Estimating Genetic Similarity Matrices using Phylogenies

Shijia Wang[1,2] Shufei Ge[2] Caroline Coljin[3]
Liangliang Wang[2] Lloyd T Elliott[2,*]

[1]School of Statistics and Data Science, LPMC and KLMDASR,
Nankai University, China

[2]Department of Statistics and Actuarial Science
Simon Fraser University, Canada

[3]Department of Mathematics
Simon Fraser University, Canada

Summer 2020

Abstract

Genetic similarity is a measure for the genetic relatedness among individuals. The standard method for computing similarity matrices involves the inner product of observed genetic variant matrices. Such an approach is inaccurate or impossible if genotypes are not available, or not densely sampled, or of poor quality (for example, 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 genetic sequences. We show that the genetic similarity matrices computed from trees are consistent with those computed from genotypes. Quantitative analysis on genetic traits and analysis of heritability and co-heritability can be conducted directly using genetic similarity matrices and so in the absence of genotype data, and the presence of phylogenetic trees derived from morphological data or geological dates, such analyses can be undertaken using our methods. We use simulation studies to demonstrate the advantages of our method, and we provide an application to ancient hominin data.

*Address correspondence to: Dr. Lloyd T. Elliott, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Dr., Burnaby, B.C. Canada V5A 1S6 (`lloyd.elliott@sfu.ca`).

1 Introduction

The computation of genetic similarities among samples (or taxa, or subjects, or individuals) is a key step in the analysis of quantitive genetic traits and the heritability of traits through variance component approaches (Chen and Witte, 2007; Malo et al., 2008; Tzeng and Zhang, 2007). For example, linear mixed models (LMMs) are popular methods for genome-wide association studies, and matrix-variate normal methods are popular for heritability and co-heritability analysis. 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 sequences, or 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 a mutation is observed in the ancestry of the subjects. We consider single nucleotide polymorphisms (SNPs), locations at which a single DNA basepair 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 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 then defined as:

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

Here G_{im} is the genotype of leaf i at marker m ($G_{im} \in \{0, 1\}$), and μ_m, σ_m^2 are the empirical mean and variance of the SNPs at marker k, and M is the total number of loci genotyped (Patterson et al., 2006). We will 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 the permutation of the loci does not affect the empirical genetic similarity.

The computation of genetic similarity among samples using equation (1) requires genotyped sequences, and several challenges may arise in the computation. Firstly, the genetic sequences may not be readily available. This may be the case when extinct species are examined. Secondly, 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 the genetic similarity matrix using a phylogeny. Our proposed approach does not require that the sequences for individuals be genotyped (or measured), instead we assume

that a phylogeny is given. The relatedness among samples is computed by integrating over all mutations occuring 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. Genetic recombination among individuals is not considered in this approach as we mainly consider genetic similarities at the species level (note that highly clonal bacteria such as *tuberculosis* may also be considered). 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).

In our simulations, the expected genetic similarity matrix is invariant to the number of samples and 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 (González-Recio et al., 2008; Ickstadt et al., 2005). The MDS approach also operates on phylogenies, but it 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 by 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 (Dahl et al., 2016; Wang et al., 2011). 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 a linear mixed model (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 which we assume to approximate the phylogeny).

1.1 Related work

Genetic similarity is a basic aspect of many approaches in genetic analysis. Identity-by-descent (Whittemore and Halpern, 1994) is a popular measure for genetic similarity, based on identification of stretches of the genetic sequence with identical ancestral origin. Also, if genetic sequences are observed, a variety of measures have been used to compute genetic similarities using categorical data clustering (González-Recio et al., 2008; Ickstadt et al., 2005) by creating dissimilarity scores for genotypes and then converting the dissimilarity scores to similarity scores. In contrast, less work has been done in assessing the statistical relationship between phylogenetic trees and genetic similarity matrices. Multi-dimensional scaling (MDS) of a distance matrix implied by the phylogeny has been used to control linear mixed models (Lees et al., 2018). This MDS approach is similar to our method, in that it 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 (the Bayesian Evolutionary Analysis Sampling Trees software; Drummond and Rambaut 2007).

