

<sup>1</sup> Simultaneous SNP selection and adjustment for  
<sup>2</sup> population structure in high dimensional prediction  
<sup>3</sup> models

<sup>4</sup> Sahir R Bhatnagar<sup>1,2</sup>, Yi Yang<sup>4</sup>, Tianyuan Lu<sup>2</sup>, Erwin Schurr<sup>6</sup>,  
<sup>5</sup> JC Loredo-Osti<sup>7</sup>, Marie Forest<sup>2</sup>, Karim Oualkacha<sup>3</sup>, and  
<sup>6</sup> Celia MT Greenwood<sup>1,2,5</sup>

<sup>7</sup> <sup>1</sup>Department of Epidemiology, Biostatistics and Occupational Health, McGill  
<sup>8</sup> University

<sup>9</sup> <sup>2</sup>Lady Davis Institute, Jewish General Hospital, Montréal, QC

<sup>10</sup> <sup>3</sup>Département de Mathématiques, Université de Québec à Montréal

<sup>11</sup> <sup>4</sup>Department of Mathematics and Statistics, McGill University

<sup>12</sup> <sup>5</sup>Departments of Oncology and Human Genetics, McGill University

<sup>13</sup> <sup>6</sup>Department of Medicine, McGill University

<sup>14</sup> <sup>7</sup>Department of Mathematics and Statistics, Memorial University

<sup>15</sup> March 7, 2020

<sup>16</sup> **Abstract**

<sup>17</sup> Complex traits are known to be influenced by a combination of environmental fac-

tors and rare and common genetic variants. However, detection of such multivariate associations can be compromised by low statistical power and confounding by population structure. Linear mixed effects models (LMM) can account for correlations due to relatedness but have not been applicable in high-dimensional (HD) settings where the number of fixed effect predictors greatly exceeds the number of samples. False positives or false negatives can result from two-stage approaches, where the residuals estimated from a null model adjusted for the subjects' relationship structure are subsequently used as the response in a standard penalized regression model. To overcome these challenges, we develop a general penalized LMM with a single random effect called `gmmix` for simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. We develop a blockwise coordinate descent algorithm with automatic tuning parameter selection which is highly scalable, computationally efficient and has theoretical guarantees of convergence. Through simulations and three real data examples, we show that `gmmix` leads to more parsimonious models compared to the two-stage approach or principal component adjustment with better prediction accuracy. Our method performs well even in the presence of highly correlated markers, and when the causal SNPs are included in the kinship matrix. `gmmix` can be used to construct polygenic risk scores and select instrumental variables in Mendelian randomization studies. **Our algorithms are available in an R package available on CRAN (<https://cran.r-project.org/package=gmmix>).**

## 38 1 Author Summary

This work addresses a recurring challenge in the analysis and interpretation of genetic association studies: which genetic variants can best predict and are independently associated with a given phenotype in the presence of population structure ? Not controlling confounding due to geographic population structure, family and/or cryptic relatedness can lead to spurious associations. Much of the existing research has therefore focused on modeling the association between a phenotype and a single genetic variant in a linear mixed model with

45 a random effect. However, this univariate approach may miss true associations due to the  
46 stringent significance thresholds required to reduce the number of false positives and also  
47 ignores the correlations between markers. We propose an alternative method for fitting  
48 high-dimensional multivariable models, which selects SNPs that are independently associ-  
49 ated with the phenotype while also accounting for population structure. We provide an  
50 efficient implementation of our algorithm and show through simulation studies and real data  
51 examples that our method outperforms existing methods in terms of prediction accuracy and  
52 controlling the false discovery rate.

## 53 2 Introduction

54 Genome-wide association studies (GWAS) have become the standard method for analyzing  
55 genetic datasets owing to their success in identifying thousands of genetic variants associated  
56 with complex diseases (<https://www.genome.gov/gwastudies/>). Despite these impressive  
57 findings, the discovered markers have only been able to explain a small proportion of the  
58 phenotypic variance; this is known as the missing heritability problem [1]. One plausible  
59 reason is that there are many causal variants that each explain a small amount of variation  
60 with small effect sizes [2]. Methods such GWAS, which test each variant or single nucleotide  
61 polymorphism (SNP) independently, may miss these true associations due to the stringent  
62 significance thresholds required to reduce the number of false positives [1]. Another major  
63 issue to overcome is that of confounding due to geographic population structure, family  
64 and/or cryptic relatedness which can lead to spurious associations [3]. For example, there  
65 may be subpopulations within a study that differ with respect to their genotype frequencies  
66 at a particular locus due to geographical location or their ancestry. This heterogeneity in  
67 genotype frequency can cause correlations with other loci and consequently mimic the signal  
68 of association even though there is no biological association [4, 5]. Studies that separate  
69 their sample by ethnicity to address this confounding suffer from a loss in statistical power

70 due to the drop in sample size.

71 To address the first problem, multivariable regression methods have been proposed which  
72 simultaneously fit many SNPs in a single model [6, 7]. Indeed, the power to detect an  
73 association for a given SNP may be increased when other causal SNPs have been accounted  
74 for. Conversely, a stronger signal from a causal SNP may weaken false signals when modeled  
75 jointly [6].

76 Solutions for confounding by population structure have also received significant attention in  
77 the literature [8, 9, 10, 11]. There are two main approaches to account for the relatedness  
78 between subjects: 1) the principal component (PC) adjustment method and 2) the linear  
79 mixed model (LMM). The PC adjustment method includes the top PCs of genome-wide  
80 SNP genotypes as additional covariates in the model [12]. The LMM uses an estimated  
81 covariance matrix from the individuals' genotypes and includes this information in the form  
82 of a random effect [3].

83 While these problems have been addressed in isolation, there has been relatively little  
84 progress towards addressing them jointly at a large scale. Region-based tests of association  
85 have been developed where a linear combination of  $p$  variants is regressed on the response  
86 variable in a mixed model framework [13]. In case-control data, a stepwise logistic-regression  
87 procedure was used to evaluate the relative importance of variants within a small genetic  
88 region [14]. These methods however are not applicable in the high-dimensional setting, i.e.,  
89 when the number of variables  $p$  is much larger than the sample size  $n$ , as is often the case in  
90 genetic studies where millions of variants are measured on thousands of individuals.

91 There has been recent interest in using penalized linear mixed models, which place a con-  
92 straint on the magnitude of the effect sizes while controlling for confounding factors such as  
93 population structure. For example, the LMM-lasso [15] places a Laplace prior on all main  
94 effects while the adaptive mixed lasso [16] uses the  $L_1$  penalty [17] with adaptively chosen  
95 weights [18] to allow for differential shrinkage amongst the variables in the model. Another

method applied a combination of both the lasso and group lasso penalties in order to select variants within a gene most associated with the response [19]. However, methods such as the LMM-lasso are normally performed in two steps. First, the variance components are estimated once from a LMM with a single random effect. These LMMs normally use the estimated covariance matrix from the individuals' genotypes to account for the relatedness but assumes no SNP main effects (i.e. a null model). The residuals from this null model with a single random effect can be treated as independent observations because the relatedness has been effectively removed from the original response. In the second step, these residuals are used as the response in any high-dimensional model that assumes uncorrelated errors. This approach has both computational and practical advantages since existing penalized regression software such as `glmnet` [20] and `gglasso` [21], which assume independent observations, can be applied directly to the residuals. However, recent work has shown that there can be a loss in power if a causal variant is included in the calculation of the covariance matrix as its effect will have been removed in the first step [13, 22].

In this paper we develop a general penalized LMM framework called `ggmix` that simultaneously selects variables and estimates their effects, accounting for between-individual correlations. We develop a blockwise coordinate descent algorithm with automatic tuning parameter selection which is highly scalable, computationally efficient and has theoretical guarantees of convergence. Our method can handle several sparsity inducing penalties such as the lasso [17] and elastic net [23]. Through simulations and three real data examples, we show that `ggmix` leads to more parsimonious models compared to the two-stage approach or principal component adjustment with better prediction accuracy. Our method performs well even in the presence of highly correlated markers, and when the causal SNPs are included in the kinship matrix. All of our algorithms are implemented in the `ggmix` R package hosted on CRAN with extensive documentation (<https://sahirbhatnagar.com/ggmix>). We provide a brief demonstration of the `ggmix` package in Appendix C.

122 The rest of the paper is organized as follows. In Section 3, we compare the performance  
123 of our proposed approach and demonstrate the scenarios where it can be advantageous to  
124 use over existing methods through simulation studies and three real data analyses. This is  
125 followed by a discussion of our results, some limitations and future directions in Section 4.  
126 Section 5 describes the `gmmix` model, the optimization procedure and the algorithm used to  
127 fit it.

## 128 3 Results

129 In this section we demonstrate the performance of `gmmix` in a simulation study and three  
130 real data applications.

### 131 3.1 Simulation Study

132 We evaluated the performance of `gmmix` in a variety of simulated scenarios. For each simula-  
133 tion scenario we compared `gmmix` to the `lasso` and the `twostep` method. For the `lasso`, we  
134 included the top 10 principal components from the simulated genotypes used to calculate the  
135 kinship matrix as unpenalized predictors in the design matrix. For the `twostep` method, we  
136 first fitted an intercept only model with a single random effect using the average information  
137 restricted maximum likelihood (AIREML) algorithm [24] as implemented in the `gaston` R  
138 package [25]. The residuals from this model were then used as the response in a regular  
139 `lasso` model. Note that in the `twostep` method, we removed the kinship effect in the first  
140 step and therefore did not need to make any further adjustments when fitting the penalized  
141 model. We fitted the `lasso` using the default settings and `standardize=FALSE` in the `glmnet`  
142 package [20], with 10-fold cross-validation (CV) to select the optimal tuning parameter. For  
143 other parameters in our simulation study, we defined the following quantities:

- 144
  - $n$ : sample size
  - $c$ : percentage of causal SNPs

- 146     •  $\beta$ : true effect size vector of length  $p$
- 147     •  $S_0 = \{j; (\beta)_j \neq 0\}$  the index of the true active set with cardinality  $|S_0| = c \times p$
- 148     • *causal*: the list of causal SNP indices
- 149     • *kinship*: the list of SNP indices for the kinship matrix
- 150     •  $\mathbf{X}$ :  $n \times p$  matrix of SNPs that were included as covariates in the model

151 We simulated data from the model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{P} + \boldsymbol{\varepsilon} \quad (1)$$

152 where  $\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\Phi)$  is the polygenic effect and  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1 - \eta)\sigma^2\mathbf{I})$  is the error term.  
 153 Here,  $\Phi_{n \times n}$  is the covariance matrix based on the *kinship* SNPs from  $n$  individuals,  $\mathbf{I}_{n \times n}$  is  
 154 the identity matrix and parameters  $\sigma^2$  and  $\eta \in [0, 1]$  determine how the variance is divided  
 155 between  $\mathbf{P}$  and  $\boldsymbol{\varepsilon}$ . The values of the parameters that we used were as follows: narrow  
 156 sense heritability  $\eta = \{0.1, 0.3\}$ , number of covariates  $p = 5,000$ , number of *kinship* SNPs  
 157  $k = 10,000$ , percentage of *causal* SNPs  $c = \{0\%, 1\%\}$  and  $\sigma^2 = 1$ . In addition to these  
 158 parameters, we also varied the amount of overlap between the *causal* list and the *kinship*  
 159 list. We considered two main scenarios:

- 160     1. None of the *causal* SNPs are included in *kinship* set.
- 161     2. All of the *causal* SNPs are included in the *kinship* set.

162 Both kinship matrices were meant to contrast the model behavior when the causal SNPs are  
 163 included in both the main effects and random effects (referred to as proximal contamina-  
 164 tion [8]) versus when the causal SNPs are only included in the main effects. These scenarios  
 165 are motivated by the current standard of practice in GWAS where the candidate marker  
 166 is excluded from the calculation of the kinship matrix [8]. This approach becomes much  
 167 more difficult to apply in large-scale multivariable models where there is likely to be overlap

168 between the variables in the design matrix and kinship matrix. We simulated random geno-  
 169 types from the BN-PSD admixture model with 1D geography and 10 subpopulations using  
 170 the `bnpsd` package [26, 27]. In Figure 1, we plot the estimated kinship matrix from a single  
 171 simulated dataset in the form of a heatmap where a darker color indicates a closer genetic  
 172 relationship.

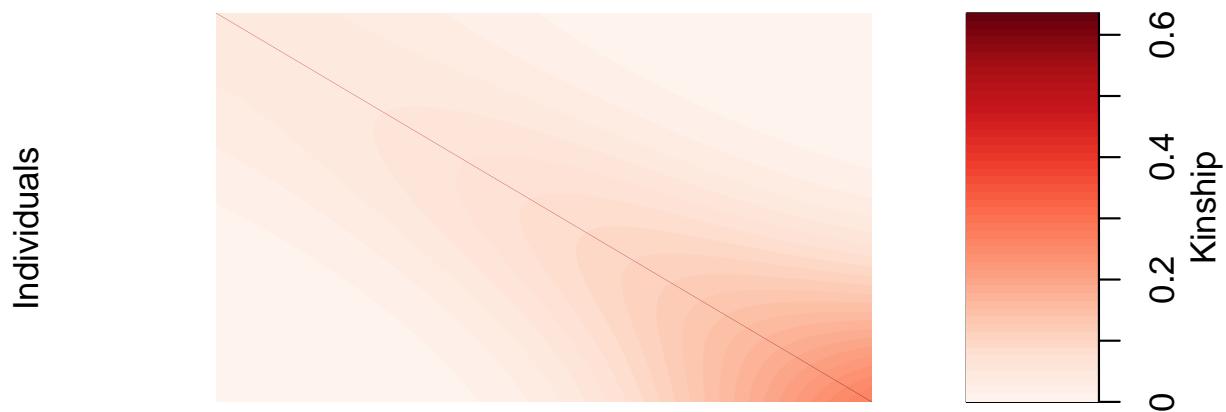


Figure 1: Example of an empirical kinship matrix used in simulation studies. This scenario models a 1D geography with extensive admixture.

173 In Figure 2 we plot the first two principal component scores calculated from the simulated  
 174 genotypes used to calculate the kinship matrix in Figure 1, and color each point by sub-  
 175 population membership. We can see that the PCs can identify the subpopulations which  
 176 is why including them as additional covariates in a regression model has been considered a  
 177 reasonable approach to control for confounding.

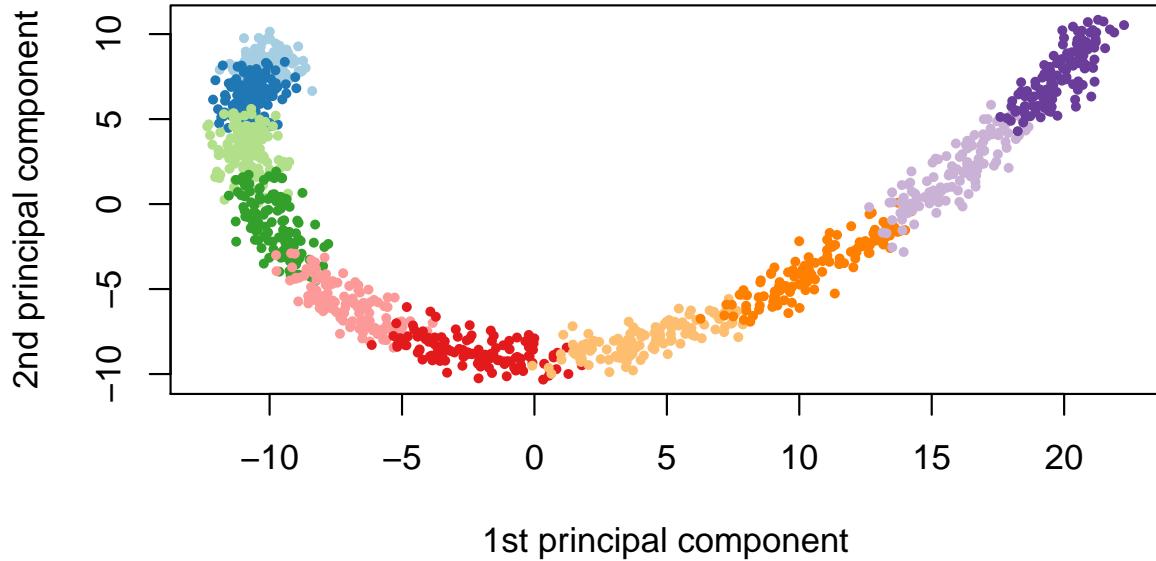


Figure 2: First two principal component scores of the genotype data used to estimate the kinship matrix where each color represents one of the 10 simulated subpopulations.

178 Using this set-up, we randomly partitioned 1000 simulated observations into 80% for training  
 179 and 20% for testing. The training set was used to fit the model and select the optimal tuning  
 180 parameter only, and the resulting model was evaluated on the test set. Let  $\hat{\lambda}$  be the esti-  
 181 mated value of the optimal regularization parameter,  $\hat{\beta}_{\hat{\lambda}}$  the estimate of  $\beta$  at regulariza-  
 182 tion parameter  $\hat{\lambda}$ , and  $\hat{S}_{\hat{\lambda}} = \{j; (\hat{\beta}_{\hat{\lambda}})_j \neq 0\}$  the index of the set of non-zero estimated coefficients.  
 183 To compare the methods in the context of true positive rate (TPR), we selected the largest  
 184 tuning parameter that would result in a false positive rate (FPR) closest to 5%, but not  
 185 more. **Note that in practice, this approach to selecting the tuning parameter is**  
 186 **generally not possible since we do not know the underlying true model in ad-**  
 187 **vance. For real data, we suggest an information criterion approach described in**  
 188 **Section 5.3.8 or a sample splitting approach such as the one we used for the UK**  
 189 **Biobank analysis shown in Section 3.2.1.** We also compared the model size ( $|\hat{S}_{\hat{\lambda}}|$ ), test

190 set prediction error based on the refitted unpenalized estimates for each selected model, the  
191 estimation error ( $\|\hat{\beta} - \beta\|_2^2$ ), and the variance components ( $\eta, \sigma^2$ ) for the polygenic random  
192 effect and error term.

