JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 6, 2021
Mary Ann Liebert, Inc.
Pp. 587–600
DOI: 10.1089/cmb.2020.0375

Estimating Genetic Similarity Matrices Using Phylogenies

SHIJIA WANG,[1,*] SHUFEI GE,[2] CAROLINE COLIJN,[3] PRISCILA BILLER,[3]
LIANGLIANG WANG,[4] and LLOYD T. ELLIOTT[4,†]

ABSTRACT

Genetic similarity is a measure of the genetic relatedness among individuals. The standard method for computing these matrices involves the inner product of observed genetic variants. Such an approach is inaccurate or impossible if genotypes are not available, or not densely sampled, or of poor quality (e.g., genetic analysis of extinct species). We provide a new method for computing genetic similarities among individuals using phylogenetic trees. Our method can supplement (or stand in for) computations based on genotypes. We provide simulations suggesting that the genetic similarity matrices computed from trees are consistent with those computed from genotypes. With our methods, quantitative analysis on genetic traits and analysis of heritability and coheritability can be conducted directly using genetic similarity matrices and so in the absence of genotype data, or under uncertainty in the phylogenetic tree. We use simulation studies to demonstrate the advantages of our method, and we provide applications to data.

Keywords: genetic similarity, infinite sites model, phylogenetic tree.

1. INTRODUCTION

The computation of genetic similarities among samples (or taxa, or subjects, or individuals) is a key step in the analysis of quantitative genetic traits and in estimating the heritability of traits through variance component approaches (Chen and Witte, 2007; Tzeng and Zhang, 2007; Malo et al., 2008). For example, linear mixed models (LMMs) are popular methods for genome-wide association studies, and matrix-variate normal methods are popular for heritability and coheritability analyses. Both of these approaches demand efficient methods to determine genetic relatedness among samples (Kang et al., 2010; Lippert et al., 2011; Listgarten et al., 2013). The genetic relatedness among samples is usually specified through a genetic similarity matrix (Patterson et al., 2006; Thompson, 2013) derived empirically from genetic

[1] School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China.
[2] Institute of Mathematical Sciences, ShanghaiTech University, Shanghai, China.
[3] Department of Mathematics, Simon Fraser University, Burnaby, Canada.
[4] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada.
*ORCID ID (https://orcid.org/0000-0003-0339-1716).
†ORCID ID (https://orcid.org/0000-0003-2187-7314).

© Shijia Wang et al., 2021; Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons License [CC-BY] (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sequences, or from a kinship matrix (Boyce, 1983; Kirkpatrick et al., 2019) derived from a pedigree. Genetic sequences are often described by a series of genome locations at which mutations may be observed in the ancestry of the subjects. We consider single-nucleotide polymorphisms (SNPs), locations at which a single DNA base pair can appear with more than one form (or, allele). The most common form is known as the major allele, and the less common forms are the minor alleles. We refer to the matrix resulting from the inner products of genetic sequences as the *empirical genetic similarity matrix*. This matrix is formed according to the following definitions. For the SNP at locus m, let G_m denote the column vector of alleles. The genetic similarity between samples i and j for a haploid sample is defined as follows:

$$K_{ij}^{G} = \frac{1}{M}\sum_{m=1}^{M}\frac{(G_{im}-\mu_m)(G_{jm}-\mu_m)}{\sigma_m^2}. \tag{1}$$

Here G_{im} is the genotype of leaf i at marker m ($G_{im} \in \{0, 1\}$), μ_m, σ_m^2 are the empirical mean and variance of the SNPs at marker m, and M is the total number of loci genotyped (Patterson et al., 2006). We assume that $G_{im}=1$ indicates the event that sample i inherits the minor allele at marker m, and $G_{im}=0$ indicates the event that sample i inherits the major allele at marker m. The minor allele is defined as the allele that occurs less often among all of the samples. Note that permutations of the loci do not affect the empirical genetic similarity. This formulation of genetic similarity is used in LMMs (Listgarten et al., 2012), and multiphenotype models (Yang et al., 2011). The matrix entry K_{ij}^{G} represents the expected genetic contribution to the correlation between phenotypes for samples i and j, and thus, K^G may be used as a fixed effect. Other methods for computing relatedness (e.g., patristic distance or identity-by-descent [IBD]) have closed forms, but are not directly exploitable in fixed effects models.

The computation of genetic similarity among samples using Equation (1) requires genotyped sequences, and several challenges may arise in the computation. First, the genetic sequences may not be readily available. This may be the case when extinct species are examined. Second, it is hard to assess the uncertainty for empirical genetic similarities in cases for which the genotyped sequences have low quality, or are homoplastic. This case may arise when examining bacterial genomes or de novo sequences. These challenges motivate us to propose an approach that does not involve genetic sequences.

In this article, we develop a method for estimating expected values for the entries of a genetic similarity matrix using a phylogeny [i.e., a closed form for the expected value of Eq. (1), conditioned on a tree]. Our proposed approach does not require that the sequences for individuals be genotyped (or measured). Our method allows analysis based on Equation (1), such as analysis using fixed effects models, to proceed for situations in which genotypes are not available but an approximation of the molecular phylogeny is, or for situations in which there is uncertainty in Equation (1) stemming from low-quality genotyping or short genomes. For example, in some studies of extinct species, genotypes might not be available but an approximate molecular phylogeny, based on morphological data and geological dates, might be. Instead of requiring individual genotypes, we assume that a phylogeny is given.

The relatedness among samples is computed by integrating over all mutations occurring in the branches of a tree, under the assumption of an infinite sites model (Ma et al., 2008), with a constant mutation rate for the evolutionary process. The infinite sites model is a popular and simple model for mutations in genome evolution, and it is a reasonable model when the genome is large. The infinite sites model postulates that new mutations are always at novel loci (and not re-entrant). Genetic recombination among samples is not considered in this approach (however, the approach for samples with recombination is clear), as we mainly consider genetic similarities at the species level. We numerically demonstrate that our proposed *expected genetic similarity matrix* is asymptotically equivalent to the empirical genetic similarity matrix, with an infinite number of genotyped loci (provided that the tree is correct).

The main application of our work is to provide a correspondence between phylogenetic trees and the normalized genetic similarity matrices (Patterson et al., 2006) that are standard in LMMs and heritability analysis. Often, many sampled molecular phylogenies are provided conditioned on genotypes (e.g., with the Bayesian Evolutionary Analysis Sampling Trees (BEAST) software; Drummond and Rambaut, 2007). Our work allows computation of the expected value of the standard and normalized genetic similarity matrix (Patterson et al., 2006) for each of the sampled trees. These matrices can then be used to average over LMMs applied to each sample (respecting the uncertainty in the genetic similarity matrix), or they can be used in Bayesian models extending the usual LMM paradigm. In contrast, work based on computing Equation (1) directly from genotypes provides a point estimate of the genetic similarity matrix (without respecting the uncertainty implied by the genotypes).

