distAngsd: Fast and Accurate Inference of Genetic Distances for Next-Generation Sequencing Data

Lei Zhao [1], Rasmus Nielsen [*,1,2,3] and Thorfinn Sand Korneliussen [*,1]

[1]Section for Geogenetics, Globe Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 København K, Denmark

[2]Department of Integrative Biology, University of California, 3040 Valley Life Sciences Building 3140, Berkeley, CA 94720-3140, USA

[3]Department of Statistics, University of California, 3040 Valley Life Sciences Building 3140, Berkeley, CA 94720-3140, USA

***Corresponding authors:** E-mails: rasmus_nielsen@berkeley.edu; tskorneliussen@sund.ku.dk

Associate editor: Koichiro Tamura

Abstract

Commonly used methods for inferring phylogenies were designed before the emergence of high-throughput sequencing and can generally not accommodate the challenges associated with noisy, diploid sequencing data. In many applications, diploid genomes are still treated as haploid through the use of ambiguity characters; while the uncertainty in genotype calling—arising as a consequence of the sequencing technology—is ignored. In order to address this problem, we describe two new probabilistic approaches for estimating genetic distances: distAngsd-geno and distAngsd-nuc, both implemented in a software suite named distAngsd. These methods are specifically designed for next-generation sequencing data, utilize the full information from the data, and take uncertainty in genotype calling into account. Through extensive simulations, we show that these new methods are markedly more accurate and have more stable statistical behaviors than other currently available methods for estimating genetic distances—even for very low depth data with high error rates.

Key words: **phylogeny reconstruction, genotype likelihood, genetic distance, high-throughput sequencing, next-generation sequencing, molecular evolution, maximum likelihood, expectation maximization.**

Introduction

While much of biology has been revolutionized by the availability of high-throughput next-generation sequencing, the state-of-art methods for calculating genetic distances used in phylogeny estimation have not changed much for the past several decades, and are still not properly modeling the uncertainty and idiosyncrasies associated with next-generation sequencing data. Modern sequencing technologies produce millions or billions of small DNA fragments through a massively parallel sequencing process. In re-sequencing studies, these fragments are then aligned to a (typically) haploid representation of the target genome. As the DNA sequences in these fragments have non-negligible error-rates, genotype likelihoods are calculated in order to model the uncertainty of genotypes arising from sequencing errors and from varying sequencing depth. These genotype likelihoods are defined as the probability of the observed sequencing data—at a particular position of the genome—as a function of the true (but unknown) genotype, which often is assumed to be diploid. Genotype likelihoods, therefore, capture all the uncertainty in the data regarding the true genotype.

In phylogenetics, the estimation of genetic distances is based on continuous time Markov chain models of nucleotide substitution. The core concept is to model nucleotide substitution as a Markov process, while allowing for differences in substitution rates among the four different nucleotides and, possibly, variation in the substitution rate among sites. The simplest model is the Jukes and Cantor model, also called the JC69 model (Jukes and Cantor 1969), which assumes equal rates of substitutions between all base pairs. A series of other models developed in the 1980s and 1990s including the K80 (Kimura 1980), F81 (Felsenstein 1981), HKY85 (Hasegawa et al. 1985), F84 (Felsenstein and Churchill 1996), TN93 (Tamura and Nei 1993) relaxes this assumption by incorporating additional parameters, such as differences in the rate of transitions and transversions, and unequal equilibrium nucleotide frequencies. The most general commonly used model is the Generalized Time Reversible (GTR) model (Tavaré 1986), which is the most parameter-rich model that is still time-reversible. Relaxing the requirement of time-reversibility of the Markov chain allows for more parameter-rich models, in particular the unrestricted UNREST model of Yang (1994) which is a fully parameterized model with 12 parameters—one for each of the 12 possible substitution types. However, because of computational simplicity and tractability, most work in phylogenetics focuses on the GTR model. In this work we will also assume the GTR model as the basic model of nucleotide

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Article

Open Access

substitution—however, the results and methodology we develop can in principle be generalized to non-reversible models.

The previously discussed substitution models were developed for haploid sequences of nucleotides. However, much data currently generated arise from the nuclear DNA of diploid organisms. Applying these models designed for haploid sequences to diploid data has been a challenge in many studies, especially due to the considerable genotyping uncertainty in much of the published sequencing data. Different approaches exist for handling heterozygous sites and uncertainty in genotype calling for phylogenetic analyses of diploid data. One approach is to construct haploid data by phasing the diploid data using computational approaches—e.g., using various extensions of PHASE (Stephens et al. 2001; Stephens and Donnelly 2003) or BEAGLE (Browning and Browning 2007). This is an approach that may work well for species where many individuals (>100) have been re-sequenced or where large reference panels of individuals from the same species are already available. Unfortunately, for most phylogenetic analyses, such data will often not be available. Other approaches include: (1) ignoring all possible heterozygous sites, i.e. treating them as missing data (Nilsson et al. 2017; Árnason et al. 2018; Maldonado et al. 2019)—in this work, we call this approach **NoAmbiguityGT**; (2) representing possible heterozygous sites as IUPAC (The International Union of Pure and Applied Chemistry, Cornish-Bowden 1985) ambiguity codes (Klicka et al. 2014; Martin et al. 2014; Uckele et al. 2021)—termed **AmbiguityGT** in this work; (3) choosing a random nucleotide from the sequenced reads at each position (Skoglund et al. 2016; Yang et al. 2020)—**RandomSEQ** in this work; (4) making a consensus call of a single nucleotide based on the raw sequencing data, i.e., converting the diploid sequencing data into a haploid sequence by, for each site, selecting (one of) the most frequent nucleotides in the sequencing data (Manthey et al. 2016; Sass et al. 2016; Yuan et al. 2016)—**ConsensusSEQ** in this work; or (5) incorporating uncertainty using genotype likelihoods or other measurements of uncertainty, e.g., **ngsDist** (Vieira et al. 2015), which was applied in Choi and Purugganan (2018), Gaunitz et al. (2018), and Hu et al. (2018). ngsDist does not perform genotype calling, but instead uses the diploid diallelic genotype likelihood to compute the average per site allelic differences by averaging over the joint posterior genotype probabilities. ngsDist was not devised for phylogenetic analyses and does not use an explicit model of nucleotide substitution.

As we will show in the results section, all existing methods have serious drawbacks in simulated scenarios. All simulated scenarios presented in the manuscript are based on the topology shown in figure 1. The first two approaches, AmbiguityGT and NoAmbituityGT, which perhaps are the most commonly used approaches, have previously been shown to cause strong biases in phylogenetic estimation (Lischer et al. 2014; Potts et al. 2014; Schrempf et al. 2016). The bias is not unidirectional: at low read depth (≤1) these two methods will overestimate the genetic distance; whereas higher read depths (≥5) will cause underestimation (see fig. 2 and supplementary fig. S8, and Section S5.1, Supplementary Material online for further discussion).

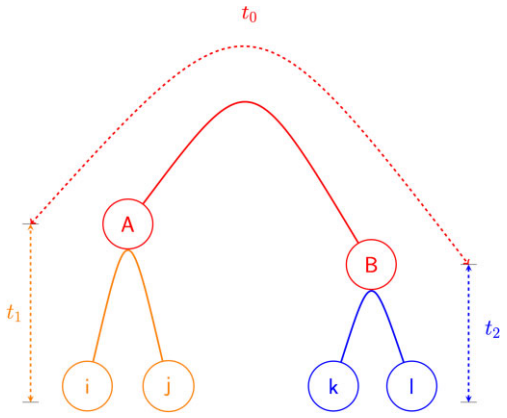

FIG. 1. Simulated divergence tree: Tree structure used for simulations. Two diploid individuals are of the genotypes *ij* and *kl*, respectively. The most recent common ancestor of *i* and *j* is A, and that of *k* and *l* is B. The divergence time of A and B is denoted as t_0, while the time from *A* to *i* or *j* is t_1, and the time from *B* to *k* or *l* is t_2. The divergence time between the two diploid individuals, *t*, is thus defined as $t = t_0 + t_1 + t_2$.

The methods RandomSEQ and ConsensusSEQ do not perform genotype calling, but instead choose a representative nucleotide among all the available data for every site. RandomSEQ samples a random allele per site and the resulting data is equivalent to a sequencing study with a deterministic read depth of 1. ConsensusSEQ is based on constructing a haploid sequence by choosing the most frequently observed nucleotide in each site. As we will later show (figs. 3, 4 and supplementary fig. S9, Supplementary Material online), both methods will be strongly affected by sequencing errors at low depths, although ConsensusSEQ can perform well at high sequencing depths.

The Bayesian approach in ngsDist assumes an infinite sites model that does not take recurrent mutations into account but more importantly it averages over the posterior probability distribution of genotypes when calculating distances. We will in the Results section show that this approach can result in highly biased estimates of genetic distances.

