Modelling complex traits with ancestral recombination graphs

Hanbin Lee*[1], Nathaniel S. Pope[2], Jerome Kelleher[3], Gregor Gorjanc[4], and Peter L. Ralph[2]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, USA
[2]Institute of Ecology & Evolution, University of Oregon, Eugene, OR, 97405, USA
[3]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK
[4]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh EH25 9RG, UK

July 14, 2025

Abstract

The ancestral recombination graph (ARG) is a powerful tool for storing and analyzing large genomic datasets, as demonstrated by the ecosystem of software tools taking advantage of the succinct tree sequence (`tskit`) format to store and analyse ARGs. Tree sequences have become standard in evolutionary genetics, and researchers are now exploring new applications to address the analytical challenges arising from the rapid growth of data. To this end, we propose ARG-LMM, a generative model which describes how complex traits emerge from ARGs, and `tslmm`, a tree sequence-based algorithm which fits ARG-LMM to simulated and inferred genealogies. The distribution of a trait is deduced from evolutionary theory, avoiding arbitrary probabilistic and statistical assumptions. The model subsumes existing quantitative genetics models by considering all past DNA coalescence and recombination events encoded in an ARG. This formulation provides a framework to study the effect of evolutionary forces on genomes and corresponding phenotypes. `tslmm` can work with ARGs encoding whole genomes of tens of thousands of individuals through efficient graph traversal algorithms with an $\mathcal{O}(N)$ time complexity, up to a logarithmic factor, a step change over most existing algorithms that require at least $\mathcal{O}(N^2)$ operations. In summary, this work presents a new quantitative genetic model (ARG-LMM) based on evolutionary theory and a new software (`tslmm`) for scalable analysis of complex traits with large genomic datasets.

1 Introduction

The ancestral recombination graph (ARG) encodes the genealogical relationship of sampled genomes in the presence of genetic drift and recombination [Hudson, 1990, Griffiths and Marjoram, 1997, Wong et al., 2023]. Recent methods are capable of reconstructing ARGs from hundreds of thousands of genomes [Rasmussen et al., 2014, Speidel et al., 2019, Ralph, 2019, Zhang et al., 2023, Deng et al., 2024], producing a rich evolutionary history from which it is possible to identify signatures of natural selection as well as demographic

*E-mail: `hblee@umich.edu`

events that happened in the past [Lewanski et al., 2024, Brandt et al., 2024]. Furthermore, the graph structure of an ARG serves as a powerful computational tool that can produce evolutionarily important statistics orders of magnitude faster than conventional approaches [Ralph et al., 2020]. The field is now seeking new applications of this powerful tool to analyze genomes outside the field of evolutionary genetics.

Using inferred genealogies to analyse complex traits is an active area of research. Zhang et al. [2023] developed ARG-Needle, a scalable ARG inference algorithm, and from the inferred trees constructed an expected genetic relatedness matrix (eGRM), which is the expectation of the single nucleotide polymorphism (SNP)-based genetic relatedness matrix (GRM) conditioned on the sample's ARG [Speed and Balding, 2015]. They used the eGRM within the linear mixed model (LMM) framework to estimate heritability and marginally test the association of variants with a trait. They also proposed testing computationally simulated variants on the trees, and found that this procedure can complement conventional imputation methods to detect genetic signals of unobserved variants [Harris, 2023].

Link et al. [2023] proposed testing GRMs restricted to particular genomic locations to detect regions associated with a trait, using Fan et al. [2022]'s method to construct local eGRMs. While Fan et al. [2022] averaged over all trees of the ARG to construct the eGRM, Link et al. [2023] used a subset of trees in the region of interest to construct localised versions. They identified disease-associated regions by fitting a linear mixed model that quantified the proportion of trait variance explained by the local eGRM. This procedure aggregates all genetic signals within a region, thereby improving statistical power when multiple genetic variants are located there. This allows the method to detect signals driven by unobserved variants, even if an imputation panel is not readily available.

These two studies employed ARGs to compute genetic relatedness matrices, but did not take full advantage of the computational potential of inferred genealogies. Both of them used existing software (e.g. GCTA) and substituted the conventional GRMs with their ARG-based version to fit relevant variance-component parameters for gene mapping. Zhu et al. [2024] pioneered the use of tree-based algorithms to speed up such tasks. They devised a novel randomized Haseman-Elston (HE) regression algorithm called ARG-RHE by combining randomized linear algebra and a tree-based matrix-vector multiplication algorithm. Similar to tree-based algorithms for computing evolutionary statistics [Ralph et al., 2020], their matrix-vector multiplication algorithm recycles information shared by adjacent trees to minimize redundant computations. Compared to traditional methods for fitting variance components that require quadratically growing operations as a function of sample size, the cost of their algorithm grows linearly, allowing biobank-scale analysis.

These recent methods which adopt ARGs to analyze complex traits all rely on the genomic random effects model [Kang et al., 2010, Yang et al., 2010, 2011]. Random effects models are widely adopted in quantitative genetics literature, and the notion of genetic relatedness ultimately stems from the random effects in complex trait analysis. From this perspective, ARG-based methods merely replace the conventional GRM with a tree-based substitute without deviating from or justifying the random effects model. However, the relationship between random effects models and the ARG, along with the tree-based methods, remains largely unexplored. Below, we derive a tree-based generative model of complex traits called ARG-LMM, to show that genetic random effects models naturally emerge from the mutational process running on a fixed genealogy. This illuminates the profound connections between the two concepts.

The key conceptual shift or our work is that SNP random effects should be viewed not as the effect of the SNP itself, but as the aggregate effect of potential mutations that could have occurred on the same edge as the SNP. This allows us to link the distribution of the random effects to the topology of the ARG, characterizing the joint distribution of genetic random effects. Common assumptions of random effects models, such as the independence across SNP effects, immediately follow from this formulation. We further explore important topics in quantitative genetics, such as heritability, genomic prediction, and pseudoreplication, through the lens of ARG-LMM.

We implement an efficient algorithm to fit the ARG-LMM model to simulated and inferred tree sequences in the `tslmm` Python package. Similar to the work of Zhu et al. [2024], we utilize an efficient matrix-vector multiplication algorithm based on tree sequences [Lehmann et al., 2025]. Our new algorithm has a number of improvements over previous approaches. It is the first method that optimizes the restricted maximum likelihood (REML) objective using a tree-based algorithm to estimate variance components, a statistically more efficient objective than the Haseman-Elston regression used in ARG-RHE. Additionally, it exactly multiplies an eGRM and a vector without randomly sampling mutations on the tree, so does not require simulating mutations. This removes randomness and computational burden, producing accurate matrix-vector products more efficiently. Similar to ARG-RHE, our algorithm also scales linearly as a function of sample size up to a logarithmic factor. Finally, `tslmm` can compute the best linear unbiased predictor (BLUP) for predicting traits of unobserved individuals in the tree sequence. The software is available at `https://github.com/hanbin973/tslmm/`.

2 Materials and Methods

2.1 Ancestral recombination graph linear mixed model

We show that complex traits follow a linear mixed model conditioned on the underlying genealogy. We define an ARG $\mathcal{T} = (\mathcal{N}, \mathcal{E})$ as a collection of nodes $\mathcal{N}$ and edges $\mathcal{E}$, where each edge spans a left/right genomic interval [Wong et al., 2023]. For conceptual clarity, we assume that each edge has a unique subtopology along its span (although not true in the case of tree sequences in general, we can enforce this by appropriate splitting of edges as in Salehi Nowbandegani et al. [2023]). Nevertheless, this splitting is merely conceptual and never takes place in the algorithms we'll demonstrate shortly, as it significantly degrades the storage efficiency of the tree sequence data structure. A node u has a time t_u; an edge e has span s_e (in base pair units) with child and parent nodes $c(e)$ and $p(e)$. The time length of an edge is therefore $l_e = t_{p(e)} - t_{c(e)}$. All expectations and variances are conditioned on $\mathcal{T}$, which removes any variation in the data due to recombination and drift. Only the mutational process is left random in this regime. This setting was first explored in Ralph [2019] and expanded in Ralph et al. [2020].

We consider complex traits which follow a linear model where $\mathbf{G} \in \mathbb{R}^{N \times P}$ is the $N \times P$ genotype matrix that contains all P positions in the genome.

$$\mathbf{y} = \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon} \tag{1}$$

$\boldsymbol{\beta} \in \mathbb{R}^P$ is a fixed vector defining a fixed effect size for each mutation at a given position. Those who are keen to conventional random effects assumptions on $\boldsymbol{\beta}$ may view our formulation as conditioning on $\boldsymbol{\beta}$, i.e. we are being agnostic about where and how $\boldsymbol{\beta}$ come from. One cannot estimate $\boldsymbol{\beta}$'s entries separately because some variants appear on the same edge. Columns of $\mathbf{G}$ corresponding to these variants will be identical, leading to perfect collinearity (or LD), making them statistically indistinguishable. We will group these indiscernible variants according to their distribution. After conditioning on $\mathcal{T}$, $\mathbf{G}$ is distributed solely according to the mutational process. Mutations occur at edge e following a Poisson process with mean proportional to $l_e s_e$ [Ralph, 2019]. We let the per unit time per base pair mutation rate u_{ep} to vary across edges and positions (p is the position on the genome). If we let $\mathbf{1}_{ep}$ be an indicator variable that is 1 if a mutation has arrived on edge e at position p and 0 otherwise, $\mathbb{E}\left[\mathbf{1}_{ep}\right] = l_e u_{ep}$.

Recurrent mutations are allowed, although mutations cannot occur below an existing mutation. This is a

slight relaxation of the infinite-sites model that allows multiple recurrent mutations in unrelated lineages but no back mutation. In tree sequence terminology, no mutation is a parent of another. Define $\mathbf{Z} \in \mathbb{R}^{N \times E}$ to be the number of chromosomes of individual n that is a descendant of edge e. $\mathbf{Z}_{ne}$ ranges from 0 to 2 in the case of diploids. We can express $\mathbf{G}$ with $\mathbf{Z}$ by

$$\mathbf{G}_{np} = \sum_{e:p \in e} \mathbf{Z}_{ne} \mathbf{1}_{ep} \quad \Leftrightarrow \quad \mathbf{G}_p = \sum_{e:p \in e} \mathbf{Z}_e \mathbf{1}_{ep} \tag{2}$$

$p \in e$ means edge e contains the position p.

Using equation (2), we can collapse variants corresponding to perfectly correlated columns of $\mathbf{G}$ into the edge in which they reside on. This is a natural choice because any mutations on the same edge are inherited by the same set of individuals. Substituting (2) to the linear equation (1) gives

$$\begin{aligned}
\mathbf{y} &= \sum_{p=1}^{P} \mathbf{G}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon} = \sum_{p=1}^{P} \sum_{e:p \in e} \mathbf{Z}_e \boldsymbol{\beta}_p \mathbf{1}_{ep} + \boldsymbol{\varepsilon} \\
&= \sum_{e=1}^{E} \sum_{p:p \in e} \mathbf{Z}_e \boldsymbol{\beta}_p \mathbf{1}_{ep} + \boldsymbol{\varepsilon} = \sum_{e=1}^{E} \mathbf{Z}_e \left(\sum_{p:p \in e} \boldsymbol{\beta}_p \mathbf{1}_{ep} \right) + \boldsymbol{\varepsilon} \\
&= \sum_{e=1}^{E} \mathbf{Z}_e \boldsymbol{\upsilon}_e + \boldsymbol{\varepsilon} \quad \because \boldsymbol{\upsilon}_e = \sum_{p:p \in e} \boldsymbol{\beta}_p \mathbf{1}_{ep} \\
&= \mathbf{Z} \boldsymbol{\upsilon} + \boldsymbol{\varepsilon}
\end{aligned} \tag{3}$$

$\boldsymbol{\upsilon}$ is a random variable because it's a linear combination of random indicator variables ($\mathbf{1}_{ep}$) weighted by effect sizes of mutations. We further decompose $\boldsymbol{\upsilon}$ into two parts $\mathbf{u} = \boldsymbol{\upsilon} - \mathbb{E}[\boldsymbol{\upsilon}]$ and $\mathbf{f} = \mathbb{E}[\boldsymbol{\upsilon}]$ to write

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{f} + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{u}$ is a random vector while $\mathbf{f}$ is a fixed vector. The expectation in $\mathbf{f}$ is taken with respect to mutations, i.e., $\mathbb{E}[\boldsymbol{\upsilon}_e] = \sum_{p:p \in e} \boldsymbol{\beta}_p \mathbb{E}[\mathbf{1}_{ep}] = \sum_{p:p \in e} \boldsymbol{\beta}_p l_e u_{ep}$. In terms of regression, $\mathbf{Z}$ is the design matrix of the regression equation (4). $\mathbf{u}$ and $\mathbf{f}$ are random and fixed effects regression coefficients, respectively. Hence, we showed that complex traits follow a linear mixed model given the ARG $\mathcal{T}$. We call this model the ancestral recombination graph linear mixed model (ARG-LMM).

2.2 ARG-LMM recovers the key properties of genomic linear mixed models

We now characterize the distribution of ARG-LMM, and how it is related to conventional genomic LMMs (GLMMs) in the literature. The columns of the random effects design matrix of GLMM ($\mathbf{Z}_{\mathrm{GLMM}}$) are a subset of columns of ARG-LMM's random effects design matrix ($\mathbf{Z}$). Only the columns of $\mathbf{Z}$ that have a genotyped variant will appear in $\mathbf{Z}_{\mathrm{GLMM}}$. In this regard, ARG-LMM subsumes GLMMs. We further show that ARG-LMM maintains a number of key assumptions of GLMMs.