In our simulations, the expected genetic similarity matrix is invariant to the number of samples and to the mutation rate in the infinite sites models. Our approach is more accurate than the multidimensional scaling (MDS) approach implemented in the software package *pyseer* (Lees et al., 2018) and Gaussian distance similarity matrices (Ickstadt et al., 2005; González-Recio et al., 2008). The MDS approach for genetic similarity matrix estimation also operates on phylogenies, but is not based on integration over all mutations.

We use our approach to compute the genetic similarity matrix for a set of ancient hominin species, with the phylogeny found using BEAST on the set of morphological phenotypes used in Dembo et al. (2016). Genetic similarity matrices are often used to partition the covariance matrix into additive components that arise from genetic, environmental, or random factors (Wang et al., 2011; Dahl et al., 2016). Given this estimated genetic similarity matrix, we compute the component of genetic covariance that is inherited along the tree (before and after speciation events) for the heights of the ancient hominin species. In particular, we apply an LMM (Yang et al., 2011) to compartmentalize the covariance matrix of the heights of the species, such that the genetic component is a scaled version of our estimate for the genetic similarity matrix. This demonstrates how our estimate can be used to determine genetic aspects without access to measured genotypes (instead, with access to a tree that we assume approximates the correct phylogeny). In addition, we apply our algorithm to evaluate the uncertainty of genetic similarities among hominin species.

1.1. Related work

Genetic similarity is a basic aspect of many approaches in genetic analysis. IBD (Whittemore and Halpern, 1994) is a popular measure for genetic similarity, based on the identification of stretches of genetic sequences that have identical ancestral origin. Kinship coefficients describe the probability that two random alleles from a pair of individuals are IBD, and these coefficients are commonly used to measure genetic similarity between a pair of individuals. There are several ways to compute kinship coefficients, including from genetic data, and also based on pedigree graphs (Maruyama and Yasuda, 1970). The idea of using trees to define similarity matrices (for use as fixed effects) is explored in Houswórth et al. (2004). However, that work uses estimates of similarity matrices based on pedigrees (and does not use the definition of genetic similarity as a normalized inner product of genetic sequences, as is standard in applications of LMMs to genome-wide association studies; Listgarten et al., 2013). Similarly, in Abney (2009), a novel graphical algorithm for computing the kinship coefficients using graph traversal is developed: ‘‘kinship graph,’’ and in Thornton et al. (2012) a robust method ‘‘REAP’’ is developed, to estimate IBD-sharing probabilities and kinship coefficients for admixtures. However, neither of these methods is designed to estimate the normalized inner product of genetic sequences [Eq. (1)].

Computational concerns for kinship matrices have also been considered: Kirkpatrick et al. (2019) propose a fast algorithm for calculation of kinship coefficients for individuals in large pedigrees. Our work provides an algorithm with similar computational complexity.

In further lines of research, if genetic sequences are observed, a variety of measures have been used to compute genetic similarities using categorical data clustering (Ickstadt et al., 2005; González-Recio et al., 2008) by creating dissimilarity scores for genotypes and then converting the dissimilarity scores to similarity scores. For example, Murray et al. (2017) present the k-mer weighted inner product, an assembly-based estimator of computation of genetic similarities among individuals that is also alignment-free. The pairwise similarity is obtained from k-mer counts. In contrast, less work has been done in assessing the statistical relationship between phylogenetic trees (rather than pedigrees or alignment-free methods) and genetic similarity matrices. The most similar approach is based on MDS, which has been applied to the control of LMMs (Lees et al., 2018). This MDS approach also provides a deterministic function that outputs a similarity matrix given a fixed tree. In typical protocols, the fixed tree used in the production of the similarity matrix may be produced from genotypes using BEAST.

2. METHODS

Let T be an unrooted binary tree with leaves $i \in \{1, \dots, N\}$. The phylogeny T represents relationships among N taxa through a tree topology τ and a set of branch lengths $\boldsymbol{e} = (e_1, e_2, \dots, e_{2N-3})$. The leaves of a sampled tree T are the samples in the study. Each interior node of T represents the most recent common ancestor of the two children of that node, and the branch lengths are proportional to the evolutionary distances between pairs of nodes. An unrooted tree represents the relatedness of leaves without making assumptions about an earliest common ancestor. In contrast, a rooted binary tree describes the relatedness for a set of leaves in the tree from a single common ancestor at the root.

In computational genetics, the infinite sites model is commonly used to model genetic variation (Kimura, 1969). In this model, we assume that polymorphism arises by single mutations of unique sites at locations within the genetic sequence, with all mutations occurring at different positions, implying that all genetic variants are biallelic. Let $\mathbf{G} = [G_1, G_2, \ldots, G_M]$ denote genotype data observed at M genetic variants, with G_m denoting a column vector of alleles for the m-th SNP for all N subjects. The genetic similarity matrix of the leaves (Patterson et al., 2006) is an $N \times N$ symmetric matrix K defined in Equation (1), in which K^G_{ij} denotes the genetic similarity between sample i and sample j. Our goal is to compute the expected value of K^G_{ij} given a tree T. By additivity of expectations, from Equation (1) we arrive at the following expectation through integration over all mutations:

$$K^T_{ij} = \mathbb{E}[K^G_{ij} | t = T] = \mathbb{E}\left[\frac{(G_i - \mu)(G_j - \mu)}{\sigma^2} | t = T\right]. \tag{2}$$

Here G_i is a random variable giving the genotype of a marker placed at a random location of the tree, and μ, σ^2 are random variables giving the mean and variance of G_i as an element of the set $\{0,\ 1\}$ (assuming haplotype data).

To compute Equation (2), we integrate the location of the marker over the tree, noting that expectation splits linearly over the union of the domain of integration. Assuming a neutral model with a constant mutation rate, the values of G_i are completely determined by the location of the marker, and they are constant over each edge of the tree. So, Equation (2) can be rewritten as a weighted sum over edges of the tree. The weights are given by $|e_d|/|T|$, where $|e_d|$ is the branch length of an edge e_d and $|T|$ is the sum of the branch lengths of all edges in the tree. The expected value on the right hand side of Equation (2) is thus given by the following:

$$K^T_{ij} = \sum_{e_d} \frac{|e_d|}{|T|} \frac{(G_{ie_d} - \mu_{e_d})(G_{je_d} - \mu_{e_d})}{\sigma^2_{e_d}}. \tag{3}$$

The values G_{ie_d}, μ_{e_d}, $\sigma^2_{e_d}$ are found by considering a mutation on each edge e_d. Note that Equation (3) is an approximation of expected value of genetic similarity found by assuming that the tree provides the correct phylogeny. Hence, it represents the genetic similarity between taxa i and j. Under our assumptions, μ_{e_d}, $\sigma^2_{e_d}$ are summary statistics for G, and K^T is a deterministic function of these variables according to Equation (1). These values can be computed by considering the following steps for each element of the sum from Equation (3).

