Deep Divergences of Human Gene Trees and Models of Human Origins

Michael G. B. Blum*,[1] and Mattias Jakobsson[2]

[1]Laboratoire des Techniques de l'Ingénierie Médicale et de la Complexité (TIMC-IMAG), Equipe Biologie Computationnelle et Mathématique (BCM), Centre National de la Recherche Scientifique (CNRS), Université Joseph Fourier (UJF), Grenoble, France
[2]Department of Evolutionary Biology, Uppsala University, Sweden

***Corresponding author:** E-mail: michael.blum@imag.fr.

Associate editor: Sarah Tishkoff

Abstract

Two competing hypotheses are at the forefront of the debate on modern human origins. In the first scenario, known as the recent Out-of-Africa hypothesis, modern humans arose in Africa about 100,000–200,000 years ago and spread throughout the world by replacing the local archaic human populations. By contrast, the second hypothesis posits substantial gene flow between archaic and emerging modern humans. In the last two decades, the young time estimates—between 100,000 and 200,000 years—of the most recent common ancestors for the mitochondrion and the Y chromosome provided evidence in favor of a recent African origin of modern humans. However, the presence of very old lineages for autosomal and X-linked genes has often been claimed to be incompatible with a simple, single origin of modern humans. Through the analysis of a public DNA sequence database, we find, similar to previous estimates, that the common ancestors of autosomal and X-linked genes are indeed very old, living, on average, respectively, 1,500,000 and 1,000,000 years ago. However, contrary to previous conclusions, we find that these deep gene genealogies are consistent with the Out-of-Africa scenario provided that the ancestral effective population size was approximately 14,000 individuals. We show that an ancient bottleneck in the Middle Pleistocene, possibly arising from an ancestral structured population, can reconcile the contradictory findings from the mitochondrion on the one hand, with the autosomes and the X chromosome on the other hand.

Key words: human origins, time to the most recent common ancestor, TMRCA, archaic admixture, African bottleneck, coalescent.

Introduction

The process by which modern humans arose has been the subject of much debate in paleoanthropology (Stringer 2002). Especially the extent of admixture between anatomically modern humans and archaic populations of *Homo* has been vigorously debated (Wolpoff et al. 2000; Templeton 2002; Garrigan and Hammer 2006; Plagnol and Wall 2006; Fagundes et al. 2007). At one end of the spectrum, the recent Out-of-Africa hypothesis posits that a group of modern humans, arising in Africa about 100,000–200,000 years ago, spread throughout the world by replacing, without admixture, the local archaic human populations (Mellars 2006). At the other end, the multiregional model posits substantial gene flow between local archaic humans and the emerging modern humans, even though it does not exclude Africa as the cradle of modern humans (Wolpoff et al. 2000). In between these two hypotheses, alternative scenarios assume archaic admixture restricted to Africa before the emerging modern humans eventually colonized the globe (Harding and McVean 2004; Gunz et al. 2009).

Substantial genetic evidence has been put forward in support of the recent Out-of-Africa hypothesis. For example, a continuous decrease of genetic diversity with increasing distance from Africa has been observed for autosomal microsatellites (Prugnolle et al. 2005; Ramachandran et al. 2005) and single nucleotide polymorphisms (Li et al. 2008), as well as a continuous increase of linkage disequilibrium (LD) with distance from Africa (Jakobsson et al. 2008). These worldwide patterns are consistent with a serial founder model in which migrant populations are formed from a subset of the previous population in the migration wave outward from Africa (DeGiorgio et al. 2009). Sex-linked markers have also provided evidence for a recent African origin because the common ancestor of all contemporary mitochondrial haplotypes existed as recently as ∼200,000 years ago (Cann et al. 1987), and the ancestor of all Y chromosomes lived ∼100,000 years ago (Thomson et al. 2000; Wilder et al. 2004).

Direct evidence against the recent Out-of-Africa hypothesis can potentially come from comparisons between ancient DNA of archaic humans, such as Neanderthals, and DNA of present-day modern humans (Noonan 2010). Recently, when releasing the draft sequence of a Neanderthal genome, Green et al. (2010) found 1–4% of Neanderthal introgression in the gene pool of non-Africans; however, previous comparisons involving ancient Neanderthal DNA did not provide evidence in favor of admixture between humans and Neanderthals (Krings et al. 1997; Noonan et al. 2006; Wall and Kim 2007). Genetic evidence that does not involve ancient DNA, in particular elevated levels of LD in modern humans, was also found to be indicative of ancient admixture (Plagnol and Wall 2006; Wall et al. 2009). It has additionally been argued that the presence of very old lineages, or deep divergences, for autosomal genes and genes on the X chromosome is incompatible with a simple, single origin of

doi:10.1093/molbev/msq265 Advance Access publication October 7, 2010

modern humans, and that these deep divergences are instead evidence in favor of archaic admixture (Harris and Hey 1999; Harding and McVean 2004; Garrigan et al. 2005; Evans et al. 2006; Garrigan and Hammer 2006; Hayakawa et al. 2006; Patin et al. 2006; Cox et al. 2008; Kim and Satta 2008).

A measure of the (deepest) divergence of a gene tree is the time to the most recent common ancestor (TMRCA). The TMRCA is the time at which the most recent common ancestor (MRCA) of all existing copies of a given gene lived. A genome-wide frequency distribution of the TMRCAs has been reported by curating the literature (Garrigan and Hammer 2006), but no systematic and consistent analysis has been performed in a single genome-wide data set. We report the first genome-wide estimation of the TMRCAs of anatomically modern humans, and we investigate if different scenarios of human evolutionary history are supported by this estimate. In particular, we investigate to what extent the ages of the autosomal and X-linked lineages are compatible with the recent Out-of-Africa hypothesis.

Materials and Methods

Sequence Data

The data comprise of 40 resequenced independent intergenic regions from the autosomes and the X chromosome (Wall et al. 2008). Each region encompasses approximately 20 kb and consists of three 2-kb sequence fragments, separated by 7 kb of unsequenced DNA ("locus trio design," see Garrigan et al. 2005). We present the results for 78 individuals from three African populations (Mandenka, Biaka, and San from Namibia), one European population (Basque), one Asiatic population (Han), and one population from Oceania (Melanesia). Two common chimpanzee sequences, available in the database, were used as outgroups.

