Replicates were considered assigned to the model that has the highest posterior probability.

corresponding to more than 600,000 in census terms. We believe that this high value basically reflects a high mtDNA variation in Gallura; under the conditions of Model 4, those levels of diversity can only be generated in a very large population or if the mutation rate is very high but not by the effects of continuous gene flow with neighboring populations, which we could not incorporate in the model. The uncertainty in the N_e estimates is also shown by the broad posterior probability distributions and by the low R^2 values, both for the current population and for the ancestral population after separation from the common ancestor to Bronze-Age nuragic people (table 2). Conversely, the posterior probability distribution is narrower, and R^2 is high (>0.5) for the modern and ancient Ogliastra's N_e estimates. Both results indicate that the summary statistics used to infer the posterior distribution of N_e in Gallura do not harbor sufficient power for an accurate estimation. The results also suggest that the Gallura population received immigrants from the mainland at a median rate of 0.005 per generation, but, once again, this value represents the effect of one of the probably multiple migration processes, that is, the only one we could model with reasonable accuracy.

The median separation of the two ancient Sardinian populations (one ancestral to both nuragic and Ogliastra people, the other to the Gallura people) is around 513 generations, or 12,825 years ago, but 95% of the values estimated from the best simulations fall in a broad interval, between 185 and 709 generations ago (4,625–17,725 years ago).

Validating the Estimated Statistics

We then ran several tests to assess the quality of our estimates. First, we calculated for each demographic parameter of each model two statistics, the relative bias and the relative RMSE, to quantify the accuracy of the estimated median values. Second, we calculated factor 2 statistic and 50% coverage, two indexes of the quality of the posterior distributions (supplementary table S1, Supplementary Material online).

Both the relative bias and the relative RMSE are generally low and do not point to any systematic over- or underestimation of the various parameters. The Models associated with the highest posterior probability (Models 1 and 4) do not have (with few exceptions) bias or RMSE higher than one. In general, the width of the 50% credible intervals is small, showing that the parameters are reasonably well estimated; most values of the parameter estimates resulting from these pseudo-observed data sets lie between 50% and 200% of the estimated median values.

We then asked whether models are different enough for us to correctly recover the true model by the logistic regression procedure (type I error). To answer, we counted the number of cases in which we recovered the true model in a set of 1,000 simulations from the prior distributions of each model. Because of the different data sets used, we had to separately compare models without, and with, immigration. We found that the data sets generated under Models 1 through 3 are correctly identified (i.e., have the highest posterior probability) in the vast majority (89% or more) of cases, and the same was the case for Models 4 through 6 (91% or more) (table 3). When Model 7 was compared with models with immigration, a slight loss of power was evident because Models 4 and 7 are very similar. Nevertheless, Model 7 was identified as the correct one in almost two-thirds of the experiments.

Finally, we ran posterior predictive tests to evaluate whether we could reproduce the observed data, under the specific demographic scenario described by each model. We found that no scenario can actually be rejected (the global P values were insignificant for all models; supplementary table S2, Supplementary Material online). When we considered each summary statistic, we found that only Model 1 showed all insignificant P values. Under Model 4, which was favored by a large majority of the tests we ran, 23 of the 25 statistics considered could be faithfully reproduced, but significant differences merged for Tajima's D in the nuragic population and for the level of haplotype sharing between ancient and modern individuals. In other words, models of genealogical continuity between Nuragic Sardinians and Ogliastra 1) showed in every case the

highest posterior probabilities, regardless of whether the model included immigration from the mainland (Models 1 and 4) and 2) generated data whose summary statistics are largely (when immigration from the mainland was considered; Model 4) or fully (in the case of no immigration; Model 1) compatible with the observed ones.

Discussion

The first human remains discovered so far in Sardinia date back to 14,000 years ago, and the first human presence in the island may be placed around 18,000 years ago (Vona 1997). The analysis of mtDNA variation in ancient and modern Sardinia and the comparison of observed and simulated patterns of mtDNA diversity clearly show that haplotypes documented in the Bronze Age, or derived from them assuming a reasonable mutation rate, are still present and common in the isolated Ogliastra community. Conversely, the modern population of Gallura seems derived from ancestors who separated in Palaeolithic times (>12,500 years ago) from the common ancestors of Bronze-Age and modern Ogliastra people and only have loose genealogical relationships, if any, with the ancient Sardinian people. Indeed, the only Bronze-Age sequence that is also observed in the modern Gallura sample is the Cambridge Reference Sequence (CRS), which is very common all over Europe. Conversely, the modern Ogliastra sample comprises not only the CRS but also two relatively rare sequences documented in the Ogliastra nuragic sites of Seulo and Perdasdefogu (Caramelli et al. 2007). All models assuming alternative genealogical links between past and present populations are much less supported by our analyses.

We assessed the quality of the analysis by a number of tests. First, we showed that in general, a large proportion of the parameters' variance is explained by the estimated summary statistics, and we identified the few parameters that could not be accurately estimated. Second, we evaluated the breadth of the empirical confidence intervals (in fact, 95% credible intervals) about the estimated parameters. Third, we showed in various ways that simulations based on the estimated parameters can in fact reasonably reproduce the observed data set. Clearly, a certain degree of uncertainty necessarily affects any analysis based on a single DNA region and on the necessarily small samples in which ancient DNA is typed. Within these unavoidable limits, we believe that the properties of the demographic models could hardly be explored in greater detail.

Under the models showing the best fit, Model 1 and Model 4, the Gallura population was larger than that of Ogliastra and grew faster through time, consistent with the trends known for the last centuries (Francalacci et al. 2003). Under Model 4, the population increase in Gallura appears partly due to immigration from the mainland, at a median rate that we estimate around 0.005 per generation (table 3). This value was calculated assuming that migration occurred at a constant rate, whereas in fact that seems unlikely; therefore, it should not be regarded as a precise measure of the actual input of genes at any moment in time. The estimated ancestral population sizes, between 1,000 and 2,000 individuals (corresponding to the sum of the two population sizes after the split), do not suggest that the Sardinian populations underwent dramatic bottlenecks, which is in good agreement with the population growth suggested by negative values of Tajima's *D* in both the modern and Bronze-Age populations.

The median values of the posterior distribution of the mutation rate are 0.0020 and 0.0014 for the whole HVR1 region (for Model 1 and Model 4, respectively), values in close agreement with those estimated in phylogenetic comparisons of humans and chimps (Pakendorf and Stoneking 2005) and hence with the relatively low values commonly accepted in studies of human mtDNA (see e.g., Hill et al. 2007). However, we also noticed that the median estimate of the mutation rate ranges from 0.0014 of Model 4 to 0.0049 of Model 3. This 3-fold difference shows how the evolutionary model considered affects the estimation of the mutation rate from ancient DNA data, an issue that has so far received little attention (but see Navascués and Emerson 2009). We think that these results illustrate how a wrong population genetic model can produce an undetected bias in the estimation of evolutionary and demographic parameters. In our study, we took different models into considerations, and so, we could notice that they yielded rather different estimates of mutation rates. On the contrary, most studies of modern DNA variation consider just a single model (i.e., constant population size or exponential increase of population size) and hence reach conclusions that may or may not hold true under different models. The R^2 values of table 2 represent the fraction of the total variance explained by the summary statistic, and show that most parameter estimates can be considered reliable (Neuenschwander et al. 2008), especially those referring to the mutation rate, and of N_e in both ancient and modern Ogliastra (all > 0.45).

In this analysis, as well as in previous diachronic analyses of genetic diversity (Belle et al. 2009), it proved difficult to reproduce the high number of different haplotypes of some modern populations. In the present study, that was the case for Gallura; considering its population extremely large by any standards (N_e > 100,000) was the only way to simulate levels of genetic diversity compatible with the observed ones. That N_e value is unrealistic, and it probably reflects the limitations of currently testable models. We could model directional gene flow from Latium (identified as a plausible source of immigrants) into Sardinia, but we had no useful information on the sources and rates of continuous immigration processes that likely occurred across the last millennia. Therefore, we had to represent our populations as essentially isolated; in this way, we disregarded the well-known fact that mtDNA diversity is not simply the product of mutations accumulating through time in isolation (see e.g., Wilkins 2006) but also reflects the input of lineages of different origins. The fact that not always could we reproduce the observed levels of Gallura's haplotype diversity seems a consequence of this inevitable approximation. Actually, the close agreement between our estimate

based on gene genealogies and the census data shows that the Ogliastra population, or at least its mtDNA pool, did evolve under strong reproductive isolation, as also indicated by previous studies (Morelli et al. 2000; Angius et al. 2001). On the contrary, the 4-fold difference between our N_e estimate for Gallura and the census value probably just means that Gallura was all but isolated, and gene flow from various sources increased its genetic diversity. Because we could not appropriately model this process, our simulations tended to reproduce the observed levels of diversity in Gallura by expanding the population size estimates, which also resulted in large variances about the median values. In short, not only the models we used are clearly a simplification of the true demographic history of these populations (even if we tried to accurately model the historical events that have likely shaped genetic diversity) but there is also an inherent limitation in using a single genetic marker to uncover complex demographic histories. Therefore, the data we are using contain enough information to estimate many, but not necessarily all, the parameters of interest, and that seems the case especially for the N_e of Gallura.

Predictably, when the simulations included variable rates of gene flow from the nearest mainland region, Latium, we could account for part of this excess variation; Model 4 (with gene flow) had indeed a greater posterior probability than Model 7. On the other hand, however, in the posterior predictive test (supplementary table S2, Supplementary Material online), only Model 1 proved to generate data that are fully compatible with the observed ones, whereas for Model 4, there were significant differences for 2 statistics of 25. In commenting this result, one has to keep in mind that there is often a trade-off between complexity of the model and its accuracy in reproducing the data. Models 4–7 have more parameters than Models 1–3, and as the number of parameters increases, their joint estimation becomes increasingly complicated. This problem is particularly serious when a single locus is considered, as in this study; however, we do not foresee any simple solution. In short, complicating the models did not fully clarify the missing details of the picture, and hence, at this stage, further complications seem unlikely to decrease significantly the estimates' uncertainty. Substantial progress in this area is to be expected only with the development of reliable methods for the typing of nuclear DNA polymorphisms in ancient samples.

Even so, this study casts new light on the nature and the extent of the genealogical links between past and present populations, a long-term source of controversy in evolutionary biology and not only there. In studies of admixture, allele frequencies of modern populations are often considered to approximate the unknown allele frequencies of the past (see e.g., Gauniyal et al. 2008; Auton et al. 2009). Although algorithms have been developed to somehow take into account the effect of genetic drift through time (Chikhi et al. 2001; Sousa et al. 2009), a genealogical continuity between the people occupying a certain region in the past and in the present is still a very common assumption. Often, such approximate admixture estimates are crucial for understanding disease susceptibility or in other medical applications (Lai et al. 2009), and so, errors in their estimation may lead to incorrect conclusions about the interaction between genes and environment in determining phenotypes of clinical relevance.

This study, albeit limited to DNA transmitted along the female lines of descent, strongly suggests that such a continuity is certainly a possibility, but not necessarily a general rule across several centuries, as previously shown, for instance, by the comparison of Etruscans and modern Tuscans (Guimaraes et al. 2009). Even when separated by short geographical distances, as in our case, modern populations may differ sharply in their genealogical relationships with prehistoric and historic inhabitants of nearby territories. However, this study also shows that it is actually possible to test for genealogical continuity across time and hence base the admixture estimation procedure upon empirical genetic information. Whenever ancient DNA data are available, a preliminary validation of the assumptions on genetic ancestry is feasible, within the framework provided by ABC methods.

In the case of Sardinia, our approach could reconstruct, and highlight the consequences of, a complex scenario in which two geographically close populations evolved under the effects of different factors. We showed that, when properly analyzed, a few tens ancient sequences are sufficient to test hypotheses on the relationships between past and modern people and to distinguish between the effects of isolation and those of even limited rates of gene flow from an external source.

Supplementary Material

Supplementary tables S1, S2, and S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Acknowledgments

This study was supported by the Italian Ministry for Universities (MIUR) Funds (PRIN 2006) to G.B.; S.G. is supported by funds (Programma Ricerca Regione Università 2007-2009) of Regione Emilia-Romagna. We thank Robert Tykot and Alessio Fonnesu for several useful informations on the Sardinian paleontological and archeological records, Enza Colonna for her help with preliminary analyses, and Giorgio Bertorelle for critical reading of the manuscript.

References

Achilli A, Olivieri A, Pala M, et al. (22 co-authors). 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.

Angius A, Melis PM, Morelli L, Petretto E, Casu G, Maestrale GB, Fraumene C, Bebbere D, Forabosco P, Pirastu M. 2001. Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. *Hum Genet.* 109:198–209.

Auton A, Bryc K, Boyko AR, et al. (13 co-authors). 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* 19:795–803.

Babalini C, Martinez-Labarga C, Tolk HV, et al. (16 co-authors). 2005. The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. *Eur J Hum Genet.* 13:902–912.

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

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

Beaumont M. 2008. Joint determination of topology, divergence time and immigration in population trees. In: Matsumura S, Forster P, Renfrew C, editors. Simulations, genetics, and human prehistory. Cambridge (UK): McDonald Institute for Archaeological Research, Cambridge. p. 135–154.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

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.

Bramanti B, Thomas MG, Haak W, et al. (16 co-authors). 2009. Genetic discontinuity between local hunter-gatherers and Central Europe's first farmers. *Science* 326:137–140.

