# Genome-wide association with uncertainty in the genetic similarity matrix

Shijia Wang[1†] Shufei Ge[2†] Benjamin Sobkowiak[5]
Liangliang Wang[3] Louis Grandjean[4] Caroline Colijn[5]
Lloyd T Elliott[3,*]

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

[2]Institute of Mathematical Sciences, ShanghaiTech University, China

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

[4]Department of Infectious Diseases, University College London, United Kingdom

[5]Department of Mathematics, Simon Fraser University, Canada
† indicates equal contribution.

**Abstract**

Genome-wide association studies are often confounded by population stratification and structure. Linear mixed models (LMMs) are a powerful class of methods for uncovering genetic effects, while controlling for such confounding. LMMs include random effects for a genetic similarity matrix, and they assume that a true genetic similarity matrix is known. However, uncertainty about the phylogenetic structure of a study population may degrade the quality of LMM results. This may happen in bacterial studies in which the number of samples or loci are small, or in studies with low quality genotyping. In this work, we develop methods for linear mixed models in which the genetic similarity matrix is unknown and is derived from MCMC estimates of the phylogeny. We apply our model to a genome-wide association study of multidrug-resistance in tuberculosis, and illustrate our methods on simulated data.

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

# 1 Introduction

Genome-wide association studies (GWASs) are designed to identify the genetic variants affecting phenotypes of interest such as multidrug-resistance in tuberculosis (Price et al., 2006; Zhang et al., 2010). Classic approaches to GWAS rely on linear association tests to quantify the relationship between phenotypes and genotypes.

Population structure (Patterson et al., 2006) in the phylogeny of bacterial genomes can lead to false positives, spurious associations, or inflated $p$-values (Novembre et al., 2008). The genealogy of tuberculosis typically exhibits strong clade structure (Cordero and Polz, 2014; Earle et al., 2016), with geographically widespread lineages, and so GWASs on TB are vulnerable to population stratification.

Linear mixed models (LMMs) use the genetic similarity among samples as a random effect. This controls for confounding from population structure, leading to improved false discovery rates. In Kang et al. (2010), the EMMAX (Efficient mixed-model association expedited) model was proposed, which computes the variance component in linear mixed models in an efficient way. In addition, factored spectrally transformed linear mixed models (FaST-LMMs) were introduced (Lippert et al., 2011; Listgarten et al., 2013), with running time and memory costs that scale linearly in the cohort size. In Dahl et al. (2016); Yang et al. (2011); Zhou and Stephens (2014), models were developed for computationally efficient linear mixed effects model with multivariate phenotypes. The efficiency of FaST-LMM methods have been further improved by subsetting the genetic variants examined, so that a set of maximally independent genetic variants are considered (Listgarten et al., 2012). Several other methods have been proposed to scale to large cohorts (such as UK Biobank; Bycroft et al. 2018, Sudlow et al. 2015). Loh et al. (2015) developed an efficient Bayesian mixed model, BOLT-LMM, that requires lower computational costs than standard LMMs, while increasing power by modeling genetic architectures via a Bayesian mixture prior on marker effect sizes. Loh et al. (2018) also proposed a much faster version of BOLT-LMM and demonstrate the method by analyzing the UK Biobank data. Jiang et al. (2021) developed a generalized linear mixed model (GLMM)-based methods for genome-wide association studies (fastGWA-GLMM) for binary phenotypes. The method is scalable to cohorts with millions of individuals.

All the above LMM methods assume that the matrix specifying the genetic similarity among the samples is known (*i.e.*, through an empirical genetic similarity matrix Patterson et al. 2006; or a kinship matrix derived from a pedigree Kirkpatrick et al. 2019). For large cohorts of human genotypes, there is often low uncertainty about estimated genetic similarity matrices. However, for some studies, such as bacterial studies, in which small numbers of samples or loci are present, or for studies in which genotyping is sparse and noisy, uncertainty about the genetic similarity matrix may degrade the quality of LMM results (for example, in S. Wang et al. 2021 heritability estimates based on genetic similarity matrices were found to have large variance, which may translate into reduced power for LMMs conducted with point estimates of the genetic similarity matrix).

In the *pyseer* (Lees et al., 2018) package, a few methods for GWAS are implemented. For example, a fixed effects model using the genetic similarity matrix represented by a multiple-dimensional scaling approximation (MDS); a linear mixed model using a kinship matrix; a whole genome model using elastic net. However, the genetic similarity matrix represented by MDS is not equal to its expectation (Patterson et al., 2006) and is biased (S. Wang et al., 2021).

Multidrug-resistant tuberculosis (MDR-TB) is a major concern for tuberculosis control (Grandjean et al., 2017). Multidrug-resistance in TB is caused by genetic variations in genes that encode drug targets and drug-converting enzymes (Coll et al., 2014). Understanding these effects is critical for improving treatment for MDR-TB patients. But population stratification (in which genetic variates correlate with structure in geographical or socioeconomic indicators), and noisy genotyping of bacterial genomes confounds such studies (Price et al., 2006; Zhang et al., 2010). In this work, we improve the control provided by linear mixed models by encoding uncertainty about genetic similarity, and report applications on TB data and simulated data.

We propose a new LMM method for genome-wide association studies, using phylogenetic trees to control for population structure. We use MCMC (Markov chain Monte Carlo) to draw samples for the phylogeny based on observed genetic sequences, and then we compute the expected genetic similarity matrix induced by each phylogeny (S. Wang et al., 2021). We then apply the linear mixed model to each sampled expected genetic similarity matrix and average the results. Simulations show that the true positive rates and false discovery rates of our method outperform both existing linear regression methods, LMM methods in which the genetic similarity matrix is estimated empirically, and *pyseer* with the genetic similarity matrix represented using consensus tree of MCMC posterior samples. We apply this method to MDR-TB in a GWAS of 467 TB subjects in a population from Lima, Peru (Grandjean et al., 2017).

# 2 Methods

## 2.1 Linear mixed effects models for genome-wide association studies

We consider a population of samples typed at given SNPs (single nucleotide polymorphisms) and with measured phenotypes. We begin this section with an exposition of the linear mixed model (Kang et al., 2010; Lippert et al., 2011). Let the study subject indices be $i = 1, 2, \ldots, N$, and let the SNP locations be indexed by $m = 1, 2, \ldots, M$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ denote a column vector of phenotypes ($y_i \in \mathbb{R}$), and let $\mathbf{G} = [G_1, G_2, \ldots, G_M]$ denote genotype data observed at the $M$ SNPs, with $G_m$ denoting a column vector of alleles for the $m$-th SNP for all $N$ subjects. For details on bacterial genetics we refer readers to Earle et al. (2016) and Coll et al. (2014). Let $G_{im} = 0$ and $G_{im} = 1$ encode the events that subject $i$ has the major allele or the minor allele at SNP $m$ respectively.

