

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

⁴ Sahir R Bhatnagar^{1,2}, Yi Yang⁴, Tianyuan Lu², Erwin Schurr⁶,
⁵ JC Loredo-Osti⁷, Marie Forest², Karim Oualkacha³, and
⁶ Celia MT Greenwood^{1,2,5}

⁷ ¹Department of Epidemiology, Biostatistics and Occupational Health,
⁸ McGill University

⁹ ²Lady Davis Institute, Jewish General Hospital, Montréal, QC

¹⁰ ³Département de Mathématiques, Université de Québec à Montréal

¹¹ ⁴Department of Mathematics and Statistics, McGill University

¹² ⁵Departments of Oncology and Human Genetics, McGill University

¹³ ⁶Department of Medicine, McGill University

¹⁴ ⁷Department of Mathematics and Statistics, Memorial University

¹⁵ December 10, 2019

¹⁶ **Abstract**

¹⁷ Complex traits are known to be influenced by a combination of environmental fac-

18 tors and rare and common genetic variants. However, detection of such multivariate
19 associations can be compromised by low statistical power and confounding by popu-
20 lation structure. Linear mixed effects models (LMM) can account for correlations due
21 to relatedness but have not been applicable in high-dimensional (HD) settings where
22 the number of fixed effect predictors greatly exceeds the number of samples. False
23 positives or false negatives can result from two-stage approaches, where the residuals
24 estimated from a null model adjusted for the subjects' relationship structure are sub-
25 sequently used as the response in a standard penalized regression model. To overcome
26 these challenges, we develop a general penalized LMM framework called **gmmix** for
27 simultaneous SNP selection and adjustment for population structure in high dimen-
28 sional prediction models. Our method can accommodate several sparsity-inducing
29 penalties such as the lasso, elastic net and group lasso, and also readily handles prior
30 annotation information in the form of weights. We develop a blockwise coordinate
31 descent algorithm which is highly scalable, computationally efficient and has theo-
32 retical guarantees of convergence. Through simulations and three real data exam-
33 ples, we show that **gmmix** leads to more parsimonious models compared to the two-
34 stage approach or principal component adjustment with better prediction accuracy.
35 **gmmix** can be used to construct polygenic risk scores and select instrumental variables
36 in Mendelian randomization studies. Our algorithms are available in an R package
37 (<https://github.com/greenwoodlab/gmmix>).

38 1 Author Summary

39 This work addresses a recurring challenge in the analysis and interpretation of genetic as-
40 sociation studies: which genetic variants can best predict and are independently associated
41 with a given phenotype in the presence of population structure ? Not controlling confound-
42 ing due to geographic population structure, family and/or cryptic relatedness can lead to
43 spurious associations. Much of the existing research has therefore focused on modeling the

44 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
52 and 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

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. Our method can accommodate several sparsity inducing penalties such as the lasso [17], elastic net [23] and group lasso [24]. `ggmix` also readily handles prior annotation information in the form of a penalty factor, which can be useful, for example, when dealing with rare variants. We develop a blockwise coordinate descent algorithm which is highly scalable and has theoretical guarantees of convergence to a stationary point. All of our algorithms are implemented in the `ggmix` R package hosted on GitHub with extensive documentation (<https://github.com/greenwoodlab/ggmix>). We provide a brief demonstration of the `ggmix` package in Appendix C.

The rest of the paper is organized as follows. In Section 3, we compare the performance

of our proposed approach and demonstrate the scenarios where it can be advantageous to use over existing methods through simulation studies and two real data analyses. This is followed by a discussion of our results, some limitations and future directions in Section 4. Section 5 describes the `ggmix` model, the optimization procedure and the algorithm used to fit it.

3 Results

In this section we demonstrate the performance of `ggmix` in a simulation study and two real data applications.

3.1 Simulation Study

We evaluated the performance of `ggmix` in a variety of simulated scenarios. For each simulation scenario we compared `ggmix` to the `lasso` and the `twostep` method. For the `lasso`, we included the top 10 principal components from the simulated genotypes used to calculate the kinship matrix as unpenalized predictors in the design matrix. For the `twostep` method, we first fitted an intercept only model with a single random effect using the average information restricted maximum likelihood (AIREML) algorithm [25] as implemented in the `gaston` R package [26]. The residuals from this model were then used as the response in a regular `lasso` model. Note that in the `twostep` method, we removed the kinship effect in the first step and therefore did not need to make any further adjustments when fitting the penalized model. We fitted the `lasso` using the default settings and `standardize=FALSE` in the `glmnet` package [20], **with 10-fold cross-validation (CV) to select the optimal tuning parameter. For other parameters in our simulation study, we defined the following quantities:**

- n : sample size

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

150 We simulated data from the model

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

151 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.
 152 **Here, $\Phi_{n \times n}$ is the covariance matrix based on the *kinship* SNPs from n individu-**
 153 **als, $\mathbf{I}_{n \times n}$ is the identity matrix and parameters σ^2 and $\eta \in [0, 1]$ determine how the variance**
 154 **is divided between \mathbf{P} and $\boldsymbol{\varepsilon}$. The values of the parameters that we used were as follows:**
 155 **narrow sense heritability $\eta = \{0.1, 0.3\}$, number of covariates $p = 5,000$, number of *kinship***
 156 **SNPs $k = 10,000$, percentage of *causal* SNPs $c = \{0\%, 1\%\}$ and $\sigma^2 = 1$. In addition to**
 157 **these parameters, we also varied the amount of overlap between the *causal* list**
 158 **and the *kinship* list. We considered two main scenarios:**

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

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

¹⁶⁶ more difficult to apply in large-scale multivariable models where there is likely to be overlap
¹⁶⁷ between the variables in the design matrix and kinship matrix. We simulated random geno-
¹⁶⁸ types from the BN-PSD admixture model with 1D geography and 10 subpopulations using
¹⁶⁹ the `bnpsd` package [27, 28]. In Figure 1, we plot the estimated kinship matrix from a single
¹⁷⁰ simulated dataset in the form of a heatmap where a darker color indicates a closer genetic
¹⁷¹ relationship.

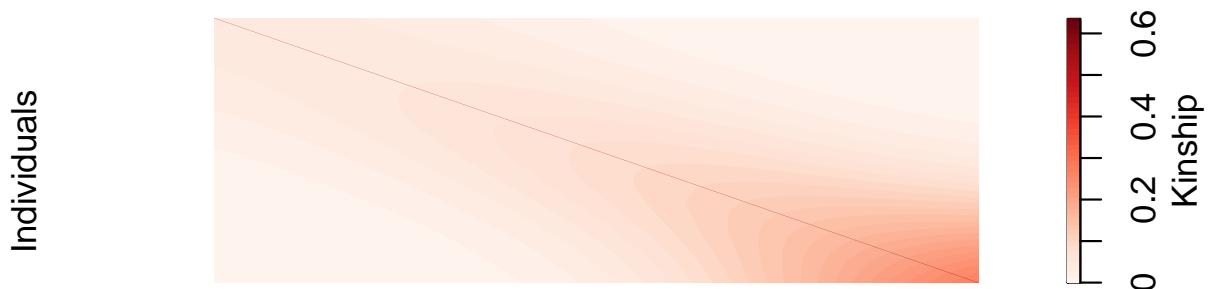


Figure 1: Example of an empirical kinship matrix used in simulation studies. This scenario models a 1D geography with extensive admixture.

¹⁷² In Figure 2 we plot the first two principal component scores calculated from the simulated
¹⁷³ genotypes used to calculate the kinship matrix in Figure 1, and color each point by sub-
¹⁷⁴ population membership. We can see that the PCs can identify the subpopulations which
¹⁷⁵ is why including them as additional covariates in a regression model has been considered a
¹⁷⁶ reasonable approach to control for confounding.

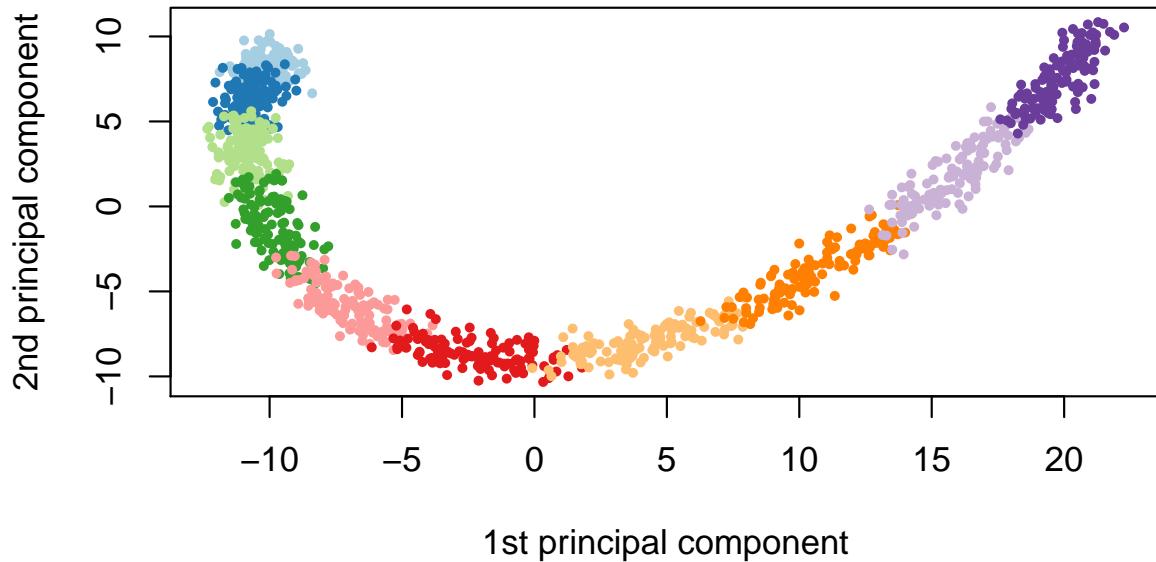


Figure 2: First two principal component scores of the genotype data used to estimate the kinship matrix where each color represents one of the 10 simulated subpopulations.

177 Using this set-up, we randomly partitioned 1000 simulated observations into 80% for training and 20% for testing. The training set was used to fit the model and select the optimal
 178 tuning parameter only, and the resulting model was evaluated on the test set. Let $\hat{\lambda}$ be the
 179 estimated value of the optimal regularization parameter, $\hat{\beta}_{\hat{\lambda}}$ the estimate of β at regular-
 180 ization 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
 181 coefficients. **To compare the methods in the context of true positive rate (TPR),**
 182 **we selected the largest tuning parameter that would result in a false positive**
 183 **rate (FPR) closest to 5%, but not more.** We also compared the model size ($|\hat{S}_{\hat{\lambda}}|$), test
 184 set prediction error based on the refitted unpenalized estimates for each selected model, the
 185 estimation error ($\|\hat{\beta} - \beta\|_2^2$), and the variance components (η, σ^2) for the polygenic random
 186 effect and error term.
 187

188 **The results are summarized in Table 1.** We see that `gmmix` outperformed the

189 **twostep** in terms of TPR, and was comparable to the **lasso**. This was the case,
190 regardless of true heritability and whether the causal SNPs were included in the calculation of
191 the kinship matrix. Across all simulation scenarios, **gmmix** had the smallest estimation error,
192 and smallest root mean squared prediction error (RMSE) on the test set while also producing
193 the most parsimonious models. Both the **lasso** and **twostep** selected more false positives,
194 even in the null model scenario. Both the **twostep** and **gmmix** overestimated the heritability
195 though **gmmix** was closer to the true value. When none of the causal SNPs were in the
196 kinship, both methods tended to overestimate the truth when $\eta = 10\%$ and underestimate
197 when $\eta = 30\%$. Across all simulation scenarios **gmmix** was able to (on average) correctly
198 estimate the error variance. The **lasso** tended to overestimate σ^2 in the null model while
199 the **twostep** overestimated σ^2 when none of the causal SNPs were in the kinship matrix.