2 Methods

Let T be an unrooted binary tree with leaves $i \in \{1, \ldots, N\}$. The phylogeny T represents relationships among N taxa through a tree topology τ and a set of branch lengths $\boldsymbol{e} = (e_1, e_2, \ldots, e_{2N-3})$. 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 distance 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 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_{ij}^G denotes the genetic similarity between sample i and sample j. Our goal is to compute the expected value of K_{ij}^G given a tree T. By additivity of expectations, from equation (1) we arrive at the following expectation through integration over all mutations:

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

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 $\{0, 1\}$.

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:

$$K_{ij}^T = \sum_{e_d} \frac{|e_d|}{|T|} \frac{(G_{ie_d} - \mu_{e_d})(G_{je_d} - \mu_{e_d})}{\sigma_{e_d}^2}. \tag{3}$$

The values G_{ie_d}, μ_{e_d}, $\sigma_{e_d}^2$ are found by considering a mutation on each edge

e_d. Under our assumptions, μ_{e_d}, $\sigma^2_{e_d}$ are summary statistics for G, and K 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 two connected components found in the graph formed by removing edge e_d from the unrooted tree T. Figure 1 shows an example of this operation. Without loss of generality, 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.

4. $\sigma^2_{e_d}$ is the induced variance of the Bernoulli distribution implied by the allele: $\mu_{e_d}(1 - \mu_{e_d})$.

Note that if c_0 and c_1 are the same size, then $\mu_{e_d} = 1 - \mu_{e_d} = 0.5$, and so $\frac{(G_{ie_d}-\mu_{e_d})(G_{je_d}-\mu_{e_d})}{\sigma^2_{e_d}}$ 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 (note that methods for computing genetic similarity from inner products have the same

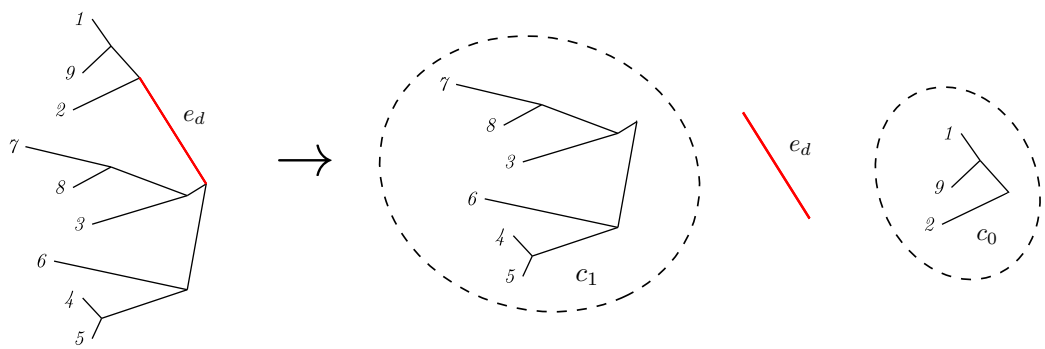

Figure 1: Segregation by an unobserved genetic variant on edge e_d. Computation of the expected mean μ_{e_d}, and variance $\sigma^2_{e_d}$, for an unobserved allele involves integration over edge e_d, and examination of the connected components of the tree formed by removing edge e_d.

asymptotic complexity). We provide an open source implementation for this method[1].

Algorithm 1 From phylogeny to expected genetic similarity matrix

Inputs: A phylogenetic tree T with a tree topology τ 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 μ_{e_d} and $\sigma^2_{e_d}$
    for $i \in \{1, 2, \ldots, N\}$ **do**
        for $j \in \{i, \ldots, N\}$ **do**
            Set $K^T_{ij} \leftarrow K^T_{ij} + \frac{|e_d|}{|T|} \frac{(G_{ie_d} - \mu_{e_d})(G_{je_d} - \mu_{e_d})}{\sigma^2_{e_d}}$.
            Set $K^T_{ji} \leftarrow K^T_{ij}$.
return K^T.