The LMM is a mixed effects model for association between SNP $G_m$ and the phenotype. Independent LMMs may be applied at each SNP as follows:

$$\mathbf{y} = G_m \beta_m + \mathbf{b}_m + \boldsymbol{\varepsilon}_m. \tag{1}$$

Here $\beta_m$ is the effect size of the fixed effect of the $m$-th SNP, $\boldsymbol{\varepsilon}_m$ is the random error vector, with $\boldsymbol{\varepsilon}_m \sim \mathrm{MVN}(\mathbf{0}, \sigma_g^2 I)$, and $\mathbf{b}_m$ is the random effect of the $m$-th SNP, with $\mathbf{b}_m \sim \mathrm{MVN}(\mathbf{0}, \sigma_e^2 \boldsymbol{\psi})$, and $\mathrm{MVN}(0, \Sigma)$ is the multivariate normal distribution with mean 0 and covariance $\Sigma$. The genetic similarity matrix $\boldsymbol{\psi}$ measures the genetic relatedness among different subjects. This is an $N \times N$ positive semi-definite matrix, and an empirical estimate of $\boldsymbol{\psi}$ is given by Patterson et al. (2006):

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

Here $\mu_m = \frac{1}{N} \sum_{i=1}^{N} G_{im}$, $\sigma_m^2 = \mu_m(1 - \mu_m)$ are the empirical mean and variance (respectively) of the genotypes of the $N$ subjects at the $m$-th SNP.

While previous LMM work approximates $\boldsymbol{\psi}$ by (2), there is often uncertainty about the true value of $\boldsymbol{\psi}$. A realisation of $\boldsymbol{\psi}$ is implied deterministically by a phylogenetic tree for the $N$ subjects (S. Wang et al., 2021). We denote this tree by $t$. In the next subsection we introduce a *li*near *m*ixed model with *u*ncertain genetic similarity matrices (LiMU), in which the genetic similarity matrix is unknown and is estimated based on the genotypes.

## 2.2 The LiMU method

The covariance matrix of the random effects for the $m$-th SNP is the positive-definite matrix $\boldsymbol{\psi}$, which measures the genetic relatedness among individuals.

The empirical estimate of genetic similarity (shown in Equation 2) is inaccurate if genotypes are not densely sampled, or are of poor quality (S. Wang et al., 2021). The inaccuracy in empirical genetic similarity estimates may lead to inconsistent estimates for parameters in linear mixed models. In this work, we propose a new linear mixed model for multivariate genome-wide association studies, by assuming an unknown genetic similarity matrix $\boldsymbol{\psi}(t)$ that depends on the underlying phylogenetic tree $t$. Phylogenetics explicitly model a rate matrix, so branch lengths are likely to give a better estimate of genetic relatedness than the inner product of sequences used in existing software packages such as GEMMA (Zhou and Stephens, 2014).

Here we consider estimating the phylogeny $t$ in a Bayesian framework. We place a proper prior distribution on the phylogenetic tree $t$ (*i.e.*, a uniform clock prior for a binary clock tree). After specifying the prior distributions, trees can be sampled conditioned on genotype data using standard software packages for phylogenetic inference (*e.g.* MrBayes; Ronquist et al. 2012). When multiple posterior samples of the phylogeny $\{t_j\}_{j=1,\dots,J}$ are available (for example, after running MrBayes for the phylogeny), we use the algorithm proposed in S. Wang et al. (2021) to compute the expected genetic similarity matrix $\{\boldsymbol{\psi}(t_j)\}_{j=1,\dots,J}$ for each posterior sample. The resulting matrices represent the uncertainty of genetic similarities among species, and we combine them with linear mixed models to account for population stratification and correct for spurious associations.

We associate each posterior sample $\{\boldsymbol{\psi}(t_j)\}_{j=1,\dots,J}$ with a linear mixed model. For each $j$ and $m$, we use restricted maximum likelihood estimation (REML) (Corbeil and Searle, 1976) to estimate parameters in each LMM

$$\mathbf{y} = G_m\beta_{mj} + \mathbf{b}_{mj} + \boldsymbol{\varepsilon}_{mj}. \tag{3}$$

Here $\boldsymbol{\varepsilon}_{mj} \sim \mathrm{MVN}(\mathbf{0}, \sigma_g^2 I)$, and $\mathbf{b}_{mj}$ is the random effect of the $m$-th SNP, and $\boldsymbol{\psi}(t_j)$ is the expected genetic similarity matrix, $\mathbf{b}_{mj} \sim \mathrm{MVN}(\mathbf{0}, \sigma_e^2\boldsymbol{\psi}(t_j))$. We compute the $p$-value for $\hat{\beta}_{mj}$, denoted by $p_{mj} = P(T_m^{\text{rep}} > T_m|\boldsymbol{\psi}(t_j))$.

Here $T_m$ is the test statistic for site $m$, and it is a function of $\boldsymbol{\psi}(t)$. $T_m^{\text{rep}}$ denotes the test statistic for a replication of site $m$. Finally, we compute the mean of $p$-values $p_{mj}$ for each site $m$, $p_m^* = \frac{1}{J}\sum_{j=1}^{J} p_{mj}$. We note that $p_m^*$ is natural way to combine a set of p-values since it is an unbiased estimator of $\int P(T_m^{\text{rep}} > T_m | \boldsymbol{\psi}(t))\pi(\boldsymbol{\psi}(t))d\boldsymbol{\psi}(t)$. This is related to posterior predictive $p$-values (Hjort et al., 2006; Meng, 1994). A permutation test is another option for finding $p$-values for this test, and may be more precise than the mean, but we found that permutation tests are not computationally tractable with this model. Algorithm 1 provides an overview of the estimation procedure of LiMU. We provide an open source software implementation for this method[1].