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

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

207 3.2 Real Data Applications

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

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 (σ^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 $\frac{1}{n-|\widehat{S}_{\lambda}|} \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda}\|_2^2$ [29] as an estimator for σ^2 . Heritability (η) 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 σ_g^2 and σ_e^2 are the variance components for the random effect and error term, respectively. η is explicitly modeled in `gmmix`. There is no positive way to calculate η 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%
TPR at FPR=5%	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, 388)	0 (0, 2) (287, 385)	0 (0, 5) (246, 317)	0 (0, 2) (245, 314)	328 278 284 279	332 276 284 285	284 279 284 285	284 329 253, 319 319
	lasso	0 (0, 6) (246, 317)	0 (0, 5) (245, 314)	0 (0, 6) (252, 321)	0 (0, 5) (244, 319)	284 279 284 285	284 279 284 285	284 329 253, 319 319	284 329 253, 319 319
	gmmix	0 (0, 0) (43, 44)	0 (0, 0) (39, 43)	0 (0, 0) (39, 43)	0 (0, 0) (38, 43)	43 (39, 49) 43 (39, 48)	43 (39, 48) 44 (38, 49)	44 (38, 49) 43 (38, 48)	44 (38, 49) 43 (38, 48)
	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)
Model Size	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)
Error Variance	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)
	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	— —	— —	— —	— —	— —	— —	— —	— —
Heritability	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.

213 3.2.1 UK Biobank

214 With more than 500,000 participants, the UK Biobank is one of the largest
215 genotyped health care registries in the world. Among these participants, 147,731
216 have been inferred to be related to at least one individual in this cohort [30].
217 Such a widespread genetic relatedness may confound association studies and bias
218 trait predictions if not properly accounted for. Among these related individuals,
219 18,150 have a documented familial relationship (parent-offspring, full siblings,
220 second degree or third degree) that was previously inferred in [31]. We attempted
221 to derive a polygenic risk score for height among these individuals. We compared
222 the `gmmix`-derived polygenic risk score to those derived by the `twostep` and `lasso`
223 methods.

224 We first estimated the pairwise kinship coefficient among the 18,150 reportedly
225 related individuals based on 784,256 genotyped SNPs using KING [32]. We
226 grouped related individuals with a kinship coefficient > 0.044 [32] into 8,300
227 pedigrees. We then randomly split the dataset into a training set, a model
228 selection set and a test set of roughly equal sample size, ensuring all individuals
229 in the same pedigree were assigned into the same set. We inverse normalized the
230 standing height after adjusting for age, sex, genotyping array, and assessment
231 center following Yengo et al. [33].

232 To reduce computational complexity, we selected 10,000 SNPs with the largest
233 effect sizes associated with height from a recent large meta-analysis [33]. Among
234 these 10,000 SNPs, 1,233 were genotyped and used for estimating the kinship
235 whereas the other 8,767 SNPs were imputed based on the Haplotype Reference
236 Consortium reference panel [34]. The distribution of the 10,000 SNPs by chro-
237 mosome and whether or not the SNP was imputed is shown in Figure B.1 in
238 Supplemental Section B. We see that every chromosome contributed SNPs to

the model with 15% coming from chromosome 6. We used `gmmix`, `twostep` and `lasso` to select SNPs most predictive of the inverse normalized height on the training set, and chose the λ 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) [35] implemented in the GEMMA package [36] to derive a polygenic risk score on the training set. Subsequently, 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).

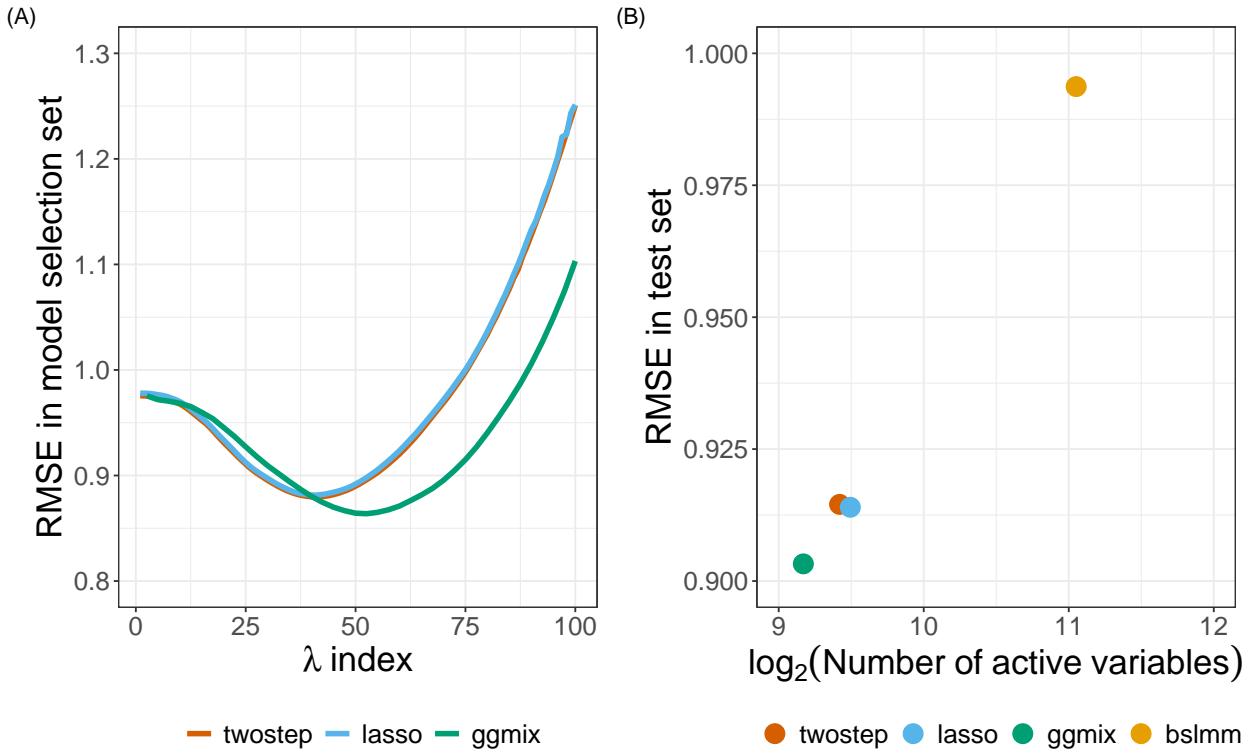


Figure 3: Model selection and testing in the UK Biobank. (A) Root-mean-square error of three methods on the model selection set with respect to a grid search of penalty factor used on the training set. (B) Performance of four methods on the test set with penalty factor optimized on the model selection 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.

257 3.2.2 GAW20

258 In the most recent Genetic Analysis Workshop 20 (GAW20), the causal modeling group in-
 259 vestigated causal relationships between DNA methylation (exposure) within some genes and
 260 the change in high-density lipoproteins Δ HDL (outcome) using Mendelian Randomization
 261 (MR) [37]. Penalized regression methods were used to select SNPs strongly associated with
 262 the exposure in order to be used as an instrumental variable (IV) [38, 39]. However, since
 263 GAW20 data consisted of families, `twostep` methods were used which could have resulted
 264 in a large number of false positives or false negatives. `ggmix` now provides an alternative

approach that could be used for selecting the IV while accounting for the family structure of the data.

We applied `gmmix` to all 200 GAW20 simulation datasets, each of 679 observations, and compared its performance to the `twostep` and `lasso` methods. Using a Factored Spectrally Transformed Linear Mixed Model (FaST-LMM) [40] adjusted for age and sex, we validated the effect of rs9661059 on blood lipid trait to be significant (genome-wide $p = 6.29 \times 10^{-9}$).

Though several other SNPs were also associated with the phenotype, these associations were probably mediated by CpG-SNP interaction pairs and did not reach statistical significance.

Therefore, to avoid ambiguity, we only focused on chromosome 1 containing 51,104 SNPs, including rs9661059. Given that population admixture in the GAW20 data was likely, we estimated the population kinship using REAP [41] after decomposing population compositions using ADMIXTURE [42]. We used 100,276 LD-pruned whole-genome genotyped SNPs for estimating the kinship. Among these, 8100 were included as covariates in our models based on chromosome 1. The causal SNP was also among the 100,276 SNPs. All methods were fit according to the same settings described in our simulation study in Section 3.1, and adjusting for age and sex. We calculated the median (inter-quartile range) number of active variables, and RMSE (standard deviation) based on five-fold CV on each simulated dataset.

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

(Table 2). We also observed that `ggmix` outperformed the `twostep` method with lower CV RMSE using the same number of SNPs. Meanwhile, it achieved roughly the same prediction accuracy as `lasso` but with fewer non-causal SNPs (Table 2). It is also worth mentioning that there was very little correlation between the causal SNP and SNPs within a 1Mb-window around it (Figure B.2 in Supplemental Section B.2), making it an ideal scenario for the `lasso` and related methods.

We also applied the `BSLMM` method by performing five-fold CV on each of the 200 simulated replicates. We found that while `BSLMM` achieved a lower CV RMSE compared to the other methods (Table 2), this higher prediction accuracy relied on approximately 80% of the 51,104 SNPs supplied. This may suggest overfitting in this dataset. It is also noteworthy that we did not adjust for age and sex in the `BSLMM` model, as the current implementation of the method in the `GEMMA` package does not allow adjustment for covariates.

Table 2: Summary of model performance based on 200 GAW20 simulations. Five-fold cross-validation root-mean-square error was reported for each simulation replicate.

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>ggmix</code>	1 (1 - 12)	0.3146 (0.0210)
<code>BSLMM</code>	40,737 (39,901 - 41,539)	0.2503 (0.0099)

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 [43]. 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 [44, 45, 46]. Therefore, a previous study developed and implemented a mixed model to find loci associated with mouse sensitivity to mycobacterial infection [47]. 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 chromosome 1 and chromosome 11 were identified.

Here we show that `gmmix` can identify these loci, as well as potentially others, in a single analysis. We reanalyzed the growth permissiveness in the spleen, as measured by colony forming units (CFUs), 6 weeks after infection from *Mycobacterium bovis* Bacille Calmette-Guerin (BCG) Russia strain as reported in [47].