The most important assumption in GLMMs is the independence of random effects. The vector of random effects has a diagonal covariance matrix, and without this assumption, computing the random effects model becomes intractable. In ARG-LMM, this independence is not an assumption but can be deduced from the nature of the mutational process. Briefly, the small mutation rate allows us to ignore the covariance between edges, because higher-order terms of the mutation rate, such as u_{ep}^2, effectively vanish. The following derivation shows how the small mutation rate leads to approximately independent random effects. We have

to show that $\mathrm{Cov}\left(\mathbf{u}_e, \mathbf{u}_{e'}\right) = 0$ for all pairs $e = e'$. This implies independence because indicator variables are independent if and only if they are uncorrelated, similar to Gaussian random variables.

We first expand $\mathrm{Cov}\left(\mathbf{u}_e, \mathbf{u}_{e'}\right)$ and reduce it to a sum of covariances of indicators.

$$\begin{aligned}
\mathrm{Cov}\left(\mathbf{u}_e, \mathbf{u}_{e'}\right) &= \mathrm{Cov}\left(\sum_{p:p\in e} \boldsymbol{\beta}_p \mathbf{1}_{ep}, \sum_{p':p'\in e'} \boldsymbol{\beta}_{p'} \mathbf{1}_{e'p'}\right) \\
&= \sum_{p\in e, p'\in e'} \boldsymbol{\beta}_p \boldsymbol{\beta}_{p'} \mathrm{Cov}\left(\mathbf{1}_{ep}, \mathbf{1}_{e'p'}\right) \\
&= \sum_{p\in e, e'} \boldsymbol{\beta}_p^2 \mathrm{Cov}\left(\mathbf{1}_{ep}, \mathbf{1}_{e'p}\right)
\end{aligned} \tag{5}$$

The last line's simplification follows from the independence of indicator variables $\mathbf{1}_{ep}$ across different positions. When e and e' have no overlapping positions, the summation is trivially zero since mutations on different positions are independent. It's also zero if the two edges are not ancestral to each other, as mutations running on separate edges are also independent. Otherwise,

$$\begin{aligned}
\mathrm{Cov}\left(\mathbf{1}_{ep}, \mathbf{1}_{e'p}\right) &= \mathbb{E}\left[\mathbf{1}_{ep}\mathbf{1}_{e'p}\right] - \mathbb{E}\left[\mathbf{1}_{ep}\right]\mathbb{E}\left[\mathbf{1}_{e'p}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{ep}\mathbf{1}_{e'p} \mid \mathbf{1}_{ep}\right]\right] - l_e u_{ep} l_{e'} u_{e'p} \\
&= \mathbb{E}\left[\mathbf{1}_{ep}\mathbb{E}\left[\mathbf{1}_{e'p} \mid \mathbf{1}_{ep}\right]\right] - l_e u_{ep} l_{e'} u_{e'p} \\
&= \mathbb{P}\left(\mathbf{1}_{ep} = 0\right) \cdot 0 \cdot l_{e'} u_{e'p} + \mathbb{P}\left(\mathbf{1}_{ep} = 1\right) \cdot 1 \cdot 0 - l_e u_{ep} l_{e'} u_{e'p} \\
&= -l_e u_{ep} l_{e'} u_{e'p} \approx 0 \quad \because \text{mutation rate is small}
\end{aligned} \tag{6}$$

Not allowing child mutations gives $\mathbb{E}\left[\mathbf{1}_{e'p} \mid \mathbf{1}_{ep} = 1\right] = 0$, provided e is an ancestor of e' without the loss of generality. Therefore, we can justify the use of diagonal covariance matrices of random effects $\mathbf{u}$ by setting all terms involving u_{ep}^2 to zero. Ignoring second or higher-order terms of small quantities like the mutation rate is a common practice in coalescent theory [Wakeley, 2008].

The model also tells us how to weigh random effects. A common choice is to assign equal weights to all SNPs [Yang et al., 2010, 2011]. Some choose to take linkage disequilibrium (LD) into account [Speed et al., 2012], while others try to adaptively learn the weights in a data-driven fashion [Zhou and Zhao, 2021]. In ARG-LMM, the weights are determined by the shape of the ARG. Expanding the definition of $\mathbf{u}$ gives

$$\mathrm{Var}\left(\mathbf{u}_e\right) = \underbrace{l_e s_e}_{\text{edge area}} \cdot \overbrace{\frac{1}{s_e} \sum_{p:p\in e} \boldsymbol{\beta}_p^2 u_{ep}}^{\substack{\text{mutation rate-weighted} \\ \text{average of squared} \\ \text{effect sizes} := \tau_e^2}} = l_e s_e \tau_e^2 \tag{7}$$

where the first term is the area of edge e and the second term is a mutation rate weighted average of squared effect sizes of variants. We used the fact that $\mathrm{Var}\left(\mathbf{1}_{ep}\right) = l_e u_{ep}(1 - l_e u_{ep}) \approx l_e u_{ep}$ for $u_{ep} \approx 0$. If we view mutations as a process of sampling balls from an urn, in which the balls are the positions p and the urn is the set of positions $\{p : p \in e\}$, τ_e^2 is the variance of effect sizes $\boldsymbol{\beta}$ with respect to this sampling process where the sampling probability of a ball with value $\boldsymbol{\beta}_p$ is proportional to the mutation rate u_{ep}.

Formula (7) provides a principled way to weigh edges in a linear mixed model. The area $l_e s_e$ is proportional to the expected numbers of mutations on edge e. The second term τ_e^2 contains the contribution of effect sizes. If an edge contains a functional region, such as a coding-region, τ_e^2 is likely to be larger than

non-coding regions. The overall weight is the combination of the two, which does not solely depend on the functional significance (measured in τ_e^2) of the edge. The overall contribution to the variance can be small despite a large τ_e^2 because the chance of having a mutation is low due to a small area ($l_e s_e$). This means that an edge comprised of functionally less important regions can still significantly contribute to the variance because of its large area. In other words, the edge's overall variance as a measure of functional significance is confounded by its area. One should adjust for the edges' area to isolate the functional significance. Practically, this formulation makes variant annotation more straightforward because edges have a clear boundary on both sides. Existing functional annotations of SNPs rely on continuous measures of physical proximity to regions of known significance. If we turn our attention to edges, we simply need to check if an edge contains the functional region or not.

The Gaussian prior is the most popular choice for the distribution of the random effects [Yang et al., 2010, 2011]. Several extensions have been proposed including spike-and-slab [Vilhjálmsson et al., 2015] and non-parametric Dirichlet-process priors [Zhou and Zhao, 2021]. According to ARG-LMM, it is actually a mixture of indicator variables with a small mean. One could possibly argue that $\mathbf{u}_e$'s distribution is approximately normal in the following sense

$$\mathbf{u}_e \Big/ \sqrt{l_e s_e} \cdot \underbrace{\sqrt{\frac{1}{s_e} \sum_{p:p\in e} \boldsymbol{\beta}_p^2 u_{ep}}}_{:=\tau_e>0} \xrightarrow{s_e\to\infty} N(0,1) \tag{8}$$

by the central limit theorem, provided the edge has a sufficiently long span (large s_e). Nevertheless, the mean of the indicator $\mathbf{1}_{ep}$ might be too small in order to achieve the Gaussian limit. Indeed, Berry-Esseen theorem's approximation bound is inversely proportional to the square of $\mathbb{E}[\mathbf{1}_{ep}] = l_e u_{ep}$ [Durrett, 1996], suggesting that the normal approximation is unlikely to be adequate, although a sharper bound might exist.

We now turn our attention to the fixed effects portion $\mathbf{f}$. Under neutral evolutionary scenarios of various kinds, we may assume that the mutation rate is constant for each position [Wakeley, 2008, Durrett, 2008], i.e., $u_{ep} = u_p$ for all $e : p \in e$ (edge e contains position p). In this case, $\mathbf{Zf}$ reduces to a constant.

$$\begin{aligned}
[\mathbf{Zf}]_n &= \sum_{e=1}^{E} \mathbf{Z}_{ne} \mathbb{E}\left[\sum_{p:p\in e} \boldsymbol{\beta}_p \mathbf{1}_{ep}\right] = \sum_{e=1}^{E} \sum_{p:p\in e} \mathbf{Z}_{ne} \boldsymbol{\beta}_p l_e u_p \\
&= \sum_{p=1}^{P} \sum_{e:p\in e} \mathbf{Z}_{ne} \boldsymbol{\beta}_p l_e u_p = \sum_{p=1}^{P} \boldsymbol{\beta}_p u_p \left(\sum_{e:p\in e} \mathbf{Z}_{ne} l_e\right) \\
&= \sum_{p=1}^{P} \boldsymbol{\beta}_p u_p \cdot 2t_{\mathrm{root},p} = \text{constant w.r.t. } n
\end{aligned} \tag{9}$$

This means that the fixed component of the genetic effect is fully accounted for by an intercept. We can further generalize equation (9)'s argument. In this setting, we stipulate that a mutation rate at a position only

depends on time, i.e., the mutation rate is a function of time $u_p(t)$.

$$
\begin{aligned}
[\mathbf{Zf}]_n &= \sum_{e=1}^{E} \mathbf{Z}_{ne} \mathbb{E}\left[\sum_{p:p\in e} \boldsymbol{\beta}_p \mathbf{1}_{ep}\right] = \sum_{e=1}^{E} \sum_{p:p\in e} \mathbf{Z}_{ne} \boldsymbol{\beta}_p \mathbb{E}\left[\mathbf{1}_{ep}\right] \\
&= \sum_{e=1}^{E} \sum_{p:p\in e} \mathbf{Z}_{ne} \boldsymbol{\beta}_p \int_{t(e(c))}^{t(e(p))} u_p(t)dt = \sum_{p=1}^{P} \boldsymbol{\beta}_p \sum_{e:p\in e} \mathbf{Z}_{ne} \int_{t(e(c))}^{t(e(p))} u_p(t)dt \\
&= 2 \cdot \sum_{p=1}^{P} \boldsymbol{\beta}_p \int_{0}^{t_{\mathrm{root},p}} u_p(t)dt = \text{constant. w.r.t. } n
\end{aligned}
\tag{10}
$$

Therefore, a non-trivial fixed effect is present only if there is heterogeneity in the mutation rate across both time and lineages. This implies that the genetic contribution to a non-constant fixed effect, driven by the mutational process and the topology of the ARG, is zero in the aforementioned scenarios. The potential causes of non-trivial fixed effects, such as natural selection, is discussed at the end of this paper. We summarize our findings in the following statement.

Ancestral recombination graph linear mixed model (ARG-LMM)
The distribution of a complex trait follows a linear mixed model conditioned on the sample's ARG $\mathcal{T} = (\mathcal{N}, \mathcal{E})$.

$$
\mathbf{y} = \mathbf{Zu} + \mathbf{u}_0 + \boldsymbol{\varepsilon} \tag{11}
$$

with a constant vector $\mathbf{u}_0$ (all entries have the same value). The random effects $\mathbf{u}$'s covariance matrix is

$$
\Sigma_{\mathbf{u}} = \mathrm{diag}\left[l_e s_e \tau_e^2\right]_{e=1,\ldots,E} \tag{12}
$$

where $\tau_e^2 = \frac{1}{s_e} \sum_{p:p\in e} \boldsymbol{\beta}_p^2 u_{ep}$. The residual $\boldsymbol{\varepsilon}$ contains the contribution of non-genetic variables. The covariance matrix $\Sigma_{\mathbf{u}}$ is the branch genetic relatedness matrix described in Lehmann et al. [2025].

2.3 ARG-LMM simulation

We present an efficient algorithm to simulate the genetic value of a complex trait according to ARG-LMM with a normal prior on $\mathbf{u}$. Here and in the remainder of the text, we assume that all τ_e^2 equals a constant τ^2. One could flexibly model τ_e^2 to be a function of edge e's annotation (e.g. including a coding region), but we leave this to future work. Nevertheless, recent edges with longer spans may roughly obey this assumption as it will cover and average over many distinct regions of varying significance. The normal prior does not reflect the Poisson nature of mutations correctly as discussed in equation (8) but preserves the first and the second moments of the mutational process.

The following equation shows that a trait can be expressed as a sum of normal random variables over

local trees.

$$
\begin{aligned}
[\mathbf{Zu}]_n &= \sum_{e=1}^{E} \mathbf{Z}_{ne}\mathbf{u}_e =_d \sum_{e=1}^{E} \mathbf{Z}_{ne} \sum_{T:e\in T} \mathbf{u}_{e,T} \quad \because \mathbf{u}_e = \sum_{e:e\in T} \mathbf{u}_{e,T} \\
&= \sum_{e=1}^{E} \sum_{T:e\in T} \mathbf{Z}_{ne}\mathbf{u}_{e,T} = \sum_{T} \sum_{e:e\in T} \mathbf{Z}_{ne}\mathbf{u}_{e,T} \\
&=_d \sum_{T} \sqrt{s_T} \sum_{e:e\in T} \mathbf{Z}_{ne}\sqrt{t_{p(e)} - t_{c(e)}} \cdot \mathbf{z}_{e,T} \quad \because \mathbf{u}_{e,T} = \sqrt{s_T(t_{p(e)} - t_{c(e)})} \cdot \mathbf{z}_{e,T} \\
&=_d \sum_{T} \sqrt{s_T} \sum_{u:u\geqslant_T n} \sqrt{t_{p(u)} - t_u} \cdot \mathbf{z}_{u,T} \quad \because \mathbf{z}_{e,T}, \mathbf{z}_{u,T} \sim N(0,1)
\end{aligned}
\tag{13}
$$

where s_T is the span of a local tree T along the genome. We introduce additional notations: $\ell_T(n) = t_{p(n)} - t_n$ and intermediate arrays $v(n)$, $z(n)$, and $x(n)$ for dynamic programming. The following algorithm computes the last expression of equation (13).

