Inferring population histories using genome-wide allele frequency data

Mathieu Gautier¹* and Renaud Vitalis¹,²

¹INRA, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez, France

²Institut des Sciences de l’Évolution, UMR 5554, CNRS, Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier cedex 5, France

* Corresponding author (E-mail: Mathieu.Gautier@supagro.inra.fr)
Abstract

The recent development of high throughput genotyping technologies has revolutionized the collection of data in a wide range of both model and non-model species. These data generally contain huge amounts of information about the past demographic history of populations.

In this study we introduce a new method to estimate divergence times on a diffusion time-scale from large SNP datasets, conditionally on a population history which is represented as a tree. We further assume that all the observed polymorphisms originate from the most ancestral (root) population, i.e. we neglect mutations that occur after the split of the most ancestral population. This method relies on a hierarchical-Bayesian model, based on Kimura’s time-dependent diffusion approximation of genetic drift. We implemented a Metropolis–Hastings within Gibbs sampler to estimate the posterior distribution of the parameters of interest in this model, which we refer to as the Kimura model. Evaluating the Kimura model on simulated population histories, we found that it provides accurate estimates of divergence time. Assessing model fit using the deviance information criterion (DIC) proved efficient for retrieving the correct tree topology among a set of competing histories. We show that this procedure is robust to low-to-moderate gene flow, as well as to ascertainment bias, providing that the most distantly related populations are represented in the discovery panel. As an illustrative example, we finally analyzed published human data consisting in genotypes for 452,198 SNPs from individuals belonging to four populations worldwide.

Our results suggest that the Kimura model may be helpful to characterize the demographic history of differentiated populations, using genome-wide allele frequency data.
Introduction

The recent development of high throughput genotyping technologies has revolutionized the collection of data in a wide range of both model and non-model species. These data, which may involve tens to hundreds of thousands of single nucleotide polymorphisms (SNPs) in humans (Jakobsson et al, 2008; Li et al, 2008) and other model species (e.g., Gautier, Laloë and Moazami-Goudarzi, 2010; Kijas et al, 2012), contain huge amounts of information about the past demographic history of populations (Wang and Nielsen, 2012). By efficiently reducing multidimensional genetic data into a few synthetic variables, multivariate analyses (see Jombart, Pontier and Dufour, 2009, for a review) such as principal component analyses (PCA) have proven useful to summarize available information about population structure (Patterson, Price and Reich, 2006; Novembre et al, 2008; Gautier, Laloë and Moazami-Goudarzi, 2010). Yet, because they are exploratory and model-free, such approaches are not aimed at making inferences about the underlying history of populations (but see McVean, 2009, for a coalescent interpretation of principal components). Alternatively, model-based approaches have been developed to infer the population structure from multilocus genotypes. One of the most popular, which has been implemented in the software package STRUCTURE (Pritchard, Stephens and Donnelly, 2000), performs a clustering of individuals into genetically homogeneous groups based on an explicit population genetic model (see also Tang et al, 2005; Alexander, Novembre and Lange, 2009). It allows the assignment of individuals into genetically homogenous clusters (sometimes interpreted as ancestral populations) but also the estimation of parameters like the (unknown) allele frequencies in each cluster, or the admixture proportions for each individual. Yet, a limitation of both multivariate analyses and clustering methods is that they are only aimed at characterizing the genetic structure of populations. They do not provide any clue regarding the historical processes that caused the observed structure.

A convenient way of representing the demographic history of populations is borrowed from phylogenetics (Felsenstein, 2003). It is based on the idea that the historical relationship between populations can be represented as a multifurcating diagram, or a "tree". The terminal nodes, or leaves, of the tree represent the present-day populations, while the internal nodes are interpreted as ancestral (unobserved) populations. The branch length between any two nodes is proportional to the amount of genetic divergence between the corresponding populations. Early attempts to characterize population trees relied on moment-based methods to infer the tree topology and to obtain estimates of the branch lengths (Saitou
and Nei, 1987). In principle, likelihood-based techniques are more efficient in using the information present in the genetic data. However, they require the definition of a stochastic model to compute the likelihood of a sample of genes, which is expressed as a function of some parameters that characterize the topology and the branch lengths of a population tree. Two categories of approaches have been followed to derive such likelihood. These differ in whether genetic drift is approximated as a backward-in-time (coalescent) process, or as a forward-in-time (diffusion) process.

The first approach, based on coalescent theory (Kingman, 1982), provides the probabilistic framework to compute the likelihood of a sample of genes, conditional on the (unknown) genealogy of that sample (Hein, Schierup and Wiuf, 2005; Wakeley, 2008). Because it can only be computed for a single genealogical history and since many such histories are compatible with the data, MCMC algorithms have been developed to integrate over unknown genealogies (Hey and Nielsen, 2004, 2007). Yet the convergence of these algorithms can be very difficult to achieve, particularly as the sample size increases and scenarios are more complex (Marjoram and Tavaré, 2006; Beaumont, 2008; Wakeley, 2008). The second approach, to which the present study belongs, is based on diffusion theory (Kimura, 1964). It consists in approximating the discrete process of genetic drift by a continuous-time diffusion process. Applications of diffusion theory provide a probability density of allele frequencies in some simple population models (Kimura, 1964; Ewens, 2004), which can then be used to compute the likelihood of a sample of genes.

In the absence of migration, genetic drift occurs independently in each branch of a population tree. Let us consider one particular branch, with effective population size $N$. At one SNP locus, the allele frequencies drift from one generation to the next so that, in the absence of mutation, the number $X(t+1)$ of copies of one allele at generation $t+1$ is a binomial random variable with index $2N$ and parameter $X(t)/2N$. In this so-called Wright–Fisher model, starting from an ancestral frequency $\pi$, the expected allele frequency $\alpha(t)$ of that allele after $t$ generations is unchanged: $E[\alpha(t)] = \pi$, and its variance is $\text{Var}_{WF}[\alpha(t)] = F\pi(1 - \pi)$, where $F = 1 - (1 - \frac{1}{2N})^t$ is a measure of divergence between the ancestral and the current population (Wright, 1969, pp 345-346). Although deriving explicit formulas for the whole distribution of allele frequencies in the Wright–Fisher model proves to be difficult (Ewens, 2004), it is possible to accurately approximate this discrete process by a continuous-time diffusion process. This diffusion approximation is based on a change of time-scale whereby infinitesimal changes in gene frequencies occur every $\delta t \equiv 1/2N$ units of time (Ewens, 2004; Crow and Kimura, 1971). Hence, the unit of time in the diffusion process corresponds to $2N$ generations in the discrete–time model. Using a diffusion
approximation to the one-locus, two-alleles Wright–Fisher process, Gutenkunst et al (2009) proposed a new method to infer population history in models involving up to three populations. Their method, implemented in the software package ∂a∂i, is based on numerical computation of the expected joint frequency spectrum within and between populations. Numerical evaluation of the diffusion approximation of the joint frequency spectrum allows considering complex evolutionary scenarios including expansions, contractions, migrations, etc. Yet, the generality of the approach comes at the cost of computational burden. Hence, although ∂a∂i handles large resequencing datasets (several Mb), it is limited by the number of populations analyzed. In the Wright–Fisher model, Kimura (1964) derived a general solution, in the absence of selection and mutation, for the distribution of allele frequencies in a finite-size population at any time \( t \). In principle, this solution may be used to compute the likelihood of a sample of genes, which paves the way for the inference of the model parameters. However, since Kimura’s (1964) expression, which depends on some hypergeometric functions, is notoriously difficult to compute (Wang and Rannala, 2004), its use in likelihood-based inference for large genetic datasets has been extremely limited so far.