Caramelli D, Vernesi C, Sanna S, et al. (15 co-authors). 2007. Genetic variation in prehistoric Sardinia. *Hum Genet.* 122:327–336.

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

Cela-Conde CJ, Ayala F. 2007. Human evolution. Trails from the past. Oxford: Oxford University Press, p. 310.

Chikhi L, Bruford MW, Beaumont MA. 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158:1347–1362.

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

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.

Currat M, Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol.* 2:e421.

Dupanloup I, Pereira L, Bertorelle G, Calafell F, Prata MJ, Amorim A, Barbujani G. 2003. A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *J Mol Evol.* 57:85–97.

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

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.

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

Francalacci P, Morelli L, Underhill PA, et al. (17 co-authors). 2003. Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. *Am J Phys Anthropol.* 121:270–279.

Fraumene C, Petretto E, Angius A, Pirastu M. 2003. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. *Hum Genet.* 114:1–10.

Gauniyal M, Chahal SM, Kshatriya GK. 2008. Genetic affinities of the Siddis of South India: an emigrant population of East Africa. *Hum Biol.* 80:251–270.

Gelman A, Carlin JS, Rubin DB. 2004. Bayesian data analysis. Boca Raton (FL): CRC Press.

Guimaraes S, Ghirotto S, Benazzo A, et al. (15 co-authors). 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.

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.

Hey J. 2006. Recent advances in assessing gene flow between diverging populations and species. *Curr Opin Genet Dev.* 16:592–596.

Hill C, Soares P, Mormina M, et al. (11 co-authors). 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.

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent RJ, Suarkia D, Schiefenhovel W, Stoneking M. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol.* 25:1362–1374.

Lai CQ, Tucker KL, Choudhry S, Parnell LD, Mattei J, Garcia-Bailo B, Beckman K, Burchard EG, Ordovas JM. 2009. Population admixture associated with disease prevalence in the Boston Puerto Rican health study. *Hum Genet.* 125:199–209.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201:786–792.

Miller N, Estoup A, Toepfer S, Bourguet D, Lapchin L, Derridj S, Kim KS, Reynaud P, Furlan L, Guillemaud T. 2005. Multiple transatlantic introductions of the western corn rootworm. *Science* 310:992.

Morelli L, Grosso MG, Vona G, Varesi L, Torroni A, Francalacci P. 2000. Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. *Hum Biol.* 72:585–595.

Navascués M, Emerson BC. 2009. Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. *Mol Ecol.* 18:4390–4397.

Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L. 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.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet.* 39:197–218.

Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol.* 18:1034–1047.

Noonan JP, Coop G, Kudaravalli S, et al. (11 co-authors). 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113–1118.

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

Pakendorf B, Stoneking M. 2005. Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet.* 6:165–183.

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.

Pugliatti M, Rosati G, Carton H, Riise T, Drulovic J, Vecsei L, Milanov I. 2006. The epidemiology of multiple sclerosis in Europe. *Eur J Neurol*. 13:700–722.

Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES. 2003. Genetic structure of Mediterranean populations revealed by Y-chromosome haplotype analysis. *Am J Phys Anthropol*. 121:157–171.

Shriner D, Liu Y, Nickle DC, Mullins JI. 2006. Evolution of intrahost HIV-1 genetic diversity during chronic infection. *Evolution* 60:1165–1176.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet*. 84:740–759.

Sokal RR, Oden NL, Wilson C. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* 351:143–145.

Sousa VC, Fritz M, Beaumont MA, Chikhi L. 2009. Approximate bayesian computation without summary statistics: the case of admixture. *Genetics* 181:1507–1519.

Tanaka MM, Francis AR, Luciani F, Sisson SA. 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* 173:1511–1520.

Tishkoff SA, Gonder MK. 2007. Human origins within and out of Africa. In: Crawford M, editor. Anthropological genetics. Cambridge (UK): Cambridge University Press. p. 358.

Verdu P, Austerlitz F, Estoup A, et al. (14 co-authors). 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol*. 19:312–318.

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

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci USA*. 102:18508–18513.

Vona G. 1997. The peopling of Sardinia (Italy): history and effects. *Int J Anthropol*. 12:71–87.

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

Wilkins JF. 2006. Unraveling male and female histories from human genetic data. *Curr Opin Genet Dev*. 16:611–617.

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

Supplemental Table 1. List of nucleotide substitutions in the ancient dataset.

	1111111111	
	6666666666	
	0011112223	
	5622892791	
	1969933881	
ANDERSON	ACTGTCCCTT	
(AL07)	....C.....	(Alghero)
(CA02)	....	(Carbonia)
(CA14)	G.........	(Carbonia)
(FL04)	...C......	(Fluminimaggiore)
(PE11)	....	(Perdasdefogu)
(PE15)	...C....C.	(Perdasdefogu)
(PE20)	....	(Perdasdefogu)
(PE23)	....	(Perdasdefogu)
(PE25)	......T...	(Perdasdefogu)
(SE01)	..C.......	(Seulo)
(SE02)	.TC..T....	(Seulo)
(SE13)	G.........	(Seulo)
(SE60)	.......T..	(Seulo)
(SE81)	....	(Seulo)
(SE84)	....	(Seulo)
(ST08)	......T...	(STeresa di Gallura)
(ST10)	....	(STeresa di Gallura)
(ST15)	...C......	(STeresa di Gallura)
(ST16)	C	(STeresa di Gallura)
(ST30)	..C.......	(STeresa di Gallura)
(ST38)	....	(STeresa di Gallura)
(ST47)	....	(STeresa di Gallura)
(ST54)	...C......	(STeresa di Gallura)

Supplemental Table 2. Parameter's Bias, Rmse, Coverage (50%) and Factor2

MODEL 1	**BIAS**	**RMSE**	**Cov (50%)**	**Factor2**
N_e Ogliastra	0.555	1.269	0.633	0.815
N_e Gallura	-0.177	0.262	0.887	0.932
Separation Time	0.343	0.885	0.875	0.992
Mutation Rate	-0.180	0.390	0.552	0.977
Ancestral N_e Ogliastra	0.968	1.832	0.489	0.605
Ancestral N_e Gallura	0.868	1.421	0.564	0.626

MODEL 4	**BIAS**	**RMSE**	**Cov (50%)**	**Factor2**
N_e Ogliastra	0.564	1.077	0.52	0.642
N_e Gallura	-0.161	0.387	0.991	0.990
Migration Rate from Latium to Gallura	0.143	0.225	0.978	0.998
Separation Time (Sardinia)	-0.099	0.187	0.839	0.986
Mutation Rate	-0.118	0.227	0.425	0.996
Ancestral N_e Ogliastra	0.229	0.753	0.624	0.738
Ancestral N_e Gallura	1.376	1.712	0.439	0.344
Ancestral N_e Latium	0.480	0.840	0.365	0.692

Supplemental Table 3. Posterior predictive test for all models analysed.
Values showing significant departures from the data are shown in bold type. The Latium sample is not included in models 1 to 3.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
[ξ]C	0.341	0.403	0.383	0.429	0.365	0.410	0.337
Hap_numb_O	0.395	0.392	0.310	0.419	0.434	0.460	0.453
seg_site_O	0.213	0.231	0.102	0.208	0.278	0.248	0.260
pair_diff_O	0.261	0.288	0.169	0.253	0.321	0.253	0.312
hap_div_O	0.315	0.429	0.236	0.303	0.431	0.410	0.348
D_taj_O	0.353	0.399	0.214	0.200	0.366	0.155	0.341
Hap_numb_G	0.342	0.315	0.147	0.347	0.231	0.243	0.221
seg_site_G	0.396	0.404	0.350	0.262	0.337	0.267	0.461
pair_diff_G	0.300	0.296	0.212	0.247	0.286	0.244	0.361
hap_div_G	0.206	0.211	0.134	0.179	0.171	0.150	0.159
D_taj_G	0.194	0.174	0.056	0.126	0.149	0.117	0.264
Hap_numb_L	-	-	-	0.428	0.454	0.446	0.425
seg_site_L	-	-	-	0.277	0.321	0.274	0.396
pair_diff_L	-	-	-	0.241	0.270	0.268	0.323
hap_div_L	-	-	-	0.372	0.370	0.387	0.325
D_taj_L	-	-	-	0.126	0.174	0.218	0.270
Hap_numb_A	0.482	0.218	0.391	0.477	0.464	0.439	0.456
seg_site_A	0.252	0.160	0.137	0.229	0.236	0.230	0.300
pair_diff_A	0.216	0.170	0.123	0.235	0.247	0.230	0.279
hap_div_A	0.343	0.453	0.494	0.383	0.314	0.397	0.328
D_taj_A	0.084	0.103	**0.018**	**0.047**	0.099	**0.020**	0.090
Fst_Hud_O/G	0.110	0.118	0.216	0.127	0.138	0.146	0.148
Hap_Shar OG/G	0.423	0.485	0.116	0.405	0.291	0.195	0.462
Hap_Shar OA/A	0.055	**0.050**	0.377	**0.017**	0.460	0.254	**0.037**
Hap_Shar GA/G	0.386	0.072	**0.032**	0.384	**0.037**	0.136	0.299

[ξ]Hap_numb = haplotype number; seg_site = number of segregating sites; pair_diff = pairwise sequence difference; hap_div = haplotype diversity; D_taj = Tajima's D; Fst_Hud = Hudson's Fst; HapShar = Haplotype sharing. Significant values (at $P<0.05$) are in boldtype.
O = Ogliastra; G = Gallura; L = Latium; A = ancient sample.

Molecular Biology and Evolution

OXFORD JOURNALS
MOLECULAR BIOLOGY AND EVOLUTION

Origins and evolution of the Etruscans' DNA

Journal:	*Molecular Biology and Evolution*
Manuscript ID:	Draft
Manuscript Type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Ghirotto, Silvia; University of Ferrara, Biology & Evolution Tassi, Francesca; University of Ferrara, Biology & Evolution Fumagalli, Erica; University of Ferrara, Biology & Evolution Colonna, Vincenza; University of Ferrara, Biology & Evolution Sandionigi, Anna; University of Florence, Evolutionary Biology Lari, Martina; University of Florence, Evolutionary Biology Vai, Stefania; University of Florence, Evolutionary Biology Petiti, Emmanuele; University of Florence, Evolutionary Biology Corti, Giorgio; National research Council, Institute for Biomedical Technologies Rizzi, Ermanno; National research Council, Institute for Biomedical Technologies De Bellis, Gianluca; National research Council, Institute for Biomedical Technologies Caramelli, David; University of Florence, Evolutionary Biology Barbujani, Guido; University of Ferrara, Biology & Evolution
Key Words:	ancient DNA, mitochondrial DNA, coalescent simulations, approximate Bayesian computations

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. To better understand the biological relationships between contemporary and ancient populations, we 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 other Etruscan sequences (Vernesi et al. 2004) we compared them with those of relevant ancient and modern human populations, namely Medieval Tuscans (Guimaraes et al. 2009), contemporary Tuscans from three sites in historical Etruria (Casentino, Murlo, Volterra) (Achilli et al. 2007) and from Florence (Turchi et al. 2008) (fig. 1), and Southwestern Anatolians (Di Benedetto et al. 2001). Once established that genealogical ties with the Etruscans are still present in some regions of Tuscany, we estimated the separation time between these Tuscan populations and a population from Southwestern Anatolia, 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.

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 bone fragments were freshly excavated and collected according to the most stringent ancient DNA criteria (Caramelli et al.

2006) 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 (Caramelli et al. 2008). 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 a single stranded library (Maricic & Paabo 2009). 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 (Andrews et al. 1999). 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.

Datasets of ancient and modern mtDNA 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 (Vernesi et al. 2004). 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 (fig. 1). In addition, this dataset includes a sample of Medieval Tuscans from Guimaraes et al. (2009).

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 (Bendall & Sykes 1995; Bandelt & Kivisild 2006). 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 (R Development Core Team 2010).

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; Beaumont, Zhang & Balding 2002) 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 (Anderson et al. 2005; available at http://iod.ucsd.edu/simplex/ssc/BayeSSc.htm). 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 (R Development Core Team 2010) using scripts from http://www.rubic.rdg.ac.uk/~mab/stuff/.

Demographic models and Priors

The three demographic models tested differ for the relationships between modern and ancient samples (fig. 2); 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 (Pakendorf & Stoneking 2005), and the high value estimated from pedigrees (Howell et al. 2003), 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. (2009). 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 (Biraben 1979).

For each modern population considered, the analysis included four steps, namely: (i) one million of gene genealogies were generated for each model and each demographic scenario (with or without bottleneck) by serial coalescent simulation; (ii) we summarized genetic diversity in the observed and simulated data by the same set of statistics (table 2); (iii) by comparing these statistics in the observed and simulated data, we selected a set (100 or 50,000, depending on the criterion chosen, see below) of simulations best reproducing variation in the data and we estimated the models' posterior probabilities (*PP*); (iv) demographic (population sizes) and evolutionary (mutation rates) parameters for the most probable model were finally estimated from the simulated data (Beaumont, Zhang & Balding 2002).

Model selection and parameter estimation

The posterior probabilities of the 24 combinations of models (3), modern populations (2) and demographic scenarios (4), were calculated either: (i) by a simple rejection procedure (AR) (Pritchard et al. 1999) for which we retained the 100 simulations associated with the shortest distance between observed and simulated statistics (Beaumont 2008); or (ii) by a weighted multinomial logistic regression (LR) (Beaumont 2008) 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 (Hamilton, Stoneking & Excoffier 2005) and according to Beaumont, Zhang, Balding (2002).