Algorithm T. (*Trait genetic value simulation*). Given a sequence of positions that are recombination breakpoints b_k for $1 \leqslant k \leqslant K$ along the genome and corresponding sequences of edges to remove (R_k) and add (A_k) at each position. Sample the genetic values g_s for $1 \leqslant s \leqslant n_S$, as long as all samples are leaves in all trees. Let T be the current tree, $\ell_T(n) = t_{p(n)} - t_n$ be the length of the edge above n in T (or zero, if n has no parent), initialize $k = 1$, $x(n) = 0$, and $v(n) = 0$ for all $n \in \mathbf{V}$. At all times, define $z(n) = \sqrt{\ell_{T_k}(n)(b_k - x(n))} \cdot \mathbf{z}_{n,T_k}$ where $\mathbf{z}_{n,T} \sim N(0,1)$ is sampled for each tree given that T_k is the current tree. The k-th tree T_k has a span $s_{T_k} = b_k - b_{k-1}$ and set $b_0 = 0$.

T1. [Remove edges] For each edge $(c,p) \in R_k$, and for each node $n \geqslant_T p$, set $v(n) \mathrel{+}= z(n)$, then $v(c) \mathrel{+}= v(n)$, and $x(n) = b_k$. Then, set $x(c) = b_k$ and remove the edge.

T2. [Add edges] For each edge $(c,p) \in R_k$, and for each node $n \geqslant_T p$, set $v(n) \mathrel{+}= z(n)$, then $v(c) \mathrel{-}= v(n)$, and $x(n) = b_k$. Then, set $x(c) = b_k$ and add the edge.

T3. [Iteration] If $k < N$, set $k = k + 1$ and return to **T1**. Otherwise, set $y_s = v(s)$ for $1 \leqslant s \leqslant n_S$ and finish.

The algorithm proceeds by storing and updating intermediate quantities v, z, and x indexed by nodes. The genetic value is recovered at the end of the algorithm by aggregating these intermediate values efficiently. Updating the intermediate values is also efficient because it only alters the entries that are ancestral to the added or deleted edge in the local tree while leaving all other entries constant.

We show that the algorithm calculates the genetic value by preserving the following equation every step from **T1** to **T3** for every node $n \in \mathcal{N}$. The equality is preserved after the completion of each step of the algorithm.

$$
\sum_{r:r\geqslant_T n} v(r) + z(r) =_d \sum_{h:h\leqslant k} \sqrt{b_h - b_{h-1}} \sum_{r:n\leqslant_{T_h} r} \sqrt{\ell_{T_h}(r)} \cdot \mathbf{z}_{r,T_h} \tag{14}
$$

We call the RHS of the algorithm as the working genetic value. At the end of the algorithm, the working genetic value is precisely equation (13) for sample $n = s$ as $s_{T_k} = b_k - b_{k-1}$. The tree is empty at termination, so the LHS reduces to $v(s) + z(s)$. $z(s)$ is zero because $x(s) = b_K$ (recall that $z(n) = \sqrt{\ell_{T_k}(n)(b_k - x(n))}$). Therefore, $v(s)$ is the desired genetic value. To avoid confusion, note that all the intermediate values $v(r)$

and $z(r)$ are random variables. The equality signs involving these variables are often identity by distribution, and may not be the precise identity in terms of their realized value. We denote this by $X =_d Y$ for a pair of random variables X and Y that are identical in distribution. The equality trivially holds at the beginning of the algorithm as both sides are zero.

Step T3 is the easiest to analyze. Tree T and the intermediate quantities v and x are left constant. Advancing the position from k to $k+1$ changes $z(s) = \sqrt{\ell_T(s)(b_k - x(s))} \cdot \mathbf{z}_{s,T_k}$ to $z'(s) = \sqrt{\ell_{T'}(s)(b_{k+1} - x(s))} \cdot \mathbf{z}_{s,T_{k+1}}$. As $T = T'$ after **T3**, the overall change is distributionally identical to adding

$$\mathcal{N}\left(0, \sum_{r:r \geqslant_{T'} n} \sqrt{\ell_{T'}(s)(b_{k+1} - b_k)}^2\right) \tag{15}$$

which is again identical to the update in the RHS:

$$\sqrt{b_{k+1} - b_k} \sum_{r:n \leqslant_{T'} r} \sqrt{\ell_{T'}(r)} \cdot \mathbf{z}_{r,T'} \tag{16}$$

A slight abuse of notation is here that k and $k+1$ are the indices of tree T and T', respectively. **T3** is the only step that changes the distribution of the variables on both sides of the (14) by adding the contribution of the k-th local tree. Note that this is the only step where the working genetic value changes. In the following steps, v, z, and x change while the working genetic value (14) remains constant.

Step T1 and **T2** leave the distribution of both sides unchanged. Without the loss of generality, we prove this for **T2**. Suppose that we added edge (c, p) where c and p are child and parent nodes of the edge, respectively. We can divide the nodes of $\mathcal{N}$ into four categories as (1) the child node c, (2) nodes below c, (3) the parent node p and nodes above p, and (4) all other nodes. The insertion changes the values of the intermediate arrays to v', z', and x' following **T2**. We denote the new tree resulting from the insertion as T'.

Observe that the difference in the LHS after **T2** is

$$\begin{aligned}
&\sum_{r:n \leqslant_{T'} r} v'(r) + z'(r) - \sum_{r:n \leqslant_T r} v(r) + z(r) \\
&= \sum_{r:n \leqslant_{T'} r, n \nleqslant_T r} v'(r) + z'(r) \\
&+ \sum_{r:n \leqslant_{T'} r, n \leqslant_T r} v'(r) + z'(r) - v(r) - z(r) \\
&- \sum_{r:n \nleqslant_{T'} r, n \leqslant_T r} v(r) + z(r) \\
&= \sum_{r:n \leqslant_{T'} r, n \nleqslant_T r} v'(r) + z'(r) + \sum_{r:n \leqslant_T r} v'(r) + z'(r) - v(r) - z(r)
\end{aligned} \tag{17}$$

for node n. The last line follows from $\{r : n \leqslant_T r\} \subset \{r : n \leqslant_{T'} r\}$ because T' has more edges than T. Nodes in each category have distinct values of the former and the latter sum of this equation.

The set $\{r : n \leqslant_{T'} r, n \nleqslant_T r\}$ of the first summation is

$$= \begin{cases} \{r : p \leqslant_{T'} r\} & n \text{ in the first and the second category} \\ \varnothing & n \text{ in the third and the fourth category} \end{cases} \tag{18}$$

This is because p and the nodes above p are the nodes that were previously not ancestors in T, but became ancestors of c and those below c after the addition of the new edge. The set is empty for the nodes in the third and the fourth category because their ancestor nodes are unchanged after edge insertion. Therefore, the first summation is

$$\begin{cases} \sum_{r:p\leqslant_{T'}r} v'(r) + z'(r) & n \text{ in the first and the second category} \\ 0 & n \text{ in the third and the fourth category} \end{cases} \tag{19}$$

The summand of the second summation $v'(r) + z'(r) - v(r) - z(r)$ is

$$= \begin{cases} v'(c) - v(c) & r \text{ in the first category} \\ 0 & r \text{ in the second, third and the fourth category} \end{cases} \tag{20}$$

For $r = c$ (r is in the first category), it follows from $z(c) = 0$ due to $\ell_T(c) = 0$ and $z'(c) = 0$ due to $x'(c) = b_k$. All the intermediate values of the second and the fourth category is unchanged by **T2**, so the summand is trivially zero. When $r = p$ (r belongs to the third category), $v'(r) = v(r) + z(r)$ and $z'(r) = 0$ by the operations $v(r) \mathrel{+}= z(r)$ and $x(r) = b_k$ by **T2**. Hence, the second summation is

$$\begin{cases} v'(c) - v(c) & n \text{ is in the first and the second category} \\ 0 & n \text{ is in the third and the fourth category} \end{cases} \tag{21}$$

because the set $\{r : n \leqslant_T r\}$ contains c if and only if n belongs to either the third or the fourth category. An expression for $v'(c) - v(c)$ follows from the operation $v(r) \mathrel{+}= z(r)$ followed by $v(c) \mathrel{-}= v(r)$ of **T2**.

$$v'(c) - v(c) = -\sum_{r:p\leqslant_T r} v(r) + z(r) \tag{22}$$

Combining the aforementioned results, we see that equation (17) is

$$\begin{cases} \sum_{r:p\leqslant_{T'}r} v'(r) + z'(r) - \sum_{r:p\leqslant_T r} v(r) + z(r) & n \text{ is in the first and the second category} \\ 0 & n \text{ is in the third and the fourth category} \end{cases} \tag{23}$$

Both cases reduces to zero because the set of nodes ancestral to p are the same in T and T', i.e., $\{r : p \leqslant_T r\} = \{r : p \leqslant_{T'} r\}$, and $v'(r) = v(r) + z(r)$ for these nodes due to **T2** and $z'(r) = 0$. The right-hand side remains the same because the operation changes nothing in its expression: it's the working sum until the previous local tree that does not contain any of the components of the current tree being modified. In summary, **T1** and **T2** modify the intermediate values v, x, and z while leaving the working genetic value constant so that **T3** correctly updates the genetic value. Note that this algorithm computes the raw genetic value prior to multiplying by τ. See **Section 2.6.2** for more details of fully simulating the trait based on this algorithm.

2.4 Average information residual maximum likelihood

Average information residual maximum likelihood (AI-REML) is a second-order optimization technique similar to the Newton-Raphson method and Fisher scoring [Gilmour et al., 1995]. An unconstrained optimization

problem of a twice differentiable multivariate function $f : \mathbb{R}^p \to \mathbb{R}$

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}) \tag{24}$$

can be solved by the following iteration

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - [\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n)]^{-1} \cdot \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n) \tag{25}$$

This is called the Newton-Raphson method.

Fisher scoring is a variant of the Newton-Raphson method when the optimization target includes a realized random variables. Let $\boldsymbol{\theta}$ be the parameter and $\mathbf{x}$ be a random variable that is generated from a distribution parameterized by $\boldsymbol{\theta}$. The optimization target (e.g. negative log-likelihood) $\ell(x, \theta)$ is a function of both $\boldsymbol{\theta}$ and x where x is a realization of $\mathbf{x}$. The optimization problem is

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \ell(x, \boldsymbol{\theta}) \tag{26}$$

The corresponding Newton-Raphson update is

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - [\nabla^2_{\boldsymbol{\theta}} \ell(x, \boldsymbol{\theta}_n)]^{-1} \nabla_{\boldsymbol{\theta}} \ell(x, \boldsymbol{\theta}_n) \tag{27}$$

Fisher scoring replaces $\nabla^2_{\boldsymbol{\theta}} \ell(x, \boldsymbol{\theta}_n)$ by its expectation

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbb{E}_{\boldsymbol{\theta}_n} \left[\nabla^2_{\boldsymbol{\theta}} \ell(\mathbf{x}, \boldsymbol{\theta}_n) \right]^{-1} \nabla_{\boldsymbol{\theta}} \ell(x, \boldsymbol{\theta}_n) \tag{28}$$

in which the random variable $\mathbf{x}$ sampled at $\boldsymbol{\theta} = \boldsymbol{\theta}_n$ replaces the realized observation x. The expectation $\mathbb{E}_{\boldsymbol{\theta}} \left[\nabla^2_{\boldsymbol{\theta}} \ell(\mathbf{x}, \boldsymbol{\theta}) \right]$ is called the *Fisher information*.

Average information is the average of these two quantities

$$\mathrm{AI}(x, \theta) = \frac{1}{2} \left(\nabla^2_{\boldsymbol{\theta}} \ell(x, \boldsymbol{\theta}) + \mathbb{E}_{\boldsymbol{\theta}} \left[\nabla^2_{\boldsymbol{\theta}} \ell(\mathbf{x}, \boldsymbol{\theta}) \right] \right) \tag{29}$$

and the AI update is

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathrm{AI}(x, \boldsymbol{\theta}_n)^{-1} \nabla_{\boldsymbol{\theta}} \ell(x, \boldsymbol{\theta}_n) \tag{30}$$

which is identical to the Newton-Raphson method and Fisher scoring except that it uses the average information.

All three methods fall into the category of second-order optimization methods. Roughly speaking, second-order methods locally approximate the optimization target as a quadratic function by utilizing curvature information encoded in the second derivative (the Hessian). While Newton-Raphson method uses the Hessian directly, Fisher scoring uses the Hessian's average over the data distribution at the current step, and AI update uses the average of the two. As quadratic approximation is more accurate than linear approximation used by first-order methods, second-order methods converge more quickly and tightly than first-order methods.