**Algorithm 1 A linear mixed model with uncertain genetic similarity matrices for genome-wide association study**

```
Inputs: Phenotype $\mathbf{y}$ and genotype $\mathbf{G}$.
Output: Significantly associated genetic variants and posterior samples of p-value.
Run MrBayes (or related software) to obtain posterior samples of phylogenetic tree $\{t_j\}_{j=1,\dots,J}$ using $\mathbf{G}$.
Compute the genetic similarity matrix $\{\boldsymbol{\psi}(t_j)\}_{j=1,\dots,J}$ using the algorithm proposed in S. Wang et al. (2021).
for $j \in \{1, 2, \dots, J\}$ do
    for $m \in \{1, 2, \dots, M\}$ do
        Use REML to estimate parameters in Equation (3) with $\boldsymbol{\psi}(t_j)$ and compute the $p$-value for site $m$, $p_{mj}$
for $m \in \{1, 2, \dots, M\}$ do
    Compute adjusted $p$-value for each site $m$ using $p_m^* = \frac{1}{J}\sum_{j=1}^{J} p_{mj}$.
Select genetic variants with p-value lower than threshold.
return Significantly associated genetic variants and the estimated $p$-values.
```

The computational cost for an MCMC step for tree construction is a linear function of $N \cdot K$. Here $K$ is the number of MCMC samples. For each thinned posterior sample, the cost for computing the genetic similarity matrix

[1] https://github.com/shijiaw/LMMTree

is $O(N^2)$, and the cost for REML of LMM is $O(N^3 \cdot M)$. The walltime of these operations can be improved by parallelizing the REML step for each genetic similarity matrix computation.

# 3 Experiments

## 3.1 Simulation

### 3.1.1 Simulation 1

In the first simulation study, we simulated datasets for four scenarios: A, B, C and D. In scenario A, we simulated 50 trees with $N = 30$ taxa, each with $M = 500$ loci; In scenario B, we simulated 50 trees with $N = 100$ taxa, each with $M = 2000$ loci; In scenario C, we simulated 50 trees with $N = 100$ taxa, each with $M = 2000$ loci; In scenario D, we simulated 50 trees with $N = 100$ taxa, each with $M = 2000$ loci. The binary trees were simulated via the *ms* software (Hudson, 2002). We used the R package *phangorn* (Schliep, 2011) to create genetic variants under the assumption of Juke Cantor model (Jukes and Cantor, 1969). We assumed that the branch lengths are in units of $2N$ generations. We standardized the genotypes, and uniformly chose one SNP to be significant. We computed a ground-truth genetic similarity matrix given the reference trees, according to S. Wang et al. (2021). In Scenarios A and B, we simulated the phenotype though the LMM described in *Section* 2.1, with $\sigma_e = 0.60$, $\sigma_g = 0.50$, effect size $\beta = 0.20$, and with $\sigma_e = 0.40$, $\sigma_g = 0.20$, effect size $\beta = 0.20$. In Scenarios C and D, we simulated the phenotype though the LM with $\sigma_e = 0.20$, $\beta = 0.20$, genotyping error 0.5% and 10% respectively. These two thresholds correspond to specific sequencing technologies used for MTB genotyping. Based on estimates of error rates from different sequencing technologies, 0.5% is towards the higher estimate for most Illumina short read sequencing technologies (Stoler and Nekrutenko, 2021), and 10% is a good estimate for technologies with higher error rates, such as Oxford Nanopore sequencing (Nicholls et al., 2019), especially given the relatively high GC content in Mtb (>60%), which can influence error rates (Delahaye and Nicolas, 2021).

We compared the true positive rate (TPR) and false discovery rate (FDR) of LiMU, LMM (using the empirical genetic similarity matrix for the kinship), the fixed effects model with the genetic similarity matrix represented by MDS implemented in the *pyseer* software, a linear mixed effects model with the expected genetic similarity matrix computed from consensus tree of Mr. Bayes (CLMM), and a linear model in a task in which associated genetic variants are recovered. To compute the similarity matrix of model implemented in *pyseer* for controlling population structure, we used the consensus tree provided by MrBayes. The estimation of linear mixed model was carried out using the *efficient mixed model association* (EMMA; Lippert et al. 2011). We examined the receiver operating characteristic (ROC) curves induced by the $p$-values for these models for simulated data in which the ground truth is known.

Figure 1 displays the ROC found for these methods for datasets $A$ (upper left panel), $B$ (upper right panel), $C$ (lower left panel) and $D$ (lower right panel). In Scenarios A and B, the ROC curve of LiMU dominates those of *pyseer*, LMM and LM at all FDR level in both Scenarios. In Scenario C, the data was generated via a linear regression model, the ROC curves found for all methods were close. In Scenario D, with higher genotyping error in data simulation, the area under all ROC curves was lower (compared with Scenario C). The area under the ROC curve (AUC) and the improvement at FDR=0.05 are listed in Table 1. Figure 2 shows the TPR at a fixed FDR level 0.05 for the four scenarios shown in Figure 1. The ROC curves provided by LiMU and LMM with the expected genetic similarity matrix computed from the consensus tree according to S. Wang et al. (2021) exhibit similar performance.

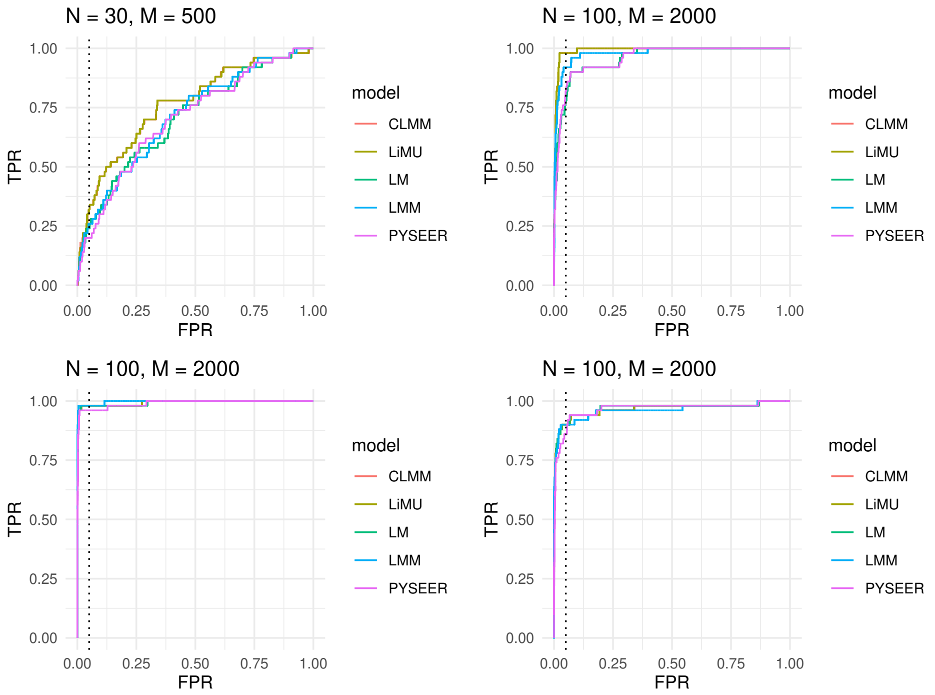

Figure 1: ROC provided by linear regression (LM), linear mixed model with empirical genetic similariy matrix (LMM), the and the fixed effects model implemented in *pyseer* software (PYSEER), and a linear mixed effects model with the expected genetic similarity matrix computed using the consensus tree of Mr. Bayes according to S. Wang et al. (2021), and a linear mixed model with unknown genetic similariy matrices (LiMU) for datasets $A$ (upper left panel), $B$ (upper right panel), $C$ (lower left panel), $D$ (lower right panel), the vertical dotted line is at FPR level of 0.05.