1. Let c_0 and c_1 be the bipartition of taxa induced by segregation over edge e_d in the unrooted tree T. Figure 1 shows an example of this operation. Without loss of generality, we assume that the number of leaves in c_0 is greater than or equal to that of c_1. This means that each leaf in c_0 will have the major allele, and each leaf in c_1 will have the minor allele.
2. The genotype G_{ie_d} is 1 if leaf i is in c_1 and G_{ie_d} is 0 if leaf i is in c_0.
3. The mean μ_{e_d} is the number of leaves in c_1, divided by the total number of leaves.

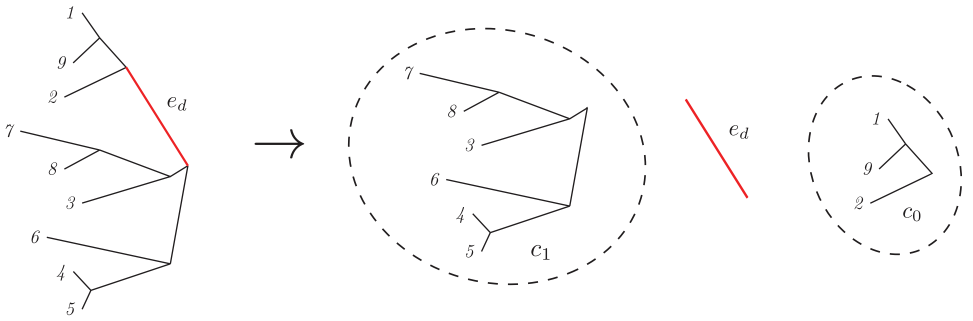

FIG. 1. Bipartition of taxa induced by an unobserved genetic variant on edge e_d. Only the membership of each taxon in the bipartitioned set is needed in the computation of the expected mean μ_{e_d}, and variance $\sigma^2_{e_d}$.

4. $\sigma^2_{e_d}$ is the variance of the Bernoulli distribution implied by the allele:

$$\sigma^2_{e_d} = \mu_{e_d}(1 - \mu_{e_d}). \tag{4}$$

Note that if c_0 and c_1 are the same size, then $\mu_{e_d} = 1 - \mu_{e_d} = 0.5$, and $\frac{(G_{ie_d} - \mu_{e_d})(G_{je_d} - \mu_{e_d})}{\sigma^2_{e_d}}$ does not depend on which allele is assigned to 0 or 1 (i.e., breaking ties for c_0 and c_1 one way or the other leaves K^T_{ij} unchanged). Equation (3) thus suggests the following $\mathcal{O}(N^3)$ algorithm (Algorithm 1) for computing the expected genetic similarity matrix, given the tree T. In this algorithm, A' denotes the transpose of the matrix A. We provide an open source software implementation for this method (https://github.com/shijiaw/Expected-Genetic-Similarity-Matrices).

Algorithm 1. From Phylogeny to Expected Genetic Similarity Matrix

```
Inputs: A phylogenetic tree $T$ with a tree topology $\tau$ and a set of branch length $\boldsymbol{e} = (e_1, e_2, \ldots, e_{2N-3})$.
Output: An $N \times N$ expected genetic similarity matrix $K^T$.
Initialize $K^T_{ij} \leftarrow 0$.
Compute the total tree distance $|T|$ by adding up branch lengths $\boldsymbol{e}$.
for $e_d \in \{e_1, e_2, \ldots, e_{2N-3}\}$ do
    Compute $\mu_{e_d}$ and $\sigma^2_{e_d}$
    for $i \in \{1, 2, \ldots, N\}$ do
        for $j \in \{i, \ldots, N\}$ do
            Set $K^T_{ij} \leftarrow K_{ij} + |e_d|/|T| \cdot (G_{ie_d} - \mu_{e_d})(G_{je_d} - \mu_{e_d})/\sigma^2_{e_d}$.
Set LowerTriangle$(K^T) \leftarrow$ LowerTriangle$(K^{T\prime})$.
return $K$.
```

3. EXPERIMENTS

In our numerical experiments, we use the *ms* software (Hudson, 2002) to simulate binary trees and genetic variation (*ms* is a coalescent simulator that generates genetic samples under the assumption of the neutral model and the infinite sites assumption). We also assume constant effective population size N. The branch lengths are thus in units of $2N$ generations. We simulate genetic sequences for haploidy of population and assume no recombination during the history (this assumption is appropriate for evolutionary timescales with species for which lateral transfer is negligible).

3.1. Simulation 1

In this simulation study, we numerically demonstrate the consistency of our algorithm. We simulate data sets for three scenarios: A, B, C. In scenario A, we simulate 100 trees, with $N = 20$ samples (or subjects, or taxa) for each tree. Also, for each tree, we simulate two sets of genetic sequences, each with $M = 1000$ or 20 loci. In scenario B, we simulate 100 trees with $N = 6$ taxa, each with $M = 1000$ loci. In scenario C, we simulate 100 trees with $N = 20$, $M = 1000$, and for each tree we simulate SNPs with a neutral mutation rate of $\mu = 7.5$ mutations on the entire sequence per generation. In the *ms* software, the mutation parameter $\theta = 2N\mu$ (Hudson, 2002).

For each scenario, we compute the empirical genetic similarity matrix using the simulated genotypes to form a ground truth. Then, we compute the expected genetic similarity matrix using Algorithm 1 and the simulated tree. Figure 2 displays scatter plots comparing our method for computing genetic similarity matrices to the ground truth [the ground truth is found by applying Equation (1) to the simulated genotypes]. Figure 2 shows that for a higher value of M, the correlation among the entries of the genetic similarity matrices computed from trees and from genotypes may be stronger than it is for low values of M (the number of loci). This figure also shows that even a small number of samples may produce unbiased estimates of the genetic similarity matrix. We also show that for situations with small numbers of samples, the estimates may still improve for higher values of M.

The entries of genetic similarity matrices computed from trees and genotypes become closer together as we increase M on the genotype from 20 to 1000, providing evidence for consistency of our methods over the range of M considered. In addition, these simulations suggest that the expected genetic similarity matrix is invariant to the number of individuals and the neutral mutation rate.