193 The results are summarized in Table 1. We see that `gmmix` outperformed the `twostep` in  
194 terms of TPR, and was comparable to the `lasso`. This was the case, regardless of true heri-  
195 tability and whether the causal SNPs were included in the calculation of the kinship matrix.

196 For the `twostep` however, the TPR at a FPR of 5%, drops, on average, from 0.84 (when  
197 causal SNPs are not in the kinship) to 0.76 (when causal SNPs are in the kinship). Across  
198 all simulation scenarios, `gmmix` had the smallest estimation error, and smallest root mean  
199 squared prediction error (RMSE) on the test set while also producing the most parsimonious  
200 models. Both the `lasso` and `twostep` selected more false positives, even in the null model  
201 scenario. Both the `twostep` and `gmmix` overestimated the heritability though `gmmix` was  
202 closer to the true value. When none of the causal SNPs were in the kinship, both methods  
203 tended to overestimate the truth when  $\eta = 10\%$  and underestimate when  $\eta = 30\%$ . Across  
204 all simulation scenarios `gmmix` was able to (on average) correctly estimate the error variance.  
205 The `lasso` tended to overestimate  $\sigma^2$  in the null model while the `twostep` overestimated  $\sigma^2$   
206 when none of the causal SNPs were in the kinship matrix.

207 Overall, we observed that variable selection results and RMSE for `gmmix` were similar regard-  
208 less of whether the causal SNPs were in the kinship matrix or not. This result is encouraging  
209 since in practice the kinship matrix is constructed from a random sample of SNPs across the  
210 genome, some of which are likely to be causal, particularly in polygenic traits.

211 In particular, our simulation results show that the principal component adjustment method  
212 may not be the best approach to control for confounding by population structure, particularly  
213 when variable selection is of interest.

Table 1: Mean (standard deviation) from 200 simulations stratified by the number of causal SNPs (null, 1%), the overlap between causal SNPs and kinship matrix (no overlap, all causal SNPs in kinship), and true heritability (10%, 30%). For all simulations, sample size is  $n = 1000$ , the number of covariates is  $p = 5000$ , and the number of SNPs used to estimate the kinship matrix is  $k = 10000$ . TPR at FPR=5% is the true positive rate at a fixed false positive rate of 5%. Model Size ( $|\widehat{S}_{\lambda}|$ ) is the number of selected variables in the training set using the high-dimensional BIC for `gmmix` and 10-fold cross validation for `lasso` and `twostep`. RMSE is the root mean squared error on the test set. Estimation error is the squared distance between the estimated and true effect sizes. Error variance ( $\sigma^2$ ) for `twostep` is estimated from an intercept only LMM with a single random effect and is modeled explicitly in `gmmix`. For the `lasso` we use  $\frac{1}{n-|\widehat{S}_{\lambda}|} \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda}\|_2^2$  [28] as an estimator for  $\sigma^2$ . Heritability ( $\eta$ ) for `twostep` is estimated as  $\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$  from an intercept only LMM with a single random effect where  $\sigma_g^2$  and  $\sigma_e^2$  are the variance components for the random effect and error term, respectively.  $\eta$  is explicitly modeled in `gmmix`. There is no positive way to calculate  $\eta$  for the `lasso` since we are using a PC adjustment.

Metric	Method	Null model				1% Causal SNPs			
		No overlap		All causal SNPs in kinship		No overlap		All causal SNPs in kinship	
		10%	30%	10%	30%	10%	30%	10%	30%
<b>TPR at FPR=5%</b>	twostep	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.84 (0.05)	0.84 (0.05)	0.76 (0.09)	0.77 (0.08)
	lasso	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.86 (0.05)	0.85 (0.05)	0.86 (0.05)	0.86 (0.05)
	gmmix	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.86 (0.05)	0.86 (0.05)	0.85 (0.05)	0.86 (0.05)
	twostep	0 (0, 5) (289)	0 (0, 2) (287)	0 (0, 5) (388)	0 (0, 2) (385)	328 (250)	332 (329)	284 (319)	284 (253)
	lasso	0 (0, 6) (246)	0 (0, 5) (317)	0 (0, 6) (314)	0 (0, 5) (245)	278 (252)	276 (321)	279 (244)	285 (319)
	gmmix	0 (0, 0) (43)	0 (0, 0) (43)	0 (0, 0) (39)	0 (0, 0) (48)	43 (39, 49) (43)	43 (39, 48) (44)	44 (38, 49) (43)	43 (38, 48) (43)
	twostep	1.02 (0.07)	1.02 (0.06)	1.02 (0.07)	1.02 (0.06)	1.42 (0.10)	1.41 (0.10)	1.44 (0.33)	1.40 (0.22)
	lasso	1.02 (0.06)	1.02 (0.06)	1.02 (0.06)	1.02 (0.06)	1.39 (0.09)	1.38 (0.09)	1.40 (0.08)	1.38 (0.08)
	gmmix	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.22 (0.10)	1.20 (0.10)	1.23 (0.11)	1.23 (0.12)
<b>Model Size</b>	twostep	0.12 (0.22)	0.09 (0.19)	0.12 (0.22)	0.09 (0.19)	2.97 (0.60)	2.92 (0.60)	3.60 (5.41)	3.21 (3.46)
	lasso	0.13 (0.21)	0.12 (0.22)	0.13 (0.21)	0.12 (0.22)	2.76 (0.46)	2.69 (0.47)	2.82 (0.48)	2.75 (0.48)
	gmmix	0.00 (0.01)	0.01 (0.02)	0.00 (0.01)	0.01 (0.02)	2.11 (1.28)	2.04 (1.22)	2.21 (1.24)	2.28 (1.34)
	twostep	0.87 (0.11)	0.69 (0.15)	0.87 (0.11)	0.69 (0.15)	14.23 (3.53)	14.13 (3.52)	1.42 (1.71)	1.28 (1.66)
	lasso	0.98 (0.05)	0.96 (0.05)	0.98 (0.05)	0.96 (0.05)	1.04 (0.13)	1.02 (0.13)	1.03 (0.14)	1.01 (0.14)
<b>Error Variance</b>	gmmix	0.85 (0.18)	0.64 (0.20)	0.85 (0.18)	0.64 (0.20)	2.00 (0.49)	1.86 (0.51)	1.06 (0.46)	0.83 (0.45)
	twostep	0.13 (0.11)	0.31 (0.15)	0.13 (0.11)	0.31 (0.15)	0.26 (0.14)	0.26 (0.14)	0.92 (0.08)	0.93 (0.08)
	lasso	—	—	—	—	—	—	—	—
	gmmix	0.15 (0.18)	0.37 (0.21)	0.15 (0.18)	0.37 (0.21)	0.18 (0.16)	0.23 (0.17)	0.59 (0.20)	0.68 (0.19)

*Note:*

Median (Inter-quartile range) is given for Model Size.

214 **3.2 Real Data Applications**

215 Three datasets with different features were used to illustrate the potential advantages of  
216 `gmmix` over existing approaches such as PC adjustment in a `lasso` regression. In the first  
217 two datasets, family structure induced low levels of correlation and sparsity in signals. In  
218 the last, a dataset involving mouse crosses, correlations were extremely strong and could  
219 confound signals.

220 **3.2.1 UK Biobank**

221 With more than 500,000 participants, the UK Biobank is one of the largest genotyped health  
222 care registries in the world. Among these participants, 147,731 have been inferred to be  
223 related to at least one individual in this cohort [29]. Such a widespread genetic relatedness  
224 may confound association studies and bias trait predictions if not properly accounted for.  
225 Among these related individuals, 18,150 have a documented familial relationship (parent-  
226 offspring, full siblings, second degree or third degree) that was previously inferred in [30]. We  
227 attempted to derive a polygenic risk score for height among these individuals. As suggested  
228 by a reviewer, the goal of this analysis was to see how the different methods performed for  
229 a highly polygenic trait in a set of related individuals. We compared the `gmmix`-derived  
230 polygenic risk score to those derived by the `twostep` and `lasso` methods.

231 We first estimated the pairwise kinship coefficient among the 18,150 reportedly related indi-  
232 viduals based on 784,256 genotyped SNPs using KING [31]. We grouped related individuals  
233 with a kinship coefficient  $> 0.044$  [31] into 8,300 pedigrees. We then randomly split the  
234 dataset into a training set, a model selection set and a test set of roughly equal sample size,  
235 ensuring all individuals in the same pedigree were assigned into the same set. We inverse  
236 normalized the standing height after adjusting for age, sex, genotyping array, and assessment  
237 center following Yengo et al. [32].

238 To reduce computational complexity, we selected 10,000 SNPs with the largest effect sizes

associated with height from a recent large meta-analysis [32]. Among these 10,000 SNPs, 1,233 were genotyped and used for estimating the kinship whereas the other 8,767 SNPs were imputed based on the Haplotype Reference Consortium reference panel [33]. The distribution of the 10,000 SNPs by chromosome and whether or not the SNP was imputed is shown in Figure B.1 in Supplemental Section B. We see that every chromosome contributed SNPs to the model with 15% coming from chromosome 6. The markers we used are theoretically independent since Yengo et al. performed a COJO analysis which should have tuned down signals due to linkage disequilibrium [32]. We used `gmmix`, `twostep` and `lasso` to select SNPs most predictive of the inverse normalized height on the training set, and chose the  $\lambda$  with the lowest prediction RMSE on the model selection set for each method. We then examined the performance of each derived polygenic risk score on the test set. Similar to Section 3.1, we adjusted for the top 10 genetic PCs as unpenalized predictors when fitting the `lasso` models, and supplied the kinship matrix based on 784,256 genotyped SNPs to `gmmix` and `twostep`.

We found that with a kinship matrix estimated using all genotyped SNPs, `gmmix` had the possibility to achieve a lower RMSE on the model selection set compared to the `twostep` and `lasso` methods (Figure 3A). An optimized `gmmix`-derived polygenic risk score that utilized the least number of SNPs was also able to better predict the trait with lower RMSE on the test set (Figure 3B).

We additionally applied a Bayesian Sparse Linear Mixed Model (BSLMM) [34] implemented in the GEMMA package [35] to derive a polygenic risk score on the training set. **A posterior probability of inclusion of each SNP was provided and prediction was based on all SNPs with a positive posterior probability.** We found that although the BSLMM-based polygenic risk score leveraged the most SNPs, it did not achieve a comparable prediction accuracy as the other three methods (Figure 3B). **Likely due to the small effect sizes of these SNPs, only 94, 35 and 1 SNPs had a posterior inclusion probability**

above 0.05, 0.10 and 0.50, respectively. The model would have further reduced prediction accuracy if the prediction was based only on these SNPs.

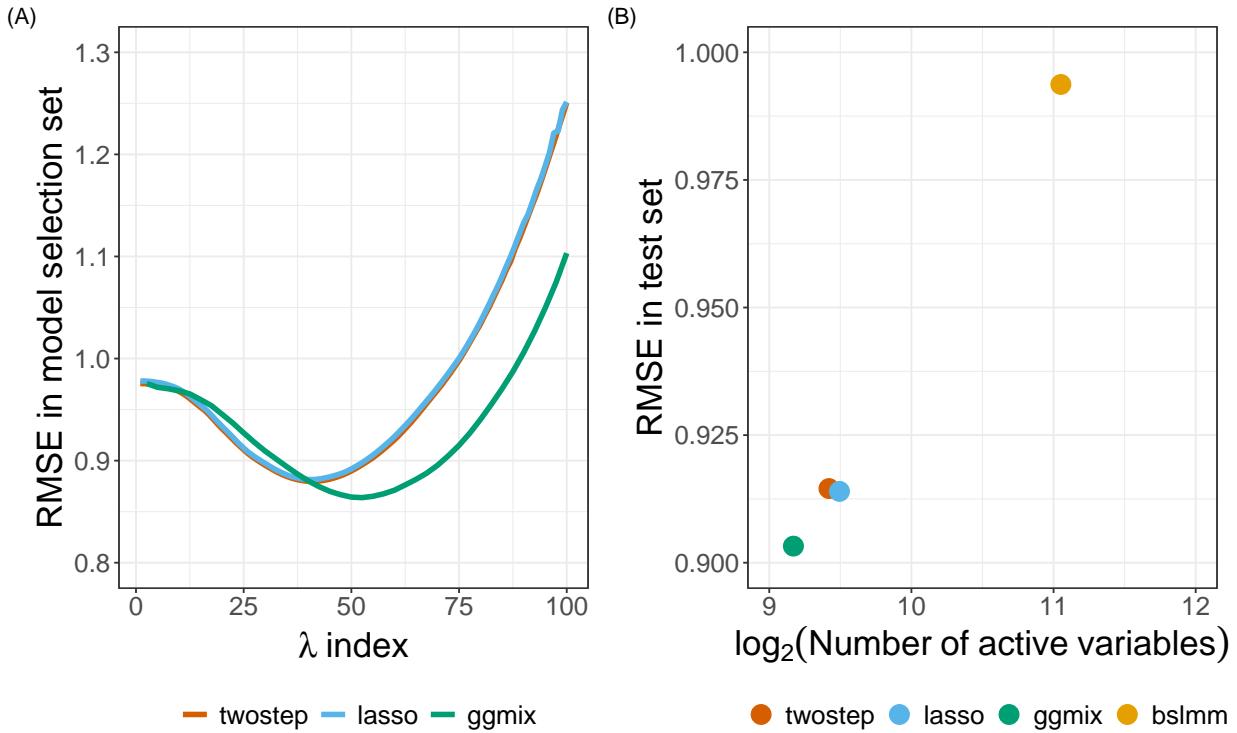


Figure 3: Model selection and testing in the UK Biobank. (A) Root-mean-square error of three methods on the model selection set with respect to a grid search of penalty factor used on the training set. (B) Performance of four methods on the test set with penalty factor optimized on the model selection set. The x-axis has a logarithmic scale. The BSLMM method optimized coefficients of each SNP through an MCMC process on the training set and was directly evaluated on the test set.

### 3.2.2 GAW20

In the most recent Genetic Analysis Workshop 20 (GAW20), the causal modeling group investigated causal relationships between DNA methylation (exposure) within some genes and the change in high-density lipoproteins  $\Delta$ HDL (outcome) using Mendelian Randomization (MR) [36]. Penalized regression methods were used to select SNPs strongly associated with the exposure in order to be used as an instrumental variable (IV) [37, 38]. However, since GAW20 data consisted of families, **twostep** methods were used which could have resulted

274 in a large number of false positives or false negatives. `ggmix` now provides an alternative  
275 approach that could be used for selecting the IV while accounting for the family structure  
276 of the data.

277 We applied `ggmix` to all 200 GAW20 simulation datasets, each of 679 observations, and  
278 compared its performance to the `twostep` and `lasso` methods. Using a Factored Spectrally  
279 Transformed Linear Mixed Model (FaST-LMM) [39] adjusted for age and sex, we validated  
280 the effect of rs9661059 on blood lipid trait to be significant (genome-wide  $p = 6.29 \times 10^{-9}$ ).  
281 Though several other SNPs were also associated with the phenotype, these associations were  
282 probably mediated by CpG-SNP interaction pairs and did not reach statistical significance.  
283 Therefore, to avoid ambiguity, we only focused on chromosome 1 containing 51,104 SNPs,  
284 including rs9661059. Given that population admixture in the GAW20 data was likely, we  
285 estimated the population kinship using REAP [40] after decomposing population composi-  
286 tions using ADMIXTURE [41]. We used 100,276 LD-pruned whole-genome genotyped SNPs  
287 for estimating the kinship. Among these, 8100 were included as covariates in our models  
288 based on chromosome 1. The causal SNP was also among the 100,276 SNPs. All methods  
289 were fit according to the same settings described in our simulation study in Section 3.1,  
290 and adjusting for age and sex. We calculated the median (inter-quartile range) number of  
291 active variables, and RMSE (standard deviation) based on five-fold CV on each simulated  
292 dataset.

293 On each simulated replicate, we calibrated the methods so that they could be easily compared  
294 by fixing the true positive rate to 1 and then minimizing the false positive rate. Hence, the  
295 selected SNP, rs9661059, was likely to be the true positive for each method, and non-causal  
296 SNPs were excluded to the greatest extent. All three methods precisely chose the correct  
297 predictor without any false positives in more than half of the replicates, as the causal signal  
298 was strong. However, when some false positives were selected (i.e. when the number of active  
299 variables  $> 1$ ), `ggmix` performed comparably to `twostep`, while the `lasso` was inclined to

300 select more false positives as suggested by the larger third quartile number of active variables  
301 (Table 2). We also observed that `gmmix` outperformed the `twostep` method with lower CV  
302 RMSE using the same number of SNPs. Meanwhile, it achieved roughly the same prediction  
303 accuracy as `lasso` but with fewer non-causal SNPs (Table 2). It is also worth mentioning  
304 that there was very little correlation between the causal SNP and SNPs within a 1Mb-  
305 window around it (Figure B.2 in Supplemental Section B.2), making it an ideal scenario for  
306 the `lasso` and related methods.