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 (Gelman et al. 2004; Ghirotto et al. 2010). 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 (Voight et al. 2005).

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 (Hey & Nielsen 2004). 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) (Hey 2005). 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; Henn et al. 2009; Soares et al. 2009) under a Hasegawa-Kishino-Yano (HKY; Hasegawa, Kishino & Yano 1985) mutational model. 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.

Results

Ancient DNA sequences

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

mtDNA HVR-I region, whereas the remaining eight gave no results (supplementary figure 1, Supplementary Material Online). Their final consensus sequences (supplementary table 1) were determined by comparing results obtained using the standard procedures (575 clones overall) and Next Generation Sequencing (127,837 reads) (supplementary fig. 2). 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

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±1.56) and by haplotype diversity (0.943±0.032), appear close to those obtained in Vernesi et al. (2004) 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 also observed in the Etruscan sample (supplementary fig. 3). A general decline of genetic resemblance with geographic distance is evident (fig. 3).

Among the 30 Etruscan individuals (ETR dataset) we observed 21 different sequences with 24 variable sites (table 2). Comparisons with 52 modern populations in the TUS and EUR datasets (listed in supplementary table 2) show that 11 of these sequences are shared with at least one of 4,910 individuals from Western Eurasia and the Southern Mediterranean shore (supplementary

table 1). The Etruscan sample falls within the range of contemporary genetic variation (EUR dataset, supplementary fig. 4 a and b). 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 (supplementary fig. 4c).

Genealogical relationships between the Etruscans and contemporary populations

We found evidence for genealogical continuity all the way from Etruscan to current times in two contemporary populations (fig. 2a); the *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 obtain incorporating in the model a population bottleneck at the time of the Medieval plague epidemics (Livi-Bacci 2007) (supplementary fig. 5), although an explicit comparison between models with and without plague favoured the latter (fig. 2b). Therefore, this event was not considered in subsequent analyses. At any rate, the relative success of the models does not depend on the presence of a bottleneck in the late Middle Age.

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) (fig. 2a); 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 (Models1-3), with parameters values randomly chosen from the correspondent prior distribution. Type I error, namely the fraction of cases in which the model generating these 12,000 pseudo-observed datasets was not recognized, was always ≤ 0.08. In particular, 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 (supplementary fig. 6). The estimated mutation rate (around 0.3 mutational events per million years per nucleotide) is in agreement with previous independent reports (Henn et al. 2009; Ghirotto et al. 2010). In general, all the parameters appear well estimated; indeed, their R^2 value are always higher than 0.1, an empirical figure generally accepted to be the value beyond which an estimate may be considered reliable (Neuenschwander et al. 2008). We note that the posterior distribution of the modern effective population sizes drives to the upper limit of the priors (supplementary fig. 6). This has also been observed in previous comparable studies (Fagundes et al. 2007; Belle et al. 2009; Laval et al. 2010) and reflects the fact that 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 Herodotous, (Achilli et al. 2007; Pellecchia et al. 2007; Brisighelli et al. 2009), we would expect that the two populations separated around 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 (Nielsen & Wakeley 2001; Hey & Nielsen 2004). Under the model that we tested, the two populations originate from a common ancestor, and may or may not exchange migrants after the split (supplementary fig. 7a). Assuming an average generation time of 25 years (Fenner 2005; Fagundes et al. 2007) and no migration after the split from the common ancestors, the most likely separation time between Tuscany and Anatolia falls around 7,600 years ago, with a 95% credible interval between 5,000 and 10,000 (fig. 4). These results are robust to changes in the proportion of members of the initial population being ancestral to the two modern populations (supplementary fig. 7b). For these tests we chose the mutation rate estimated from the data in the previous ABC analyses (very close to the figure accounting for the time-dependency of the mitochondrial molecular clock (Henn et al. 2009) ($\mu = 0.003$). Tests were also run using the value incorporating a correction for the effects of purifying selection (Soares et al. 2009) ($\mu = 0.0014$), always finding that it results in a further increase of the estimated separation times

(supplementary fig.7b). Only assuming an implausibly high mutation rate, twice as large as estimated in Henn et al. (2009), was it possible to obtain separation times <5,000 years (supplementary fig 7b). 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, characterised by plausible population sizes and mutation rates. However, this applies to Volterra, and especially Casentino, but not to other communities dwelling in areas rich with Etruscan archaeological remains (Murlo), nor to the bulk of the current Tuscan population, here represented by a forensic sample of the inhabitants of Florence. The IM analysis shows that there might have been a genealogical link between modern Tuscans and the inhabitants of what Herodotus considered the Etruscans' homeland, Anatolia. However, that link does not suggest an oriental origin for the Etruscans, because, even under the unrealistic assumption of complete reciprocal isolation for millennia between Tuscany and Anatolia, the likely separation of the two gene pools must be placed long before the onset of the Etruscan culture.

There are several reasons to be confident that Etruscan sequences are authentic. As for the ones typed in this study: (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 (supplementary table 1); (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 ≥ 5bp 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.

ABC and other recent Bayesian inference methods are making it possible to test complex evolutionary models against genetic data (Gelman et al. 2004; Bertorelle, Benazzo & Mona 2010). 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 we could not incorporate 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.

As also suggested by the analysis of skull diversity (Claassen & Wree 2004), 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 (Barker 2006; Lacan et al. 2011), and do not appear related with the onset of the Etruscan culture (fig.4). 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 evident only in a few relatively isolated localities, such as Casentino and Volterra.

ACKNOWLEDGMENTS. This study was supported by the Italian Ministry for Universities (MIUR) 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. 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.

References

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. 80:759-768.

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.

Bandelt HJ. 2004. Etruscan artifacts. Am.J.Hum.Genet. 75:919-920; author reply 923-917.

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.

Barker G. 2006. The Agricultural revolution in prehistory: Why did foragers become farmers?. Oxford: Oxford University Press.

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

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

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.

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.

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.

Biraben J-N. 1979. Essai sur l' evolution du nombre des hommes. Population 34:13-25.

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.

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

Caramelli D, Milani L, Vai S, et al. 2008. A 28,000 years old Cro-Magnon mtDNA sequence differs from all potentially contaminating modern sequences. PLoS One 3:e2700.

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.

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.

Fagundes NJ, Ray N, Beaumont MA, 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.

Gelman A, Carlin J, Stern H, Rubin D. 2004. Bayesian Data Analysis. Boca Raton, Florida: 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.

Guimaraes S, Ghirotto S, Benazzo A, et al. 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.

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.

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.

Hey J. 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol. 3:e193.

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.

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C. 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.

Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, Guilaine J, Crubezy E, Ludes B. 2011. Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc. Natl. Acad. Sci. USA 108:9788-9791.

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.

Livi-Bacci M. 2007. A concise history of world population. Oxford: Blackwell.

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.

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

Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L. 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.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885-896.

Pakendorf B, Stoneking M. 2005. Mitochondrial DNA and human evolution. Annu. Rev. Genomics. Hum. Genet. 6:165-183.

Pellecchia M, Negrini R, Colli L, et al. 2007. The mystery of Etruscan origins: novel clues from Bos taurus mitochondrial DNA. Proc. Biol. Sci. 274:1175-1179.

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.

R Development Core Team. R: A Language and Environment for Statistical Computing. 2010. R Foundation for Statistical Computing, Vienna, Austria.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. 84:740-759.

Turchi C, Buscemi L, Previdere C, Grignani P, Brandstatter A, Achilli A, Parson W, Tagliabracci A. 2008. Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int. J. Legal. Med. 122:199-204.

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

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. USA 102:18508-18513.

Figure Legends

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).

Figure 2. 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 (Beaumont 2008). (*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 (Livi-Bacci 2007).

Figure 3. Genetic distances (percent F_{st} values) between the Etruscan and modern population samples. Different colors represent different levels of genetic differentiation from the Etruscans.

The map was obtained using ArcGIS v10 (ESRI; Redlands, CA, USA) with the Kriging interpolation procedure.

Figure 4. Separation time between the gene pools of Southwestern Anatolians and contemporary Tuscans (Casentino and Volterra) estimated by the IM model. Means, upper bound and lower bound of the 95% credible intervals in 4 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.

Tables

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

Table 2. Statistics summarizing intra (A) and inter (B) population genetic diversity. These values were used in the ABC analysis.

A	Etruscans	Medievals	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		Etruscans	Medievals	Casentino	Murlo	Volterra	Florence
Fst	Etruscans	0.000	0.015	0.020	0.010	0.012	0.014
	Medievals	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

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.

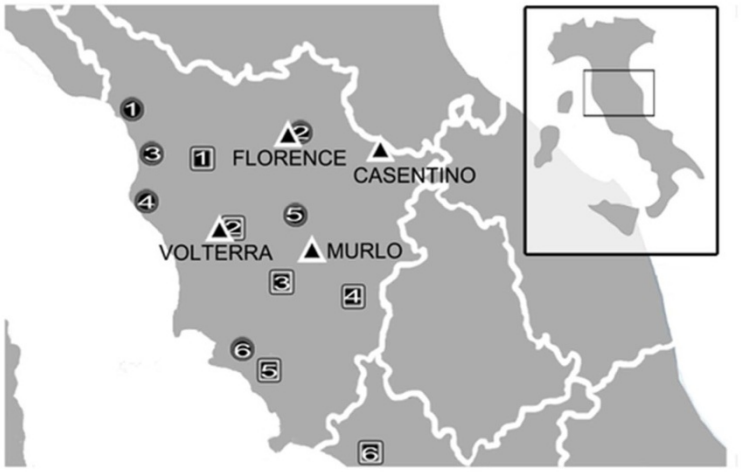
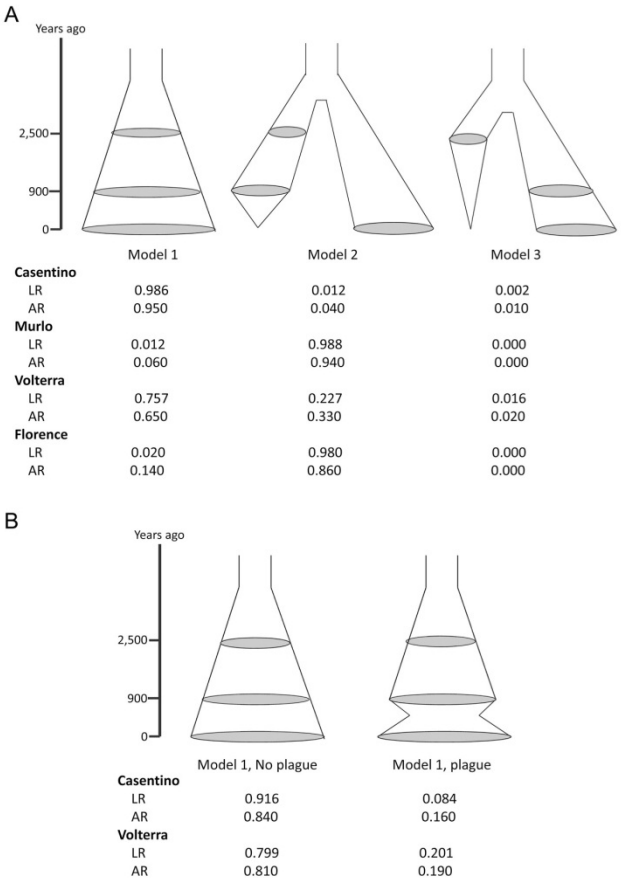


Figure 1. Geographic location of the samples considered in the ABC analysis. Triangles, Contemporary Tuscans (n=370); Circles, Medieval Tuscans: 1. Massa Carrara (n=3); 2. Florence, (n=10); 3. Pisa, (n=6); 4. Livorno, (n=3); 5. Siena, (n=4); 6. Grosseto (n=1); Squares, Etruscans: 1. Castelfranco di Sotto (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).
45x29mm (300 x 300 DPI)



	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 2. 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 (Beaumont 2008). (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 (Livi-Bacci 2007).
168x239mm (300 x 300 DPI)

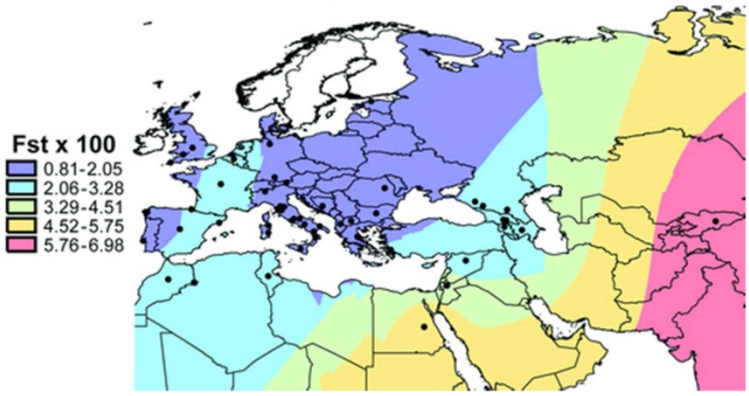
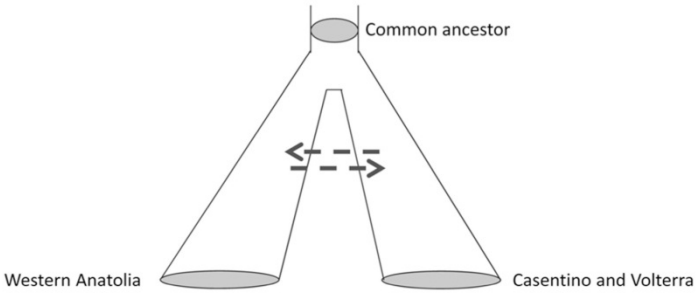


Figure 3. Genetic distances (percent Fst values) between the Etruscan and modern population samples. Different colors represent different levels of genetic differentiation from the Etruscans. The map was obtained using ArcGIS v10 (ESRI; Redlands, CA, USA) with the Kriging interpolation procedure.
42x22mm (300 x 300 DPI)



	Run 1	Run 2	Run 3	Run 4
Mean	7,775	7,980	7,663	6,692
95% lower bound	5,572	5,945	5,665	6,002
95% upper bound	9,958	10,556	10,444	9,861

Figure 4. Separation time between the gene pools of Southwestern Anatolians and contemporary Tuscans (Casentino and Volterra) estimated by the IM model. Means, upper bound and lower bound of the 95% credible intervals in 4 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.
82x49mm (300 x 300 DPI)

Supporting Figures - Legends

Supp Fig 1: Amplicons of the 10 sequences from Casenovole. DNA sequences from the 575 clones analysed for the 10 Casenovole Estruscan 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.