By taking the consensus between the “main model” and the “conditional model” of the original study, we regarded markers D1Mit435 on chromosome 1 and D11Mit119 on chromosome 11 as two true positive loci. We directly estimated the kinship between mice using genotypes at 625 microsatellite markers. The estimated kinship entered directly into `gmmix` and `twostep`. For the `lasso`, we calculated and included the first 10 principal components of the estimated kinship. To evaluate the robustness of different models, we bootstrapped the 189-sample dataset and repeated the analysis 200 times. **We then conceived a two-fold criteria to evaluate performance of each model. We first examined whether a model could pick up both true positive loci using some λ . If the model failed to pick up both loci simultaneously with any λ , we counted as modeling failure on the corresponding bootstrap replicate; otherwise, we counted as modeling success and recorded which other loci were picked up given the largest λ .** Con-

sequently, similar to the strategy used in the GAW20 analysis, we optimized the models by tuning the penalty factor such that these two true positive loci were picked up, while the number of other active loci was minimized. Significant markers were defined as those captured in at least half of the successful bootstrap replicates (Figure 4).

We demonstrated that `gmmix` recognized the true associations more robustly than `twostep` and `lasso`. In almost all (99%) bootstrap replicates, `gmmix` was able to capture both true positives, while the `twostep` failed in 19% of the replicates and the `lasso` failed in 56% of the replicates by missing at least one of the two true positives (Figure 4). **The robustness of `gmmix` is particularly noteworthy due to the strong correlations between all microsatellite markers in this dataset (Figure B.3 in Supplemental Section B.2).** These strong correlations with the causal markers, partially explain the poor performance of the `lasso` as it suffers from unstable selections in the presence of correlated variables (e.g. [48]).

We also identified several other loci that might also be associated with susceptibility to mycobacterial infection (Table 3). Among these new potentially-associated markers, D2Mit156 was found to play a role in control of parasite numbers of *Leishmania tropica* in lymph nodes [49]. An earlier study identified a parent-of-origin effect at D17Mit221 on CD4M levels [50]. This effect was more visible in crosses than in parental strains. In addition, D14Mit131, selected only by `gmmix`, was found to have a 9% loss of heterozygosity in hybrids of two inbred mouse strains [51], indicating the potential presence of putative suppressor genes pertaining to immune surveillance and tumor progression [52]. This result might also suggest association with anti-bacterial responses yet to be discovered.

We did not apply the `BSLMM` method because the microsatellite marker-based genotypes could not be converted to a `BIMBAM` or `PLINK` format that the package demands.

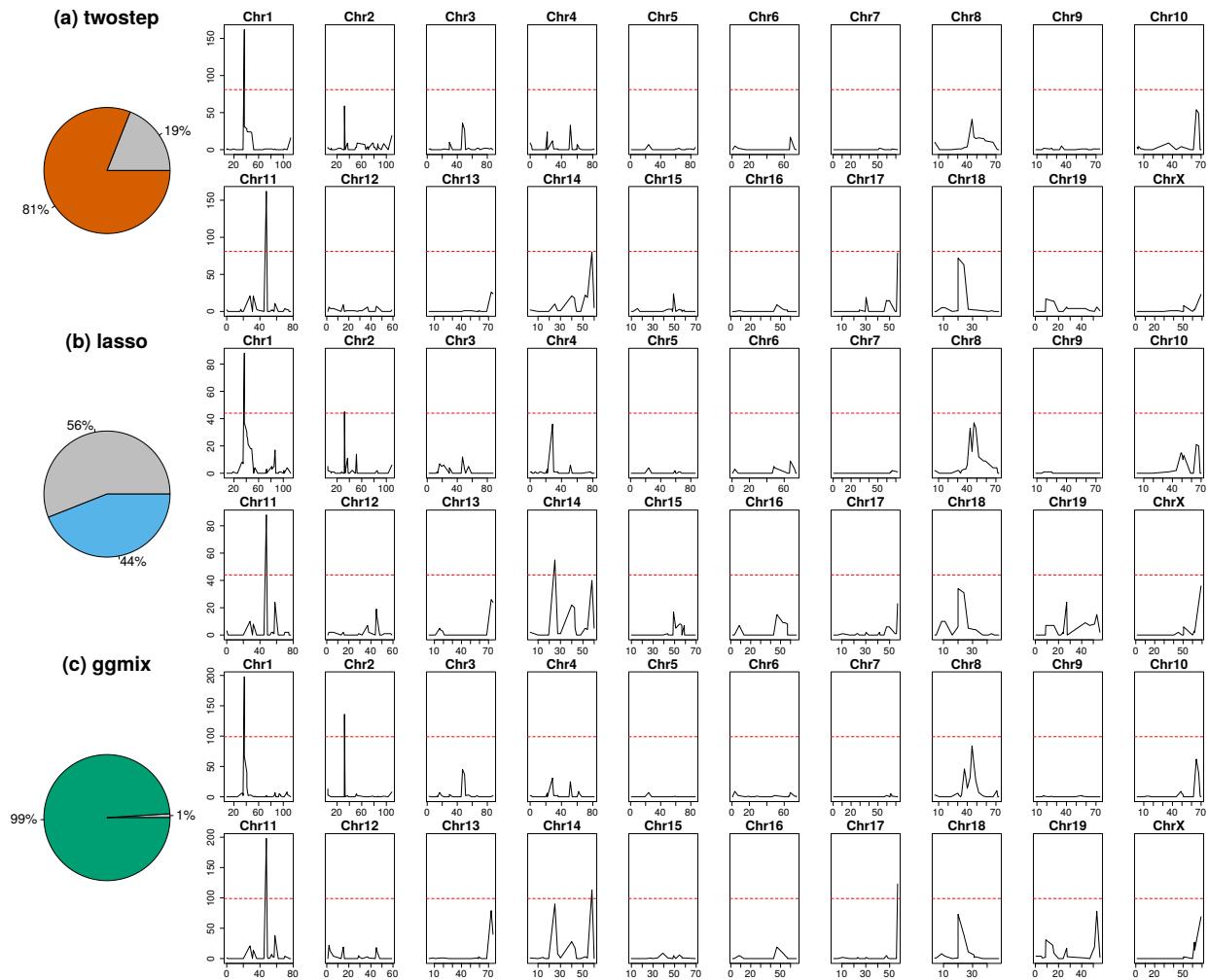


Figure 4: Comparison of model performance on the mouse cross data. Pie charts depict model robustness where grey areas denote bootstrap replicates on which the corresponding model is unable to capture both true positives using any penalty factor, whereas colored areas denote successful replicates. Chromosome-based 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.

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

360 4 Discussion

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

372 Through an extensive simulation study and three real data analyses that mimic many ex-
 373 perimental designs in genetics, we show that the current approaches of PC adjustment and

374 two-stage procedures are not necessarily sufficient to control for confounding by population
375 structure leading to a high number of false positives. Our simulation results show that **gmmix**
376 outperforms existing methods in terms of sparsity and prediction error even when the causal
377 variants are included in the kinship matrix (Table 1). Many methods for single-SNP analyses
378 avoid this proximal contamination [8] by using a leave-one-chromosome-out scheme [53], i.e.,
379 construct the kinship matrix using all chromosomes except the one on which the marker be-
380 ing tested is located. However, this approach is not possible if we want to model many SNPs
381 (across many chromosomes) jointly. For example, in the UK Biobank, we found that with
382 a kinship matrix estimated using all genotyped SNPs, **gmmix** had achieved a lower RMSE
383 on the model selection set compared to the **twostep** and **lasso** methods. Furthermore, an
384 optimized **gmmix**-derived polygenic risk score that utilized the least number of SNPs was also
385 able to better predict the trait with lower RMSE on the test set. In the GAW20 example,
386 we showed that while all methods were able to select the strongest causal SNP, **gmmix** did so
387 with the least amount of false positives while also maintaining good predictive ability. In the
388 mouse crosses example, we showed that **gmmix** is robust to perturbations in the data using
389 a bootstrap analysis. Indeed, **gmmix** was able to consistently select the true positives across
390 bootstrap replicates, while **twostep** failed in 19% of the replicates and **lasso** failed in 56% of
391 the replicates by missing of at least one of the two true positives. Our re-analysis of the data
392 also lead to some potentially new findings, not found by existing methods, that may warrant
393 further study. This particular example had many markers that were strongly correlated with
394 each other (Figure B.3 of Supplemental Section B.2). However, in 200 bootstraps, we still
395 observed that the two true positive loci (indicated) were the most often selected while none
396 of the nearby markers were picked up in more than 50% of the bootstrap replicates. This
397 shows that our method does recognize the true positives in the presence of highly correlated
398 markers. Nevertheless, we think the issue of variable selection for correlated SNPs warrants
399 further study. The recently proposed Precision Lasso [48] seeks to address this problem in
400 the high-dimensional fixed effects model.

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

411 Although we derive a CGD algorithm for the ℓ_1 penalty, our approach can also be easily ex-
412 tended to other penalties such as the elastic net and group lasso with the same guarantees of
413 convergence. A limitation of `gmmix` is that it first requires computing the covariance matrix
414 with a computation time of $\mathcal{O}(n^2k)$ followed by a spectral decomposition of this matrix in
415 $\mathcal{O}(n^3)$ time where k is the number of SNP genotypes used to construct the covariance matrix.
416 This computation becomes prohibitive for large cohorts such as the UK Biobank [54] which
417 have collected genetic information on half a million individuals. When the matrix of geno-
418 types used to construct the covariance matrix is low rank, there are additional computational
419 speedups that can be implemented. While this has been developed for the univariate case [8],
420 to our knowledge, this has not been explored in the multivariable case. We are currently
421 developing a low rank version of the penalized LMM developed here, which reduces the time
422 complexity from $\mathcal{O}(n^2k)$ to $\mathcal{O}(nk^2)$. There is also the issue of how our model scales with
423 an increasing number of covariates (p). Due to the coordinate-wise optimization procedure,
424 we expect this to be less of an issue, but still prohibitive for $p > 1e5$. The `biglasso` pack-
425 age [55] uses memory mapping strategies for large p , and this is something we are exploring
426 for `gmmix`.

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

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

434 There are other applications in which our method could be used as well. For example, there
435 has been a renewed interest in polygenic risk scores (PRS) which aim to predict complex
436 diseases from genotypes. `gmmix` could be used to build a PRS with the distinct advantage
437 of modeling SNPs jointly, allowing for main effects as well as interactions to be accounted
438 for. Based on our results, `gmmix` has the potential to produce more robust and parsimonious
439 models than the `lasso` with better predictive accuracy. Our method is also suitable for fine
440 mapping SNP association signals in genomic regions, where the goal is to pinpoint individual
441 variants most likely to impact the underlying biological mechanisms of disease [56].