307 We also applied the `BSLMM` method by performing five-fold CV on each of the 200 simu-  
308 lated replicates. We found that while `BSLMM` achieved a lower CV RMSE compared to the  
309 other methods (Table 2), this higher prediction accuracy relied on approximately 80% of the  
310 51,104 SNPs **with a positive posterior inclusion probability**. This may suggest over-  
311 fitting in this dataset. **We additionally tried imposing a stricter posterior inclusion**  
312 **probability threshold (0.05, 0.10 and 0.50) in order to improve feature selection.**  
313 **These thresholds however, resulted in overly sparse models as most SNPs had a**  
314 **low posterior probability.** It is also noteworthy that we did not adjust for age and sex  
315 in the `BSLMM` model, as the current implementation of the method in the `GEMMA` package  
316 does not allow adjustment for covariates.

Table 2: Summary of model performance based on 200 GAW20 simulations for the `twostep`, `lasso`, `gmmix` and `BSLMM` model with different posterior inclusion probability (PIP) thresholds. Five-fold cross-validation root-mean-square error (RMSE) was reported for each simulation replicate. Prediction performance was not reported for `BSLMM` with PIP greater than 0.05, 0.10 and 0.50 because some of the replications contained no active SNPs.

Method	Median number of active variables (Inter-quartile range)	RMSE (SD)
<code>twostep</code>	1 (1 - 11)	0.3604 (0.0242)
<code>lasso</code>	1 (1 - 15)	0.3105 (0.0199)
<code>gmmix</code>	1 (1 - 12)	0.3146 (0.0210)
<code>BSLMM</code> (PIP > 0)	40,737 (39,901 - 41,539)	0.2503 (0.0099)
<code>BSLMM</code> (PIP > 0.05)	2 (1 - 4)	
<code>BSLMM</code> (PIP > 0.10)	0 (0 - 1)	
<code>BSLMM</code> (PIP > 0.50)	0 (0 - 0)	

### 3.2.3 Mouse Crosses and Sensitivity to Mycobacterial Infection

Mouse inbred strains of genetically identical individuals are extensively used in research. Crosses of different inbred strains are useful for various studies of heritability focusing on either observable phenotypes or molecular mechanisms, and in particular, recombinant congenic strains have been an extremely useful resource for many years [42]. However, ignoring complex genetic relationships in association studies can lead to inflated false positives in genetic association studies when different inbred strains and their crosses are investigated [43, 44, 45]. Therefore, a previous study developed and implemented a mixed model to find loci associated with mouse sensitivity to mycobacterial infection [46]. The random effects in the model captured complex correlations between the recombinant congenic mouse strains based on the proportion of the DNA shared identical by descent. Through a series of mixed model fits at each marker, new loci that impact growth of mycobacteria on

330 chromosome 1 and chromosome 11 were identified.

331 Here we show that `gmmix` can identify these loci, as well as potentially others, in a single  
332 analysis. We reanalyzed the growth permissiveness in the spleen, as measured by colony  
333 forming units (CFUs), 6 weeks after infection from *Mycobacterium bovis* Bacille Calmette-  
334 Guerin (BCG) Russia strain as reported in [46].

335 By taking the consensus between the “main model” and the “conditional model” of the original  
336 study, we regarded markers D1Mit435 on chromosome 1 and D11Mit119 on chromosome 11  
337 as two true positive loci. We directly estimated the kinship between mice using genotypes at  
338 625 microsatellite markers. The estimated kinship entered directly into `gmmix` and `twostep`.

339 For the `lasso`, we calculated and included the first 10 principal components of the estimated  
340 kinship. To evaluate the robustness of different models, we bootstrapped the 189-sample  
341 dataset and repeated the analysis 200 times. We then conceived a two-fold criteria to evaluate  
342 performance of each model. We first examined whether a model could pick up both true  
343 positive loci using some  $\lambda$ . If the model failed to pick up both loci simultaneously with any  
344  $\lambda$ , we counted as modeling failure on the corresponding bootstrap replicate; otherwise, we  
345 counted as modeling success and recorded which other loci were picked up given the largest  
346  $\lambda$ . Consequently, similar to the strategy used in the GAW20 analysis, we optimized the  
347 models by tuning the penalty factor such that these two true positive loci were picked up,  
348 while the number of other active loci was minimized. Significant markers were defined as  
349 those captured in at least half of the successful bootstrap replicates (Figure 4).

350 We demonstrated that `gmmix` recognized the true associations more robustly than `twostep`  
351 and `lasso`. In almost all (99%) bootstrap replicates, `gmmix` was able to capture both true  
352 positives, while the `twostep` failed in 19% of the replicates and the `lasso` failed in 56% of  
353 the replicates by missing at least one of the two true positives (Figure 4). The robustness  
354 of `gmmix` is particularly noteworthy due to the strong correlations between all microsatellite  
355 markers in this dataset (Figure B.3 in Supplemental Section B.2). These strong correlations

356 with the causal markers, partially explain the poor performance of the `lasso` as it suffers  
357 from unstable selections in the presence of correlated variables (e.g. [47]).

358 We also identified several other loci that might also be associated with susceptibility to my-  
359 cobacterial infection (Table 3). Among these new potentially-associated markers, D2Mit156  
360 was found to play a role in control of parasite numbers of *Leishmania tropica* in lymph  
361 nodes [48]. An earlier study identified a parent-of-origin effect at D17Mit221 on CD4M  
362 levels [49]. This effect was more visible in crosses than in parental strains. In addition,  
363 D14Mit131, selected only by `gmmix`, was found to have a 9% loss of heterozygosity in hy-  
364 brids of two inbred mouse strains [50], indicating the potential presence of putative suppressor  
365 genes pertaining to immune surveillance and tumor progression [51]. This result might also  
366 suggest association with anti-bacterial responses yet to be discovered.

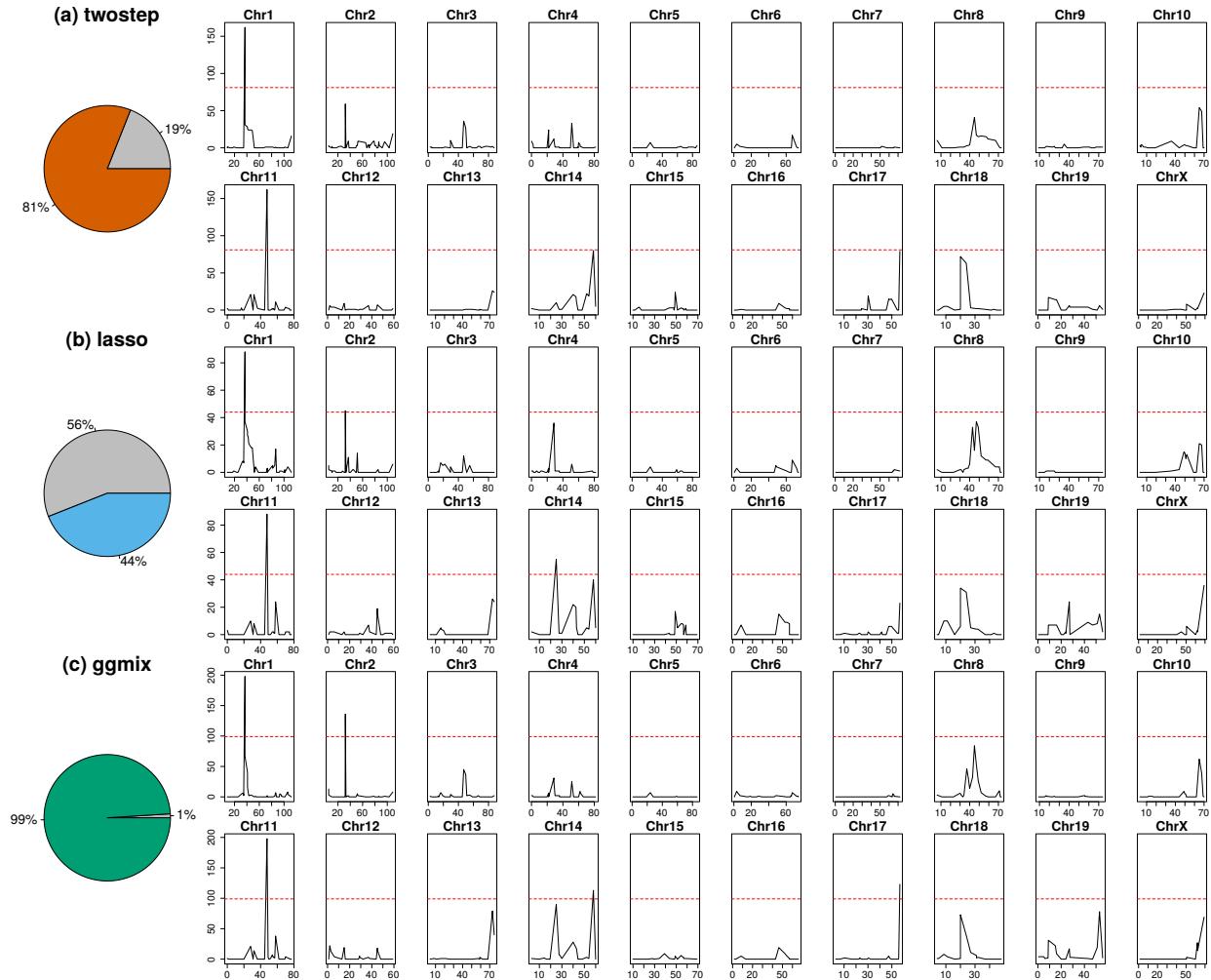


Figure 4: Comparison of model performance on the mouse cross data. Pie charts depict model robustness where grey areas denote bootstrap replicates on which the corresponding model is unable to capture both true positives using any penalty factor, whereas colored areas denote successful replicates. Chromosome-based signals record in how many successful replicates the corresponding loci are picked up by the corresponding optimized model. Red dashed lines delineate significance thresholds.

Table 3: Additional loci significantly associated with mouse susceptibility to myobacterial infection, after excluding two true positives. Loci needed to be identified in at least 50% of the successful bootstrap replicates that captured both true positive loci.

<b>Method</b>	<b>Marker</b>	<b>Position in cM</b>	<b>Position in bp</b>
<b>twostep</b>	N/A	N/A	N/A
<b>367 lasso</b>	D2Mit156	Chr2:31.66	Chr2:57081653-57081799
	D14Mit155	Chr14:31.52	Chr14:59828398-59828596
<b>ggmix</b>	D2Mit156	Chr2:31.66	Chr2:57081653-57081799
	D14Mit131	Chr14:63.59	Chr14:120006565-120006669
	D17Mit221	Chr17:59.77	Chr17:90087704-90087842

## 368 4 Discussion

369 We have developed a general penalized LMM framework called **ggmix** which simultaneously  
 370 selects SNPs and adjusts for population structure in high dimensional prediction models. We  
 371 compared our method to the **twostage** procedure, where in the first stage, the dependence  
 372 between observations is adjusted for in a LMM with a single random effect and no covariates  
 373 (i.e. null model). The residuals from this null model can then be used in any model for  
 374 independent observations because the relatedness has been effectively removed from the  
 375 original response. We also compared our method to the **lasso** and **BSLMM** which are closely  
 376 related to **ggmix** since they also jointly model the relatedness and SNPs in a single step.  
 377 The key differences are that the **lasso** uses a principal component adjustment and **BSLMM** is  
 378 a Bayesian method focused on phenotype prediction.

379 Through an extensive simulation study and three real data analyses that mimic many ex-  
 380 perimental designs in genetics, we show that the current approaches of PC adjustment and  
 381 two-stage procedures are not necessarily sufficient to control for confounding by population  
 382 structure leading to a high number of false positives. Our simulation results show that **ggmix**

383 outperforms existing methods in terms of sparsity and prediction error even when the causal  
384 variants are included in the kinship matrix (Table 1). Many methods for single-SNP analyses  
385 avoid this proximal contamination [8] by using a leave-one-chromosome-out scheme [52], i.e.,  
386 construct the kinship matrix using all chromosomes except the one on which the marker  
387 being tested is located. However, this approach is not possible if we want to model many  
388 SNPs (across many chromosomes) jointly to create, for example, a polygenic risk score. For  
389 the purposes of variable selection, we would also want to model all chromosomes together  
390 since the power to detect an association for a given SNP may be increased when other causal  
391 SNPs have been accounted for. Conversely, a stronger signal from a causal SNP may weaken  
392 false signals when modeled jointly [6], particularly when the markers are highly correlated  
393 as in the mouse crosses example.

394 In the UK Biobank, we found that with a kinship matrix estimated using all genotyped SNPs,  
395 `gmmix` had achieved a lower RMSE on the model selection set compared to the `twostep` and  
396 `lasso` methods. Furthermore, an optimized `gmmix`-derived polygenic risk score that utilized  
397 the least number of SNPs was also able to better predict the trait with lower RMSE on  
398 the test set. In the GAW20 example, we showed that while all methods were able to select  
399 the strongest causal SNP, `gmmix` did so with the least amount of false positives while also  
400 maintaining good predictive ability. In the mouse crosses example, we showed that `gmmix` is  
401 robust to perturbations in the data using a bootstrap analysis. Indeed, `gmmix` was able to  
402 consistently select the true positives across bootstrap replicates, while `twostep` failed in 19%  
403 of the replicates and `lasso` failed in 56% of the replicates by missing of at least one of the  
404 two true positives. Our re-analysis of the data also lead to some potentially new findings, not  
405 found by existing methods, that may warrant further study. This particular example had  
406 many markers that were strongly correlated with each other (Figure B.3 of Supplemental  
407 Section B.2). Nevertheless, we observed that the two true positive loci were the most often  
408 selected while none of the nearby markers were picked up in more than 50% of the 200  
409 bootstrap replicates. This shows that our method does recognize the true positives in the

410 presence of highly correlated markers. Nevertheless, we think the issue of variable selection  
411 for correlated SNPs warrants further study. The recently proposed Precision Lasso [47] seeks  
412 to address this problem in the high-dimensional fixed effects model.

413 We emphasize here that previously developed methods such as the LMM-lasso [15] use a two-  
414 stage fitting procedure without any convergence details. From a practical point of view, there  
415 is currently no implementation that provides a principled way of determining the sequence  
416 of tuning parameters to fit, nor a procedure that automatically selects the optimal value of  
417 the tuning parameter. To our knowledge, we are the first to develop a coordinate gradient  
418 descent (CGD) algorithm in the specific context of fitting a penalized LMM for population  
419 structure correction with theoretical guarantees of convergence. Furthermore, we develop  
420 a principled method for automatic tuning parameter selection and provide an easy-to-use  
421 software implementation in order to promote wider uptake of these more complex methods  
422 by applied practitioners.

423 Although we derive a CGD algorithm for the  $\ell_1$  penalty, our approach can also be easily ex-  
424 tended to other penalties such as the elastic net and group lasso with the same guarantees of  
425 convergence. A limitation of `gmmix` is that it first requires computing the covariance matrix  
426 with a computation time of  $\mathcal{O}(n^2k)$  followed by a spectral decomposition of this matrix in  
427  $\mathcal{O}(n^3)$  time where  $k$  is the number of SNP genotypes used to construct the covariance matrix.  
428 This computation becomes prohibitive for large cohorts such as the UK Biobank [53] which  
429 have collected genetic information on half a million individuals. When the matrix of geno-  
430 types used to construct the covariance matrix is low rank, there are additional computational  
431 speedups that can be implemented. While this has been developed for the univariate case [8],  
432 to our knowledge, this has not been explored in the multivariable case. We are currently  
433 developing a low rank version of the penalized LMM developed here, which reduces the time  
434 complexity from  $\mathcal{O}(n^2k)$  to  $\mathcal{O}(nk^2)$ . There is also the issue of how our model scales with  
435 an increasing number of covariates ( $p$ ). Due to the coordinate-wise optimization procedure,

436 we expect this to be less of an issue, but still prohibitive for  $p > 1e5$ . The `biglasso` pack-  
437 age [54] uses memory mapping strategies for large  $p$ , and this is something we are exploring  
438 for `gmmix`.

439 As was brought up by a reviewer, the simulations and real data analyses presented here  
440 contained many more markers used to estimate the kinship than the sample size ( $n/k \leq 0.1$ ).

441 In the single locus association test, Yang et al. [22] found that proximal contamination was  
442 an issue when  $n/k \approx 1$ . We believe further theoretical study is needed to see if these results  
443 can be generalized to the multivariable models being fit here. Once the computational  
444 limitations of sample size mentioned above have been addressed, these theoretical results  
445 can be supported by simulation studies.

446 There are other applications in which our method could be used as well. For example, there  
447 has been a renewed interest in polygenic risk scores (PRS) which aim to predict complex  
448 diseases from genotypes. `gmmix` could be used to build a PRS with the distinct advantage  
449 of modeling SNPs jointly, allowing for main effects as well as interactions to be accounted  
450 for. Based on our results, `gmmix` has the potential to produce more robust and parsimonious  
451 models than the `lasso` with better predictive accuracy. Our method is also suitable for fine  
452 mapping SNP association signals in genomic regions, where the goal is to pinpoint individual  
453 variants most likely to impact the underlying biological mechanisms of disease [55].