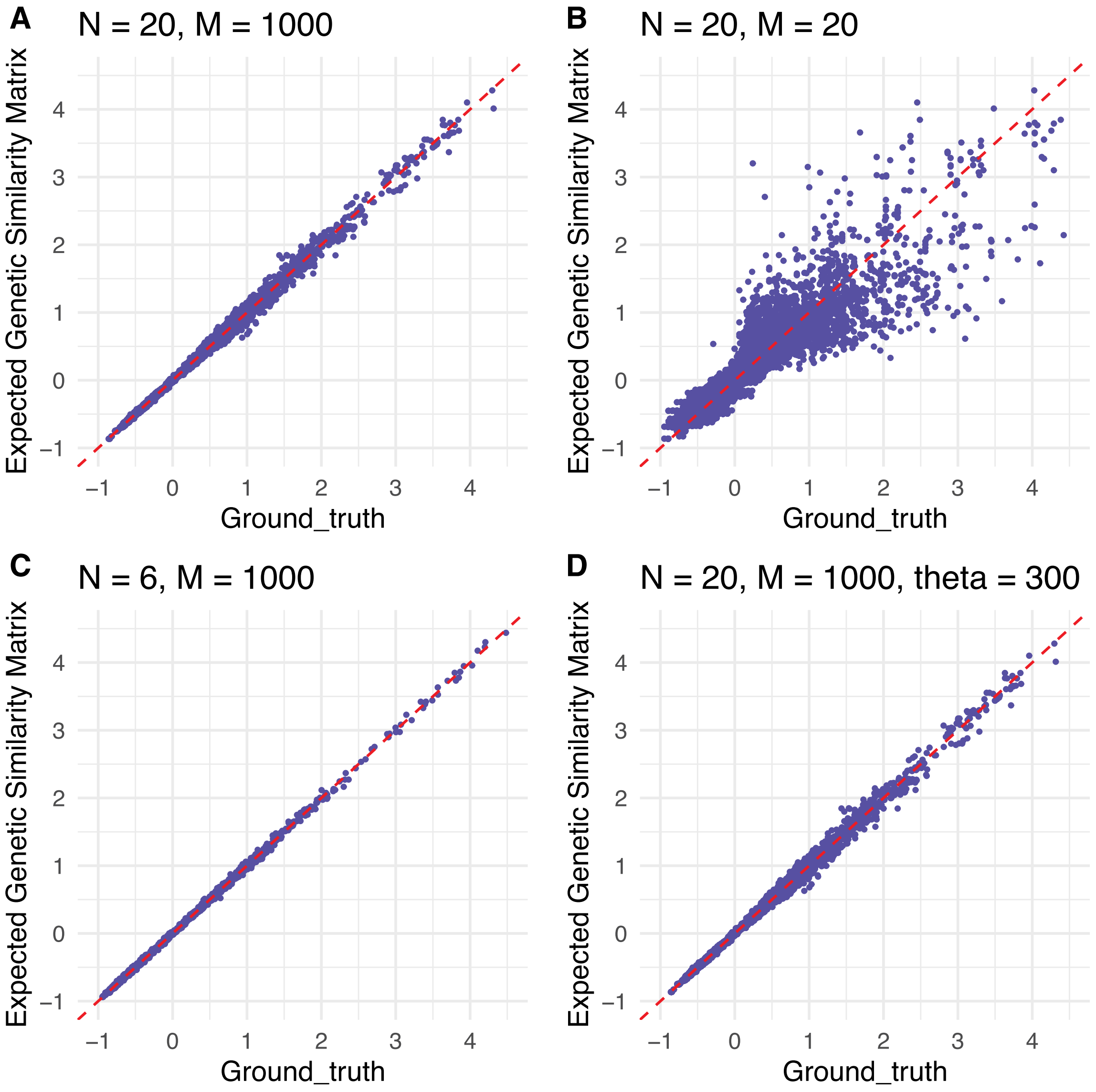

FIG. 2. Comparison of simulated genetic similarity matrices from trees and genotypes. Scatter plots are provided for entrywise differences between the genetic similarity matrices produced by our method and the ground truth [from Equation (1), applied to the simulated genotypes]. The top left and top right panels show the entrywise differences for the two conditions for scenario **(A)**, and the bottom left and bottom right panels show the entrywise differences for scenarios **(B, C),** (respectively).

We conduct an additional set of experiments to investigate the difference between entries of the expected genetic similarity matrix and the ground truth as *(A):* a function of number of loci for each sequence; *(B):* a function of number of samples (or subjects, or taxa). For *(A)*, we simulate 100 trees, with $N = 50$ samples (or subjects, or taxa) for each tree. In addition, we simulate eight sets of genetic sequences, each with M = 10, 30, 100, 300, 1000, 3000, 10,000, or 30,000 loci, for each tree. Figure 3 displays the difference between the expected genetic similarity matrices and the ground truth with different numbers of loci for each fixed number of sequences. The violin plots denote the entrywise difference between the expected genetic similarity matrices and the ground truth ($K^G_{ij} - K^T_{ij}$). This figure indicates that for a higher value of M, the correlation among the entries of the genetic similarity matrices computed from trees and from genotypes is larger than it is for low values of M.

For *(B)*, we simulate six sets of trees, each with $N = 10$, 30, 100, 300, 1000, 3000 taxa for each tree, and with 100 trees for each set. We simulate genetic sequences, with 1000 loci, for each tree. Table 1 shows the summary statistics (mean, standard deviation, 2.5%, 25%, 50%, 75%, 97.5% quantiles) for entrywise

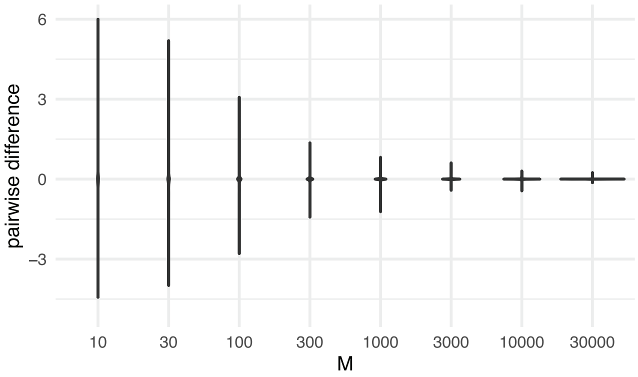

FIG. 3. Comparison of simulated genetic similarity matrices from trees and genotypes. The violins are provided for entrywise differences between the genetic similarity matrices produced by our method and the ground truth.

differences between the genetic similarity matrices produced by our method and the ground truth ($K^G_{ij} - K^T_{ij}$). This table indicates that the correlation among the entries of the genetic similarity matrices computed from trees and from genotypes is invariant to the number of taxa (N). This invariance is further supported in Supplementary Figure S1 through identical violin plots.

3.2. Simulation 2

In our second simulation study, we compare our tree-based genetic similarity matrix approach with other approaches that measure genetic similarities using phylogenies. The first approach we compare with is a Gaussian distance similarity matrix (González-Recio et al., 2008; Ickstadt et al., 2005), in which we first compute the distance between two samples i and j, d_{ij}, by computing the length of the shortest path between them on the tree. Then we convert the distance to a similarity matrix through the equation $K^S_{ij} = \exp(-\lambda d_{ij}/|T|)$. Here λ is a fixed bandwidth. We refer to Supplementary Appendix SA1 for more details of this approach. This approach is similar to Ickstadt et al. (2005) and González-Recio et al. (2008), but with distances computed through the phylogeny. The second approach we compare with is the MDS approach implemented in the software package *pyseer* (Lees et al., 2018). In the MDS approach, the similarity between each pair of samples is calculated based on the shared branch length between the pair's most recent common ancestor and the root. MDS is then performed on the resulting similarity matrix. We denote the genetic similarity matrix computed via *pyseer* by K^{MDS}_{ij} (note that we do not perform the MDS itself, and instead compare the similarity matrices directly).