Supp Fig 2: 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%.

Supp Fig 3: Allele sharing (*A*) and Fst (x 100) (*B*) in 52 modern populations of Western Eurasia and the Mediterranean basin. Population labels and sample sizes are provided in supplementary table 2.

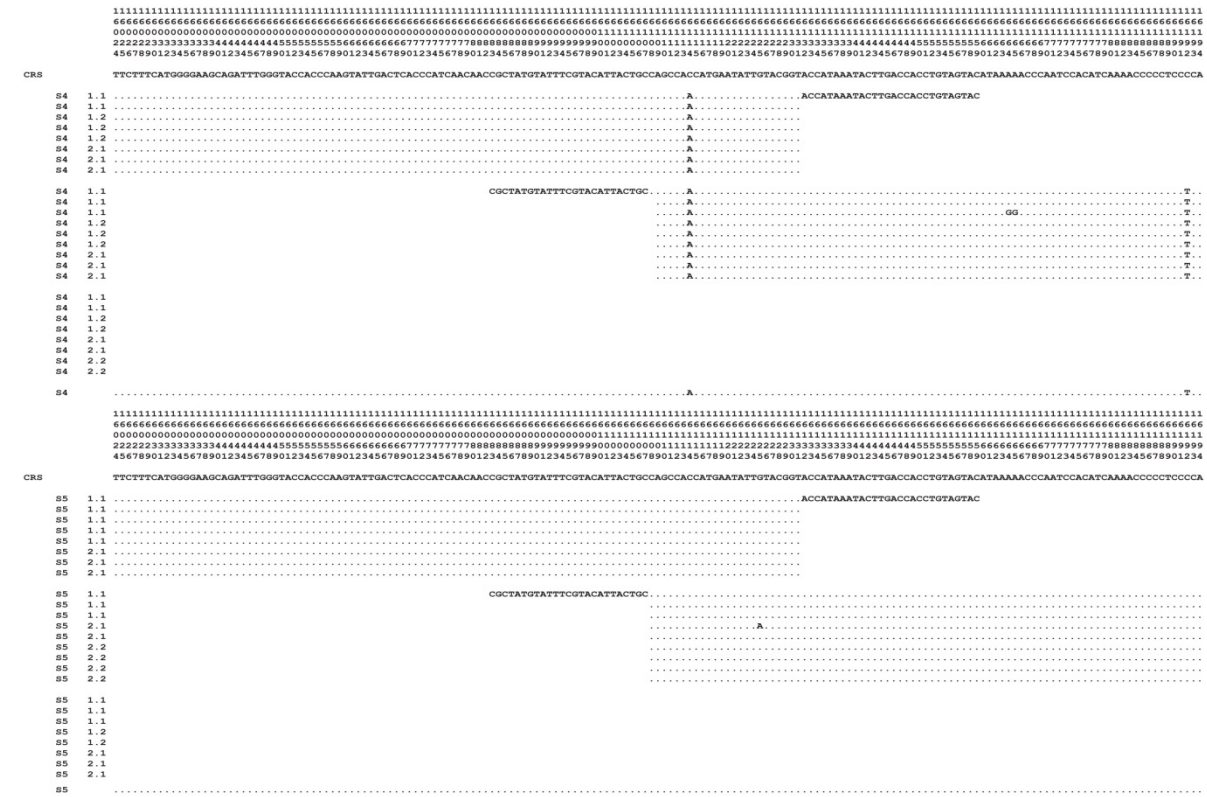
Supp Fig 4: Multi Dimensional Scaling summarizing genetic affinities between the Etruscans and (*A*) 52 modern populations of Western Eurasia and the Mediterranean basin; (*B*) Medieval and modern Italian populations; (*C*) 9 ancient populations of Europe. Population labels and sample sizes are provided in supplementary table 2.

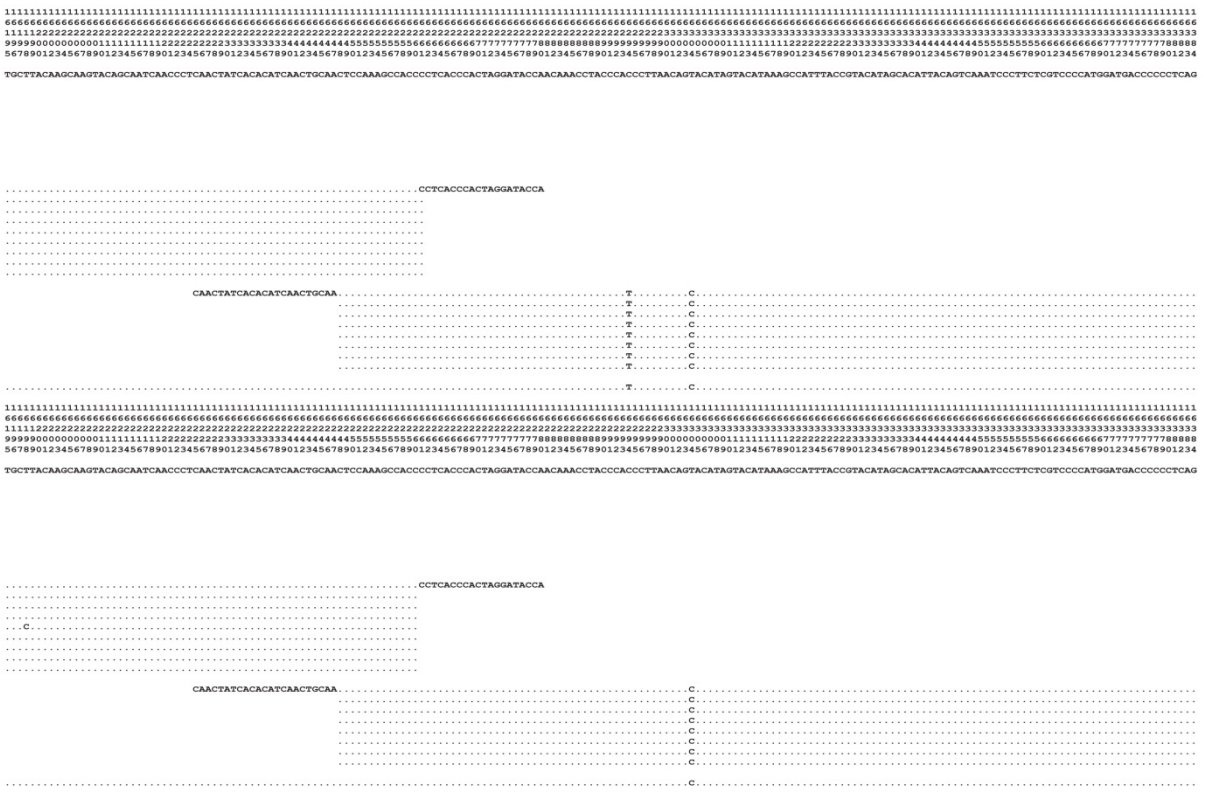
Supp Fig 5: Results of model selection, with or without a bottleneck representing the plague epidemics at 625 BP, in Casentino, Murlo and Volterra. Dashed lines represent the presence of plague epidemic that killed one third of the population. For each sample we report the posterior probabilities calculated comparing Models 1-3, either considering or disregarding this demographic event.

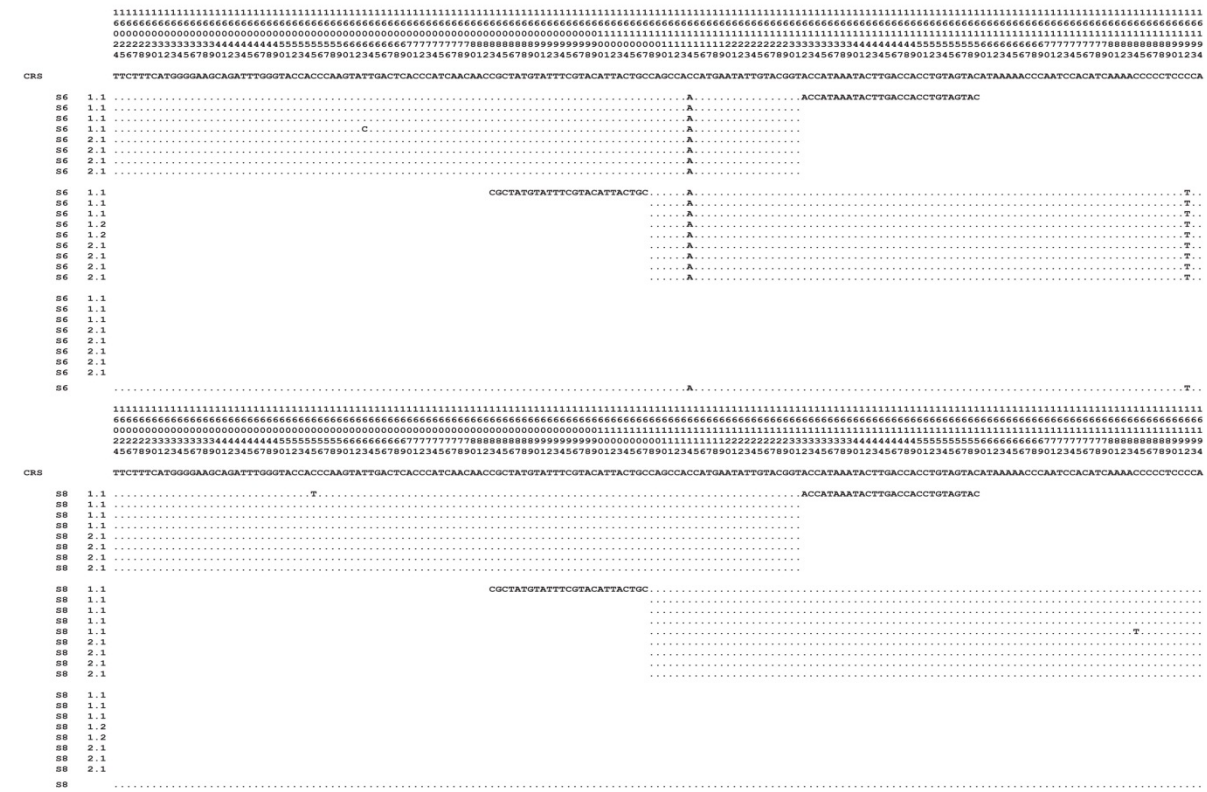
Supp Fig 6: Parameter estimates and posterior distributions under Model 1, for Casentino (*A*) and Volterra (*B*). Upper panels: Prior distributions (all the priors were uniform), median and mode estimates, the 95% of the highest posterior density (lower and upper bound), and coefficient of determination 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.

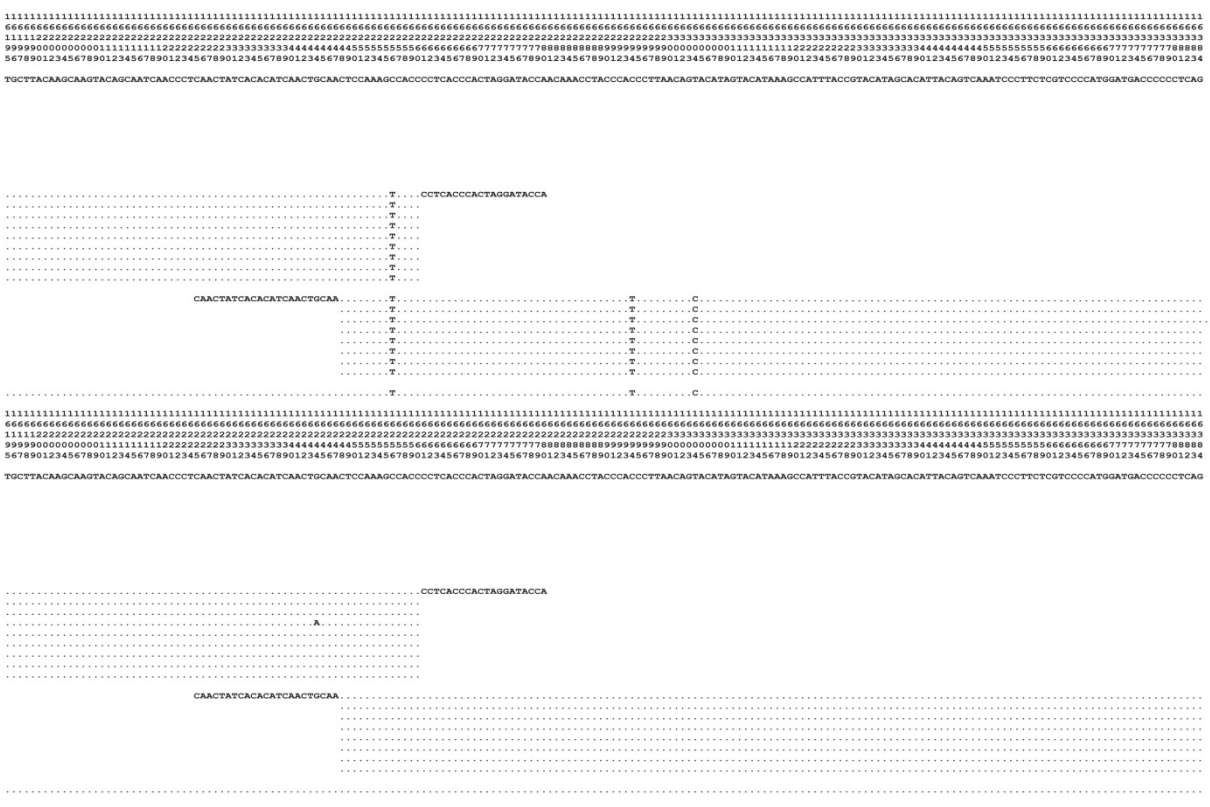
Supp Fig 7: IM model (*A*) and IM estimates (*B*) for the separation time between the Anatolian and Tuscan gene pools. Different mutation rates and proportions of the ancestral population founding the descendant populations were considered.