442 **5 Materials and Methods**

443 **5.1 Model Set-up**

444 Let $i = 1, \dots, N$ be a grouping index, $j = 1, \dots, n_i$ the observation index within a group
445 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
446 the observed vector of responses or phenotypes, \mathbf{X}_i an $n_i \times (p + 1)$ design matrix (with
447 the column of 1s for the intercept), \mathbf{b}_i a group-specific random effect vector of length
448 n_i and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$ the individual error terms. Denote the stacked vectors $\mathbf{Y} =$
449 $(\mathbf{y}_1, \dots, \mathbf{y}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_N)^T \in \mathbb{R}^{N_T \times 1}$, and the
450 stacked matrix

⁴⁵¹ $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_N^T) \in \mathbb{R}^{N_T \times (p+1)}$. Furthermore, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T \in \mathbb{R}^{(p+1) \times 1}$ be a vec-
⁴⁵² tor of fixed effects regression coefficients corresponding to \mathbf{X} . We consider the following
⁴⁵³ linear mixed model with a single random effect [57]:

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

⁴⁵⁴ where the random effect \mathbf{b} and the error variance $\boldsymbol{\varepsilon}$ are assigned the distributions

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

⁴⁵⁵ Here, $\boldsymbol{\Phi}_{N_T \times N_T}$ is a known positive semi-definite and symmetric covariance or kinship ma-
⁴⁵⁶ trix calculated from SNPs sampled across the genome, $\mathbf{I}_{N_T \times N_T}$ is the identity matrix and
⁴⁵⁷ parameters σ^2 and $\eta \in [0, 1]$ determine how the variance is divided between \mathbf{b} and $\boldsymbol{\varepsilon}$. Note
⁴⁵⁸ that η is also the narrow-sense heritability (h^2), defined as the proportion of phenotypic
⁴⁵⁹ variance attributable to the additive genetic factors [1]. The joint density of \mathbf{Y} is therefore
⁴⁶⁰ multivariate normal:

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

⁴⁶¹ The LMM-Lasso method [15] considers an alternative but equivalent parameterization given
⁴⁶² 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 $\delta = \sigma_e^2/\sigma_g^2$, σ_g^2 is the genetic variance and σ_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 $\delta \in [0, \infty)$ [57]. We define the complete parameter

vector as $\Theta := (\beta, \eta, \sigma^2)$. The negative log-likelihood for (4) is given by

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

⁴⁶³ 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 Φ , 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 Λ_i . \mathbf{V} can then be further simplified [57]

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

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

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

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

464 where $\tilde{\mathbf{Y}} = \mathbf{U}^T \mathbf{Y}$, $\tilde{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$, \tilde{Y}_i denotes the i^{th} element of $\tilde{\mathbf{Y}}$, \tilde{X}_{ij} is the i, j^{th} entry of $\tilde{\mathbf{X}}$
465 and $\mathbf{1}$ is a column vector of N_T ones.

466 5.2 Penalized Maximum Likelihood Estimator

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

472 function (14). The penalty term is a necessary constraint because in our applications, the
473 sample size is much smaller than the number of predictors. We define the following objective
474 function:

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

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

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

480 This is the general set-up for our model. In Section 5.3 we provide more specific details on
481 how we solve (16). We note here that the main difference between the proposed model, and
482 the `lmmlasso` [58], is that we are limiting ourselves to a single unpenalized random effect.
483 Another key difference is that we rotate the response vector Y and the design matrix X
484 by the eigen vectors of the kinship matrix, resulting in a diagonal covariance matrix which
485 greatly speeds up the computation.

486 5.3 Computational Algorithm

487 We use a general purpose block coordinate gradient descent algorithm (CGD) [59] to solve (16).
488 At each iteration, we cycle through the coordinates and minimize the objective function with
489 respect to one coordinate only. For continuously differentiable $f(\cdot)$ and convex and block-
490 separable $P(\cdot)$ (i.e. $P(\beta) = \sum_i P_i(\beta_i)$), Tseng and Yun [59] show that the solution gener-
491 ated by the CGD method is a stationary point of $Q_\lambda(\cdot)$ if the coordinates are updated in a
492 Gauss-Seidel manner i.e. $Q_\lambda(\cdot)$ is minimized with respect to one parameter while holding

⁴⁹³ all others fixed. The CGD algorithm has been successfully applied in fixed effects models
⁴⁹⁴ (e.g. [60], [20]) and linear mixed models with an ℓ_1 penalty [58]. In the next section we
⁴⁹⁵ provide some brief details about Algorithm 1. A more thorough treatment of the algorithm
⁴⁹⁶ 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 ϵ ;

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

⁴⁹⁷ **5.3.1 Updates for the β parameter**

⁴⁹⁸ Recall that the part of the objective function that depends on β has the form

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

⁴⁹⁹ 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 β_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

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

500 $\text{sign}(x)$ is the signum function

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

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

502 **5.3.2 Updates for the η parameter**

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

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

504 We use a bound constrained optimization algorithm [61] implemented in the `optim` function

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

506 **5.3.3 Updates for the σ^2 parameter**

507 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:

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

508 **5.3.4 Regularization path**

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

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

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

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

512 The equations in (24) are equivalent to

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

513 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
 514 of $\tilde{\mathbf{X}}^T$, and $\boldsymbol{\gamma} \in \mathbb{R}^p$ is the subgradient function of the ℓ_1 norm evaluated at $(\hat{\beta}_1, \dots, \hat{\beta}_p)$.

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

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

519 We can solve the KKT system of equations in (26) (with a numerical solution for η) in order

520 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
 521 can solve for the smallest value of λ such that the entire vector $(\widehat{\beta}_1, \dots, \widehat{\beta}_p)$ is 0:

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

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

525 **5.3.5 Warm Starts**

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

530 **5.3.6 Prediction of the random effects**

531 We use an empirical Bayes approach (e.g. [62]) to predict the random effects \mathbf{b} . Let the
 532 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

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

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:

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

533 5.3.7 Phenotype prediction

534 Here we describe the method used for predicting the unobserved phenotype \mathbf{Y}^* in a set of
 535 individuals with predictor set \mathbf{X}^* that were not used in the model training e.g. a testing
 536 set. Let q denote the number of observations in the testing set and $N - q$ the number of
 537 observations in the training set. We assume that a `gmmix` model has been fit on a set of
 538 training individuals with observed phenotype \mathbf{Y} and predictor set \mathbf{X} . We further assume
 539 that \mathbf{Y} and \mathbf{Y}^* are jointly multivariate Normal:

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

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

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

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

542 The phenotype prediction is thus given by:

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

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

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

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

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

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

545 5.3.8 Choice of the optimal tuning parameter

546 In order to choose the optimal value of the tuning parameter λ , we use the generalized
547 information criterion [63] (GIC):

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

548 where \hat{df}_λ is the number of non-zero elements in $\hat{\boldsymbol{\beta}}_\lambda$ [64] plus two (representing the variance
549 parameters η and σ^2). Several authors have used this criterion for variable selection in mixed
550 models with $a_n = \log N_T$ [58, 65], which corresponds to the BIC. We instead choose the high-
551 dimensional BIC [66] given by $a_n = \log(\log(N_T)) * \log(p)$. This is the default choice in our
552 **ggmix** R package, though the interface is flexible to allow the user to select their choice of
553 a_n .

554 **Availability of data and material**

- 555 1. The UK Biobank data is available upon successful project application.
- 556 2. The GAW20 data is freely available upon request from <https://www.gaworkshop.org/data-sets>.
- 557 3. Mouse cross data is available from GitHub at <https://github.com/sahirbhatnagar/ggmix/blob/pgen/RealData/mice.RData>.
- 558 4. The entire simulation study is reproducible. Source code available at <https://github.com/sahirbhatnagar/ggmix/tree/pgen/simulation>. This includes scripts for `ggmix`,
559 `lasso` and `twostep` methods.
- 560 5. The R package `ggmix` is freely available from GitHub at <https://github.com/greenwoodlab/ggmix>.
- 561 6. A website describing how to use the package is available at <https://sahirbhatnagar.com/ggmix/>.
- 562
- 563
- 564
- 565
- 566
- 567

567 **Competing interests**

568 The authors declare that they have no competing interests.

569 **Author's contributions**

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

574 **Acknowledgements**

575 SRB was supported by the Ludmer Centre for Neuroinformatics and Mental Health and
576 the Canadian Institutes for Health Research PJT 148620. This research was enabled in
577 part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada
578 (www.computecanada.ca). The funders had no role in study design, data collection and
579 analysis, decision to publish, or preparation of the manuscript.

580 **Supporting Information**

581 Contains the following sections:

- 582 A **Block Coordinate Descent Algorithm** - a detailed description of the algorithm
583 used to fit our `ggmix` model.
- 584 B **Additional Real Data Analysis Results** - supporting information for the GAW20
585 and UK Biobank analyses
- 586 C **ggmix Package Showcase** - a vignette describing how to use our `ggmix` R package

587 **References**

- 588 [1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding
589 the missing heritability of complex diseases. *Nature*. 2009;461(7265):747. [3](#), [24](#)
- 590 [2] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common
591 SNPs explain a large proportion of the heritability for human height. *Nature genetics*.
592 2010;42(7):565. [3](#)

- 593 [3] Astle W, Balding DJ, et al. Population structure and cryptic relatedness in genetic
594 association studies. *Statistical Science*. 2009;24(4):451–471. [3](#), [4](#)
- 595 [4] Song M, Hao W, Storey JD. Testing for genetic associations in arbitrarily structured
596 populations. *Nature genetics*. 2015;47(5):550–554. [3](#)
- 597 [5] Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population
598 structure on large genetic association studies. *Nature genetics*. 2004;36(5):512. [3](#)
- 599 [6] Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in
600 genome-wide and re-sequencing association studies. *PLoS genetics*. 2008;4(7):e1000130.
601 [4](#)
- 602 [7] Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies.
603 *Bioinformatics*. 2010;27(4):516–523. [4](#)
- 604 [8] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear
605 mixed models for genome-wide association studies. *Nature methods*. 2011;8(10):833–
606 835. [4](#), [7](#), [21](#), [22](#)
- 607 [9] Kang HM, Sul JH, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, et al. Variance
608 component model to account for sample structure in genome-wide association studies.
609 *Nature genetics*. 2010;42(4):348. [4](#)
- 610 [10] Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-
611 model method for association mapping that accounts for multiple levels of relatedness.
612 *Nature genetics*. 2006;38(2):203. [4](#)
- 613 [11] Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell
614 HJ, et al. Comparison of methods to account for relatedness in genome-wide association
615 studies with family-based data. *PLoS Genet*. 2014;10(7):e1004445. [4](#)

- 616 [12] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
617 components analysis corrects for stratification in genome-wide association studies.
618 *Nature genetics*. 2006;38(8):904. 4
- 619 [13] Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, et al. Adjusted
620 sequence kernel association test for rare variants controlling for cryptic and family
621 relatedness. *Genetic epidemiology*. 2013;37(4):366–376. 4, 5
- 622 [14] Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the
623 relative effects of polymorphisms within a gene using case/control or family data:
624 application to HLA in type 1 diabetes. *The American Journal of Human Genetics*.
625 2002;70(1):124–141. 4
- 626 [15] Rakitsch B, Lippert C, Stegle O, Borgwardt K. A Lasso multi-marker mixed
627 model for association mapping with population structure correction. *Bioinformatics*.
628 2013;29(2):206–214. 4, 22, 24
- 629 [16] Wang D, Eskridge KM, Crossa J. Identifying QTLs and epistasis in structured plant
630 populations using adaptive mixed LASSO. *Journal of agricultural, biological, and en-*
631 *vironmental statistics*. 2011;16(2):170–184. 4
- 632 [17] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal*
633 *Statistical Society Series B (Methodological)*. 1996;p. 267–288. 4, 5
- 634 [18] Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical*
635 *association*. 2006;101(476):1418–1429. 5
- 636 [19] Ding X, Su S, Nandakumar K, Wang X, Fardo DW. A 2-step penalized regression
637 method for family-based next-generation sequencing association studies. In: *BMC pro-*
638 *ceedings*. vol. 8. BioMed Central; 2014. p. S25. 5