### 3.1.2 Simulation 2

In the second simulation study, we first examined the area under the ROC curve (AUC) provided by the four methods discussed above as a function of $\sigma_e$. We simulated 50 trees with $N = 15$ taxa, each with $M = 100$ loci. We considered 11 levels of $\sigma_e$ equally distanced between 0 and 1. For each level of $\sigma_e$, we simulated the phenotype though the LMM described in *Section* 2.1, with $\beta = 0.1$, $\sigma_g = 0.1$. Hence, we have 50 data sets for each level of $\sigma_e$. The

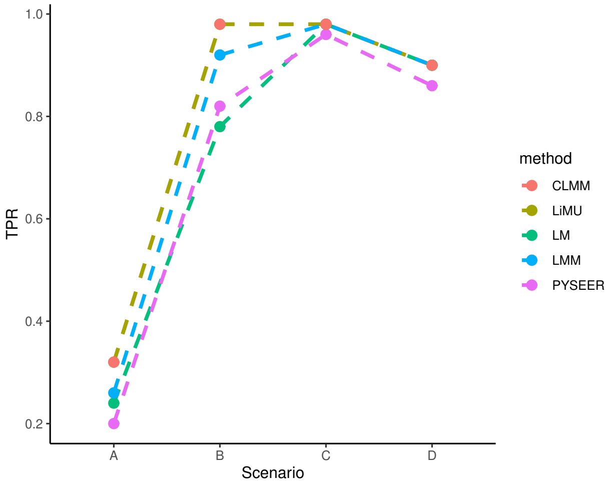

Figure 2: TPR at a fixed FDR level 0.05 for the four scenarios shown in Figure 1.

| | AUC | | | | | TPR (FDR = 0.05) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | LM | PYSEER | LMM | LiMU | CLMM | LM | PYSEER | LMM | LiMU | CLMM |
| A | 0.706 | 0.703 | 0.713 | **0.760** | **0.760** | 0.24 | 0.20 | 0.26 | **0.32** | **0.32** |
| B | 0.958 | 0.958 | 0.980 | **0.992** | **0.992** | 0.78 | 0.82 | 0.92 | **0.98** | **0.98** |
| C | 0.993 | 0.990 | **0.997** | 0.993 | 0.993 | **0.98** | 0.96 | **0.98** | **0.98** | **0.98** |
| D | **0.969** | 0.965 | 0.961 | 0.966 | 0.966 | **0.90** | 0.86 | **0.90** | **0.90** | **0.90** |

Table 1: AUC and TPR at FDR 0.05 for simulation Scenarios A, B and C. LiMU shows improved AUC and TPR in Scenarios A and B. The AUC and TPR are close for all methods in Scenario C.

rest of the setup for this simulation was the same as the previous simulation study. We also report the compute time required for each step of LiMU in Table 2. The experiments are conducted on a 2.3 GHz Intel Core i9 processor. Half a million iterations of Mr.Bayes run costs 6.524 seconds, computation of the genetic similarity matrix for one thinned posterior sample takes $6.84 \cdot 10^{-3}$ seconds, and one run of REML with a sample for the genetic similarity matrix takes 0.613 seconds.

We examined the area under curves (AUC) induced by the p-values of linear regression (LM), linear mixed model with empirical genetic similariy

Table 2: Timing for each step of LiMU. The experiments are conducted on a 2.3 GHz Intel Core i9 processor.

| | Mr.Bayes | GSM | REML |
|---|---|---|---|
| Time (Sec) | 6.524 | $6.84 \cdot 10^{-3}$ | 0.613 |

matrix (LMM), the *pyseer* software (pyseer) and a linear mixed model with unknown genetic similariy matrices (LiMU), for simulated data in which the ground truth was known. Figure 3 displays the area under curve (AUC) as a function of $\sigma_e$. When $\sigma_e$ is small, the AUC provided by the four methods are similar. The AUCs of LiMU and rest of the methods start to diverge once we increase $\sigma_e$, showing that LiMU outperforms the other methods significantly when the heritability is high.

In addition, we examined the area under curve (AUC) as a function of $\sigma_e$ provided by LiMU with p-value summarized by four different statistics (*max, mean, min, median*). Figure 4 indicates that the AUC provided by *mean* and *median* statistics are similar, and are higher than *max* statistics with a high level of $\sigma_e$, the AUC provided by the *min* statistic is lower than the other three.

#### 3.1.3 Simulation 3

In the third simulation study, we design experiments with more sophisticated setups. We create 50 data sets, in each of them we randomly sample $N = 50$ genetic sequences among the 467 tuberculosis subjects that we analyzed in Section 3.2. We first run Mr. Bayes to obtain a consensus phylogeny for each data set, and the consensus tree is used to simulate phenotype. The amount of uncertainty in the phylogeny can be quantified using the R package *treespace* (Jombart et al., 2017; Team et al., 2013). Figure 5 shows the density plot for the first two components provided by metric multidimensional scaling (MDS) (Williams, 2000). The MDS is conducted on the pairwise distance between the 100 thinned posterior samples for one of the 50 data sets using *treespace*.