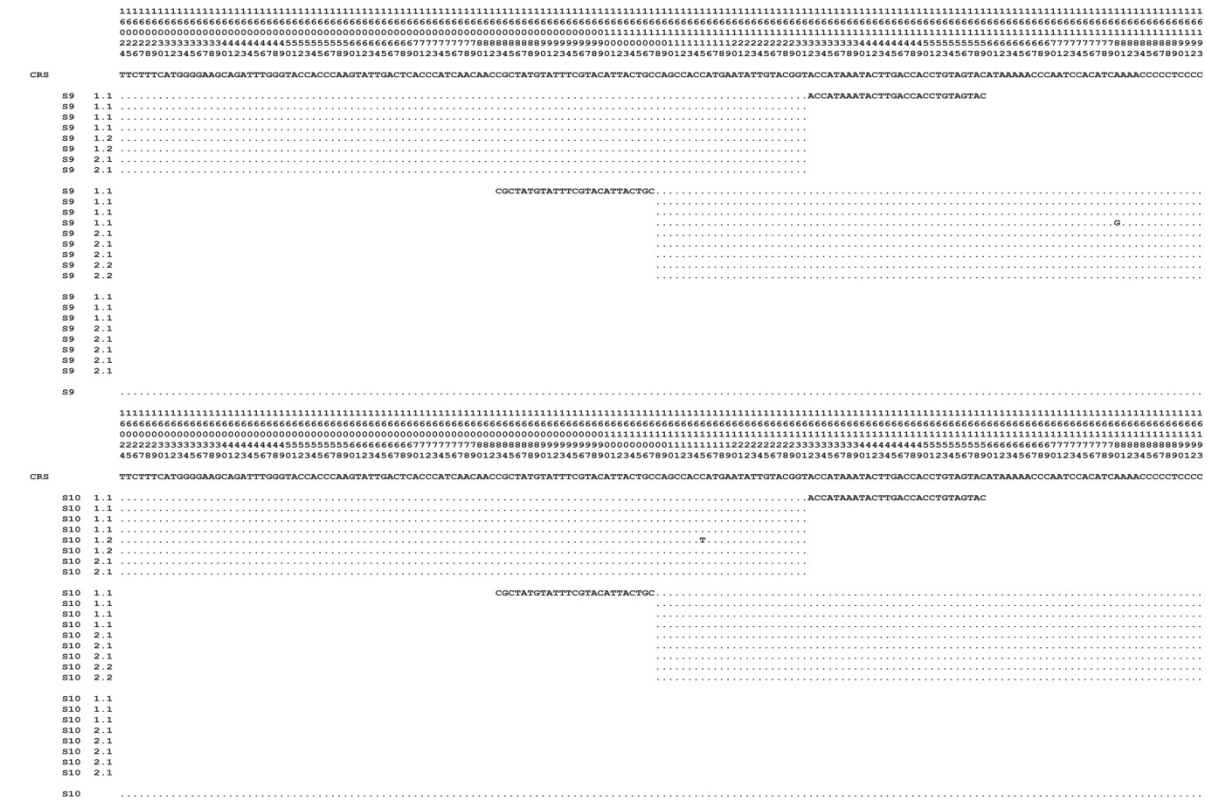
Supp Fig 1

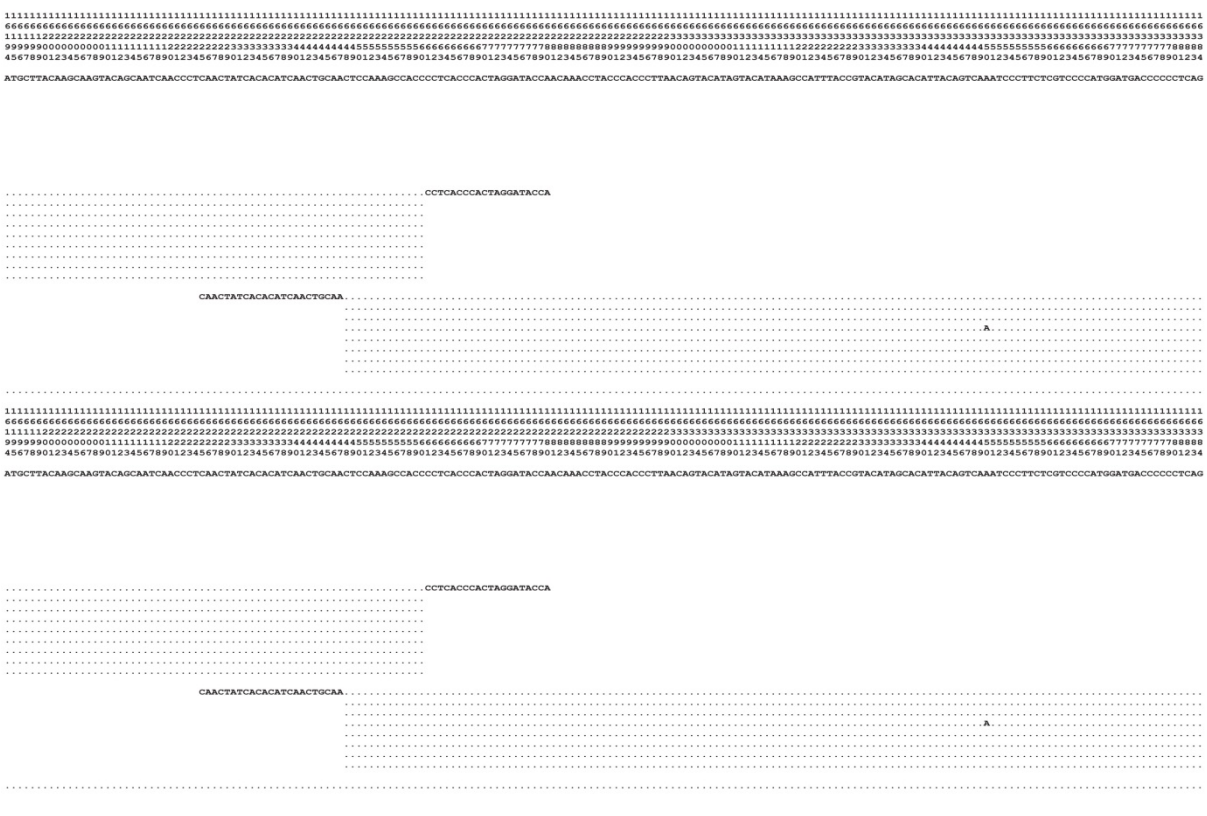


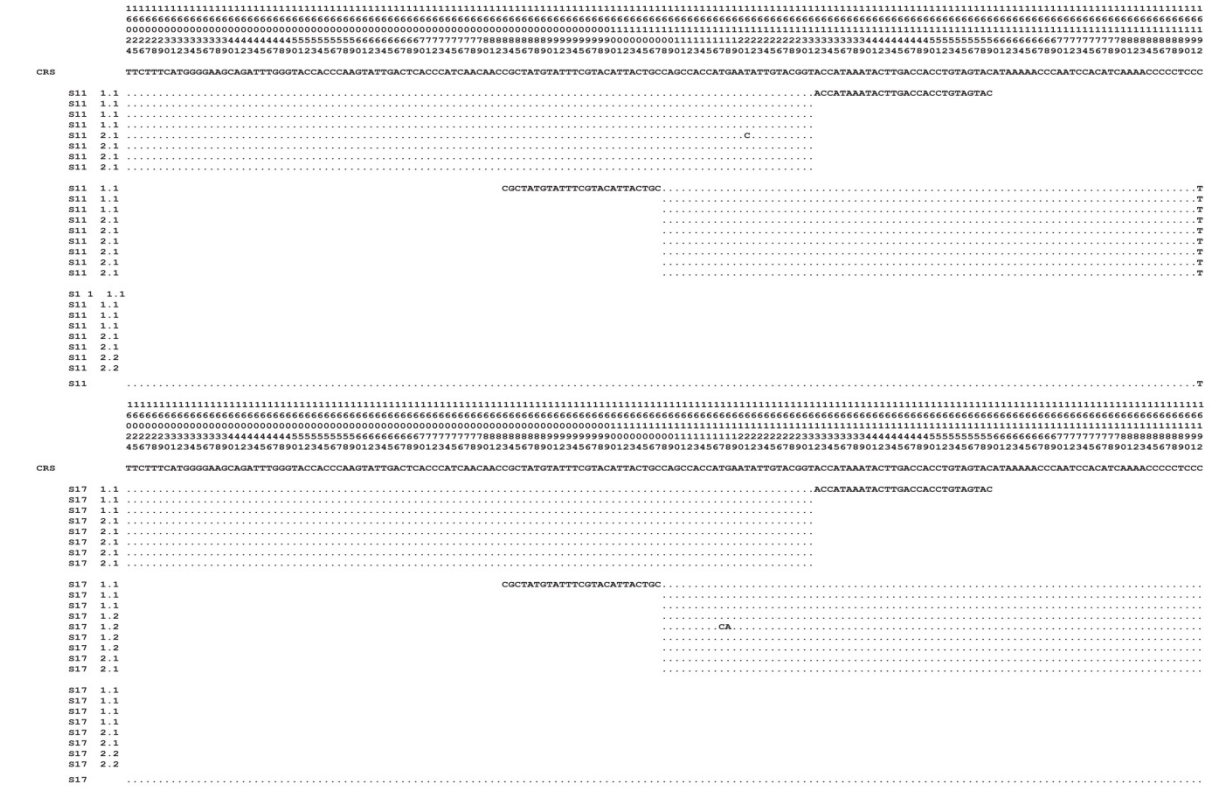


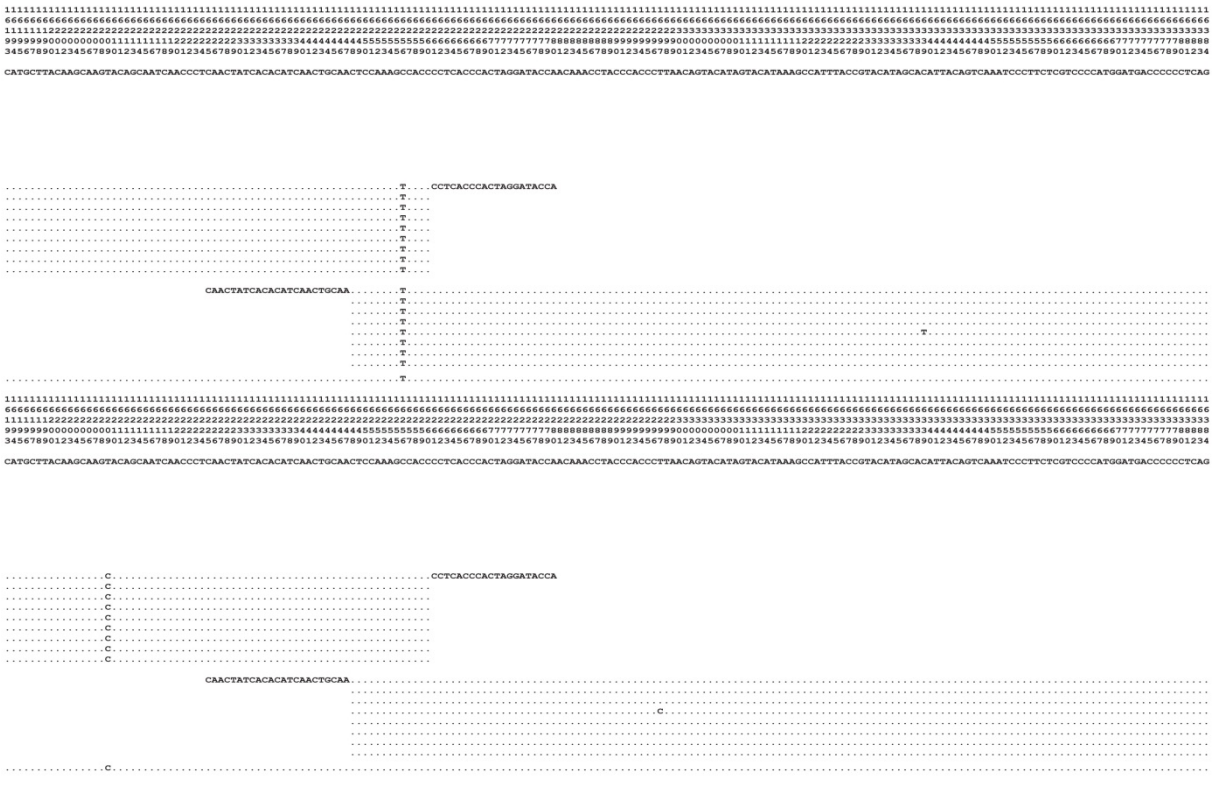






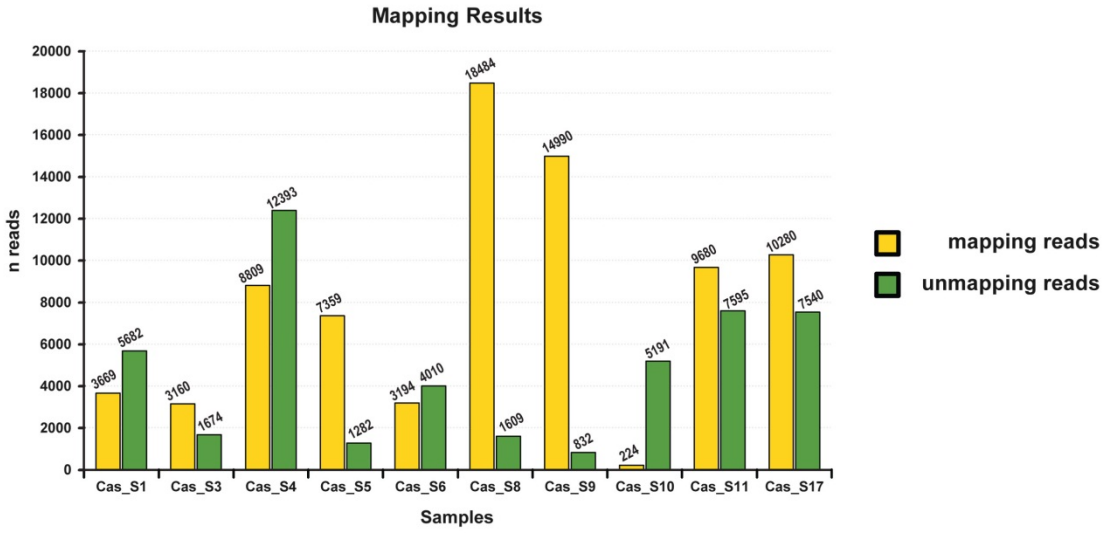




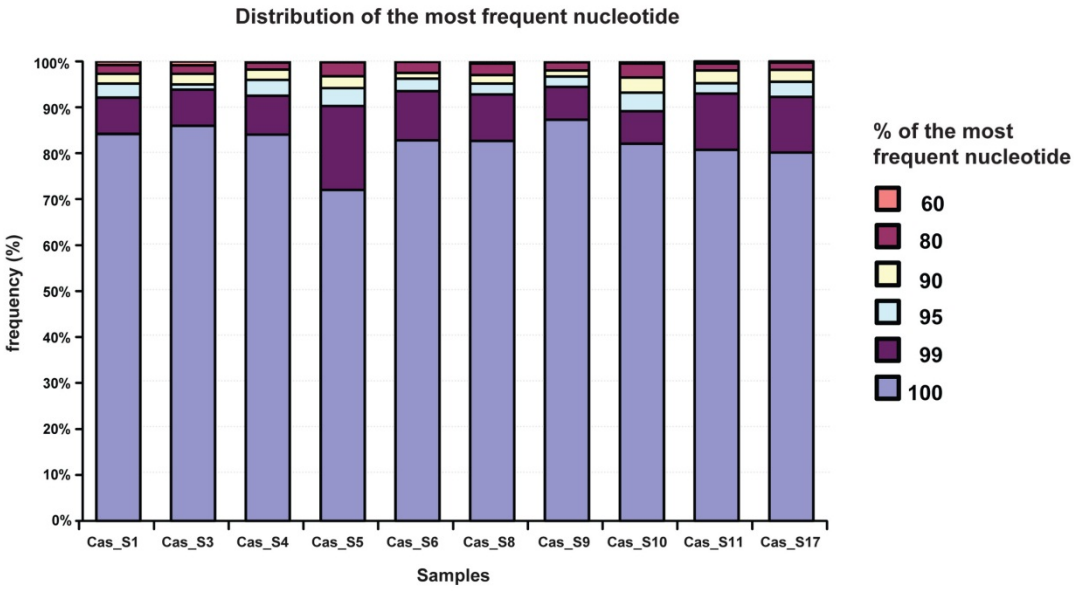


Supp Fig 2

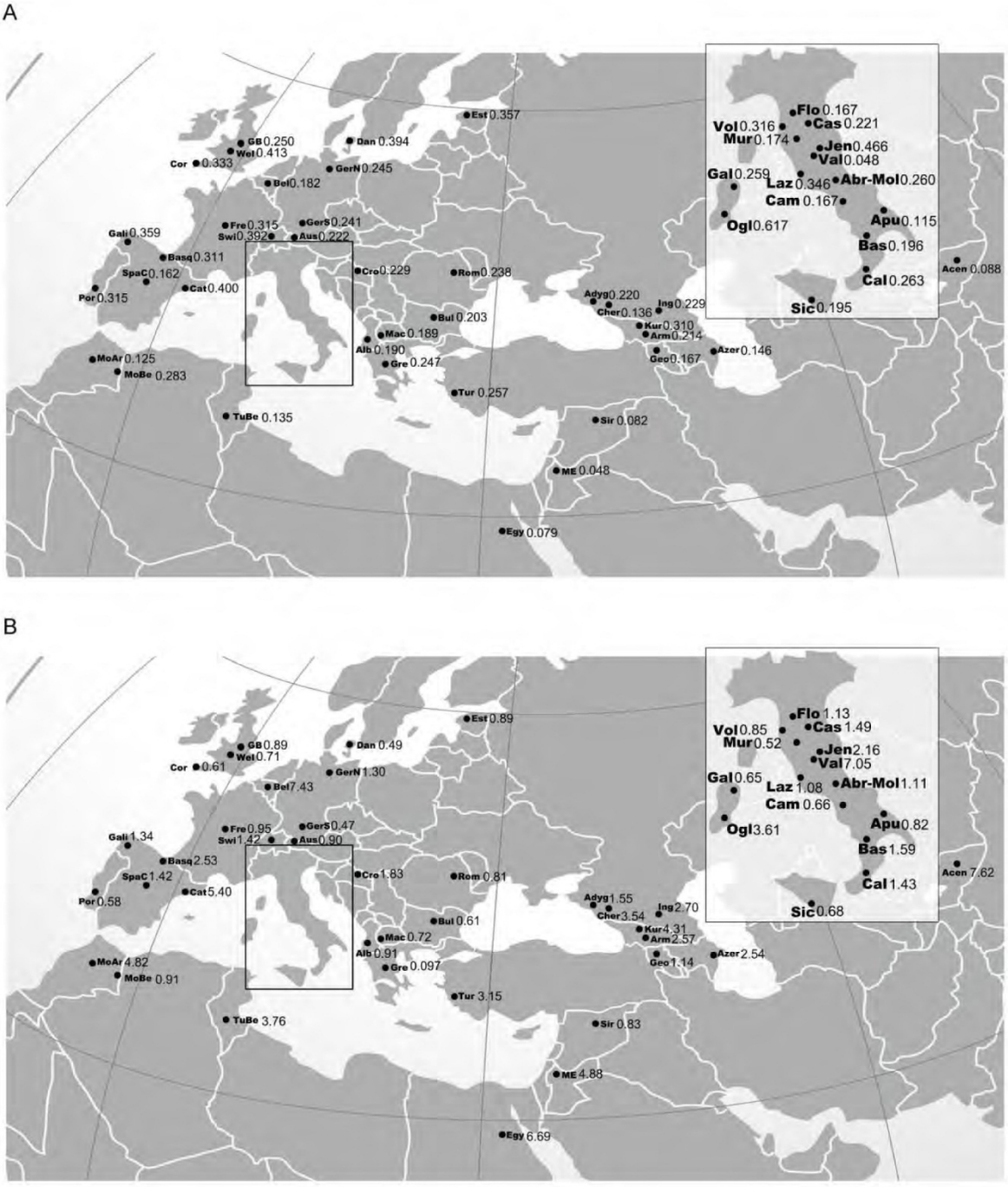
A


Mapping Results
n reads
Samples
Cas_S1
Cas_S3
Cas_S4
Cas_S5
Cas_S6
Cas_S8
Cas_S9
Cas_S10
Cas_S11
Cas_S17
3669
5682
3160
1674
8809
12393
7359
1282
3194
4010
18484
1609
14990
832
224
5191
9680
7595
10280
7540
mapping reads
unmapping reads

B

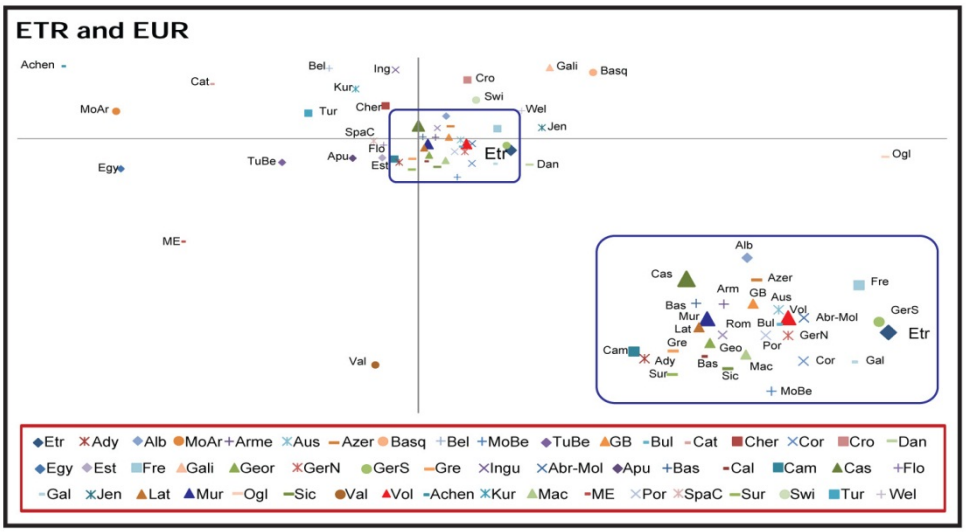

Distribution of the most frequent nucleotide
frequency (%)
Samples
Cas_S1
Cas_S3
Cas_S4
Cas_S5
Cas_S6
Cas_S8
Cas_S9
Cas_S10
Cas_S11
Cas_S17
% of the most frequent nucleotide
60
80
90
95
99
100

Supp Fig 3

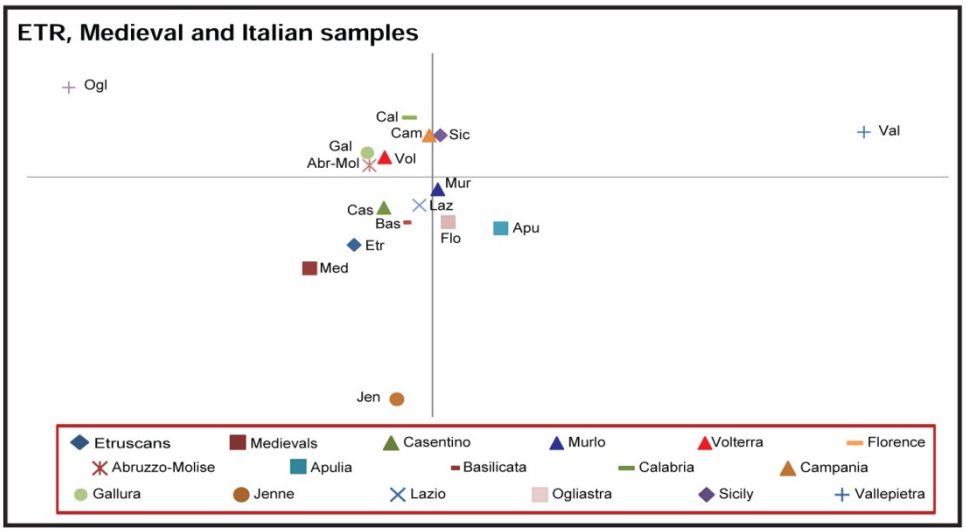

A
Est 0.357
GB 0.250
Wet 0.413
Dan 0.394
Cor 0.333
GerN 0.245
Bel 0.182
Fre 0.315