- 639 [20] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models
640 via coordinate descent. *Journal of statistical software*. 2010;33(1):1. 5, 6, 29, 33, 47
- 641 [21] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalized learning
642 problems. *Statistics and Computing*. 2015;25(6):1129–1141. 5
- 643 [22] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in
644 the application of mixed-model association methods. *Nature genetics*. 2014;46(2):100.
645 5, 23
- 646 [23] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of*
647 *the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301–320.
648 5
- 649 [24] Yuan M, Lin Y. Model selection and estimation in regression with grouped vari-
650 ables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
651 2006;68(1):49–67. 5
- 652 [25] Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algo-
653 rithm for variance parameter estimation in linear mixed models. *Biometrics*. 1995;p.
654 1440–1450. 6
- 655 [26] Dandine-Roulland C. *gaston: Genetic Data Handling (QC, GRM, LD, PCA) and*
656 *Linear Mixed Models*; 2018. R package version 1.5.3. Available from: <https://CRAN.R-project.org/package=gaston>. 6
- 658 [27] Ochoa A, Storey JD. FST and kinship for arbitrary population structures I: Generalized
659 definitions. *bioRxiv*. 2016;. 8
- 660 [28] Ochoa A, Storey JD. FST and kinship for arbitrary population structures II: Method
661 of moments estimators. *bioRxiv*. 2016;. 8

- 662 [29] Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regres-
663 sion. *Statistica Sinica*. 2016;p. 35–67. 11
- 664 [30] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank
665 resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203. 12
- 666 [31] Biobank U. Genotyping and quality control of UK Biobank, a large-scale, ex-
667 tensively phenotyped prospective resource. Available at: biobank.ctsu.ox.ac.
668 uk/crystal/docs/genotyping_qc.pdf Accessed April. 2015;1:2016. 12
- 669 [32] Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relation-
670 ship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867–
671 2873. 12
- 672 [33] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-
673 analysis of genome-wide association studies for height and body mass index in 700000
674 individuals of European ancestry. *Human molecular genetics*. 2018;27(20):3641–3649.
675 12
- 676 [34] 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*.
677 2016;48(10):1279. 12
- 679 [35] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear
680 mixed models. *PLoS genetics*. 2013;9(2):e1003264. 13
- 681 [36] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association stud-
682 ies. *Nature genetics*. 2012;44(7):821. 13
- 683 [37] Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology con-
684 tribute to understanding environmental determinants of disease? *International journal
685 of epidemiology*. 2003;32(1):1–22. 14

- 686 [38] Cherlin S, Howey RA, Cordell HJ. Using penalized regression to predict phenotype
687 from SNP data. In: BMC proceedings. vol. 12. BioMed Central; 2018. p. 38. **14**
- 688 [39] Zhou W, Lo SH. Analysis of genotype by methylation interactions through sparsity-
689 inducing regularized regression. In: BMC proceedings. vol. 12. BioMed Central; 2018.
690 p. 40. **14**
- 691 [40] Howey RA, Cordell HJ. Application of Bayesian networks to GAW20 genetic and blood
692 lipid data. In: BMC proceedings. vol. 12. BioMed Central; 2018. p. 19. **15**
- 693 [41] Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Esti-
694 mating kinship in admixed populations. The American Journal of Human Genetics.
695 2012;91(1):122–138. **15**
- 696 [42] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
697 unrelated individuals. Genome research. 2009;19(9):1655–1664. **15**
- 698 [43] Fortin A, Diez E, Rochefort D, Laroche L, Malo D, Rouleau GA, et al. Recombinant
699 congenic strains derived from A/J and C57BL/6J: a tool for genetic dissection of com-
700 plex traits. Genomics. 2001;74(1):21–35. **17**
- 701 [44] Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, et al. A
702 high-resolution association mapping panel for the dissection of complex traits in mice.
703 Genome research. 2010;20(2):281–290. **17**
- 704 [45] Flint J, Eskin E. Genome-wide association studies in mice. Nature Reviews Genetics.
705 2012;13(11):807. **17**
- 706 [46] Cheng R, Lim JE, Samocha KE, Sokoloff G, Abney M, Skol AD, et al. Genome-wide
707 association studies and the problem of relatedness among advanced intercross lines and
708 other highly recombinant populations. Genetics. 2010;185(3):1033–1044. **17**

- 709 [47] Di Pietrantonio T, Hernandez C, Girard M, Verville A, Orlova M, Belley A, et al.
710 Strain-specific differences in the genetic control of two closely related mycobacteria.
711 PLoS pathogens. 2010;6(10):e1001169. 17
- 712 [48] Wang H, Lengerich BJ, Aragam B, Xing EP. Precision Lasso: accounting for cor-
713 relations and linear dependencies in high-dimensional genomic data. Bioinformatics.
714 2018;35(7):1181–1187. 18, 21
- 715 [49] Sohrabi Y, Havelková H, Kobets T, Šíma M, Volkova V, Grekov I, et al. Mapping the
716 Genes for Susceptibility and Response to Leishmania tropica in Mouse. PLoS neglected
717 tropical diseases. 2013;7(7):e2282. 18
- 718 [50] Jackson AU, Fornés A, Galecki A, Miller RA, Burke DT. Multiple-trait quantitative
719 trait loci analysis using a large mouse sibship. Genetics. 1999;151(2):785–795. 18
- 720 [51] C Stern1 M, Benavides F, A Klingelberger E, J Conti2 C. Allelotype analysis of chemi-
721 cally induced squamous cell carcinomas in F1 hybrids of two inbred mouse strains with
722 different susceptibility to tumor progression. Carcinogenesis. 2000;21(7):1297–1301. 18
- 723 [52] Lasko D, Cavenee W, Nordenskjöld M. Loss of constitutional heterozygosity in human
724 cancer. Annual review of genetics. 1991;25(1):281–314. 18
- 725 [53] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al.
726 Efficient Bayesian mixed-model analysis increases association power in large cohorts.
727 Nature genetics. 2015;47(3):284. 21
- 728 [54] Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank:
729 Current status and what it means for epidemiology. Health Policy and Technology.
730 2012;1(3):123–126. 22
- 731 [55] Zeng Y, Breheny P. The biglasso package: a memory-and computation-efficient solver
732 for lasso model fitting with big data in R. arXiv preprint arXiv:170105936. 2017;. 22

- 733 [56] Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human molecular
734 genetics. 2015;24(R1):R111–R119. [23](#)
- 735 [57] Pirinen M, Donnelly P, Spencer CC, et al. Efficient computation with a linear mixed
736 model on large-scale data sets with applications to genetic studies. The Annals of
737 Applied Statistics. 2013;7(1):369–390. [24](#), [25](#)
- 738 [58] Schelldorfer J, Bühlmann P, DE G, VAN S. Estimation for High-Dimensional Lin-
739 ear Mixed-Effects Models Using L1-Penalization. Scandinavian Journal of Statistics.
740 2011;38(2):197–214. [28](#), [29](#), [36](#), [47](#)
- 741 [59] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable mini-
742 mization. Mathematical Programming. 2009;117(1):387–423. [28](#), [47](#), [50](#)
- 743 [60] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. Journal
744 of the Royal Statistical Society: Series B (Statistical Methodology). 2008;70(1):53–71.
745 [29](#), [47](#)
- 746 [61] Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained
747 optimization. SIAM Journal on Scientific Computing. 1995;16(5):1190–1208. [30](#)
- 748 [62] Wakefield J. Bayesian and frequentist regression methods. Springer Science & Business
749 Media; 2013. [33](#)
- 750 [63] Nishii R. Asymptotic properties of criteria for selection of variables in multiple regres-
751 sion. The Annals of Statistics. 1984;p. 758–765. [36](#)
- 752 [64] Zou H, Hastie T, Tibshirani R, et al. On the degrees of freedom of the lasso. The
753 Annals of Statistics. 2007;35(5):2173–2192. [36](#)
- 754 [65] Bondell HD, Krishna A, Ghosh SK. Joint Variable Selection for Fixed and Random
755 Effects in Linear Mixed-Effects Models. Biometrics. 2010;66(4):1069–1077. [36](#)

- 756 [66] Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likeli-
757 hood. Journal of the Royal Statistical Society: Series B (Statistical Methodology).
758 2013;75(3):531–552. [36](#)
- 759 [67] Xie Y. Dynamic Documents with R and knitr. vol. 29. CRC Press; 2015. [57](#)

760 A Block Coordinate Descent Algorithm

761 We use a general purpose block coordinate descent algorithm (CGD) [59] to solve (16). At
 762 each iteration, the algorithm approximates the negative log-likelihood $f(\cdot)$ in $Q_\lambda(\cdot)$ by a
 763 strictly convex quadratic function and then applies block coordinate decent to generate a
 764 decent direction followed by an inexact line search along this direction [59]. For continuously
 765 differentiable $f(\cdot)$ and convex and block-separable $P(\cdot)$ (i.e. $P(\beta) = \sum_i P_i(\beta_i)$), [59] show
 766 that the solution generated by the CGD method is a stationary point of $Q_\lambda(\cdot)$ if the coor-
 767 dinates are updated in a Gauss-Seidel manner i.e. $Q_\lambda(\cdot)$ is minimized with respect to one
 768 parameter while holding all others fixed. The CGD algorithm can thus be run in parallel and
 769 therefore suited for large p settings. It has been successfully applied in fixed effects models
 770 (e.g. [60], [20]) and [58] for mixed models with an ℓ_1 penalty. Following Tseng and Yun [59],
 771 the CGD algorithm is given by Algorithm 2.

772 The Armijo rule is defined as follows [59]:

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

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

773

774 Common choices for the constants are $\delta = 0.1$, $\varrho = 0.001$, $\gamma = 0$, $\alpha_{init}^{(k)} = 1$ for all k [58].

775 Below we detail the specifics of Algorithm 2 for the ℓ_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;

₇₇₆ **A.1 ℓ_1 penalty**

₇₇₇ The objective function is given by

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

₇₇₈ **A.1.1 Descent Direction**

₇₇₉ For simplicity, we remove the iteration counter (k) from the derivation below.

₇₈₀ For $\Theta_j^{(k)} \in \{\beta_1, \dots, \beta_p\}$, let

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

₇₈₁ where

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

₇₈₂ Since $G(d)$ is not differentiable at $-\Theta_j$, we calculate the subdifferential $\partial G(d)$ and search

₇₈₃ for d with $0 \in \partial G(d)$:

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

₇₈₄ where

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

₇₈₅ We consider each of the three cases in (49) below

1. $d > -\Theta_j$

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

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

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

The solution can be written compactly as

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

where $\text{mid} \{a, b, c\}$ denotes the median (mid-point) of a, b, c [59].

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

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

787 We see all three cases lead to the same solution for (48). Therefore the descent direction for
788 $\Theta_j^{(k)} \in \{\beta_1, \dots, \beta_p\}$ for the ℓ_1 penalty is given by

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

789 **A.1.2 Solution for the β parameter**

790 If the Hessian $\nabla^2 f(\Theta^{(k)}) > 0$ then $H^{(k)}$ defined in (41) is equal to $\nabla^2 f(\Theta^{(k)})$. Using $\alpha_{init} = 1$,
791 the largest element of $\{\alpha_{init}^{(k)} \delta^r\}_{r=0,1,2,\dots}$ satisfying the Armijo Rule inequality is reached for
792 $\alpha^{(k)} = \alpha_{init}^{(k)} \delta^0 = 1$. The Armijo rule update for the β parameter is then given by

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

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

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

794 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 β 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

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

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

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

⁷⁹⁶ and $(x)_+ = \max(x, 0)$.