We simulate 100 trees, with $N = 20$ taxa in each tree. For each tree, we simulate genetic sequences with $M = 1000$ loci. We compute the genetic similarities via K^T_{ij} (our method, based on the tree), K^S_{ij}, K^G_{ij} (the inner product from the sampled genotypes), and K^{MDS}_{ij}. Figure 4 displays the comparison of genetic similarity matrices using trees. The violins denote the differences between $K^G_{ij} - K^T_{ij}$ (red), $K^G_{ij} - K^S_{ij}$ (green), and $K^G_{ij} - K^{\mathrm{MDS}}_{ij}$ (blue). The empirical and expected genetic similarity matrices are closer to each other than they are to the Gaussian distance similarity matrix, and are closer to each other than they are to the MDS similarity matrix.

Note that the simulation conditions of the comparison between our method and the empirical genetic similarity matrix are the same as the upper left panel of Figure 2. For the $K^G_{ij} - K^S_{ij}$ condition, we note that

Table 1. Comparison of Simulated Genetic Similarity Matrices from Trees and Genotypes As a Function of Number of Taxa

No. of taxa	*Mean*	*SD*	2.5%	25%	50%	75%	97.5%
10	0.000	0.028	−0.054	−0.013	0.000	0.012	0.057
30	0.000	0.033	−0.055	−0.011	0.000	0.012	0.054
100	0.000	0.034	−0.049	−0.011	0.000	0.011	0.050
300	0.000	0.035	−0.039	−0.010	0.000	0.010	0.044
1000	0.000	0.034	−0.041	−0.009	0.000	0.010	0.036
3000	0.000	0.036	−0.055	−0.008	0.002	0.012	0.038

The summary statistics (mean, SD, 2.5%, 25%, 50%, 75%, 97.5% quantiles) for entrywise differences between the genetic similarity matrices produced by our method and the ground truth are close. This observation is further supported in Supplementary Appendix SA2.
SD, standard deviation.

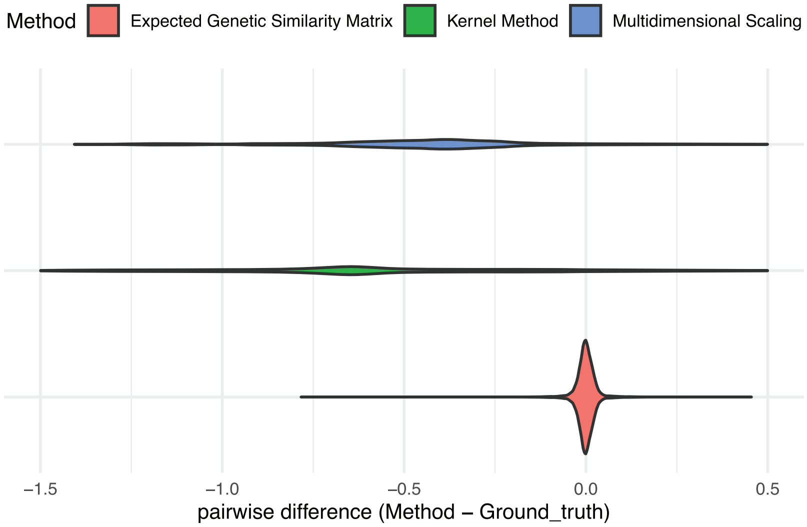

FIG. 4. Comparison of expected genetic similarity matrix approaches: $K_{ij}^G - K_{ij}^T$ (our method, bottom), $K_{ij}^G - K_{ij}^S$ (middle), and $K_{ij}^G - K_{ij}^{MDS}$ (top). The entries of the expected genetic similarity matrices (the red violin, thin) are close to the empirical genetic similarity matrix. They are closer to the empirical genetic similarity matrix than are the entries of the Gaussian distance similarity matrices (the green violin), or the multidimensional scaling similarity matrices (the blue violin).

the median difference is far from zero (as the range of the Gaussian distance method is positive). However, K_{ij}^G and K_{ij}^S are still correlated ($\rho = 0.27$, $p < 0.001$). Here ρ denotes the correlation between entries of K_{ij}^G and K_{ij}^S, and p denotes the p-value for the null hypothesis that ρ is equal to 0. For the $K_{ij}^G - K_{ij}^{\mathrm{MDS}}$ condition, the median difference is still far from zero, but less far than it is for the $K_{ij}^G - K_{ij}^S$ condition. The $K_{ij}^G - K_{ij}^S$ still involves correlation between the entries of the similarity matrices ($\rho = 0.56$, $p < 0.001$).

3.3. Ancient hominin data

We consider two experiments on ancient hominin data. In the first experiment, we apply our method to compute the genetic similarities between eight hominin species. Fossil remains of a previously unknown ancient hominin species (*Homo naledi*) were discovered in South Africa (Berger et al., 2015) and the DNA of this ancient hominin species remains unsequenced. While some ancient hominin species have been sequenced, often geological, paleontological, morphological, or anthropological observations are used to assess the evolutionary relationships among extinct species (Dembo et al., 2016). In this way, a tree *approximating* the molecular phylogeny can still be constructed. We assume that the branch lengths of the molecular phylogeny are proportional to the phylogeny inferred from paleontological dates (and in experiments described later in this section we assume that molecular phylogeny is proportional to morphological trees). Recent work in hominin phylogeny suggests that morphological trees broadly match molecular phylogeny (Wiens, 2004; Wood and Boyle, 2017). However, convergent evolution and high substitution rates modulate the validity of this assumption (Berger et al., 2017; Scally et al., 2012). Our method allows us to compute genetic similarity matrices that could then be used in the analysis of heritability or coheritability (as is done in Dahl et al., 2016), in the absence of any genetic sequences.

We use our proposed approach to compute the expected genetic similarity matrix for *Homo sapiens* and seven extinct cousins or ancestors of humans (*Homo habilis*, *Homo rudolfensis*, *Georgian Homo. erectus*, *African Homo erectus sapiens*, *Asian Homo erectus sapiens*, *Homo antecessor*, and *Homo neanderthalensis*). The phylogenetic tree for the species that we consider is provided and computed in Dembo et al. (2016) using geological dates. While Dembo et al. (2016) reconstruct the phylogenetic tree of 24 species, we consider the 7 hominin species most closely related to humans, as well as humans themselves (yielding 8 species). Figure 5 shows the heatmap for the expected genetic similarity matrix for the eight hominins considered in this article. The genetic similarities between *H. rudolfensis* and *Georgian H. erectus*, *H. rudolfensis* and *H. habilis*, and *Georgian H. erectus* and *H. habilis* are larger than the rest.

