The Annals of Applied Statistics
2016, Vol. 10, No. 4, 2047–2074
DOI: 10.1214/16-AOAS964

INFERRING ROOTED POPULATION TREES USING ASYMMETRIC NEIGHBOR JOINING

BY YONGLIANG ZHAI AND ALEXANDRE BOUCHARD-CÔTÉ[1]

University of British Columbia

We introduce a new inference method to estimate evolutionary distances for any two populations to their most recent common ancestral population using single-nucleotide polymorphism allele frequencies. Our model takes fixation into consideration, making it nonreversible, and guarantees that the distribution of reconstructed ancestral frequencies is contained on the interval [0, 1]. To scale this method to large numbers of populations, we introduce the asymmetric neighbor joining algorithm, an efficient method for reconstructing rooted bifurcating nonclock trees. Asymmetric neighbor joining provides a scalable rooting method applicable to any nonreversible evolutionary modeling setups. We explore the statistical properties of asymmetric neighbor joining, and demonstrate its accuracy on synthetic data. We validate our method by reconstructing rooted phylogenetic trees from the Human Genome Diversity Panel data. Our results are obtained without using an outgroup, and are consistent with the prevalent recent single-origin model.

1. Introduction. Recovering the path of human migration has long been a fascinating research area for researchers in many areas including archeology, anthropology, linguistics and biology [Lipo (2006)]. With the rapid expansion of genetic data available [Li et al. (2008), Pickrell et al. (2012)], statistical methods play an increasingly important role in inferring the history of human populations [Felsenstein (1983), Mau, Newton and Larget (1999), Li, Pearl and Doss (2000), Huelsenbeck et al. (2001)].

Information on human migrations can be informed by the reconstruction of a *phylogenetic tree*, a connected acyclic graph consisting of a set of *vertices* and a set of *edges* [Semple and Steel (2003)]. See Figure 1 for illustrations of different types of phylogenetic trees. The phylogenetic tree cannot be observed directly and is often the main parameter of interest to estimate, as data is usually only available at current populations or species, which are called *leaves* of the phylogenetic tree, in the form of genetic information such as molecular polymorphism and deoxyribonucleic acid (DNA) sequences.

The phylogenetic tree can be reconstructed using a range of methods. Distance-based methods measure the similarity between current populations [Nei (1972),

Received November 2015; revised June 2016.

[1]Supported in part by a Discovery Grant from the National Science and Engineering Research Council.

Key words and phrases. Asymmetric neighbor-joining algorithm, fixation and drift, phylogenetics, population histories, rooted tree inference, single-nucleotide polymorphism.

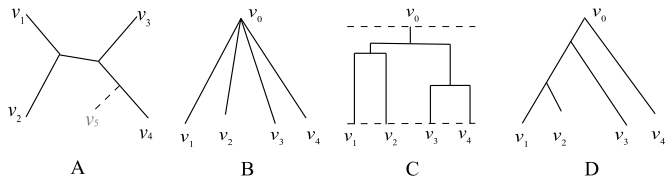


FIG. 1. (A): *An unrooted bifurcating phylogenetic tree of four leaves with an outgroup* v_5. (B): *A star-shaped phylogenetic tree with root* v_0. (C): *A rooted tree assuming a molecular clock assumption*. (D): *A rooted bifurcating phylogenetic tree without the molecular clock assumption*.

Weir and Cockerham (1984)] and then construct the tree based on the similarity using unrooted tree construction algorithms such as neighbor joining (NJ) [Saitou and Nei (1987)]. Probability-based methods aim to model the evolutionary process along the branches of the tree [Felsenstein (1981), Hasegawa, Kishino and Yano (1985), Tavaré (1986)] as a stochastic process. These stochastic processes can be either the basis of evolutionary distances or of a joint likelihood suitable for maximum likelihood [Felsenstein (1981)] or Bayesian inference [Li, Pearl and Doss (2000)].

Determining the *root* of a phylogenetic tree is an important step in a range of phylogenetic studies [Huelsenbeck, Bollback and Levine (2002)]. For example, the root of human population trees provides clues on the origin and the path of human migration [Gray, Drummond and Greenhill (2009)], and the root of the tree of life is used to study the origins of life via reconstruction methods. However, most classic methods estimate unrooted trees only. Distance-based methods depend on symmetric similarity measures and classic probability-based methods depend on reversible continuous time Markov chain (CTMC) models. Both approaches ignore the direction of evolution, and thus are unable to identify the root of the tree directly.

There are three main methods for constructing a rooted phylogenetic tree: assuming a molecular clock assumption (or a relaxation of the molecular clock assumption), adding an outgroup, or using nonreversible models. The clock tree assumes that all leaves are equally distant from the root [see Figure 1(C)]. This is unrealistic, as it ignores evolution rate variation which depends on factors such as the population size [Swofford et al. (1996), Huelsenbeck, Bollback and Levine (2002)]. There is a large literature on relaxing the clock assumption while modeling rooting, but these methods are generally computationally expensive [Battistuzzi et al. (2010)].

The outgroup criterion is also widely used to root an unrooted phylogenetic tree: the idea is to find a taxon (species or population) sufficiently far from those under study and to use the attachment point of this taxon to the other taxa to infer the root [see Figure 1(A)]. However, finding a reasonable outgroup is not always easy. Previous works have shown that the outgroup criterion can lead to errors when the outgroup is distant [Wheeler (1990), Pearson et al. (2013)], or when the traits are changing at a high rate [Outlaw and Ricklefs (2011)], or when the outgroup does

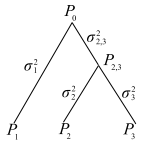
not exist or is unknown, for example, when studying the tree of life [Iwabe et al. (1989)], or linguistic families [Gray, Drummond and Greenhill (2009)].

We propose a novel nonreversible model for rooted bifurcating tree [see Figure 1(D)] inference, with application to the evolutionary process of single nucleotide polymorphism (SNP) frequencies. Our model is motivated by modeling the main forces that shape the patterns of variations found in SNP frequencies across populations: mutation, which created variations in ancestral populations; as well as drift and fixation, which determine the allele frequencies in modern populations. Among these forces, drift over time of allele frequencies, which is usually modeled as a reversible process, has been the focus of classical probability models. Brownian motion, in particular, provides a simple and tractable approximation of the Wright–Fisher model [Edwards and Cavalli-Sforza (1964), Felsenstein (1973, 1981), Pickrell and Pritchard (2012)].

While the reversible Brownian approximation is accurate for frequencies bounded away from zero and one, it breaks down at the extremities. The Brownian approximation fails in this regime because it ignores *fixation*: the simple observation that if all individuals in a population share the same allele, then the drift is fixed for a significant time period (until new mutations create subsequent drifts). Moreover, the distribution of reconstructed ancestral states according to the Brownian motion assigns positive probability to frequencies smaller than zero and greater than one. This limitation has motivated the development of more sophisticated models which track allele counts in a more detailed way through coalescent theory. This brings fixation back into the model, thus nonreversibility, but at the cost of a significant increase in computational requirements and model complexity. As a consequence, most current probability models either ignore fixation or otherwise are limited to inference in small sets of populations at a time [RoyChoudhury, Felsenstein and Thompson (2008)].

In this work, we present a tree reconstruction method that models both the fixation and drift found in SNP frequency data while keeping the computational requirements and the model complexity manageable. Our method is based on a likelihood model that considers only valid ancestral frequencies. Being nonreversible, the method has the additional advantage of being able to identify the placement of the root without using outgroups. Our approach consists two steps: first, relevant evolutionary parameters are estimated under the nonreversible model based on the maximum likelihood principle, then, a rooted phylogenetic tree is constructed via a novel efficient rooted-tree reconstruction algorithm called the asymmetric neighbor-joining (ANJ) algorithm. See Figure 2 for an illustration (with details in Sections 3 and 4). ANJ is inspired by the popular neighbor-joining (NJ) algorithm, but utilizes the additional information provided by our nonreversible model.

The main advantage of using ANJ versus using full maximum likelihood or Bayesian methods for estimating phylogenetic tree is computational efficiency, as it is well known that optimization over all possible trees is computationally hard [Roch (2006), Nichols and Warnow (2008)]. However, it is worth mentioning that



$$\Leftrightarrow \quad \mathbf{A} = \begin{pmatrix} \infty & \sigma_1^2 & \sigma_1^2 \\ \sigma_2^2 + \sigma_{2,3}^2 & \infty & \sigma_2^2 \\ \sigma_3^2 + \sigma_{2,3}^2 & \sigma_3^2 & \infty \end{pmatrix}$$

FIG. 2. *An example of a rooted bifurcating tree T for 3 populations and the asymmetric dissimilarity matrix $\mathbf{A}$ with relevant parameters derived from this tree where $a_{i,j}$ is the sum of evolutionary distance from the ith leave to the most recent common ancestor of the ith leaf and the jth leaf. The first step of our inference method is to estimate $\mathbf{A}$ based on a nonreversible model using maximum likelihood. The second step is to estimate T based on $\widehat{\mathbf{A}}$ using ANJ.*

incorporating maximum likelihood methods and Bayesian methods with our proposed nonreversible model is also feasible. The ANJ algorithm presented here is broadly applicable to any nonreversible setup: ANJ can turn collections of pairwise rooted trees obtained from any nonreversible model into a joint rooted nonclock tree. It is therefore an interesting alternative to the outgroup criterion.

We perform a series of simulation studies to assess the accuracy and robustness of the methods. The results show that our method can recover true topologies and root placements with higher probability than models that ignore fixation. Branch lengths are also shown to be accurate. Our method also has a reasonable computational cost.

Using our method, we reanalyze the SNP data from 53 populations from the Human Genome Diversity Panel (HGDP) [Cann et al. (2002)], reconstructing the full tree with particular attention to the location of the root. The internal organization of the tree is similar to Li et al. (2008), but also corrects some problems found in this previous work.