Instead, several approximations to the pure-drift divergence process have been sought to facilitate the computation of the likelihood of population trees and estimate the underlying parameters. Cavalli-Sforza and Edwards (1967) approximated genetic drift as a Brownian motion process through arc-sine square root transformation of allele frequencies. Very efficient algorithms based on this approximation have been developed, which have been extensively used in the context of maximum-likelihood inference of population trees (see Felsenstein, 2003, pp. 410–414). Yet, the Brownian motion approximation is only valid for small divergence times. Recently, Síren, Marttinen and Corander (2011) proposed to reconstruct population histories from a combination of analytical, numerical and Monte Carlo integration techniques based on a beta approximation of the allele frequency distribution, with expectation \( \pi \) and variance \( F\pi(1 − \pi) \). Even more recently, Pickrell and Pritchard (2012) developed a statistical model for inferring population splits and mixtures, based on a multivariate generalization of the Gaussian approximation of the allele frequency distribution (Coop et al, 2010). Approximating the allele frequency distribution by a Gaussian distribution with mean \( \pi \) and variance \( F\pi(1 − \pi) \) was originally suggested by Nicholson et al (2002). Yet, although the Gaussian and the beta distributions have the same expectation and variance as predicted in the Wright–Fisher model, neither can be derived from first principles in this model. In that respect, both models are reminiscent of Cavalli-Sforza and Edwards’s (1967) approach, in the sense that they are based on mathematically convenient instrumental distributions, rather than on the diffusion
approximation of the process at play.

With the advent of computing power, though, the calculation of complex expressions in likelihood-based inference techniques is now within reach, even for large datasets. In this study we therefore use Kimura’s (1964) diffusion approximation to estimate divergence times from large SNP datasets, conditionally on a population history which is represented as a tree. We further assume that all the observed polymorphisms originate from the most ancestral (root) population, i.e. we neglect mutation. This method is based upon a hierarchical-Bayesian model, for which we implemented a Metropolis–Hastings within Gibbs sampler to estimate the posterior distribution of the parameters of interest, namely the branch lengths throughout the population tree. We evaluated our method using simulated population histories, accounting for various departures from the model assumptions (gene flow and ascertainment bias). Because the true population history is usually unknown, we further investigated the use of the deviance information criterion (DIC) (Spiegelhalter et al, 2002) for choosing between alternative population histories. As an application example, we finally re-analyzed published human data consisting in genotypes for 452,198 SNPs from individuals belonging to four human populations worldwide (Jakobsson et al, 2008).

New Approaches

The Kimura statistical model

In the following, we derive a hierarchical-Bayesian model for integrating gene frequencies in a population tree. Consider a sample made of $J$ populations sharing a common history. Each population has a label, $k$, which varies from 1 to $J$ for the sampled populations, and from $J + 1$ to $r$ for the internal nodes of the tree, where $r$ represents the population at the root of the tree. For a bifurcating tree, there are $J - 1$ internal nodes and therefore $r = 2J - 1$. For a star-shaped phylogeny, where all sampled populations derive from a single ancestral population, $r = J + 1$. In the following, we note $a(k)$ the ancestral population of population $k$. The directed acyclic graph (DAG) of the model is provided in Figure 1, where the annotations are given for illustrative purposes in the special case of a bifurcating tree with three populations. The data consist in $I$ SNP loci, which are biallelic markers with an ancestral and a derived allelic state. In the following, we consider a reference allele, which is arbitrarily defined (e.g., by randomly drawing the ancestral or the derived state). Let $n_{ij}$ be the total number of genes sampled at the
ith locus \(1 \leq i \leq I\) in the \(j\)th population \(1 \leq j \leq J\), i.e. twice the number of genotyped individuals in a diploid population. Let \(x_{ij}\) be the observed count of the reference allele at the \(i\)th locus in the \(j\)th sampled population. Assuming Hardy-Weinberg Equilibrium (HWE), the conditional distribution of \(x_{ij}\) given \(n_{ij}\) and the true (yet unknown) allele frequency \(\alpha_{ij}\) is binomial:

\[
x_{ij} \mid n_{ij}, \alpha_{ij} \sim_{iid} B(\alpha_{ij}, n_{ij})
\]

Let us now consider the second level of the hierarchical model (see Figure 1), which integrates over the distribution of the reference allele frequencies \(\alpha_{ik}\) at the \(i\)th SNP in the \(k\)th population \((k < r)\). In the absence of mutation, assuming that population \(k\) with effective size \(N_k\) diverged from \(a(k)\) for \(t_k\) discrete non-overlapping generations, the distribution of \(\alpha_{ik}\), conditional upon the allele frequency \(\alpha_{ia(k)}\) in the parental population, and upon the branch length \(\tau_k \equiv t_k/(2N_k)\), reads:

\[
\begin{align*}
\pi(\alpha_{ik} \mid \alpha_{ia(k)}, \tau_k) &= (1 - w_{ik}^2) \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} T_{l-1}^1(w_{ik}) T_{l-1}^1(z_i) e^{-\frac{1}{2}l(l+1)\tau_k} & 0 < \alpha_{ik} < 1 \\
P(\alpha_{ik} = 0 \mid \alpha_{ia(k)}, \tau_k) &= (1 - \alpha_{ia(k)}) + \frac{(1-z_i)^2}{2} \sum_{l=1}^{\infty} (-1)^l \frac{2l+1}{l(l+1)} T_{l-1}^1(z_i) e^{-\frac{1}{2}l(l+1)\tau_k} \quad (2) \\
P(\alpha_{ik} = 1 \mid \alpha_{ia(k)}, \tau_k) &= \alpha_{ia(k)} + \frac{(1-z_i)^2}{2} \sum_{l=1}^{\infty} (-1)^l \frac{2l+1}{l(l+1)} T_{l-1}^1(z_i) e^{-\frac{1}{2}l(l+1)\tau_k}
\end{align*}
\]

(see formulae 4.9 and 4.16 in Kimura, 1964). In equation (2), \(w_{ik} = 1 - 2\alpha_{ik}\) and \(z_i = 1 - 2\alpha_{ia(k)}\); \(T_{l-1}^1(x)\) denotes the Gegenbauer polynomial, which can be computed using the recursion: \(T_0^1(x) = 1\), \(T_1^1(x) = 3x\), ..., and \(T_n^1(x) = \frac{1}{n} \left[ (2x(n + \frac{1}{2})T_{n-1}^1(x) - (n + 1)T_{n-2}^1(x)) \right]\) for \(n \geq 2\). In practice, the Gegenbauer polynomial was computed using an iterative algorithm checking for convergence for two consecutive iterations. Convergence was generally achieved within 30 iterations, except for low divergence (e.g., \(\tau_k < 0.02\)) whereby a few hundreds iterations could be needed.