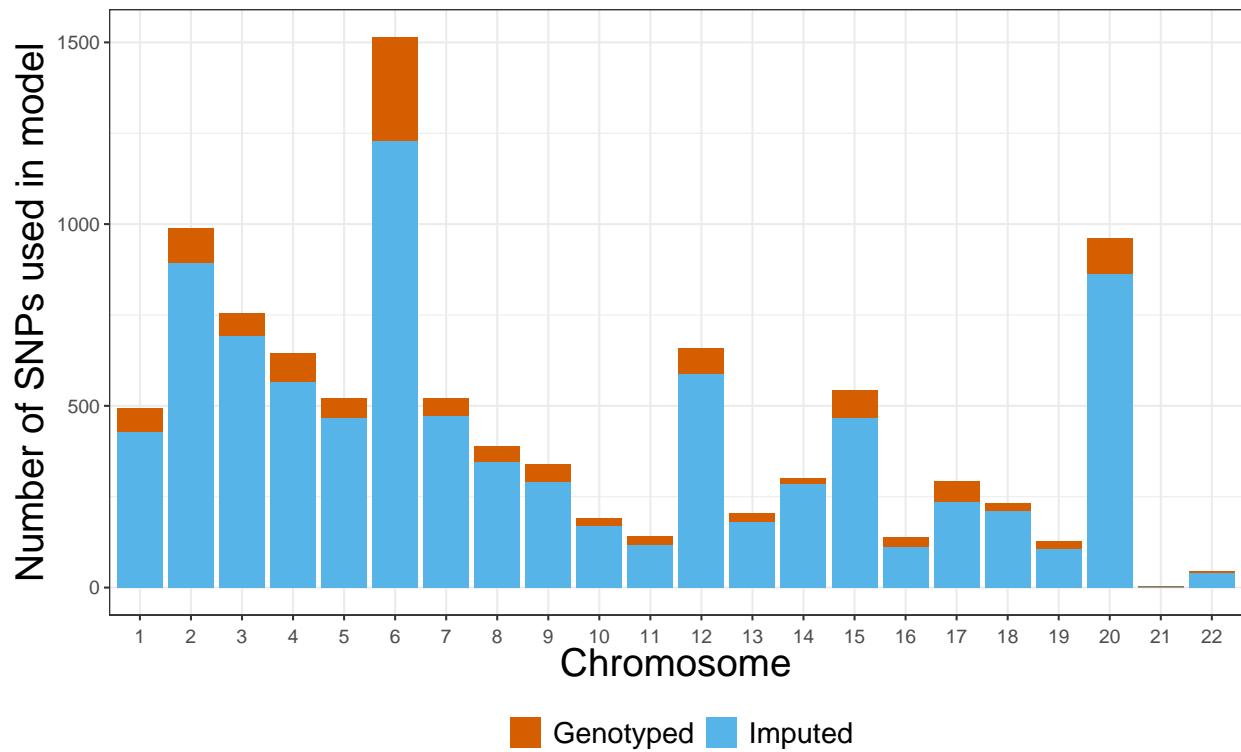
797 B Additional Real Data Analysis Results**798 B.1 Distribution of SNPs used in UK Biobank analysis**

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

799 **B.2 LD structure among the markers in the GAW20 and the**
 800 **mouse dataset**

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

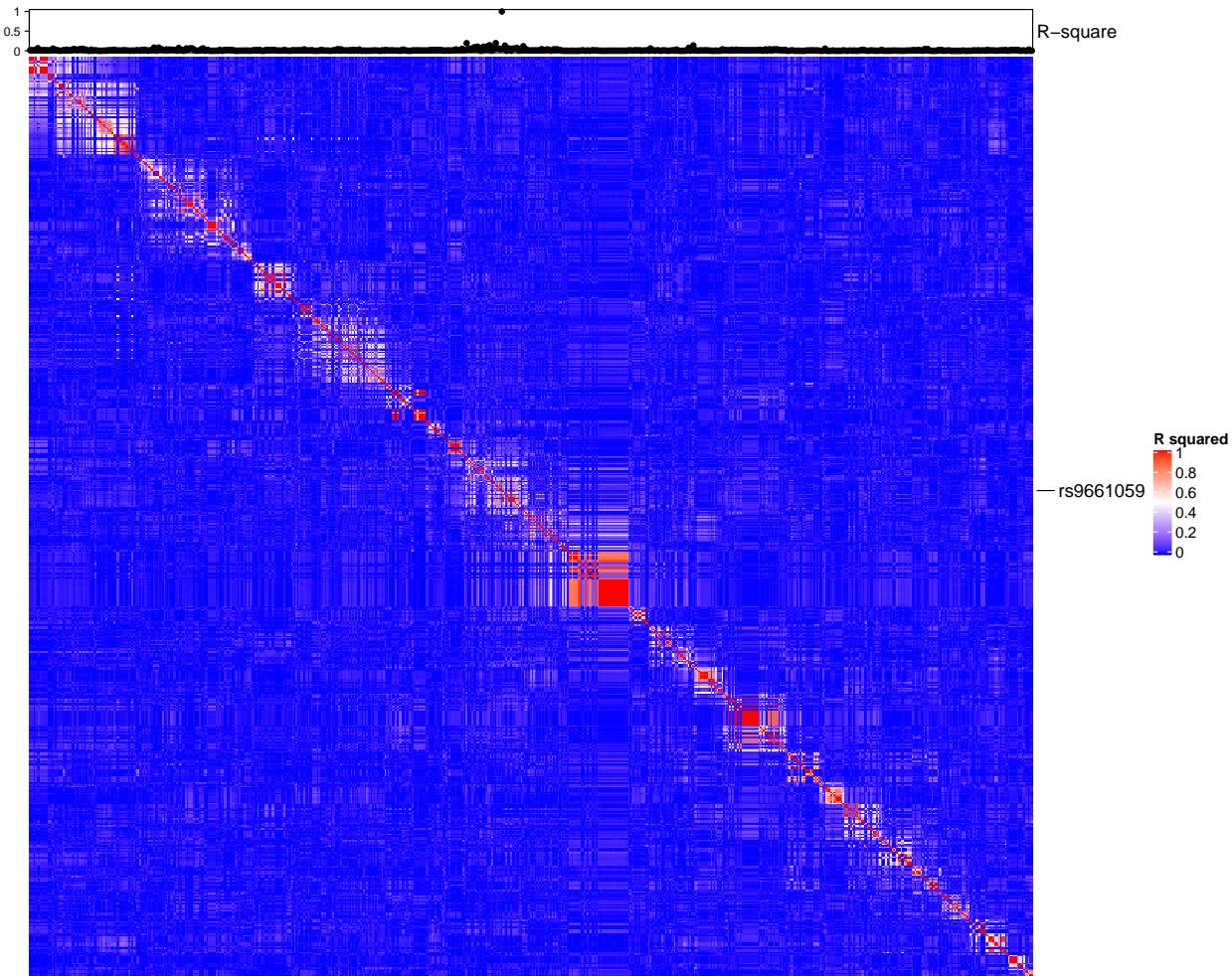


Figure B.2: LD structure among the markers in the GAW20 dataset

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

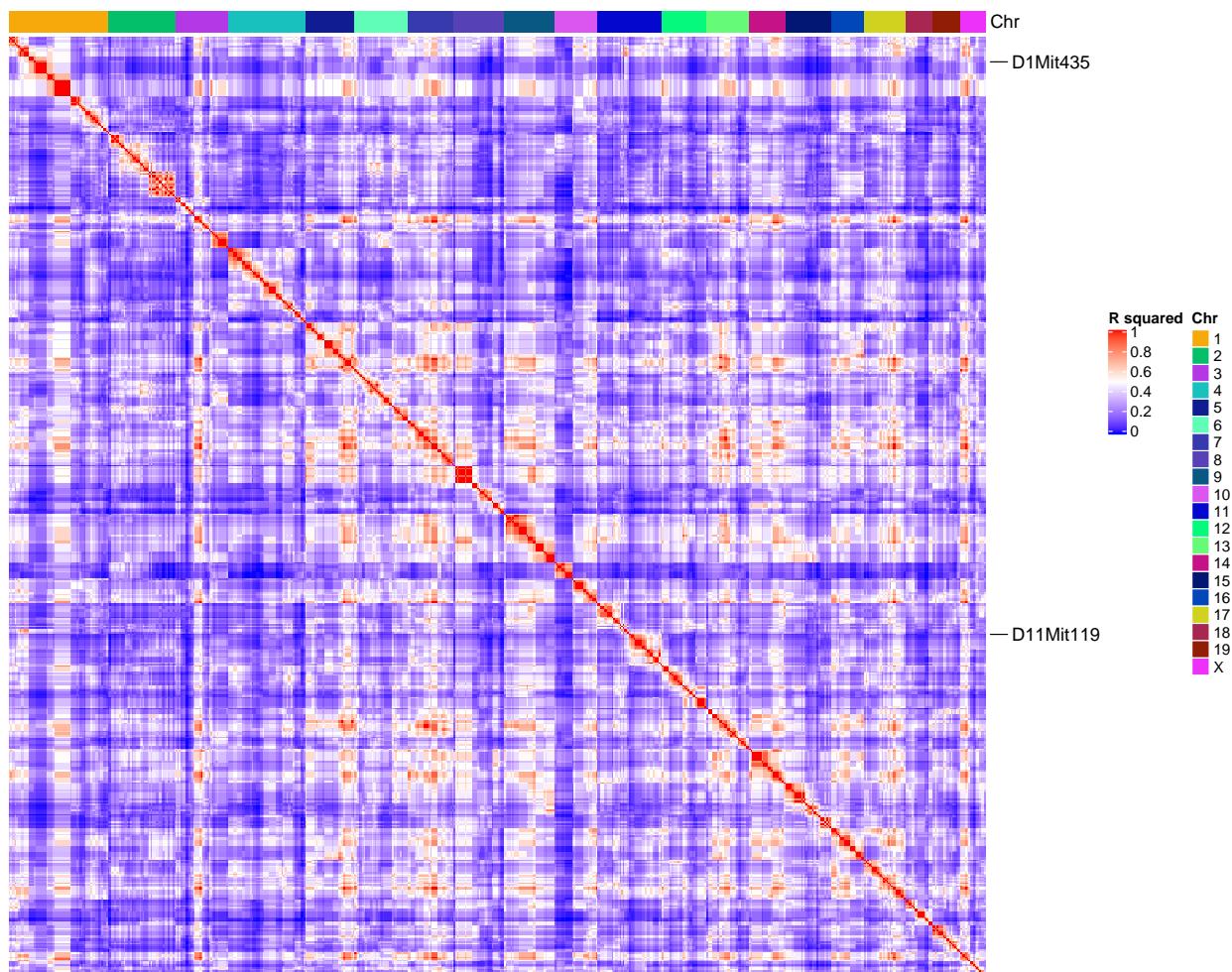


Figure B.3: LD structure among the markers in the mouse dataset

811 C ggmix Package Showcase

- 812 In this section we briefly introduce the freely available and open source `ggmix` package in R.
 813 More comprehensive documentation is available at <https://sahirbhatnagar.com/ggmix>.
 814 Note that this entire section is reproducible; the code and text are combined in an `.Rnw`¹ file
 815 and compiled using `knitr` [67].