454 

## 5 Materials and Methods

455 

### 5.1 Model Set-up

456 Let  $i = 1, \dots, N$  be a grouping index,  $j = 1, \dots, n_i$  the observation index within a group  
457 and  $N_T = \sum_{i=1}^N n_i$  the total number of observations. For each group let  $\mathbf{y}_i = (y_1, \dots, y_{n_i})$  be  
458 the observed vector of responses or phenotypes,  $\mathbf{X}_i$  an  $n_i \times (p + 1)$  design matrix (with  
459 the column of 1s for the intercept),  $\mathbf{b}_i$  a group-specific random effect vector of length

460  $n_i$  and  $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$  the individual error terms. Denote the stacked vectors  $\mathbf{Y} =$   
461  $(\mathbf{y}_i, \dots, \mathbf{y}_N)^T \in \mathbb{R}^{N_T \times 1}$ ,  $\mathbf{b} = (\mathbf{b}_i, \dots, \mathbf{b}_N)^T \in \mathbb{R}^{N_T \times 1}$ ,  $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_i, \dots, \boldsymbol{\varepsilon}_N)^T \in \mathbb{R}^{N_T \times 1}$ , and the  
462 stacked matrix

463  $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_N^T) \in \mathbb{R}^{N_T \times (p+1)}$ . Furthermore, let  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T \in \mathbb{R}^{(p+1) \times 1}$  be a vec-  
464 tor of fixed effects regression coefficients corresponding to  $\mathbf{X}$ . We consider the following  
465 linear mixed model with a single random effect [56]:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{b} + \boldsymbol{\varepsilon} \quad (2)$$

466 where the random effect  $\mathbf{b}$  and the error variance  $\boldsymbol{\varepsilon}$  are assigned the distributions

$$\mathbf{b} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1 - \eta)\sigma^2\mathbf{I}) \quad (3)$$

467 Here,  $\boldsymbol{\Phi}_{N_T \times N_T}$  is a known positive semi-definite and symmetric covariance or kinship ma-  
468 trix calculated from SNPs sampled across the genome,  $\mathbf{I}_{N_T \times N_T}$  is the identity matrix and  
469 parameters  $\sigma^2$  and  $\eta \in [0, 1]$  determine how the variance is divided between  $\mathbf{b}$  and  $\boldsymbol{\varepsilon}$ . Note  
470 that  $\eta$  is also the narrow-sense heritability ( $h^2$ ), defined as the proportion of phenotypic  
471 variance attributable to the additive genetic factors [1]. The joint density of  $\mathbf{Y}$  is therefore  
472 multivariate normal:

$$\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\boldsymbol{\Phi} + (1 - \eta)\sigma^2\mathbf{I}) \quad (4)$$

473 The LMM-Lasso method [15] considers an alternative but equivalent parameterization given  
474 by:

$$\mathbf{Y}|(\boldsymbol{\beta}, \delta, \sigma_g^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_g^2(\boldsymbol{\Phi} + \delta\mathbf{I})) \quad (5)$$

where  $\delta = \sigma_e^2/\sigma_g^2$ ,  $\sigma_g^2$  is the genetic variance and  $\sigma_e^2$  is the residual variance. We instead  
consider the parameterization in (4) since maximization is easier over the compact set  $\eta \in$

[0, 1] than over the unbounded interval  $\delta \in [0, \infty)$  [56]. We define the complete parameter vector as  $\Theta := (\beta, \eta, \sigma^2)$ . The negative log-likelihood for (4) is given by

$$-\ell(\Theta) \propto \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \log(\det(\mathbf{V})) + \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta) \quad (6)$$

475 where  $\mathbf{V} = \eta\Phi + (1 - \eta)\mathbf{I}$  and  $\det(\mathbf{V})$  is the determinant of  $\mathbf{V}$ .

Let  $\Phi = \mathbf{U}\mathbf{D}\mathbf{U}^T$  be the eigen (spectral) decomposition of the kinship matrix  $\Phi$ , where  $\mathbf{U}_{N_T \times N_T}$  is an orthonormal matrix of eigenvectors (i.e.  $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ ) and  $\mathbf{D}_{N_T \times N_T}$  is a diagonal matrix of eigenvalues  $\Lambda_i$ .  $\mathbf{V}$  can then be further simplified [56]

$$\begin{aligned} \mathbf{V} &= \eta\Phi + (1 - \eta)\mathbf{I} \\ &= \eta\mathbf{U}\mathbf{D}\mathbf{U}^T + (1 - \eta)\mathbf{U}\mathbf{I}\mathbf{U}^T \\ &= \mathbf{U}\eta\mathbf{D}\mathbf{U}^T + \mathbf{U}(1 - \eta)\mathbf{I}\mathbf{U}^T \\ &= \mathbf{U}(\eta\mathbf{D} + (1 - \eta)\mathbf{I})\mathbf{U}^T \\ &= \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T \end{aligned} \quad (7)$$

where

$$\tilde{\mathbf{D}} = \eta \mathbf{D} + (1 - \eta) \mathbf{I} \quad (8)$$

$$\begin{aligned} &= \eta \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_{N_T} \end{bmatrix} + (1 - \eta) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 + \eta(\Lambda_1 - 1) & & & \\ & 1 + \eta(\Lambda_2 - 1) & & \\ & & \ddots & \\ & & & 1 + \eta(\Lambda_{N_T} - 1) \end{bmatrix} \\ &= \text{diag}\{1 + \eta(\Lambda_1 - 1), 1 + \eta(\Lambda_2 - 1), \dots, 1 + \eta(\Lambda_{N_T} - 1)\} \end{aligned} \quad (9)$$

Since (8) is a diagonal matrix, its inverse is also a diagonal matrix:

$$\tilde{\mathbf{D}}^{-1} = \text{diag} \left\{ \frac{1}{1 + \eta(\Lambda_1 - 1)}, \frac{1}{1 + \eta(\Lambda_2 - 1)}, \dots, \frac{1}{1 + \eta(\Lambda_{N_T} - 1)} \right\} \quad (10)$$

From (7) and (9),  $\log(\det(\mathbf{V}))$  simplifies to

$$\begin{aligned} \log(\det(\mathbf{V})) &= \log \left( \det(\mathbf{U}) \det(\tilde{\mathbf{D}}) \det(\mathbf{U}^T) \right) \\ &= \log \left\{ \prod_{i=1}^{N_T} (1 + \eta(\Lambda_i - 1)) \right\} \\ &= \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) \end{aligned} \quad (11)$$

since  $\det(\mathbf{U}) = 1$ . It also follows from (7) that

$$\begin{aligned}\mathbf{V}^{-1} &= \left( \mathbf{U} \tilde{\mathbf{D}} \mathbf{U}^T \right)^{-1} \\ &= (\mathbf{U}^T)^{-1} \left( \tilde{\mathbf{D}} \right)^{-1} \mathbf{U}^{-1} \\ &= \mathbf{U} \tilde{\mathbf{D}}^{-1} \mathbf{U}^T\end{aligned}\tag{12}$$

since for an orthonormal matrix  $\mathbf{U}^{-1} = \mathbf{U}^T$ . Substituting (10), (11) and (12) into (6) the negative log-likelihood becomes

$$\begin{aligned}-\ell(\boldsymbol{\Theta}) &\propto \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{U} \tilde{\mathbf{D}}^{-1} \mathbf{U}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &\quad (13)\end{aligned}$$

$$\begin{aligned}&= \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} (\mathbf{U}^T \mathbf{Y} - \mathbf{U}^T \mathbf{X}\boldsymbol{\beta})^T \tilde{\mathbf{D}}^{-1} (\mathbf{U}^T \mathbf{Y} - \mathbf{U}^T \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) \\ &= \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} \sum_{i=1}^{N_T} \frac{\left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2}{1 + \eta(\Lambda_i - 1)}\end{aligned}\tag{14}$$

where  $\tilde{\mathbf{Y}} = \mathbf{U}^T \mathbf{Y}$ ,  $\tilde{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$ ,  $\tilde{Y}_i$  denotes the  $i^{\text{th}}$  element of  $\tilde{\mathbf{Y}}$ ,  $\tilde{X}_{ij}$  is the  $i, j^{\text{th}}$  entry of  $\tilde{\mathbf{X}}$

and  $\mathbf{1}$  is a column vector of  $N_T$  ones.

## 5.2 Penalized Maximum Likelihood Estimator

We define the  $p + 3$  length vector of parameters  $\boldsymbol{\Theta} := (\Theta_0, \Theta_1, \dots, \Theta_{p+1}, \Theta_{p+2}, \Theta_{p+3}) = (\boldsymbol{\beta}, \eta, \sigma^2)$  where  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ ,  $\eta \in [0, 1]$ ,  $\sigma^2 > 0$ . In what follows,  $p + 2$  and  $p + 3$  are the indices in  $\boldsymbol{\Theta}$  for  $\eta$  and  $\sigma^2$ , respectively. In light of our goals to select variables associated with the response in high-dimensional data, we propose to place a constraint on the magnitude of the regression coefficients. This can be achieved by adding a penalty term to the likelihood function (14). The penalty term is a necessary constraint because in our applications, the

485 sample size is much smaller than the number of predictors. We define the following objective  
 486 function:

$$Q_\lambda(\Theta) = f(\Theta) + \lambda \sum_{j \neq 0} v_j P_j(\beta_j) \quad (15)$$

487 where  $f(\Theta) := -\ell(\Theta)$  is defined in (14),  $P_j(\cdot)$  is a penalty term on the fixed regression  
 488 coefficients  $\beta_1, \dots, \beta_{p+1}$  (we do not penalize the intercept) controlled by the nonnegative  
 489 regularization parameter  $\lambda$ , and  $v_j$  is the penalty factor for  $j$ th covariate. These penalty  
 490 factors serve as a way of allowing parameters to be penalized differently. Note that we do  
 491 not penalize  $\eta$  or  $\sigma^2$ . An estimate of the regression parameters  $\widehat{\Theta}_\lambda$  is obtained by

$$\widehat{\Theta}_\lambda = \arg \min_{\Theta} Q_\lambda(\Theta) \quad (16)$$

492 This is the general set-up for our model. In Section 5.3 we provide more specific details on  
 493 how we solve (16). **We note here that the main difference between the proposed**  
**494 model, and the lmlasso [57], is that we rotate the response vector  $Y$  and the**  
**495 design matrix  $X$  by the eigen vectors of the kinship matrix.** This results in a  
 496 **diagonal covariance matrix making our method orders of magnitude faster and**  
 497 **usable for high-dimensional genetic data.** A secondary difference is that we are  
 498 **limiting ourselves to a single unpenalized random effect.**

### 499 5.3 Computational Algorithm

500 We use a general purpose block coordinate gradient descent algorithm (CGD) [58] to solve (16).  
 501 At each iteration, we cycle through the coordinates and minimize the objective function with  
 502 respect to one coordinate only. For continuously differentiable  $f(\cdot)$  and convex and block-  
 503 separable  $P(\cdot)$  (i.e.  $P(\beta) = \sum_i P_i(\beta_i)$ ), Tseng and Yun [58] show that the solution generated  
 504 by the CGD method is a stationary point of  $Q_\lambda(\cdot)$  if the coordinates are updated in a  
 505 Gauss-Seidel manner i.e.  $Q_\lambda(\cdot)$  is minimized with respect to one parameter while holding  
 506 all others fixed. The CGD algorithm has been successfully applied in fixed effects models

507 (e.g. [59], [20]) and linear mixed models with an  $\ell_1$  penalty [57]. In the next section we  
 508 provide some brief details about Algorithm 1. A more thorough treatment of the algorithm  
 509 is given in Appendix A.

---

**Algorithm 1:** Block Coordinate Gradient Descent

---

Set the iteration counter  $k \leftarrow 0$ , initial values for the parameter vector  $\Theta^{(0)}$  and convergence threshold  $\epsilon$ ;

**for**  $\lambda \in \{\lambda_{\max}, \dots, \lambda_{\min}\}$  **do**

**repeat**

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg \min_{\boldsymbol{\beta}} Q_\lambda \left( \boldsymbol{\beta}, \eta^{(k)}, \sigma^2^{(k)} \right)$$

$$\eta^{(k+1)} \leftarrow \arg \min_{\eta} Q_\lambda \left( \boldsymbol{\beta}^{(k+1)}, \eta, \sigma^2^{(k)} \right)$$

$$\sigma^2^{(k+1)} \leftarrow \arg \min_{\sigma^2} Q_\lambda \left( \boldsymbol{\beta}^{(k+1)}, \eta^{(k+1)}, \sigma^2 \right)$$

$$k \leftarrow k + 1$$

**until** convergence criterion is satisfied:  $\|\Theta^{(k+1)} - \Theta^{(k)}\|_2 < \epsilon$ ;

**end**

---

510 **5.3.1 Updates for the  $\beta$  parameter**

511 Recall that the part of the objective function that depends on  $\beta$  has the form

$$Q_\lambda(\Theta) = \frac{1}{2} \sum_{i=1}^{N_T} w_i \left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2 + \lambda \sum_{j=1}^p v_j |\beta_j| \quad (17)$$

512 where

$$w_i := \frac{1}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \quad (18)$$

Conditional on  $\eta^{(k)}$  and  $\sigma^2^{(k)}$ , it can be shown that the solution for  $\beta_j$ ,  $j = 1, \dots, p$  is given by

$$\beta_j^{(k+1)} \leftarrow \frac{\mathcal{S}_\lambda \left( \sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) \right)}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \quad (19)$$

where  $\mathcal{S}_\lambda(x)$  is the soft-thresholding operator

$$\mathcal{S}_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$$

513  $\text{sign}(x)$  is the signum function

$$\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

514 and  $(x)_+ = \max(x, 0)$ . We provide the full derivation in Appendix A.1.2.

515 **5.3.2 Updates for the  $\eta$  parameter**

516 Given  $\beta^{(k+1)}$  and  $\sigma^{2(k)}$ , solving for  $\eta^{(k+1)}$  becomes a univariate optimization problem:

$$\eta^{(k+1)} \leftarrow \arg \min_{\eta} \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^{2(k)}} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j^{(k+1)}\right)^2}{1 + \eta(\Lambda_i - 1)} \quad (20)$$

517 We use a bound constrained optimization algorithm [60] implemented in the `optim` function

518 in R and set the lower and upper bounds to be 0.01 and 0.99, respectively.

519 **5.3.3 Updates for the  $\sigma^2$  parameter**

520 Conditional on  $\beta^{(k+1)}$  and  $\eta^{(k+1)}$ ,  $\sigma^{2(k+1)}$  can be solved for using the following equation:

$$\sigma^{2(k+1)} \leftarrow \arg \min_{\sigma^2} \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j\right)^2}{1 + \eta(\Lambda_i - 1)} \quad (21)$$

There exists an analytic solution for (21) given by:

$$\sigma^2(k+1) \leftarrow \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j^{(k+1)}\right)^2}{1 + \eta^{(k+1)}(\Lambda_i - 1)} \quad (22)$$

#### 5.3.4 Regularization path

In this section we describe how determine the sequence of tuning parameters  $\lambda$  at which to fit the model. Recall that our objective function has the form

$$Q_\lambda(\Theta) = \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2} \sum_{i=1}^{N_T} w_i \left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2 + \lambda \sum_{j=1}^p v_j |\beta_j| \quad (23)$$

The Karush-Kuhn-Tucker (KKT) optimality conditions for (23) are given by:

$$\begin{aligned} \frac{\partial}{\partial \beta_1, \dots, \beta_p} Q_\lambda(\Theta) &= \mathbf{0}_p \\ \frac{\partial}{\partial \beta_0} Q_\lambda(\Theta) &= 0 \\ \frac{\partial}{\partial \eta} Q_\lambda(\Theta) &= 0 \\ \frac{\partial}{\partial \sigma^2} Q_\lambda(\Theta) &= 0 \end{aligned} \quad (24)$$

525 The equations in (24) are equivalent to

$$\begin{aligned}
 & \sum_{i=1}^{N_T} w_i \tilde{X}_{i1} \left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right) = 0 \\
 & \frac{1}{v_j} \sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right) = \lambda \gamma_j, \\
 & \gamma_j \in \begin{cases} \text{sign}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_j = 0 \end{cases}, \quad \text{for } j = 1, \dots, p \\
 & \frac{1}{2} \sum_{i=1}^{N_T} \frac{\Lambda_i - 1}{1 + \eta(\Lambda_i - 1)} \left( 1 - \frac{\left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \right) = 0 \\
 & \sigma^2 - \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left( \tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2}{1 + \eta(\Lambda_i - 1)} = 0
 \end{aligned} \tag{25}$$

526 where  $w_i$  is given by (18),  $\tilde{\mathbf{X}}_{-1}^T$  is  $\tilde{\mathbf{X}}^T$  with the first column removed,  $\tilde{\mathbf{X}}_1^T$  is the first column  
 527 of  $\tilde{\mathbf{X}}^T$ , and  $\gamma \in \mathbb{R}^p$  is the subgradient function of the  $\ell_1$  norm evaluated at  $(\hat{\beta}_1, \dots, \hat{\beta}_p)$ .