Although it is possible in theory to integrate the binomial sampling over the first level of the hierarchy (i.e., over the \(\alpha_{ik}\) for \(k = 1\) to \(k = J\); see the Supplementary Materials), we found that the evaluation of the resulting formula was numerically instable and was not efficient computationally (data not shown).

Since the allele frequencies in the ancestral population of the full sample \((k = r)\) are unknown, we assumed that the prior distribution of the frequency \(\alpha_{ir}\) of the reference allele for the \(i\)th SNP follows a
beta distribution:

\[ \alpha_{ir} \sim_{iid} \text{Beta}(1.0, 1.0) \] (3)

This prior is non-informative in the sense of the indifference principle, which assigns equal probabilities to all possibilities.

Last, the divergence parameter \( \tau_k \)'s are assumed to be sampled from a uniform distribution:

\[ \tau_k \sim_{iid} \mathcal{U}(0, 10) \] (4)

Assuming that genetic drift occurs independently in each branch of the tree, we may characterize the gene frequency hierarchically along the tree from the most ancestral population toward the leaves. The full model (see Figure 1) then takes the form:

\[
\pi(\alpha, \tau \mid x) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} P(x_{ij} \mid \alpha_{ij}) \prod_{j=1}^{J-1} \tau_j \prod_{i=1}^{I} \pi(\alpha_{ij} \mid \alpha_{ia(j)}, \tau_j) \prod_{i=1}^{I} \pi(\alpha_{ir})
\] (5)

**Results**

**Precision of the Kimura model for estimating differentiation in the Wright–Fisher model**

In order to analyze the precision of the Kimura model based on Kimura’s diffusion approximation, we first analyzed simulated data consisting of allele counts at 5,000 SNPs genotyped from four populations having diverged simultaneously from a single ancestral population (star-shaped phylogeny). Since the truncated Gaussian model (Nicholson et al, 2002) and the beta model (Balding and Nichols, 1995) have been used to approximate genetic drift in star-shaped population histories (see, e.g., Gautier, Hocking and Foulley, 2010), we analyzed these simulated datasets using previously described programs based on these two models (Gautier, Hocking and Foulley, 2010). For the truncated Gaussian and the beta models, branch lengths were interpreted in terms of a differentiation parameter \( F_k \in (0, 1) \) which corresponds to a population-specific \( F_{ST} \) (Excoffier, 2007; Weir and Hill, 2002). Following the notations and assumptions made above, \( F_k = 1 - [1 - 1/(2N_k)]^t \) (Wright, 1969, pp. 345–346). Hence, providing \( N_k \) is not too small (e.g., \( N_k > 50 \)) and \( \tau_k \) is not too large (e.g., \( \tau_k < 0.15 \)), then \( F_k \approx \tau_k \) (see Figure 2B–C). Twenty-three
temporal samples (from \( t = 0.01 \) to \( t = 1 \) unit of time) were taken to assess the precision of the estimations as a function of the level of differentiation. The Kimura model provided very good estimates of divergence times (\( \tau_k \equiv t / (2N_k) \)), irrespectively of the level of differentiation (Figure 2A). Conversely, and as already observed (Gautier, Hocking and Foulley, 2010), both the truncated Gaussian model (Figure 2B) and the beta model (Figure 2C) provided upwardly biased estimates for \( F_k \) when \( \tau_k \) exceeded 0.2. For the beta model, though, biased estimates of \( F_k \) were close to \( \tau_k \) as long as \( \tau_k < 0.5 \) (Figure 2C).

We further evaluated the performance of the Kimura model to estimate branch lengths in bifurcating trees. To that end, allele count data for 5,000 SNPs were simulated for three populations (denoted P1, P2 and P3) related by the same topology ((P1,P2),P3) but with varying branch lengths. Fifty replicates per history were performed. As illustrated in Figure 3, the Kimura model provided accurate estimates of divergence times even if a slight-to-moderate bias was observed for the internal branch (leading to population P4 in the simulated scenarios). The bias was more pronounced with smaller divergence time (Figure 3D).

Performance of the Kimura model to characterize population history

Fifty datasets consisting in allele counts at 5,000 SNPs were simulated using \textit{ms} coalescent-based simulations for a three-populations history noted T1 in Figure 4. Each dataset was then analyzed conditionally on the four possible tree topologies represented in Figure 4 and denoted by T1, T2, T3 and S. As mentioned above, T1 corresponded to the true (simulated) history, T2 and T3 corresponded to incorrect histories and S corresponded to a star-shaped (also incorrect) history.

We found that, across replicated simulated datasets, the DIC always provided a clear support in favor of the correct population history (Table 1). However, the power of the DIC to identify the correct tree was strongly dependent on the information available in the datasets. Indeed, as illustrated in Table 2, for small datasets (1,000 SNPs), the DIC criterion provided support for an incorrect tree topology in a substantial number of cases, while for reasonably larger datasets (including 5,000 or 25,000 SNPs), the DIC was always supporting the correct tree topology. Similarly, the average difference in DIC across replicates between the correct and alternative tree topologies increased sharply with the number of SNPs in the datasets. Finally, for any given number of SNPs, the power of the DIC to support the correct population history slightly decreased with the number of populations considered, and the complexity of the tree topology, as illustrated in Figures S1 and S2. Nevertheless, a realistic number of 25,000 SNPs
(considering currently available datasets in both model and non-model species) was sufficient to obtain
very satisfactory results on more complex tree topologies, as illustrated in Figure S2 for six-population
trees.

As shown in Figure 4A, and in agreement with the above results (see Figure 3B), the different branch
lengths were correctly estimated when the analyses were performed conditionally on the correct population
tree T1. However, when the analyses were performed conditionally on incorrect tree topologies (T2 and
T3), and although the length of the branch leading to population P3 was correctly estimated, the lengths
of the branches leading to P1 and P2 were upwardly biased, and the length of the internal branch (leading
to P4) tended toward zero. Therefore, although the method provided correct estimates of the divergence
of P3 (the most diverged population in the simulations), all other branch lengths in the population tree
were poorly estimated. Interestingly, all else being equal, the internal branch leading to P4 tended to
be shrunk, and therefore the lengths of the branches leading to P1, P2 and P3 were close to the values
obtained in the analyses run conditionally on a star-shaped population history (S).