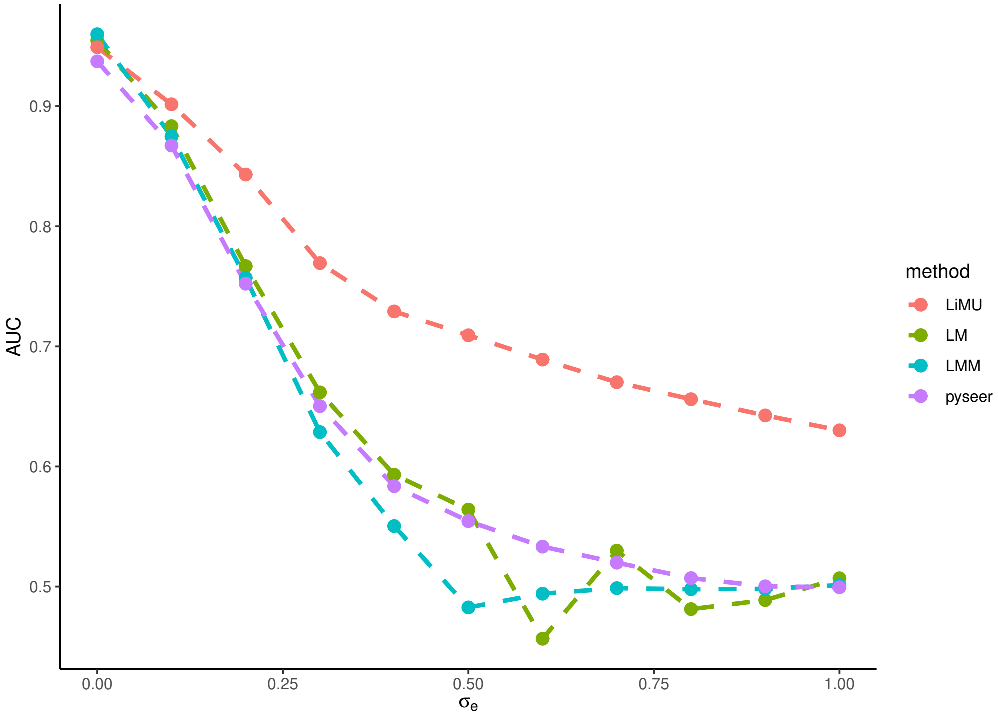

Figure 3: Area under curve (AUC) as a function of $\sigma_e$ provided by linear regression (LM), linear mixed model with empirical genetic similariy matrix (LMM), the *pyseer* software (pyseer) and a linear mixed model with unknown genetic similariy matrices (LiMU). With a small value of $\sigma_e$, the AUC provided by the four methods are close. LiMU and rest methods start to diverge once we increase $\sigma_e$. LiMU works better with higher heritability.

We investigate the effects of polygenic traits using LiMU, linear regression (LM), linear mixed model with empirical genetic similariy matrix (LMM), and the fixed effects model implemented in the *pyseer* software (PYSEER). The genetic sequences are obtained by randomly sampling $M$ markers from the original TB sequences. We examine three levels of $M$, $M = 100, 300, 1000$. For each level of $M$, we simulate phenotypes in four scenarios using multiple significant loci, and in each scenario we simulate 50 data sets with different random seeds. The number of significant loci is set to $d$ = 1, 2, 5 and 10

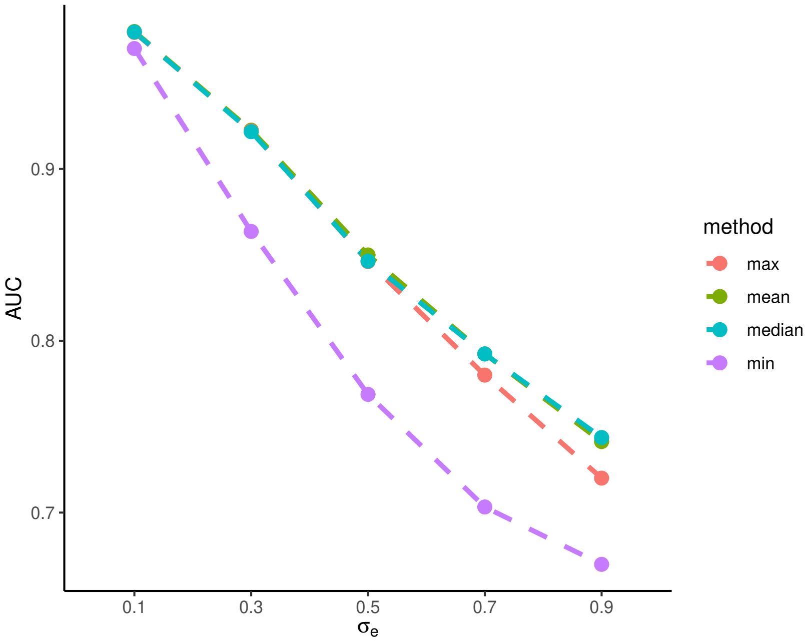

Figure 4: Area under curve (AUC) as a function of $\sigma_e$ provided by LiMU with p-value summarized by four different statistics (*i.e. max*, *mean*, *min*, *median*). The AUC provided by *mean* and *median* statistics are close, and are higher than *max* statistics with a high level of $\sigma_e$. The AUC provided by *min* statistic is lower than the other three.

in the four scenarios respectively. The effective sizes are sampled from a uniform distribution. Figure 6 shows the AUC as a function of $M$ provided by LiMU, linear regression (LM), linear mixed model with empirical genetic similariy matrix (LMM), and the fixed effects model implemented in the *pyseer* software (PYSEER) in four scenarios with size of effects 1 (Upper Left), 2 (Upper Right), 5 (Lower Left) and 10 (Lower Right). The AUCs provided by LMM and LiMU are higher than those provided by LM and PYSEER. For all cases, LiMU approaches LMM with a higher value of $M$. For a small value of $M$, LiMU performs slightly better than LMM when there is only a single significant locus, and LMM performs better than LiMU when

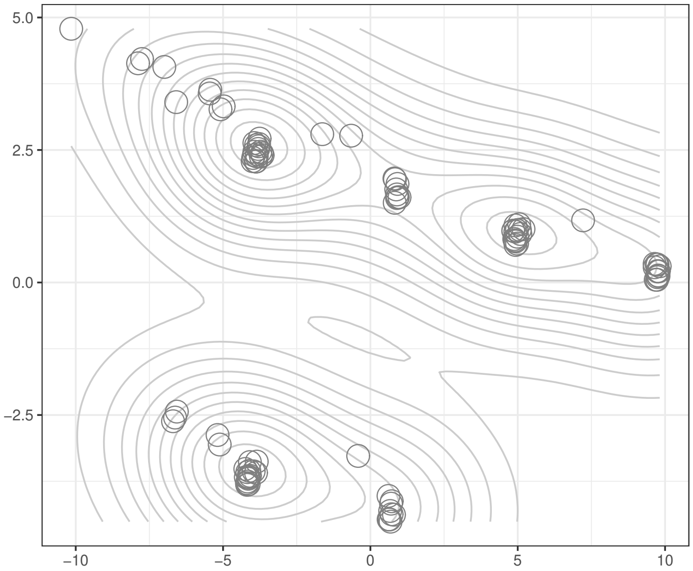

Figure 5: The density plot for the first two components provided by metric multidimensional scaling (MDS). The MDS is conducted on the pairwise distance between the 100 thinned posterior samples for one of the 50 data sets using R package *treespace*.

there are multiple significant loci.