528 Therefore  $\widehat{\Theta}$  is a solution in (16) if and only if  $\widehat{\Theta}$  satisfies (25) for some  $\gamma$ . We can determine  
 529 a decreasing sequence of tuning parameters by starting at a maximal value for  $\lambda = \lambda_{max}$   
 530 for which  $\hat{\beta}_j = 0$  for  $j = 1, \dots, p$ . In this case, the KKT conditions in (25) are equivalent  
 531 to

$$\begin{aligned}
 & \frac{1}{v_j} \sum_{i=1}^{N_T} \left| w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \tilde{X}_{i1} \beta_0 \right) \right| \leq \lambda, \quad \forall j = 1, \dots, p \\
 & \beta_0 = \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{i1} \tilde{Y}_i}{\sum_{i=1}^{N_T} w_i \tilde{X}_{i1}^2} \\
 & \frac{1}{2} \sum_{i=1}^{N_T} \frac{\Lambda_i - 1}{1 + \eta(\Lambda_i - 1)} \left( 1 - \frac{\left( \tilde{Y}_i - \tilde{X}_{i1} \beta_0 \right)^2}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \right) = 0 \\
 & \sigma^2 = \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left( \tilde{Y}_i - \tilde{X}_{i1} \beta_0 \right)^2}{1 + \eta(\Lambda_i - 1)}
 \end{aligned} \tag{26}$$

532 We can solve the KKT system of equations in (26) (with a numerical solution for  $\eta$ ) in order

533 to have an explicit form of the stationary point  $\widehat{\Theta}_0 = \left\{ \widehat{\beta}_0, \mathbf{0}_p, \widehat{\eta}, \widehat{\sigma}^2 \right\}$ . Once we have  $\widehat{\Theta}_0$ , we  
534 can solve for the smallest value of  $\lambda$  such that the entire vector  $(\widehat{\beta}_1, \dots, \widehat{\beta}_p)$  is 0:

$$\lambda_{max} = \max_j \left\{ \left| \frac{1}{v_j} \sum_{i=1}^{N_T} \widehat{w}_i \widetilde{X}_{ij} (\widetilde{Y}_i - \widetilde{X}_{i1} \widehat{\beta}_0) \right| \right\}, \quad j = 1, \dots, p \quad (27)$$

535 Following Friedman et al. [20], we choose  $\tau \lambda_{max}$  to be the smallest value of tuning parameters  
536  $\lambda_{min}$ , and construct a sequence of  $K$  values decreasing from  $\lambda_{max}$  to  $\lambda_{min}$  on the log scale.  
537 The defaults are set to  $K = 100$ ,  $\tau = 0.01$  if  $n < p$  and  $\tau = 0.001$  if  $n \geq p$ .

538 **5.3.5 Warm Starts**

539 The way in which we have derived the sequence of tuning parameters using the KKT con-  
540 ditions, allows us to implement warm starts. That is, the solution  $\widehat{\Theta}$  for  $\lambda_k$  is used as the  
541 initial value  $\Theta^{(0)}$  for  $\lambda_{k+1}$ . This strategy leads to computational speedups and has been  
542 implemented in the `ggmix` R package.

543 **5.3.6 Prediction of the random effects**

544 We use an empirical Bayes approach (e.g. [61]) to predict the random effects  $\mathbf{b}$ . Let the  
545 maximum a posteriori (MAP) estimate be defined as

$$\widehat{\mathbf{b}} = \arg \max_{\mathbf{b}} f(\mathbf{b} | \mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2) \quad (28)$$

where, by using Bayes rule,  $f(\mathbf{b}|\mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2)$  can be expressed as

$$\begin{aligned} f(\mathbf{b}|\mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2) &= \frac{f(\mathbf{Y}|\mathbf{b}, \boldsymbol{\beta}, \eta, \sigma^2)\pi(\mathbf{b}|\eta, \sigma^2)}{f(\mathbf{Y}|\boldsymbol{\beta}, \eta, \sigma^2)} \\ &\propto f(\mathbf{Y}|\mathbf{b}, \boldsymbol{\beta}, \eta, \sigma^2)\pi(\mathbf{b}|\eta, \sigma^2) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) - \frac{1}{2\eta\sigma^2}\mathbf{b}^T\boldsymbol{\Phi}^{-1}\mathbf{b} \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \frac{1}{\eta}\mathbf{b}^T\boldsymbol{\Phi}^{-1}\mathbf{b} \right] \right\} \end{aligned} \quad (29)$$

Solving for (28) is equivalent to minimizing the exponent in (29):

$$\hat{\mathbf{b}} = \arg \min_{\mathbf{b}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \frac{1}{\eta}\mathbf{b}^T\boldsymbol{\Phi}^{-1}\mathbf{b} \right\} \quad (30)$$

Taking the derivative of (30) with respect to  $\mathbf{b}$  and setting it to 0 we get:

$$\begin{aligned} 0 &= -2(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \frac{2}{\eta}\boldsymbol{\Phi}^{-1}\mathbf{b} \\ &= -(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{b} + \left( \frac{1}{\eta}\boldsymbol{\Phi}^{-1} \right) \mathbf{b} \\ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) &= \left( \mathbf{I}_{N_T \times N_T} + \frac{1}{\eta}\boldsymbol{\Phi}^{-1} \right) \mathbf{b} \\ \hat{\mathbf{b}} &= \left( \mathbf{I}_{N_T \times N_T} + \frac{1}{\hat{\eta}}\boldsymbol{\Phi}^{-1} \right)^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \left( \mathbf{I}_{N_T \times N_T} + \frac{1}{\hat{\eta}}\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T \right)^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \quad (31)$$

546 where  $(\hat{\boldsymbol{\beta}}, \hat{\eta})$  are the estimates obtained from Algorithm 1.

### 547 5.3.7 Phenotype prediction

548 Here we describe the method used for predicting the unobserved phenotype  $\mathbf{Y}^*$  in a set of  
 549 individuals with predictor set  $\mathbf{X}^*$  that were not used in the model training e.g. a testing  
 550 set. Let  $q$  denote the number of observations in the testing set and  $N - q$  the number of  
 551 observations in the training set. We assume that a `gmmix` model has been fit on a set of

552 training individuals with observed phenotype  $\mathbf{Y}$  and predictor set  $\mathbf{X}$ . We further assume  
 553 that  $\mathbf{Y}$  and  $\mathbf{Y}^*$  are jointly multivariate Normal:

$$\begin{bmatrix} \mathbf{Y}^* \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \boldsymbol{\mu}_{1(q \times 1)} \\ \boldsymbol{\mu}_{2(N-q) \times 1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11(q \times q)} & \boldsymbol{\Sigma}_{12_{q \times (N-q)}} \\ \boldsymbol{\Sigma}_{21_{(N-q) \times q}} & \boldsymbol{\Sigma}_{22_{(N-q) \times (N-q)}} \end{bmatrix} \right) \quad (32)$$

554 Then, from standard multivariate Normal theory, the conditional distribution  $\mathbf{Y}^* | \mathbf{Y}, \eta, \sigma^2, \boldsymbol{\beta}, \mathbf{X}, \mathbf{X}^*$   
 555 is  $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$  where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{Y} - \boldsymbol{\mu}_2) \quad (33)$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \quad (34)$$

556 The phenotype prediction is thus given by:

$$\boldsymbol{\mu}_{q \times 1}^* = \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \boldsymbol{\Sigma}_{12} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}) \quad (35)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \boldsymbol{\Sigma}_{12} \mathbf{U} \tilde{\mathbf{D}}^{-1} \mathbf{U}^T (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}) \quad (36)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \boldsymbol{\Sigma}_{12} \mathbf{U} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \quad (37)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \eta \sigma^2 \boldsymbol{\Phi}^* \mathbf{U} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \quad (38)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \eta \boldsymbol{\Phi}^* \mathbf{U} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \quad (39)$$

557 where  $\boldsymbol{\Phi}^*$  is the  $q \times (N - q)$  covariance matrix between the testing and training individu-  
 558 als.

559 **5.3.8 Choice of the optimal tuning parameter**

560 In order to choose the optimal value of the tuning parameter  $\lambda$ , we use the generalized  
561 information criterion [62] (GIC):

$$GIC_\lambda = -2\ell(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\eta}) + a_n \cdot \hat{df}_\lambda \quad (40)$$

562 where  $\hat{df}_\lambda$  is the number of non-zero elements in  $\hat{\boldsymbol{\beta}}_\lambda$  [63] plus two (representing the variance  
563 parameters  $\eta$  and  $\sigma^2$ ). Several authors have used this criterion for variable selection in mixed  
564 models with  $a_n = \log N_T$  [57, 64], which corresponds to the BIC. We instead choose the high-  
565 dimensional BIC [65] given by  $a_n = \log(\log(N_T)) * \log(p)$ . This is the default choice in our  
566 `ggmix` R package, though the interface is flexible to allow the user to select their choice of  
567  $a_n$ .

568 **Availability of data and material**

- 569 1. The UK Biobank data is available upon successful project application.
- 570 2. The GAW20 data is freely available upon request from <https://www.gaworkshop.org/data-sets>.
- 571 3. Mouse cross data is available from GitHub at <https://github.com/sahirbhatnagar/ggmix/blob/pgen/RealData/mice.RData>.
- 572 4. The entire simulation study is reproducible. Source code available at <https://github.com/sahirbhatnagar/ggmix/tree/pgen/simulation>. This includes scripts for ggmix, lasso and twostep methods.
- 573 5. The R package `ggmix` is freely available from CRAN at <https://cran.r-project.org/package=ggmix>.
- 574 6. A website describing how to use the package is available at <https://sahirbhatnagar.com/ggmix/>.
- 575
- 576
- 577
- 578
- 579
- 580

581 **Competing interests**

582 The authors declare that they have no competing interests.

583 **Author's contributions**

584 SRB, KO, YY and CMTG conceived the idea. SRB developed the algorithms, software  
585 and simulation study. TL completed the real data analysis. ES and JCLO provided data  
586 and interpretations. SRB, TL and CMTG wrote a draft of the manuscript then all authors  
587 edited, read and approved the final manuscript.

588 **Acknowledgements**

589 SRB was supported by the Ludmer Centre for Neuroinformatics and Mental Health and  
590 the Canadian Institutes for Health Research PJT 148620. This research was enabled in  
591 part by support provided by Calcul Québec ([www.calculquebec.ca](http://www.calculquebec.ca)) and Compute Canada  
592 ([www.computecanada.ca](http://www.computecanada.ca)). The funders had no role in study design, data collection and  
593 analysis, decision to publish, or preparation of the manuscript.

594 **Supporting Information**

595 Contains the following sections:

596     **A Block Coordinate Descent Algorithm** - a detailed description of the algorithm  
597         used to fit our `gmmix` model.

598     **B Additional Real Data Analysis Results** - supporting information for the GAW20  
599         and UK Biobank analyses

600     **C gmmix Package Showcase** - a vignette describing how to use our `gmmix` R package

601 **References**

602 [1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding  
603 the missing heritability of complex diseases. *Nature*. 2009;461(7265):747. [3](#), [25](#)

604 [2] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common  
605 SNPs explain a large proportion of the heritability for human height. *Nature genetics*.  
606 2010;42(7):565. [3](#)

607 [3] Astle W, Balding DJ, et al. Population structure and cryptic relatedness in genetic  
608 association studies. *Statistical Science*. 2009;24(4):451–471. [3](#), [4](#)

- 609 [4] Song M, Hao W, Storey JD. Testing for genetic associations in arbitrarily structured  
610 populations. *Nature genetics*. 2015;47(5):550–554. 3
- 611 [5] Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population  
612 structure on large genetic association studies. *Nature genetics*. 2004;36(5):512. 3
- 613 [6] Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in  
614 genome-wide and re-sequencing association studies. *PLoS genetics*. 2008;4(7):e1000130.  
615 4, 22
- 616 [7] Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies.  
617 *Bioinformatics*. 2010;27(4):516–523. 4
- 618 [8] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear  
619 mixed models for genome-wide association studies. *Nature methods*. 2011;8(10):833–  
620 835. 4, 7, 22, 23
- 621 [9] Kang HM, Sul JH, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, et al. Variance  
622 component model to account for sample structure in genome-wide association studies.  
623 *Nature genetics*. 2010;42(4):348. 4
- 624 [10] Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-  
625 model method for association mapping that accounts for multiple levels of relatedness.  
626 *Nature genetics*. 2006;38(2):203. 4
- 627 [11] Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell  
628 HJ, et al. Comparison of methods to account for relatedness in genome-wide association  
629 studies with family-based data. *PLoS Genet*. 2014;10(7):e1004445. 4
- 630 [12] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Princi-  
631 pal components analysis corrects for stratification in genome-wide association studies.  
632 *Nature genetics*. 2006;38(8):904. 4

- 633 [13] Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, et al. Ad-  
634 justed sequence kernel association test for rare variants controlling for cryptic and family  
635 relatedness. *Genetic epidemiology*. 2013;37(4):366–376. 4, 5
- 636 [14] Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the  
637 relative effects of polymorphisms within a gene using case/control or family data:  
638 application to HLA in type 1 diabetes. *The American Journal of Human Genetics*.  
639 2002;70(1):124–141. 4
- 640 [15] Rakitsch B, Lippert C, Stegle O, Borgwardt K. A Lasso multi-marker mixed  
641 model for association mapping with population structure correction. *Bioinformatics*.  
642 2013;29(2):206–214. 4, 23, 25
- 643 [16] Wang D, Eskridge KM, Crossa J. Identifying QTLs and epistasis in structured plant  
644 populations using adaptive mixed LASSO. *Journal of agricultural, biological, and en-*  
645 *vironmental statistics*. 2011;16(2):170–184. 4
- 646 [17] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal*  
647 *Statistical Society Series B (Methodological)*. 1996;p. 267–288. 4, 5
- 648 [18] Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical*  
649 *association*. 2006;101(476):1418–1429. 4
- 650 [19] Ding X, Su S, Nandakumar K, Wang X, Fardo DW. A 2-step penalized regression  
651 method for family-based next-generation sequencing association studies. In: *BMC pro-*  
652 *ceedings*. vol. 8. BioMed Central; 2014. p. S25. 5
- 653 [20] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models  
654 via coordinate descent. *Journal of statistical software*. 2010;33(1):1. 5, 6, 30, 34, 47
- 655 [21] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning  
656 problems. *Statistics and Computing*. 2015;25(6):1129–1141. 5

- 657 [22] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in  
658 the application of mixed-model association methods. *Nature genetics*. 2014;46(2):100.  
659 5, 24
- 660 [23] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of*  
661 *the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301–320.  
662 5
- 663 [24] Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algo-  
664 rithm for variance parameter estimation in linear mixed models. *Biometrics*. 1995;p.  
665 1440–1450. 6
- 666 [25] Dandine-Roulland C. *gaston: Genetic Data Handling (QC, GRM, LD, PCA) and*  
667 *Linear Mixed Models*; 2018. R package version 1.5.3. Available from: <https://CRAN.R-project.org/package=gaston>. 6
- 668
- 669 [26] Ochoa A, Storey JD. FST and kinship for arbitrary population structures I: Generalized  
670 definitions. *bioRxiv*. 2016;. 8
- 671 [27] Ochoa A, Storey JD. FST and kinship for arbitrary population structures II: Method  
672 of moments estimators. *bioRxiv*. 2016;. 8
- 673 [28] Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regres-  
674 sion. *Statistica Sinica*. 2016;p. 35–67. 11
- 675 [29] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank  
676 resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203. 12
- 677 [30] Biobank U. Genotyping and quality control of UK Biobank, a large-scale, ex-  
678 tensively phenotyped prospective resource. Available at biobank ctsu ox ac  
679 uk/crystal/docs/genotyping\_qc pdf Accessed April. 2015;1:2016. 12

- 680 [31] Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867–  
681 2873. 12
- 683 [32] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-  
684 analysis of genome-wide association studies for height and body mass index in 700000  
685 individuals of European ancestry. *Human molecular genetics*. 2018;27(20):3641–3649.  
686 12, 13
- 687 [33] McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*.  
688 2016;48(10):1279. 13
- 690 [34] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear  
691 mixed models. *PLoS genetics*. 2013;9(2):e1003264. 13
- 692 [35] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association stud-  
693 ies. *Nature genetics*. 2012;44(7):821. 13
- 694 [36] Davey Smith G, Ebrahim S. ‘Mendelian randomization’: can genetic epidemiology con-  
695 tribute to understanding environmental determinants of disease? *International journal  
696 of epidemiology*. 2003;32(1):1–22. 14
- 697 [37] Cherlin S, Howey RA, Cordell HJ. Using penalized regression to predict phenotype  
698 from SNP data. In: *BMC proceedings*. vol. 12. BioMed Central; 2018. p. 38. 14
- 699 [38] Zhou W, Lo SH. Analysis of genotype by methylation interactions through sparsity-  
700 inducing regularized regression. In: *BMC proceedings*. vol. 12. BioMed Central; 2018.  
701 p. 40. 14
- 702 [39] Howey RA, Cordell HJ. Application of Bayesian networks to GAW20 genetic and blood  
703 lipid data. In: *BMC proceedings*. vol. 12. BioMed Central; 2018. p. 19. 15

- 704 [40] Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. *The American Journal of Human Genetics*.  
705 2012;91(1):122–138. 15
- 706
- 707 [41] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in  
708 unrelated individuals. *Genome research*. 2009;19(9):1655–1664. 15
- 709
- 710 [42] Fortin A, Diez E, Rochefort D, Laroche L, Malo D, Rouleau GA, et al. Recombinant  
711 congenic strains derived from A/J and C57BL/6J: a tool for genetic dissection of com-  
plex traits. *Genomics*. 2001;74(1):21–35. 17
- 712
- 713 [43] Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, et al. A  
714 high-resolution association mapping panel for the dissection of complex traits in mice.  
Genome research. 2010;20(2):281–290. 17
- 715
- 716 [44] Flint J, Eskin E. Genome-wide association studies in mice. *Nature Reviews Genetics*.  
2012;13(11):807. 17
- 717
- 718 [45] Cheng R, Lim JE, Samocha KE, Sokoloff G, Abney M, Skol AD, et al. Genome-wide  
719 association studies and the problem of relatedness among advanced intercross lines and  
other highly recombinant populations. *Genetics*. 2010;185(3):1033–1044. 17
- 720
- 721 [46] Di Pietrantonio T, Hernandez C, Girard M, Verville A, Orlova M, Belley A, et al.  
722 Strain-specific differences in the genetic control of two closely related mycobacteria.  
PLoS pathogens. 2010;6(10):e1001169. 17, 18
- 723
- 724 [47] Wang H, Lengerich BJ, Aragam B, Xing EP. Precision Lasso: accounting for cor-  
725 relations and linear dependencies in high-dimensional genomic data. *Bioinformatics*.  
2018;35(7):1181–1187. 19, 23
- 726
- [48] Sohrabi Y, Havelková H, Kobets T, Šíma M, Volkova V, Grekov I, et al. Mapping the