GerS 0.241
Swi 0.392
Aus 0.222
Gali 0.359
Basq 0.311
SpaC 0.162
Cat 0.400
Por 0.315
Cro 0.229
Rom 0.238
Bul 0.203
Mac 0.189
Alb 0.190
Gre 0.247
MoAr 0.125
MoBe 0.283
TuBe 0.135
Tur 0.257
Sir 0.082
ME 0.048
Egy 0.079
Adyg 0.220
Cher 0.136
Ing 0.229
Kur 0.310
Arm 0.214
Geo 0.167
Azer 0.146
Flo 0.167
Vol 0.316
Cas 0.221
Mur 0.174
Jen 0.466
Val 0.048
Gal 0.259
Laz 0.346
Abr-Mol 0.260
Cam 0.167
Ogl 0.617
Apu 0.115
Bas 0.196
Cal 0.263
Acen 0.088
Sic 0.195
B
Est 0.89
GB 0.89
Wet 0.71
Dan 0.49
Cor 0.61
GerN 1.30
Bel 7.43
Fre 0.95
GerS 0.47
Swi 1.42
Aus 0.90
Gali 1.34
Basq 2.53
SpaC 1.42
Cat 5.40
Por 0.58
Cro 1.83
Rom 0.81
Bul 0.61
Mac 0.72
Alb 0.91
Gre 0.097
MoAr 4.82
MoBe 0.91
TuBe 3.76
Tur 3.15
Sir 0.83
ME 4.88
Egy 6.69
Adyg 1.55
Cher 3.54
Ing 2.70
Kur 4.31
Arm 2.57
Geo 1.14
Azer 2.54
Flo 1.13
Vol 0.85
Cas 1.49
Mur 0.52
Jen 2.16
Val 7.05
Gal 0.65
Laz 1.08
Abr-Mol 1.11
Cam 0.66
Ogl 3.61
Apu 0.82
Bas 1.59
Cal 1.43
Acen 7.62
Sic 0.68