TMRCA Estimation

We use the method of Thomson et al. (2000) for estimating TMRCAs, and we reiterate the basic motivation for the estimator (for an extensive description, see Thomson et al. 2000; Hudson 2007). Thomson's estimator requires an outgroup that provides information on the ancestral state at every polymorphic position and assumes an infinite sites model.

Let T be the time to the MRCA for a sample of n lineages, and let x_i be the number of mutations that have occurred between the MRCA and lineage i. We assume that the number of mutations along a branch, x_i, is Poisson distributed with mean equal to the product of the mutation rate (u) and the branch length (T). Then, Thomson's estimator of TMRCA is

$$\hat{T} = \sum_{i=1}^{n} x_i/(nu).$$

To estimate the mutation rate u, we assume a molecular divergence of 6 My between human and chimpanzee (Glazko and Nei 2003), and a generation time of 25 years. Computing the mean number of nucleotide differences between two chimp sequences and the human sequences, we find a mean mutation rate of 9.90×10^{-10}/bp/year, on the same order as Fagundes et al. (2007).

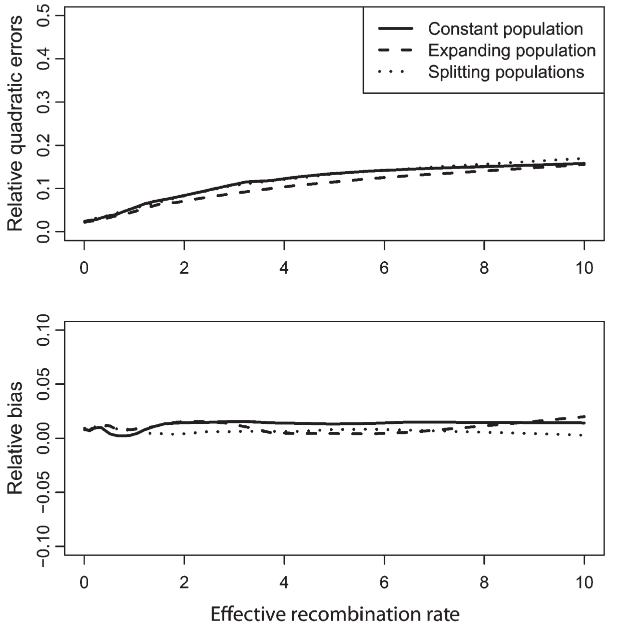

FIG. 1. Average errors of the TMRCA estimate for $n = 100$ individuals. We compare estimated TMRCAs $\hat{T}$ to true averaged TMRCAs T_{av}. The true TMRCAs are averaged using the median along the 20 kb sequences. The relative bias is defined as the average over simulations of $(\hat{T} - T_{av})/T_{av}$, and the relative quadratic error is the average of $(\hat{T} - T_{av})^2/T_{av}^2$. The effective recombination rate is equal to four times the ancestral population size times the recombination rate per generation. In all simulations, we use the mutation rate estimated from the data.

To compute Thomson's estimator $\hat{T}$, we reduce the data set to human polymorphic sites. Because we only consider sites that are polymorphic within humans, the mutations unique to the outgroup sequences are excluded from the computation of the TMRCA estimator (and have no effect on the TMRCA estimates). We reconstruct the ancestral state of each polymorphic site and, under the infinitely-many-sites mutation model, the state of the MRCA of the within-species sample is assumed to be equal to the outgroup allelic state. For 20 out of 1,588 sites, the ancestral state could not be deduced, and, for these sites, the most frequent allele was assumed to be the ancestral variant. The choice of the infinitely-many-sites model is motivated by the small estimated mutation rate ($u = 9.90 \times 10^{-10}$/bp/year), which makes the probability of two mutations hitting the same site negligible.

We investigate the effect of recombination on Thomson's estimator using simulations, and we consider an effective recombination rate ranging from 0 to 10. This range encompasses previous estimates of the recombination rate in humans (The International HapMap Consortium 2005; Voight et al. 2005; Coop et al. 2008b). Using, for instance, a standard estimate of 1 cM/Mb (i.e. 10^{-8}/bp) for the recombination rate (The International HapMap Consortium 2005) gives an effective recombination rate of $\rho = 8$ for a 20-kb region, assuming an effective population size of 10,000 individuals. To quantify the effect of recombination on $\hat{T}$, we compute the relative bias and mean square error of $\hat{T}$ as a function of the recombination rate (fig. 1).

We also investigate if Thomson's estimator $\hat{T}$ remains accurate when the population experienced demographic changes, and we consider three different demographic models: a) a population with a constant size of $N = 10,000$ individuals; b) an expanding population where the population experienced a 10-fold expansion starting at a time distributed uniformly between 0 and 200, 000 years ago from a population size of $N = 10,000$ individuals; and c) a population splitting model. For the last model c, we assume that an ancestral population (of size $N = 10,000$ individuals) split at a time distributed uniformly between 0 and 200, 000 years ago into two subpopulations. One of the two subpopulations contains $N = 10,000$ individuals and the other population contains $N = 10,000 \times p$ individuals with p chosen from a uniform distribution between 0 and 0.5.

Approximate Bayesian Computation

We use approximate Bayesian computation (ABC) to find the range of demographic parameters that yield TMRCAs similar to the empirical estimates of TMRCAs. For each scenario of human evolution (described in the Results and Discussion), the ABC statistical procedure can be described as follows:

- Generate the demographic parameters according to the prior distributions given in supplementary table 1, Supplementary Material online.
- Simulate sequence data with the software *ms* (Hudson 2002). One simulation comprises of generating 20 autosomal and 20 X-linked sequence regions with the same number of samples and the same sequence lengths as in the empirical data.
- Compute the summary statistics (see below) for the simulated sequences, and compute the Euclidean distance between observed and simulated summary statistics.

After performing a total of 100,000 simulations, we retain the 500 simulations that provide the best match to the data. We use an Epanechnikov kernel to assign larger weights to simulations that provide the best match (Beaumont et al. 2002). To account for the imperfect match between simulated and observed summary statistics, we then use regression adjustment as described by Blum and François (2010). After completion of the algorithm, the posterior distribution of the parameters consists of the set of accepted parameters after adjustment.