## 3.2 Association study for MDR-TB in Lima, Peru

We carried out a genome-wide association study using the LiMU method to control for population structure for 467 tuberculosis subjects (of which 158 had multidrug-resistant strains) collected in Lima, Peru. These data were previously studied in Grandjean et al. (2017), and in that work many homoplastic variant sites were identified to be significantly correlated, indicating *epistasis*. Our analysis further refines these results with the LiMU control for population structure. We removed genotypes with minor allele frequency below 0.005, yielding $9,848$ SNPs. We compared LM, the fixed effects model with the genetic similarity matrix represented by MDS implemented in software *pyseer*,

so basically here, LMM does really well.
But in Figure 3, LiMU does the best
I think it would help if the text clearly said why (at the point where this figure is discussed)

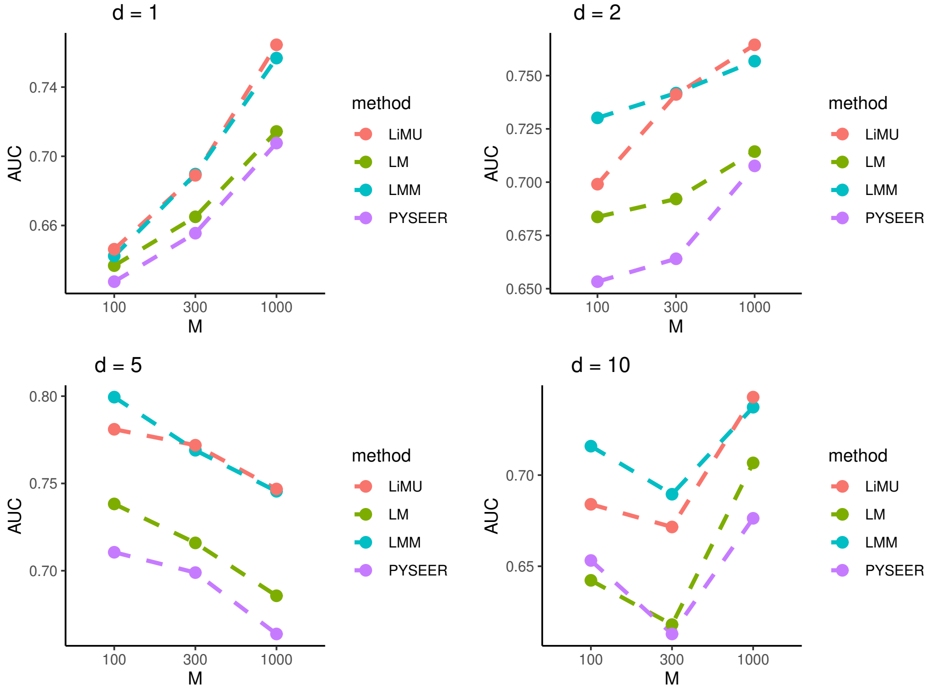

Figure 6: Area under curve (AUC) as a function of $M$ provided by LiMU, linear regression (LM), linear mixed model with empirical genetic similariy matrix (LMM), and the fixed effects model implemented in *pyseer* software (PYSEER) in four scenarios with number of significant loci $d = 1$ (Upper Left), 2 (Upper Right), 5 (Lower Left) and 10 (Lower Right).

LMM and the LiMU. For the LMM, we use the empirical genetic similarity matrix, and the inference was carried out by the EMMA method. For LiMU, we first ran MrBayes to get posterior samples of trees. We ran MrBayes with one million iterations, with burnin given by the first half of the chain, and we collected 50 thinned posterior tree samples. Given the expected genetic similarity matrix based on each sampled tree, we use the EMMA method to infer the LMM parameters. The genetic similarity matrix of the fixed effects model in *pyseer* was computed using the consensus tree provided by the MrBayes analysis. We consider multidrug-resistance (MDR) as the phenotype of interest, and form a binary variable indicating MDR or non-MDR. All samples identified as resistant to either Rifampicin, or Isoniazid (but not resistant to other drugs) are included in the non-MDR set.

We compared our methods to a classical linear regression GWAS with $t$-tests. This linear analysis identified 100 genetic variants that significantly associated with multidrug-resistance after Bonferroni (BF) correction, with $p$-value $< 0.05/9,848$. The LiMU identifies 23 significantly associated genetic variants (red pluses in Figure 8) after BF correction ($p$-values $< 0.05/9,848$). LMM identifies 8 associated genetic variants (blue triangles) after BF correction. Figure 7 shows a Venn diagram of base pair positions for hits provided by LMM and LiMU. *pyseer* identifies 96 associated genetic variants (grey crosses) after BF correction. Both LMM and LiMU significantly correct hits found through linear regression, suggesting that many of these hits are due to population structure. Figure 8 displays the Manhattan plot for these GWASs. Table 3 shows the $p$-values of the 3 most significant hits identified by LiMU. We also summarize the posterior $p$-values through posterior median and geometric mean, yielding values that are close to the $p$-values summarized by mean (the $-\log_{10}$ $p$-values found by posterior median and geometric mean match those found by the posterior mean to two decimal places).

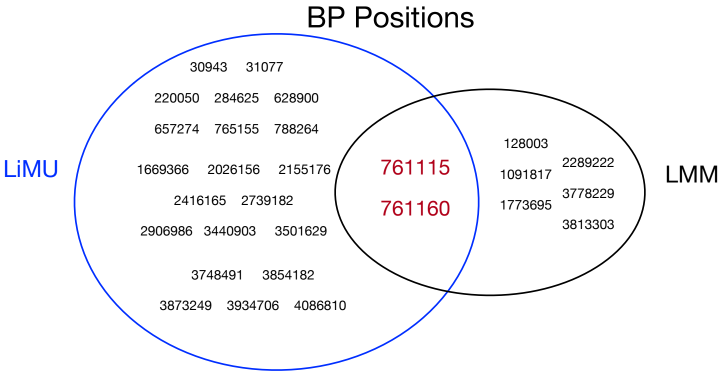

Figure 7: The base pair positions of LiMU and LMM hits.