Motivated by the statistical limitations of previous methods, we develop two new methods, **distAngsd-geno** and **distAngsd-nuc**, which are both maximum-likelihood (ML) estimators of the genetic distance. They differ from previous methods as they do not attempt genotype or haploid calling, but instead model the sequencing uncertainty while also using an explicit nucleotide substitution model. Through extensive simulations, and by applying our methods to real sequencing data, we show that the two novel methods outperform previous methods by having significantly less biases and smaller variances—especially in the context of low read depth, or small genome sizes. When using the methods for estimating phylogenetic trees on a

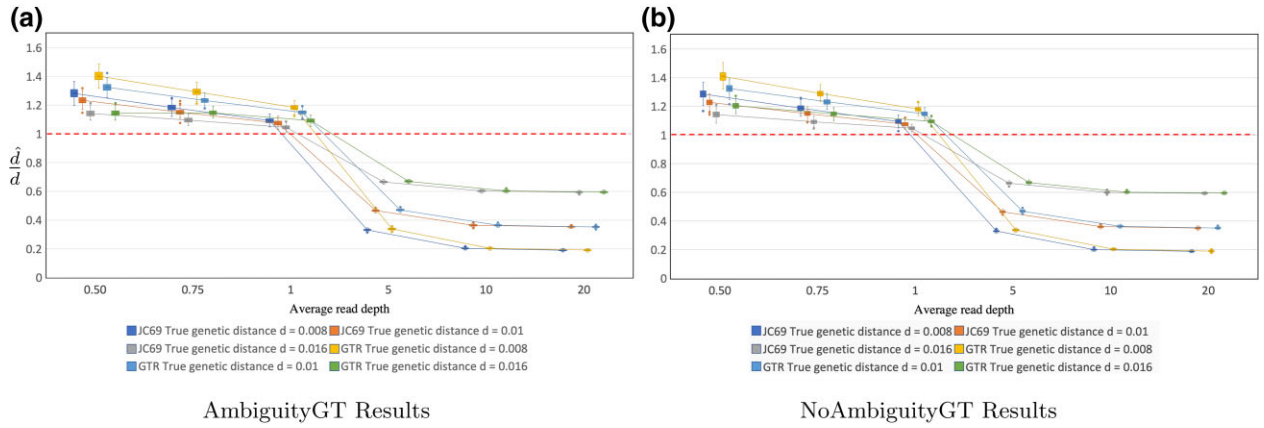

FIG. 2. The (scaled) genetic distance estimations ($\hat{d}/d$) for (*a*) AmbiguityGT and (*b*) NoAmbiguityGT based on the simulation results. The average read depth = 0.5, 0.75, 1, 5, 10, 20. True genetic distance $d = 0.008, 0.01, 0.016$. JC69 and GTR models were used. $t_1 = 0.004$, $t_2 = 0.0025$, see figure 1. The default simulation setting was applied.

real data set, we show that the phylogenetic trees estimated using the new methods have higher phylogenetic concordance with existing taxonomic categories than trees estimated using previous methods.

Materials and Methods

Both distAngsd-geno and distAngsd-nuc share similar characteristics. They both pre-compute a genome-wide estimate of the joint distribution of either genotypes (distAngsd-geno) or nucleotides (distAngsd-nuc) in two individuals. This is accomplished by using genotype likelihoods or by directly using the quality scores associated with each nucleotide across all reads. The estimation of these joint distributions is free of assumptions regarding evolutionary models. Inference of genetic distance, *d*, using maximum likelihood based on models of molecular evolution, e.g., the JC69 model or the more realistic GTR model, then proceeds by treating the inferred joint distributions as pseudo-data (i.e., $\hat{\mathbf{M}}$ and $\hat{\mathbf{N}}$ below). The inference procedure, therefore, has two steps: (1) Maximum-likelihood inference of genome-wide joint distribution of alleles or genotypes, and (2) maximum-likelihood inference of genetic distances based on the results of step (1). This two-step approach ignores the statistical uncertainty introduced in step (1) and merely uses point estimates of joint distributions. However, for genome-wide data, the variances in the estimates of genotype or allele frequencies should be negligible, and as we will show, this procedure does in fact have good statistical properties for realistic parameters' values.

Estimation of Joint Distributions Using EM

As accurate genotype calling is not possible with low read depth sequencing data, e.g., Nielsen et al. (2011), we employ a likelihood approach to estimate the joint global distribution using the expectation maximization algorithm (EM) (Dempster et al. 1977).

In distAngsd-geno, the joint genotype distribution is represented by a 10 by 10 matrix **M** where 10 is the total number of possible unphased genotypes in a diploid individual, i.e., {AA, AC, AG, AT, CC, CG, CT, GG, GT, TT}. Both rows and columns of **M** are indexed by these genotypes, and every element of **M**, $\mathbf{M}(g_i, g_j)$, represents the proportion of the informative sites (sites where we have data for both individuals) where the true genotypes are given by g_i and g_j. Similarly in distAngsd-nuc, a 4 by 4 matrix **N** representing the joint distribution of nucleotides is estimated. Here, **N** is indexed by nucleotides, i.e., {A,C,G,T}, rather than the 10 possible genotypes.

The **M** matrix used in distAngsd-geno is inferred via Algorithm 1, while distAngsd-nuc applies Algorithm 2 in order to estimate **N**. The input of Algorithm 1 are the genotype likelihoods, while the input in Algorithm 2 are the likelihoods of all nucleotides given by the Phred-scaled

ALGORITHM 1: EM estimation for **M**. EM algorithm for estimating the 10 × 10 joint genotype distribution matrix **M**, where $\mathbf{M}_t$ is the matrix **M** in the *t*th iteration and the function $\mathrm{GL}_k(s, g_i)$ is the genotype likelihood of genotype g_i at site *s* in sample *k* ($k \in \{1,2\}$).

Input: Genotype likelihoods of sites;

Output: 10 × 10 matrix **M**;

initialization: $\mathbf{M}_0(g_i,g_j) \leftarrow \frac{1}{100}$, $t \leftarrow 0$;

while *elements in* **M** *do not converge* **do**

  $\mathbf{M}_{t+1}(g_i,g_j) \leftarrow 0$;

  for $s \leftarrow 1$ **to** #*sites* **do**

    if *s is an informative site* **then**

      $\mathbf{M}_{t+1}(g_i,g_j) \leftarrow \mathbf{M}_{t+1}(g_i,g_j) + \frac{\mathbf{M}_t(g_i,g_j)\mathrm{GL}_1(s,g_i)\mathrm{GL}_2(s,g_j)}{\sum_{g_i,g_j}\mathbf{M}_t(g_i,g_j)\mathrm{GL}_1(s,g_i)\mathrm{GL}_2(s,g_j)}$;

    end

  end

  $\mathbf{M}_{t+1}(g_i,g_j) \leftarrow \frac{\mathbf{M}_{t+1}(g_i,g_j)}{\#\textit{informative sites}}$;

  $t \leftarrow t+1$;

end

Algorithm 2: EM estimation for matrix **N**. EM algorithm for estimating the 4×4 joint nucleotide distribution matrix **N**, where $\mathbf{N}_t$ is the matrix **N** in the tth iteration, and the function $q_l(s, k_l, n_i)$ is the likelihood function of the nucleotide n_i in the k_l'th read at site s in sample l, $l = 1$ or 2. q_l can be obtained from the read quality scores.

Input: Reads data and their quality scores;
Output: 4×4 matrix N;
initialization: $\mathbf{N}_0(n_i, n_j) \leftarrow \frac{1}{16}$, $t \leftarrow 0$;
While *elements in* **N** *do not converge* **do**
  $\mathbf{N}_t(n_i, n_j) \leftarrow 0$;
  for $s \leftarrow 1$ **to** $\#sites$ **do**
    for $k_1 \leftarrow 1$ **to** $\#reads_{s,1}$ **do**
      for $k_2 \leftarrow 1$ **to** $\#reads_{s,2}$ **do**
        $\mathbf{N}_{t+1}(n_i, n_j) \leftarrow \mathbf{N}_{t+1}(n_i, n_j) + \frac{\mathbf{N}_t(n_i, n_j) q_1(s, k_1, n_i) q_2(s, k_2, n_j)}{\sum_{n_i, n_j} \mathbf{N}_t(n_i, n_j) q_1(s, k_1, n_i) q_2(s, k_2, n_j)}$;
      end
    end
  end
  $\mathbf{N}_{t+1}(n_i, n_j) \leftarrow \frac{\mathbf{N}_{t+1}(n_i, n_j)}{\sum_{s=1}^{\#sites} \#reads_{s,1} \times \#reads_{s,2}}$;
  $t \leftarrow t + 1$;
end

base quality score in the read data. The objective function (i.e., eq. S3 for **M** and eq. S4 for **N**) is essentially the same as equation (1) in Korneliussen et al. (2014), which is the two dimensional equivalent of the likelihood function presented in Keightley and Halligan (2011), Li (2011), and Nielsen et al. (2012).