To compute the summary statistics, we consider, for each of the 40 sequence-regions, the mean number of mutations, $\sum_{i=1}^{n} x_i/n$, between the set of sequences and the ancestral sequence. To reduce the number of summary statistics, we compute, separately for the X-linked and the autosomal markers, the three quartiles of the 20 mean numbers of mutations. This procedure results in a total of six summary statistics.

Choice of Priors for the Mutation and Recombination Rates

For the mutation rate, we choose an empirical Bayes approach in which the prior depends on the data (Casella 1985). We choose a Gamma distribution for the mutation rate, and we estimate the parameters so that the Gamma distribution fits the empirical distribution of the 40 estimated mutation rates. We obtain a Gamma distribution with a shape parameter of 15.18, and a scale parameter of 6.50×10^{-11}. This results in a median mutation rate of 9.65×10^{-10} mutations/bp/year and 95% of the simulated mutation rates are between 5.54×10^{-10} and 1.54×10^{-9} mutations/bp/year. For the crossing-over rate, we consider a log-normal distribution with a mean and a standard deviation (on a log scale) of −18.148 and 0.5802 (Voight et al. 2005). We assume homogeneous cross-over rate along each sequence region.

Computation of the Relative Model Probabilities

We introduce an indicator variable $Y = \{1, 2, 3, 4\}$ that indicates, for each simulation, which one of the four different demographic scenarios (described in the Results and Discussion) generated the data. We perform the same number of simulations for each scenario. We then regress the indicator variable Y by the six summary statistics using local multinomial logistic regression to obtain the model probabilities $P(Y|s)$ as a function of the summary statistics s (Fagundes et al. 2007; Beaumont 2008). Local logistic regression differs from standard logistic regression because larger weights are given to the simulations for which the summary statistics are close to the observed summary statistics. By computing the logistic regression equation for the observed summary statistics $P(Y|s = s_{\text{obs}})$, we obtain the relative model probabilities. To perform local multinomial logistic regression, we use the R package VGAM (Yee 2010).

Posterior Predictive Simulations

To perform goodness of fit of the scenarios of human evolution (fig. 2), we perform posterior predictive checks (Gelman et al. 2003). We sample, with replacement, 10,000 multivariate demographic parameters at random from the posterior distribution obtained with the ABC algorithm. Using *ms* (Hudson 2002), we simulate, for each multivariate demographic parameter, gene trees along a 20 kb sequence and compute the median of the (potentially different) simulated TMRCAs found along the sequence. For each scenario, this results in a total of 10,000 median TMRCAs that are displayed in figure 2.

Results and Discussion

Genome-wide Estimation of the TMRCAs

The data comprise 40 resequenced independent intergenic regions from the autosomes and the X chromosome provided by a public DNA sequence database that has been designed for the purpose of analyzing human prehistory (Wall et al. 2008). To compute genome-wide estimation of the

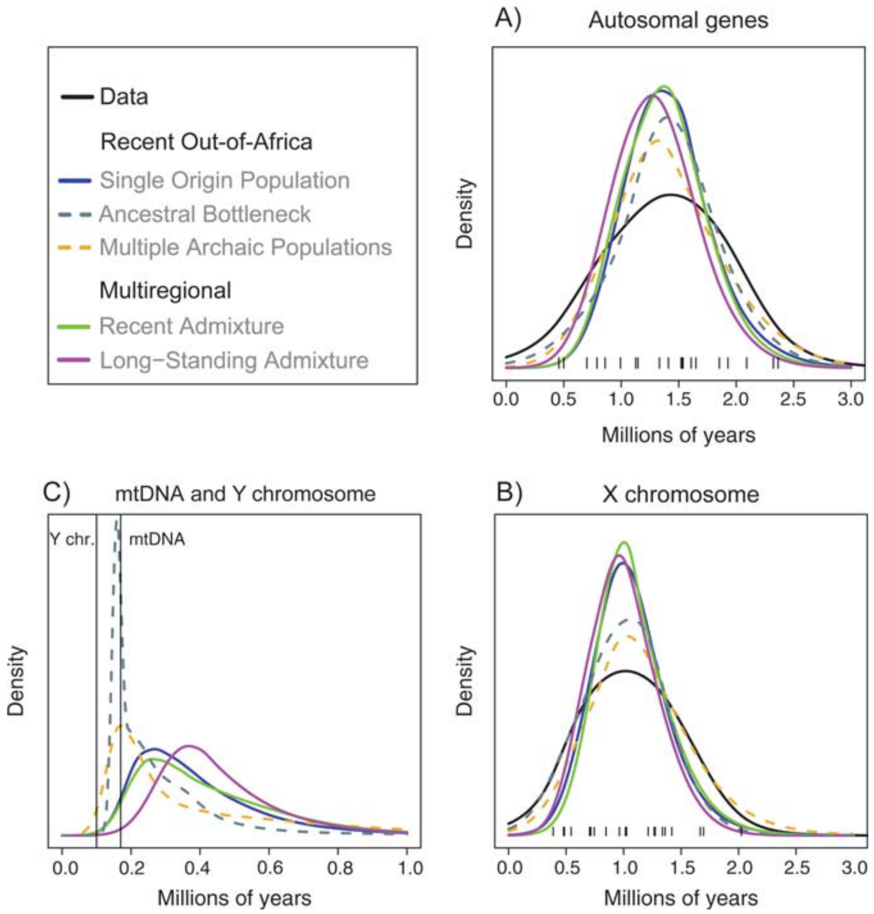

FIG. 2. Estimates of the TMRCAs and posterior predictive distribution for the different models of human evolution. (*A*) The autosomes, (*B*) the X chromosome, and (*C*) mtDNA and the Y chromosome. In C, the thick vertical lines correspond to TMRCA estimates, from the literature, for the Y chromosome and the mtDNA (Ingman et al. 2000; Wilder et al. 2004). The short vertical lines in A and B show the TMRCA estimates for each individual locus. The simulated TMRCAs have been obtained by computing the median of the TMRCAs along the 20 kb simulated sequences.