[1] `https://github.com/shijiaw/Expected-Genetic-Similarity-Matrices`

3 Experiments

In our numerical experiments, we use the *ms* software (Hudson, 2002) to simulate binary trees and genetic variation (*ms* is software for generating samples under a neutral model). We assume an infinite sites model with a neutral evolution model and constant effective population size N. The branch lengths are in units of $4N$ generations.

3.1 Simulation 1

In this simulation study, we numerically demonstrate the consistency of our algorithm. We simulate datasets 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 $\theta = 4N\mu$ (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 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. 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 the number of loci (M) on the genotype from 20 to 100, 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.

3.2 Simulation 2

In our second simulation study, we compare our tree based genetic similarity matrix approach to other approaches that measure genetic similarity using phylogenies. The first approach is a Gaussian distance similarity matrix. 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 $K^S_{ij} = \exp(-\lambda d_{ij}/|T|)$. Here λ is a fixed bandwidth. We refer to Appendix A of the Supplementary Material for more details of this approach. This approach is similar to Ickstadt et

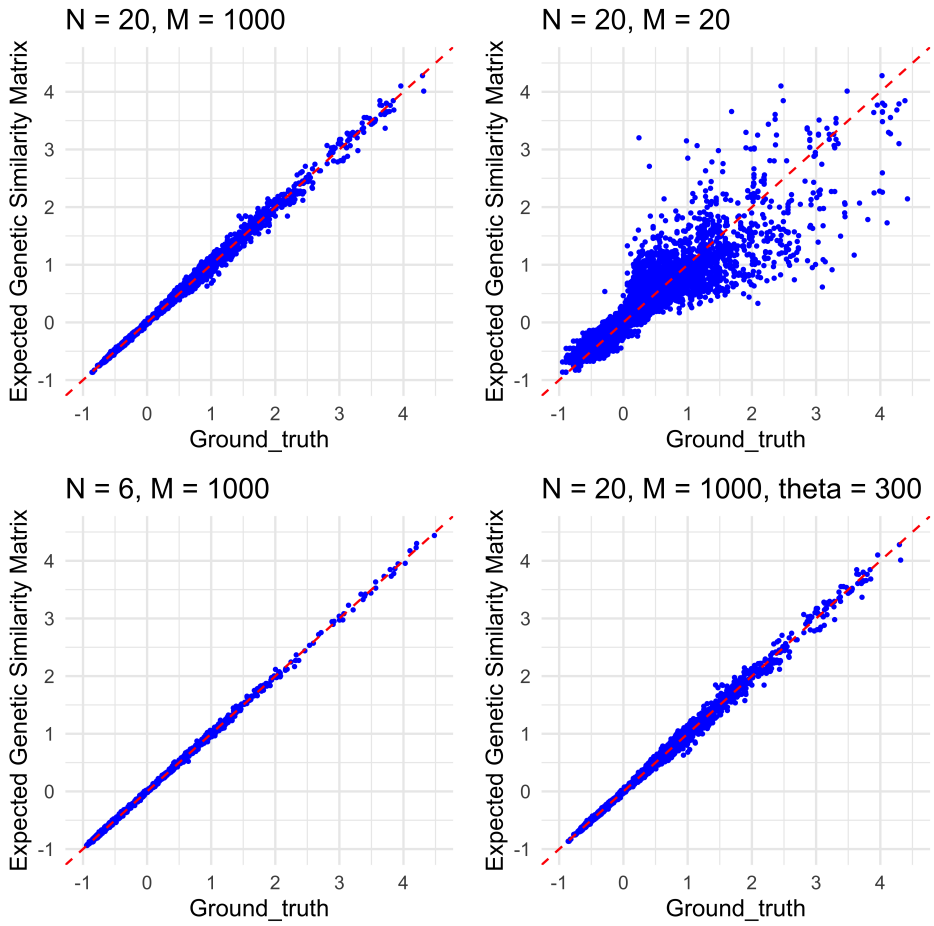

Figure 2: Comparison of simulated genetic similarity matrices from trees and genotypes. Scatter plots are provided for entrywise differences between the genetic similarity matrices produced by our method and the ground truth (from equation 1, applied to the simulated genotypes). The *top left* and *top 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 and C (respectively).