Log-likelihood of distAngsd-geno Inference

The log-likelihood function in distAngsd-geno given the inferred pseudo-observation matrix $\hat{\mathbf{M}}$ is defined as,

$$l(d|\hat{\mathbf{M}}) = \sum_{g_1 g_2} \hat{\mathbf{M}}(g_1, g_2) \left\{ \sum_{i,j=1}^{2} \log \left[\pi_{g_{1,i}} \mathrm{P}_{g_{1,i} g_{2,j}}(d) \right] \right\}. \quad (1)$$

Here, the sum is over all possible unphased diallelic genotype nucleotide configurations and $g_{1,i}$ and $g_{2,j} \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$ are the ith and jth nucleotides of genotypes g_1 and g_2, respectively. $\mathrm{P}_{g_{1,i} g_{2,j}}(d)$ is the transition probability from nucleotide $g_{1,i}$ to $g_{2,j}$ given the genetic distance d between the two samples. $\pi_{g_{1,i}}$ is the stationary distribution of nucleotide $g_{1,i}$.

Log-likelihood of distAngsd-nuc Inference

The log-likelihood function of distAngsd-nuc given the inferred pseudo-observation matrix $\hat{\mathbf{N}}$ is,

$$l(d \mid \hat{\mathbf{N}}) = \sum_{n_1, n_2} \hat{\mathbf{N}}(n_1, n_2) \log \left[\pi_{n_1} \mathrm{P}_{n_1, n_2}(d) \right], \quad (2)$$

where $\mathrm{P}_{n_1, n_2}(d)$ is the transition probability from nucleotide n_1 to n_2 given the genetic distance d between the two samples. π_{n_1} is the stationary distribution of nucleotide n_1.

Parameters of the Substitution Model

The forms of $\mathrm{P}_{\cdot,\cdot}(\cdot)$, $\pi_{\cdot}$ and d in both equations (1) and (2) will be determined by the nucleotide substitution model (e.g., the JC69 or GTR model, see supplementary Section S3, Supplementary Material online for details). The genetic distance between two individuals, d, is a product of the divergence time $t = t_0 + t_1 + t_2$ (as shown in fig. 1) and the substitution rate, μ, $d = t\mu$. Notice that this method, like other phylogenetic methods, estimate d but cannot independently estimate t and μ. We then scale the nucleotide substitution rate matrices of the JC69 model and GTR model so that the mean number of substitutions occurring per base per scaled time unit is fixed to be 1, and $d = t$ represents the mean number of substitutions per site in the time interval. In the following, we will not distinguish between estimating d and estimating t. Also, notice that, similarly to classical phylogenetic methods, we here ignore complications from varying coalescence time and incomplete lineage sorting (See supplementary Section S6, Supplementary Material online for some further investigations on the effects on the estimation method of incomplete lineage sorting) between the two individuals. If the method is applied in settings where the coalescent process causes significant variation in t, the method is expected to estimate an average value of t for the two individuals.

The detailed derivation of the EM algorithm for both proposed methods are given in supplementary Section S2, Supplementary Material online.

Simulations and Comparisons to Previous Methods

Simulations

We will focus on the JC69 model and GTR model assuming all sites are variable. (i.e., no site is invariable). However, we also explore models where a proportion of sites are invariable in supplementary Section S3.5, Supplementary Material online.

We simulate data following Algorithm 3 assuming a diploid species. While distAngsd-geno assumes diploidy, distAngsd-nuc is applicable to species with other ploidies, but we here only compare the methods for the diploid case. The simulated phylogeny is described in figure 1. The genotype likelihood model used in the simulations is the canonical genotype likelihood model (see also McKenna et al. 2010 and supplementary Section S1, Supplementary Material online):

$$\mathrm{P}(r_k = l \,|\, g = ij) = \begin{cases} 1 - e_k, & \text{If } l = i = j, \\ \frac{e_k}{3}, & \text{If } l \neq i \text{ and } l \neq j, \\ \frac{1}{2} - \frac{e_k}{3}, & \text{Otherwise.} \end{cases}$$

Here $i, j, l \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$, and e_k is the error rate of the kth read in the focal site. $\mathrm{P}(r_k = l \,|\, g = ij)$ is the probability that the kth read at the focal site is called as nucleotide l given the true genotype in the site is ij.

Algorithm 3: Simulation of sequencing data. Simulation scheme of two individuals with a shared ancestor. A simulation scheme considers both variable and invariable sites can be found in supplementary alg. S1, Supplementary Material online.

Input: Substitution rate matrix R (the stationary distribution of π. is known), divergence time t, t_1 and t_2, error rate e, mean read depth RD, **genome length l;**

Output: Read data and Genotype likelihoods across sites;

Ancestral sequences construction:

The ancestral nucleotide for sample A and B, denoted by a_1 and a_2, are simulated given R, π., divergence time $t - t_1 - t_2$, and l. Sites are assumed to evolve independently;

Sequence construction and calling:

initialization: $GL_1(s, g) \leftarrow GL_2(s, g) \leftarrow 0$;

for $j \leftarrow 1$ **to 2 do**

  Two sequences q_{j1} and q_{j2} are simulated from a_j given R, l, and time t_j;

  for *every site s in sample j* **do**

    Generate read depth $n \sim$ **Poisson**(RD).

    Sample n reads from the true genotype at the site s of q_{j1} and q_{j2} with symmetric errors at rate e, to generate data $r_1, \ldots, r_n$.

    Define genotype likelihoods as $GL_j(s, g) = \sum_{k=1}^{n} \log[\,\textbf{Prob}(r_k|g)]$.

  end

end

In the simulations, it is assumed that the error rates are identical across different reads and sites (i.e., $e_k = e$ for all k). The read depths are assumed to be i.i.d., Poisson random variables in all sites, and only parameter needed to define this distribution is the average read depth. It would also be possible to simulate varying error rates by sampling them from a different distribution, but as long as the base error rates are calibrated correctly, variation in the error rate among sites should not affect the behavior of the new methods qualitatively.

Parameter Settings

Unless otherwise stated, we simulate data with fixed values of $t_1 = 0.004$, and $t_2 = 0.0025$ (see fig. 1), but vary the total divergence time t_0, such that $t = t_0 + t_1 + t_2 = 0.008$, 0.01 and 0.016.

In the main text, most of the simulations are conducted with the genome length and base calling error set to 1 Mb and 0.2%. And for each scenario, 200 replicates are simulated. We will refer to these as the **default simulation setting**.

The parameters of the GTR model are $a = 2.0431$, $b = 0.0821$, $c = 0.0000$, $d = 0.0670$, $e = 0.0000$, $\pi_T = 0.2184$, $\pi_C = 0.2606$, $\pi_A = 0.3265$, $\pi_G = 0.1946$. These values are suggested by the table 1.3 in Yang (2006) and were originally inferred from the human and orangutan 12S rRNA genes.

Other parameter values with larger base calling error, longer genetic distance, etc., are explored in the Supplementary Material online but are not presented in the main text.

Comparisons with the Previous Methods

To assess the performance of the new methods, we compare their performance against the five previous methods (RandomSEQ, ConsensusSEQ, NoAmbiguityGT, AmbiguityGT, and ngsDist).

1. RandomSEQ and ConsensusSEQ

The likelihood function for both RandomSEQ and ConsensusSEQ is given by:

$$l(d \mid \hat{\mathbf{C}}_1) = \sum_{n_1,n_2} \hat{\mathbf{C}}_1(n_1, n_2) \log\left[\pi_{n_1} \mathrm{P}_{n_1,n_2}(d)\right],$$

where $\hat{\mathbf{C}}_1$ is the joint nucleotide counts matrix of the size 4×4 obtained by either sampling the nucleotide (RandomSEQ) or by using the Consensus (ConsensusSEQ). The consensus sequence is created by choosing the most common base in each site, choosing randomly if multiple bases are equally common, and ignoring the site if no bases are observed in one or both of the samples. The definitions of $\mathrm{P}_{n_1,n_2}(d)$ and π_{n_1} are the same as those in equation (2).