TMRCA, we consider the statistic of Thomson et al. (2000), which has been used for instance to date the ancestor of the human Y chromosome (Thomson et al. 2000) and the human FOXP2 gene (Coop et al. 2008a). For each DNA fragment, the TMRCA estimate is obtained by computing the number of mutations between each sample (gene copy) and a reconstructed ancestral sequence, and averaging across the gene-copies of the sample. Generally, the bias of this estimator has been shown to be small (Hudson 2007), and we demonstrate that it is also robust to recombination (see fig. 1). To affix a time scale in years to the TMRCA estimates, we assume a molecular divergence of 6 My between human and chimpanzee (Glazko and Nei 2003), and a generation time of 25 years. For each region of 20 kb, we compute one TMRCA estimate so that this estimate captures an average TMRCA for the possibly different TMRCAs found along the 20 kb region.

Figure 2 displays the distribution of the TMRCAs for the 20 autosomal loci (fig 2*A*) and the 20 loci located on the X chromosome (fig 2*B*). The median of the TMRCA is approximately 1,500,000 years for the autosomes (first quartile = 950,000 and third quartile = 1,700,000) and approximately 1,000,000 years for the X chromosome (first quartile = 700,000 and third quartile = 1,350,000). These numbers are close to what Garrigan and Hammer (2006) found by collecting TMRCA estimates from the literature. However, they are at odds with numerical predictions obtained by Fagundes et al. (2007) for the recent Out-of-Africa model. They found that the distribution of autosomal TMRCAs should peak around the time when modern humans emerged (100,000–200,000 years ago) and that 50% of the TMRCAs should be more recent than 650,000 years. In contrast, we find that only two out of 20 autosomal TMRCAs are more recent than 650,000 years.

A number of authors have argued that deep genealogical histories are incompatible with the recent Out-of-Africa hypothesis, and instead claimed that these deep genealogies are evidence in favor of archaic admixture (Harris and Hey 1999; Harding and McVean 2004; Garrigan et al. 2005; Evans et al. 2006; Garrigan and Hammer 2006; Hayakawa et al. 2006; Patin et al. 2006; Cox et al. 2008; Kim and Satta 2008). In the following sections, we investigate to what extent the genome-wide distribution of the TMRCAs is compatible with the recent Out-of-Africa model. We also consider alternative models that assume archaic admixture and check if they provide a better fit to the TMRCA distribution.

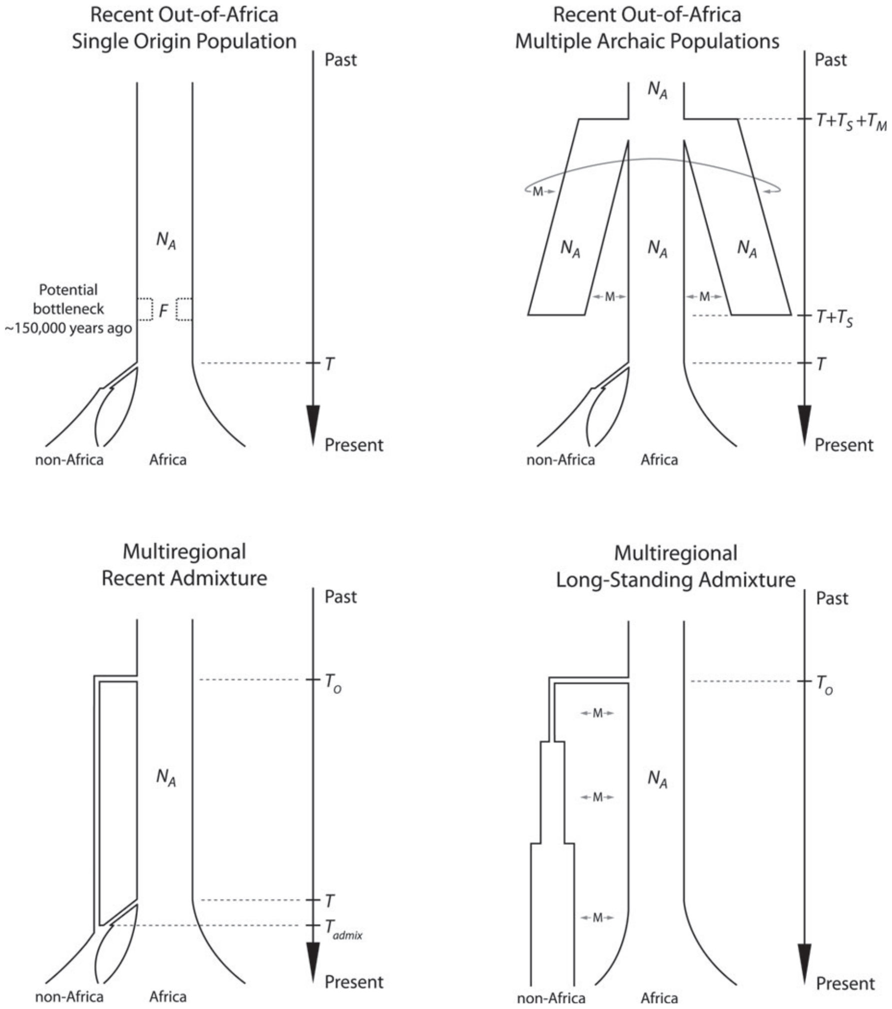

FIG. 3. Four different scenarios of human evolution. (1) *"Single origin population."* A recent Out-of-Africa scenario in which modern humans descended from one subpopulation of archaic humans that was a separate population for a long time in Africa. The recent Out-of-Africa scenario potentially includes a bottleneck before the exodus from Africa, ∼150,000 years ago. (2) *"Multiple archaic populations."* A recent Out-of-Africa scenario in which different archaic African populations were connected by gene flow, even though only one archaic population eventually colonized the globe (Harding and McVean 2004; Garrigan and Hammer 2006). (3) *"Recent admixture."* A multiregional scenario in which archaic and modern humans were isolated during 300–600,000 years and admixed recently in Eurasia, 30–70,000 years ago (Plagnol and Wall 2006). (4) *"Long-standing admixture."* A multiregional scenario with continuous and long-standing admixture between archaic and modern humans. Ancestral population size: N_A, time of the migration out of Africa: T, inbreeding coefficient during the bottleneck: F, time of structuring of archaic African population: $T + T_S + T_M$, ending of structured archaic African population: $T + T_S$, time of archaic humans exiting Africa: T_0, time of admixture: T_{admix}, and migration rate: M.