al. 2005 and González-Recio et al. 2008, but with distances computed through the phylogeny. The second approach we compare to is the multidimensional scaling (MDS) approach implemented in the software package *pyseer* (Lees et al., 2018). In the MDS approach, the similarity between each pair is calculated based on the shared branch length between the pair's most recent common ancestor and the root. Multidimensional scaling is then performed on the similarity matrix, and the resulting MDS components can be used in linear-mixed models. We denote the genetic similarity matrix computed via *pyseer* by K_{ij}^{MDS} (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_{ij}^{T} (our method, based on the tree), K_{ij}^{S}, K_{ij}^{G} (the inner product from the sampled genotypes) and K_{ij}^{MDS}. Figure 3 displays the comparison of genetic similarity matrices using trees. The violins denote the difference between $K_{ij}^{G} - K_{ij}^{T}$ (red), $K_{ij}^{G} - K_{ij}^{S}$ (green) and $K_{ij}^{G} - K_{ij}^{MDS}$ (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 multidimensional scaling similarity matrix.

Note that the simulation conditions of the comparison between our method and the empirical genetic similarity matrix are the same as upper left panel of Figure 2. For the $K_{ij}^{G} - K_{ij}^{S}$ condition, we note that the median difference is far from zero (as the range of the kernel method is positive). However,

K_{ij}^G and K_{ij}^S are still correlated ($\rho = 0.27$, $p < 0.001$). For the $K_{ij}^G - K_{ij}^{MDS}$ 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

In this experiment, we apply our method to compute the genetic similarities between 8 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 or anthropological observations are used to assess the evolutionary relationships among extinct species (Dembo et al., 2016). In this way, a tree approximating the phylogenetic tree of life can still be constructed. Our method allows us to compute genetic similarity matrices which could then be used in analysis of heritability or coheritability (as is done in Dahl et al. 2016), in absence of any genetic sequences.

We use our proposed approach to compute the expected genetic similarity matrix for *H. sapiens* and 7 extinct cousins or ancestors of humans (*H. habilis, H. rudolfensis, Georgian H. erectus, African H. erectus s.s., Asian H. erectus s.s., H. antecessor, H. neanderthalensis*). The phylogenetic tree for these species are provided by Dembo et al. 2016 using geological dates. While Dembo et al. 2016 reconstruct the phylogenetic tree of 24 species, we examine 7

