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 ... n_k, of individuals studied t_0, t_1, t_2 ... 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 ... 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.

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

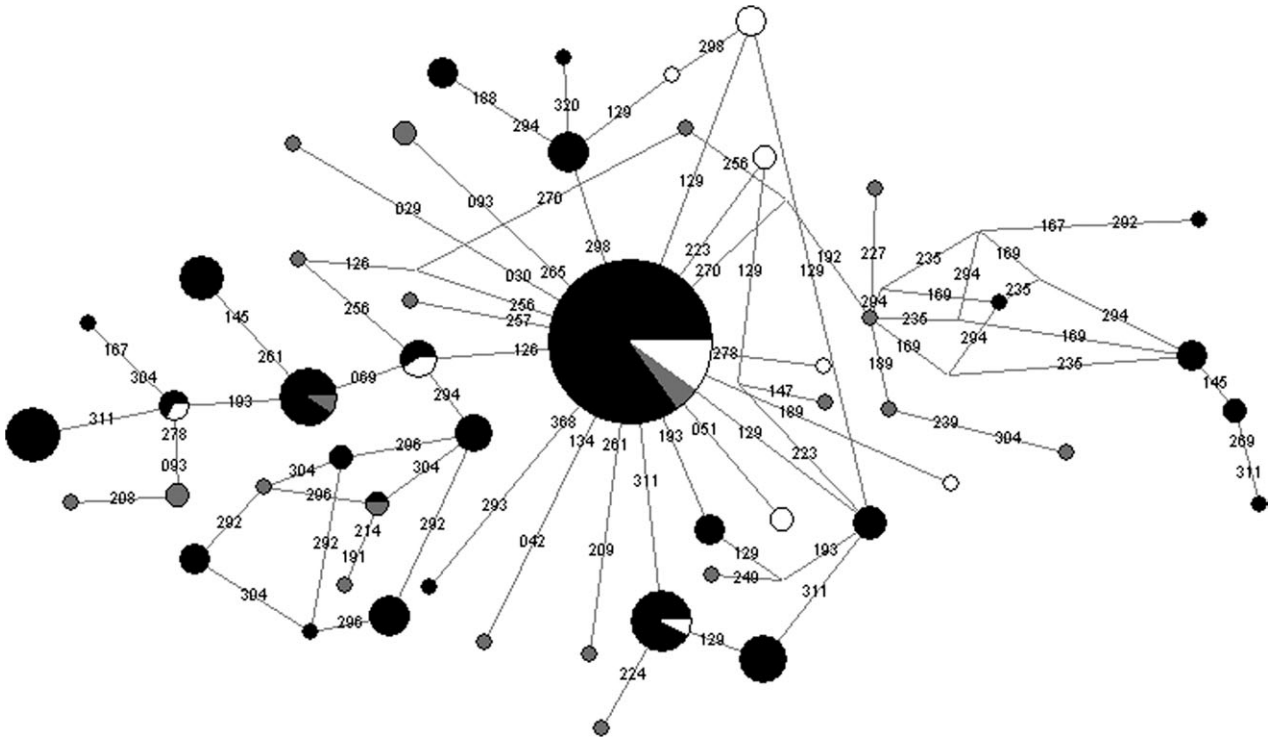

FIG. 1. Median-joining network of the DNA sequences considered. Ancient samples are represented by white areas in the pies, Gallura by gray areas and Ogliastra by black areas. Figures on the edges of the network indicate the position of the nucleotide substitution in the mtDNA reference sequence minus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

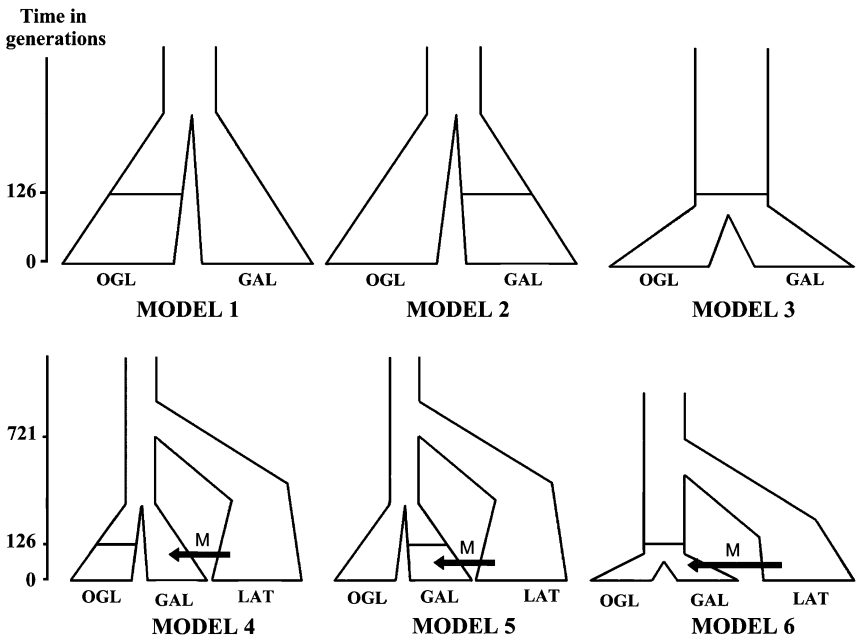

FIG. 3. A schematic summary of the six models tested. Numbers on the *y* axis are generations from the present. The horizontal line at generation 126 represents the nuragic population; the arrows represent gene flow at a rate *M*, which was estimated from the data, from the mainland into Gallura. OGL,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

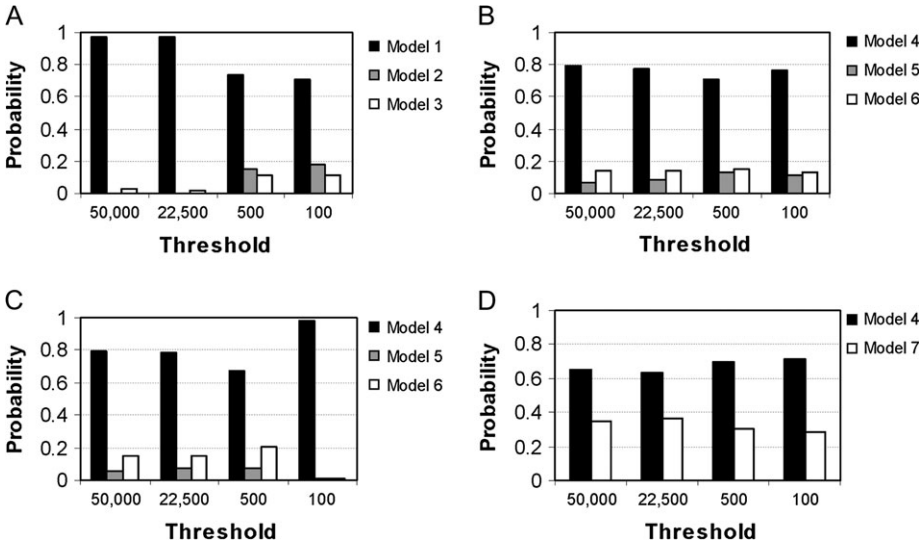

FIG. 4. Posterior probabilities of the models evaluated over different thresholds. Thresholds on the *x* axis indicate the number of simulations retained under the AR (acceptance–rejection sampling: 500 and 100) and the LR (logistic regression: 50,000 and 22,500) approaches. (*A*) Models without immigration; (*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.