Despite this advantage, quadratic methods are often computationally more expensive than first-order methods because computing the Hessian or related quantities can be very demanding. This is also true for the restricted maximum likelihood target of ARG-LMM. We write ARG-LMM as

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \mathbf{V}) \quad \text{where} \quad \mathbf{V} = \tau^2 \mathbf{Z} \Sigma_{\mathbf{u}} \mathbf{Z}^T + \sigma^2 \mathbf{I} \tag{31}$$

The restricted maximum likelihood (REML) objective is (up to a constant factor)

$$\ell(\mathbf{y}, \mathbf{X}; \tau^2, \sigma^2) = (N - P)\log(2\pi) + \log\det(\mathbf{V}) + \log\det(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}) + \mathbf{y}^T\mathbf{P}\mathbf{y} \tag{32}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. The first derivatives are

$$\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2} &= \operatorname{tr}(\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})) - \mathbf{y}^T\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}\mathbf{y} \\
&= \operatorname{tr}(\mathbf{P}) - \mathbf{y}^T\mathbf{P}\mathbf{P}\mathbf{y} \\
\frac{\partial \ell}{\partial \tau^2} &= \operatorname{tr}(\mathbf{P}(\nabla_{\tau^2}\mathbf{V})) - \mathbf{y}^T\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}\mathbf{y} \\
&= \operatorname{tr}(\mathbf{P}\mathbf{C}) - \mathbf{y}^T\mathbf{P}\mathbf{C}\mathbf{P}\mathbf{y}
\end{aligned} \tag{33}$$

because $\nabla_{\sigma^2}\mathbf{V} = \mathbf{I}$ and $\nabla_{\tau^2}\mathbf{V} = \mathbf{C}(:= \mathbf{Z}\Sigma_{\mathbf{u}}\mathbf{Z}^T)$, and the second derivatives are

$$\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2 \partial \sigma^2} &= -2\mathbf{y}^T\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}\mathbf{y} + \operatorname{tr}(\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})) \\
\frac{\partial \ell}{\partial \sigma^2 \partial \tau^2} &= -2\mathbf{y}^T\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}\mathbf{y} + \operatorname{tr}(\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V})) \\
\frac{\partial \ell}{\partial \tau^2 \partial \tau^2} &= -2\mathbf{y}^T\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}\mathbf{y} + \operatorname{tr}(\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V}))
\end{aligned} \tag{34}$$

The expectations of the second derivatives are

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \ell}{\partial \sigma^2 \partial \sigma^2}\right] &= -\operatorname{tr}(\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})) \\
\mathbb{E}\left[\frac{\partial \ell}{\partial \sigma^2 \partial \tau^2}\right] &= -\operatorname{tr}(\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V})) \\
\mathbb{E}\left[\frac{\partial \ell}{\partial \tau^2 \partial \tau^2}\right] &= -\operatorname{tr}(\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V}))
\end{aligned} \tag{35}$$

Finally, averaging the Hessian (34) and the Fisher information (35) yields the average information

$$\begin{aligned}
\mathrm{AI}(\sigma^2, \sigma^2) &= -\mathbf{y}^T\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}\mathbf{y} \\
&= -\mathbf{y}^T\mathbf{P}\mathbf{P}\mathbf{P}\mathbf{y} \\
\mathrm{AI}(\sigma^2, \tau^2) &= -\mathbf{y}^T\mathbf{P}(\nabla_{\sigma^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}\mathbf{y} \\
&= -\mathbf{y}^T\mathbf{P}\mathbf{P}\mathbf{C}\mathbf{P}\mathbf{y} \\
\mathrm{AI}(\tau^2, \tau^2) &= -\mathbf{y}^T\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}(\nabla_{\tau^2}\mathbf{V})\mathbf{P}\mathbf{y} \\
&= -\mathbf{y}^T\mathbf{P}\mathbf{C}\mathbf{P}\mathbf{C}\mathbf{P}\mathbf{y}
\end{aligned} \tag{36}$$

At the n-th iteration, the working fixed effects estimates $\mathbf{b}^{(n)}$ is obtained from solving the generalized least square (GLS) from the current variance component parameters $\tau^{(n)}$ and $\sigma^{(n)}$. We denote the working covariance matrix as $\mathbf{V}^{(n)}$ computed from $\tau^{(n)}$ and $\sigma^{(n)}$.

$$\mathbf{b}^{(n)} = (\mathbf{X}^T\mathbf{V}^{(n)-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{(n)-1}\mathbf{y} \tag{37}$$

Table 1: The default hyperparameter settings of `tslmm`.

Parameter name	Default value
Conjugate gradient convergence threshold	10^{-5}
Number of test vectors for gradient estimation	50
Rank of the randomized Nystrom preconditioner	500
Number of threads in matrix-vector multiplication	1
Number of test vectors for Haseman-Elston regression	50
REML convergence criteria	0.05
Minimum REML steps before termination	15

$\mathbf{V}^{-1}$ in $\mathbf{P}$ is computed by the conjugate gradient algorithm. To speed up the conjugate gradient algorithm, `tslmm` deploys the randomized Nystrom preconditioner [Frangella et al., 2021]. Traces in equation (33) are obtained from `XTrace` [Epperly et al., 2024]. All three algorithms in the paragraph access the genetic relatedness matrix through `tskit`'s matrix-vector product algorithm described in Lehmann et al. [2025]. Note that Harville [1977] showed that REML coincides with minimum norm quadratic unbiased estimation (MINQUE), which is a moment-based estimator. Hence, normality is not a requirement for consistent point estimation, provided the covariance matrix $\mathbf{V}$ is correct.

2.5 Genomic prediction

The best linear unbiased predictor (BLUP) is $\widehat{\mathbf{g}}$ of the genetic value $\mathbf{g}$ is

$$\tau^2\mathbf{C} \cdot (\tau^2\mathbf{C} + \sigma^2\mathbf{I})^{-1} \cdot (\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{38}$$

In practice, we replace the true parameters τ^2 and σ^2 with their estimates obtained by AI-REML. As before, we multiply $\mathbf{C}$ through `tskit` and solve $\tau^2\mathbf{C} + \sigma^2\mathbf{I}$ by conjugate gradient.

2.6 `tslmm` evaluation experiments

2.6.1 Default hyperparameter settings

Several hyperparameters affect the runtime and accuracy of `tslmm`. The number of test vectors is the dimension of the test vectors in randomized trace estimation [Epperly et al., 2024]. The preconditioner's rank is also the number of test vectors in randomized singular value decomposition used to construct the preconditioner [Frangella et al., 2021]. `tskit`'s matrix-vector product Python API is free from the global interpreter lock (GIL), so is capable of multi-threading. In the benchmark, we limited the number of threads to 1. When the parameter estimates change less than 5%, `tslmm` continues for an additional 15 steps and then terminates. It returns the average value of the last 15 iterations. We listed these hyperparameters and their default values in **Table 1**.

2.6.2 Trait simulation

We first simulated the genetic value from tree sequences using the algorithm described in the **ARG-LMM simulation** section. To make the genetic and non-genetic components comparable, we divided the square root of $4 \times \sum_T s_T t_{\mathrm{root},T}$ from the genetic value. This is because the upper bound of an individual's genetic

value is

$$\mathrm{Var}\left(h_{n1}+h_{n2}\right) \leqslant 2\mathrm{Var}\left(h_{n1}\right)+2\mathrm{Var}\left(h_{n2}\right)=4\times\tau^2\sum_T s_T t_{\mathrm{root},T} \tag{39}$$

as explained in equation (42). The final trait output was

$$\tau\cdot\{\text{genetic value}\}+\sigma\cdot\{\text{random noise}\} \tag{40}$$

where the random noise was sampled from a standard normal distribution.

2.6.3 Runtime measurement

The genomes we simulated were 100Mb (=10^8bp) long. The recombination rate was set to 10^{-8}. We adopted an exponentially growing panmictic model with a growth rate of 0.01 per generation. The sample sizes were $N \in \{2\times10^3, 5\times10^3, 1\times10^4, 2\times10^4, 5\times10^4, 1\times10^5\}$. We simulated the traits and fitted the variance components with `tslmm` 5 times to obtain the runtime. We set $\tau^2=1$ and $\sigma^2=1$, which corresponds to the case where the most genetically variable individual's trait has a roughly 50:50 ratio of genetic and non-genetic composition.

2.6.4 Variance component estimation in simulated trees

The same genomes from **Runtime measurement** section was used. We recorded the genetic variance component τ^2 for each iteration. $\sigma^2=1$ was fixed and $\tau^2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. We repeated the procedure 5 times to quantify the variability of the estimates.

2.6.5 Variance component estimation in inferred trees

We simulated mutations with mutation rate 10^{-8} on the tree sequences used in the **Runtime measurement** section. We used `tsinfer` and `tsdate` to infer and date the tree sequences. These tree sequences were used to fit the variance components with `tslmm`. $\sigma^2=1$ was fixed and $\tau^2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Note that the trait was simulated from the true tree sequence. The inferred tree sequence was used for estimation. We repeated the procedure 5 times to quantify the variability of the estimates.

2.6.6 Genomic prediction benchmark

For each of the simulated and inferred tree sequences in the previous section, we simulated traits with $\sigma^2=1$ and $\tau^2 \in \{0, 0.4, 0.8, 1.2, 1.6, 2\}$. Traits were generated from the true simulated tree sequences. Variance components were estimated using both true and simulated trees. Only half of the individuals were used in estimation. Best linear unbiased predictor (BLUP) of each individual was generated using the algorithm described in **Genomic prediction**. We computed the sample Pearson correlation between the actual trait values and the BLUPs produced from true and inferred tree sequences. The correlation was computed in the leftover individuals not used in estimation. The average of the correlation over 5 replicates was recorded.

2.7 American admixture simulation

We used the demographic model in `stdpopsim` that was initially documented by Browning et al. [2018]. We set both the recombination rate and the mutation rate to 10^{-8} and simulated 500 individuals from each

population in {AFR, EUR, ASIA, AMR}. The genome was 10^8 long. We calculated the genetic relatedness between the haplotypes of each individual using the function `TreeSequence.genetic_relatedness()`. The options were `mode=branch`, `centre=False`, and `proportion=False`.

2.8 Pseudoreplication experiment

We simulated a 10^4 long tree sequence of 2000 individuals. It followed a constant size demography and the recombination rate was set to 10^{-8}. For each simulation, samples were split into two equal size subsets, and the variance components were estimated separately in each subset. The variance components were set to $\sigma^2 = 1$ and $\tau^2 = 3$.

2.9 Pedigree experiment

We simulated a pedigree using `SLiM` version 4 [Haller and Messer, 2023], under an individual-based model with spatially-restricted mating [Chevy et al., 2025]. The simulation takes place in a square spatial domain, where the population size is stochastic and emergent from local dispersal and competition between proximal individuals. Average dispersal distance and strength of competition were chosen so as to result in a strong signal of isolation by distance, so that nearby individuals are typically close relatives. The spatial pedigree was simulated for 150 time units (where a time unit corresponds to the expected lifetime of an individual). Subsequently, we used `msprime` with the `fixed_pedigree` model [Anderson-Trocmé et al., 2023] to simulate partial genealogies (conditional on the pedigree) for the 13,000 individuals alive at the end of the simulation, along a sequence of length 10^8 with a recombination rate of 10^{-8}. Finally, we completed the partial genealogies backwards in time with the standard coalescent model, using the average population size from the spatial simulation to determine the rate of coalescence. We repeated the coalescent simulations to acquire multiple realizations of the ARG from the pedigree. Finally, for each simulated ARG, we simulated neutral mutations at a rate of 10^{-8}.

For a fair comparison, we simulated the trait by `tstrait` that assigns effect sizes to materialized mutations, instead of **Algorithm T** that smooths out the effects of the mutations [Tagami et al., 2024]. One percent of the mutations were set to be causal and their effect sizes were drawn from a standard normal distribution. The resulting genetic value was divided by the square root of $4 \times \sum_T s_T t_{\text{root},T}$ similar to **Section 2.6.2**. Finally, we added a random noise following the standard normal distribution to the genetic value.

We extracted the parent-child relationship from the simulated tree sequence. The extracted pedigree was fed to the `getASubset` function in the R `pedigreeTools` package to compute the pedigree relatedenss matrix (pGRM) [Vazquez et al., 2018]. We passed the pGRM to the `lmebreed` function in the `lme4breeding` package to obtain the pedigree-based BLUP prediction [Covarrubias-Pazaran, 2024]. The site GRM (sGRM) was computed by the `arg-needle-lib` with the tree sequence generated in the previous step [Zhang et al., 2023]. As described in their paper, Zhang et al. [2023] normalizes the GRM according to the allele frequency. We fed the sGRM to `lmebreed` to obtain the site-based predictions.

3 Results

3.1 `tslmm` is a scalable implementation of ARG-LMM

`tslmm` is a new member of the `tskit` ecosystem for complex trait analysis [Kelleher et al., 2016, Ralph et al., 2020]. Currently, `tslmm` only supports a simplified version of ARG-LMM that has a single genetic variance component parameter τ^2. It assumes that $\tau_e^2 = \tau^2$ for all $e \in \mathcal{E}$ although it accounts for the varying edge

areas $l_e s_e$ that constitute the edge variance. This is likely a good approximation when the edges have a large s_e. When the edges cover a long range in the genome, varying effect sizes and the mutation rates of the positions will average out and lead to similar values of τ_e^2.

The software fits the variance components of ARG-LMM via restricted maximum likelihood (REML). REML is unbiased and statistically the most efficient estimator in variance component estimation. Fitting REML requires inverting the GRM that amounts to $\mathcal{O}(N^3)$ operations. To this end, we lower the computational burden to $\mathcal{O}(N)$ by utilizing an efficient matrix-vector multiplication algorithm implemented in `tskit`. The matrix-vector product algorithm is similar to the algorithm previously described in Kelleher et al. [2016] and Ralph et al. [2020] that recycles information in consecutive local trees to avoid redundant calculations. The algorithm keeps several bookkeeping variables indexed by nodes for dynamic programming. As we iterate over the local trees of the tree sequence, bookkeeping variables are updated only minimally as adjacent local trees are almost the same except for a few edges. After the traversal is complete, the algorithm aggregates the bookkeeping variables to obtain the matrix-vector product. The algorithm is almost the same as the trait simulation algorithm in **Materials and Methods**, and was documented earlier in Lehmann et al. [2025].