There has been earlier work on the construction of tractable methods taking fixation into account, in particular the work of Nicholson et al. (2002) which our model builds on, but this previous work has been restricted to star-shaped trees of closely related populations [see Figure 1(C)]. There is also a rich literature on exact calculation of marginal densities of stochastic processes taking fixation into account [for example, Song and Steinrücken (2012)] and on simulation of these processes [Jenkins and Spano (2015)]. These methods can in principle be used to estimate bifurcating trees (for example, using a particle MCMC framework [Wang, Bouchard-Côté and Doucet (2015)]), but at a significant computational cost. While there has been considerable progress on improving the scalability of these methods via advanced Monte Carlo and numerical methods [RoyChoudhury, Felsenstein and Thompson (2008), Bryant et al. (2012)], these other methods would not easily process the datasets used in the present paper, where the number of populations is large. For example, in a recent paper, Bryant et al. (2012) reduced the time complexity of a multispecies coalescent model likelihood calculation to $O(LnN^2 \log N)$ for L SNPs, where n is the number of populations and N is the total number of individuals (typically, $N \gg n$). Our method models frequencies of

populations directly, which is beneficial when the number of individuals sampled in each population are large. The time complexity of our algorithm is $O(Ln^2 + N)$.

2. Evolutionary models. In this section, we briefly review evolutionary models for frequency data and propose a new approximative model for genetic drift, called the normal approximation with general fixation.

We focus on independent bi-allelic sites in this paper.

2.1. *Review of evolutionary models for frequency data.* Evolutionary models play a fundamental role in phylogenetics. The Wright–Fisher model and its approximations are widely used evolutionary models for frequency data. Consider one locus with two alleles. Suppose that there are N_0 copies of one allele in N haploid individuals. If the number of haploid individuals N is constant, then, according to the Wright–Fisher model, the number of copies of this allele in the next generation, N_1, follows a binomial distribution,

$$N_1|N_0 \sim \text{Binomial}(N, N_0/N).$$

More generally, let p_0 be the frequency of a binary allele in an ancestral population. If the number of individuals in each generation is constant, then the frequency of this allele after t generations, $p(t)$, can be modeled by a sequence of binomial distributions using the Wright–Fisher model. We write $p(t)$ as p for simplicity throughout this paper.

The distribution of p depends on both the number of generations t and the initial frequency $p_0 = N_0/N$. For modeling the evolutionary history of all human populations, the number of generations from the most recent common ancestor of human populations to current human populations is large. As a result, calculating the distribution of p is computationally expensive.

Edwards and Cavalli-Sforza (1964), Felsenstein (1973, 1981) used a Brownian motion approximation for the Wright–Fisher model, that is,

$$p \sim \text{Normal}\big(p_0, \sigma_t^2\big), \qquad 0 < p_0 < 1, \tag{2.1}$$

where σ_t^2 is a measure of drift strength combining the effects of t and N. Larger N and smaller t lead to smaller σ_t^2, that is, smaller drift. Note that t and N are not identifiable in the normal approximation of the Wright–Fisher model and the variance σ_t^2 depends on the frequency p_0, which can be further stabilized by a transformation [Felsenstein (1973)] for frequencies which are not close to 0 and 1.

A main drawback of the Brownian motion approximation is that p is not bounded in [0, 1]. The normal approximation works well if the evolution distance from the parent population to the child population is short, but the probability of breaking the boundary 0 or 1 is not negligible when the evolution distance is long.

Nicholson et al. (2002) proposed a normal approximation with fixation for the Wight–Fisher model which takes the probability of fixation in child populations into consideration:

$$p \sim \text{Normal}\big(p_0, p_0(1 - p_0)\sigma^2\big), \qquad 0 < p_0 < 1, \tag{2.2}$$

constrained in [0, 1] with atom mass on 0 and 1 equal to the total mass of relevant distribution on $(-\infty, 0)$ and $(1, \infty)$, respectively. In other words, p is a mixture of continuous and discrete random variable which has a Normal density in $(0, 1)$, and point masses at 0 and 1 respectively modeling the probability of $p < 0$ and $p > 1$ derived from the Normal density. In model (2.2), the component of the variance which depends on p_0, $p_0(1 - p_0)$, is separated from the other component σ^2, which measures the strength of the genetic drift.

Balding and Nichols (1995) used a beta distribution which matches the first two moments of the Brownian motion model as an approximation to the Wright–Fisher diffusion for genetic drift in island populations. This model is also widely used in modeling the effects of genetic drift of SNP frequencies [Sirén, Marttinen and Corander (2011), Sirén, Hanage and Corander (2013)].

The work of Nicholson et al. (2002) and Balding and Nichols (1995) both impose restrictions on the allele frequencies at internal nodes. More precisely, while both methods allow the full range of frequencies at the leaves, $[0, 1]$, the frequencies of all alleles at any internal nodes are restricted to $(0, 1)$. While SNPs are by definition polymorphic at the root of the tree, requiring each of them to be polymorphic in all internal nodes is limiting for the purpose of medium or large population tree reconstruction.

One should also keep in mind that the Wright–Fisher model is itself an approximation for the genetic drift rather than an exact model of the genetic drift.

2.2. *Normal approximation with general fixation.* We generalize an approximative model to the Wright–Fisher model for genetic drift [Nicholson et al. (2002)] by extending the domain of the ancestral frequencies p_0 from $(0, 1)$ into $[0, 1]$. We denote this evolution model as

$$p \sim \text{FixNormal}\big(p_0, \sigma^2\big), \qquad 0 \le p_0 \le 1,$$

where σ^2 is the drift parameter which models the strength of genetic drift. We call this model the normal approximation with general fixation.

The density f_p of p under the FixNormal model has the following form when $0 < p_0 < 1$:

$$f_p\big(p|p_0, \sigma^2\big) = \phi\bigg(\frac{p - p_0}{\sigma_0\sigma}\bigg), \qquad 0 < p < 1, \tag{2.3a}$$

$$f_p\big(0|p_0, \sigma^2\big) = \int_{-\infty}^{0} \phi\bigg(\frac{t - p_0}{\sigma_0\sigma}\bigg)\,dt = \Phi\bigg(\frac{-p_0}{\sigma_0\sigma}\bigg), \tag{2.3b}$$

$$f_p\big(1|p_0, \sigma^2\big) = \int_{1}^{\infty} \phi\bigg(\frac{t - p_0}{\sigma_0\sigma}\bigg)\,dt = 1 - \Phi\bigg(\frac{1 - p_0}{\sigma_0\sigma}\bigg), \tag{2.3c}$$

where $\sigma_0 = \sqrt{p_0(1 - p_0)}$, ϕ and Φ are the probability density and cumulative distribution functions of the standard normal random variable. Note that the above density is defined with respect to a reference measure composed of a uniform

measure on $(0, 1)$ superposed with a unit point mass at each boundary point $\{0, 1\}$. This part of the density f_p is the same as that of the model of Nicholson et al. (2002).

Under the assumption that mutations are rare, we further assume that fixation is not reversible. In other words, once fixed, the frequency at one location will not change. Under this assumption,

$$\text{(2.4a)} \quad f_p(0|0,\sigma^2)=1, \qquad f_p(p|0,\sigma^2)=0, \qquad 0<p\le 1, \sigma^2>0,$$

$$\text{(2.4b)} \quad f_p(1|1,\sigma^2)=1, \qquad f_p(p|1,\sigma^2)=0, \qquad 0\le p<1, \sigma^2>0.$$

Since our model does not require the variation observed at current populations to be present in ancestral populations, it allows fixation in both current populations and ancestral populations. This simple relaxation makes the new model a reasonable choice for modeling SNP frequencies among distantly related populations, and also has an important impact on the likelihood model and inference methods.

2.3. *Comparisons to the Wright–Fisher model.* We briefly comment on the differences of our evolutionary model and the Wright–Fisher model. See the Supplementary Material [Zhai and Bouchard-Côté (2016)] (Section 1) for simulation studies. When σ^2 is small or when p_0 is close to 0.5, the difference between the distributions of p from our model and the Wright–Fisher model is negligible. When p is close to 0 or 1, and σ^2 is very large, there is a larger difference between the two models.

However, this limitation is not a major concern in the types of applications we are interested in. First, we are more interested in σ^2 rather than p. We do not need to estimate both p and σ^2. Since we marginalize over the values of p, the effects of those p which are close to 0 will be offset by those p which are close to 1, especially after the symmetric transformation introduced in the next section of our inference method. Second, in real problems, reasonable values of σ^2 are not too large (mostly between 0 and 2 in our data analysis). One should also keep in mind that the Wright–Fisher model is itself an approximation for the genetic drift rather than an exact model of the genetic drift.

3. Likelihood model and inference method.

3.1. *Inference method for two populations.* In this paper, we assume the phylogenetic tree is bifurcating (see Figure 2 for an example with three leaves). We first focus on a pair of populations and propose a new likelihood-based method to estimate evolutionary distances from each of the two current populations to their MRCA.

We denote the SNP frequencies in two current populations and their MRCA as $p_{i,1}$, $p_{i,2}$ and $p_{i,0}$ $(i = 1, 2, \ldots, L)$, respectively. We only consider independent SNP sites in this paper. We use two sets of independent normal approximations

with general fixation for the genetic drift after the separation of two populations, that is,

$$p_{i,1} \sim \text{FixNormal}(p_{i,0}, \sigma_1^2) \quad \text{and} \quad p_{i,2} \sim \text{FixNormal}(p_{i,0}, \sigma_2^2). \tag{3.1}$$

Note that $p_{i,1}$ and $p_{i,2}$ are conditionally independent given $p_{i,0}$ ($i = 1, 2, \ldots, L$). We also assume that the variances do not depend on the SNP sites, that is, all SNP sites share the same σ_1^2 and σ_2^2. The log-likelihood function for two populations is given by

$$l(\mathbf{p}_0, \sigma_1^2, \sigma_2^2) = \sum_{i=1}^{L} \{\ln f_p(p_{i,1}|p_{i,0}, \sigma_1^2) + \ln f_p(p_{i,2}|p_{i,0}, \sigma_2^2)\}, \tag{3.2}$$

where $\mathbf{p}_0 = (p_{1,0}, p_{2,0}, \ldots, p_{L,0})$, and f_p is the density function defined in (2.3) and (2.4).