TMRCA Distributions Predicted by Different Scenarios of Human evolution

We compare observed TMRCAs to simulated TMRCAs for four different scenarios of modern human origin (fig. 3 and supplementary table 1, Supplementary Material online). The two first scenarios are versions of the Out-of-Africa model, and the last two scenarios are versions of the multiregional model:

1. "Single origin population." A recent Out-of-Africa scenario in which modern humans descended from one subpopulation of archaic humans that was a separate population for a long time in Africa,
2. "Multiple archaic populations." A recent Out-of-Africa scenario in which different archaic African populations were connected by gene flow, even though only one archaic population eventually colonized the globe (Harding and McVean 2004; Garrigan and Hammer 2006; Campbell and Tishkoff 2008),
3. "Recent admixture." A multiregional scenario in which archaic and modern humans were isolated during 300–600,000 years and admixed recently in Eurasia, 30–70,000 years ago (Plagnol and Wall 2006), and
4. "Long-standing admixture." A multiregional scenario with continuous and long-standing admixture between archaic and modern humans.

To find the range of demographic parameter values that yield TMRCAs similar to the 40 estimates from the empirical data, we use ABC (Beaumont et al. 2002; Csilléry et al.

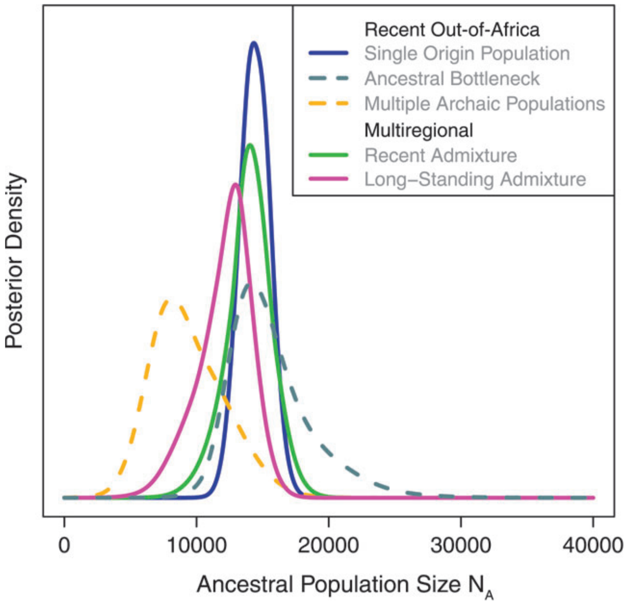

FIG. 4. Posterior distribution of the ancestral effective population size N_A.

2010) and coalescent simulations (Hudson 2002). The empirical genetic data are summarized by six summary statistics that capture the divergence of gene trees. For each DNA fragment, we compute, for the human polymorphic sites, the mean number of mutations between the gene copies of the sample and the reconstructed ancestral sequence (see Materials and Methods). We compute—separately for the X chromosome and the autosomes—the three quartiles of these averaged number of mutations.

We estimate the ancestral effective population size of archaic humans (fig. 4 and supplementary table 2, Supplementary Material online) and find, for the recent Out-of-Africa scenario, that the most likely value is 14,500 (95% credibility interval [CI] = 12,000–17,000) similar to previous estimates (Takahata 1993; Harding et al. 1997; Wall 2003; Voight et al. 2005). In scenario 2, which assumes a structured population in Africa before the emergence of modern humans, we find a slightly lower estimate on the order of 8,000 individuals (95% CI = 5,000–15,000) reflecting that several archaic African populations contribute to the modern gene pool.

In addition to parameter inference, the ABC approach also offers a convenient way to assign a probability to each of the scenarios (Fagundes et al. 2007; François et al. 2008; Verdu et al. 2009). We find that the four different models are almost equally supported by the divergence of gene trees because the four posterior probabilities range between 20% and 30% (supplementary fig. 1, Supplementary Material online). These even probabilities reflect that the relatively ancient lineages found in the autosomes and X-linked genes neither favor nor disfavor the models with archaic admixture (models 2–4).

To check if the different scenarios of human evolution provide a good fit to the data, we compare the empirical TMRCA estimates to the TMRCAs predicted by the different scenarios. All four models predict, on average, lineages as old as 1,500,000 years for autosomal fragments (fig. 2A) and as old as 1,000,000 years for X-linked fragments (fig. 2B). In short, we find that both the simple replacement model and the models with archaic admixture are perfectly compatible with the deep divergences found in the empirical data.

However, not all aspects of the empirical TMRCAs are well captured by the modeled scenarios of human evolution. The variance of the empirical TMRCAs is larger than the variance predicted by three of the four different models of human evolution (see fig. 2 and supplementary table 3, Supplementary Material online), and this large variance has been interpreted as the result of archaic sub-structure in Africa (Harding and McVean 2004). Indeed, the "multiple archaic populations" (scenario 2) shows similar variance of TMRCAs as the empirical data, but the inflated variance of the empirical TMRCA estimates can also be due to variation in mutation or recombination rate across the 40 sequence regions (McVean et al. 2004).

The Mitochondrion and the Y chromosome

We also investigate the distribution of TMRCAs that is expected for the Y chromosome and the mitochondrial chromosome (fig. 2C). The models of human evolution typically predict older TMRCAs compared with the estimated 170,000 years for mitochondrial DNA (mtDNA) (Ingman et al. 2000), and the upper estimate of 100,000 years for the Y chromosome (Tang et al. 2002; Wilder et al. 2004; Shi et al. 2010). For mtDNA, a TMRCA of 170,000 years is within the range of values predicted by the "multiple archaic populations" scenario ($P(\mathrm{TMRCA} < 170,000) = 0.21$), but the mitochondrial TMRCA estimate is difficult to reconcile with the remaining three scenarios ($P < 4 \times 10^{-2}$). For the Y chromosome, a TMRCA of 100,000 years is clearly at odds with three of the models ($P < 6 \times 10^{-4}$), but for the "multiple archaic populations" scenario with archaic African admixture, the proportion of simulated gene trees with TMRCAs younger than 100,000 years is larger than for the other three models, albeit quite small ($P = 1.5 \times 10^{-2}$). Although assuming an archaic structured population in Africa, as in scenario 2, reduces the Y chromosome and mtDNA TMRCAs, the model cannot fully explain a Y chromosome ancestor living as recently as 100,000 years ago. In scenario 2, we consider three archaic populations, and increasing the number of archaic populations will further decrease the effective population size of each subpopulation, which will decrease the predicted haploid TMRCAs, bringing them in line with the empirical estimates. However, because the mtDNA and the Y chromosome are nonrecombining units, their young TMRCAs can also be explained by recent selective sweeps, caused by directional selection at any gene within the nonrecombining units (Kreitman 2000; Jobling and Tyler-Smith 2003).