We reformulated the average information REML (AI-REML) algorithm [Gilmour et al., 1995] with the new matrix-vector multiplication algorithm. AI-REML is a second-order optimization scheme that deploys the target's curvature information, improving speed and convergence compared to first-order methods. Traditional implementations are computationally intractable because they need to invert the GRM. We present an algorithm that computes the average information at a cost that is no more expensive than calculating the gradient. We generated datasets of 100Mb genomes with different numbers of diploid individuals to measure the time required for `tslmm` to fit the variance components. We plotted the runtime with respect to the number of individuals, nodes, and edges (**Figure 1**).

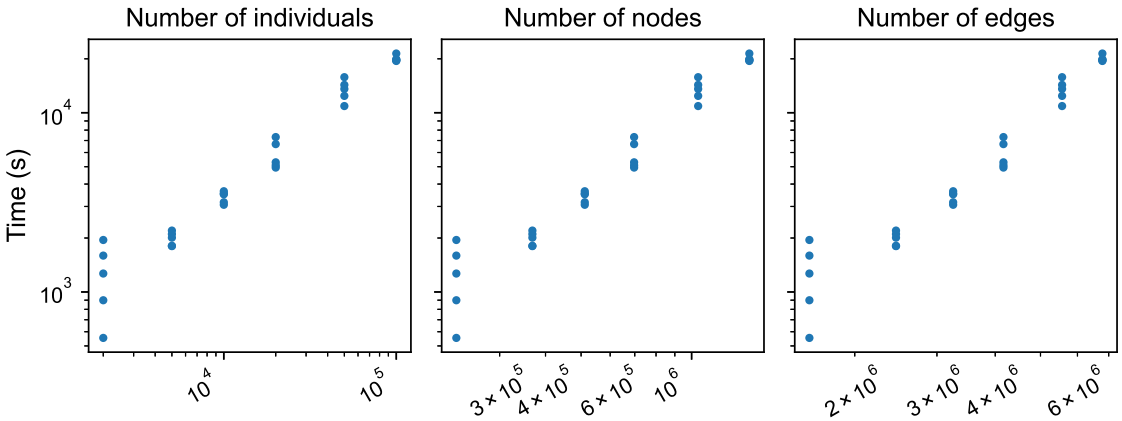

Figure 1: Runtime benchmark. We simulated 100Mb genomes with varying number of individuals. 5 simulations were done for each sample size.

When the true ARG is available, `tslmm` accurately recovers the variance component parameters (**Figure 2A**). In practice, one does not have access to the true ARG, so instead uses the inferred version. Even with an inferred ARG, `tslmm` produces accurate estimates, although the estimates are noisier than using the true ARG (**Figure 2B**). We also find that there is only a slight decrease in the accuracy of the best linear unbiased predictor (BLUP) predictions when using an inferred ARG compared to the true ARG (**Figure 3**).

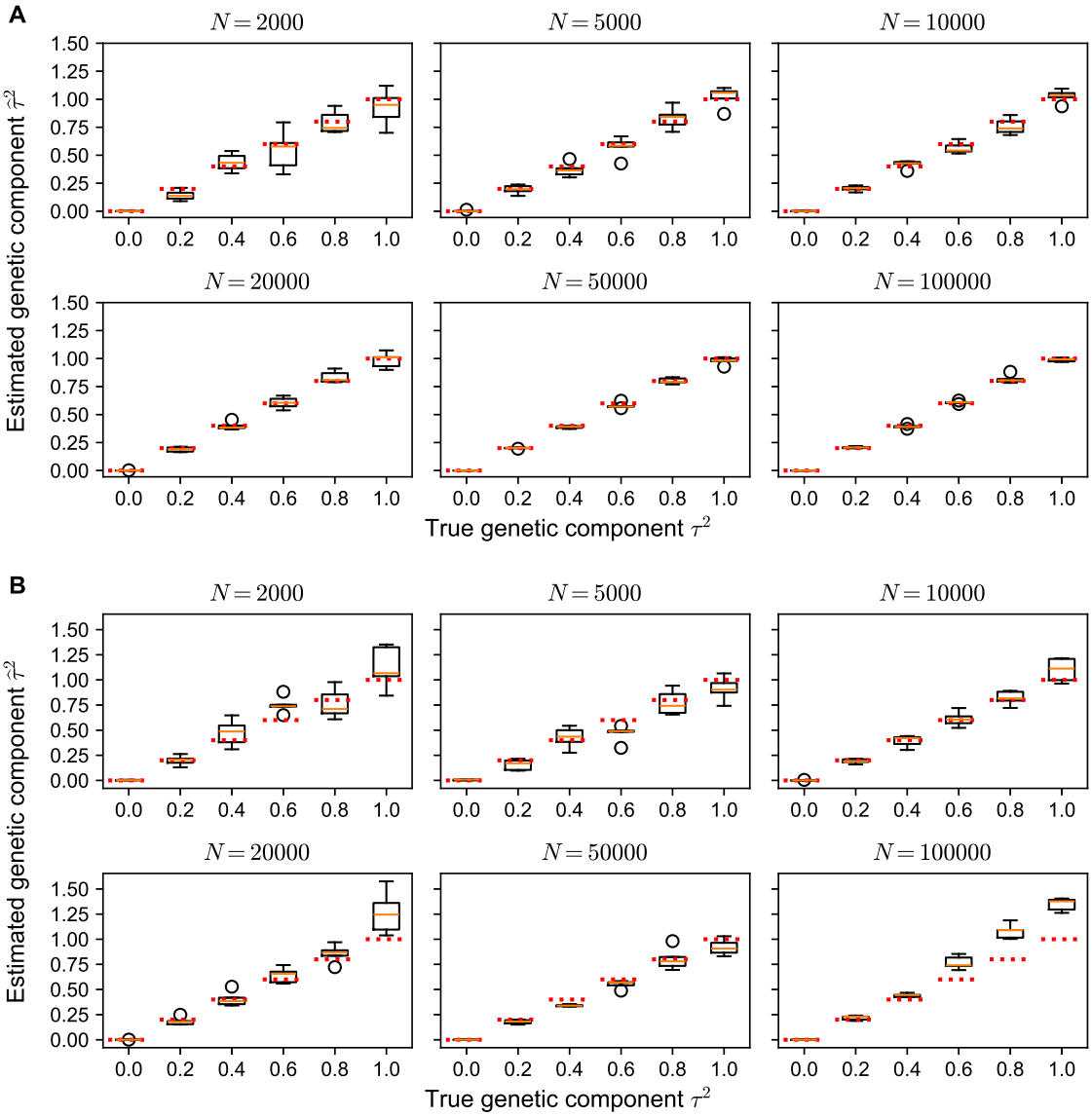

Figure 2: `tslmm`'s variance component estimates from true and inferred tree sequences. Each simulation was repeated 5 times. (A) Results from true tree sequences. (B) Results from inferred tree sequence.

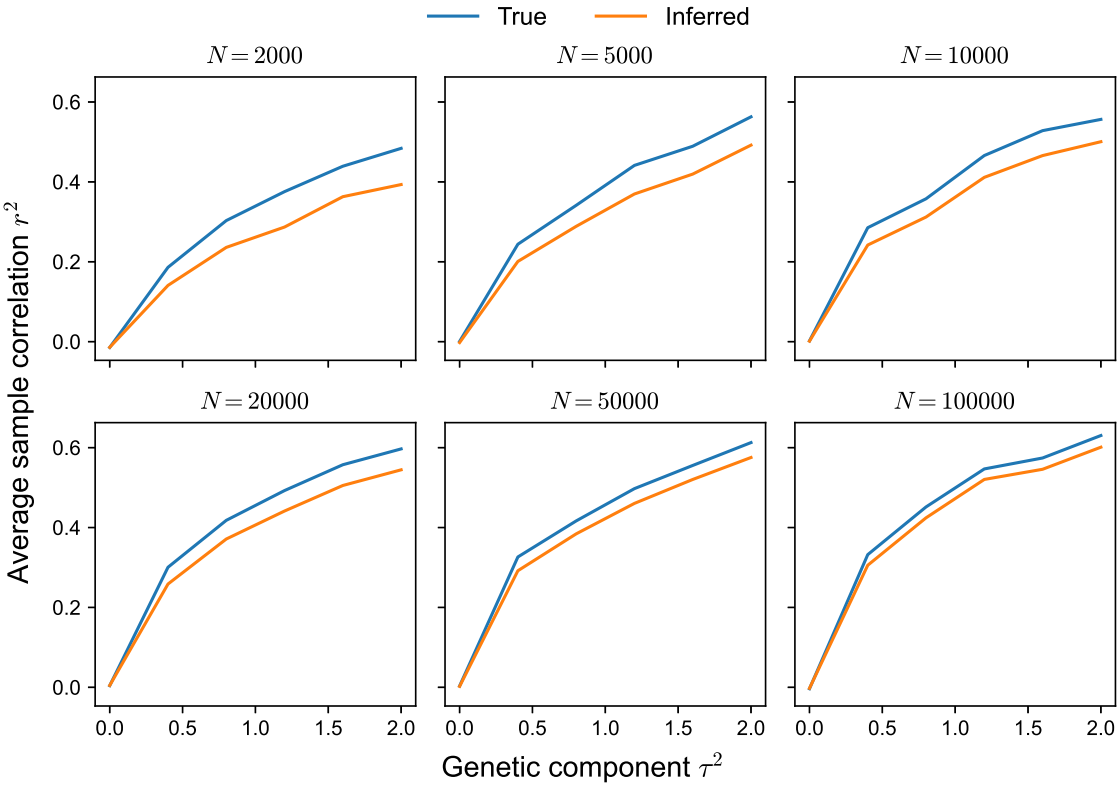

Figure 3: Average sample Pearson correlation of actual trait values and `tslmm`'s BLUP predictions from true and inferred trees.

3.2 Narrow-sense heritability beyond haploids

Narrow-sense heritability is frequently estimated from genomic linear mixed models (GLMMs) by taking the quotient of the genetic and environmental variance components. GLMMs commonly assume that the GRM's diagonal is 1 and the environmental variance is constant across individuals [Yang et al., 2010, 2011, Wu and Sankararaman, 2018, Zhang et al., 2023]. This is a direct consequence of normalizing the genotype matrix with respect to the sample frequency. As a result, the relative contributions of genetic and environmental variances remain constant for all individuals. Hence, narrow-sense heritability, which is the ratio of the two variances, serves as a common measure that applies to all individuals. ARG-LMM tells a different story: except for haploids, a trait's genetic component is different across individuals. Hence, no single quantity can effectively summarize the relative importance of genetic and non-genetic effects in organisms with two or more ploidy. In haploids, all the diagonal elements of the ARG-GRM are equal to a weighted average of the local time of the root node [Zhang et al., 2023]. Thus, we can divide the GRM by this constant so that the diagonals are equal to 1. However, with larger ploidy, the diagonals of the GRM are the variance of the sum of haploid genetic effects.

Suppose that h_{n1} and h_{n2} are the haplotype genetic effects of an diploid individual n. Then,

$$\mathrm{Var}\left(h_{n1} + h_{n2}\right) = \mathrm{Var}\left(h_{n1}\right) + \mathrm{Var}\left(h_{n2}\right) + 2\mathrm{Cov}\left(h_{n1}, h_{n2}\right) \tag{41}$$

The first two variances are equal to the product of τ^2 and the branch relatedness between the two chromosomes.

$$\mathrm{Var}\left(h_{n1}\right) = \mathrm{Var}\left(h_{n2}\right) = \tau^2 \sum_T s_T t_{\mathrm{root},T} \tag{42}$$

The sum runs over all local trees, s_T is the span of the local tree T, and $t_{\mathrm{root},T}$ is the root time of T. Nonetheless, the all individuals have a different value for the last term. The covariance between the two haplotypes $\mathrm{Cov}\left(h_{n1}, h_{n2}\right)$, the self-relatedness, is highly variable. In **Figure 4**, we plotted the distribution of self-relatedness of diploid individuals simulated according to the demographic history of Americans [Browning et al., 2018].