For a pair of populations, there are $2L$ samples and $L + 2$ unknown parameters, σ_1^2, σ_2^2, which are univariate, and $\mathbf{p}_0$, which is L-dimensional. It is hard to estimate $p_{i,0}$ accurately from two samples $p_{i,1}$ and $p_{i,2}$ even if σ_1^2 and σ_2^2 are known. However, to estimate σ_1^2 and σ_2^2, which are shared by all loci, we can avoid estimating $\mathbf{p}_0$ by modeling its distribution. The effects of model misspecification are illustrated in the simulation studies.

3.2. *Modeling ancestral allele frequencies.* For one SNP locus, we can write the density function of (p_1, p_2) as

$$\begin{aligned} f(p_1, p_2|\sigma_1^2, \sigma_2^2) &= \int_0^1 \pi(\mathrm{d}p_0) f_p(p_1|p_0, \sigma_1^2) f_p(p_2|p_0, \sigma_2^2) \\ &= \int_0^1 \pi^*(\mathrm{d}p_0) f_p(p_1|p_0, \sigma_1^2) f_p(p_2|p_0, \sigma_2^2) \\ &\quad + P(p_0 = 0) f_p(p_1|0, \sigma_1^2) f_p(p_2|0, \sigma_2^2) \\ &\quad + P(p_0 = 1) f_p(p_1|1, \sigma_1^2) f_p(p_2|1, \sigma_2^2), \end{aligned} \tag{3.3}$$

where π is the distribution of p_0 defined on [0, 1], and π^* is the continuous part of π defined on (0, 1). Let $m_0 = P(p_0 = 0)$, $m_1 = P(p_0 = 1)$, and $m_2 = \int_0^1 \pi^*(p_0)\, dp_0$ with a constraint $m_0 + m_1 + m_2 = 1$.

If an allele is randomly chosen to measure the frequency, it is reasonable to assume that π is symmetric with respect to 0.5.[1] Under the symmetric assumption of π, we can further combine m_0 and m_1 into one parameter $m_f \equiv 2m_0 = 2m_1$ because $m_0 = m_1$ when π is symmetric. Then $m_{nf} \equiv 1 - m_f$ is the proportion of

[1] If the allele used to measure the frequency is not randomly chosen (for example, it is chosen to be the minor allele in the discovery panel of the SNP ascertainment process as HGDP), we can perform a symmetric transformation by inverting half of the frequencies from p into $1 - p$.

unfixed SNPs in the ancestral population. The density of (p_1, p_2) in (3.3) can be simplified as

$$
\begin{aligned}
f(p_1, p_2|\sigma_1^2, \sigma_2^2) &= \int_0^1 \pi^*(p_0) f_p(p_1|p_0, \sigma_1^2) f_p(p_2|p_0, \sigma_2^2)\, dp_0 \\
&\quad + \frac{m_f}{2}\{f_p(p_1|0, \sigma_1^2) f_p(p_2|0, \sigma_2^2) \\
&\quad + f_p(p_1|1, \sigma_1^2) f_p(p_2|1, \sigma_2^2)\}.
\end{aligned} \tag{3.4}
$$

The log-likelihood of (σ_1^2, σ_2^2) for L frequencies $\mathbf{p}_1 = (p_{1,1}, p_{2,1}, \ldots, p_{L,1})$ and $\mathbf{p}_2 = (p_{1,2}, p_{2,2}, \ldots, p_{L,2})$ is given by

$$
l(\sigma_1^2, \sigma_2^2|\mathbf{p}_1, p_2) = \sum_{i=1}^{L} \ln f(p_{i,1}, p_{i,2}|\sigma_1^2, \sigma_2^2). \tag{3.5}
$$

We can maximize the log-likelihood (3.5) to find the maximum likelihood estimate $(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2)$.

In real data analysis, we do not observe $p_{i,1}$ and $p_{i,2}$. We observe the number of sampled individuals $n_{i,1}$ and $n_{i,2}$ with the number of certain alleles $x_{i,1}$ and $x_{i,2}$. We can integrate out p_1 and p_2 in (3.4) after adding two binomial probability mass functions. In this paper, we focus on population frequencies directly and simply use $\widehat{p}_{i,l} = x_{i,l}/n_{i,l}$ $(i = 1, 2, \ldots, L$ and $l = 1, 2)$ as data in the inference.

In this paper, we use the Uniform$(0, 1)$ distribution multiplied by a factor m_{nf} to model π^*. If there is extra information on the frequencies of certain ancestral populations, it could also be easily incorporated into our model. Other distributions motivated by population genetics such as $\pi^*(x) \propto 1/\{x(1-x)\}$ for $x \in (0, 1)$ [Ewens (1973)] can also be used.

For the values of m_f, we propose an empirical Bayes estimator to estimate m_f from data which works well in our simulation studies and data analysis. We estimate

$$
\widehat{m}_0 = \frac{1}{L}\sum_{i=1}^{L} I(p_{i,1} = 0) I(p_{i,2} = 0) \quad \text{and} \quad \widehat{m}_1 = \frac{1}{L}\sum_{i=1}^{L} I(p_{i,1} = 1) I(p_{i,2} = 1),
$$

where I is an indicator function and estimate

$$
\widehat{m}_f = \widehat{m}_0 + \widehat{m}_1. \tag{3.6}
$$

Note that $\widehat{m}_0 \geq m_0$ since, if $p_{i,0} = 0$, then $p_{i,1} = p_{i,2} = 0$, but $p_{i,1} = p_{i,2} = 0$ does not imply $p_{i,0} = 0$ $(i = 1, 2, \ldots, L)$. Similarly, $\widehat{m}_1 \geq m_1$, and thus $\widehat{m}_f \geq m_f$. The bias $\widehat{m}_f - m_f$ increases when σ_1^2 and σ_2^2 are larger. It is also possible to estimate m_f as a parameter together with σ_1^2 and σ_2^2, which is not the focus of this paper.

In Section 5, we show that using Uniform$(0, 1)$ with fixation rate m_f works well for a variety of distributions on p_0 even when the model is misspecified, and that the choice of the weight m_f has a negligible effect in the estimation of σ_1^2, σ_2^2 as long as m_f is nonzero.

3.3. *Inference method for n populations*. Suppose that we have SNP frequencies at L independent SNP loci for n populations. We denote the SNP frequency of the ith locus in the jth population by p_{ij} ($i = 1, 2, \ldots, L$, and $j = 1, 2, \ldots, n$).

3.3.1. *Asymmetric dissimilarity matrix*. Our inference method for n populations is a generalization of our method for 2 populations. For n populations, we can estimate the branch lengths of any two populations to their MRCA. This information is encoded in an asymmetric dissimilarity matrix $\mathbf{A}$. In this section, we discuss the relationship between asymmetric dissimilarity matrices and bifurcating trees.

DEFINITION 1. A matrix $\mathbf{A}$ is said to be an asymmetric dissimilarity matrix representation of a rooted bifurcating tree T of n populations if $\mathbf{A}$ is an $n \times n$ matrix, with $a_{i,i} = \infty$ ($i = 1, 2, \ldots, n$), and $a_{i,j}$ is equal to the sum of branch lengths from population i to the most recent common ancestral population of population i and population j ($i, j = 1, 2, \ldots, n$ and $i \neq j$).

It is easy to see that, for a given rooted bifurcating tree T, the asymmetric dissimilarity matrix $\mathbf{A}$ is uniquely defined. Figure 2 provides an example for 3 populations. In the tree T, branch lengths are defined on the topology of the tree, and they measure the evolutionary distance from the parent population to the child population. These branch lengths are modeled by the drift parameter σ^2 in our evolutionary model.

From the SNP frequencies of n populations, we can calculate an $n \times n$ asymmetric dissimilarity matrix $\widehat{\mathbf{A}}_L$ by setting

$$\begin{aligned} \widehat{a}_{ii} &= \infty, \qquad 1 \le i \le n, \\ \widehat{a}_{ij} &= \widehat{\sigma}^2_{ij}, \qquad 1 \le i, j \le n, \text{ and } i \neq j, \end{aligned}$$

where $\widehat{\sigma}^2_{ij}$ and $\widehat{\sigma}^2_{ji}$ are estimates of the branch lengths, respectively, from the ith population and the jth population to their MRCA using our inference method proposed in Section 3.1.

Before claiming that $\widehat{\mathbf{A}}_L$ is a reasonable estimate of $\mathbf{A}$, we need one more assumption on the additivity of the branch lengths in our model to maintain the compatibility of the model and the bifurcating tree structure. To illustrate the compatibility of the model and the tree, we take three populations as an example. In Figure 2, there are three populations P_1, P_2 and P_3, and two internal populations P_0 and $P_{2,3}$. Based on the assumption that after the separation of two populations each branch evolves independently, we model the genetic drift by

$$\text{(3.7a)} \quad p_{i,1} \sim \text{FixNormal}\big(p_{i,0}, \sigma_1^2\big), \qquad p_{i,23} \sim \text{FixNormal}\big(p_{i,0}, \sigma_{2,3}^2\big),$$

$$\text{(3.7b)} \quad p_{i,2} \sim \text{FixNormal}\big(p_{i,23}, \sigma_2^2\big), \qquad p_{i,3} \sim \text{FixNormal}\big(p_{i,23}, \sigma_3^2\big).$$

Under this set of assumptions, we can estimate σ_2^2 and σ_3^2 by $\widehat{\sigma}_{23}^2$ and $\widehat{\sigma}_{32}^2$, but we cannot estimate $\sigma_{2,3}^2$ using our inference method, as we do not have data for internal population $P_{2,3}$.

If we further assume

$$\begin{aligned} p_{i,2} &\sim \text{FixNormal}(p_{i,0}, \sigma_2^2 + \sigma_{2,3}^2), \\ p_{i,3} &\sim \text{FixNormal}(p_{i,0}, \sigma_3^2 + \sigma_{2,3}^2), \end{aligned} \tag{3.8}$$

we can estimate $\sigma_{2,3}^2 + \sigma_2^2$ by $\widehat{\sigma}_{21}^2$, which is the total branch length from P_2 to P_0 since P_0 is the MRCA of P_1 and P_2. Similarly, we can estimate $\sigma_{2,3}^2 + \sigma_3^2$ by $\widehat{\sigma}_{31}^2$.