- 727 Genes for Susceptibility and Response to *Leishmania tropica* in Mouse. PLoS neglected  
728 tropical diseases. 2013;7(7):e2282. 19
- 729 [49] Jackson AU, Fornés A, Galecki A, Miller RA, Burke DT. Multiple-trait quantitative  
730 trait loci analysis using a large mouse sibship. Genetics. 1999;151(2):785–795. 19
- 731 [50] C Stern1 M, Benavides F, A Klingelberger E, J Conti2 C. Allelotype analysis of chemi-  
732 cally induced squamous cell carcinomas in F1 hybrids of two inbred mouse strains with  
733 different susceptibility to tumor progression. Carcinogenesis. 2000;21(7):1297–1301. 19
- 734 [51] Lasko D, Cavenee W, Nordenskjöld M. Loss of constitutional heterozygosity in human  
735 cancer. Annual review of genetics. 1991;25(1):281–314. 19
- 736 [52] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al.  
737 Efficient Bayesian mixed-model analysis increases association power in large cohorts.  
738 Nature genetics. 2015;47(3):284. 22
- 739 [53] Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank:  
740 Current status and what it means for epidemiology. Health Policy and Technology.  
741 2012;1(3):123–126. 23
- 742 [54] Zeng Y, Breheny P. The biglasso package: a memory-and computation-efficient solver  
743 for lasso model fitting with big data in R. arXiv preprint arXiv:170105936. 2017;. 24
- 744 [55] Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human molecular  
745 genetics. 2015;24(R1):R111–R119. 24
- 746 [56] Pirinen M, Donnelly P, Spencer CC, et al. Efficient computation with a linear mixed  
747 model on large-scale data sets with applications to genetic studies. The Annals of  
748 Applied Statistics. 2013;7(1):369–390. 25, 26

- 749 [57] Schelldorfer J, Bühlmann P, DE G, VAN S. Estimation for High-Dimensional Lin-  
750 ear Mixed-Effects Models Using L1-Penalization. Scandinavian Journal of Statistics.  
751 2011;38(2):197–214. 29, 30, 37, 47
- 752 [58] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable mini-  
753 mization. Mathematical Programming. 2009;117(1):387–423. 29, 47, 50
- 754 [59] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. Journal  
755 of the Royal Statistical Society: Series B (Statistical Methodology). 2008;70(1):53–71.  
756 30, 47
- 757 [60] Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained  
758 optimization. SIAM Journal on Scientific Computing. 1995;16(5):1190–1208. 31
- 759 [61] Wakefield J. Bayesian and frequentist regression methods. Springer Science & Business  
760 Media; 2013. 34
- 761 [62] Nishii R. Asymptotic properties of criteria for selection of variables in multiple regres-  
762 sion. The Annals of Statistics. 1984;p. 758–765. 37
- 763 [63] Zou H, Hastie T, Tibshirani R, et al. On the “degrees of freedom” of the lasso. The  
764 Annals of Statistics. 2007;35(5):2173–2192. 37
- 765 [64] Bondell HD, Krishna A, Ghosh SK. Joint Variable Selection for Fixed and Random  
766 Effects in Linear Mixed-Effects Models. Biometrics. 2010;66(4):1069–1077. 37
- 767 [65] Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likeli-  
768 hood. Journal of the Royal Statistical Society: Series B (Statistical Methodology).  
769 2013;75(3):531–552. 37
- 770 [66] Xie Y. Dynamic Documents with R and knitr. vol. 29. CRC Press; 2015. 57

## 771 A Block Coordinate Descent Algorithm

772 We use a general purpose block coordinate descent algorithm (CGD) [58] to solve (16). At  
 773 each iteration, the algorithm approximates the negative log-likelihood  $f(\cdot)$  in  $Q_\lambda(\cdot)$  by a  
 774 strictly convex quadratic function and then applies block coordinate decent to generate a  
 775 decent direction followed by an inexact line search along this direction [58]. For continuously  
 776 differentiable  $f(\cdot)$  and convex and block-separable  $P(\cdot)$  (i.e.  $P(\beta) = \sum_i P_i(\beta_i)$ ), [58] show  
 777 that the solution generated by the CGD method is a stationary point of  $Q_\lambda(\cdot)$  if the coor-  
 778 dinates are updated in a Gauss-Seidel manner i.e.  $Q_\lambda(\cdot)$  is minimized with respect to one  
 779 parameter while holding all others fixed. The CGD algorithm can thus be run in parallel and  
 780 therefore suited for large  $p$  settings. It has been successfully applied in fixed effects models  
 781 (e.g. [59], [20]) and [57] for mixed models with an  $\ell_1$  penalty. Following Tseng and Yun [58],  
 782 the CGD algorithm is given by Algorithm 2.

783 The Armijo rule is defined as follows [58]:

Choose  $\alpha_{init}^{(k)} > 0$  and let  $\alpha^{(k)}$  be the largest element of  $\{\alpha_{init}^k \delta^r\}_{r=0,1,2,\dots}$  satisfying

$$Q_\lambda(\Theta_j^{(k)} + \alpha^{(k)} d^{(k)}) \leq Q_\lambda(\Theta_j^{(k)}) + \alpha^{(k)} \varrho \Delta^{(k)} \quad (45)$$

where  $0 < \delta < 1$ ,  $0 < \varrho < 1$ ,  $0 \leq \gamma < 1$  and

$$\Delta^{(k)} := \nabla f(\Theta_j^{(k)}) d^{(k)} + \gamma (d^{(k)})^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d^{(k)}) - \lambda P(\Theta_j^{(k)}) \quad (46)$$

784

785 Common choices for the constants are  $\delta = 0.1$ ,  $\varrho = 0.001$ ,  $\gamma = 0$ ,  $\alpha_{init}^{(k)} = 1$  for all  $k$  [57].

786 Below we detail the specifics of Algorithm 2 for the  $\ell_1$  penalty.

**Algorithm 2:** Coordinate Gradient Descent Algorithm to solve (16)

---

Set the iteration counter  $k \leftarrow 0$  and choose initial values for the parameter vector

$$\Theta^{(0)};$$

**repeat**

 Approximate the Hessian  $\nabla^2 f(\Theta^{(k)})$  by a symmetric matrix  $H^{(k)}$ :

$$H^{(k)} = \text{diag} \left[ \min \left\{ \max \left\{ \left[ \nabla^2 f(\Theta^{(k)}) \right]_{jj}, c_{min} \right\} c_{max} \right\} \right]_{j=1,\dots,p} \quad (41)$$

**for**  $j = 1, \dots, p$  **do**

 Solve the descent direction  $d^{(k)} := d_{H^{(k)}}(\Theta_j^{(k)})$  ;

**if**  $\Theta_j^{(k)} \in \{\beta_1, \dots, \beta_p\}$  **then**

$$d_{H^{(k)}}(\Theta_j^{(k)}) \leftarrow \arg \min_d \left\{ \nabla f(\Theta_j^{(k)})d + \frac{1}{2}d^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d) \right\} \quad (42)$$

**end**
**end**

Choose a stepsize;

$$\alpha_j^{(k)} \leftarrow \text{line search given by the Armijo rule}$$

Update;

$$\widehat{\Theta}_j^{(k+1)} \leftarrow \widehat{\Theta}_j^{(k)} + \alpha_j^{(k)} d^{(k)}$$

Update;

$$\widehat{\eta}^{(k+1)} \leftarrow \arg \min_{\eta} \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2(k)} \sum_{i=1}^{N_T} \frac{\left( \widetilde{Y}_i - \sum_{j=0}^p \widetilde{X}_{ij+1} \beta_j^{(k+1)} \right)^2}{1 + \eta(\Lambda_i - 1)} \quad (43)$$

Update;

$$\widehat{\sigma^2}^{(k+1)} \leftarrow \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left( \widetilde{Y}_i - \sum_{j=0}^p \widetilde{X}_{ij+1} \beta_j^{(k+1)} \right)^2}{1 + \eta^{(k+1)}(\Lambda_i - 1)} \quad (44)$$

$$k \leftarrow k + 1$$

**until** convergence criterion is satisfied;

---

<sup>787</sup> **A.1  $\ell_1$  penalty**

<sup>788</sup> The objective function is given by

$$Q_\lambda(\Theta) = f(\Theta) + \lambda|\beta| \quad (47)$$

<sup>789</sup> **A.1.1 Descent Direction**

<sup>790</sup> For simplicity, we remove the iteration counter ( $k$ ) from the derivation below.

<sup>791</sup> For  $\Theta_j^{(k)} \in \{\beta_1, \dots, \beta_p\}$ , let

$$d_H(\Theta_j) = \arg \min_d G(d) \quad (48)$$

<sup>792</sup> where

$$G(d) = \nabla f(\Theta_j)d + \frac{1}{2}d^2 H_{jj} + \lambda|\Theta_j + d|$$

<sup>793</sup> Since  $G(d)$  is not differentiable at  $-\Theta_j$ , we calculate the subdifferential  $\partial G(d)$  and search

<sup>794</sup> for  $d$  with  $0 \in \partial G(d)$ :

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda u \quad (49)$$

<sup>795</sup> where

$$u = \begin{cases} 1 & \text{if } d > -\Theta_j \\ -1 & \text{if } d < -\Theta_j \\ [-1, 1] & \text{if } d = \Theta_j \end{cases} \quad (50)$$

<sup>796</sup> We consider each of the three cases in (49) below

1.  $d > -\Theta_j$

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda = 0$$

$$d = \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}}$$

Since  $\lambda > 0$  and  $H_{jj} > 0$ , we have

$$\frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} > \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} = d \stackrel{\text{def}}{>} -\Theta_j$$

The solution can be written compactly as

$$d = \text{mid} \left\{ \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \right\}$$

797 where  $\text{mid} \{a, b, c\}$  denotes the median (mid-point) of  $a, b, c$  [58].

2.  $d < -\Theta_j$

$$\begin{aligned} \partial G(d) &= \nabla f(\Theta_j) + dH_{jj} - \lambda = 0 \\ d &= \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} \end{aligned}$$

Since  $\lambda > 0$  and  $H_{jj} > 0$ , we have

$$\frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} < \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} = d \stackrel{\text{def}}{<} -\Theta_j$$

Again, the solution can be written compactly as

$$d = \text{mid} \left\{ \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \right\}$$

3.  $d_j = -\Theta_j$

There exists  $u \in [-1, 1]$  such that

$$\begin{aligned} \partial G(d) &= \nabla f(\Theta_j) + dH_{jj} + \lambda u = 0 \\ d &= \frac{-(\nabla f(\Theta_j) + \lambda u)}{H_{jj}} \end{aligned}$$

For  $-1 \leq u \leq 1$ ,  $\lambda > 0$  and  $H_{jj} > 0$  we have

$$\frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \leq d \stackrel{\text{def}}{=} -\Theta_j \leq \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}$$

The solution can again be written compactly as

$$d = \text{mid} \left\{ \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \right\}$$

798 We see all three cases lead to the same solution for (48). Therefore the descent direction for  
799  $\Theta_j^{(k)} \in \{\beta_1, \dots, \beta_p\}$  for the  $\ell_1$  penalty is given by

$$d = \text{mid} \left\{ \frac{-(\nabla f(\beta_j) - \lambda)}{H_{jj}}, -\beta_j, \frac{-(\nabla f(\beta_j) + \lambda)}{H_{jj}} \right\} \quad (51)$$

800 **A.1.2 Solution for the  $\beta$  parameter**

801 If the Hessian  $\nabla^2 f(\Theta^{(k)}) > 0$  then  $H^{(k)}$  defined in (41) is equal to  $\nabla^2 f(\Theta^{(k)})$ . Using  $\alpha_{init} = 1$ ,  
802 the largest element of  $\{\alpha_{init}^{(k)} \delta^r\}_{r=0,1,2,\dots}$  satisfying the Armijo Rule inequality is reached for  
803  $\alpha^{(k)} = \alpha_{init}^{(k)} \delta^0 = 1$ . The Armijo rule update for the  $\beta$  parameter is then given by

$$\beta_j^{(k+1)} \leftarrow \beta_j^{(k)} + d^{(k)}, \quad j = 1, \dots, p \quad (52)$$

804 Substituting the descent direction given by (51) into (52) we get

$$\beta_j^{(k+1)} = \text{mid} \left\{ \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}, 0, \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}} \right\} \quad (53)$$

805 We can further simplify this expression. Let

$$w_i := \frac{1}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \quad (54)$$

Re-write the part depending on  $\beta$  of the negative log-likelihood in (14) as

$$g(\boldsymbol{\beta}^{(k)}) = \frac{1}{2} \sum_{i=1}^{N_T} w_i \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} - \tilde{X}_{ij} \beta_j^{(k)} \right)^2 \quad (55)$$

The gradient and Hessian are given by

$$\nabla f(\beta_j^{(k)}) := \frac{\partial}{\partial \beta_j^{(k)}} g(\boldsymbol{\beta}^{(k)}) = - \sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} - \tilde{X}_{ij} \beta_j^{(k)} \right) \quad (56)$$

$$H_{jj} := \frac{\partial^2}{\partial \beta_j^{(k)} \partial \beta_j^{(k)}} g(\boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2 \quad (57)$$

Substituting (56) and (57) into  $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}$

$$\begin{aligned} & \beta_j^{(k)} + \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} - \tilde{X}_{ij} \beta_j^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \\ &= \beta_j^{(k)} + \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} - \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2 \beta_j^{(k)}}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \\ &= \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \end{aligned} \quad (58)$$

Similarly, substituting (56) and (57) in  $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}}$  we get

$$\frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) - \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \quad (59)$$

Finally, substituting (58) and (59) into (53) we get

$$\begin{aligned}\beta_j^{(k+1)} &= \text{mid} \left\{ \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) - \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2}, 0, \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \right\} \\ &= \frac{\mathcal{S}_\lambda \left( \sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left( \tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) \right)}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2}\end{aligned}\tag{60}$$

Where  $\mathcal{S}_\lambda(x)$  is the soft-thresholding operator

$$\mathcal{S}_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$$

$\text{sign}(x)$  is the signum function

$$\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

and  $(x)_+ = \max(x, 0)$ .

808 **B Additional Real Data Analysis Results**

809 **B.1 Distribution of SNPs used in UK Biobank analysis**

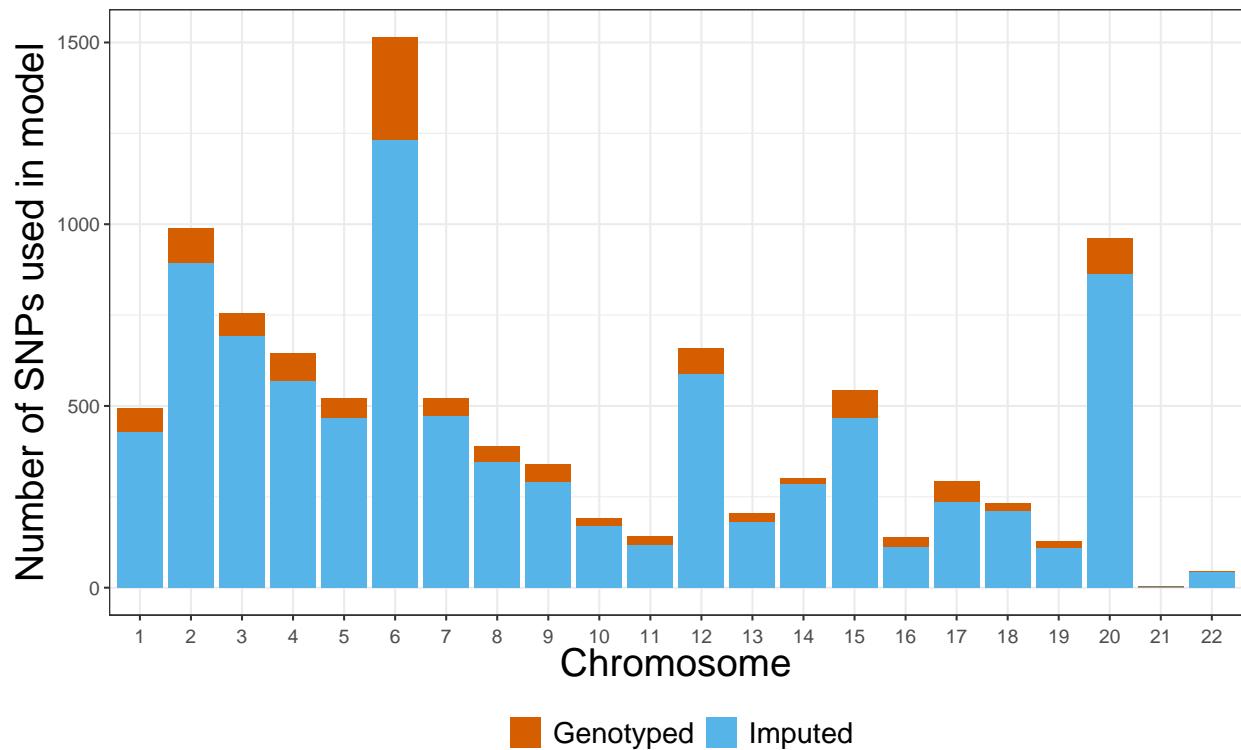


Figure B.1: Distribution of SNPs used in UK Biobank analysis by chromosome and whether or not the SNP was imputed.