Robustness to model misspecification

Inference of branch lengths in the presence of gene flow

Overall, we found that the DIC always provided a clear support for the correct population tree, across
replicated simulated datasets, for low-to-moderate levels of migration (\(M = 0.1\) and \(M = 1\), Table 1)
and in most instances (>90%) for a high level of migration (\(M = 10\)). As for the estimation of branch
lengths, we obtained similar results as above when analyzing datasets simulated with a low amount of
gene flow between populations (\(M = 0.1\), Figure 4B). However, and as expected, increasing the level
of migration from moderate (\(M = 1\), Figure 4C) to high levels (\(M = 10\), Figure 4D) tended to bias
downward the branch length estimates. The magnitude of the bias increased with \(M\) and was more
pronounced for the internal branches. In the most extreme case (\(M = 10\), Figure 4D) values, the results
obtained conditionally on the correct tree topology (T1) were similar to those obtained conditionally on
the incorrect topologies T2, T3 and S.
Inference of branch lengths with ascertained SNPs

In order to investigate the sensitivity of the Kimura model to ascertainment bias, we simulated datasets based on the three-populations history T1 considered above. To mimic ascertainment bias, we defined "ghost" individuals within each of the three sampled populations that were used exclusively for discovery and then discarded from further analyses. We considered three different ascertainment schemes, which differed by the origins of the discovery panels used to ascertain SNPs (see the Model and Methods section). As shown in Figure 5A, when the three populations contributed evenly to the SNP discovery panel (ascertainment scheme AS1), we obtained similar results as in the absence of ascertainment bias (compare Figure 4A and Figure 5A). However, we observed a moderate upward bias in the estimate of the branch leading to P3 and a downward bias for the internal branch length estimate (leading to P4), which was larger for analyses performed conditionally on incorrect population histories. When only populations P1 and P3 contributed to the SNP discovery panel (ascertainment scheme AS2), and although the correct population tree was still recognized based on the DIC criterion (see Table 1), more substantial downward (resp. upward) biases were observed for estimates of the branch lengths leading to P1 and P3 (resp. P2 and P4) (see Figure 5B). Finally, when only populations P1 and P2 contributed to the SNP discovery panel (ascertainment scheme AS3), severe biases were observed for estimates of the branch lengths for populations P3 and P4 (Figure 5C). In this latter ascertainment scheme, all else being equal, the branch length estimates were very similar across analyses run with different prior tree topologies. In particular, the analyses run conditionally on incorrect topologies T2 and T3 provided similar branch length estimates as in the analyses run conditionally on a star-shaped population history (S), with a shrunk internal branch leading to P4. It was only for this ascertainment scheme (AS3) that the DIC failed to support the correct population tree (Table 1).

Inference of branch lengths in the presence of recent mutations

In SNP genotyping assays, the SNP ascertainment schemes similar to the AS1 and the AS2 schemes ensure that ancestral SNPs (i.e., SNPs that predate the divergence of the populations under study) are largely over-represented. An over-representation of ancestral SNPs may also occur if the common ancestral population has undergone a strong bottleneck, as for instance in humans or in most domesticated species. However, overlooking derived SNPs (i.e., SNPs that arose after the divergence of the populations under study from the root population) is expected to affect the robustness of our model, if new mutations
occur after the split of the most ancestral population.

Removing the bottleneck (i.e., the -en option in the ms command) in our simulations resulted indeed in higher downward biases for the estimation of branch lengths (Figure S3) and poorer identification of the correct underlying tree (Table S2) as divergence times increased. However, some ascertainment schemes (particularly AS1) may improve performance probably if they result in an enrichment of the datasets in ancestral SNPs (Figure S4).

**Example application on a large human dataset**

As an illustration example, we ran our model based on Kimura’s diffusion approximation on a large human dataset. The data consisted in allele counts at 452,198 autosomal SNPs from four human populations of African (YRI and BIA), European (CEU) and East Asian (JPT) ancestry (Jakobsson et al, 2008). In order to evaluate the extent to which summarizing the history of these four populations by a bifurcating a tree is not overly simplistic, we first performed four-population tests for treeness (also referred to as quartet tests for migration) on the three possible (labeled) unrooted tree topologies (see Reich et al, 2009; Keinan et al, 2007). As detailed in Table S1, the four-population tests for treeness supported the ((YRI,BIA),(CEU,JPT)) unrooted tree topology ($P < 0.15$) and rejected the two alternative ones ($P < 10^{-20}$). This result is in agreement with Keinan et al (2007), who considered individuals from the same populations (YRI and CEU) and closely related ones (Mbuti Pygmy and Chinese from Beijing) but a different set of SNPs. This result further suggests that migration between YRI or BIA on the one hand and CEU or JPT on the other hand has been maintained at a low level since the population split, and has not seriously distorted the joint allele frequency spectrum.

In order to infer more precisely the relations between the four worldwide populations YRI, BIA, CEU and JPT, we then ran our analyses conditionally on six alternative rooted population trees. As represented in Figure 6, these trees corresponded to the five possible rooted topologies derived from the ((YRI,BIA),(CEU,JPT)) unrooted topology plus a star-shaped phylogeny. Based on the DIC criterion, the tree T1 unambiguously received the strongest support, in agreement with the widely accepted Out-Of-Africa model of human evolution.
Discussion

An efficient model to estimate divergence times

In this study, we developed a new hierarchical-Bayesian model to estimate divergence times (on a diffusion time-scale) conditionally on a population tree, using genome-wide allele frequency data. Since genetic drift occurs independently in each branch of the tree in the absence of gene flow, the allele frequency at one particular locus can be modelled hierarchically along the tree from the most ancestral population toward the leaves. Our model is based on Kimura’s diffusion approximation (Kimura, 1964), which arises as an explicit solution of the time-dependent Wright–Fisher model. The parameters of interest in this model are the branch lengths, measured as \( \tau_k \equiv t/(2N_k) \). This definition calls for two remarks. First, it is evident from Kimura’s diffusion approximation that disentangling divergence time \( t \) from effective population size \( N_k \) is not possible, because of the non-identifiability of these two parameters. Estimating \( t \), which is generally the parameter of biological interest, would therefore require informative priors on \( N_k \), which in practice can be derived from other analyses (see, e.g., Gautier et al, 2007).

Second, the population size \( N_k \) needs not to be constant. Indeed, we expect Kimura’s diffusion approximation to be robust to the (unknown) demography of the population (see, e.g., Ewens, 2004, chapter 5). For instance, with demographic fluctuations, and if we assume that population \( k \) has size \( N_{k,i} \) in each generation \( i \), then \( \tau_k = N_{k,0}^{-1} + \sigma^2 \sum_{i=1}^t N_{k,i}^{-1} \) (see Nicholson et al, 2002), where \( \sigma^2 \) represents the variance in the number of descendant across generations (e.g., \( \sigma^2 = 1 \) in the Wright–Fisher model).