We emphasize that, in our model, just as in the previous work [Balding and Nichols (1995), Nicholson et al. (2002), Pickrell and Pritchard (2012)], assumptions of the form of (3.7a–b) do not imply that $p_{i,2}$ and $p_{i,3}$ are exactly distributed as shown in (3.8). However, in the Supplementary Material (Section 2), we show that approximation (3.8) is reasonable in practice.

4. Asymmetric neighbor-joining algorithm. We propose a new algorithm, called asymmetric neighbor joining (ANJ), to construct a rooted bifurcating tree based on asymmetric dissimilarity matrices $\mathbf{A}$ or its estimate $\widehat{\mathbf{A}}_L$. ANJ can be applied to the matrix derived using our methods and also to any asymmetric dissimilarity matrix derived by other methods. ANJ shares many of NJ's desirable properties.

ANJ proceeds in three steps: search step, estimation step and update step. At the search step, we propose a new cherry-picking algorithm to identify a pair of populations which forms a *cherry* in the tree, that is, two distinct leaves in a tree which are adjacent to a common vertex [Semple and Steel (2003)]. At the estimation step, we estimate the branch lengths from two leaves of the cherry to the root of the cherry. At the update step, we combine the two populations in the cherry into one new internal population, and update the distances related to this internal population.

Proofs can be found in the Supplementary Material (Section 7).

4.1. *Identifying a cherry.* First, we show that a cherry, which is a pair of adjacent external vertices on the tree T, can be identified by comparing elements in an asymmetric dissimilarity matrix $\mathbf{A}$ induced by the tree T in Proposition 1.

DEFINITION 2. The two-matrix row minimization condition (2-MRMC) is satisfied for the ith and jth row of a matrix $\mathbf{A}$ if a_{ij} is the smallest entry in the ith row of $\mathbf{A}$ and a_{ji} is the smallest entry in the jth row of $\mathbf{A}$.

PROPOSITION 1. *For any square matrix* $\mathbf{A}$ *that is induced by a rooted bifurcating tree* T, *the* 2*-MRMC is satisfied for* (P_i, P_j) *if and only if* P_i *and* P_j *form a cherry.*

Algorithm 1 Cherry-picking algorithm

```
q ← 1
(n_i, n_j) ← index of the smallest element in A
while (q < n) do
    if a_{n_j,n_i} = min A_{n_j,·} then
        return (n_i, n_j), break
    else
        n_i ← n_j
        n_j ← index of the smallest element in the n_i-th row of A
    end if
    q ← q + 1
end while
```

Our cherry-picking algorithm is summarized in Algorithm 1. Its properties are summarized in Theorem 4.1 and Proposition 2.

THEOREM 4.1. *For a given bifurcating phylogenetic tree T with asymmetric dissimilarity matrix $\mathbf{A}$, the cherry-picking algorithm will always return a cherry, and the algorithm terminates before the iteration variable q equals n.*

PROPOSITION 2. *If $\widehat{\mathbf{A}}_L$ is a consistent estimator of an asymmetric dissimilarity matrix $\mathbf{A}$, then the probability of identifying a cherry of the tree T using our cherry-picking algorithm on $\widehat{\mathbf{A}}_L$ converges to 1 as the number of loci L goes to infinity.*

By consistency, we mean in Proposition 2 that $\lim_{L\to\infty} P(|\widehat{\mathbf{A}}_L - A| \geq \varepsilon) = 0$ for any $\varepsilon > 0$ where the norm of a matrix is defined as the maximum of all elements.

Proposition 2 and Theorem 4.1 show that a cherry can be found using our cherry-picking algorithm by searching for a pair (P_i, P_j) which satisfies the 2-MRMC condition. However, when we apply the algorithm on a matrix $\widehat{\mathbf{A}}_L$ obtained from finite data, it is possible that there is no pair (P_i, P_j) satisfying the 2-MRMC condition. If no pair satisfies the 2-MRMC condition, the algorithm stops when $q = n$, and selects the two populations on hold as a cherry. A consequence of the proof of Proposition 2 is that the probability of this happening goes to zero as $L \to \infty$.

4.2. *Estimation step and update step.* Let $\mathbf{P} = \{P_1, P_2, \ldots, P_n\}$ denote the set of all populations. Assume that (P_i, P_j) is the selected cherry, and denote the internal node linking P_i and P_j as $P_{i,j}$. In our ANJ algorithm, we estimate the distance from P_i and P_j to $P_{i,j}$ respectively by

$$\widehat{d}_{i,ij} = a_{i,j} \quad \text{and} \quad \widehat{d}_{j,ij} = a_{j,i}. \tag{4.1}$$

We update the distance from any other population P_m $(m \neq i, j)$ to the new internal node $P_{i,j}$ by

$$a_{m,ij} = (a_{m,i} + a_{m,j})/2, \tag{4.2}$$

and update the distance from $P_{i,j}$ to P_m by

$$a_{ij,m} = (a_{i,m} + a_{j,m} - \widehat{d}_{i,ij} - \widehat{d}_{j,ij})/2. \tag{4.3}$$

We remove the ith and jth columns and the ith and jth rows of $\mathbf{A}$, add a new column and a new row for $P_{i,j}$, using $a_{m,ij}$ and $a_{ij,m}$, and set $a_{ij,ij} = \infty$ to form a reduced matrix $r(\mathbf{A})$. If $\widehat{\mathbf{A}}_L$ is used instead of $\mathbf{A}$, replace all $a_{\cdot}$ with $\widehat{a}_{\cdot}$ in the above equations to form a reduced estimated matrix $r(\widehat{\mathbf{A}}_L)$.

We define the reduced populations $r(\mathbf{P}) = \mathbf{P} - \{P_i, P_j\} + \{P_{i,j}\}$, and the reduced true tree $r(T)$ as a rooted bifurcating tree with leaf populations $r(\mathbf{P})$, that is, a subtree of T without descendants of $P_{i,j}$. We denote the asymmetric dissimilarity matrix of tree $r(T)$ as $\mathbf{A}^{r(T)}$.

Lemmas 1 and 2 establish the relationships between $\mathbf{A}^{r(T)}$, $r(\mathbf{A})$ and $r(\widehat{\mathbf{A}}_L)$.

LEMMA 1. *If a matrix* $\mathbf{A}$ *is an asymmetric dissimilarity matrix representation of a rooted bifurcating tree* T, *then, after a cherry of* T *is removed, the reduced matrix* $r(\mathbf{A})$ *is an asymmetric dissimilarity matrix representation of the reduced tree* $r(T)$, *that is,* $r(\mathbf{A}) = \mathbf{A}^{r(T)}$, *when the leaf populations* $r(\mathbf{P})$ *are arranged in the same order.*

LEMMA 2. *If* $\widehat{\mathbf{A}}_L$ *is a consistent estimator of an asymmetric dissimilarity matrix* $\mathbf{A}$, *then the reduced estimated matrix* $r(\widehat{\mathbf{A}}_L)$ *is a consistent estimator of* $r(\mathbf{A})$ (*i.e.*, $\mathbf{A}^{r(T)}$) *when the leaf populations* $r(\mathbf{P})$ *are arranged in the same order.*

We repeat the search, estimation and update steps until we have only one node left, which is the MRCA for all populations being considered.

Alternatively, estimation and update steps analogous to NJ and unweighted neighbor joining (UNJ) [Gascuel (1997)] can also be used. However, we found these alternatives not to work as well. This can be explained by the fact that ANJ uses only short branch lengths to construct the tree and short branch length estimates are more accurate than long branch length estimates using our method (see Section 5.2). If more accurate branch length estimates can be obtained using other methods for long branches, the asymmetric versions of NJ and UNJ may perform better than ANJ by utilizing more branch length estimates to reconstruct the tree.

Algorithm 2 summarizes our algorithm for constructing a rooted bifurcating tree from an asymmetric dissimilarity matrix $\mathbf{A}$. A bifurcating tree can be fully recovered from the recorded pairs (P_i, P_j) and the branch lengths $(\widehat{d}_{i,ij}, \widehat{d}_{j,ij})$. The properties of ANJ are summarized in Theorem 4.2 and Proposition 3.

Algorithm 2 Constructing a tree from an asymmetric dissimilarity matrix

```
index = {1, 2, ..., n}
for k = 1 to n − 1 do
    pick up a cherry (P_i, P_j) using the Cherry–Picking algorithm (Algorithm 1)
    estimate d̂_{i,j} and d̂_{j,i}
    record (P_i, P_j) and (d̂_{i,j}, d̂_{i,j})
    combine (P_i, P_j) into a new node P_{n+k}
    remove P_i, P_j and associated rows and columns from A (or Â_L)
    index ← index − {i, j}
    for l in index do
        update the distance from P_l and P_{n+k} to their MRCA
    end for
    Add a new row and a new column in A for P_{n+k} using updated distances
    index ← index + {n + k}
end for
```

THEOREM 4.2. *If a matrix* $\mathbf{A}$ *is an asymmetric dissimilarity matrix representation of a rooted bifurcating tree* T, *then asymmetric neighbor joining can recover* T *correctly from* $\mathbf{A}$, *in terms of both topology and branch lengths*.

PROPOSITION 3. *If* $\widehat{\mathbf{A}}_L$ *is a consistent estimator of an asymmetric dissimilarity matrix* $\mathbf{A}$, *then the probability of recovering the topology of the true tree* T *using asymmetric neighbor joining on* $\widehat{\mathbf{A}}_L$ *converges to* 1 *when the number of loci* $L \to \infty$. *Moreover, the branch length estimates are also consistent*.