2. AmbiguityGT and NoAmbiguityGT

AmbiguityGT and NoAmbiguityGT perform genotype calling and, thereby, have an implicit assumption of known ploidy level. We assume diploid genotypes obtained through standard genotype calling where a posterior probability is computed for each of the 10 possible genotypes under the assumption of a uniform prior. In the NoAmbiguityGT approach, heterozygous sites are discarded, i.e. the data are treated as haploid data ignoring heterozygous sites. However, in the AmbiguityGT method heterozygous sites are represented as ambiguity characters and the likelihood function is calculated by assigning equal likelihood to each of the two nucleotides, i.e., heterozygous sites are treated as sites that are haploid but with uncertainty regarding the nucleotide in the site. This is the standard way of dealing with missing data or uncertainty regarding nucleotide state in phylogenetic inference. We can think of the AmbiguityGT approach as being based on the

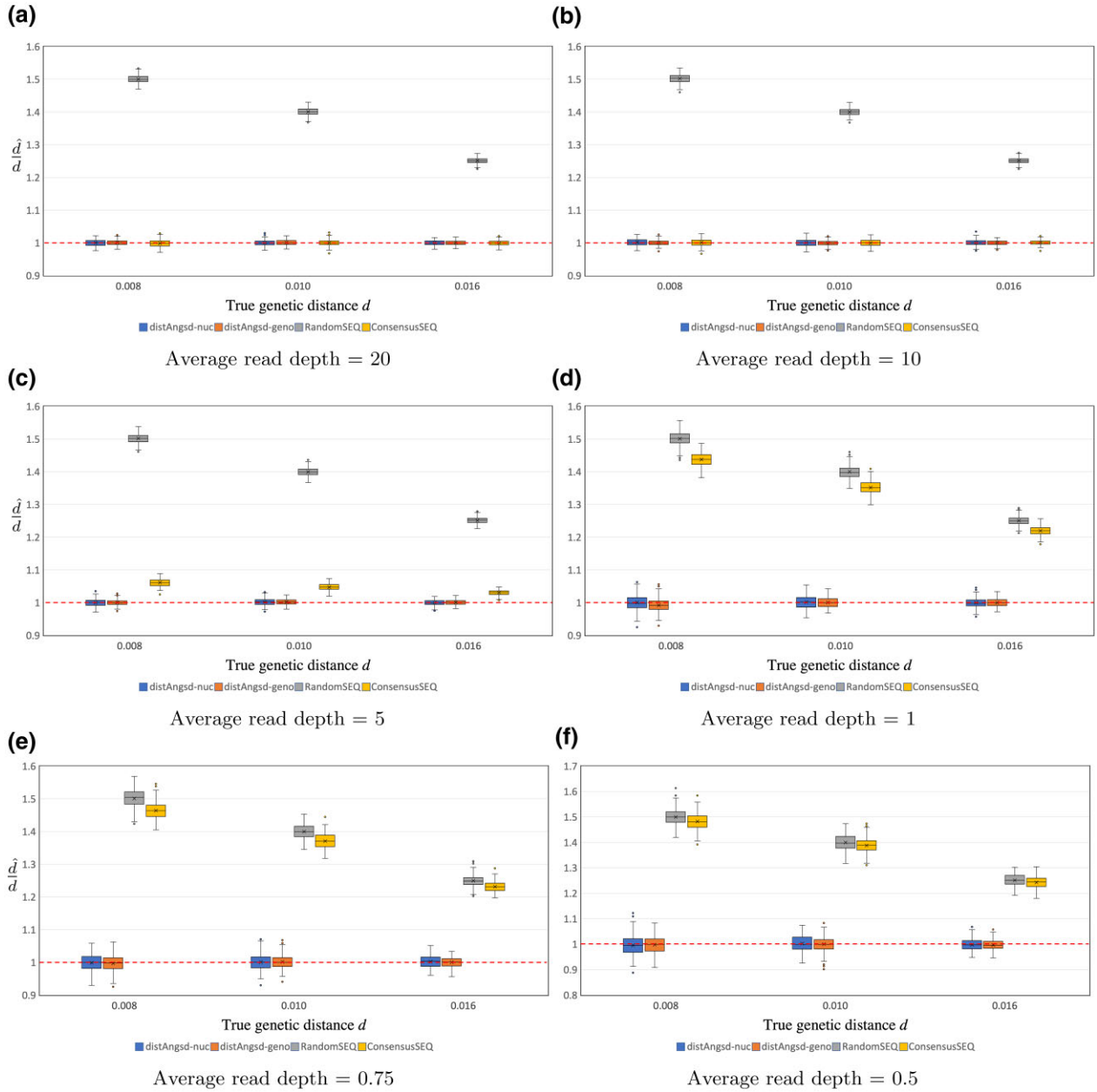

FIG. 3. The (scaled) genetic distance estimations ($\hat{d}/d$) for distAngsd-nuc, distAngsd-geno, RandomSEQ, and ConsensusSEQ based on the simulation results under JC69 model. Average read depth = 20, 10, 5, 1, 0.75, 0.5 for (*a*)–(*f*), respectively. True genetic distance $d = 0.008, 0.01, 0.016$. $t_1 = 0.004$, $t_2 = 0.0025$, see figure 1. The default simulation setting was applied.

following likelihood function:

$$l(d \mid \hat{\mathbf{C}}_2) = \sum_{g_1,g_2} \hat{\mathbf{C}}_2(g_1, g_2)\{\log[\mathrm{P}_{g_1,g_2}(d)]\}, \qquad (3)$$

where $\log[\mathrm{P}_{g_1=ij,g_2=kl}(d)]$ equals $\log[\sum_{n_1 \in M_1} \pi_{n_1} \sum_{n_2 \in M_2} \mathrm{P}_{n_1,n_2}(d)]$, $\hat{\mathbf{C}}_2$ is a count matrix of the called genotype pairs, and M1 and M2 are the sets of nucleotides observed in the two sites. M_i will contain two nucleotides if sample i is heterozygous in that site, and one nucleotide if homozygous. We elaborate more on the details of these likelihood functions in supplementary Section S5.1, Supplementary Material online.

3. ngsDist

All methods mentioned above including RandomSEQ, ConsensusSEQ, AmbiguityGT, and NoAmbiguityGT were implemented in distAngsd, and we used this implementation for the inferences for the simulated data.

We also compare the new methods with ngsDist (Vieira et al. 2015), which is the only pre-existing method that models the uncertainty of the data to estimate genetic distances. However, this method assumes a di-allelic model. We, therefore, convert the 10-genotype likelihoods simulated by Algorithm 3 to 3-genotype likelihoods, by first inferring the major and minor alleles as the most frequently, and second most frequently observed nucleotides,