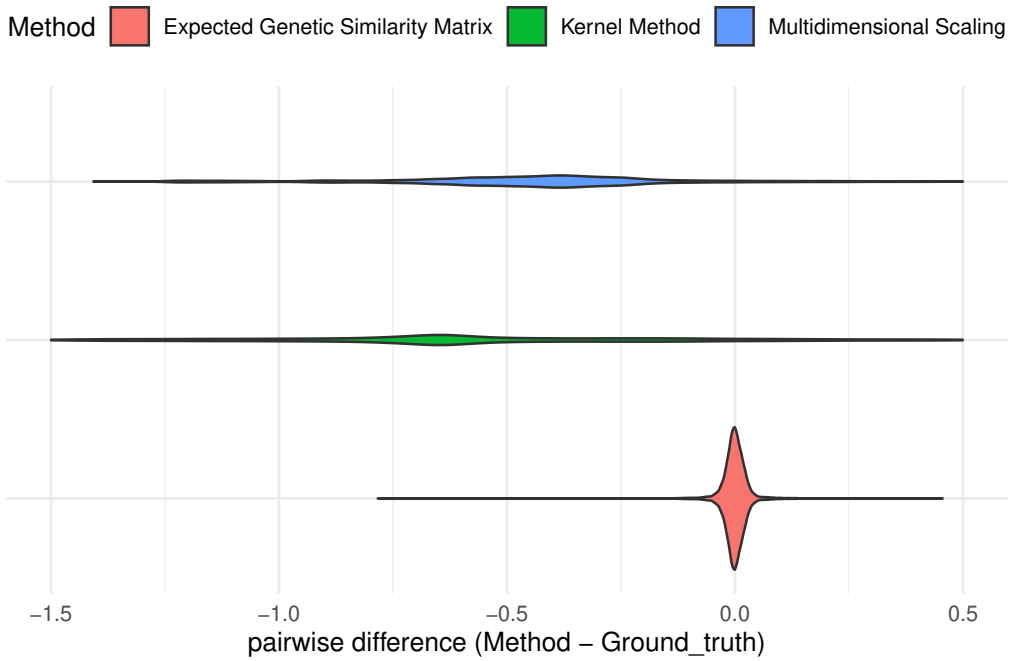

Figure 3: Comparison of expected genetic similarity matrix approaches: $K_{ij}^G - K_{ij}^T$ (our method, *bottom*), $K_{ij}^G - K_{ij}^S$ *(middle)* and $K_{ij}^G - K_{ij}^{MDS}$ *(top)*. The entries of the expected genetic similarity (the red violin, thin) are close to the empirical genetic similarity 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).

hominin species closely related to humans, as well as humans themselves (yielding 8 species). Figure 4 shows the heatmap for the expected genetic similarity matrix for the 8 hominins considered in this article. The genetic similarities between *H. rudolfensis* and *Georgian H. erectus*, *H. rudolfensis* and *H. habilis*, *Georgian H. erectus* and *H. habilis* are larger than the rest.

We also provide the specific values of the heatmap in Figure 4 of the similarity in Table 1 of Appendix B of the Supplementary Material. This table could be used in conjunction with models such as Dahl et al. 2016 to conduct heritability or co-heritability analysis on traits of the species considered.

To that end, we consider a heritability analysis on the average heights of genetic males in the 8 hominin species. We gather these average heights from the following references: Carretero et al. (1999); Helmuth (1998); Lordkipanidze et al. (2007); Man and Jawa (2016); McHenry and Coffing (2000). We estimate variance components from a linear mixed model, which is standard model for methods such as genome-wide complex trait analysis (GCTA; Yang et al. 2011). The form of this model is as follows:

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

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

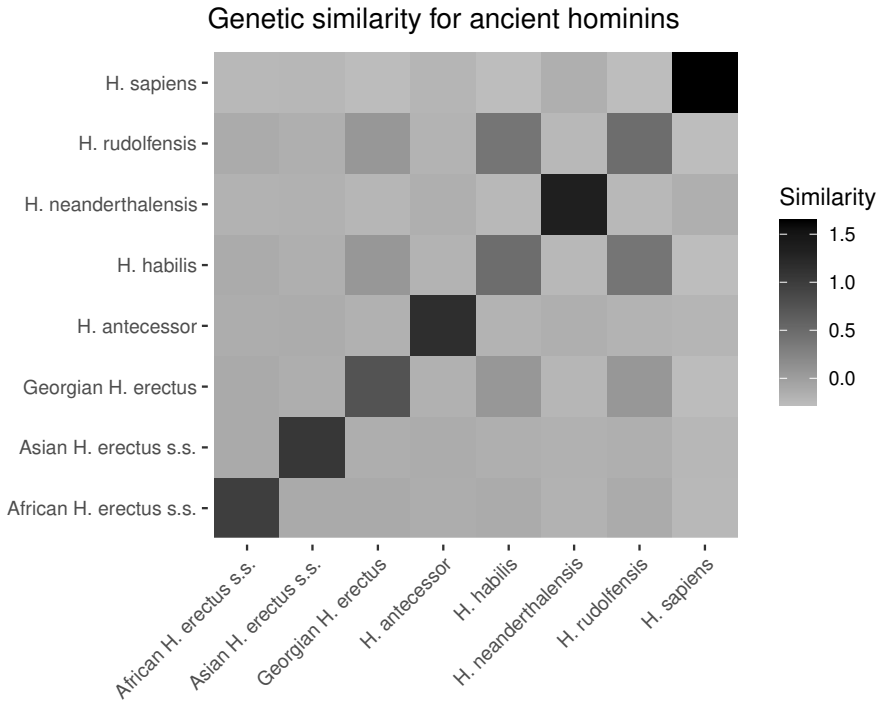

Figure 4: Genetic similarity matrix for 8 hominin species as a heat map, computed using geological dates and Algorithm 1.

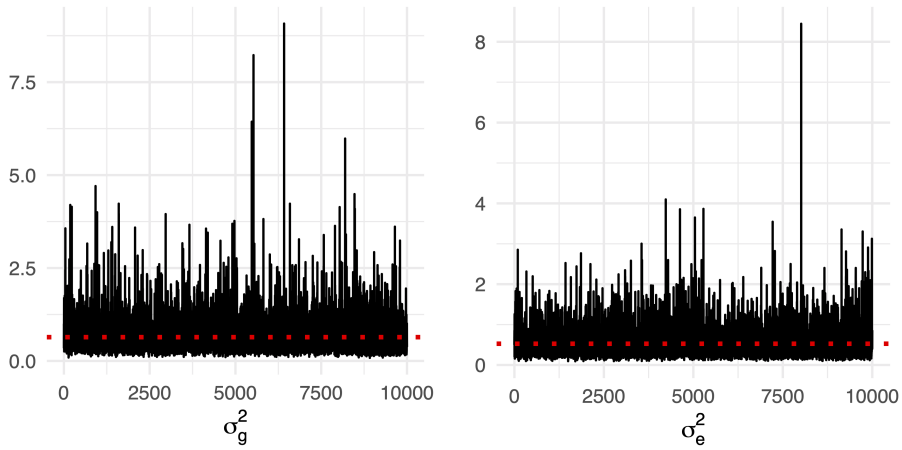

Figure 5: MCMC trace plots of σ_g^2 and σ_e^2. The red dashed lines indicate posterior means for σ_g^2 and σ_e^2 with initial $5,000$ iterations as burn in.

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 linear mixed model with K^T (our expected genetic similarity matrix) computed earlier in this Section. We conducted inference using MCMC (Gibbs Sampling) with $10,000$ iterations, of which $5,000$ are discarded for burnin. Figure 5 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 initial $5,000$ iterations discarded. Figure 6 shows the histograms for posterior samples of σ_g^2, σ_e^2 and h^2 with initial $5,000$ iterations 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 $\widehat{\sigma}_g^2 = 0.640\ (0.185, 1.754)$

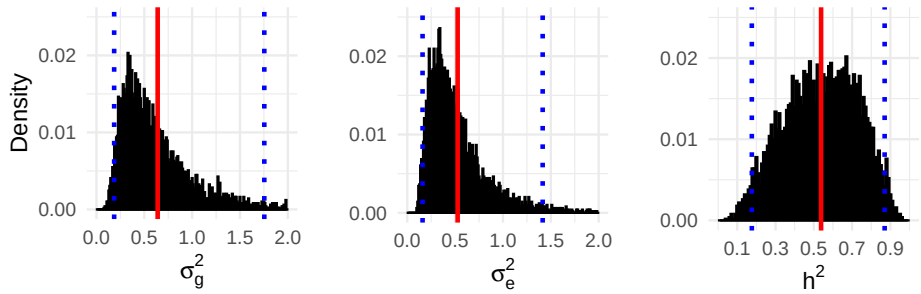

Figure 6: Histograms for posterior samples of σ_g^2, σ_e^2 and h^2. The red lines indicate posterior means for σ_g^2, σ_e^2 and h^2, the blue dash lines indicate the 95% credible intervals.