It is worth mentioning that if our inference method is used to estimate $\widehat{\mathbf{A}}_L$, then $\widehat{\mathbf{A}}_L$ is not a consistent estimator of $\mathbf{A}$, as we do not know the true marginal distributions of SNP frequencies in any of the internal populations. However, $\widehat{\mathbf{A}}_L$ can still serve as a useful estimate of $\mathbf{A}$ since the asymptotic biases of the estimates are roughly proportional to the magnitude of the true branch lengths in the model, as shown in the next section, which means the branch lengths are scaled but still maintain their general relative lengths.

5. Simulation studies. We conduct simulation studies to investigate the performance of our inference method, both in the well-specified and misspecified cases. We first conduct simulation studies for two populations to investigate the consistency of our branch length estimators. Then we conduct simulation studies for more than two populations to investigate the performance of our inference method in recovering true tree topology and estimating branch lengths of larger trees.

For a pair of two populations, our method gives consistent estimates of branch lengths when the distribution of SNP frequencies in the ancestral population used

in our likelihood model is the true distribution used in simulation. When the two distributions are different, the estimates of branch lengths are not guaranteed to be consistent, but still reflect the relative lengths of the branches. For more than two populations, our method identifies the root of the phylogenetic tree correctly in all simulation runs.

Source code is available at https://github.com/yzhai220/anj.

5.1. *Simulation for two populations*. First, we generate three sets of SNP frequencies of an ancestral population P_0 using a mixed distribution with a continuous part from uniform(0, 1), beta(0.5, 0.5) or beta(2, 2), and discrete parts on 0 and 1 with mass m_0 and m_1. The proportion of frequencies which are fixed in P_0 is denoted $m_f = m_0 + m_1$. In the simulation, we set $m_0 = m_1 = 0.1$. We generate the values $\mathbf{p}_0$ of L independent SNPs with $L = 100, 1000, 10{,}000$ in this simulation study.

Second, we generate SNP frequencies for the leaf populations P_1 and P_2, $\mathbf{p}_1$ and $\mathbf{p}_2$, based on $\mathbf{p}_0$ using the normal approximation with general fixation model with parameters σ_1^2 and σ_2^2, respectively, as shown in (3.1). We set $\sigma_1^2 = 0.1, 0.5, 1.0$, and $\sigma_2^2 = 2\sigma_1^2$ so that it is not a clock tree.

The box plots of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ from 100 simulation runs are shown in Figure 3 [Uniform(0, 1) used for $\mathbf{p}_0$], Figure S7 [Beta(0.5, 0.5) used for $\mathbf{p}_0$ in the Supplementary Material] and Figure S9 [Beta(2, 2) used for $\mathbf{p}_0$ in the Supplementary Material].

From Figure 3, we can see that the estimates $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ are very accurate. The variances of the estimates decrease when the number of SNPs increases for all true values used. The effect of using $\widehat{m}_f$ versus using m_f on the estimates $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ is generally small. There are slight biases for larger values of σ_1^2 and σ_2^2, which can be explained by the fact that, for larger values of branch lengths, $\widehat{m}_f$ tends to overestimate m_f.

From Figure S7 of the Supplementary Material, we can see that the estimates of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ are highly accurate for the smaller branch length σ_1^2, but they underestimate the larger branch length σ_2^2 when the continuous part of $\mathbf{p}_0$ is generated from Beta(0.5, 0.5) and estimated using Uniform(0, 1). This can be explained by the shape of the Beta(0.5, 0.5) density. The effects of using empirical Bayes estimators $\widehat{m}_f$ (3.6) are similar to the effects observed in Figure 3. Although biases are introduced, the relative magnitudes of estimated branch lengths are preserved.

5.2. *Simulation for more than two populations*. We investigate the performance of our methods for datasets containing more than two populations. In order to obtain realistic simulation scenarios, we use trees estimated from real data as references. The two reference trees used to simulate population SNP frequencies are based on an estimated subtree of 8 African populations, and an estimated subtree of 7 American and Oceanic populations (with an extra outgroup, namely, the

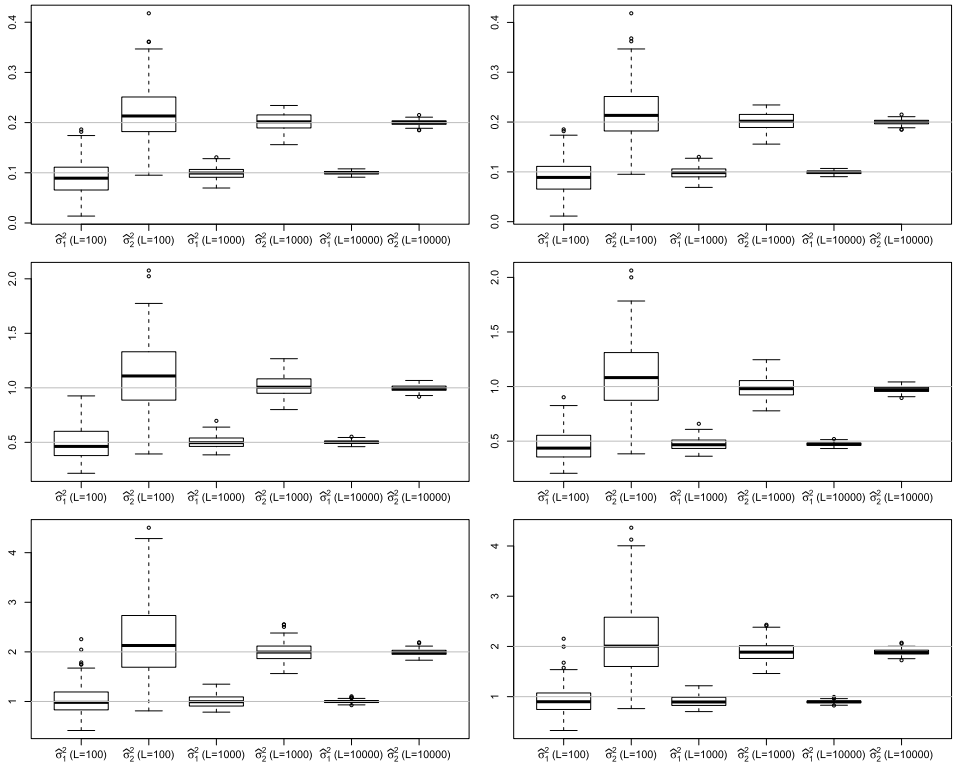


FIG. 3. *Boxplots of* $\widehat{\sigma}_1^2$ *and* $\widehat{\sigma}_2^2$ *in* 100 *simulation runs for two populations when* $\mathbf{p}_0$ *is generated from* Uniform(0, 1) *with point mass* 0.1 *at* 0 *and* 1. *In the three left panels, the true value* $m_f = 0.2$ *is used in estimation, and in the three right panels, the empirical Bayes estimate* $\widehat{m}_f$ (3.6) *is used in estimation. Grey lines indicate true values of* σ_1^2 *and* $\sigma_2^2 = 2\sigma_1^2$.

Hazara population). The estimated tree comes from our data analysis results using the HGDP data of 53 human populations. Results on the full tree are presented in the next section.

We first generate $\mathbf{p}_0$, the SNP frequencies at the root of the subtree, using a mixed distribution with a continuous part from Uniform(0, 1) and discrete parts on 0 and 1. Then we generate SNP frequencies of internal nodes and leaf populations based on the true tree from the root to leaves using our normal approximation with a general fixation model. We keep only the SNP frequencies of leaf populations for inference. We also conduct simulation studies when the Wright–Fisher model is used to generate data. See the Supplementary Material for results.

We set $m_0 = m_1 = 0.1$, $L = 5000$ in the simulation for African populations, and $L = 2000$ for American and Oceanic populations.

We construct a rooted bifurcating tree based on the asymmetric dissimilarity matrix for the leaf populations using our inference method and our asymmetric neighbor-joining algorithm. For comparison, we also construct an unrooted bifur-

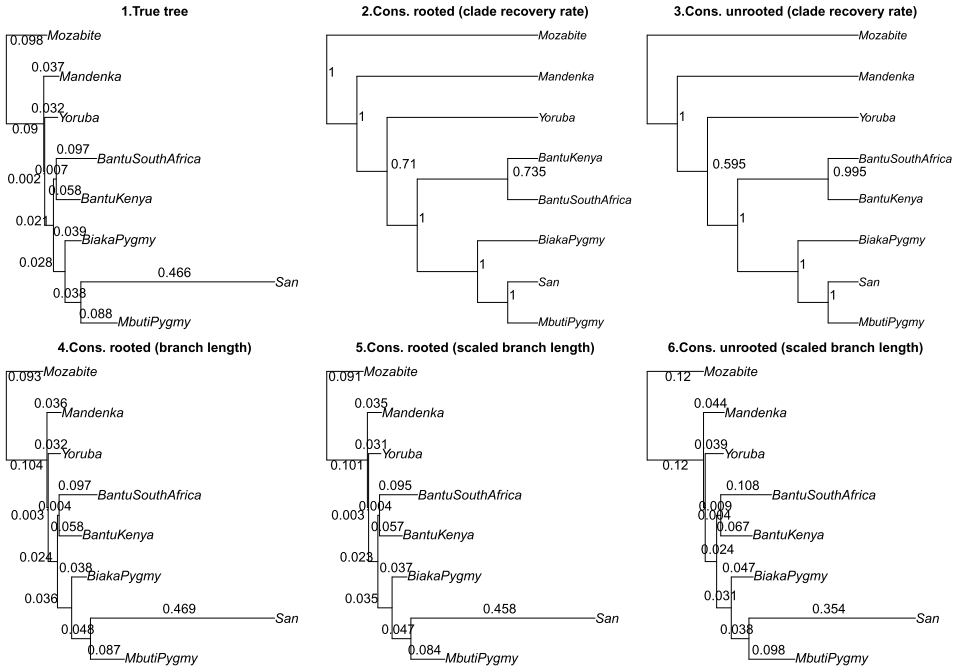


FIG. 4. *Simulation results for African populations. Panel* 1 *shows the true tree and its branch lengths. Panels* 2 *and* 3 *show the consensus rooted tree using our method and the consensus unrooted tree using NJ, respectively, with the numeric annotations being successful recovery rates of clades in* 200 *simulation runs. Panels* 4 *and* 5 *show the consensus rooted trees with unscaled and scaled estimated branch lengths, and Panel* 6 *shows the consensus unrooted tree with scaled branch lengths.*

cating tree based on a symmetric dissimilarity matrix for the leaf populations using Nei's symmetric dissimilarity measure [Nei (1972)] and NJ, a classical method that is still widely used because of its efficiency and accuracy. We further root those unrooted trees from NJ results with an outgroup (Mozabite for African populations and Hazara for American and Oceanic populations) by setting the root at the midpoint from the outgroup to the clade containing all other populations for illustration and comparison purposes. We use empirical Bayes estimates $\widehat{m}_f$ (3.6) for each pair of populations in our likelihood model.