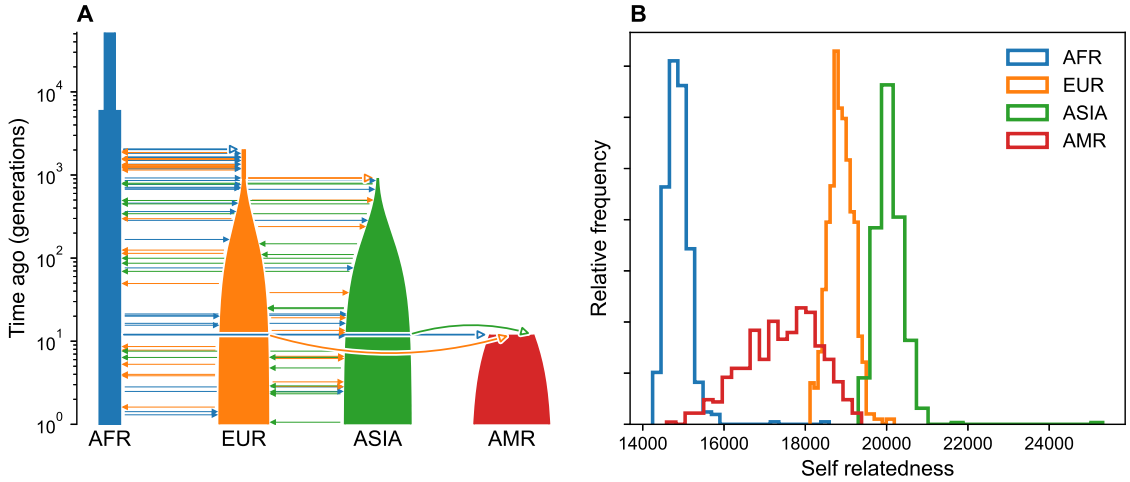

Figure 4: (A) The demographic history of Americans adopted from Browning et al. [2018]. (B) The distribution of self-relatedness of individuals simulated from the demographic process in the left panel. AFR: African, EUR: European, ASIA: Asian, AMR: Americans.

For example, an individual with parents of different ancestry will have lower relatedness between the two copies than an individual with parents of the same ancestry. As a result, some individuals will have larger genetic variance with respect to the mutational process, while some will have a smaller genetic variance. As a result, the proportion of genetic variance relative to the overall trait variance varies significantly across individuals. Therefore, we cannot obtain a single summary of the relative contribution of genetic and non-genetic factors for everyone in the sample, let alone in the population.

This observation translates to different upper bounds of the polygenic prediction accuracy across populations. Empirical studies have shown that polygenic prediction is more accurate in the European ancestry, and this has been mainly attributed to the larger sample size of European cohorts and lack of appropriate methods [Martin et al., 2019, Ding et al., 2023]. As shown in **Figure 4**, Europeans and East Asians have larger genetic variance due to the initial bottleneck at the time they were found. Therefore, we predict that

polygenic prediction in the African ancestry will not fully catch up with European and East Asian individuals despite the growing number of samples. A GRM's non-constant diagonal is not unique to branch GRM. In the presence of inbreeding, the genetic variance of an individual conditioned on the pedigree is enlarged by a factor of $1 + F_i$ where F_i is the inbreeding coefficient of individual i [Falconer, 1995]. In this context, heritability given the pedigree is the relative contribution of genetic factors to trait variance only in individuals with $F_i = 0$, which mainly includes the founders of the pedigree. This leads to a higher predictive accuracy of linear mixed models for individuals subject to higher levels of inbreeding [Satoh, 2024]. In terms of genealogies, two chromosomes of an inbred individual share a larger common branch area, leading to larger self relatedness. Thus, our result based on ARG-LMM shows that higher predictive accuracy in an inbred population still persists after conditioning on the ARG.

3.3 Shared evolutionary history and pseudoreplication

Rosenberg and VanLiere [2009] showed that shared genealogy of non-overlapping samples can induce pseudoreplication in genetic association studies. They ran coalescent simulations to generate two non-overlapping samples and conducted trait-SNP association tests. They found that due to the shared genealogical process underlying the two samples, test statistics separately produced from the two samples are correlated. They conclude that genetic association studies conducted on non-overlapping samples do not necessarily provide independent evidence of genetic signals.

Both the genealogy and the mutational process varied in Rosenberg and VanLiere [2009]'s simulation. We show that their argument remains valid in the setting of fixed genealogies. The covariance between statistics $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ obtained from two separate datasets can be expressed as

$$\begin{aligned}\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2\right) &= \mathrm{Cov}\left(\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_1 \mid \mathbf{u}\right], \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_2 \mid \mathbf{u}\right]\right) + \mathbb{E}\left[\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2 \mid \mathbf{u}\right)\right] \\ &= \mathrm{Cov}\left(\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_1 \mid \mathbf{u}\right], \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_2 \mid \mathbf{u}\right]\right)\end{aligned} \tag{43}$$

The second term is zero because genetic random effect $\mathbf{u}$ is the only source of covariation in the data. This is summarized in the left panel of **Figure 5A** that depicts the data-generating process of ARG-LMM as a directed acyclic graph (DAG).

The first term represents the covariance of the two statistics. We can clearly see that the estimates $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ are correlated in the left panel of **Figure 5** through the path $\widehat{\boldsymbol{\theta}}_1 \leftarrow \mathbf{y}_1 \leftarrow \mathbf{u} \rightarrow \mathbf{y}_2 \rightarrow \widehat{\boldsymbol{\theta}}_2$. The degree of pseudoreplication ultimately depends on the extent of shared genealogy. We estimated the variance components with `tslmm` from two non-overlapping set of individuals. Estimates of the non-genetic variance component σ^2 were independent (**Figure 6A**) while those of genetic variance component τ^2 appear highly correlated (**Figure 6B**). Hence, non-overlapping samples are insufficient to ensure independence of the estimates. To guarantee this independence, the incorrect data-generating process in the right panel of **Figure 5**, which assumes that two samples have a completely separated mutational history, is demanded.

3.4 Comparing pedigree and ARG-based genetic predictions

Pedigree plays a central role in livestock and crop improvement through the prediction of genetic (breeding) values. It encodes the genetic inheritance pattern, and hence can be used to compute the expected amount of shared genetic materials with respect to the Mendelian sampling process. The resulting genetic relatedenss matrix, hereafter the pedigree GRM, is passed to the linear mixed model to produce the BLUP for genetic values.

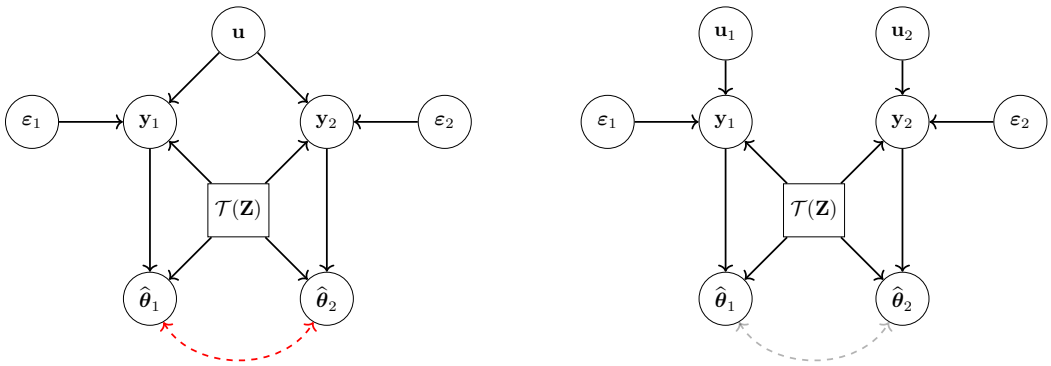

Figure 5: The dependence structure of variables in equation (11). The subscripts $\cdot_1$ and $\cdot_2$ denote indices of two non-overlapping groups of individuals. (A) The genetic effect $\mathbf{u}$ is shared by the two groups due to shared mutational history. Subsequently, the estimates from the two groups are correlated (red arrow). (B) One assumes that the two groups have their own separate mutational histories $\mathbf{u}_1$ and $\mathbf{u}_2$. As a result, the estimates are independent (grey arrow). Rectangle around the variables means that the variable is fixed (conditioned on).

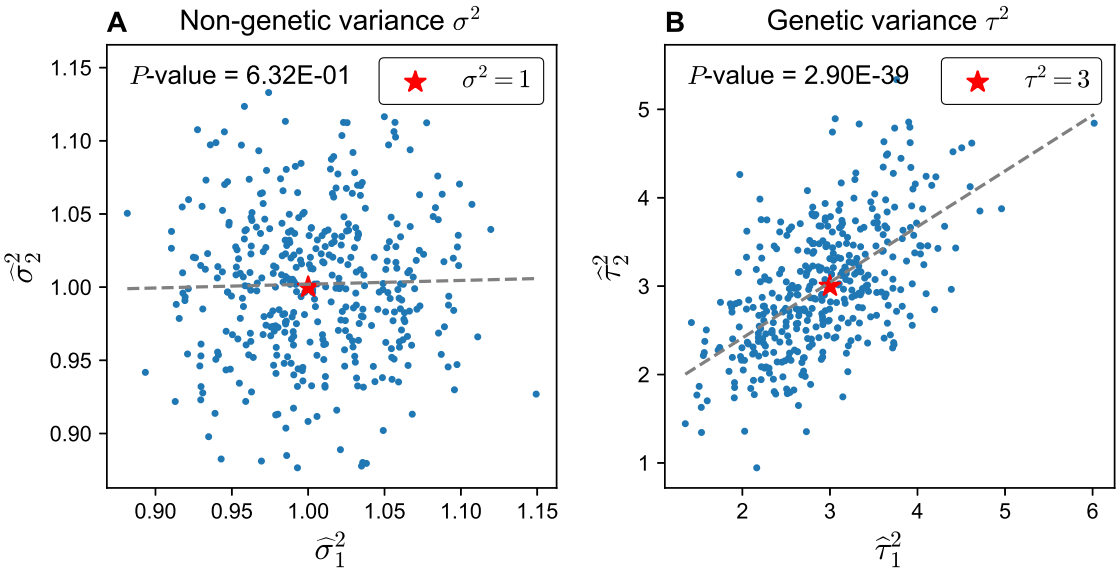

Figure 6: The correlation between variance component estimates. (A) Estimates of σ^2 from two non-overlapping sets of individuals over 400 simulations runs. (B) Estimates of τ^2 from two non-overlapping individuals over 400 simulation runs. The subscripts indicate the non-overlapping groups. The true simulation parameters are $\sigma^2 = 1$ and $\tau^2 = 3$.

Conditioning on the pedigree fixes the mating process while leaves all subsequent processes, such as recombination and mutation, to vary. The approach we take in this paper, which fixes the ARG, further conditions on recombination and genetic drift. The relatedness matrix conditioned on the ARG, the branch GRM, encodes more information about the trait. To test this speculation, we compared the prediction accuracy of ARG-LMM and pedigree-based LMM. Three mixed models, based on the true ARG, an inferred ARG (`tsinfer+tsdate`), and the pedigree, respectively, were compared. ARG-based predictions matched the true genetic value better than the pedigree-based prediction (**Figure 7**).

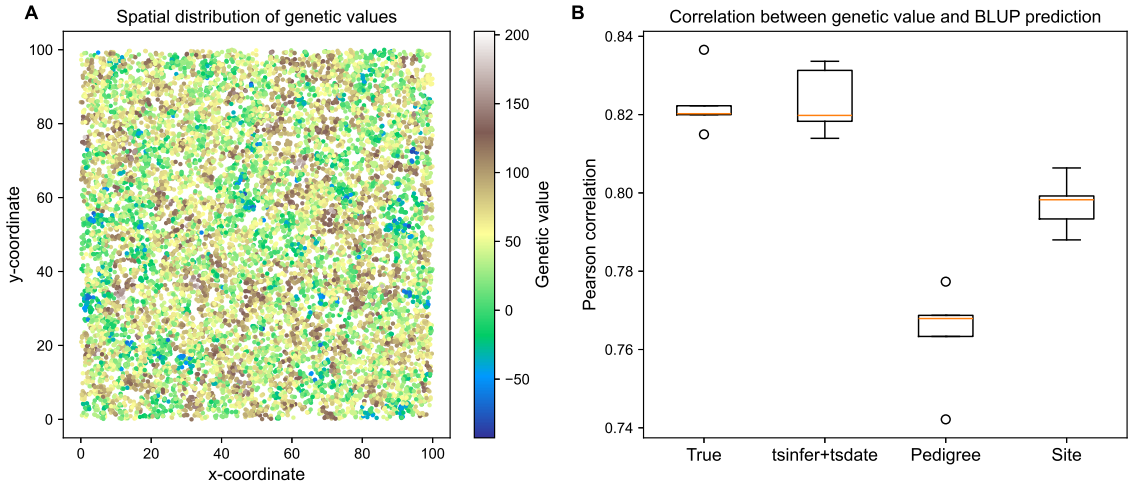

Figure 7: ARG-based versus pedigree-based predictions in a spatially structure population. (A) The distribution of genetic values of individuals on a 2-dimensional plane. Adjacent individuals have similar genetic values, exhibiting spatial autocorrelation. (B) The Pearson correlation between BLUP predictions of the methods and the true genetic value was plotted over 5 replicates. The x-axis labels refer to the type of GRMs. The ARG-based predictions were computed by `tslmm` with true and inferred tree sequences.

The flowchart in **Figure 8** illustrates how a trait is generated from a sequence of random processes. At each step, some amount of random noise is injected due to the stochastic nature of the genetic processes. Thus, it is easier to recover one entity from an adjacent step than more distant ones. For example, Fan et al. [2022] showed that principal components of the branch GRM can accurately recover population structure. Lehmann et al. [2025] found that the pedigree GRM reflects the structure more clearly than the branch GRM. This follows from the flowchart because the pedigree is closer to the underlying demographic structure than the ARG. The opposite holds for trait prediction. The ARG-based trait prediction is superior to the pedigree-based one because the ARG is closer to the trait than the pedigree.