Table 4 reports the timing (in seconds) for the three main steps of LiMU (*i.e.* 1 million iterations of Mr. Bayes, computation of the genetic similarity

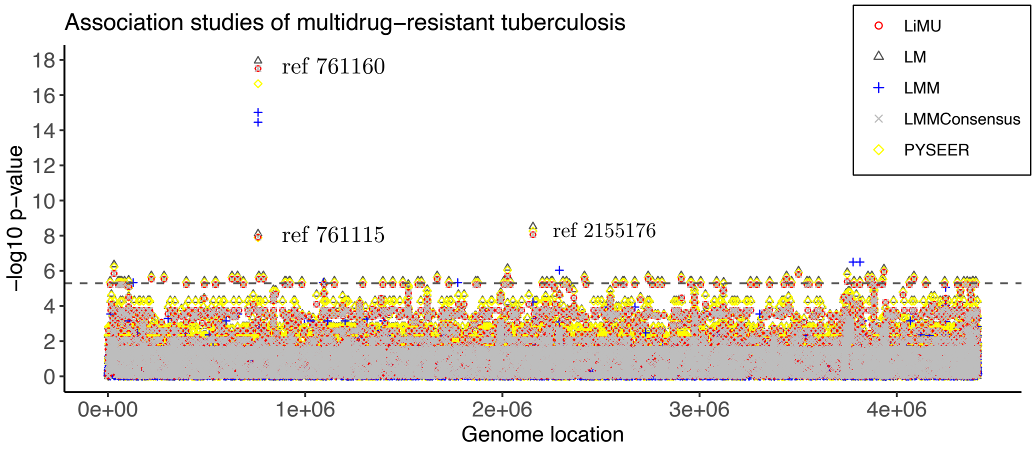

Figure 8: Manhattan plot of genome-wide association studies carried out by linear model (LM), *pyseer* (PYSEER), linear mixed model (LMM) and linear mixed model with uncertain genetic similarity matrices (LiMU) for the TB data. The dashed horizontal line indicates the threshold after Bonferroni (BF) correction. The base pair positions of LiMU hits are also provided.

matrix and REML for one LMM fit). One million iterations of Mr. Bayes run costs $8,414.99$ second, computation of the genetic similarity matrix for one thinned posterior sample takes 0.437 seconds, and one run of REML takes $4,799.37$ seconds.

The hits with BP 761160 and 761115 are non-synonymous mutations (these alter the amino acid produced) in the rpoB gene, which is associated with rifampin resistance (Goldstein, 2014; Lipin et al., 2007). The majority of mutations that confer resistance to rifampin occur within an 81bp region of rpoB, referred to as the rifampin-resistance determining region (RRDR). And while neither of the sites identified here occur within the RRDR region, there is still a chance that strains carrying these mutations may be rifampin-resistant, or they may be compensatory mutations (Lempens et al., 2018; Ma et al., 2021).

In addition, the hit with BP 2155176 is a non-synonymous mutation in the katG gene, which is associated with isoniazid resistance. Rifampin and

Table 3: Negative log $P$-values of LM, LMM, LiMU and *pyseer* for base positions 761115, 761160, 2155176.

| BP Position | 761115 | 761160 | 2155176 |
|---:|---:|---:|---:|
| LM | 8.12 | 17.95 | 8.53 |
| LMM | 14.45 | 15.01 | 4.23 |
| LiMU | 7.92 | 17.51 | 8.06 |
| *pyseer* | 7.88 | 16.64 | 8.26 |

Table 4: Timing (in Seconds) for each step in LiMU. The experiments are conducted on a 2.3 GHz Intel Core i9 processor.

| | Mr.Bayes | GSM | REML |
|---|---|---|---|
| Time (Sec) | $8,414.99$ | 0.437 | $4,799.37$ |

isoniazid are first-line antimicrobials used to treat TB and strains resistant to both are termed multi-drug resistant TB (MDR-TB) (Lipin et al., 2007).

There was also a hit identified by LiMU for a non-synonymous mutation within rpoC, which has been previously shown be involved in compensation of fitness costs associated with rifampin-resistance (De Vos et al., 2013). In addition, there were hits within various PPE and PE-PGRS family genes, and while the exact function of many of these genes is not well understood, there is evidence that many are involved in the host-pathogen interaction and infection (Qian et al., 2020). However, there can be technical challenges with assembly and variant calling at these loci because of a high GC content and excess of repetitive sequences (Ates, 2020), and further work would be required to validate the variation found in these genes.

# 4 Discussion

Standard linear mixed models (LMMs) for genome-wide association studies often assume a single known genetic similarity matrix as a random effect (typically computed as the symmetric matrix resulting from inner products of genetic variants). However, such an approach is inaccurate if genotypes are not densely sampled, or are of poor quality (S. Wang et al., 2021): in S. Wang et al. (2021), it was found that uncertainty in genetic similarity matrices (measured in standard deviation) varied from 0.223 to 0.031 as number of markers varied from 20 to 1000. Uncertainty about the genetic similarity matrix may degrade the quality of LMM estimates.

We have developed a linear mixed effects model for genome-wide association studies incorporating uncertainty about the genetic similarity matrices, in which the genetic similarity matrix is induced by a phylogeny based on the genotype. To account for the uncertainty of phylogeny, we considered a Bayesian framework for the underlying tree and derive the posterior samples through Markov chain Monte Carlo methods (*i.e.* MrBayes). Our proposed method, LiMU is computational more expensive than standard LMMs as we require multiple runs of standard LMM, and use Bayesian sampling methods to obtain posterior tree samples. However, LiMU allows us to consider the uncertainty in the genetic similarity matrix (or phylogeny).

In LiMU, we first estimate posterior samples for the phylogeny, and then estimate parameters of the LMM conditioned on the trees. Our method can utilize any Bayesian phylogenetic inference methods that exist in current literature (Bouckaert et al., 2014; Ronquist et al., 2012; L. Wang et al., 2020; S. Wang and Wang, 2021). In addition, our method is flexible in the sense that the estimates of phylogenies could be obtained from DNA, RNA, or any data source arising from trees (including phylolinguistic data, for example).

Our simulations demonstrate the consistency of our methods, and improved false positive rates over the LMM and *pyseer*. The ROC curve and AUC provided by LiMU dominate those provided by the LMM, *pyseer* and a linear model. Our simulations further show that the advantage of LiMU is seen most clearly when the heritibility of the phenotype is high. There is more uncertainty in the phylogeny for smaller data sets, and in this case LiMU is preferable. We also demonstrate that LiMU is robust to model misspecification and high genotyping error (LiMU outperforms other methods in simulations with high genotyping error, and with simulations in which the phenotypes are not sampled from an LMM). We recommend that LiMU be used for datasets with high genotyping error or small numbers of markers. Our experiments involved less than 10,000 markers. If the number of markers is much larger, then the genetic similarity matrix will have less uncertainty and LiMU results may approach those of the LMM.