We compute consensus trees [Felsenstein (2004), Paradis, Claude and Strimmer (2004), Paradis (2012)] from 200 simulation runs. The true tree, the consensus rooted tree of estimated rooted trees using our method and the consensus unrooted tree of estimated unrooted trees using NJ are shown in Figure 4 (for African populations) and Figure 5 (for American and Oceanic populations).

Note that, in all comparisons in this section, we set a higher standard for results using our methods than results using traditional methods. When we compare estimated trees with the true tree, we use estimated rooted trees using ANJ directly, and use estimated unrooted trees using NJ with extra information on rooting, that

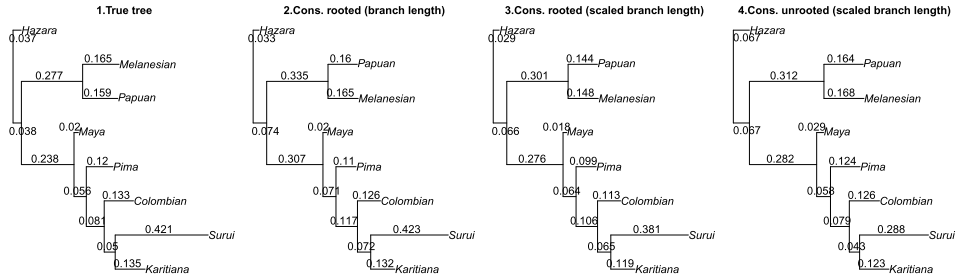


FIG. 5. *Simulation results for American and Oceanic populations. Panel* 1 *shows the true tree and its branch lengths. Panels* 2 *and* 3 *show the consensus rooted tree with unscaled and scaled estimated branch lengths. Panel* 4 *shows the consensus unrooted tree with scaled estimated branch lengths.*

is, a correct outgroup specified as the root. In other words, ANJ estimates one more node than NJ in all simulation runs, but, on the other hand, we set the correct location of this node for trees estimated by NJ so that results using both methods can be compared. We scale the total branch lengths of the estimated trees so that they have the same total branch length as the true tree. Results are reported for both unscaled and scaled trees.

Note that our method correctly identifies the location of the root of the phylogenetic tree in all simulation runs. In Figure 4, the branch lengths in the true tree of African populations (panel 1) are short, with the shortest edge length only 0.002. Our method achieves a high rate of clade recovery on this data (panel 2 shows the results obtained using our method, and panel 3 shows the results obtained using NJ, which uses additional information). The consensus branch lengths estimated by our method (panels 4 and 5) achieve a high level of accuracy, even without rescaling them (panel 4). This contrasts with the NJ method, which is less accurate in terms of branch length estimates, even after rescaling (panel 6).

In Figure 5, our method recovers the true tree topology in each of 200 simulation runs for American and Oceanic populations. We do not show panels 2 and 3 of Figure 4 in Figure 5 as the recovery rates are 1 for all clades. Panels in Figure 5 show similar results to those in Figure 4. Panels 1, 2 and 3 show that branch lengths which are shorter or closer to leaf populations are estimated more accurately than branch lengths which are longer and closer to the root. This observation is consistent with our results in previous sections that our estimates are more accurate for shorter branch lengths and branches which are close to leaf populations. However, our method still works well for longer branch lengths in terms of estimating the relative magnitude.

Table 1 summarizes the mean distances between the estimated trees and the true trees under several different distance measures for phylogenetic trees: the proportion of trees which have the same topology as the true tree, the Penny–Hendy (PH) distance [Penny and Hendy (1985)], the Kuhner–Felsenstein (KF) distance [Kuhner and Felsenstein (1994)] and the geodesic distances [Billera, Holmes and

TABLE 1
Mean distances between estimated trees and true trees calculated from 200 *simulations*

	African-ANJ	**African-NJ**	**American-ANJ**	**American-NJ**
Correct rate	0.555	0.590	1	1
PH	1.110	0.820	0	0
KF (unscaled)	0.016	0.045	0.109	0.331
KF (scaled)	0.015	0.013	0.065	0.062
Geodesic (unscaled)	0.030	0.459	0.123	0.584
Geodesic (scaled)	0.027	0.092	0.098	0.157

Vogtmann (2001), Chakerian and Holmes (2012)]. The PH distance equals two times the number of clades being mistakenly specified. The KF distance measures the number of misspecified clades as well as differences of internal branch lengths. The geodesic distance measures the number of misspecified clades and differences of all branch lengths.

From Table 1, we can see that for African populations the probability of recovering the full true tree topology using ANJ is similar to that using NJ. ANJ performs better in terms of KF for unscaled trees and in terms of geodesic distances for both scaled and unscaled trees, but performs worse than NJ in terms of PH. For American and Oceanic populations, both methods recover the true tree topology all the time, but our method performs better in all distance measures considered expect for KF for scaled trees.

More importantly, our method identifies the root of the tree correctly in all cases, an inference which cannot be obtained using symmetric dissimilarity measures, unless extra information in the form of an outgroup is available.

5.3. *Computational cost.* For a pair of populations, the main computational cost of our method is the calculation of the likelihood for various parameters (σ_1^2, σ_2^2), which is required in the optimization of the likelihood. Calculation of the likelihood for a fixed (σ_1^2, σ_2^2) involves a univariate numerical integration for each SNP location, which is vectorized and numerically solved using Gaussian quadratures with 40 points in our simulation and data analysis. The numerical optimization cost can be reduced by choosing a reasonable starting point. It can also be further improved by learning those parameters using a small portion of SNPs and then using those estimates as starting points with more SNPs.

In our simulation studies, calculating branch lengths for a pair of populations takes around 0.14 seconds ($L = 100$), 1.40 seconds ($L = 1000$) and 14.7 seconds ($L = 10{,}000$) for $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.2$ and Uniform(0, 1). The results vary for different parameters but remain similar. All analyses are run on a 2.8 GHz CPU Mac Mini (Intel dual core i5) using R version 3.1.0.

For more than two populations, we estimate the asymmetric dissimilarity matrix using our method applied to each pair of populations. The computational cost

is of order $\binom{n}{2}$ times the cost for one pair. Overall, this gives a running time of $O(Ln^2)$ for the computation of the asymmetric dissimilarity matrix, and it can be also easily paralleled. In our simulation studies, calculating the asymmetric dissimilarity matrix for eight African populations takes 37.5 seconds per run on average ($L = 2000$), and 41.3 seconds per run on average ($L = 2000$) for seven American and Oceanian populations. The computational cost for the construction of the trees from an asymmetric dissimilarity matrix using our algorithm is negligible compared to the cost of forming the asymmetric dissimilarity matrix.

6. Data analysis: Human genome diversity panel. We apply our methods on a subset of the data from the Human Genome Diversity Panel (HGDP). The HGDP contains 650,000 common SNP loci data for 938 unrelated individuals from 53 populations from Africa, Europe, the Middle East, Asia and the Americas. The number of individuals in one population varies from 5 to 46.

Li et al. (2008) obtained an unrooted phylogenetic tree for 51 populations (with Han and Han North China combined as Han, and Bantu South Africa and Bantu Kenya combined as Bantu) using `contml`, which is implemented based on the pruning algorithm proposed by Felsenstein (1973). By genotyping two chimpanzee samples, Li et al. (2008) defined the putative ancestral allele for most of the SNPs, and used the chimpanzee data to locate the root of the tree, that is, the MRCA of all human populations, at the node linking San and all other populations. Pickrell and Pritchard (2012) analyzed the same data sets to investigate population admixtures. Pickrell and Pritchard (2012) obtained a similar phylogenetic tree as Li et al. (2008), but not all populations within the same geographical regions are clearly grouped together in their results, even after adjusting for population admixtures. Pickrell and Pritchard (2012) also removed two Oceanian populations (Melanesian and Papuan) from the analysis because, according to the authors, including these two populations will lead to confusing results. Our method does not suffer from this issue and reconstructs a high-quality phylogenetic tree from the full dataset.

6.1. *Human population tree.* To reduce correlation among SNPs, we subsample the SNPs to one for every tenth loci, and we focus on a subset of the original data set containing 131,329 SNPs. As a result, our data contains frequencies of 13,133 SNPs in 53 populations, that is, the data matrix $\mathbf{X} = (x_{ij})$ is a $53 \times 13{,}133$ matrix with each row representing one population and each column representing one SNP locus. The results obtained are consistent when other subsets of SNPs are used.