We also provide the specific values of the heatmap in Figure 5 of the similarities in Supplementary Table S1 of Supplementary Appendix SA3. This table could be used in conjunction with models such as

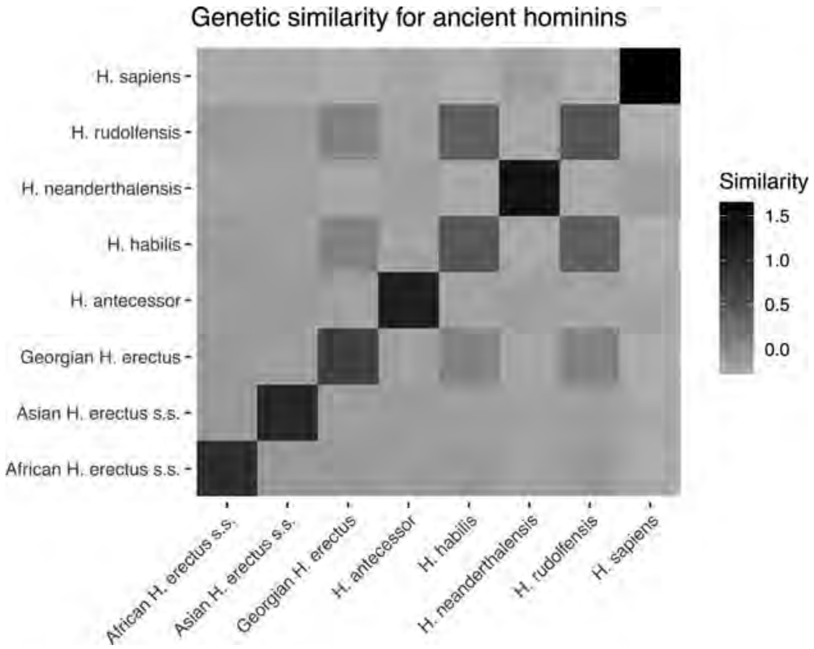

FIG. 5. Genetic similarity matrix for eight hominin species as a heat map, computed using geological dates and Algorithm 1.

Dahl et al. (2016) to conduct heritability or coheritability analysis on traits of the species considered. To that end, we consider a heritability analysis on the average heights of genetic males in the eight hominin species. We gather these average heights of ancient hominins from the following references: Carretero et al. (1999); Helmuth (1998); Lordkipanidze et al. (2007); and McHenry and Coffing (2000). We estimate variance components from an LMM (a standard model for methods such as genome-wide complex trait analysis; Yang et al., 2011). The form of this model is as follows:

$$\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{\epsilon};\ \boldsymbol{b} \sim \mathcal{N}(0, \sigma_g^2 K^T),\ \ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_e^2 I).$$

Here $\boldsymbol{y}$ denotes the standardized phenotypes (genetic male heights), and I is an identity matrix. The random effects $\boldsymbol{b}$ are the genetic effects. The term $\boldsymbol{\epsilon}$ encodes the environmental effect.

Narrow sense heritability, denoted h^2, is the fraction of the variance of y due to the genetic component. This quantity is defined as follows:

$$h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2).$$

We fit the LMM with K^T (our expected genetic similarity matrix) computed earlier in this section. We conducted inference using Markov chain Monte Carlo (Gibbs sampling) (MCMC) for the parameters of the LMM, using 10,000 iterations, of which the first 5000 are discarded for burn-in. Figure 6 shows the MCMC trace plots of σ_g^2 and σ_e^2. The red dashed lines indicate posterior means for σ_g^2 and σ_e^2 with the burn-in discarded. Figure 7 shows the histograms for posterior samples of σ_g^2, σ_e^2, and h^2 with the burn-in discarded. The red lines indicate the posterior means, and the blue dash lines indicate the 95% credible intervals.

We obtain posterior means and 95% credible intervals $\hat{\sigma}_g^2 = 0.640$ (0.185, 1.754) and $\hat{\sigma}_e^2 = 0.526$ (0.160, 1.415). Therefore, from $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ we estimate that the narrow sense heritability for this trait is $\widehat{h^2} = \hat{\sigma}_g^2 / (\hat{\sigma}_e^2 + \hat{\sigma}_g^2) = 0.538$, with 95% credible interval given by (0.176, 0.870). Note that the validity of these estimates depends crucially on the assumptions of the model: First, that the paleontological tree is proportional to the phylogeny, and also that the assumptions of the neutral model and infinite sites model hold. For humans, the narrow-sense heritability of height suggested by Yang et al. (2015) is between 0.6 and 0.7.

3.4. Uncertainty assessment in hominin species

In this section, we present a second experiment on hominin data. We apply our approach to evaluate the uncertainty for genetic similarities for 24 hominin species. The DNA sequences of some hominin species

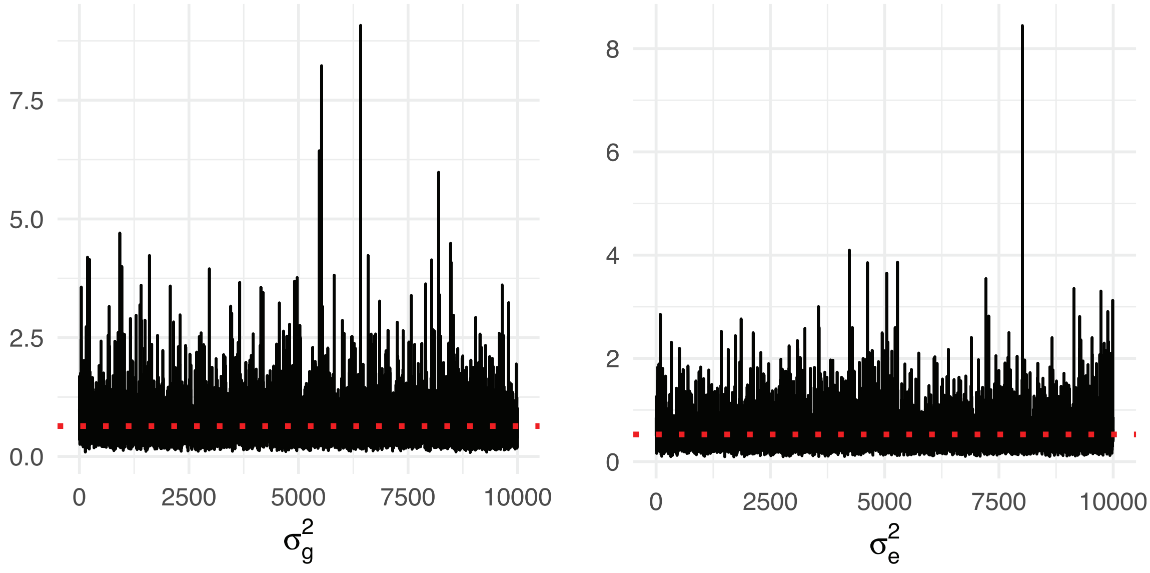

FIG. 6. MCMC trace plots of σ_g^2 and σ_e^2. The red dashed lines indicate posterior means for σ_g^2 and σ_e^2 with initial 5000 iterations as burn-in. MCMC, Markov chain Monte Carlo.

remain unavailable (e.g., *Sahelanthropus tchadensis*). We compute the genetic similarity matrix for a posterior sample of phylogenetic trees for the hominin species. We used 10,000 posterior tree samples (after thinning) created through MCMC with BEAST2 (Bouckaert et al., 2014). Details of the BEAST2 run are provided in Supplementary Appendix SA4. The morphological data used are provided in Dembo et al. (2016), but BEAST2 was used rather than Mr. Bayes 3.2.4 (Ronquist et al., 2012).