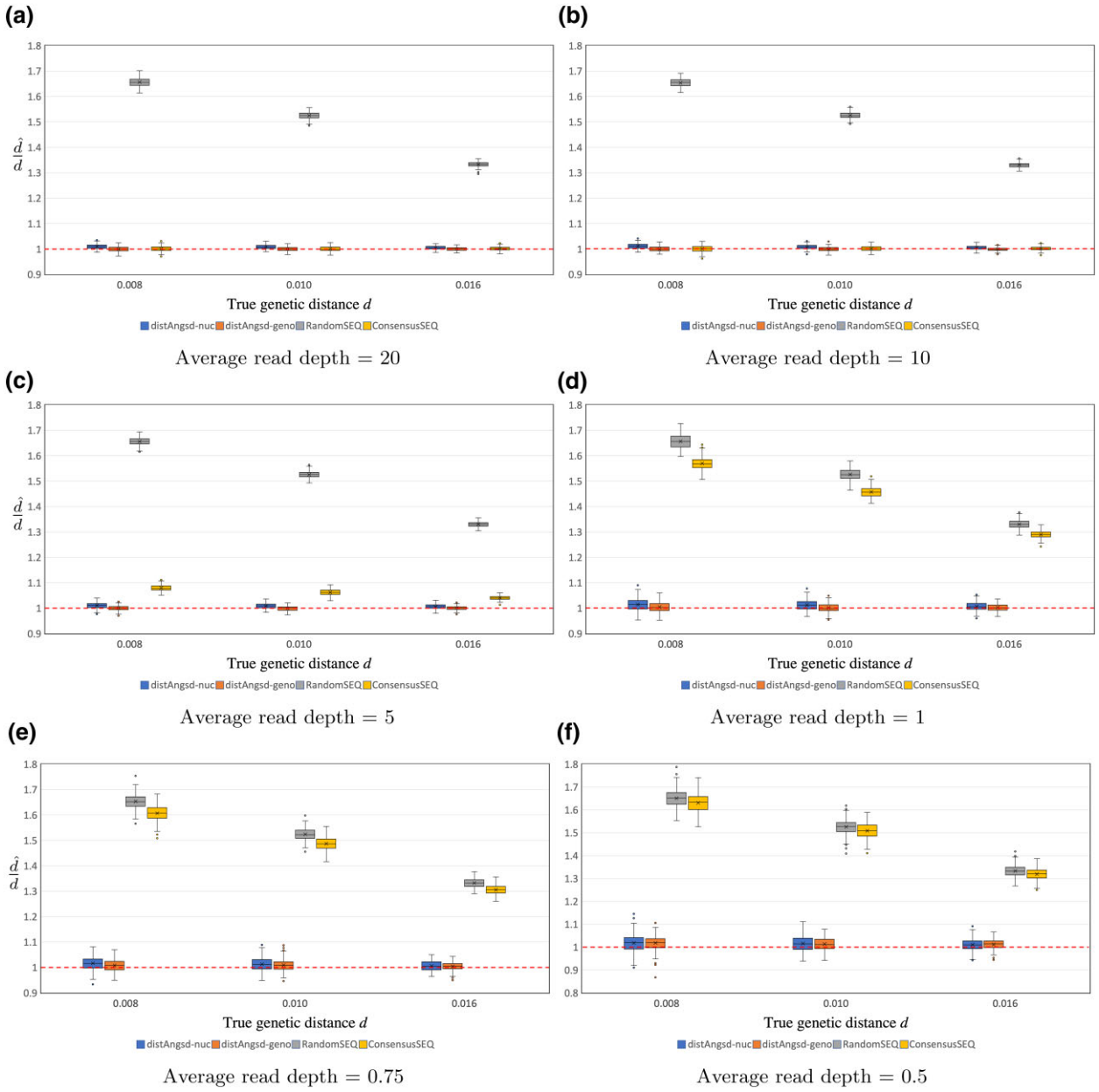

Fig. 4. The (scaled) genetic distance estimations ($\hat{d}/d$) for distAngsd-nuc, distAngsd-geno, RandomSEQ, and ConsensusSEQ based on the simulation results under GTR model. Average read depth = 20, 10, 5, 1, 0.75, 0.5 for (*a*) to (*f*), respectively. True genetic distance $d = 0.008, 0.01, 0.016$. $t_1 = 0.004$, $t_2 = 0.0025$, see figure 1. The default simulation setting was applied.

respectively. The resulting 3-genotype likelihoods are then used as input to ngsDist to calculate the pairwise distances (see supplementary Section S5.2, Supplementary Material online for details).

Experimental Data Analyses

We apply the new distAngsd-geno method in a phylogenetic analysis of a previously published dataset of RADseq reads from oak trees (Fitz-Gibbon et al. 2017). This data set is comprised of 83 oak samples representing 16 taxa located around the USA. We obtained the raw fastq sequences and followed the reference mapping approach described in detail in Fitz-Gibbon et al. (2017). The sites retained for downstream analyses are based on the callable-locus BED files shared by the authors. For each site, we use the raw genotype likelihood (bcftools genotype likelihood model). The data processing pipeline differs from that of the original authors by allowing heterozygote sites, which had been masked in the original analyses. These sites are either encoded as IUPAC ambiguity codes for the AmbiguityGT method, or used in the form of raw genotype likelihoods for our new methods. We obtained nucleotide consensus sequences for the ConsensusSEQ method by using the bcftools consensus command.

The fastq files for the 83 samples were aligned to the reference genome (Reference of the genome v0.5, Sork et al. 2016; available from https://valleyoak.ucla.edu/genomicresources/) to obtain per sample BAM files. For the distAngsd-geno

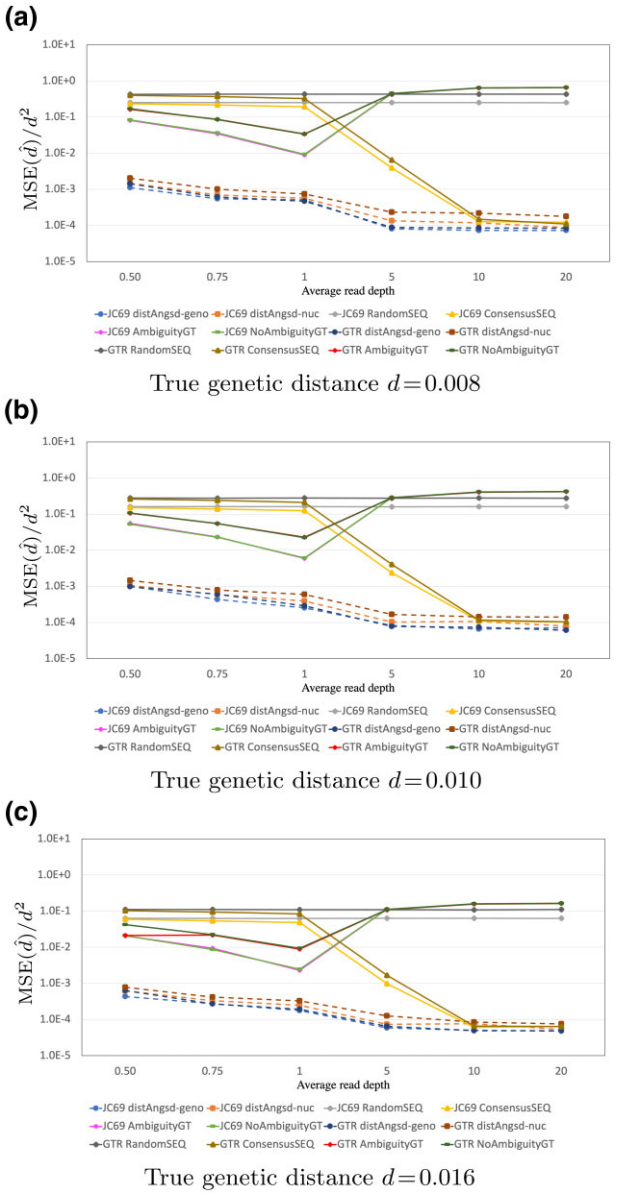

FIG. 5. The (scaled) mean squared errors for six different methods (distAngsd-geno, distAngsd-nuc, RandomSEQ, ConsensusSEQ, AmbiguityGT, and NoAmbiguityGT) in the simulation results. True genetic distance d: (a) $d = 0.008$. (b) $d = 0.010$. (c) $d = 0.016$. $t_1 = 0.004$, $t_2 = 0.0025$, see figure 1. The default simulation setting was applied. Solid lines correspond to the results of previous methods, while the dashed ones represent those of the two proposed methods.

inference, VCF files were generated by applying bcftools mpileup to all pairs of samples. Similarly, bcftools mpileup commands were applied to obtain per sample information for use in the AmbiguityGT and ConsensusSEQ analyses. Genotype calling was performed with "bcftools call" followed by "bcftools consensus -I" to obtain the putative genotype calls with IUPAC ambiguity codes for the AmbiguityGT analyses whereas we did not apply "bcftools call" but solely used the "bcftools consensus" for the ConsensusSEQ analyses (see supplementary Section S10, Supplementary Material online for details).

The distAngsd-geno, ConsensusSEQ, and AmbiguityGT methods were applied on each pair of oak samples using the JC69 substitution model. Based on the resulting 83 × 83 pairwise distance matrices, we performed neighbor joining estimation for each method using the program PhyD (Criscuolo and Gascuel 2008) with the BioNJ (Gascuel 1997) algorithm. The trees were then plotted with FigTree v1.4.4. (http://tree.bio.ed.ac.uk/software/figtree/).

To compare the methods in low-coverage scenarios, we down-sampled each of the 83 samples to lower coverage. The original 83 samples have an average read depth of 20.32 and lowest depth of 10.34. Each sample is down-sampled to read depths of 10, 5, 1, 0.75, 0.5, and 0.25. Phylogenetic trees were then estimated using the same methods as previously described.

Results

We implemented the new methods in a program: **distAngsd** (https://github.com/lz398/distAngsd). This program is threaded, scales linearly in the number of cores allocated, and is supplied as an open source **c/c++** program under the GPL license hosted on Github. Importantly, it is user friendly and allows for various standard formats that will enable researchers to integrate these methods in standard data-analysis pipelines. In the following, we compare the performance of the new methods to previous methods using extensive simulations and an application to a real data set.

Simulation-based Inference

The quantities we compared across different methods in this section are mainly the scaled genetic distances, $\hat{d}/d$ and the scaled mean squared error, $\mathrm{MSE}(\hat{d})/d^2$. The distributions of the genetic distance estimates, $\hat{d}$, are presented as boxplots illustrating the variability among simulation replicates. The deviations of the mean of $\hat{d}/d$ from 1 reflects the estimation biases, and the magnitudes of the boxes illustrate the variances of the estimates. $\mathrm{MSE}(\hat{d})/d^2$ is an overall measure of the accuracy of the estimates combining both bias and variance. Given the same true genetic distance d, the lower $\mathrm{MSE}(\hat{d})/d^2$ is, the better the inference of the method is.