Since the HGDP records frequencies of the minor SNP, we symmetrize the SNP frequencies so that the prior over the ancestral frequencies can be assumed to be symmetric:

$$x^*_{ij} = \begin{cases} x_{ij}, & \text{for } i = 1, 2, \ldots, 53, \text{ and } j = 1, 3, \ldots, 13{,}133, \\ 1 - x_{ij}, & \text{for } i = 1, 2, \ldots, 53, \text{ and } j = 2, 4, \ldots, 13{,}132. \end{cases}$$

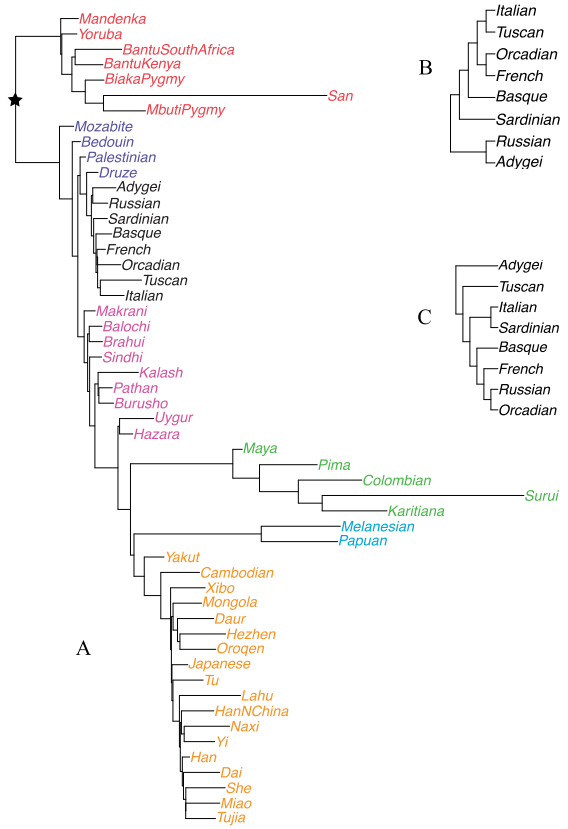


FIG. 6. (A): *Rooted phylogenetic tree for* 53 *human populations in HGDP using our method.* (★) *indicates the root of our tree. Populations are colored according to regions as specified in Li et al.* (2008). (B): *Zoomed dendrogram for European populations using our method.* (C): *Zoomed dendrogram for European populations reproduced using* `contml` [*Felsenstein* (1989)], *the same subtree as in Li et al.* (2008).

We calculate the asymmetric dissimilarity matrix for the 53 populations using the data $\mathbf{X}^* = (x^*_{ij})$. We then construct a rooted bifurcating phylogenetic tree using asymmetric neighbor joining on the asymmetric dissimilarity matrix proposed in this paper (see Figure 6). Our results are obtained using less SNPs than Li et al. (2008) and without chimpanzee data.

From Figure 6, we can see that populations from the same continents are grouped together, which is consistent with our knowledge of human migrations. The pattern of human populations moving out of African can be observed clearly from Figure 6(A) and is consistent with the prevalent consensus [Cavalli-Sforza and Feldman (2003)]. American and Asian populations are farthest from the root, while African and Middle Eastern populations are closest to the root.

Our results are more accurate than the results of Li et al. (2008) and Pickrell and Pritchard (2012), in terms of estimating phylogenetic trees for closely related populations. For example, from historical records, the Tuscan and Italian populations are closely related, and this relationship is correctly identified using our method [see Figure 6(B)], but not in Li et al. (2008) [see Figure 6(C)]. Similarly, we can say the same for the Adygei and Russian populations. Our result clearly identifies closely related populations within the same regions into subgroups [see colors in Figure 6(A)], while this is not the case in Pickrell and Pritchard (2012), as they failed to group all Middle Eastern populations together and failed to incorporate two Oceanian populations.

More importantly, our method identifies the root of the human population tree without using extra data from Chimpanzee or Neandertal as in Li et al. (2008) and Pickrell and Pritchard (2012). In our results, the root is located between a clade containing current middle and southern African populations, and a clade containing current northern African and non-African populations [see ★ in Figure 6(A)]. However, results in this region of the tree should be treated with caution as both models ignore effects such as recent admixtures which have a large effect on the SNP frequencies in African populations [Pickrell et al. (2012)].

It takes 7.88 hours to calculate the asymmetric dissimilarity matrix for 52 populations (without parallel computation) and less than 1 second to reconstruct the rooted tree based on the dissimilarity matrix.

6.2. *Assessing uncertainty of the estimated tree.* Assessing uncertainty of the estimated tree is not a simple task. The uncertainty of the estimated tree comes from not only the topology of the estimated tree, but also the associated branch lengths.

The bootstrap method has been widely used to assess the uncertainty of the estimated tree [Felsenstein (1983), Zharkikh and Li (1995)]. Summarizing bootstrap estimates in a meaningful way is not straightforward if more than one tree topology is present in the bootstrap estimated trees. A similar problem arises in the Bayesian phylogenetic framework when summarizing sample trees from the posterior distritbution [Yang and Rannala (1997)].

The consensus tree is a popular tool to summarize a set of phylogenetic trees in practice. The consensus tree usually consists of only the topology [Zharkikh and Li (1995), Gray and Atkinson (2003), Smeulders et al. (2011), Song and Steinrücken (2012)]. One question is how the average branch lengths should be calculated for trees of different topology since the branch lengths are conditional on the phylogenetic tree and are meaningful only for specific trees [Yang, Goldman and Friday (1995)]. Some software calculate the mean branch lengths of the consensus tree by averaging branch length estimates from only trees that have the same topology as the consensus tree [Sukumaran and Holder (2010), Revell (2012)]. This approach is debatable, as it ignores all other estimated trees of different topology [Felsenstein (2004)]. For the same reason, a sound definition of variances of branch

length estimates is only available for the mean estimated tree if all estimated trees share the same topology.

On the other hand, Billera, Holmes and Vogtmann (2001) investigated the geometry of the space of phylogenetic trees and proposed the geodesic distance, which enables averaging phylogenetic trees and the construction of a convex hull of a family of trees as a measure of variability of the estimated tree. Efficient algorithms have been developed to calculate the Fréchet mean tree and the Fréchet variance of the estimated tree [Owen and Provan (2011), Benner, Bačák and Bourguignon (2014)] based on the geodesic distance. The Fréchet mean tree is defined as the tree that achieves the minimum mean square geodesic distance to all trees considered, and the Fréchet variance is the mean square geodesic distance between the Fréchet mean tree to all trees considered. The Fréchet variance can be interpreted as the uncertainty measure for the estimated tree as a whole. However, this approach does not straightforwardly provide uncertainty measures for individual nodes or individual branch length estimates. As a consequence, it could be argued that the Fréchet variance is not easy to interpret.

We illustrate how to use the bootstrap method to assess the uncertainty of the estimated phylogenetic tree based on our methods. For each bootstrap run, we sample SNPs with replacement and then estimate the phylogenetic tree using the bootstrap sample. We then construct the consensus tree of all 5000 bootstrap trees with uncertainty measure on each node, and the Fréchet mean tree of all bootstrap trees with the Fréchet variance. The geodesic mean tree and its variance are calculated using `TrAP` [Benner, Bačák and Bourguignon (2014)]. To provide an example where the uncertainty is relatively large, we show a reconstruction using only 1314 SNPs (i.e., every hundredth of the 131,329 SNPs) from six sample populations across the world.

Figure 7 shows that the root of the estimated tree is placed at the node separating African populations with all other populations with probability 1 in this analysis. The consensus tree [Figure 7(3)] and the Fréchet mean tree [Figure 7(4)] share the same topology and almost identical branch length estimates. Uncertainty of the estimated tree topology is observed on the node of the consensus tree linking

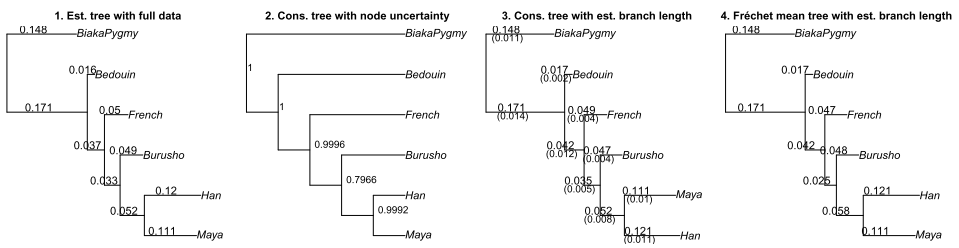


FIG. 7. (1): *Estimated tree based on* 1314 *SNPs*. (2): *Consensus tree with estimated node uncertainty from* 5000 *bootstrap samples*. (3): *Consensus tree with estimated branch lengths and their standard deviations in brackets*. (4): *Fréchet mean tree with estimated branch length. The Fréchet variance of this Fréchet mean tree is* 0.0013.

Han, Papuan and Maya [0.7968 in Figure 7(2)], meaning that this node does not appear in approximately 20% of the bootstrap estimate trees. Uncertainty of the estimated branch lengths are summarized by the standard deviations of the branch length estimates in the consensus tree [Figure 7(3)].

Note that uncertainty of the estimated tree rapidly reduces when the number of SNPs used in the analysis increases. When all 13,133 SNPs are used, the Fréchet variance decreases to less than 0.0001 from 0.0013, and the proportion of bootstrap estimate trees that share the same topology as the consensus tree increases to 99.82% from 79.66% (see Figure S11 in the Supplementary Material).

In conclusion, reasonably accurate human population trees can be identified using a relatively small number of SNPs based on our method. The uncertainty of this root placement is extremely low even when the number of SNPs used in the analysis is small. This result shows that the information on the root of the human population tree is stored in the gene of human populations, and can be recovered using nonreversible models on the evolutionary process of SNP frequencies of human populations, which does not depend on the choice of outgroups.

7. Discussion. We have presented a simple and effective method for reconstructing *rooted* trees from multi-population SNP frequencies. By modeling fixation in a principled way, and generalizing neighbor-joining methods, we can perform joint tree and rooting inference without requiring an outgroup. We have shown in simulations and real worldwide genome-wide data that this rooting can be more accurate despite using less data. The method is also computationally efficient with time complexity $O(Ln^2)$, where L is the number of sites and n is the number of populations.

One key element we have ignored thus far is the effect of admixture, which could potentially bias the analysis [Pickrell et al. (2012)]. To take admixture into account, one could design rooted network inference algorithms from dissimilarity matrices. Note also that it may be feasible to detect admixtures automatically by looking at iterations of Algorithm 2 where no population pair satisfies the 2-MRMC (this property is only guaranteed to hold for matrices derived from a tree). We have not detected such cases in our data analysis, suggesting that the tree approximation might be reasonable at the tree scales of the HGDP dataset. We have also ignored sampling errors, but since the pairwise asymmetric distance calculations are based on a probability model, it is possible to include an error model at the cost of a more expensive optimization. Similarly, ascertainment bias may also be incorporated into the likelihood model via a correction term [Nicholson et al. (2002)]. The effect of selection is ignored in our model. It is generally unclear how strong the effect of selection has been in human populations, although Hernandez et al. (2011) showed that strong selective sweeps appeared rare.