A Bottleneck When Anatomically Modern Humans Emerged

An alternative explanation for young haploid TMRCA involves a demographic bottleneck concomitant with the

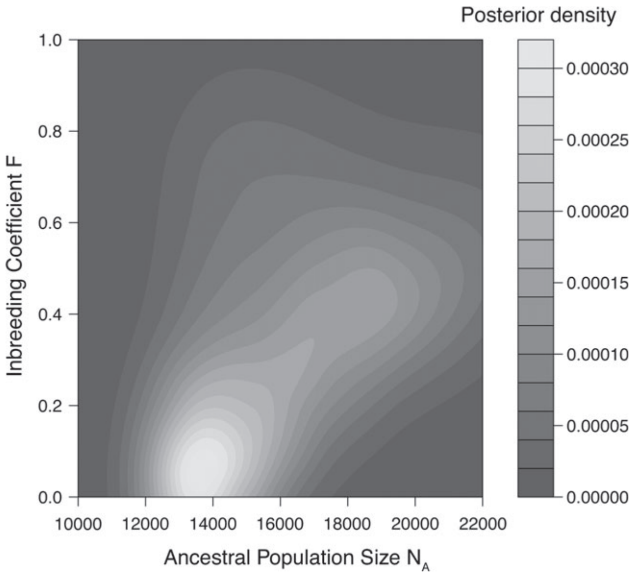

FIG. 5. Joint posterior distribution of the ancestral effective population size N_A and the inbreeding coefficient F during the ancestral bottleneck.

emergence of anatomically modern humans. The earliest known suite of derived morphological traits associated with modern humans appears in fossil remains from Ethiopia dated to 150–190 kya (White et al. 2003; McDougall et al. 2005). A model of the origin and spread of modern humans proposed by Lahr and Foley (1994) assumes that the emerging modern humans experienced bottlenecks within Africa during the penultimate glacial age (130–190 kya) when cold, dry climates caused isolation of populations within Africa (see also Ambrose 1998). Here, we consider a bottleneck that occurred 150,000 years ago, and we measure the bottleneck intensity by the inbreeding coefficient during the bottleneck

$$F = 1 - \left(1 - \frac{1}{2bN_A}\right)^D,$$

where bN_A is the diploid population size during the bottleneck and D corresponds to the duration (in generations) of the bottleneck. To give a reference value, the inbreeding coefficient F corresponding to the Out-of-Africa bottleneck was inferred at $F = 0.175$ (Akey et al. 2004; Voight et al. 2005). Considering an uniform prior between 0 and 1 for F, we find that there is a large range of parameter values compatible with the autosomal and X-linked TMRCAs (fig. 5). For instance, both pairs of parameter values ($F = 0, N_A = 14,000$) and ($F = 0.4, N_A = 18,000$) are clearly within the range of the bivariate posterior distribution found with our ABC approach. The relatively large bivariate posterior range of the inbreeding coefficient F and of the ancestral size N_A shows that a strong bottleneck can still produce an average autosomal TMRCA of 1.5 My provided that the ancestral size was large enough before the bottleneck.

To investigate if a bottleneck 150,000 years ago in Africa can account for both recent haploid and old autosomal ancestors, we simulate TMRCAs for the different chromosomes by sampling the demographic parameters from the posterior distribution that was obtained using ABC. We find that the recent mtDNA TMRCA is clearly within the range of predicted values ($P = 0.37$, fig. 2*C*) as well as the old X-linked and autosomal TMRCAs (fig. 2*A* and *B*). The fact that a bottleneck can accommodate an 8-fold discrepancy between autosomal and mtDNA TMRCA can be seen by plotting TMRCAs as a function of F and N_A (fig. 6). Setting the inbreeding coefficient at $F = 0.4$, for instance, figure 6 shows that the mean mtDNA TMRCA is smaller than 200,000 years, for a large range of values of N_A, whereas the 20 kb averages of TMRCAs, for autosomal and X-linked markers, are older than 1.5 My and 1 My when $N_A > 16,000$ individuals. Finally, as for the other models of human evolution, the TMRCA of 100,000 years for the Y chromosome remains unexpectedly young ($P = 4 \times 10^{-4}$).

We find that both the "multiple archaic populations" model and a sudden bottleneck, 150,000 years ago in Africa, can account for the 8-fold discrepancy between autosomal and mtDNA TMRCA. Although modeling different patterns of human evolution, these two scenarios are different versions of a bottleneck in the human lineage before the succeeding migration out of Africa. Previous attempts to detect an ancestral African bottleneck have often been inconclusive and the genetic evidence in favor of the Out-of-Africa bottleneck are more salient (Marth et al. 2004; Voight et al. 2005). However, it is more difficult to detect this potential middle-Pleistocene bottleneck compared with the Out-of-Africa bottleneck because it is more ancient (Depaulis et al. 2003). Additionally, the unexpected large levels of African LD that has been interpreted as evidence of archaic admixture (Plagnol and Wall 2006; Wall et al. 2009) may potentially be explained by an ancient bottleneck (Schmegner et al. 2005). It is therefore important to consider an ancestral bottleneck, possibly following ancestral substructure and gene flow, when investigating demographic models of human evolution (Schaffner et al. 2005; Gutenkunst et al. 2009; Laval et al. 2010).