1. AmbiguityGT and NoAmbiguityGT

As shown in figure 2*a* and *b*, the AmbiguityGT and NoAmbiguityGT approaches, which mimic the currently most commonly used methods, are generally found to be highly biased, and are affected by two oppositely directed biases: For example, for $d = 0.01$ under the JC69 model with average read depth is 0.5, AmbiguityGT and NoAmbiguityGT overestimate the true distance 1.234 and 1.228 times, respectively. However, if the average read depth is large, e.g., 20, both methods underestimate the true distance ($\hat{d}/d = 0.353$ for AmbiguityGT and $\hat{d}/d = 0.351$ for NoAmbiguityGT). Similar patterns of overestimation for low read depth and underestimation for high read depth are

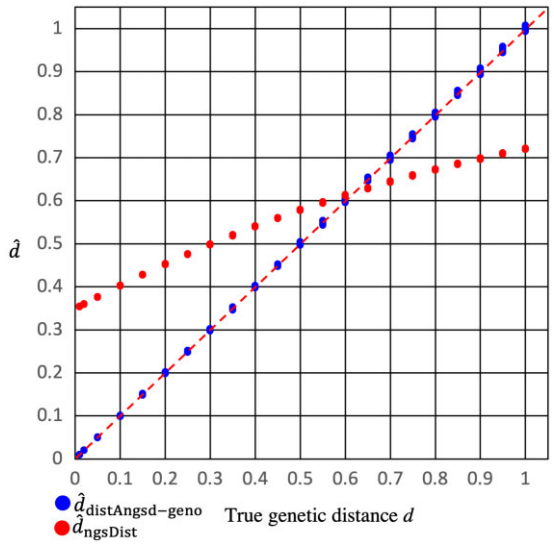

FIG. 6. Comparison of estimated genetic distances by distAngsd-geno and ngsDist for different true distances ranging from 0.01 to 1. The JC69 model was used. $t_1 = 0.004$, $t_2 = 0.0025$, see figure 1. At each true d value, 200 replicates were simulated and estimated. The default simulation setting was applied and the average read depth is set to 1. Blue points: distAngsd-geno. Red points: ngsDist.

observed across simulation settings (see also supplementary figs. S8 and S12, and table S4, Supplementary Material online).

The reasons that AmbiguityGT and NoAmbiguityGT are biased upwards for low depth and downwards for high depth are as follows: When the average read depth is low, most heterozygous sites appear as homozygous, however, sequencing errors will appear as additional fixed differences that cause overestimation of the divergence time. As the sequencing depth increases, the genotype calling becomes more accurate and less affected by sequencing errors. However, for the NoAmbiguityGT methods, the removal of heterozygous sites leads to underestimation of the genetic distance. Instead of estimating $t = t_0 + t_1 + t_2$ in figure 1, effectively only t_0 is being estimated. There will be a similar effect for the AmbiguityGT method, which also ends up effectively just estimating t_0. This effect is explored in more detail in supplementary S5.1, Supplementary Material online.

Since the biases of AmbiguityGT and NoAmbiguityGT are quite large, we do not include these methods in future comparisons.

2. RandomSEQ and ConsensusSEQ

Simulation results using the same parameter settings as in the previous section, for distAngsd-geno, distAngsd-nuc, RandomSEQ, and ConsensusSEQ are plotted in figures 3 (JC69) and 4 (GTR) with different read depths in different panels.

The results are qualitatively similar for both the GTR and JC69 models across all methods (including distAngsd-geno, distAngsd-nuc, RandomSEQ, and ConsensusSEQ, and the previously mentioned AmbiguityGT and NoAmbiguityGT methods see fig. 2). However, the biases tend to be larger under the GTR model due to the asymmetric nature of the GTR model. We summarize the common features of the results of distAngsd-geno, distAngsd-nuc, RandomSEQ, and ConsensusSEQ under both nucleotide substitution models as follows:

Both distAngsd-geno and distAngsd-nuc have higher accuracy and precision (smaller biases and variances) than RandomSEQ and ConsensusSEQ. This is especially clear for low depth scenarios (figs. 3*e,f* and 4*e,f*) and shorter genome length scenarios (see also supplementary figs. S13 and S14, Supplementary Material online). RandomSEQ and ConsensusSEQ are strongly affected by errors at lower sequencing depths. This sensitivity to errors also affects RandomSEQ at higher sequencing depths, while the performance of ConsensusSEQ improves strongly with increasing depth. It should be noted that when the mean read depth is small and the base calling error is relatively large (relative to the true genetic divergence), both of the new methods can still be biased. However, these biases are much smaller than those observed for the previous methods (e.g., see fig. 4). We refer readers to supplementary Section S4, Supplementary Material online for a general discussion on bias analyses.

As expected, since the new methods take advantage of the full information of the sequencing data (see discussion in supplementary Section S5.1, Supplementary Material online), they also have the smallest variance and the smallest mean squared error (MSE, see fig. 5, and supplementary fig. S15, Supplementary Material online for shorter genome scenarios). Furthermore since distAngsd-geno also uses prior knowledge of ploidy level, it always has the least variance in our diploid simulations.

Supplementary figures S13 and S14, Supplementary Material online correspond to figures 3 and 4, with the only difference being the number of variable sites used in the actual simulation (1 MB in the main text, 0.1 MB in Supplementary Material online).

3. ngsDist

ngsDist is based on a model which measures pairwise differences and does not distinguish between different types of nucleotide substitutions, we therefore compare the results of ngsDist with disAngsd-geno only under the symmetric JC69 model. ngsDist is also the only previous method that makes use of genotype likelihoods. We should also emphasize that ngsDist was not developed for phylogenetic analysis and the results presented is therefore not the recommended scenarios for ngsDist. We therefore present the result separately (fig. 6). With an average read depth of 1, ngsDist overestimates genetic distance d when a relatively small true d is simulated (35-fold difference when true $d = 0.01$). This is due to ngsDist averaging over the posterior genotype distribution when calculating genetic distances (eq. 2 in Vieira et al. 2015). Averaging over the posterior leads to overestimation of genetic distances when the true distances are small, because uncertainty regarding the genotype is effectively interpreted as a high probability of nucleotide differences. For example, for a uniform prior on diallelic genotypes and with no data, so that only the prior contributes to the posterior, the expected number

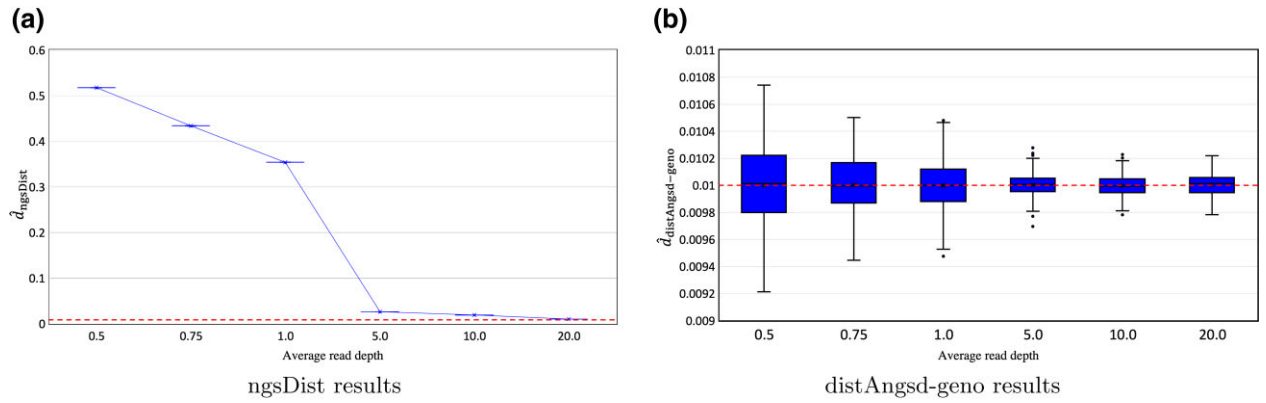

FIG. 7. Comparison of estimated genetic distances by ngsDist and distAngsd-geno for different read depths ranging from 0.5 to 20. (*a*) Genetic distance was estimated by ngsDist. (*b*) Genetic distance was estimated by distAngsd-geno. JC69 model was used. True genetic distance (d) was assumed to be 0.01, $t_1 = 0.004$, $t_2 = 0.0025$, see figure 1. The default simulation setting was applied.

of nucleotide differences per site is 0.5. If the sequencing depth is low, the genetic distance will therefore be biased towards large values especially when the true genetic distance is small. Incorporating statistical uncertainty by averaging over a posterior can, in general, lead to highly biased estimates and will be very sensitive to the choice of prior. DistAngsd-geno remains approximately unbiased but has increasing variance as depth decreases. In contrast, ngsDist has an obvious depth-dependent bias in the tested scenarios. As the depth becomes smaller, the genotype prior increasingly influences the estimated genetic distance (see fig. 7 for the genome length 1 Mb case and supplementary fig. S6, Supplementary Material online for the genome length 0.1 Mb scenario).

Inference based on Experimental Data

To compare the methods on real data, we used a previously published data set of RADseq data from oak trees (Fitz-Gibbon et al. 2017). We estimate neighbor joining trees using genetic distances based on distAngsd-geno and two more popular previous methods, ConsensusSEQ and AmbiguityGT under the JC69 model (fig. 8). We used the JC69 model to facilitate a more fair comparison among methods.

We only compare distAngsd-geno to the ConsensusSEQ and AmbiguityGT methods since (1) they perform equally as well or better than RandomSEQ and NoAmbiguity in simulations. (2) ngsDist overestimates the pairwise genetic distances and is not appropriate for phylogenetic inference (figs. 6 and 7).