We apply our method to a genome-wide association study of 467 multidrug-resistant TB (with around 10,000 markers) in a population from Lima, Peru. In our real data analysis, LiMU found fewer hits than a linear model without random effects. The hits we found involve non-synonymous mutations in the rpoB and katG genes, and a non-synonymous mutation within rpoC, that is associated with rifampin-resistance. These genes are known to be involved with multi-drug resistance or host-pathogen interaction and infection. Our simulations suggest that the false positive rate of LiMU is lower than that of the LMM, and so these hits are likely to be true positives. Also, the hit we identified at BP position 2155176 was not found by the LMM.

Our current approach is limited to sequences without recombination. We could extend to data with recombination events in genealogies. The ancestral recombination graph (ARG) describes the coalescence and recombination events among individuals (Rasmussen et al., 2014). The ARG is composed of a set of coalescent trees separated by break points. To compute expected genetic similarity matrix for samples given an ARG, we could first compute the expected genetic similarity matrices for each of the coalescent trees in the ARG, then compute weighted average for those expected genetic similarity matrices. The weights would be proportional to the number of loci between each consecutive pair of break points. Finally, we could apply LiMU to this weighted set of trees, computing their expected genetic similarity matrices.

# Acknowledgements

# References

Ates, L. S. 2020. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. *Molecular Microbiology*, *113*(1), 4–21.

Bouckaert, R., Heled, J., Kühnert, D., et al. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *10*(4).

Bycroft, C., Freeman, C., Petkova, D., et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209.

Coll, F., McNerney, R., Guerra-Assuncao, J. A., et al. 2014. A robust SNP barcode for typing mycobacterium tuberculosis complex strains. *Nature Communications*, *5*(4812).

Corbeil, R. R., and Searle, S. R. 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, *18*(1), 31–38.

Cordero, O. X., and Polz, M. F. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology*, *12*(4).

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

Delahaye, C., and Nicolas, J. 2021. Sequencing DNA with nanopores: Troubles and biases. *PloS one*, *16*(10), e0257521.

De Vos, M., Müller, B., Borrell, S., et al. 2013. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrobial agents and chemotherapy*, *57*(2), 827–832.

Earle, S. G., Wu, C.-H., Charlesworth, J., et al. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, *1*(5).

Goldstein, B. P. 2014. Resistance to rifampicin: a review. *The Journal of antibiotics*, *67*(9), 625–630.

Grandjean, L., Gilman, R. H., Iwamoto, T., et al. 2017. Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. *PLOS One*, *12*(12).

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. 2006. Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157–1174.

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

Jiang, L., Zheng, Z., Fang, H., and Yang, J. 2021. A generalized linear mixed model association tool for biobank-scale data. *Nature genetics*, *53*(11), 1616–1621.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. 2017. treespace: Statistical exploration of landscapes of phylogenetic trees. *Molecular ecology resources*, *17*(6), 1385–1392.

Jukes, T. H., and Cantor, C. R. 1969. Evolution of protein molecules. *Mammalian Protein Metabolism*, *3*(21).

Kang, H. M., Sul, J. H., Service, S. K., et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4).

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

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

Lempens, P., Meehan, C. J., Vandelannoote, K., et al. 2018. Isoniazid resistance levels of Mycobacterium tuberculosis can largely be predicted by high-confidence resistance-conferring mutations. *Scientific reports*, *8*(1), 1–9.

Lipin, M., Stepanshina, V., Shemyakin, I., and Shinnick, T. 2007. Association of specific mutations in katg, rpob, rpsl and rrs genes with spoligotypes of multidrug-resistant mycobacterium tuberculosis isolates in russia. *Clinical microbiology and infection*, *13*(6), 620–626.

Lippert, C., Listgarten, J., Liu, Y., et al. 2011. FaST linear mixed models for genome-wide association studies. *Nature methods*, *8*(10), 833–835.

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

Listgarten, J., Lippert, C., Kadie, C. M., et al. 2012. Improved linear mixed models for genome-wide association studies. *Nature Methods*, *9*(6).

Loh, P.-R., Kichaev, G., Gazal, S., et al. 2018. Mixed-model association for biobank-scale datasets. *Nature genetics*, *50*(7), 906–908.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., et al. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, *47*(3), 284–290.

Ma, P., Luo, T., Ge, L., et al. 2021. Compensatory effects of M. tuberculosis rpoB mutations outside the rifampicin resistance-determining region. *Emerging microbes and infections*, *10*(1), 743–752.

Meng, X.-L. 1994. Posterior predictive *p*-values. *The annals of statistics*, *22*(3), 1142–1160.

Nicholls, S. M., Quick, J. C., Tang, S., and Loman, N. J. 2019. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience*, *8*(5), giz043.

Novembre, J., Johnson, T., Bryc, K., et al. 2008. Genes mirror geography within Europe. *Nature*, *456*(7218).

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

Price, A. L., Patterson, N. J., Plenge, R. M., et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8).

Qian, J., Chen, R., Wang, H., and Zhang, X. 2020. Role of the PE/PPE family in host–pathogen interactions and prospects for anti-tuberculosis vaccine and diagnostic tool design. *Frontiers in cellular and infection microbiology*, *10*, 743.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, *10*(5).

Ronquist, F., Teslenko, M., van der Mark, P., et al. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*, 539-542.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593.

Stoler, N., and Nekrutenko, A. 2021. Sequencing error profiles of illumina sequencing instruments. *NAR genomics and bioinformatics*, *3*(1), lqab019.

Sudlow, C., Gallacher, J., Allen, N., et al. 2015. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), e1001779.

Team, R. C., et al. 2013. R: A language and environment for statistical computing.

Wang, L., Wang, S., and Bouchard-Côté, A. 2020. An annealed sequential Monte Carlo method for Bayesian phylogenetics. *Systematic Biology*, *69*(1), 155-183.

Wang, S., Ge, S., Colijn, C., et al. 2021. Estimating genetic similarity matrices using phylogenies. *Journal of Computational Biology*.

Wang, S., and Wang, L. 2021. Particle Gibbs sampling for Bayesian phylogenetic inference. *Bioinformatics*, *37*(5), 642–649.

Williams, C. 2000. On a connection between kernel PCA and metric multidimensional scaling. *Advances in neural information processing systems*, *13*.

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

Zhang, Z., Ersoz, E., Lai, C.-Q., et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4).

Zhou, X., and Stephens, M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*(4).