4 Discussion

We studied how complex traits are conditionally distributed on the sample's ancestral recombination graph. This generative model turns out to be a linear mixed model, unveiling the connection between ARGs and linear mixed models. We show that random effects that were previously thought to be SNP effects can be interpreted as the aggregate effect of all potential variants on an edge, and the randomness is driven by

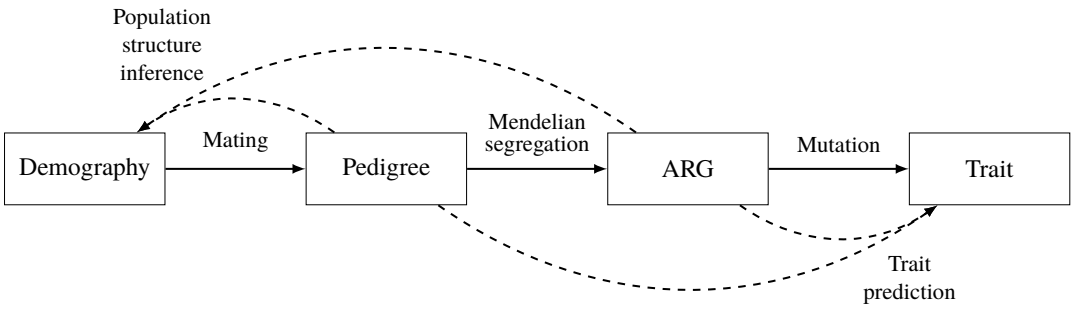

Figure 8: The order in which the genetic processes materialize in a population of mating individuals. The pedigree emerges from the demography through mating, the ARG realizes from the pedigree via Mendelian segregation, and the trait is determined by mutations landing on the ARG.

the mutational process running on the trees, providing a population genetic basis for the random effects assumption. This led to profound implications on heritability, genomic prediction, and pseudoreplication in genetic studies. We developed an efficient algorithm to fit ARG-LMM to simulated and inferred tree sequences in `tslmm` and demonstrated its statistical and computational performance, along with simulations that confirm the theoretical predictions of ARG-LMM with the new software. In this discussion, we outline important problems in quantitative genetics that are potentially related to ARG-LMM and directions for future research.

ARG-LMM might explain earlier studies that reported the relationship between natural selection and confounding in genome-wide association studies (GWAS). The concerns around confounding and false discoveries caused by population structure in genome-wide association studies led to the development of numerous methods such as genomic control, principal component (PC) adjustment, and linear mixed models [Devlin and Roeder, 1999, Price et al., 2006, Kang et al., 2010]. Principal component adjustment and linear mixed models continue to be popular in literature, and recently proposed methods tend to include the elements of both approaches [Loh et al., 2015, Jiang et al., 2019].

Early discussions centered on determining the most suitable method for adjusting population structure [Price et al., 2010, Sul and Eskin, 2013]. Shortly after, people started to recognize the close relationship between the two methods, as both methods used the same genetic relatedness matrix [Hoffman, 2013]. Linear mixed model includes the entire relatedness matrix, and PC adjustment takes the top eigenvectors of the relatedness matrix after eigendecomposition. For this reason, some authors argued that PC adjustment approximates the linear mixed model [Hoffman, 2013, Yao and Ochoa, 2023]. A recent analysis shows that this claim is only partially correct because the linear mixed model and PC adjustment weigh the eigenvectors differently [Schraiber et al., 2024]. These analyses are based on the eigenspectrum of the genetic relatedness matrix.

An alternative viewpoint comes from interpreting genetic effects as either fixed or random. It is common to test a single SNP at a time, and other SNPs that may be potentially associated with the trait are omitted from the regression [Sul et al., 2018]. If an omitted causal variant is correlated with the tested variant, the tested variant appears to be associated with the trait even in the absence of true causation [Vilhjálmsson and Nordborg, 2012, Sul et al., 2018, Veller and Coop, 2024]. Thus, one must account for the omitted variants in a univariate regression. PC adjustment can be thought of as including the omitted variants in the regression by reducing their dimension via principal component analysis [Hastie et al., 2009]. PC covariates produced from dimension reduction are included as fixed effects in the regression. Liner mixed models integrate over

the omitted variable by treating them as random variables [Sul et al., 2018].

ARG-LMM offers a more population genetic explanation. We have seen earlier in equation (9) that the genetic component of fixed effects is trivial, i.e., it can be accounted solely by an intercept, if the evolution is neutral. Therefore, by contraposition, non-neutral forces such as natural selection may create non-trivial fixed effects. It is currently unclear how exactly $\mathbf{Zf}$ will look like under varying scenarios of selection. Also, the sheer dimensionality of $\mathbf{f}$ makes it impossible to fit a full model, including $\mathbf{Z}$. A possible solution is to reduce the dimension of $\mathbf{Z}$ with PCA and use it instead of the full $\mathbf{Z}$. Since $\mathbf{Z}$ is the superset of the genotype matrix, this is equivalent to fitting the model with PC covariates derived from the genotype matrix. Previous simulation studies have found that linear mixed models without PC covariates are intact in scenarios without selection [Price et al., 2010]. Equation (9) theoretically supports this result. On the other hand, it produces an excess amount of false discoveries when selection is present, being in line with our conjecture that fixed effects comes from non-neutral forces [Sul and Eskin, 2013, Price et al., 2013]. It would be interesting to more thoroughly investigate the behavior of ARG-LMM under selection in future works.

The time zero of evolutionary change must be specified in any evolutionary analysis. For example, the notion of identity-by-descent (IBD) and relatedness are valid given a particular reference time point because any genetic materials will eventually coalesce if we go sufficiently further back in time [Thompson, 2013]. The idea of fixing a time point in which the evolutionary process began has a long history that dates back to Wright [1949] and has been expanded in subsequent studies in the context of quantitative genetics [Weir, 1990, Weir and Hill, 2002, Ochoa and Storey, 2021, Hou and Ochoa, 2023]. Choosing this time point is also important in ARG-LMM's generative model. The current paper makes an implicit choice of which we condition at the time of the oldest root node in the tree sequence, i.e., the grand TMRCA. However, one could choose another time point below the grand TMRCA. All mutations that occurred above this time point but below the grand TMRCA will be conditioned and become fixed effects. These effects are not constant across individuals even if the evolution is neutral. If this time is set below the grand TMRCA, one should include all the genetic effects above this time as fixed effects in the regression model. The relatedness driven by the random effects will decay as the reference time gets closer to the present and will eventually vanish at the present. This reiterates the previous arguments that relatedness is a relative measure with respect to a reference time point that is set by the researcher, and no single measure of relatedness exists in an absolute sense [Thompson, 2013]. We plan to add an option in `tslmm` to set this time point and automatically handle the evolution of relatedness and fixed effects as a function of time zero.

From a theoretical side, we studied the distribution of complex traits conditioned on the ARG. This fixes all evolutionary processes except for mutation [Ralph, 2019, Ralph et al., 2020]. One could choose to characterize how traits are distributed by conditioning less [Veller and Coop, 2024], or from a forwards-in-time perspective unlike the coalescent and the ARG [Simons et al., 2018, 2022]. For example, one could condition on the pedigree of a sample, thereby allowing recombination and Mendelian segregation to vary. One could condition even less, like fixing only the macroscopic demographic events [Spence et al., 2023]. All these choices will lead to different probabilistic characterizations of complex traits. Thus, ARG-LMM is a powerful, but not the only approach to study complex traits from a population genetic perspective.

We list a number of future improvements for `tslmm`. `tslmm` is currently limited to continuous traits, so adding support for dichotomous traits will be valuable, as many diseases are recorded that way. Edge and Coop [2018] reconstructed temporal trajectories of the mean polygenic score of traits using inferred ARGs. A follow-up study evaluated a wider range of ARG inference methods for this task [Peng et al., 2024]. Currently, `tslmm` can compute the genetic values of non-sample nodes, but averaging over all existing lineages needs a new algorithm. Along side with the genetic component parameter τ_e^2, one can immediately extend the definition to incorporate the notion of genetic correlation of a pair of traits [Lee et al., 2012, Vattikuti et al.,

2012]. If $\mathbf{u}_1$ and $\mathbf{u}_2$ are the edge effects of the two traits, the covariance is

$$\mathrm{Cov}\left(\mathbf{u}_1, \mathbf{u}_2\right) = l_e s_e \cdot \frac{1}{s_e} \sum_{p:p \in e} \boldsymbol{\beta}_{1p} \boldsymbol{\beta}_{2p} u_{ep} \tag{44}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the effect sizes of the positions for the two traits respectively. Finally, we plan to utilize the powerful trait simulator to develop a simulation-based inference framework for complex traits. This will allow us to fit larger models with multiple variance components in an amortized fashion, enabling ultra-fast processing of biobank datasets [Cranmer et al., 2020].

5 Data availability

`tslmm` is a Python package available at `https://github.com/hanbin973/tslmm`. The results and the figures of the paper can be reproduced using the scripts at `https://github.com/hanbin973/tslmm_paper`.

6 Acknowledgments

An early idea of the paper was discussed with Doc Edge (University of Southern California). Yan Wong (University of Oxford) and Jonathan Terhorst (University of Michigan) read a draft and gave helpful comments and suggestions. We thank the participants of the Probabilistic Modeling in Genomics 2025 conference for valuable questions and discussions. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

7 Funding

HL acknowledges the support from the Department of Statistics, University of Michigan, Ann Arbor through the the Departmental Fellowship. NP, JK, and PR acknowledge support from the NIH NHGRI (research grant HG012473). GG acknowledges support from the BBSRC ISP grant to The Roslin Institute (BBS/E/D/30002275, BBS/E/RL/230001A, and BBS/E/RL/230001C), BBSRC research grant BB/T014067/1, and the NRC research grant 346741.

8 Conflicts of interest

The author(s) declare no conflicts of interest.

References

L. Anderson-Trocmé, D. Nelson, S. Zabad, A. Diaz-Papkovich, I. Kryukov, N. Baya, M. Touvier, B. Jeffery, C. Dina, H. Vézina, et al. On the genes, genealogies, and geographies of quebec. *Science*, 380(6647): 849–855, 2023.

D. Y. C. Brandt, C. D. Huber, C. W. K. Chiang, and D. Ortega-Del Vecchyo. The promise of inferring the past using the Ancestral Recombination Graph (ARG). *Genome Biology and Evolution*, page evae005, 2024. doi: 10.1093/gbe/evae005. URL https://doi.org/10.1093/gbe/evae005.

S. R. Browning, B. L. Browning, M. L. Daviglus, R. A. Durazo-Arvizu, N. Schneiderman, R. C. Kaplan, and C. C. Laurie. Ancestry-specific recent effective population size in the americas. *PLOS Genetics*, 14(5): e1007385, May 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007385. URL http://dx.doi.org/10.1371/journal.pgen.1007385.

E. T. Chevy, J. Min, V. Caudill, S. E. Champer, B. C. Haller, C. T. Rehmann, C. C. Smith, S. Tittes, P. W. Messer, A. D. Kern, et al. Population genetics meets ecology: A guide to individual-based simulations in continuous landscapes. *Ecology and Evolution*, 15(4):e71098, 2025.

G. Covarrubias-Pazaran. lme4breeding: Relationship-based mixed-effects models, May 2024. URL http://dx.doi.org/10.32614/CRAN.package.lme4breeding.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, May 2020. ISSN 1091-6490. doi: 10.1073/pnas.1912789117. URL http://dx.doi.org/10.1073/pnas.1912789117.

Y. Deng, R. Nielsen, and Y. S. Song. Robust and accurate bayesian inference of genome-wide genealogies for large samples. *bioRxiv*, Mar. 2024. doi: 10.1101/2024.03.16.585351. URL http://dx.doi.org/10.1101/2024.03.16.585351.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec. 1999. ISSN 1541-0420. doi: 10.1111/j.0006-341x.1999.00997.x. URL http://dx.doi.org/10.1111/j.0006-341x.1999.00997.x.

Y. Ding, K. Hou, Z. Xu, A. Pimplaskar, E. Petter, K. Boulier, F. Privé, B. J. Vilhjálmsson, L. M. Olde Loohuis, and B. Pasaniuc. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618 (7966):774–781, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06079-4. URL http://dx.doi.org/10.1038/s41586-023-06079-4.

R. Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996. ISBN 0-534-24318-5.

R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, New York, NY, 2nd edition, 2008. ISBN 978-0387781686.

M. D. Edge and G. Coop. Reconstructing the history of polygenic scores using coalescent trees. *Genetics*, 211(1):235–262, Nov. 2018. ISSN 1943-2631. doi: 10.1534/genetics.118.301687. URL http://dx.doi.org/10.1534/genetics.118.301687.

E. N. Epperly, J. A. Tropp, and R. J. Webber. Xtrace: Making the most of every sample in stochastic trace estimation. *SIAM Journal on Matrix Analysis and Applications*, 45(1):1–23, Jan. 2024. ISSN 1095-7162. doi: 10.1137/23m1548323. URL http://dx.doi.org/10.1137/23M1548323.

D. S. Falconer. *Introduction to Quantitative Genetics*. Longman, 4th edition, 1995.

C. Fan, N. Mancuso, and C. W. K. Chiang. A genealogical estimate of genetic relationships. *The American Journal of Human Genetics*, 109(5):812–824, 2022. doi: 10.1016/j.ajhg.2022.03.016. URL https://www.sciencedirect.com/science/article/pii/S0002929722001124.

Z. Frangella, J. A. Tropp, and M. Udell. Randomized nyström preconditioning, 2021. URL https://arxiv.org/abs/2110.02820.

A. R. Gilmour, R. Thompson, and B. R. Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51(4):1440, Dec. 1995. ISSN 0006-341X. doi: 10.2307/2533274. URL http://dx.doi.org/10.2307/2533274.