810 **B.2 LD structure among the markers in the GAW20 and the mouse**  
 811 **dataset**

812 We illustrate the LD structure among the markers in the GAW20 dataset and the mouse  
 813 dataset separately in Figures B.2 and B.3, respectively. In Figure B.2, we show the pairwise  
 814  $r^2$  for 655 SNPs within a 1Mb-window around the causal SNP rs9661059 (indicated) that we  
 815 focused on. The dotplot above the heatmap denotes  $r^2$  between each SNP and the causal  
 816 SNP. It is clear that although strong correlation does exist between some SNPs, none of these  
 817 nearby SNPs is correlated with the causal SNP. The only dot denoting an  $r^2 = 1$  represents  
 818 the causal SNP itself.

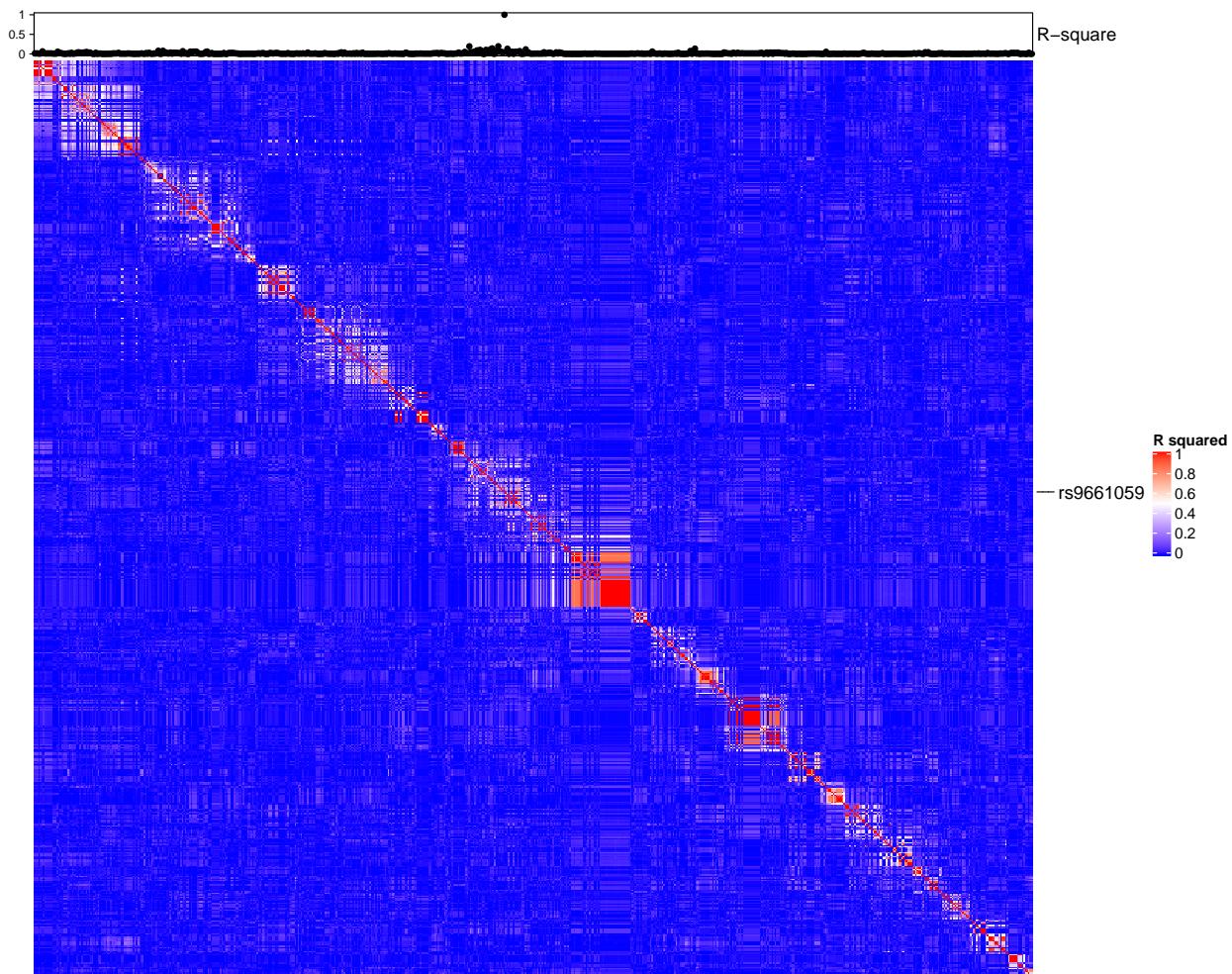


Figure B.2: LD structure among the markers in the GAW20 dataset

- 819 In Figure B.3, we show the pairwise  $r^2$  for all microsatellite markers in the mouse dataset.  
820 It is clear that many markers are considerably strongly correlated with each other, as we  
821 expected.

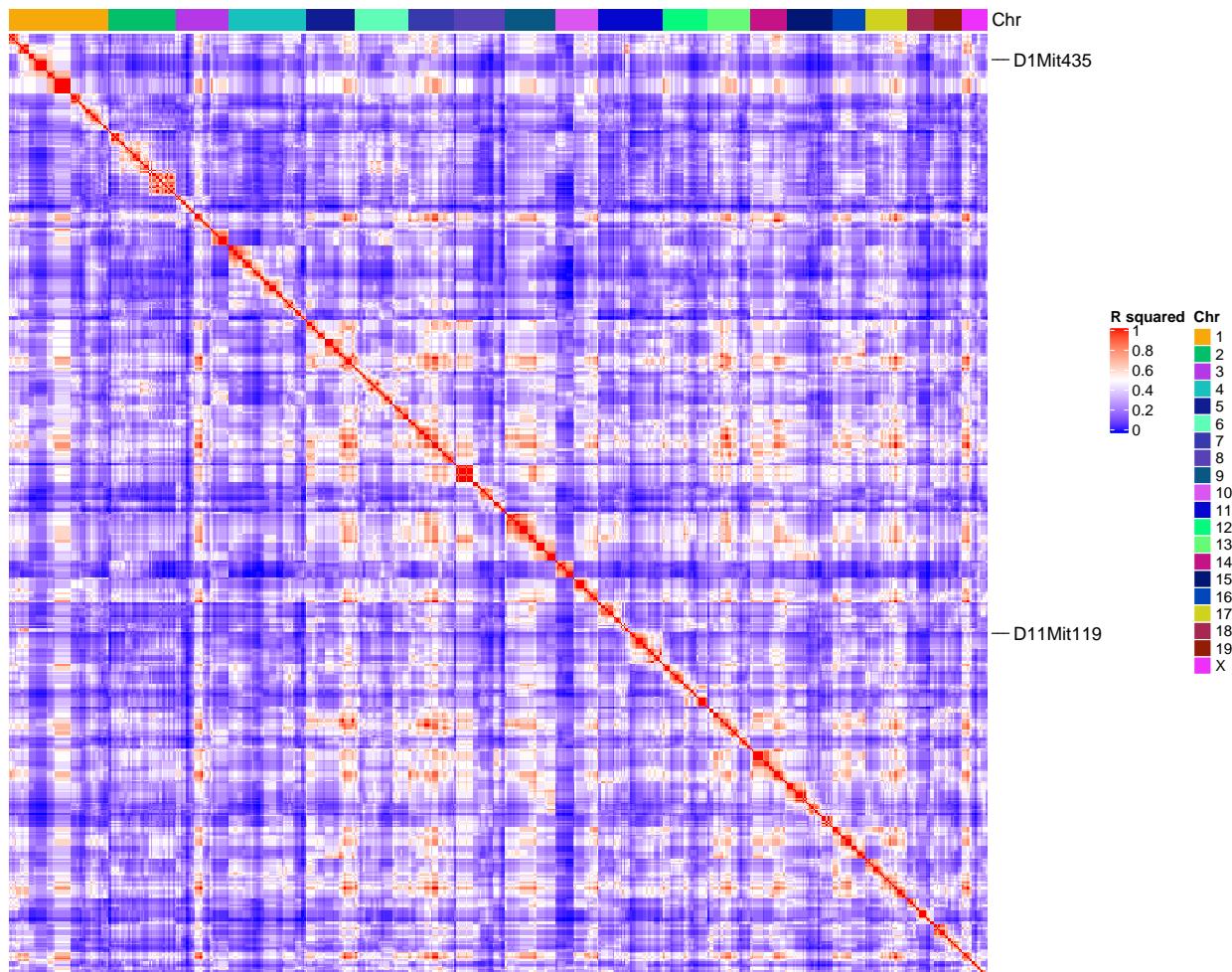


Figure B.3: LD structure among the markers in the mouse dataset

## 822 C ggmix Package Showcase

823 In this section we briefly introduce the freely available and open source `ggmix` package in R.  
 824 More comprehensive documentation is available at <https://sahirbhatnagar.com/ggmix>.  
 825 Note that this entire section is reproducible; the code and text are combined in an `.Rnw`<sup>1</sup> file  
 826 and compiled using `knitr` [66].

### 827 C.1 Installation

828 The package can be installed from [GitHub](#) via

```
install.packages("pacman")
pacman::p_load_gh('sahirbhatnagar/ggmix')
```

829 To showcase the main functions in `ggmix`, we will use the simulated data which ships with  
 830 the package and can be loaded via:

```
## library(ggmix)
data("admixed")
names(admixed)

## [1] "ytrain"      "ytune"        "ytest"        "xtrain"
## [5] "xtune"        "xtest"        "xtrain_lasso" "xtune_lasso"
## [9] "xtest_lasso"  "Xkinship"     "kin_train"    "kin_tune_train"
## [13] "kin_test_train" "mu_train"     "causal"       "beta"
## [17] "not_causal"   "kinship"     "coancestry"  "PC"
## [21] "subpops"
```

831 For details on how this data was simulated, see `help(admixed)`.

832 There are three basic inputs that `ggmix` needs:

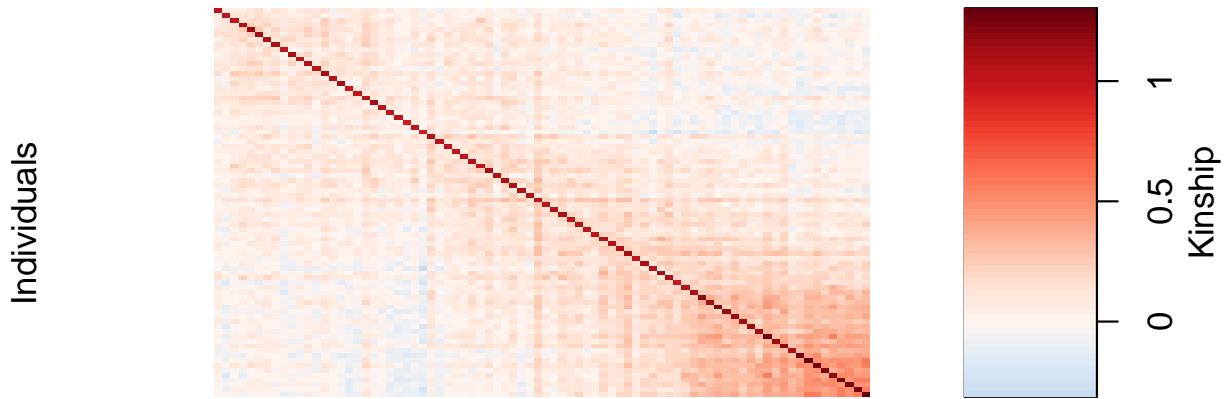
- 833 1.  $Y$ : a continuous response variable
- 834 2.  $X$ : a matrix of covariates of dimension  $N \times p$  where  $N$  is the sample size and  $p$  is the  
 835 number of covariates
- 836 3.  $\Phi$ : a kinship matrix

---

<sup>1</sup>scripts available at <https://github.com/sahirbhatnagar/ggmix/tree/pgen/manuscript>

837 We can visualize the kinship matrix in the admixed data using the popkin package:

```
# need to install the package if you don't have it
# pacman::p_load_gh('StoreyLab/popkin')
popkin::plot_popkin(admixed$kin_train)
```



838

## 839 C.2 Fit the linear mixed model with Lasso Penalty

840 We will use the most basic call to the main function of this package, which is called `ggmix`.

841 This function will by default fit a  $L_1$  penalized linear mixed model (LMM) for 100 distinct

842 values of the tuning parameter  $\lambda$ . It will choose its own sequence:

```
fit <- ggmix(x = admixed$xtrain,
```

```

y = admixed$ytrain,
kinship = admixed$kin_train)

names(fit)

## [1] "result"      "ggmix_object"  "n_design"     "p_design"    "lambda"
## [6] "coef"        "b0"          "beta"        "df"         "eta"
## [11] "sigma2"      "nlambda"      "cov_names"   "call"

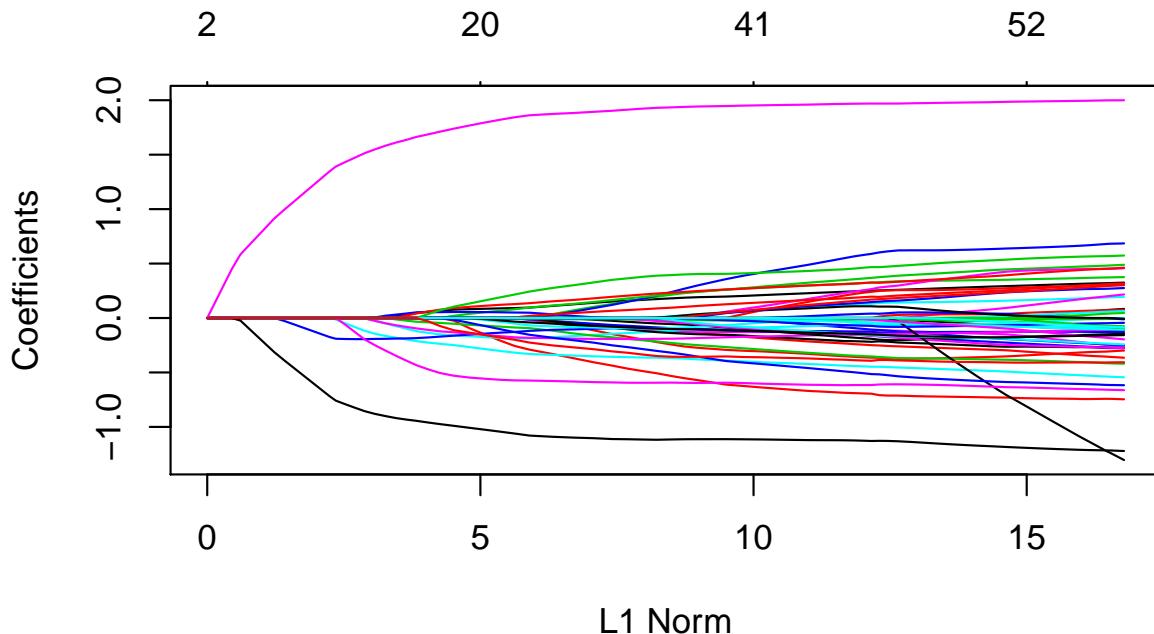
class(fit)

## [1] "lassofullrank" "ggmix_fit"

```

- 843 We can see the solution path for each variable by calling the `plot` method for objects of  
 844 class `ggmix_fit`:

```
plot(fit)
```



- 845
- 846 We can also get the coefficients for given value(s) of lambda using the `coef` method for  
 847 objects of class `ggmix_fit`:

```
# only the first 5 coefficients printed here for brevity
```

```

coef(fit, s = c(0.1,0.02))[1:5, ]

## 5 x 2 Matrix of class "dgeMatrix"
##           1         2
## (Intercept) -0.03715135  0.247105426
## X23        0.00000000  0.098030248
## X36        0.00000000 -0.013022250
## X38        0.00000000  0.005378361
## X40        0.00000000  0.004028934

```

848 Here, `s` specifies the value(s) of  $\lambda$  at which the extraction is made. The function uses linear  
 849 interpolation to make predictions for values of `s` that do not coincide with the lambda  
 850 sequence used in the fitting algorithm.