Not surprisingly, Kimura’s diffusion approximation outperforms the two alternative models that have been proposed so far for estimating divergence in star-shaped population histories (Nicholson et al, 2002; Falush, Stephens and Pritchard, 2003; Gautier, Hocking and Foulley, 2010). One of these alternative models assumed that the allele frequency distribution in each population might be well approximated by a truncated Gaussian distribution (Nicholson et al, 2002), while the second model relied on a beta distribution for allele frequencies (Balding and Nichols, 1995; Balding, 2003). Yet, neither the truncated Gaussian nor the beta models are based on the analysis of the Wright–Fisher model of genetic drift. Both models are indeed only aimed at modelling the first two moments of the expected distribution of allele frequency, conditionally on the frequency in the ancestral population and the amount of divergence. Note, however, that the beta model (which has been referred to as the F-model by Falush, Stephens and Pritchard, 2003) arises as the diffusion approximation of genetic drift in the migration-drift equilibrium.
island model (Balding and Nichols, 1995; Balding, 2003).

An important difference between the truncated Gaussian and the beta models stems from the fact that only the former allows for variation to be lost in some populations: although the support of the truncated Gaussian is by construction on the \([0, 1]\) real line, the beta model does not include probability masses in 0 and 1, and therefore ignores the possibility of allele loss or fixation. Furthermore, the truncation made on the Gaussian distribution by Nicholson et al (2002) to account for the masses in 0 and 1 reduces the variance of the distribution of allele frequencies, which becomes rapidly smaller than that expected under pure-drift divergence, as divergence increases and the ancestral allele frequency departs from 0.5 (see Figure S5). Indeed, while the expected variance in the Wright–Fisher model equals \(\text{Var}_{WF} = F\pi(1-\pi)\), the variance of the truncated Gaussian distribution considered by Nicholson et al (2002) equals

\[
\text{Var} = \left[1 + \frac{a\phi(a) - b\phi(b)}{\Phi(b) - \Phi(a)} - \left(\frac{\phi(a) - \phi(b)}{\Phi(b) - \Phi(a)}\right)^2\right] \text{Var}_{WF}
\] (6)

where \(a = -\pi/\sqrt{c\pi(1-\pi)}\) and \(b = (1-\pi)/\sqrt{c\pi(1-\pi)}\). In equation (6), \(\phi(x)\) and \(\Phi(x)\) represent, respectively, the probability density function and the cumulative distribution function of the standard normal distribution. The mismatch between the variance of the allele frequency distribution in the truncated Gaussian model and that expected in the Wright–Fisher model may very well explain the observed upward bias in estimates of divergence with the truncated Gaussian model as divergence time increases (see Figure 2).

For all the reasons cited above, the Kimura model clearly outperforms the truncated Gaussian and the beta models for the estimation of divergence time under pure-drift scenarios (see Figure 2), even for small effective population sizes, over the whole parameter space (see, e.g., Figure S6). This improved accuracy comes at the expense of computational burden, though, even if the use of a recursive algorithm to estimate the density function was proven to be efficient. For instance, a typical analysis of one of the datasets from Figures 3 and 4 (i.e., with 5,000 SNPs and 3 sampled populations, conditionally on a bifurcating tree) took about 39 min on a desktop computer equipped with a 3.4 GHz processor. Because computation times are approximately proportional to the product of the number of SNPs and the total number of populations in a given scenario (including internal nodes), analyzing large datasets of several tens to hundreds of thousands SNPs therefore remains tractable even with standard computers.

Providing a prior knowledge on plausible alternative trees (and thus population histories) is available,
the DIC model comparison criterion (Spiegelhalter et al., 2002) was proven to be efficient. The DIC had a better behavior than the pseudo Bayes Factor (pBF) (Gelfand and Dey, 1994), another commonly used measure to assess model fit (see the supplementary information and compare Table S3 and Table 2). However, the identification of the correct underlying topology requires, of course, that the correct tree lies among the set of tested trees. As a consequence, our approach might only be viewed as complementary to tree inference methods, to which the two aforementioned approaches proposed by Síren, Marttinen and Corander (2011) and Pickrell and Pritchard (2012) belong. Síren, Marttinen and Corander (2011)'s approach may indeed be used to characterize the posterior optimal tree topology, but remains in practice limited to the analysis of a few hundred SNPs due to the computational burden. Similarly, Pickrell and Pritchard’s (2012) approach provides a graph representation of the relationships between populations, but since it relies on some extensions of the truncated Gaussian model it may therefore provide accurate branch length estimates for recent divergences only. In any case, we recommend running three- or four-population tests for treeness (Reich et al., 2009; Keinan et al., 2007), in order to evaluate the extent to which summarizing a population history by a bifurcating tree is a reasonable assumption (see the example application on a large human dataset, above).

**Influence of dataset properties on the method performance**

Increasing the number of SNPs had generally no effect on the accuracy of branch length estimates. However, based on the DIC criterion, increasing the number of markers improved substantially the choice of the correct population tree topology. Typically, about 5,000 SNPs seemed sufficient to resolve three-population trees in most instances. The number of SNPs required to resolve alternative tree topologies increased with tree complexity, i.e., with increasing numbers of sampled populations.

With the recent development of high throughput genotyping technologies, typical datasets may involve hundreds of thousands of SNPs. In this context, the implicit assumption of conditional independence of markers, which is made in our and others model, and amounts to assume the exchangeability of markers, might be violated. First, the residual linkage disequilibrium (LD) within the genome creates local dependency among neighboring SNPs. The expected squared correlation coefficient $r^2$ between two SNPs (Hill and Robertson, 1968) is well approximated by $\mathbb{E}(r^2) \approx 1/\rho$ for large values of the population recombination rate $\rho \equiv 4Nr$, where $r$ represents the frequency of crossing over events per generation (Ohta and Kimura, 1969; McVean, 2007; Sved, 1971). Hence, the extent of pairwise LD
declines rapidly towards negligible values as the genetic distance increases, which typically reduces the data to several tens of thousands of “effective” SNPs, as confirmed by empirical studies (see, e.g., Duggal et al, 2008). Although the correlation structure among SNP allele frequencies is not explicitly accounted for in the models, we expect LD to have a limited effect on divergence time estimates in population trees. Nevertheless, LD-based pruning techniques (Purcell et al, 2007), which aim at generating subsets of SNP data in approximate linkage equilibrium, might represent a valuable approach to overcome these difficulties.

Second, the SNP exchangeability assumption also implicitly requires that SNPs are not located within genomic regions targeted by selection. However, the model is expected to remain robust to such departure from neutrality, provided that only a small fraction of SNPs are indeed affected by selective effects (see, e.g., Gautier, Hocking and Foulley, 2010). Conversely, evaluating the local adjustment of the model at each locus (e.g., using posterior predictive checking) may provide a means to identify outlier SNPs, while simultaneously taking into account the demographic history of the sampled populations (see, e.g., Gautier, Hocking and Foulley, 2010).