816 C.1 Installation

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

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

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

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

## [1] "y"                 "x"                 "causal"
## [4] "beta"              "kin"               "Xkinship"
## [7] "not_causal"        "causal_positive" "causal_negative"
## [10] "x_lasso"
```

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

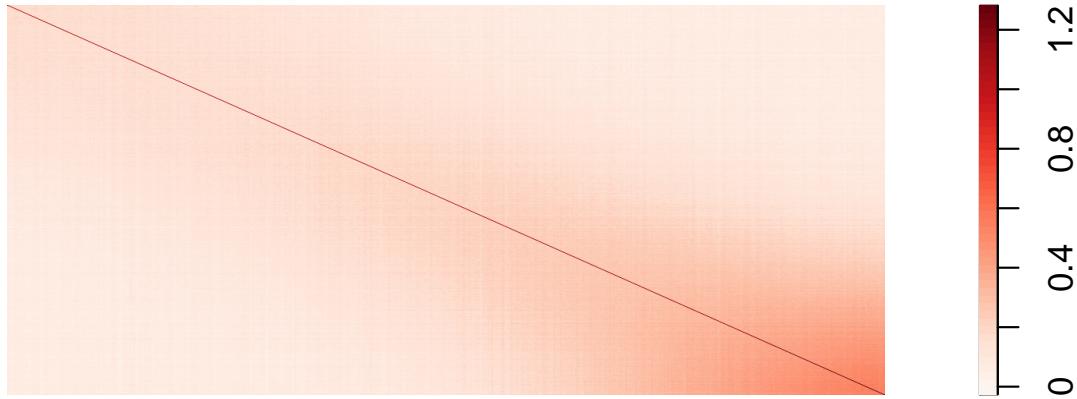
- 821 There are three basic inputs that `ggmix` needs:

- 822 1. Y : a continuous response variable
- 823 2. X : a matrix of covariates of dimension $N \times p$ where N is the sample size and p is the
 824 number of covariates
- 825 3. Φ : a kinship matrix

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

826 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::plotPopkin(admixed$kin)
```



827

828 C.2 Fit the linear mixed model with Lasso Penalty

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

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

831 values of the tuning parameter λ . It will choose its own sequence:

```
fit <- ggmix(x = admixed$x, y = admixed$y, kinship = admixed$kin)
```

```

names(fit)

## [1] "result"      "ggmix_object" "n_design"     "p_design"
## [5] "lambda"       "coef"        "b0"          "beta"
## [9] "df"           "eta"         "sigma2"      "nlambda"
## [13] "cov_names"    "call"

class(fit)

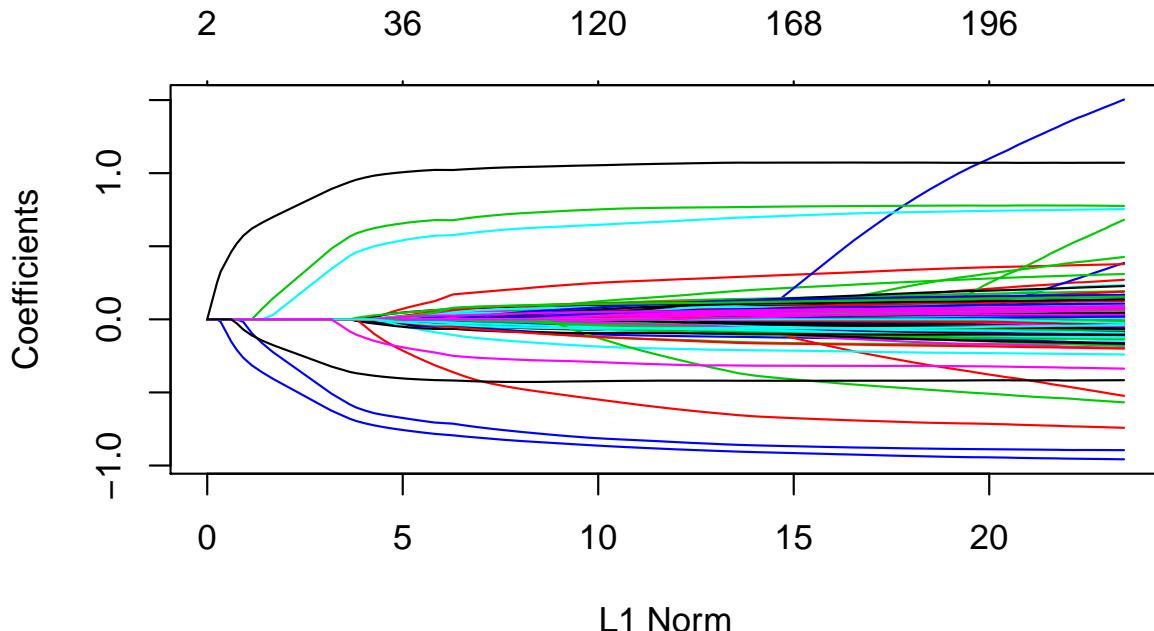
## [1] "lassofullrank" "ggmix_fit"

```

832 We can see the solution path for each variable by calling the `plot` method for objects of

833 class `ggmix_fit`:

```
plot(fit)
```



834

835 We can also get the coefficients for given value(s) of lambda using the `coef` method for

836 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) -0.3824525 -0.030224599
## X62        0.0000000  0.000000000
## X185       0.0000000  0.001444518
## X371       0.0000000  0.009513475
## X420       0.0000000  0.000000000

```

- 837 Here, `s` specifies the value(s) of λ at which the extraction is made. The function uses linear
 838 interpolation to make predictions for values of `s` that do not coincide with the lambda
 839 sequence used in the fitting algorithm.
- 840 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$x[1:5,])

##           1         2
## id1 -1.19165061 -1.3123392
## id2 -0.02913052  0.3885923
## id3 -2.00084875 -2.6460043
## id4 -0.37255277 -0.9542463
## id5 -1.03967831 -2.1377268

```

841 C.3 Find the Optimal Value of the Tuning Parameter

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

```
# pass the fitted object from ggmix to the gic function:
```

```

hdbic <- gic(fit)
class(hdbic)

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

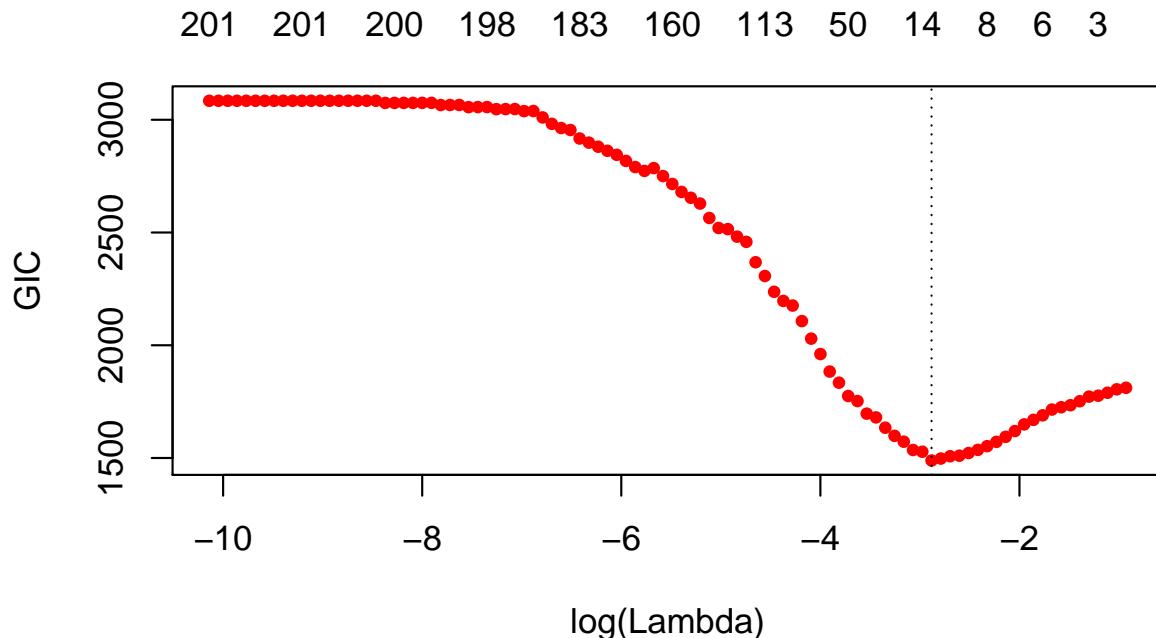
# we can also fit the BIC by specifying the an argument
bicfit <- gic(fit, an = log(length(admixed$y)))

```

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

846 `ggmix_gic`:

```
plot(hdbic)
```



847

848 The optimal value for λ according to the HDBIC, i.e., the λ that leads to the minium HDBIC

849 is:

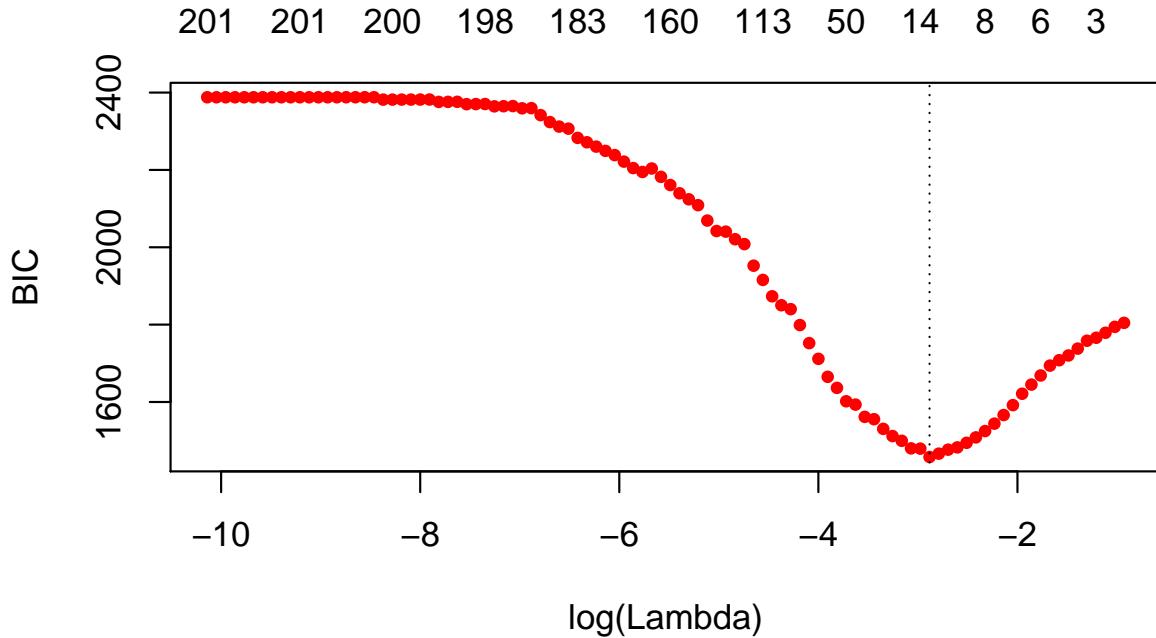
```

hdbic[["lambda.min"]]
## [1] 0.05596623