851 We can also get predictions ( $X\hat{\beta}$ ) using the `predict` method for objects of class `ggmix_fit`:

```

# need to provide x to the predict function
# predict for the first 5 subjects
predict(fit, s = c(0.1,0.02), newx = admixed$xtest[1:5,])

##           1         2
## id26   2.30208546  2.45597763
## id39   0.87334032  1.62931898
## id45  -0.12296837 -0.06075786
## id52  -0.03715135 -0.97519671
## id53  -0.21046107 -0.23151040

```

### 852 C.3 Find the Optimal Value of the Tuning Parameter

853 We use the Generalized Information Criterion (GIC) to select the optimal value for  $\lambda$ . The  
 854 default is  $a_n = \log(\log(n)) * \log(p)$  which corresponds to a high-dimensional BIC (HD-  
 855 BIC):

```

# pass the fitted object from ggmix to the gic function:
hdbic <- gic(fit)
class(hdbic)

## [1] "ggmix_gic"      "lassofullrank" "ggmix_fit"

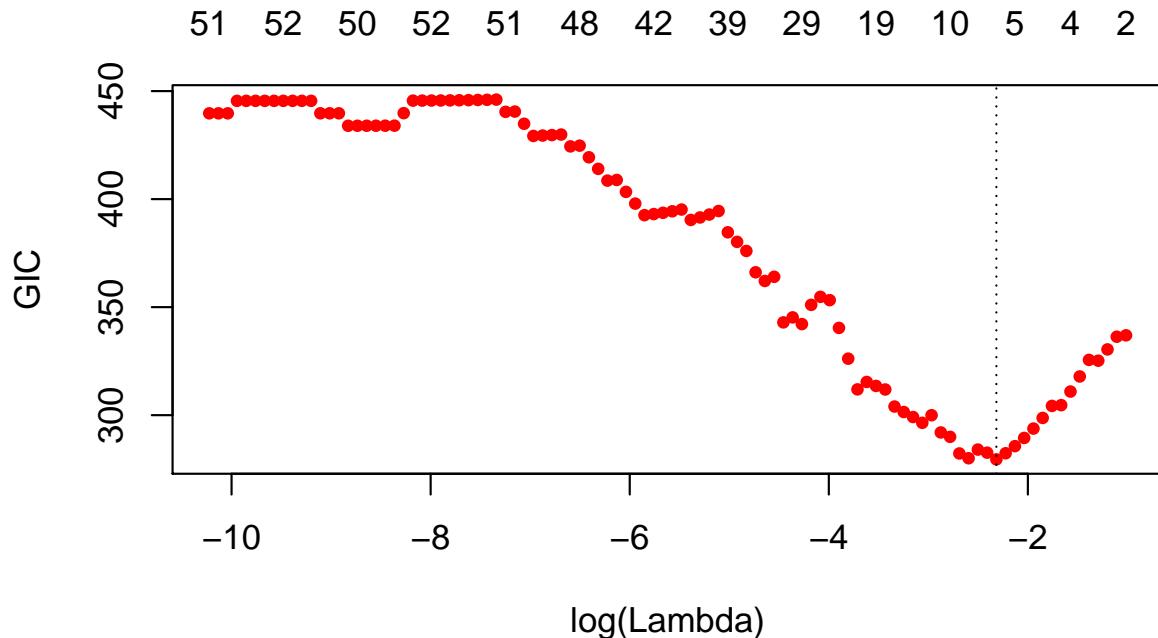
# we can also fit the BIC by specifying the an argument
bicfit <- gic(fit, an = log(length(admixed$ytrain)))

```

856 We can plot the HDBIC values against  $\log(\lambda)$  using the `plot` method for objects of class

857 `ggmix_gic`:

```
plot(hdbic)
```



858 The optimal value for  $\lambda$  according to the HDBIC, i.e., the  $\lambda$  that leads to the minium HDBIC

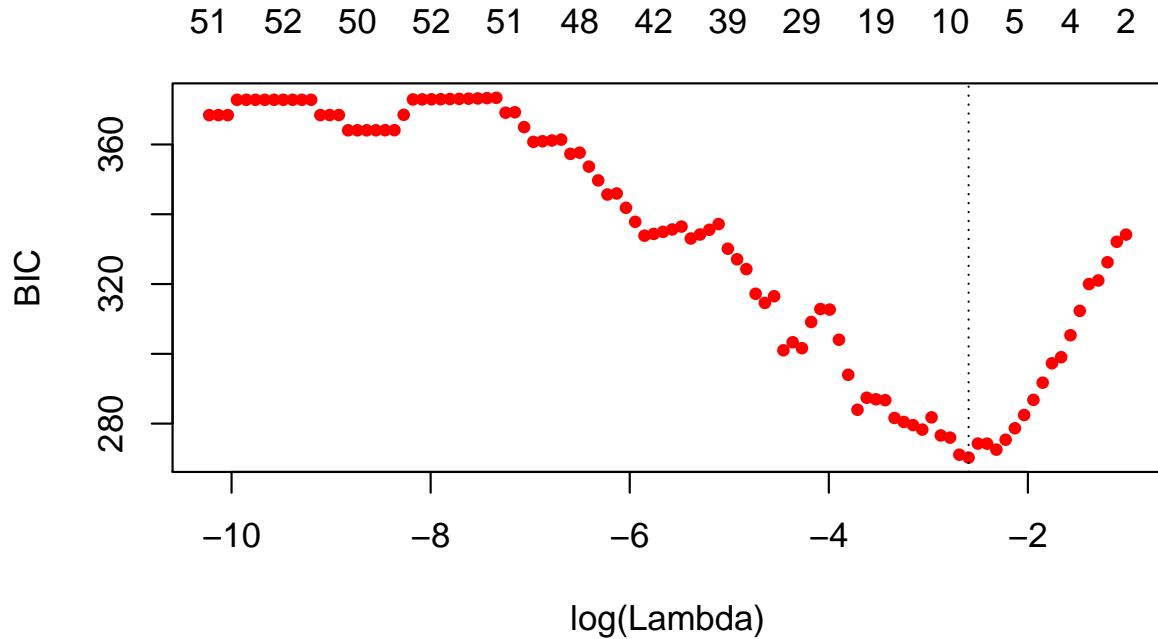
860 is:

```
hdbic[["lambda.min"]]
```

```
## [1] 0.09862269
```

861 We can also plot the BIC results:

```
plot(bicfit, ylab = "BIC")
```



862

```
bicfit[["lambda.min"]]
## [1] 0.07460445
```

#### 863 C.4 Get Coefficients Corresponding to Optimal Model

864 We can use the object outputted by the `gic` function to extract the coefficients corresponding  
 865 to the selected model using the `coef` method for objects of class `ggmix_gic`:

```
coef(hdbic)[1:5, , drop = FALSE]
## 5 x 1 sparse Matrix of class "dgCMatrix"
##           1
## (Intercept) -0.03660806
## X23         .
## X36         .
## X38         .
## X40         .
```

866 We can also extract just the nonzero coefficients which also provide the estimated variance  
 867 components  $\eta$  and  $\sigma^2$ :

```

coef(hdbic, type = "nonzero")

##          1
## (Intercept) -0.03660806
## X302       -0.17607392
## X524        1.34951500
## X538       -0.72052613
## eta         0.99000000
## sigma2      1.60476289

```

- 868 We can also make predictions from the `hdbic` object, which by default will use the model  
 869 corresponding to the optimal tuning parameter:

```

predict(hdbic, newx = admixed$xtest[1:5,])

##          1
## id26   2.31027410
## id39   0.86922183
## id45  -0.12814532
## id52  -0.03660806
## id53  -0.21268198

```

## 870 C.5 Extracting Random Effects

- 871 The user can compute the random effects using the provided `ranef` method for objects of  
 872 class `ggmix_gic`. This command will compute the estimated random effects for each subject  
 873 using the parameters of the selected model:

```

ranef(hdbic)[1:5]

## [1] -2.4889655  1.1834200 -0.5641832 -0.9310334 -0.3458703

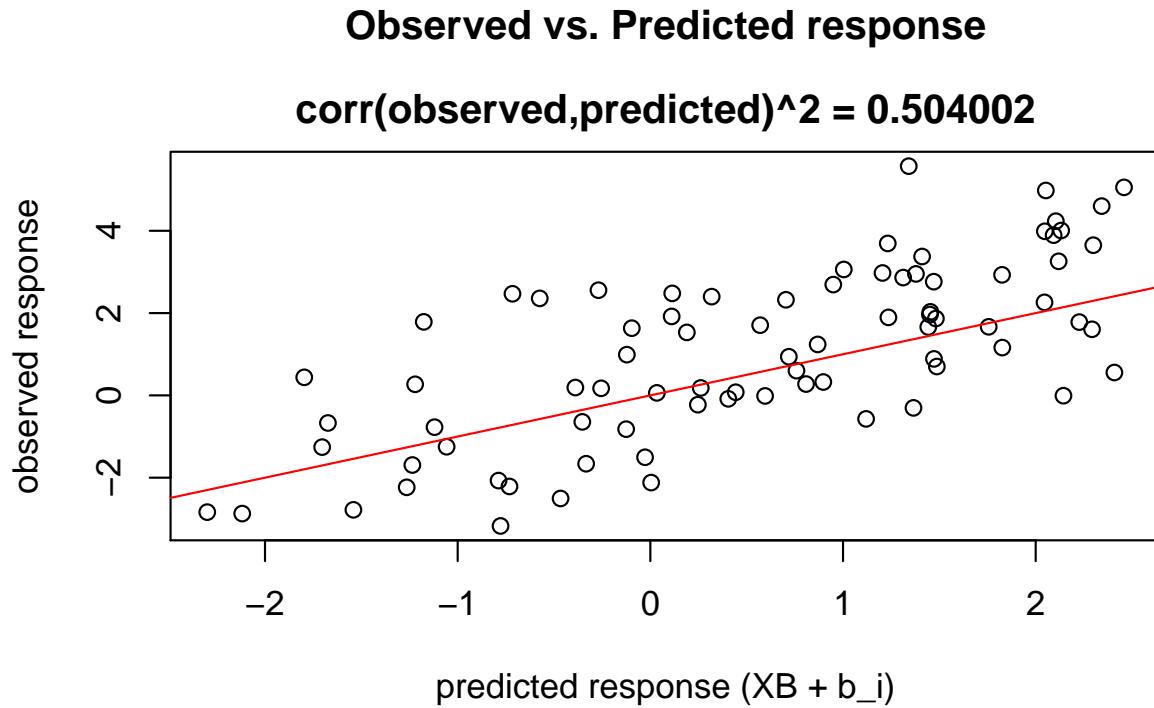
```

## 874 C.6 Diagnostic Plots

- 875 We can also plot some standard diagnostic plots such as the observed vs. predicted response,  
 876 QQ-plots of the residuals and random effects and the Tukey-Anscombe plot. These can be  
 877 plotted using the `plot` method on a `ggmix_gic` object as shown below.

## 878 C.6.1 Observed vs. Predicted Response

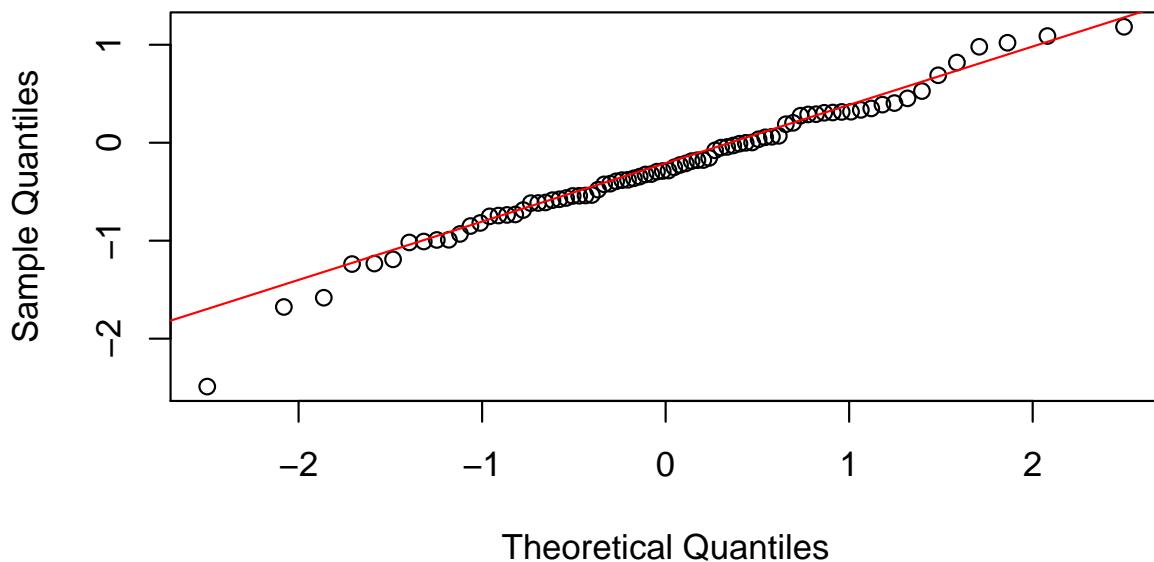
```
plot(hdbic, type = "predicted", newx = admixed$xtrain, newy = admixed$ytrain)
```



879

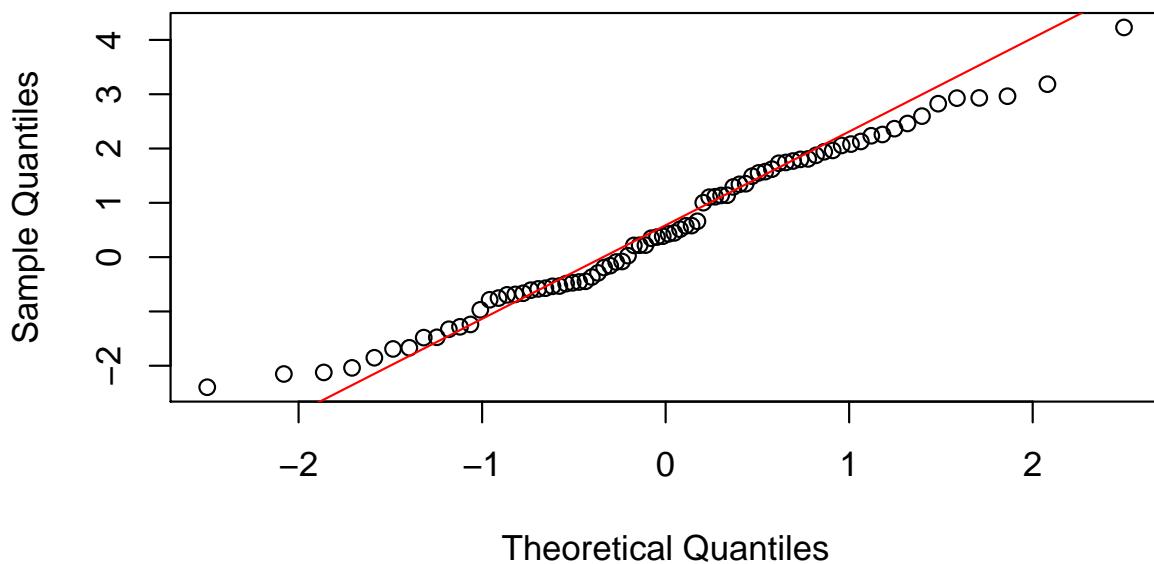
## 880 C.6.2 QQ-plots for Residuals and Random Effects

```
plot(hdbic, type = "QQranef", newx = admixed$xtrain, newy = admixed$ytrain)
```

**QQ-Plot of the random effects at lambda = 0.10**

881

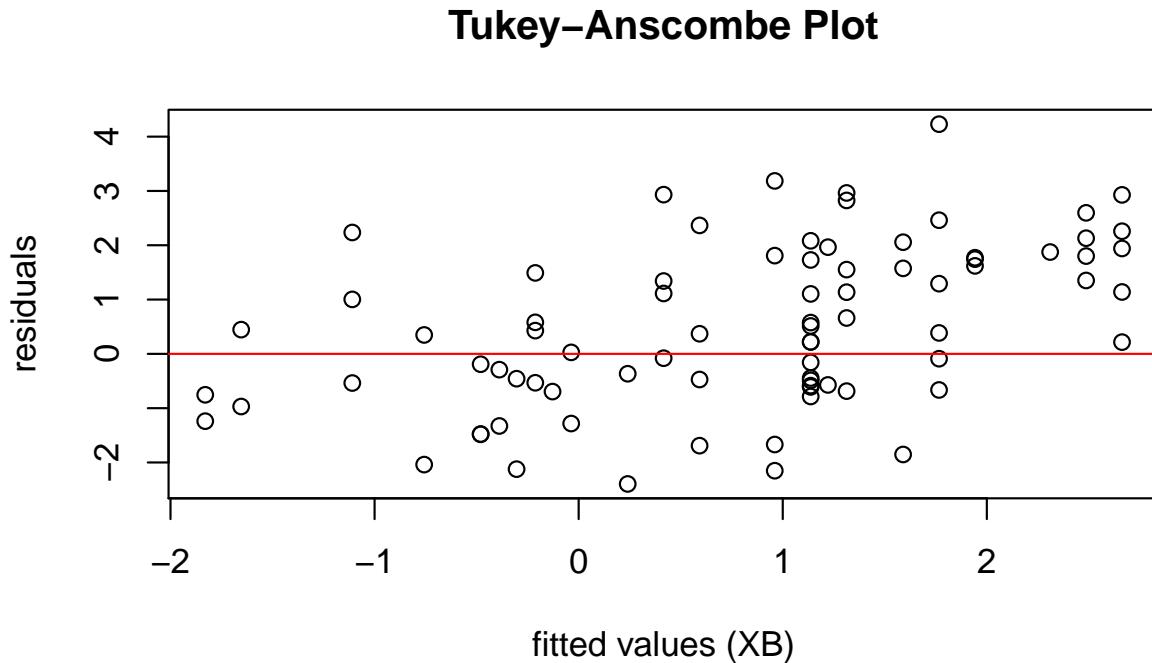
```
plot(hdbic, type = "QQresid", newx = admixed$xtrain, newy = admixed$ytrain)
```

**QQ-Plot of the residuals at lambda = 0.10**

882

## 883 C.6.3 Tukey-Anscombe Plot

```
plot(hdbic, type = "Tukey", newx = admixed$xtrain, newy = admixed$ytrain)
```



884