Distances estimated using distAngsd-geno results in trees that are more concordant with existing taxonomic assignments, as they are closer at identifying *Quercus berberidifolia*, *Quercus durata var. gabrielensis* and *Quercus durata var. durata* (fig. 8) as monophyletic groups.

To further examine the relative performance of the methods on real data, we down-sample to obtain mean read depths of 10, 5, 1, 0.75, 0.5, and 0.25. The original mean depth per sample was 20.32 with a lowest mean depth of any sample of 10.34. We then, again, estimated neighbor-joining trees using distances inferred using distAngsd-geno, ConsensusSEQ, and AmbiguityGT. To compare the compatibility of the results with the existing taxonomy, we developed a compatibility measurement, m, that measures the amount of taxonomic compatibility observed in the trees, i.e., higher values correspond to better performance (see supplementary Section S8, Supplementary Material online). Results for all scenarios and methods can be found in table 1.

Clearly, trees estimated using distAngsd-geno are more compatible with existing taxonomy than both ConsensusSEQ and AmbiguityGT, particularly as the read depth decreases.

Discussion

We have here presented two novel methods for estimating genetic distances: distAngsd-geno and distAngsd-nuc. Both of these methods incorporate the uncertainty that is inherently associated with high-throughput sequencing data. The uncertainty is either modeled through standard diploid genotype likelihoods or, equivalently, through the application of the base quality scores in a framework that facilitates inferences in a polyploid or unknown ploidy context. Both methods can estimate the genetic distance between samples with high sequencing error rates and low average read depth.

The key characteristic of both methods is to decompose the likelihood optimization process into two parts: (1) estimation of global pseudo-observations (i.e. joint distributions of genotypes or nucleotides between each pair of samples); (2) the maximum-likelihood estimation of genetic distance (and other related parameters) based on the previously calculated pseudo-observations. This decomposition reduces the complexity of the maximum-likelihood estimation.

However, we note that this decomposition can introduce biases (See supplementary Section S4, Supplementary Material online), particularly when the sequences compared are short. The proposed methods are not intended for data consisting of very short sequences. However, the bias for our

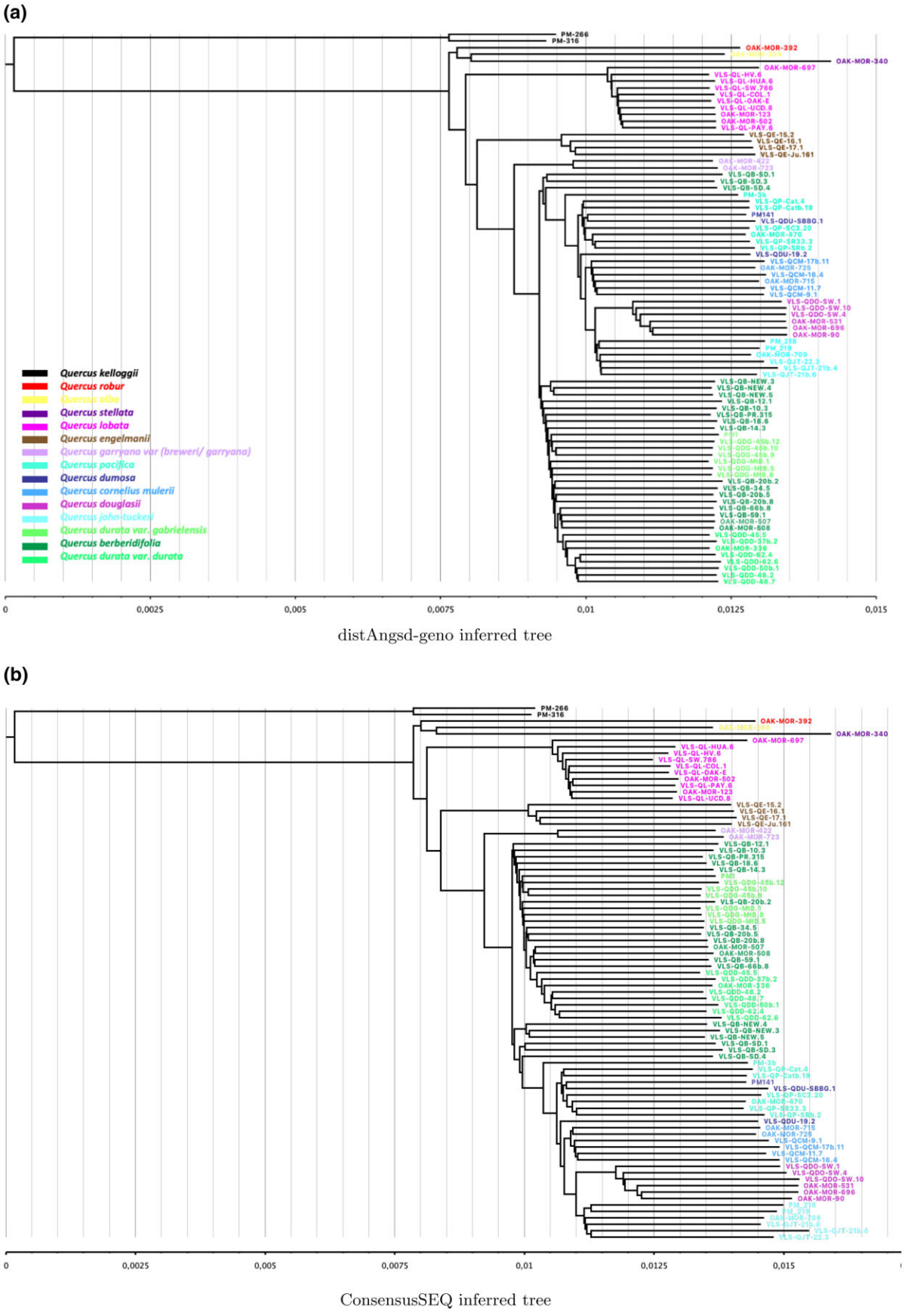

FIG. 8. Neighbor-Joining Tree of 83 oak samples inferred using genetic distances estimated by (*a*) distAngsd-geno (*b*) ConsensusSEQ , and (*c*) AmbiguityGT. All inferences are based on the JC69 model. The colors of the leaf-node labels correspond to the species shown in the legend. All three trees are estimated as unrooted trees but roots are placed between the known outgroup *Quercus kelloggii* and other oak species.

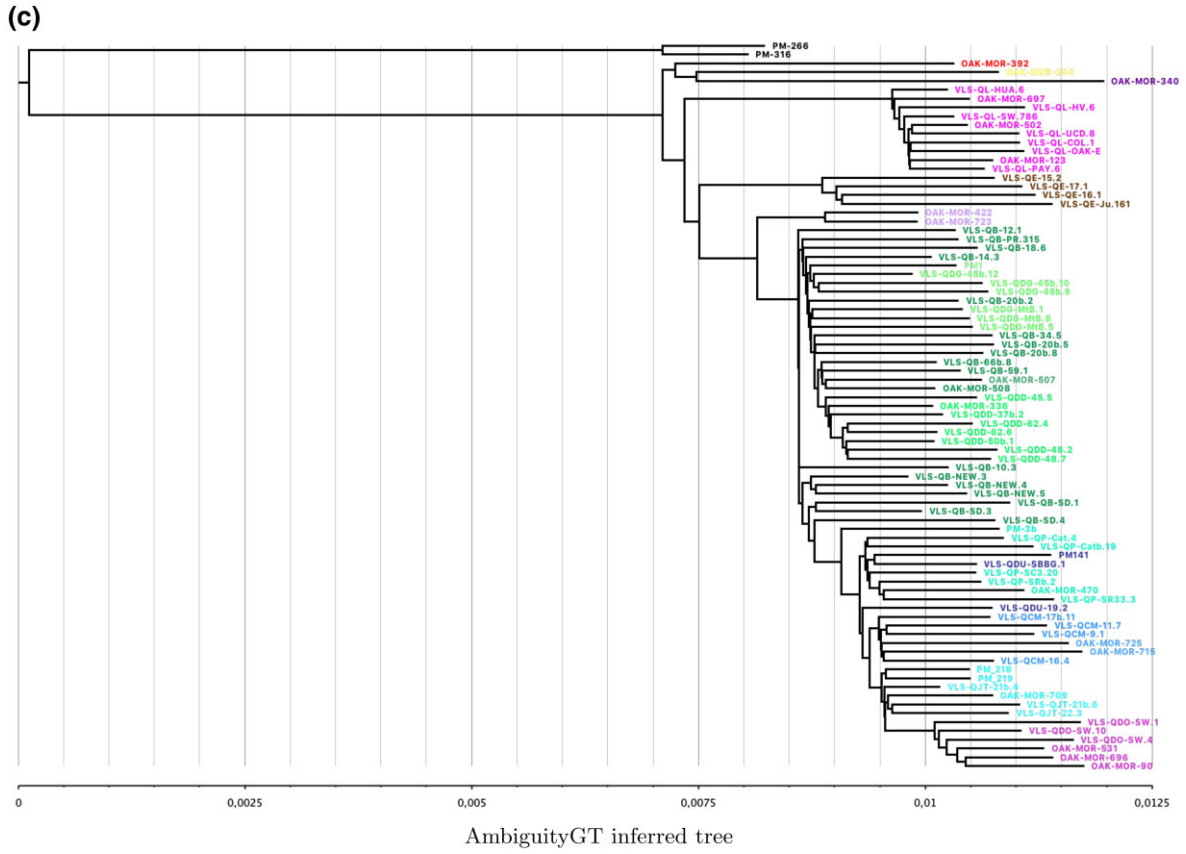

FIG. 8. Continued

Table 1. The Compatibility Measurement m between the Inferred Pairwise Distance Trees and the Prior Species Knowledge.