A Simple Population Genetic Prediction

Although we perform coalescent simulations, standard population genetic theory can predict the TMRCAs found for the autosomes and the X chromosome in the Out-of-Africa model when there is no ancestral bottleneck or admixture. For a population of constant diploid size N in which the effective number of males and females is the same, the expected waiting time (in generations) before the coalescence of a (reasonably large) sample of genes is approximately $4N$ for the autosomes and $3N$ for the X-linked genes (see, e.g., Hein et al. 2005). Using a generation time of 25 years and an effective population size of 14,000 individuals, the computation leads to an average TMRCA of 1,400,000 years for autosomal genes and 1,050,000 years for X-linked genes, which are both very close to our estimates from the sequence data. This theoretical argument shows that the difference between the TMRCAs of the autosomes and the X-linked genes is easily explained by the difference in effective population size. The same argument yields an average TMRCA

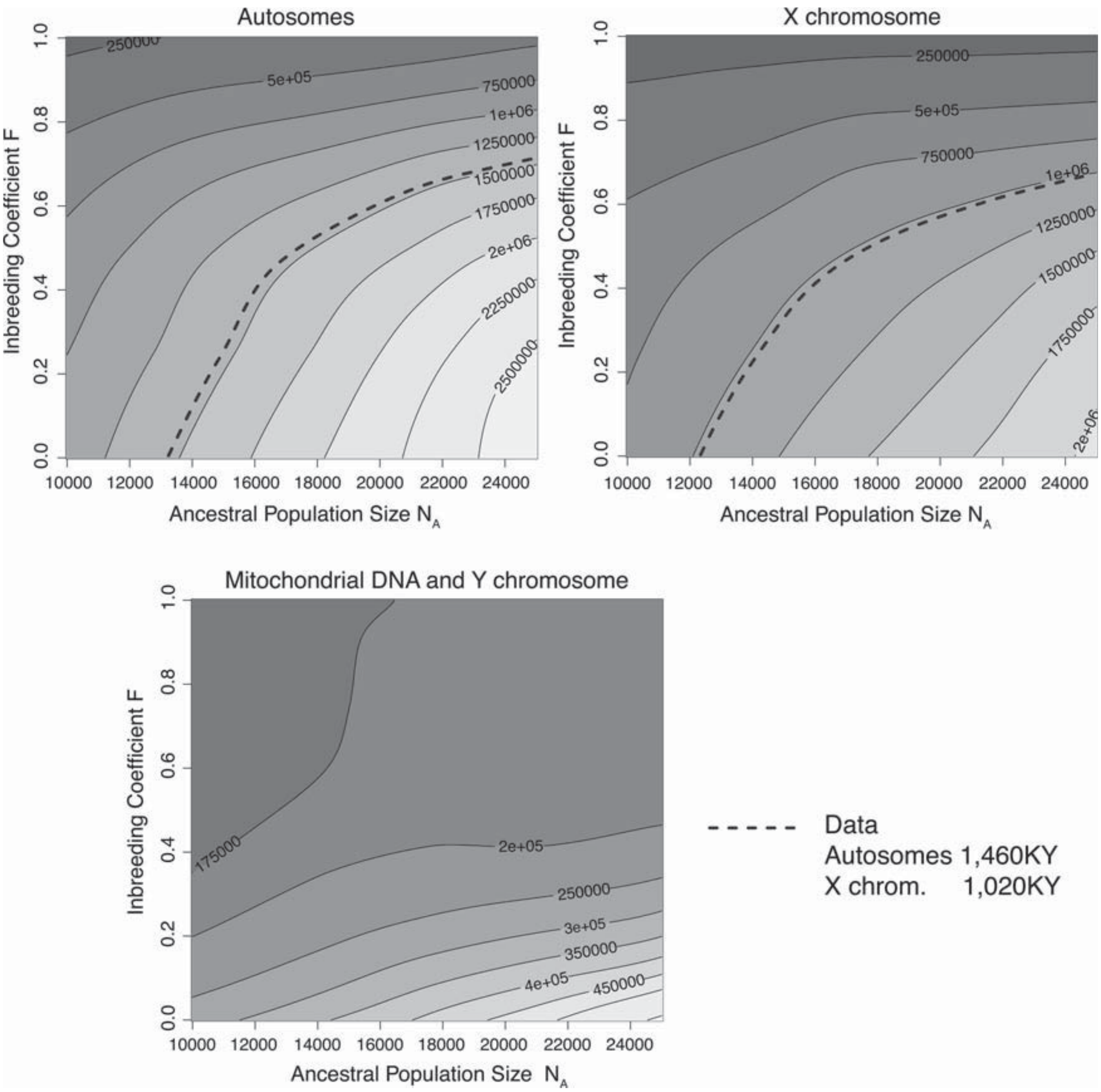

Fig. 6. Mean TMRCA as a function of the inbreeding coefficient F and the ancestral population size N_A for autosomal, X-linked, and haploid genes.

of 350,000 years for both the mtDNA and Y chromosome, clearly deviating from the estimates based on empirical data (Ingman et al. 2000; Wilder et al. 2004; Shi et al. 2010).

Conclusion

We provide 20 autosomal and 20 X-linked estimates of TMRCAs of a sample of contemporary humans and find that the autosomal ancestors of modern humans lived ∼1,500,000 years ago and that the X-linked ancestors lived ∼1,000,000 years ago. The ranges of values for the TMRCAs are quite large: 450,000–2,400,000 years for the autosomes, and 380,000–2,000,000 for the X chromosome. These values are in the same range as previous estimates for autosomal and X-linked genes (see, e.g., Templeton 2002; Garrigan and Hammer 2006; Tishkoff and Gonder 2006). We investigate to what extent the recent Out-of-Africa model reproduces the pattern of estimated TMRCAs, and when setting the ancestral effective size of humans to ∼14,000, this model reproduces the old TMRCAs of the empirical data. Deep divergences in human gene trees are therefore not incompatible with the recent Out-of-Africa hypothesis, and the observation of deep gene genealogies should not be taken as evidence for the multiregional hypothesis. Finally, we show that an ancestral bottleneck in Africa, possibly arising in a structured population, can account for the unexpectedly large discrepancy between young mtDNA and Y chromosome ancestors and old autosomal and X-linked ancestors.

Supplementary Material

Supplementary figure 1 and tables 1–3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a grant from the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) to Mattias Jakobsson and Michael G. B. Blum. Mattias Jakobsson would like to thank the Swedish Research Council and the Erik Philip-Sörensen's foundation for financial support.

References

Akey J, Eberle M, Rieder M, Carlson C, Shriver M, Nickerson D, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2:e286.