Third, SNP exchangeability requires the absence of ascertainment bias. This assumption is valid for SNP genotyping datasets obtained by means of next-generation sequencing (NGS) technologies (e.g., Baird et al, 2008) but not for SNP genotyping assays which remain common in most model species. Extending the model to distinguish demographic from ascertainment processes is theoretically possible (Guillot and Foll, 2009), although it might lead to additional computational burden in practice. Yet, our simulation evaluation showed that the Kimura model was robust to ascertainment bias, if the discovery panel was made of individuals sampled from all the populations, or from the most distant populations. This suggests that the discovery panel needs to contain at least some information about the history of the sampled populations as a whole. Both the AS1 and the AS2 ascertainment schemes considered in the present study are representative of the procedures used in humans, which generally rely on the sequencing of a small subset of individuals from very diverse origins. Conversely, when the discovery panel is only made of individuals from the most recently diverged populations (ascertainment scheme AS3), there is virtually no information for the branches issued from the most ancestral population. More generally, our results suggest that demographic inference should be interpreted with caution, and we recommend accounting for SNP ascertainment bias if the analysed populations are only barely related to the discovery panel. The resulting bias may mainly be related to the over-representation of derived SNPs (i.e., SNPs
that did not exist in the most ancestral population) and is therefore expected to be more pronounced if the populations represented in the discovery panel have rapidly expanded since divergence (e.g., see Figure S3).

Overlooking derived SNPs is expected to affect the robustness of any inference method based on models that neglect recent mutations (e.g., Nicholson et al, 2002; Coop et al, 2010; Gautier, Hocking and Foulley, 2010; Siren, Marttinen and Corander, 2011), although this limitation is generally not explicitly stated. New mutations may occur after the split of the most ancestral population, e.g., if population sizes are large, or if divergence is ancient. Accordingly, the models of population divergence that neglect recent mutations should be used with caution on datasets enriched with derived SNPs.

Finally, we investigated the robustness of our model to departure from pure-drift divergence, by analyzing datasets with simulated with low-to-high levels of gene flow. Interestingly, the correct population tree was generally well supported, even for moderate levels of divergence. Of course, branch lengths were generally biased downward, and the bias increased as the migration rate increased. Alternatively, owing to the flexibility of Bayesian hierarchical modeling, it should be straightforward to account (and to test) for admixture in the model by modifying the priors on the ancestral allele frequencies for the presumably admixed populations.

**Material and Methods**

**Parameter estimation**

Our aim is to estimate divergence times (on a diffusion time-scale) from genome-wide allele frequency data, conditionally on a population tree from the hierarchical-Bayesian model described above. To that end, the full posterior distribution of the parameters is estimated by means of a Metropolis–Hastings within Gibbs MCMC algorithm. In this algorithm, each parameter of interest is updated iteratively. The starting values of the chains are taken as standard moment-based estimates of the parameter of interest. At each iteration $t$ of the algorithm, the $I \times J$ parameters $\alpha_{ik}$ (for $k \leq J$), the $I \times (r - J)$ parameters $\alpha_{ik}$ (for $J < k < r$), the $I$ parameters $\alpha_{ir}$ and the $(r - 1)$ parameters $\tau_k$ are successively updated in that order, following the steps briefly described in the Supplementary Materials.

In practice, in order to deal with probability masses in $\alpha_{ik} = 0$ and $\alpha_{ik} = 1$, we introduced a latent (continuous) variable $\beta_{ik}$ with support $[-1, 2]$, so that $\alpha_{ik} = \min(1, \max(0, \beta_{ik}))$. The model was then
redefined using $\beta_{ik}$ with probability density function $\pi(\beta_{ik} | \alpha_{ia(k)}, \tau_k)$:

$$
\begin{align*}
\pi(\beta_{ik} | \alpha_{ia(k)}, \tau_k) &= P(\alpha_{ik} = 0 | \alpha_{ia(k)}, \tau_k) & \text{if } -1 \leq \beta_{ik} \leq 0 \text{ and } \alpha_{ia(k)} \in (0, 1) \\
\pi(\beta_{ik} | \alpha_{ia(k)}, \tau_k) &= \pi(\alpha_{ik} | \alpha_{ia(k)}, \tau_k) & \text{if } 0 < \beta_{ik} < 1 \text{ and } \alpha_{ia(k)} \in (0, 1) \\
\pi(\beta_{ik} | \alpha_{ia(k)}, \tau_k) &= P(\alpha_{ik} = 1 | \alpha_{ia(k)}, \tau_k) & \text{if } 1 \leq \beta_{ik} \leq 2 \text{ and } \alpha_{ia(k)} \in (0, 1) \\
\pi(\beta_{ik} | \alpha_{ia(k)} = 0, \tau_k) &= 1 & \text{if } \beta_{ik} \leq 0 \\
\pi(\beta_{ik} | \alpha_{ia(k)} = 0, \tau_k) &= 0 & \text{if } \beta_{ik} > 0 \\
\pi(\beta_{ik} | \alpha_{ia(k)} = 1, \tau_k) &= 1 & \text{if } \beta_{ik} \geq 1 \\
\pi(\beta_{ik} | \alpha_{ia(k)} = 1, \tau_k) &= 0 & \text{if } \beta_{ik} < 1
\end{align*}
$$

The proposal distributions for the Metropolis–Hastings updates of the parameters $\beta_{ij}$ and $\tau_j$ are provided in the Supplementary Materials. In order to achieve good convergence of the MCMC, these proposal distributions were adjusted during pilot runs (typically, 25 runs of 1,000 steps were run for each Markov chain). After each pilot run, the proposals were adjusted to increase or decrease the acceptance rate, in order to obtain acceptance rates lying between 0.2 and 0.4 (see, e.g., Gilks, Richardson and Spiegelhalter, 1996). Then, each MCMC was run for 25,000 iterations after a 5,000 iterations burn-in period. Samples were taken from the chain every 25 iterations (thinning) to reduce autocorrelations.

**Model assessment**

Because the tree topology is usually unknown, we were interested in characterizing, for a given dataset, the strength of evidence for alternative tree topologies. To do so, we used the deviance information criterion (DIC), which is a standard criterion for model selection among a finite set of models (Spiegelhalter et al, 2002). It relies on a measure of deviance defined as $D = -2\log (\pi(y | \theta)) + 2\log (\pi(y))$, where $\log (\pi(y | \theta))$ represents the log-likelihood of the data $y$ under the model specified by the parameters $\theta$, and $\pi(y)$ is some fully specified standardizing term, which is function of the data alone. The DIC is then defined as $\text{DIC} = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta})$ where $\bar{D}$ is the posterior mean deviance, which can be interpreted as a Bayesian measure of fit. The effective dimension of the hierarchical model $p_D$ is such that $p_D = \bar{D} - D(\bar{\theta})$, where $D(\bar{\theta})$ is the Bayesian deviance evaluated at the posterior mean of the parameters $\theta$. A DIC difference larger than 10 units between any two models is generally regarded as
strong evidence (in term of predictive ability) in favor of the model with the smallest DIC. Here, the data \( y \) correspond to the allele counts \( \{x_{ij}\} \), and the parameters \( \theta \) correspond to the subset of parameters in the hierarchical models, upon which the data immediately depend, i.e. the \( \alpha_{ij} \)s with \( 1 \leq j \leq J \). Since the latter parameters are not straightforward to integrate out in the model based on Kimura’s diffusion approximation (see above and Supplementary Materials), the DIC was simply computed for all the models considered here as:

\[
\text{DIC} = \frac{2}{T} \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} \log \left( \binom{n_{ij}}{x_{ij}} \alpha_{ij}(t)^{x_{ij}}(1 - \alpha_{ij}(t))^{n_{ij}-x_{ij}} \right) - \sum_{i=1}^{I} \sum_{j=1}^{J} \log \left( \binom{n_{ij}}{x_{ij}} \frac{\alpha_{ij}}{\alpha_{ij}^{\text{post}}}(1 - \frac{\alpha_{ij}}{\alpha_{ij}^{\text{post}}})^{n_{ij}-x_{ij}} \right)
\]

In equation (8), \( \alpha_{ij}(t) \) is the \( t \)th sampled value of the parameter \( \alpha_{ij} \) along the MCMC, out of a total of \( T \) value, and \( \frac{\alpha_{ij}}{\alpha_{ij}^{\text{post}}} \) is the posterior mean of \( \alpha_{ij} \).