	Downsampling Depth		Full	10	5	1	0.75	0.5	0.25
Criterion $\lvert\Omega_1 \cap \Omega_2\rvert = 0$	**distAngsd-geno**	***m***	**17**	**17**	**17**	**16**	**16**	**17**	**13**
	ConsensusSEQ		**17**	**17**	**17**	**16**	**16**	**16**	**13**
	AmbiguityGT		**16**	**16**	**16**	**16**	**16**	**16**	**12**
Criterion $\lvert\Omega_1 \cap \Omega_2\rvert \leq 1$	**distAngsd-geno**	***m***	**75**	**75**	**73**	**74**	**73**	**70**	**57**
	ConsensusSEQ		**74**	**75**	**76**	**72**	**70**	**66**	**58**
	AmbiguityGT		**75**	**75**	**74**	**74**	**70**	**67**	**56**

NOTE:—The trees are inferred based on the original samples as well as the downsampled samples. The original 83 samples have mean read depth 20.32 with lowest depth 10.34, and each sample is downsampled to mean read depth 10, 5, 1, 0.75, 0.5, and 0.25.

proposed methods is still smaller—by a large margin—compared to any previous method.

We also simulated and inferred genetic distances based on tree topologies different from figure 1 (See supplementary fig. S7, Supplementary Material online). Such topologies can occur due to incomplete lineage sorting, i.e. when coalescence times are shorter between alleles from different individuals than from the same individual. We find that for a realistic range of divergence levels, the results of the new methods still offer higher accuracy than the previous methods (See supplementary Section S6, Supplementary Material online). The biases of the previous methods for topologies resulting from incomplete linage sorting are similar to those observed for the standard topology in figure 1, and the general conclusions from the standard simulation scenario carries over to the case of incomplete lineage sorting. One exception is for extremely large divergence levels (see supplementary fig. S10, Supplementary Material online), where the new methods will be increasingly biased.

While the distAngsd-geno results presented here assume diploidy, it can in principle be extended to any ploidy level. However, unlike the distAngsd-geno, the distAngsd-nuc does not require prior knowledge of ploidy level, and the size of joint distribution matrix of nucleotides N remains a 4×4 regardless of ploidy level. We observe that distAngsd-geno produces more accurate results for our diploid simulation scenarios but speculate that in the context of unknown ploidy distAngsd-nuc will yield more robust estimates and would therefore be more suitable.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution online*.

Acknowledgments

We thank two anonymous reviewers and editor Tom Whitehead for their careful reading and insightful comments and suggestions. We thank Prof. Victoria L. Sork and Dr. Sorel Fitz-Gibbon for kindly sharing the oak data with us. We thank Prof. Rasmus Heller for advice of processing RAD sequences. We thank Prof. Ziheng Yang for sharing code for matrix exponentiation. We thank Dr. Abigail Daisy Ramsøe for English polishing. L.Z. was funded by Lundbeck Foundation Centre for Disease Evolution: R302-2018-2155. T.S.K. was funded by a Carlsberg Foundation Young Researcher Fellowship awarded by the Carlsberg Foundation in 2019 (CF19-0712). R.N. is supported by NIH grant R01GM138634.

Data availability

Raw oak data were from previously published dataset (described in Fitz-Gibbon et al. 2017). Derived data supporting the findings of this study are available on request.

References

Árnason Ú, Lammers F, Kumar V, Nilsson MA, Janke A. 2018. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Sci Adv*. **4**(4): eaap9873.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. **81**:1084–1097.

Choi JY, Purugganan MD. 2018. Multiple origin but single domestication led to Oryza sativa. *G3-Genes Genom Genet*. **8**:797–803.

Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984. *Nucleic Acids Res*. **13**(9):3021–3030.

Criscuolo A, Gascuel O. 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC Bioinform*. **9**:166.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B*. **39**:1–38.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. **17**:368–376.

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*. **13**: 93–104.

Fitz-Gibbon S, Hipp AL, Pham KK, Manos PS, Sork VL. 2017. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (Quercus section Quercus). *Genome*. **60**:743–755.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. **14**: 685–695.

Gaunitz C, Fages A, Hanghøj K, Albrechtsen A, Khan N, Schubert M, Seguin-Orlando A, Owens IJ, Felkel S, Bignon-Lau O, *et al*. 2018. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science* **360**:111–114.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. **22**:160–174.

Hu XJ, Yang J, Xie XL, Lv FH, Cao YH, Li WR, Liu MJ, Wang YT, Li JQ, Liu YG, *et al*. 2018. The genome landscape of Tibetan sheep reveals adaptive introgression from argali and the history of early human settlements on the Qinghai-Tibetan Plateau. *Mol Biol Evol*. **36**:283–303.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. *Mammalian protein metabolism*. New York: Academic Press. p. 21–132.

Keightley PD, Halligan DL. 2011. Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics* **188**(4):931–940.

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

Klicka J, Keith Barker F, Burns KJ, Lanyon SM, Lovette IJ, Chaves JA, Bryson RW. 2014. A comprehensive multilocus assessment of sparrow (Aves: Passerellidae) relationships. *Mol Phylogenet Evol*. **77**:177–182.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinform*. **15**:356.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**:2987–2993.

Lischer HE, Excoffier L, Heckel G. 2014. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of microtus voles. *Mol Biol Evol*. **31**: 817–831.

Maldonado LL, Arrabal JP, Rosenzvit MC, Oliveira GCD, Kamenetzky L. 2019. Revisiting the phylogenetic history of helminths through genomics, the case of the new *Echinococcus oligarthrus* genome. *Front Genet*. **10**:708.

Manthey JD, Campillo LC, Burns KJ, Moyle RG. 2016. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: Piranga). *Syst Biol*. **65**:640–650.

Martin FN, Blair JE, Coffey MD. 2014. A combined mitochondrial and nuclear multilocus phylogeny of the genus Phytophthora. *Fungal Genet Biol*. **66**:19–32.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al*. 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**(9): 1297–1303.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing Data. *PLoS ONE* **7**:e37558.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. **12**: 443–451.

Nilsson MA, Zheng Y, Kumar V, Phillips MJ, Janke A. 2017. Speciation generates mosaic genomes in kangaroos. *Genome Biol Evol*. **10**: 33–44.

Potts AJ, Hedderson TA, Grimm GW. 2014. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Syst Biol*. **63**:1–16.

Sass C, Iles WJD, Barrett CF, Smith SY, Specht CD. 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ*. **4**:e1584.

Schrempf D, Minh BQ, Maio ND, von Haeseler A, Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J Theor Biol*. **407**:362–370.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, *et al*. 2016. Genomic

insights into the peopling of the Southwest Pacific. *Nature* **538**: 510–513.
Sork VL, Fitz-Gibbon ST, Puiu D, Crepeau M, Gugger PF, Sherman R, Stevens K, Langley CH, Pellegrini M, Salzberg SL. 2016. First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née (Fagaceae). *G3 (Bethesda)*. **6**(11): 3485–3495.
Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*. **73**:1162–1169.
Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. **68**:978–989.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. **10**:512–526.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci*. **17**:57–86.
Uckele KA, Adams RP, Schwarzbach AE, Parchman TL. 2021. Genome-wide RAD sequencing resolves the evolutionary history of serrate leaf Juniperus and reveals discordance with chloroplast phylogeny. *Mol Phylogenet Evol*. **156**:107022.
Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. 2015. Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biol J Linn Soc*. **117**:139–149.
Yang MA, Fan X, Sun B, Chen C, Lang J, Ko YC, Tsang Ch, Chiu H, Wang T, Bao Q, *et al.* 2020. Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**:282–288.
Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J Mol Evol*. **39**:105–111.
Yang Z. 2006. *Computational molecular evolution*. Oxford (UK): Oxford University Press.
Yuan H, Jiang J, Jiménez FA, Hoberg EP, Cook JA, Galbreath KE, Li C. 2016. Target gene enrichment in the cyclophyllidean cestodes, the most diverse group of tapeworms. *Mol Ecol Resour*. **16**:1095–1106.