Ambrose S. 1998. Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *J Hum Evol.* 34:623–651.

Beaumont M. 2008. Simulation, genetics and human prehistory. In: Matsumura S, Forster P, Renfrew C, editors. Joint determination of topology, divergence time, and immigration in population trees. Cambridge: McDonald Institute for Archaeological Research. p. 134–154.

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

Blum MGB, François O. 2010. Non-linear regression models for approximate Bayesian computation. *Stat Comput.* 20:63–73.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet.* 9:403–433.

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

Casella G. 1985. An introduction to empirical Bayes data analysis. *Am Stat.* 39:83–87.

Coop G, Bullaughey K, Luca F, Przeworski M. 2008a. The timing of selection at the human FOXP2 gene. *Mol Biol Evol.* 25:1257–1259.

Coop G, Wen X, Ober C, Pritchard J, Przeworski M. 2008b. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319:1395–1398.

Cox MP, Mendez FL, Karafet TM, Pilkington MM, Kingan SB, Destro-Bisol G, Strassmann BI, Hammer MF. 2008. Testing for archaic hominin admixture on the x chromosome: Model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. *Genetics* 178:427–437.

Csilléry K, Blum MGB, Gaggiotti OE, François O. 2010. Approximate Bayesian computation in practice. *Trends Ecol Evol.* 25:410–418.

DeGiorgio M, Jakobsson M, Rosenberg NA. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci USA.* 106:16057–16062.

Depaulis F, Mousset S, Veuille M. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *J Mol Evol.* 57:S190–S200.

Evans P, Mekel-Bobrov N, Vallender E, Hudson R, Lahn B. 2006. Evidence that the adaptive allele of the brain size gene microcephalin introgressed into homo sapiens from an archaic homo lineage. *Proc Natl Acad Sci USA.* 103:18178–18183.

Fagundes NJR, 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.

François O, Blum MGB, Jakobsson M, Rosenberg NA. 2008. Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet.* 4:e1000075.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet.* 7:669–680.

Garrigan D, Mobasher Z, Kingan SB, Wilder JA, Hammer MF. 2005. Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics* 177:1849–1856.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. Bayesian data analysis, 2nd ed. *Texts in statistical science*. Boca Raton (FL): Chapman & Hall/CRC.

Glazko G, Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol.* 20:424–434.

Green RE, Krause J, Briggs AW, et al. (56 co-authors). 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.

Gunz P, Bookstein FL, Mitteroecker P, Stadlmayr A, Seidler H, Weber GW. 2009. Early modern human diversity suggests subdivided population structure and a complex out-of-Africa scenario. *Proc Natl Acad Sci USA.* 106:6094–6098.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.

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

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. 1997. Archaic African *and* Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet.* 60:772–789.

Harris EE, Hey J. 1999. X chromosome evidence for ancient human histories. *Proc Natl Acad Sci USA.* 96:3320–3324.

Hayakawa T, Aki I, Varki YSA, Satta Y, Takahata N. 2006. Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. *Genetics* 172:1139–1146.

Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution. Oxford: Oxford University Press.

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

Hudson RR. 2007. The variance of coalescent time estimates from DNA sequences. *J Mol Evol.* 64:702–705.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708–713.

Jakobsson M, Scholz SW, Scheet P, et al. (24 co-authors). 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.

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

Kim H, Satta Y. 2008. Population genetic analysis of the n-acylsphingosine amidohydrolase gene associated with mental activity in humans. *Genetics* 178:1505–1515.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet.* 1:539–559.

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

Lahr M, Foley R. 1994. Multiple dispersals and modern human origins. *Evol Anthropol.* 3:48–60.

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.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166:351–372.

McDougall I, Brown F, Fleagle JG. 2005. Stratigraphic placement and age of modern humans from kibish, Ethiopia. *Nature* 433:733–736.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.

Mellars P. 2006. Why did modern humans populations disperse from Africa ca. 60,000 years ago. *Proc Natl Acad Sci USA.* 103: 9381–9386.

Noonan JP. 2010. Neanderthal genomics and the evolution of modern humans. *Genome Res*. 20:547–553.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Platt D, Paabo S, Pritchard JK, Rubin EM. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113–1118.

Patin E, Barreiro LB, Sabeti PC, et al. (15 co-authors). 2006. Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. *Am J Hum Genet*. 78:423–436.

Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet*. 2:e105.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr Biol*. 15:159–160.

Ramachandran S, Deshpande, O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA*. 102:15942–15947.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*. 15:1576–1583.

Schmegner C, Hoegel J, Vogel W, Assum G. 2005. Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the european population. *Hum Genet*. 118:276–286.

Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, van der Gaag K, de Knijff P, Kayser M, Xue, Y, Tyler-Smith C. 2010. A Worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol*. 27: 385–393.

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

Takahata N. 1993. Allelic genealogy and human evolution. *Mol Biol Evol*. 10:2–22.

Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW. 2002. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* 161:447–459.

Templeton A. 2002. Out of Africa again and again. *Nature* 416:45–51.

The International HapMap Consortium 2005. A haplotype map of the human genome. *Nature* 437:1299–1319.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci USA*. 97:7360–7365.

Tishkoff S, Gonder M. 2006. Human origins within and out of Africa. In: Crawford M, editor. *Anthropological genetics*: theory methods and applications. Cambridge: Cambridge University Press. p. 337–379.

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.

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.

Wall J. 2003. Estimating ancestral population sizes and divergence times. *Genetics* 163:395–404.

Wall J, Cox M, Mendez, F, Woerner A, Severson T, Hammer M. 2008. A novel DNA sequence database for analyzing human demographic history. *Genome Res*. 18:1354–1361.

Wall JD, Kim SK. 2007. Inconsistencies in neanderthal genomic DNA sequences. *PLoS Genet*. 3:e175.

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*. 26:1823–1827.

White TD, Asfaw B, DeGusta D, Gilbert H, Richards G, Suwa, G, Howell FC. 2003. Pleistocene *homo sapiens* from middle awash, Ethiopia. *Nature* 423:742–747.

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.

Wolpoff M, Hawks J, Caspari R. 2000. Multiregional, not multiple origins. *Am J Phys Anthropol*. 112:129–136.

Yee TW. 2010. The VGAM package for categorical data analysis. *J Stat Softw*. 32:1–34.