**Simulated datasets**

In order to analyze the precision of our model for estimating the level of differentiation over generations, relatively to the beta (Balding and Nichols, 1995) and the truncated Gaussian (Nicholson et al., 2002) models, we used a Wright–Fisher forward-in-time simulation algorithm consisting in successive binomial sampling over generations, as described in Gautier, Hocking and Foulley (2010). In these simulations, we considered four populations diverging simultaneously from a single ancestral population (star-shaped history), each made of 1,000 haploid individuals. We simulated 5,000 SNPs, and the initial reference allele frequencies (in the most ancestral population) were sampled from a uniform distribution: \( U(0.001, 0.999) \). The sample sizes were set to 100 genes per population to allow the accurate estimation of population differentiation parameters in the first generations \( (t < 50) \). All SNPs were retained in that set of simulations, even if fixed.

In order to test the performance of our model in more general settings, we performed additional stochastic simulations, independent from the model assumptions. To that end, simulations were carried out using the coalescent algorithm implemented in the \( ms \) software package (Hudson, 2002). We used the \(-s\) option, which randomly puts a single mutation on the simulated genealogies. Each of these simulated datasets also consisted in 5,000 SNPs, genotyped in 50 diploid individuals (100 genes) per population. As an example, for the tree topology described in Figure 1, which assumes that P1 and P2 derived from population P4 \( \tau_1 = \tau_2 = 0.1 \) units of time ago (on a diffusion time-scale) and that P3 and P4 derived
from the most ancestral one \( \tau_3 = 0.3 \) and \( \tau_4 = 0.1 \) units of time ago, respectively, we used the following \( ms \) command line:

\[
ms 300 5000 -I 3 100 100 100 0 -ej 0.05 2 1 -ej 0.1 3 1 -en 0.1 1 25 -s 1
\]

Note that we assumed that the most ancestral (root) population went through a bottleneck before splitting into P3 and P4, in order to limit the occurrence of SNPs that arose after the divergence of the populations under study from the most ancestral (root) population (see above) which were referred to as derived SNPs.

We then investigated the sensitivity of our model to misspecification, in particular in the presence of gene flow, ascertainment bias and derived SNPs. We tested the effect of model misspecification both on model choice (when comparing alternative tree topologies) and branch length estimates. We first examined the consequence of gene flow on the inference of population histories. To that end, we introduced a migration parameter \( M \equiv 4Nm \) (where \( m \) represents the immigration rate in each generation along the simulated tree). Four values were investigated, namely \( M = 0 \) (corresponding to the simulation model above), \( M = 0.1 \) (slight departure from the pure-drift model), \( M = 1 \) (moderate departure) and \( M = 10 \) (strong departure).

The analysis of SNP data is usually complicated by the discovery protocols applied to ascertain SNPs. Typically, SNPs are called from the genetic material of a small sample of individuals, referred to as the discovery panel. Only then are the ascertained SNPs genotyped in the samples of interest. This procedure results in samples that contain less alleles at low-frequency than expected in the absence of ascertainment (Nielsen, 2000). In order to analyze the consequences of SNP ascertainment bias on the inference of divergence times, we simulated SNP datasets mimicking different ascertainment schemes. For that purpose, we simulated three-populations trees with the \( ms \) program, introducing diploid ”ghost” individuals that were used exclusively for discovery and then discarded from further analyses. Three different ascertainment schemes were considered. In the first (AS1), SNPs were retained if polymorphic in the six ghost individuals (12 genes) originating from P1, P2 and P3. In the second (AS2), SNPs were retained if polymorphic in the four ghost individuals (8 genes) originating from P1 and P3. In the third (AS3), SNPs were retained if polymorphic in the four ghost individuals originating from P1 and P2.
Real dataset

As an illustrative example, we analyzed a subset of the human data from Jakobsson et al (2008) consisting in allele counts at 452,198 autosomal SNPs from four human populations: the Yorubas from Nigeria (YRI, $2n = 72$), the Biaka Pygmies from Congo (BIA, $2n = 64$), the U.S. European Americans from Utah with Northern and Western European ancestry (CEU, $2n = 96$) and the Japanese (JPT, $2n = 32$). The 452,198 SNPs that we retained from the total dataset fulfilled the following conditions: (i) to pass the quality check performed by Jakobsson et al (2008), (ii) to be polymorphic in the total pooled sample and (iii) to be genotyped in at least 95% of individuals from each population.

Acknowledgments

We wish to thank Khalid Belkhir and Remy Dernat (Institut des Sciences de l’Évolution, Montpellier) for making the program available in the MBB platform (http://mbb.univ-montp2.fr/MBB/index.php). We are grateful to three anonymous reviewers for their helpful suggestions and corrections. Part of this research was supported by the French ”Agence Nationale de la Recherche” (grant number 09-BLAN-0145-01).
References


Figure Legends

Figure 1: Directed acyclic graph (DAG) of the hierarchical-Bayesian model for an example tree with \( J = 3 \) populations. The topology ((P1,P2),P3) is represented in grey. \( Y_{ij} \) represents the observed count of the reference allele at the \( i \)th SNP in population \( j \), and \( \alpha_{ij} \) is its (unknown) frequency in that population. The parameter \( \tau_j \equiv t/N_j \) is the length (on a diffusion time-scale) of the branch leading to population \( j \). Following our notations (see the main text), the observed populations (P1, P2 and P3) are indexed from \( j = 1 \) to \( j = J = 3 \). The (unobserved) population ancestral to P1 and P2 is indexed by \( a(1) = a(2) = 4 \) and the (unobserved) population that is ancestral to P3 and P4 (root population) is indexed by \( r = a(3) = a(4) = 5 \).