The resulting matrices represent the uncertainty of genetic similarities among the hominin species. Figure 8 displays the uncertainty of genetic similarities between *Ardipithecus ramidus* and *Australopithecus anamensis*, *Gorilla gorilla* and *Pan troglodytes*, *Parantropus robustus* and *Homo naledi*, and *P. troglodytes* and *Australopithecus afarensis*. We refer readers to Supplementary Appendix SA5 (Supplementary Figs. S2–S4) for the posterior mean and the 95% credible interval for a heatmap of the genetic similarity matrix for all 24 hominin species.

4. DISCUSSION

We provide an unbiased estimate for the genetic similarity matrix of samples, conditioned on a phylogenetic tree. This can be used to perform heritability and coheritability estimates based on models such as Dahl et al. (2016), and can be used as a building block for LMM-like models in which uncertainty about

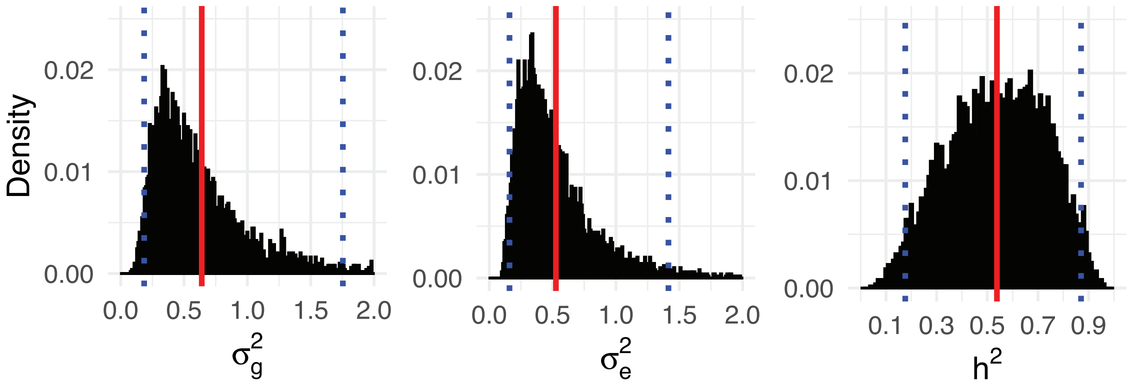

FIG. 7. Histograms for posterior samples of σ_g^2, σ_e^2, and h^2. The red lines indicate posterior means for σ_g^2, σ_e^2, and h^2, and the blue dash lines indicate the 95% credible intervals.

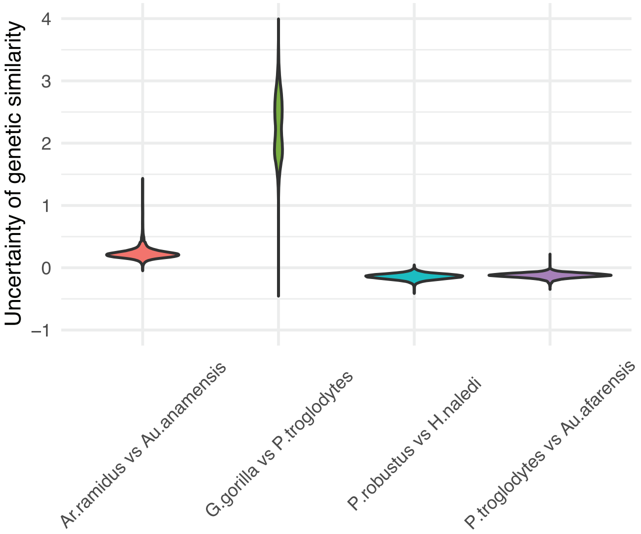

FIG. 8. Uncertainty of genetic similarities between *Ardipithecus ramidus* and *Australopithecus anamensis*, *Gorilla gorilla* and *Pan troglodytes*, *Parantropus robustus* and *Homo naledi*, and *P. troglodytes* and *Australopithecus afarensis*.

inferred trees is modeled jointly with LMM regression parameters. As a proof-of-concept, we provide estimations of the genetic component for human and ancient hominin heights. To our knowledge, this is the first work to describe the integrals and expectations involved. The assumptions include haploidy (and no recombination), neutral models of evolution, and the infinite sites model. These assumptions can be challenged by data in which samples undergo nontree-like evolution, incomplete lineage sorting, hybridization, or homoplasy. However, these assumptions are appropriate for evolutionary timescales with limited lateral transfer (e.g., in hominins, where lateral transfer such as from Neanderthal to humans accounts for only a small proportion of the modern human genome; Sánchez-Quinto et al., 2012), or for trees derived from highly clonal bacteria such as *tuberculosis*.

Our numerical experiments demonstrate consistency between the empirically calculated genetic similarity matrix and our proposed algorithm. We also apply our method to compute genetic blue for eight hominin species, based on phylogenetic trees inferred from geological dates. Although our methods are based on the expected value of genetic similarity (i.e., molecular phylogeny), we assume that geological dates are proportional to molecular phylogeny. Our method can be used to provide genetic similarity matrices for heritability analyses (Dahl et al., 2016) in cases for which genotypes are not available. We also conduct a heritability analysis on the average heights of genetic males for hominin species. Even though the evolution of hominin species involves recombination and some lateral transfer, which is not accounted in our analysis, the main effects in molecular phylogeny at evolutionary timescales involve mutation at the species level.

In cases where genotypes are available, our work can be applied if phenotypes or geography suggests a mismatch between phylogeny and genotype. This can happen in application of LMMs for multivariate genome-wide association studies with low-quality or low-coverage genotypes. This can also happen when genetic variation is low, but the examined phenotype is still under selection, causing homoplasy, which could bias estimates of a similarity matrix based on genotypes. When multiple samples of the phylogeny are available (e.g., after running MCMC inference for the phylogeny), our algorithm can be used to compute the expected genetic similarity matrix for each posterior sample. The resulting matrices represent the uncertainty of genetic similarities among species, and could be combined with LMMs to better identify population stratification and correct for spurious association.

Our current approach is limited to the computation of genetic similarities among species, or samples without recombination. One future direction for this work would be to consider genealogies with recombination events. The ancestral recombination graph (ARG) describes the coalescence and recombination events among individuals (Rasmussen et al., 2014). The ARG is composed of a set of coalescent trees separated by break points. To compute the expected genetic similarity matrix for samples given an ARG, we could first compute the expected genetic similarity matrices for each of the coalescent trees in the ARG, and then compute weighted average for those expected genetic similarity matrices. The weights would be proportional to the number of loci between each consecutive pair of break points. This would constitute a straightforward extension of our method. Another future direction would be to incorporate more advanced evolutionary, mutation, or selection models (relaxing the infinite sites and neutral assumptions). Finally, there are strong similarities between genetic and linguistic evolution (Cavalli-Sforza, 1997; Colonna et al., 2010). These two subjects both involve evolution and variation in a similar manner, and phylogenetic methods have been applied to construct trees for language data (Atkinson and Gray, 2005; Jordan, 2007). Our work could be extended to compute similarities for languages using linguistic trees.