Supp Fig 4

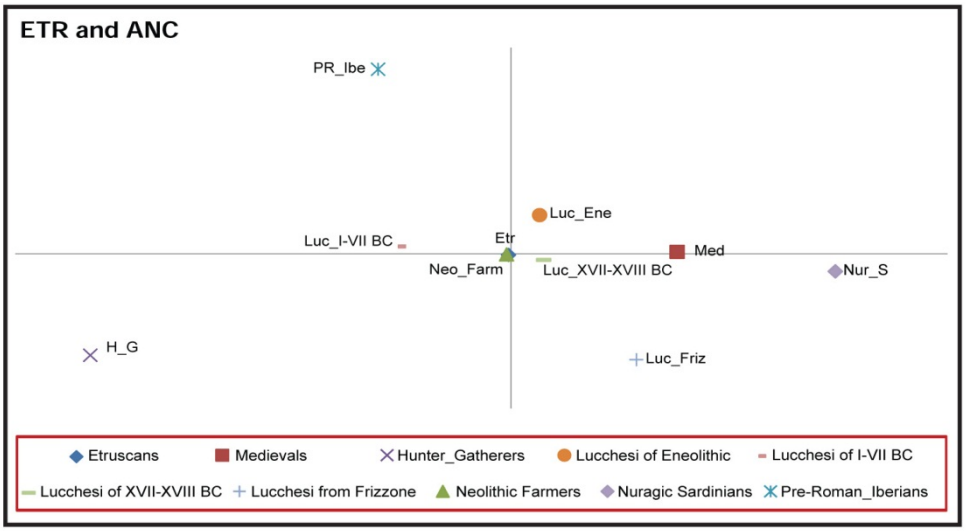
A



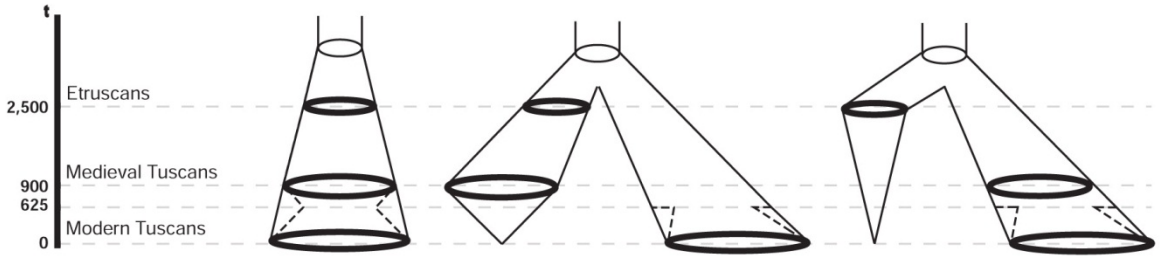
B



C



Supp Fig 5



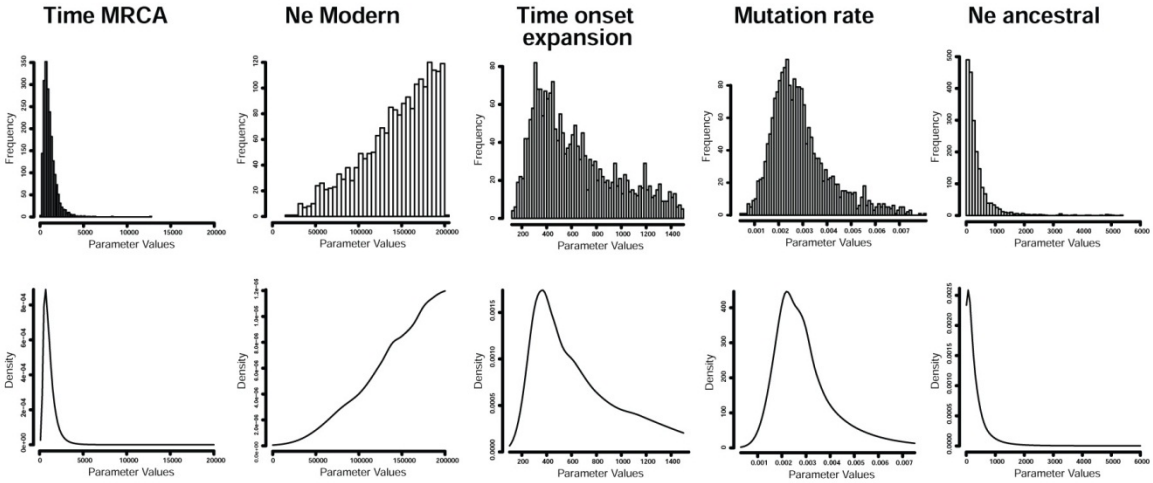
Probabilities

		MODEL 1		MODEL 2		MODEL 3	
		CASENTINO					
	N	NoPlague	Plague	NoPlague	Plague	NoPlague	Plague
LR	50,000	**0.986**	**0.998**	0.012	0.000	0.002	0.002
AR	100	**0.949**	**0.990**	0.040	0.000	0.010	0.000
		MURLO					
		NoPlague	Plague	NoPlague	Plague	NoPlague	Plague
LR	50,000	0.012	0.350	**0.988**	**0.647**	0.000	0.004
AR	100	0.060	0.360	**0.940**	**0.620**	0.000	0.020
		VOLTERRA					
	N	NoPlague	Plague	NoPlague	Plague	NoPlague	Plague
LR	50,000	**0.757**	**0.958**	0.227	0.024	0.016	0.017
AR	100	**0.650**	**0.900**	0.330	0.070	0.020	0.030

Supp Fig 6

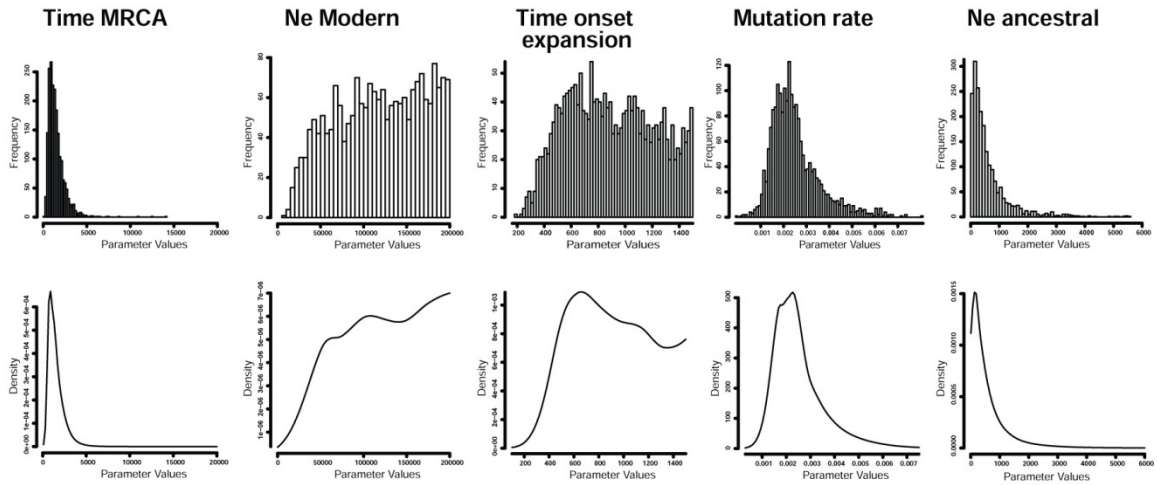
A

	Priors	Median	Mode	95% HPD-LowB	95% HPD-UppB	R^2
Time MRCA	*	23,400	15,725	62,100	4,100	0.55
Ne Modern	(100 - 200,000)	152,645	200,000	200,000	66,833	0.40
Time Onset Expansion	(2,550 - 37,500)	13,575	9,175	33,450	4,125	0.28
Mutation Rate	(0.0003 - 0.0075)	0.0027	0.0022	0.0057	0.0009	0.64
Ne Ancestral	(5 - 6,000)	216	67	1,187	5	0.43



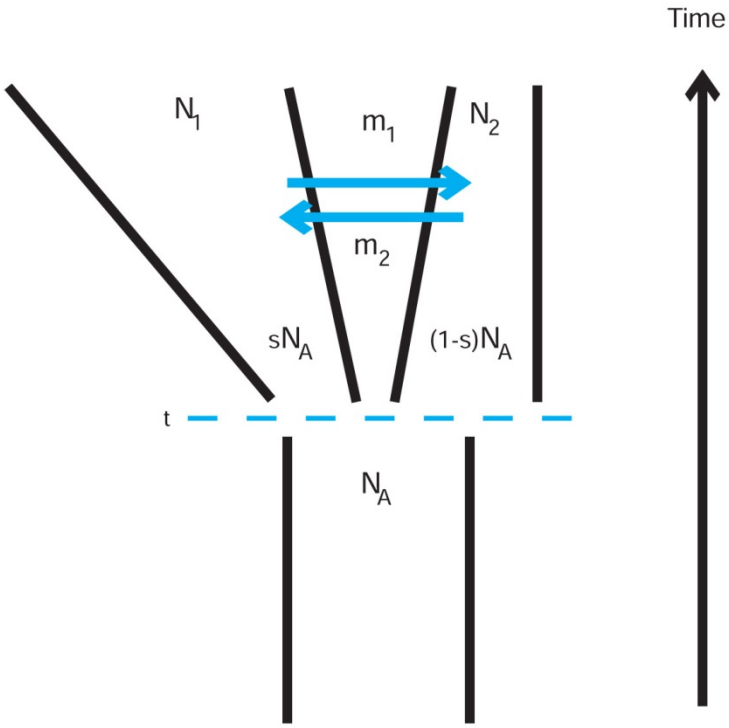
B

	Priors	Median	Mode	95% HPD-LowB	95% HPD-UppB	R^2
Time MRCA		31,425	20,400	6,525	81,125	0.56
Ne Modern	(100 - 200,000)	119,430	200,000	35,406	200,000	0.38
Time Onset Expansion	(2,550 - 37,500)	21,625	16,450	10,300	37,500	0.28
Mutation Rate	(0.0003 - 0.0075)	0.0023	0.0023	0.0008	0.0048	0.64
Ne Ancestral	(5 - 6,000)	388	150	5	1,953	0.43



Supp Fig 7

A



B

Separation Time	Run 1	Run 2	Run3	Run 4
Mutation Rate = 0.0014				S = 0.5
Mean	16,180	16,939	16,861	15,739
95% LowB CI	12,500	12,461	12,020	11,780
95% UppB CI	20,620	22,020	22,700	21,580
Mutation Rate = 0.003				S = 0.5
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
Mutation Rate = 0.006				S = 0.5
Mean	3,775	3,915	3,748	3,831
95% LowB CI	2,683	2,954	2,749	2,935
95% UppB CI	4,989	5,017	4,839	4,905
Mutation Rate = 0.003		S = 0.95		
Mean	9,183	9,146		
95% LowB CI	6,698	6,474		
95% UppB CI	12,583	13,088		

Supplementary table 1: Upper panel: Consensus HVRI 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 (*5*). 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 supplementary fig. 1 are between parentheses. Lower panel: Sequences of all the investigators who had direct contact with the ancient specimens.