```

850 We can also plot the BIC results:

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



851

```
bicfit[["lambda.min"]]
## [1] 0.05596623
```

852 C.4 Get Coefficients Corresponding to Optimal Model

853 We can use the object outputted by the `gic` function to extract the coefficients corresponding
854 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) -0.2668419
## X62         .
## X185         .
## X371         .
## X420         .
```

855 We can also extract just the nonzero coefficients which also provide the estimated variance

856 components η and σ^2 :

```
coef(hdbic, type = "nonzero")

##           1
## (Intercept) -0.26684191
## X336      -0.67986393
## X7638      0.43403365
## X1536      0.93994982
## X1943      0.56600730
## X2849      -0.58157979
## X56       -0.08244685
## X4106      -0.35939830
## eta        0.26746240
## sigma2     0.98694300
```

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

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

##           1
## id1 -1.3061041
## id2  0.2991654
## id3 -2.3453664
## id4 -0.4486012
## id5 -1.3895793
```

859 C.5 Extracting Random Effects

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

```
ranef(hdbic)[1:5]

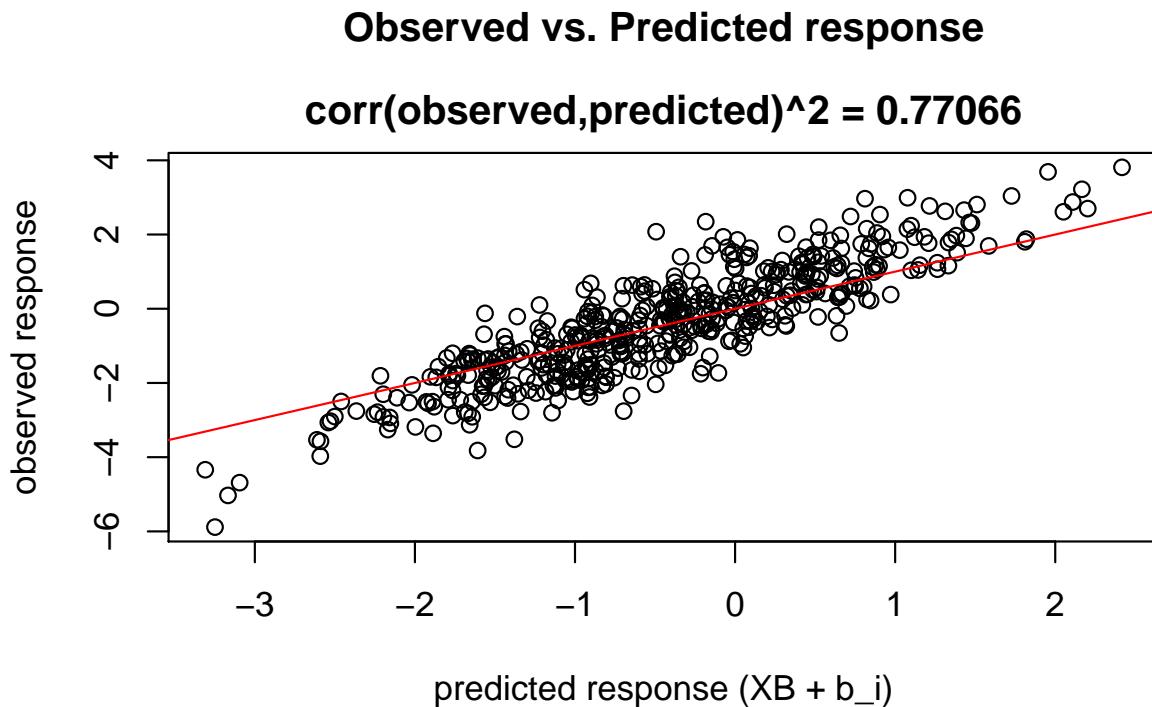
## [1] -0.02548691 -0.10011680  0.13020240 -0.30650997  0.16045768
```

863 C.6 Diagnostic Plots

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

867 C.6.1 Observed vs. Predicted Response

```
plot(hdbic, type = "predicted", newx = admixed$x, newy = admixed$y)
```

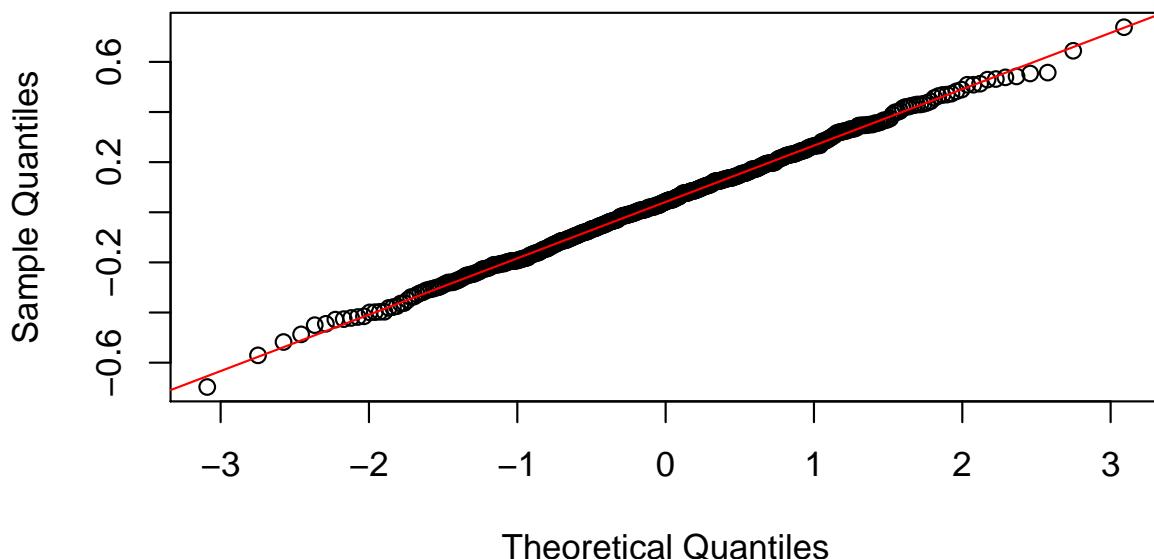


868

869 C.6.2 QQ-plots for Residuals and Random Effects

```
plot(hdbic, type = "QQranef", newx = admixed$x, newy = admixed$y)
```

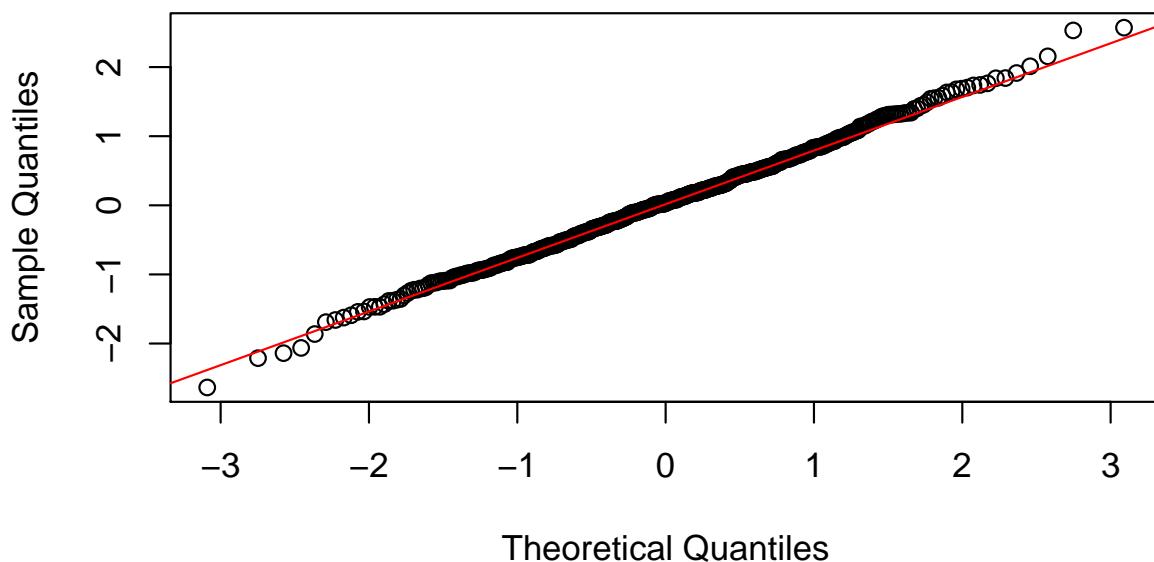
QQ-Plot of the random effects at lambda = 0.06



870

```
plot(hdbic, type = "QQresid", newx = admixed$x, newy = admixed$y)
```

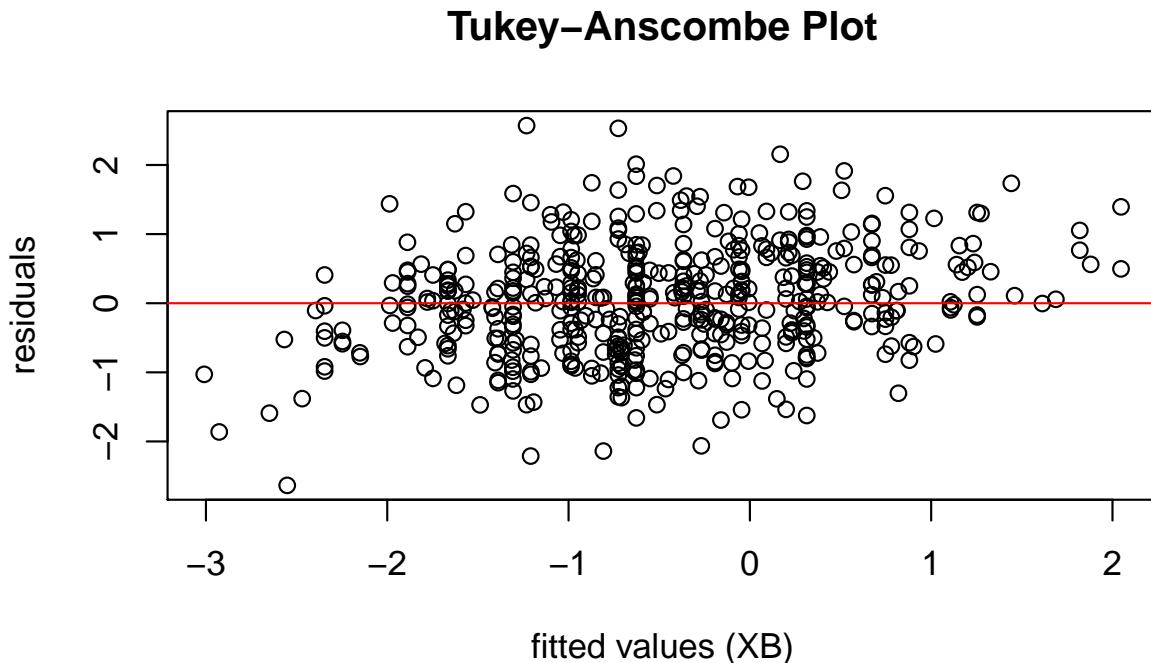
QQ-Plot of the residuals at lambda = 0.06



871

872 C.6.3 Tukey-Anscombe Plot

```
plot(hdbic, type = "Tukey", newx = admixed$x, newy = admixed$y)
```



873