ACKNOWLEDGMENT

We would like to thank Alexandre Bouchard Côté for helpful discussion.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

FUNDING INFORMATION

This research was supported by NSERC grant numbers RGPIN/05484-2019 and DGECR/00118-2019.

SUPPLEMENTARY MATERIAL

Supplementary Data
Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Table S1
Supplementary Table S2
Supplementary Appendix SA1
Supplementary Appendix SA2
Supplementary Appendix SA3
Supplementary Appendix SA4
Supplementary Appendix SA5

REFERENCES

Abney, M. 2009. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics* 25, 1561–1563.

Atkinson, Q.D., and Gray, R.D. 2005. Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Syst. Biol.* 54, 513–526.

Berger, L.R., Hawks, J., de Ruiter, D.J., et al. 2015. *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. *eLife* 4, e09560.

Berger, L.R., Hawks, J., Dirks, P.H., et al. 2017. *Homo naledi* and Pleistocene hominin evolution in subequatorial Africa. *eLife* 6, e24234.

Bouckaert, R., Heled, J., Kühnert, D., et al. 2014. BEAST 2: A Software Platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10, e1003537.

Boyce, A. 1983. Computation of inbreeding and kinship coefficients on extended pedigrees. *J. Heredity* 74, 400–404.

Carretero, J.M., Lorenzo, C., and Arsuaga, J.L. 1999. Axial and appendicular skeleton of *Homo* antecessor. *J. Hum. Evol.* 37, 459–499.

Cavalli-Sforza, L.L. 1997. Genes, peoples, and languages. *Proc. Natl Acad. Sci. U. S. A.* 94, 7719–7724.

Chen, G.K., and Witte, J.S. 2007. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am. J. Hum. Genet.* 81, 397–404.

Colonna, V., Boattini, A., Guardiano, C., et al. 2010. Long-range comparison between genes and languages based on syntactic distances. *Hum. Heredity* 70, 245–254.

Dahl, A., Iotchkova, V., Baud, A., et al. 2016. A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* 48, 466–472.

Dembo, M., Radovčić, D., Garvin, H.M., et al. 2016. The evolutionary relationships and age of *Homo naledi*: An assessment using dated Bayesian phylogenetic methods. *J. Hum. Evol.* 97, 17–26.

Drummond, A., and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214.

González-Recio, O., Gianola, D., Long, N., et al. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. *Genetics* 178, 2305–2313.

Helmuth, H. 1998. Body height, body mass and surface area of the Neandertals. *Z. Morphol. Anthropol.* 82:1–12.

Housworth, E.A., Martins, E.P., and Lynch, M. 2004. The phylogenetic mixed model. *Am. Nat.* 163, 84–96.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18, 337–338.

Ickstadt, K., Selinski, S., and Müller, T. 2005. Cluster analysis: A comparison of different similarity measures for SNP data. Technical Report, University of Dortmund.

Jordan, F. 2007. A comparative phylogenetic approach to Austronesian cultural evolution [unpublished doctoral dissertation]. University College London, London.

Kang, H.M., Sul, J.H., Service, S.K., et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 893–903.

Kirkpatrick, B., Ge, S., and Wang, L. 2019. Efficient computation of the kinship coefficients. *Bioinformatics* 35, 1002–1008.

Lees, J.A., Galardini, M., Bentley, S.D., et al. 2018. pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* 34, 4310–4312.

Lippert, C., Listgarten, J., Liu, Y., et al. 2011. FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835.

Listgarten, J., Lippert, C., and Heckerman, D. 2013. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet.* 45, 470–471.

Listgarten, J., Lippert, C., Kadie, C.M., et al. 2012. Improved linear mixed models for genome-wide association studies. *Nat. Methods* 9, 525–526.

Lordkipanidze, D., Jashashvili, T., Vekua, A., et al. 2007. Postcranial evidence from early Homo from Dmanisi, Georgia. *Nature* 449, 305–310.

Ma, J., Ratan, A., Raney, B.J., et al. 2008. The infinite sites model of genome evolution. *Proc. Natl Acad. Sci. U. S. A.* 105, 14254–14261.

Malo, N., Libiger, O., and Schork, N.J. 2008. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J. Hum. Genet.* 82, 375–385.

Maruyama, T., and Yasuda, N. 1970. Use of graph theory in computation of inbreeding and kinship coefficients. *Biometrics* 209–219.

McHenry, H.M., and Coffing, K. 2000. *Australopithecus* to *homo*: Transformations in body and mind. *Annu. Rev. Anthropol.* 29, 125–146.

Murray, K.D., Webers, C., Ong, C.S., et al. 2017. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS Comput. Biol.* 13, e1005727.

Patterson, N., Price, A.L., and Reich, D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2, e190.

Rasmussen, M.D., Hubisz, M.J., Gronau, I., et al. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* 10, e1004342.

Ronquist, F., Teslenko, M., Van Der Mark, P., et al. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542.

Sánchez-Quinto, F., Botigué, L.R., Civit, S., et al. 2012. North African populations carry the signature of admixture with Neandertals. *PLoS One* 7, e47765.

Scally, A., Dutheil, J.Y., Hillier, L.W., et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483, 169–175.

Thompson, E.A. 2013. Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics* 194, 301–326.

Thornton, T., Tang, H., Hoffmann, T.J., et al. 2012. Estimating kinship in admixed populations. *Am. J. Hum. Genet.* 91, 122–138.

Tzeng, J.-Y., and Zhang, D. 2007. Haplotype-based association analysis via variance-components score test. *Am. J. Hum. Genet.* 81, 927–938.

Wang, X., Guo, X., He, M., et al. 2011. Statistical inference in mixed models and analysis of twin and family data. *Biometrics* 67, 987–995.

Whittemore, A.S., and Halpern, J. 1994. Probability of gene identity by descent: Computation and applications. *Biometrics* 50, 109–117.

Wiens, J.J. 2004. The role of morphological data in phylogeny reconstruction. *Syst. Biol.* 53, 653–661.

Wood, B., and Boyle, E. 2017. Hominins: Context, origins, and taxic diversity. *In On Human Nature*. Academic Press.

Yang, J., Bakshi, A., Zhu, Z., et al. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120.

Yang, J., Lee, S.H., Goddard, M.E., et al. 2011. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82.

Address correspondence to:
Dr. Lloyd T. Elliott
Department of Statistics and Actuarial Science
Simon Fraser University
8888 University Drive
Burnaby V5A 1S6
Canada

E-mail: lloyd.elliott@sfu.ca