Sequence label	**Century (BC)**	**HVR1 motif (16024-16384)**	**Haplotype**	**Reference**
Tarq_1	3rd	069, 126, 193	**Hap1**	This study
Tarq_2	4th-3rd	CRS	**Hap2**	This study
Tarq_3	6th	270	**Hap3**	This study
Tarq_4	3rd-2nd	CRS	**Hap2**	This study
Tarq_5	3rd	126, 229, 362	Hap4	(Vernesi et. al 2004)
Tarq_6	5th	126, 193	**Hap5**	(Vernesi et. al 2004)
Tarq_7	3rd	126, 193, 228, 229, 278	Hap6	(Vernesi et. al 2004)
Tarq_8	5th	278, 334	Hap7	(Vernesi et. al 2004)
Tarq_9	3rd	098, 311, 327	Hap8	(Vernesi et. al 2004)
Cas_1 (S1)	3rd	192	**Hap2**	This study
Cas_2 (S10)	3rd	CRS	**Hap2**	This study
Cas_3 (S11)	3rd	192, 256	**Hap9**	This study
Cas_4 (S17)	3rd	209	**Hap10**	This study
Cas_5 (S3)	3rd	192, 256	**Hap9**	This study
Cas_6 (S4)	3rd	114A, 192,294, 304	Hap11	This study
Cas_7 (S5)	3rd	304	**Hap12**	This study
Cas_8 (S6)	3rd	114A, 192, 256, 294, 304	Hap13	This study
Cas_9 (S8)	3rd	CRS	**Hap2**	This study
Cas_10 (S9)	3rd	CRS	**Hap2**	This study
Vol_1	6th-5th	193, 219	**Hap14**	(Vernesi et. al 2004)
Vol_2	2nd-1st	189, 274, 334, 356	Hap15	(Vernesi et. al 2004)
Vol_3	6th-5th	261	**Hap16**	(Vernesi et. al 2004)
Pie_1	?	193, 219, 256, 270, 291	Hap17	(Vernesi et. al 2004)
Sot_1	?	189, 356	**Hap18**	(Vernesi et. al 2004)
MM_1	7th-6th	CRS	**Hap2**	(Vernesi et. al 2004)
MM_2	6th	126	**Hap5**	(Vernesi et. al 2004)
MM_3	6th	126, 193	**Hap5**	(Vernesi et. al 2004)
MM_4	6th	095G, 126, 189	Hap19	(Vernesi et. al 2004)
MM_5	7th-6th	066, 126, 193, 219	Hap20	(Vernesi et. al 2004)
MM_6	6th	311	**Hap21**	(Vernesi et. al 2004)

Researcher	**Task**	**HVR1 haplotype**
E.P	Excavation	16165 G, 16222 T
S.V	Ancient DNA Laboratory analysis	16311 C
A.S	Ancient DNA Laboratory analysis	16145 A
M.L	Ancient DNA Laboratory analysis	16261 T, 16311 C
D.C.	Ancient DNA Laboratory analysis	16193 T, 16278 T

Supplementary table 2: Detailed description of the samples in the EUR and ANC datasets.

Population	ID	Region	n	Reference
EUR				
Adygei	Ady	Caucasus	50	(Macaulay et al. 1999)
Albanians	Alb	Europe, SouthEast	84	(Belledi et al. 2000; Bosch et al. 2006)
Arabs, Maroc	MoAr	North Africa	32	(Rando et al. 1998)
Armenians	Arme	Caucasus	42	(Nasidze & Stoneking 2001)
Austrians	Aus	Europe, Central	117	(Handt et al. 1994; Parson et al. 1998)
Azerbaijani	Azer	Caucasus	41	(Nasidze & Stoneking 2001)
Basques	Basq	Europe, West	106	(Bertranpetit et al. 1995; Corte-Real et al. 1996)
Belgians	Bel	Europe, Central	33	(Decorte et al. 1996)
Berbers, Maroc	MoBe	North Africa	60	(Pinto et al. 1996; Rando et al. 1998)
Berbers, Tunisia	TuBe	North Africa	155	(Fadhlaoui-Zid et al. 2004)
British	GB	Europe, North	100	(Piercy et al. 1993)
Bulgarians	Bul	Europe,SouthEast	882	(Calafell et al. 1996; Karachanak et al. 2011)
Catalans	Cat	Europe, West	15	(Corte-Real et al. 1996)
Cherkessians	Cher	Caucasus	44	(Nasidze & Stoneking 2001)
Cornish	Cor	Europe, NorthWest	69	(Richards et al. 1996)
Croatians	Cro	Europe, SouthEast	96	(Babalini et al. 2005)
Danes	Dan	Europe, North	32	(Richards et al. 1996)
Egyptians	Egy	North Africa	124	(Krings et al. 1999; Stevanovitch et al. 2004)
Estonians	Est	Europe, North	28	(Sajantila et al. 1995)
French	Fre	Europe, Central	111	(Cali et al. 2001)
Galicians	Gali	Europe, West	92	(Salas et al. 1998)
Georgians	Geo	Caucasus	102	(Comas et al. 2000; Nasidze & Stoneking 2001)
Germans, North	GerN	Europe, North	108	(Richards et al. 1996)
Germans, South	GerS	Europe, Central	249	(Richards et al. 1996; Lutz et al. 1998)
Greeks	Gre	Europe, SouthEast	73	(Vernesi et al. 2001; Bosch et al. 2006)
Ingush	Ingu	Caucasus	35	(Nasidze & Stoneking 2001)
Italians, Abruzzo_Molise	Abr-Mol	Italy, Central	73	(Babalini et al. 2005)
Italians, Apulia	Apu	Italy, South	26	(Babalini et al. 2005)
Italians, Basilicata	Bas	Italy, South	92	(Ottoni et al. 2009)
Italians, Calabria	Cal	Italy, South	95	(Ottoni et al. 2009)
Italians, Campania	Cam	Italy, South	48	(Babalini et al. 2005)
Italians, Casentino	Cas	Italy, Central	122	(Achilli et al. 2007)
Italians, Florence	Flo	Italy, Central	48	(Turchi et al. 2008)
Italians, Gallura	Gal	Italy,Sardinia	27	(Morelli et al. 2000)
Italians, Jenne	Jen	Italy, Central	103	(Messina et al. 2010)
Italians, Latium	Lat	Italy, Central	52	(Babalini et al. 2005)
Italians, Murlo	Mur	Italy, Central	86	(Achilli et al. 2007)
Italians, Ogliastra	Ogl	Italy,Sardinia	175	(Fraumene et al. 2003)
Italians, Sicily	Sic	Italy, South	154	(Ottoni et al. 2009)
Italians, Vallepietra	Val	Italy, Central	21	(Morelli et al. 2000)
Italians, Volterra	Vol	Italy, Central	114	(Achilli et al. 2007)
Kazakhs, Kirghizs,Uyghurs	Achen	Central Asia	205	(Comas et al. 1998)
Kurds	Kur	Near East	29	(Comas et al. 1998)
Macedonians	Mac	Europe, SouthEast	37	(Bosch et al. 2006)
Middle East	ME	Near East	42	(Di Rienzo & Wilson 1991)
Portuguese	Por	Europe, West	54	(Corte-Real et al. 1996)
Romanians	Rom	Europe, SouthEast	105	(Bosch et al. 2006)
Spaniards, Central	SpaC	Europe, West	74	(Corte-Real et al. 1996; Pinto et al. 1996)
Surians	Sur	Near East	49	(Vernesi et al. 2001)
Swiss	Swi	Europe, Central	72	(Pult et al. 1994)
Turks, Western Anatolia	Tur	Near East	35	(Di Benedetto et al. 2001)
Welsh	Wel	Europe, North	92	(Richards et al. 1996)
ANC				
Medieval Tuscans	Med	Italy, Central	27	(Guimaraes et al. 2009)
Neolithic Farmers	Neo_Farm	Europe, Central	71	(Haak et al. 2005; Lacan et al. 2011)
Hunter-Gatherers Europeans	H_G	Europe, Central	20	(Bramanti et al. 2009)

Pre-Roman Iberian	PR_Ibe	Europe, West	17	(Sampietro et al. 2005)
Lucchesi of Eneolithic	Luc_Ene	Italy, Central	10	unpublished data
Lucchesi from Frizzone	Luc_Friz	Italy, Central	8	unpublished data
Lucchesi of I-VII BC	Luc_ I-VIIBC	Italy, Central	4	unpublished data
Lucchesi of XVI-XVIII BC	Luc_ XVI-XVIII BC	Italy, Central	10	unpublished data
Nuragic Sardians	Nur_S	Italy,Sardinia	23	(Caramelli et al. 2007)

References

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. 80:759-768.

Babalini C, Martinez-Labarga C, Tolk HV, et al. 2005. The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. Eur.J.Hum.Genet. 13:902-912.

Belledi M, Poloni ES, Casalotti R, Conterio F, Mikerezi I, Tagliavini J, Excoffier L. 2000. Maternal and paternal lineages in Albania and the genetic structure of Indo-European populations. Eur.J.Hum.Genet. 8:480-486.

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D. 1995. Human mitochondrial DNA variation and the origin of Basques. Ann.Hum.Genet. 59:63-81.

Bosch E, Calafell F, Gonzalez-Neira A, et al. 2006. Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann.Hum.Genet. 70:459-487.

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

Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L. 1996. From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. Ann.Hum.Genet. 60:35-49.

Cali F, Le Roux MG, D'Anna R, Flugy A, De Leo G, Chiavetta V, Ayala GF, Romano V. 2001. MtDNA control region and RFLP data for Sicily and France. Int.J.Legal.Med. 114:229-231.

Caramelli D, Vernesi C, Sanna S, et al. 2007. Genetic variation in prehistoric Sardinia. Hum.Genet. 122:327-336.

Comas D, Calafell F, Bendukidze N, Fananas L, Bertranpetit J. 2000. Georgian and kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. Am.J.Phys.Anthropol. 112:5-16.

Comas D, Calafell F, Mateu E, et al. 1998. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. Am.J.Hum.Genet. 63:1824-1838.

Corte-Real HB, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, Sykes BC. 1996. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann.Hum.Genet. 60:331-350.

Decorte R, Jehaes E, Xiao FX, Cassiman J-J. 1996. Genetic analysis of single hair shafts by automated sequence analysis of the mitochondrial d-loop region. Advances in Forensic Haemogenetics 6:17-19.

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.

Di Rienzo A, Wilson AC. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc.Natl.Acad.Sci.USA 88:1597-1601.

Fadhlaoui-Zid K, Plaza S, Calafell F, Ben Amor M, Comas D, Bennamar El gaaied A. 2004. Mitochondrial DNA heterogeneity in Tunisian Berbers. Ann.Hum.Genet. 68:222-233.

Fraumene C, Petretto E, Angius A, Pirastu M. 2003. Striking differentiation of sub-populations within a genetically homogeneous isolate (Ogliastra) in Sardinia as revealed by mtDNA analysis. Hum.Genet. 114:1-10.

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

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

Handt O, Richards M, Trommsdorff M, et al. 1994. Molecular genetic analyses of the Tyrolean Ice Man. Science 264:1775-1778.

Karachanak S, Carossa V, Nesheva D, et al. 2011. Bulgarians vs the other European populations: a mitochondrial DNA perspective. Int.J.Legal.Med.

Krings M, Geisert H, Schmitz RW, Krainitzki H, Paabo S. 1999. DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc.Natl.Acad.Sci.USA 96:5581-5585.

Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, Guilaine J, Crubezy E, Ludes B. 2011. Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc.Natl.Acad.Sci.USA 108:9788-9791.

Lutz S, Weisser HJ, Heizmann J, Pollak S. 1998. Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. Int.J.Legal.Med. 111:67-77.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A. 1999. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am.J.Hum.Genet. 64:232-249.

Messina F, Scorrano G, Labarga CM, Rolfo MF, Rickards O. 2010. Mitochondrial DNA variation in an isolated area of Central Italy. Ann.Hum.Biol. 37:385-402.

Morelli L, Grosso MG, Vona G, Varesi L, Torroni A, Francalacci P. 2000. Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. Hum.Biol. 72:585-595.

Nasidze I, Stoneking M. 2001. Mitochondrial DNA variation and language replacements in the Caucasus. Proc.Biol.Sci. 268:1197-1206.

Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, Contini I, Biondi G, Rickards O. 2009. Human mitochondrial DNA variation in Southern Italy. Ann.Hum.Biol. 36:785-811.

Parson W, Parsons TJ, Scheithauer R, Holland MM. 1998. Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case. Int.J.Legal.Med. 111:124-132.

Piercy R, Sullivan KM, Benson N, Gill P. 1993. The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. Int.J.Legal.Med. 106:85-90.

Pinto F, Gonzalez AM, Hernandez M, Larruga JM, Cabrera VM. 1996. Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. Ann.Hum.Genet. 60:321-330.

Pult I, Sajantila A, Simanainen J, Georgiev O, Schaffner W, Paabo S. 1994. Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. Biol.Chem.Hoppe-Seyler 375:837-840.

Rando JC, Pinto F, Gonzalez AM, Hernandez M, Larruga JM, Cabrera VM, Bandelt HJ. 1998. Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. Ann.Hum.Genet. 62:531-550.

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, Sykes B. 1996. Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am.J.Hum.Genet. 59:185-203.

Sajantila A, Lahermo P, Anttinen T, et al. 1995. Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res. 5:42-52.

Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo A. 1998. mtDNA analysis of the Galician population: a genetic edge of European variation. Eur.J.Hum.Genet. 6:365-375.

Sampietro ML, Caramelli D, Lao O, Calafell F, Comas D, Lari M, Agusti B, Bertranpetit J, Lalueza-Fox C. 2005. The genetics of the pre-Roman Iberian Peninsula: a mtDNA study of ancient Iberians. Ann.Hum.Genet. 69:535-548.

Stevanovitch A, Gilles A, Bouzaid E, Kefi R, Paris F, Gayraud RP, Spadoni JL, El-Chenawi F, Beraud-Colomb E. 2004. Mitochondrial DNA sequence diversity in a sedentary population from Egypt. Ann.Hum.Genet. 68:23-39.

Turchi C, Buscemi L, Previdere C, Grignani P, Brandstatter A, Achilli A, Parson W, Tagliabracci A. 2008. Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int.J.Legal.Med. 122:199-204.

Vernesi C, Di Benedetto G, Caramelli D, Secchieri E, Simoni L, Katti E, Malaspina P, Novelletto A, Marin VT, Barbujani G. 2001. Genetic characterization of the body attributed to the evangelist Luke. Proc.Natl.Acad.Sci.USA 98:13460-13463.