Figure 2: Estimating branch lengths in star phylogenies with four populations. The posterior means of each of the four branch lengths were estimated with the Kimura model in terms of divergence on a diffusion time-scale \( \tau_k \) (A). With the truncated Gaussian and the beta models (graphs B and C, respectively) divergence was measured as \( F_k \), which corresponds to a population-specific \( F_{ST} \). The datasets consisted in 5,000 SNPs sampled from four populations diverging simultaneously from their common ancestor (star-shaped phylogeny). The data were collected at 24 different time points after divergence (see the main text for further details). In (A), (B) and (C), the dashed lines indicate the (true) divergence time on the diffusion time-scale (here, \( \tau_k \equiv t/(2N_e) \) where \( N_e = 500 \) is the (diploid) effective population size). In (B) and (C), the dotted lines indicates the (true) value of the population-specific \( F_{ST} \) \( (F_k = 1 - [1/(2N_e)]^t) \). Note that for small \( \tau_k \) (e.g., \( \tau_k < 0.15 \)), \( F_k \approx \tau_k \).

Figure 3: Performance of the Kimura model for estimating branch lengths in population trees. All histories share the same topology ((P1,P2),P3), but differ in divergence times. For each history, 50 datasets made of 5,000 SNPs were simulated. The boxplots summarize the distributions of the 50 posterior means of \( \tau_k \) for each of the four branches. The branches are identified by the index of their terminal population and the horizontal dashed lines indicate the (true) simulated values.

Figure 4: Performance of the Kimura model to identify the correct population tree for different levels of gene flow. In all cases, data were simulated according to the T1 history, repre-
sented at the left-hand side of the upper panel. The analyses were then performed conditionally on each of the four possible histories represented in the upper panel (T1, T2, T3 and S). Four levels of gene flow were considered: $M = 0$ (A), $M = 0.1$ (B), $M = 1$ (C) and $M = 10$ (D). In each case, 50 datasets made of 5,000 SNPs were simulated. The boxplots summarize the distributions of the 50 posterior means of $\tau_k$ for each branch. The branches are identified by the index of their terminal population and the horizontal dashed lines indicate the (true) simulated values.

**Figure 5:** Performance of the Kimura model to identify the correct population tree for different SNP ascertainment schemes. In all cases, data were simulated according to the T1 history, represented at the left-hand side of the upper panel. The analyses were then performed conditionally on each of the four possible histories represented in the upper panel (T1, T2, T3 and S). Three different SNP ascertainment schemes were considered: AS1 (A), AS2 (B) and AS3 (C) (see the main text for further details on the ascertainment schemes). In each case, 50 datasets made of 5,000 SNPs were simulated. The boxplots summarize the distributions of the 50 posterior means of $\tau_k$ for each branch. The branches are identified by the index of their terminal population and the horizontal dashed lines indicate the (true) simulated values.

**Figure 6:** Estimation of divergence times conditionally on different histories relating four human populations: the Yorubas from Nigeria (YRI), the Biaka Pygmies from Congo (BIA), the U.S. European Americans from Utah with Northern and Western European ancestry (CEU) and the Japanese (JPT). The dataset from Jakobsson et al (2008) consisted in allele counts at 452,198 autosomal SNPs. We analyzed the data using the Kimura model conditionally on five rooted bifurcating topologies (each of which was derived by placing the root on one of the five branches of the most likely unrooted tree, following the four-population test for treeness) and a star phylogeny. The estimated posterior means of the divergence time (on a diffusion time-scale) are provided for each branch. The topology that received the strongest support based on the DIC is highlighted in gray (upper-left panel).
Table 1. Strength of evidence for the different topologies considered in Figures 4 and 5, based on the DIC. This Table reports the median (minimum ; maximum) of the distribution of the difference between the DIC ($\Delta_{DIC}$) of a wrong topology (T2, T3 or S) and the correct one (T1), over 50 replicated simulations. $n_{10}$ gives the number of simulations (out of 50) where $\Delta_{DIC} > 10$.

<table>
<thead>
<tr>
<th>Topology</th>
<th>DIC$<em>{T2}$ - DIC$</em>{T1}$ median (min. ; max.)</th>
<th>$n_{10}$</th>
<th>DIC$<em>{T3}$ - DIC$</em>{T1}$ median (min. ; max.)</th>
<th>$n_{10}$</th>
<th>DIC$<em>{S}$ - DIC$</em>{T1}$ median (min. ; max.)</th>
<th>$n_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 4A ($M = 0$)</td>
<td>72.0 (45.9 ; 94.7)</td>
<td>50</td>
<td>68.2 (47.3 ; 100.6)</td>
<td>50</td>
<td>69.8 (35.9 ; 94.2)</td>
<td>50</td>
</tr>
<tr>
<td>Figure 4B ($M = 0.1$)</td>
<td>72.7 (44.3 ; 103.2)</td>
<td>50</td>
<td>73.5 (44.6 ; 106.1)</td>
<td>50</td>
<td>65.9 (38.5 ; 107.8)</td>
<td>50</td>
</tr>
<tr>
<td>Figure 4C ($M = 1$)</td>
<td>61.0 (30.7 ; 87.0)</td>
<td>50</td>
<td>61.5 (33.2 ; 94.5)</td>
<td>50</td>
<td>57.1 (26.8 ; 87.0)</td>
<td>50</td>
</tr>
<tr>
<td>Figure 4D ($M = 10$)</td>
<td>26.5 (-13.2 ; 58.8)</td>
<td>40</td>
<td>27.6 (-5.2 ; 56.5)</td>
<td>43</td>
<td>30.6 (-0.7 ; 63.2)</td>
<td>41</td>
</tr>
<tr>
<td>Figure 5A (AS1)</td>
<td>65.8 (36.9 ; 88.5)</td>
<td>50</td>
<td>68.7 (31.4 ; 97.1)</td>
<td>50</td>
<td>69.7 (36.9 ; 99.5)</td>
<td>49</td>
</tr>
<tr>
<td>Figure 5B (AS2)</td>
<td>81.3 (44.0 ; 106.6)</td>
<td>50</td>
<td>98.8 (66.4 ; 126.5)</td>
<td>50</td>
<td>101 (65.5 ; 135)</td>
<td>50</td>
</tr>
<tr>
<td>Figure 5C (AS3)</td>
<td>10.6 (-22.2 ; 41.8)</td>
<td>25</td>
<td>8.80 (-26.3 ; 38.2)</td>
<td>22</td>
<td>10.3 (-17.6 ; 44)</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 2. Effect of the number of SNPs on the DIC. Data were simulated according to the same history as in Figure 3A, with 1,000, 5,000 and 25,000 SNPs. Each column reports the median (minimum : maximum) of the distribution of the difference (\(\Delta_{DIC}\)) between the DIC of a wrong tree (T2, T3 or S) and the correct one (T1), over 50 replicated simulations. \(n_{10}\) gives the number of simulations (out of 50) where \(\Delta_{DIC} > 10\)
Figure 1
A) Kimura model

$$\tau_k = \frac{t}{2N_k}$$

$$F_k = 1 - \left(1 - \frac{1}{Ne}\right)^t$$

B) Truncated Gaussian model

$$\tau_k = \frac{t}{2N_k}$$

$$F_k = 1 - \left(1 - \frac{1}{Ne}\right)^t$$

C) Beta model

$$\tau_k = \frac{t}{2N_k}$$

$$F_k = 1 - \left(1 - \frac{1}{Ne}\right)^t$$

Figure 2
Figure 3
Figure 5
Figure 6