A more fundamental limitation is that our model ignores the internal structure of each leaf population. This is mostly of concern for inferring relationships among closely related populations, where it may be difficult to cluster individuals into well-separated subpopulations. In such cases, computationally expensive coales-

cent and ancestral graph methods [Bryant et al. (2012)] may be preferable. For medium- and large-scale tree inference, however, we have demonstrated that our method is an accurate and computationally affordable alternative.

Acknowledgments. The authors thank Dr. John Petkau, Dr. Jiahua Chen and Dr. Bonnie Kirkpatrick for their suggestions on this paper. The authors are grateful to the referees, the Associate Editor and the Editor for their valuable comments.

SUPPLEMENTARY MATERIAL

Supplement to: "Inferring rooted population trees using asymmetric neighbor joining" (DOI: 10.1214/16-AOAS964SUPP; .pdf). We provide additional simulation studies and proofs on the properties of the algorithms in the supplementary material [Zhai and Bouchard-Côté (2016)].

REFERENCES

BALDING, D. J. and NICHOLS, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96** 3–12.

BATTISTUZZI, F. U., FILIPSKI, A., HEDGES, S. B. and KUMAR, S. (2010). Performance of relaxed-clock methods in estimating evolutionary divergence times and their credibility intervals. *Mol. Biol. Evol.* **27** 1289–1300.

BENNER, P., BAČÁK, M. and BOURGUIGNON, P.-Y. (2014). Point estimates in phylogenetic reconstructions. *Bioinformatics* **30** i534–i540.

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931

BRYANT, D., BOUCKAERT, R., FELSENSTEIN, J., ROSENBERG, N. A. and ROYCHOUDHURY, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29** 1917–1932.

CANN, H. M., DE TOMA, C., CAZES, L., LEGRAND, M.-F., MOREL, V., PIOUFFRE, L., BODMER, J., BODMER, W. F., BONNE-TAMIR, B., CAMBON-THOMSEN, A. et al. (2002). A human genome diversity cell line panel. *Science* **296** 261–262.

CAVALLI-SFORZA, L. L. and FELDMAN, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* **33 Suppl** 266–275.

CHAKERIAN, J. and HOLMES, S. (2012). Computational tools for evaluating phylogenetic and hierarchical clustering trees. *J. Comput. Graph. Statist.* **21** 581–599. MR2970909

EDWARDS, A. and CAVALLI-SFORZA, L. (1964). Reconstruction of evolutionary trees. *Systematics Association Publ.* **6** 67–76.

EWENS, W. J. (1973). Conditional diffusion processes in population genetics. *Theor. Popul. Biol.* **4** 21–30.

FELSENSTEIN, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25** 471–492.

FELSENSTEIN, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution* **35** 1229–1242.

FELSENSTEIN, J. (1983). Statistical inference of phylogenies. *J. R. Stat. Soc., A* **146** 246–272.

FELSENSTEIN, J. (1989). PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5** 164–166.

FELSENSTEIN, J. (2004). *Inferring Phytogenies*. Sinauer, Sunderland, Massachusetts.

GASCUEL, O. (1997). Concerning the NJ algorithm and its unweighted version, UNJ. In *Mathematical Hierarchies and Biology* (*Piscataway*, *NJ*, 1996). *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **37** 149–170. AMS, Providence, RI. MR1600536

GRAY, R. D. and ATKINSON, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426** 435–439.

GRAY, R. D., DRUMMOND, A. J. and GREENHILL, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323** 479–483.

HASEGAWA, M., KISHINO, H. and YANO, T.-A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22** 160–174.

HERNANDEZ, R. D., KELLEY, J. L., ELYASHIV, E., MELTON, S. C., AUTON, A., MCVEAN, G., SELLA, G., PRZEWORSKI, M. et al. (2011). Classic selective sweeps were rare in recent human evolution. *Science* **331** 920–924.

HUELSENBECK, J. P., BOLLBACK, J. P. and LEVINE, A. M. (2002). Inferring the root of a phylogenetic tree. *Syst. Biol.* **51** 32–43.

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. and BOLLBACK, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294** 2310–2314.

IWABE, N., KUMA, K-I., HASEGAWA, M., OSAWA, S. and MIYATA, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* **86** 9355–9359.

JENKINS, P. A. and SPANO, D. (2015). Exact simulation of the Wright–Fisher diffusion. preprint. Available at arXiv:1506.06998.

KUHNER, M. K. and FELSENSTEIN, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11** 459–468.

LI, S., PEARL, D. K. and DOSS, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **95** 493–508.

LI, J. Z., ABSHER, D. M., TANG, H., SOUTHWICK, A. M., CASTO, A. M., RAMACHANDRAN, S., CANN, H. M., BARSH, G. S., FELDMAN, M., CAVALLI-SFORZA, L. L. and MYERS, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319** 1100–1104.

LIPO, C. P. (2006). *Mapping Our Ancestors*: *Phylogenetic Approaches in Anthropology and Prehistory*. Transaction Publishers. New Brunswick and London.

MAU, B., NEWTON, M. A. and LARGET, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55** 1–12. MR1705672

NEI, M. (1972). Genetic distance between populations. *Amer. Nat.* **106** 283–292.

NICHOLS, J. and WARNOW, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* **2** 760–820.

NICHOLSON, G., SMITH, A. V., JÓNSSON, F., GÚSTAFSSON, O., STEFÁNSSON, K. and DONNELLY, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 695–715. MR1979384

OUTLAW, D. C. and RICKLEFS, R. E. (2011). Rerooting the evolutionary tree of malaria parasites. *Proc. Natl. Acad. Sci. USA* **108** 13183–13187.

OWEN, M. and PROVAN, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (*TCBB*) **8** 2–13.

PARADIS, E. (2012). *Analysis of Phylogenetics and Evolution with R*, 2nd ed. Springer, New York. MR2883250

PARADIS, E., CLAUDE, J. and STRIMMER, K. (2004). Ape: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** 289–290.

PEARSON, T., HORNSTRA, H. M., SAHL, J. W., SCHAACK, S., SCHUPP, J. M., BECKSTROM-STERNBERG, S. M., O'NEILL, M. W., PRIESTLEY, R. A., CHAMPION, M. D., BECKSTROM-STERNBERG, J. S., KERSH, G. J., SAMUEL, J. E., MASSUNG, R. F. and KEIM, P. (2013). When outgroups fail; phylogenomics of rooting the emerging pathogen, Coxiella burnetii. *Syst. Biol.* **62** 752–762.

PENNY, D. and HENDY, M. (1985). The use of tree comparison metrics. *Syst. Zool.* **34** 75–82.

PICKRELL, J. K. and PRITCHARD, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8** e1002967.

PICKRELL, J. K., PATTERSON, N., BARBIERI, C., BERTHOLD, F., GERLACH, L., GÜLDEMANN, T., KURE, B., MPOLOKA, S. W., NAKAGAWA, H., NAUMANN, C. et al. (2012). The genetic prehistory of southern Africa. *Nature Communications* **3** 1–6.

REVELL, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3** 217–223.

ROCH, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (*TCBB*) **3** 92.

ROYCHOUDHURY, A., FELSENSTEIN, J. and THOMPSON, E. A. (2008). A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* **180** 1095–1105.

SAITOU, N. and NEI, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** 406–425.

SEMPLE, C. and STEEL, M. (2003). *Phylogenetics*. *Oxford Lecture Series in Mathematics and Its Applications* **24**. Oxford Univ. Press, Oxford. MR2060009

SIRÉN, J., HANAGE, W. P. and CORANDER, J. (2013). Inference on population histories by approximating infinite alleles diffusion. *Mol. Biol. Evol.* **30** 457–468.

SIRÉN, J., MARTTINEN, P. and CORANDER, J. (2011). Reconstructing population histories from single nucleotide polymorphism data. *Mol. Biol. Evol.* **28** 673–683.

SMEULDERS, M. J., BARENDS, T. R. M., POL, A., SCHERER, A., ZANDVOORT, M. H., UDVARHELYI, A., KHADEM, A. F., MENZEL, A., HERMANS, J., SHOEMAN, R. L. et al. (2011). Evolution of a new enzyme for carbon disulphide conversion by an acidothermophilic archaeon. *Nature* **478** 412–416.

SONG, Y. S. and STEINRÜCKEN, M. (2012). A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* **190** 1117–1129.

SUKUMARAN, J. and HOLDER, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics* **26** 1569–1571.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. and HILLIS, D. M. (1996). Phylogenetic inference. In *Molecular Systematics* (M. D. Hillis and C. Moritz, eds.) 407–514. Sinauer Associates, Sunderland.

TAVARÉ, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some Mathematical Questions in Biology—DNA Sequence Analysis* (*New York*, 1984). *Lectures Math. Life Sci.* **17** 57–86. AMS, Providence, RI. MR0846877

WANG, L., BOUCHARD-CÔTÉ, A. and DOUCET, A. (2015). Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *J. Amer. Statist. Assoc.* **110** 1362–1374. MR3449032

WEIR, B. S. and COCKERHAM, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38** 1358–1370.

WHEELER, W. C. (1990). Nucleic acid sequence phylogeny and random outgroups. *Cladistics* **6** 363–367.

YANG, Z., GOLDMAN, N. and FRIDAY, A. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Systematic Biology* **44** 384–399.

YANG, Z. and RANNALA, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14** 717–724.

ZHAI, Y. and BOUCHARD-CÔTÉ, A. (2016). Supplement to "Inferring rooted population trees using asymmetric neighbor joining." DOI:10.1214/16-AOAS964SUPP.

ZHARKIKH, A. and LI, W. H. (1995). Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.* **4** 44–63.

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
3182 EARTH SCIENCES BUILDING
2207 MAIN MALL
VANCOUVER, BC V6T1Z4
CANADA
E-MAIL: y.zhai@stat.ubc.ca
bouchard@stat.ubc.ca