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

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 co-heritability 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 inferred trees is modelled 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 non-tree-like evolution, incomplete lineage sorting, hybridization or homoplasy.

Our numerical experiments demonstrate the consistency between the empirically calculated genetic similarity matrix, and our proposed algorithm. We also apply our method to compute the genetic similarity for 8 hominin species, based on phylogenetic trees inferred from geological time. 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.

In cases where genotypes are available, our work can be applied if phenotype or geography suggest a mismatch between phylogeny and genotype. This can happen in application of linear mixed models 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.

Our current approach is limited to the computation of genetic similarities

among species, or samples without recombination. One future direction for this work is to consider genealogies with recombination events. Another future direction is to incorporate more advanced evolutionary, mutation or selection models (relaxing the infinite sites and neutral assumptions).

Acknowledgements

We would like to thank Alexandre Bouchard Côté for helpful discussion. This research was supported by NSERC grant numbers RGPIN/05484-2019 and DGECR/00118-2019.

References

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.

Boyce, A. 1983. Computation of inbreeding and kinship coefficients on extended pedigrees. *Journal of Heredity*, *74*(6).

Carretero, J. M., Lorenzo, C., and Arsuaga, J. L. 1999. Axial and appendicular skeleton of *Homo* antecessor. *Journal of Human Evolution*, *37*(3-4).

Chen, G. K., and Witte, J. S. 2007. Enriching the analysis of genomewide association studies with hierarchical modeling. *The American Journal of Human Genetics*, *81*(2).

Dahl, A., Iotchkova, V., Baud, A., et al. 2016. A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, *4*(48).

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. *Journal of Human Evolution*, *97*.

Drummond, A., and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1).

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*(4).

Helmuth, H. 1998. Body height, body mass and surface area of the Neandertals. *Zeitschrift für Morphologie und Anthropologie*, *82*(1).

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, *18*(2).

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

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. *Nature Genetics*, *42*(4).

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, *61*(4).

Kirkpatrick, B., Ge, S., and Wang, L. 2019. Efficient computation of the kinship coefficients. *Bioinformatics*, *35*(6).

Lees, J. A., Galardini, M., Bentley, S. D., et al. 2018. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310–4312.

Lippert, C., Listgarten, J., Liu, Y., et al. 2011. FaST linear mixed models for

genome-wide association studies. *Nature Methods*, *8*(10).

Listgarten, J., Lippert, C., and Heckerman, D. 2013. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics*, *45*(5).

Lordkipanidze, D., Jashashvili, T., Vekua, A., et al. 2007. Postcranial evidence from early Homo from Dmanisi, Georgia. *Nature*, *449*(7160), 305–310.

Ma, J., Ratan, A., Raney, B. J., et al. 2008. The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, *105*(38).

Malo, N., Libiger, O., and Schork, N. J. 2008. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, *82*(2).

Man, J., and Jawa, M. 2016. Java Man. *Homo*, *54*.

McHenry, H. M., and Coffing, K. 2000. *Australopithecus* to *homo*: transformations in body and mind. *Annual review of Anthropology*, *29*(1).

Patterson, N., Price, A. L., and Reich, D. 2006. Population structure and eigenanalysis. *PLOS genetics*, *2*(12).

Thompson, E. A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–326.

Tzeng, J.-Y., and Zhang, D. 2007. Haplotype-based association analysis via variance-components score test. *The American Journal of Human Genetics*, *81*(5).

Wang, X., Guo, X., He, M., and Zhang, H. 2011. Statistical inference in

mixed models and analysis of twin and family data. *Biometrics*, *67*(3).

Whittemore, A. S., and Halpern, J. 1994. Probability of gene identity by descent: Computation and applications. *Biometrics*, *50*(1).

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. *Nature Genetics*, *47*(10).

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1).