R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In *Progress in population genetics and human evolution (Minneapolis, MN, 1994)*, volume 87 of *IMA Vol. Math. Appl.*, pages 257–270. Springer, New York, 1997. URL http://www.math.canterbury.ac.nz/~r.sainudiin/recomb/ima.pdf.

B. C. Haller and P. W. Messer. Slim 4: multispecies eco-evolutionary modeling. *The American Naturalist*, 201(5):E127–E139, 2023.

K. Harris. Using enormous genealogies to map causal variants in space and time. *Nature Genetics*, 55:730–731, 2023. doi: 10.1038/s41588-023-01389-9. URL https://doi.org/10.1038/s41588-023-01389-9.

D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, June 1977. ISSN 1537-274X. doi: 10.1080/01621459.1977.10480998. URL http://dx.doi.org/10.1080/01621459.1977.10480998.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2nd ed. edition, 2009. ISBN 978-0-387-84857-0; 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7.

G. E. Hoffman. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS ONE*, 8(10):e75707, Oct. 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0075707. URL http://dx.doi.org/10.1371/journal.pone.0075707.

Z. Hou and A. Ochoa. Genetic association models are robust to common population kinship estimation biases. *GENETICS*, 224(1), Feb. 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad030. URL http://dx.doi.org/10.1093/genetics/iyad030.

R. Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1): 44, 1990. URL http://web.eve.ucdavis.edu/pbg298/pdfs/Hudson_OxfordSurveysEvolBiol_1991.pdf.

L. Jiang, Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755, Nov. 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0530-8. URL http://dx.doi.org/10.1038/s41588-019-0530-8.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, Mar. 2010. ISSN 1546-1718. doi: 10.1038/ng.548. URL http://dx.doi.org/10.1038/ng.548.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5):e1004842, May 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004842. URL http://dx.doi.org/10.1371/journal.pcbi.1004842.

S. Lee, J. Yang, M. Goddard, P. Visscher, and N. Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, July 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts474. URL http://dx.doi.org/10.1093/bioinformatics/bts474.

B. Lehmann, H. Lee, L. Anderson-Trocmé, J. Kelleher, G. Gorjanc, and P. L. Ralph. On args, pedigrees, and genetic relatedness matrices. Mar. 2025. doi: 10.1101/2025.03.03.641310. URL http://dx.doi.org/10.1101/2025.03.03.641310.

A. L. Lewanski, M. C. Grundler, and G. S. Bradburd. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *PLOS Genetics*, 20(1): 1–24, 2024. doi: 10.1371/journal.pgen.1011110. URL https://doi.org/10.1371/journal.pgen.1011110.

V. Link, J. G. Schraiber, C. Fan, B. Dinh, N. Mancuso, C. W. K. Chiang, and M. D. Edge. Tree-based QTL mapping with expected local genetic relatedness matrices. *The American Journal of Human Genetics*, 110 (12):2077–2091, 2023. doi: 10.1016/j.ajhg.2023.10.017. URL https://www.sciencedirect.com/science/article/pii/S0002929723003956.

P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, Feb. 2015. ISSN 1546-1718. doi: 10.1038/ng.3190. URL http://dx.doi.org/10.1038/ng.3190.

A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, Mar. 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0379-x. URL http://dx.doi.org/10.1038/s41588-019-0379-x.

A. Ochoa and J. D. Storey. Estimating fst and kinship for arbitrary population structures. *PLOS Genetics*, 17 (1):e1009241, Jan. 2021. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009241. URL http://dx.doi.org/10.1371/journal.pgen.1009241.

D. Peng, O. J. Mulder, and M. D. Edge. Evaluating arg-estimation methods in the context of estimating population-mean polygenic score histories. *bioRxiv*, 2024. doi: 10.1101/2024.05.24.595829. URL https://www.biorxiv.org/content/early/2024/12/20/2024.05.24.595829.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006. ISSN 1546-1718. doi: 10.1038/ng1847. URL http://dx.doi.org/10.1038/ng1847.

A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, June 2010. ISSN 1471-0064. doi: 10.1038/nrg2813. URL http://dx.doi.org/10.1038/nrg2813.

A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. Response to sul and eskin. *Nature Reviews Genetics*, 14(4):300–300, Feb. 2013. ISSN 1471-0064. doi: 10.1038/nrg2813-c2. URL http://dx.doi.org/10.1038/nrg2813-c2.

P. Ralph, K. Thornton, and J. Kelleher. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics*, 215(3):779–797, 2020. doi: 10.1534/genetics.120.303253. URL https://doi.org/10.1534/genetics.120.303253.

P. L. Ralph. An empirical approach to demographic inference with genomic data. *Theoretical Population Biology*, 127:91–101, June 2019. ISSN 0040-5809. doi: 10.1016/j.tpb.2019.03.005. URL http://dx.doi.org/10.1016/j.tpb.2019.03.005.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, May 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004342. URL http://dx.doi.org/10.1371/journal.pgen.1004342.

N. A. Rosenberg and J. M. VanLiere. Replication of genetic associations as pseudoreplication due to shared genealogy. *Genetic Epidemiology*, 33(6):479–487, Feb. 2009. ISSN 1098-2272. doi: 10.1002/gepi.20400. URL http://dx.doi.org/10.1002/gepi.20400.

P. Salehi Nowbandegani, A. W. Wohns, J. L. Ballard, E. S. Lander, A. Bloemendal, B. M. Neale, and L. J. O'Connor. Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. *Nature Genetics*, 55(9):1494–1502, 2023.

M. Satoh. Methods of calculating prediction error variance and prediction accuracy for restricted best linear unbiased prediction of breeding values. *Journal of Animal Breeding and Genetics*, 142(4):363–372, Nov. 2024. ISSN 1439-0388. doi: 10.1111/jbg.12910. URL http://dx.doi.org/10.1111/jbg.12910.

J. G. Schraiber, M. D. Edge, and M. Pennell. Unifying approaches from statistical genetics and phylogenetics for mapping phenotypes in structured populations. *PLOS Biology*, 22(10):e3002847, Oct. 2024. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002847. URL http://dx.doi.org/10.1371/journal.pbio.3002847.

Y. B. Simons, K. Bullaughey, R. R. Hudson, and G. Sella. A population genetic interpretation of gwas findings for human quantitative traits. *PLOS Biology*, 16(3):e2002985, Mar. 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2002985. URL http://dx.doi.org/10.1371/journal.pbio.2002985.

Y. B. Simons, H. Mostafavi, C. J. Smith, J. K. Pritchard, and G. Sella. Simple scaling laws control the genetic architectures of human complex traits. *bioRxiv*, Oct. 2022. doi: 10.1101/2022.10.04.509926. URL http://dx.doi.org/10.1101/2022.10.04.509926.

D. Speed and D. J. Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015. doi: 10.1038/nrg3821. URL https://www.nature.com/articles/nrg3821.

D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, Dec. 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2012.10.010. URL http://dx.doi.org/10.1016/j.ajhg.2012.10.010.

L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019. doi: 10.1038/s41588-019-0484-x. URL http://dx.doi.org/10.1038/s41588-019-0484-x.

J. P. Spence, T. Zeng, H. Mostafavi, and J. K. Pritchard. Scaling the discrete-time wright–fisher model to biobank-scale datasets. *GENETICS*, 225(3), Sept. 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad168. URL http://dx.doi.org/10.1093/genetics/iyad168.

J. H. Sul and E. Eskin. Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics*, page 300–300, Feb. 2013. doi: 10.1038/nrg2813-c1. URL http://dx.doi.org/10.1038/nrg2813-c1.

J. H. Sul, L. S. Martin, and E. Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLOS Genetics*, 14(12):e1007309, Dec. 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007309. URL http://dx.doi.org/10.1371/journal.pgen.1007309.

D. Tagami, G. Bisschop, and J. Kelleher. <tt>tstrait</tt>: a quantitative trait simulator for ancestral recombination graphs. *Bioinformatics*, 40(6), May 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae334. URL http://dx.doi.org/10.1093/bioinformatics/btae334.

E. A. Thompson. Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, June 2013. ISSN 1943-2631. doi: 10.1534/genetics.112.148825. URL http://dx.doi.org/10.1534/genetics.112.148825.

S. Vattikuti, J. Guo, and C. C. Chow. Heritability and genetic correlations explained by common snps for metabolic syndrome traits. *PLoS Genetics*, 8(3):e1002637, Mar. 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002637. URL http://dx.doi.org/10.1371/journal.pgen.1002637.

A. I. Vazquez, D. Bates, S. Avadhanam, and P. Perez Rodriguez. pedigreetools: Versatile functions for working with pedigrees, Dec. 2018. URL http://dx.doi.org/10.32614/CRAN.package.pedigreeTools.

C. Veller and G. M. Coop. Interpreting population- and family-based genome-wide association studies in the presence of confounding. *PLOS Biology*, 22(4):e3002511, Apr. 2024. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002511. URL http://dx.doi.org/10.1371/journal.pbio.3002511.

B. J. Vilhjálmsson and M. Nordborg. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2, Nov. 2012. ISSN 1471-0064. doi: 10.1038/nrg3382. URL http://dx.doi.org/10.1038/nrg3382.

B. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, T. Hayeck, H.-H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson, A. L. Price, S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. DeLisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodrguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, J. Grove, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lnnqvist,

M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, P. B. Mortensen, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Mller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S.-Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H.-C. So, C. C. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Sderman, S. Thirumalai, D. Toncheva, P. A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, J. Q. Wu, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St. Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O'Donovan, P. Kraft, D. J. Hunter, M. Adank, H. Ahsan, K. Aittomäki, L. Baglietto, S. Berndt, C. Blomquist, F. Canzian, J. Chang-Claude, S. J. Chanock, L. Crisponi, K. Czene, N. Dahmen, I. d. S. Silva, D. Easton, A. H. Eliassen, J. Figueroa, O. Fletcher, M. Garcia-Closas, M. M. Gaudet, L. Gibson, C. A. Haiman, P. Hall, A. Hazra, R. Hein, B. E. Henderson, A. Hofman, J. L. Hopper, A. Irwanto, M. Johansson, R. Kaaks, M. G. Kibriya, P. Lichtner, S. Lindström, J. Liu, E. Lund, E. Makalic, A. Meindl, H. Meijers-Heijboer, B. Müller-Myhsok, T. A. Muranen, H. Nevanlinna, P. H. Peeters, J. Peto, R. L. Prentice, N. Rahman, M. J. Sánchez, D. F. Schmidt, R. K. Schmutzler, M. C. Southey, R. Tamimi, R. Travis, C. Turnbull, A. G. Uitterlinden, R. B. van der Luijt, Q. Waisfisz, Z. Wang, A. S. Whittemore, R. Yang, and W. Zheng. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592, Oct. 2015. ISSN 0002-9297. doi: 10.1016/j.ajhg.2015.09.001. URL `http://dx.doi.org/10.1016/j.ajhg.2015.09.001`.

J. Wakeley. *Coalescent Theory*. Roberts & Company, Greenwood Village, CO, Aug. 2008.

B. S. Weir. *Genetic Data Analysis*. Sinauer Associates, Sunderland, MA, June 1990.

B. S. Weir and W. G. Hill. Estimating f-statistics. *Annual Review of Genetics*, 36(1):721–750, Dec. 2002. ISSN 1545-2948. doi: 10.1146/annurev.genet.36.050802.093940. URL `http://dx.doi.org/10.1146/annurev.genet.36.050802.093940`.

Y. Wong, A. Ignatieva, J. Koskela, G. Gorjanc, A. W. Wohns, and J. Kelleher. A general and efficient representation of ancestral recombination graphs. *bioRxiv*, 2023. doi: 10.1101/2023.11.03.565466. URL `https://www.biorxiv.org/content/early/2023/11/04/2023.11.03.565466`.

S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, Jan. 1949. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1949.tb02451.x. URL `http://dx.doi.org/10.1111/j.1469-1809.1949.tb02451.x`.

Y. Wu and S. Sankararaman. A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics*, 34(13):i187–i194, June 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty253. URL http://dx.doi.org/10.1093/bioinformatics/bty253.

J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, June 2010. ISSN 1546-1718. doi: 10.1038/ng.608. URL http://dx.doi.org/10.1038/ng.608.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, Jan. 2011. ISSN 0002-9297. doi: 10.1016/j.ajhg.2010.11.011. URL http://dx.doi.org/10.1016/j.ajhg.2010.11.011.

Y. Yao and A. Ochoa. Limitations of principal components in quantitative genetic association models for human studies. *eLife*, 12, May 2023. ISSN 2050-084X. doi: 10.7554/elife.79238. URL http://dx.doi.org/10.7554/eLife.79238.

B. C. Zhang, A. Biddanda, A. F. Gunnarsson, F. Cooper, and P. F. Palamara. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, 55(5):768–776, 2023. doi: 10.1038/s41588-023-01379-x. URL https://doi.org/10.1038/s41588-023-01379-x.

G. Zhou and H. Zhao. A fast and robust bayesian nonparametric method for prediction of complex traits using summary statistics. *PLOS Genetics*, 17(7):e1009697, July 2021. ISSN 1553-7404. doi: 10.1371/journal.pgen.1009697. URL http://dx.doi.org/10.1371/journal.pgen.1009697.

J. Zhu, G. Kalantzis, A. Pazokitoroudi, A. F. Gunnarsson, H. Loya, H. Chen, S. Sankararaman, and P. F. Palamara. Fast variance component analysis using large-scale ancestral recombination graphs. *BioRxiv*, Aug. 2024. doi: 10.1101/2024.08.31.610262. URL http://dx.doi.org/10.1101/2024.08.31.610262.