MegaBayesianAlphabet: Mega-scale Bayesian Regression methods for genome-wide prediction and association studies with thousands of traits

Jiayi Qu[*], **Daniel Runcie**[†,1] **and Hao Cheng**[†,1]

[*]Department of Animal Science, University of California Davis, Davis, CA 95616, [†]Department of Plant Sciences, University of California Davis, Davis, CA 95616

ABSTRACT Large-scale phenotype data are expected to increase the accuracy of genome-wide prediction and the power of genome-wide association analyses. However, genomic analyses of high-dimensional, highly correlated data are challenging. We developed MegaBayesianAlphabet to simultaneously analyze genetic variants underlying thousands of traits using the flexible priors of the Bayesian Alphabet family. As a demonstration, we implemented the BayesC prior in the R package MegaLMM and applied it to both simulated and real data sets. Our analyses show that the resulting model MegaBayesC can effectively use high-dimensional phenotypic data to improve the accuracy of genetic value prediction, the reliability of marker discovery, and the accuracy of marker effect size estimation in genome-wide analyses.

KEYWORDS multi-trait; genomic prediction; genome-wide association studies; high-throughput phenotyping; Bayesian regression models

Introduction

The advent of high density genome-wide single nucleotide polymorphism (SNP) arrays in the past decades has provided exciting new material for the genetic analysis of complex traits. Linear mixed models that can integrate such large-scale genomic data are widely used for genomic prediction (Meuwissen *et al.* 2001; VanRaden 2008) and genome-wide association studies (Visscher *et al.* 2017). Recent advance in multi-omics methodologies now provide opportunities to generate large-scale transcriptomic, metabolomic, and epigenomic profiles as well. The integration of these high-dimensional phenotypes into association studies can increase power to detect causal variants. For example, gene expression profiling in thousands of genes has been used for the identification of genes that affect transcriptional variation (i.e., eQTLs) (Gibson and Weir 2005; McGraw *et al.* 2011), and integrative approaches combining genomic and gene expression data can have higher power to capture the true pathway associations underlying human diseases and complex traits (Xiong *et al.* 2012). In addition, recent developments of high-throughput phenotyping platforms have made the collection of thousands to millions of physiological measurements affordable to breeders (Araus *et al.* 2018). For example, images collected through thermal and hyperspectral cameras are used to increase the accuracy in genomic prediction for grain yield in wheat (Rutkoski *et al.* 2016). To further improve genomic prediction and to understand the underlying genetic mechanism, statistical models that enable the joint analysis of high-dimensional traits are required to establish the connection between phenomics and genomics.

Genomic analyses of high-dimensional, highly correlated data present analytic and computational challenges. The multivariate linear mixed model (MvLMM) is a widely-used statistical model for the genetic analyses of two or more correlated traits (Henderson and Quaas 1976). However, most algorithms used to fit MvLMMs require repeated inversions of genetic and residual covariance matrices among all traits, with a computational burden that grows cubically to quintically as the number of traits increases (Zhou and Stephens 2014). MvLMMs are also susceptible to over-fitting unless sample sizes are very large. Re-parameterizing MvLMMs as Bayesian sparse factor models can alleviate much of this computational burden (Runcie and Mukherjee 2013; Runcie *et al.* 2021) and can significantly improve the accuracy of genomic prediction (Runcie *et al.* 2021). For example, `BSFG` and `MegaLMM` are based on the assumption that the covariances among large sets of traits can be explained by a small set of latent factors (e.g., through gene regulatory

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 6th May, 2022
[1]Corresponding author: Department of Plant Sciences, Department of Animal Science, University of California Davis, 1 Shield Ave., Davis, CA 95616. E-mail: deruncie@ucdavis.edu, qtlcheng@ucdavis.edu.

networks), which is consistent with the discovery that variation in gene expressions of human diseases are mainly regulated by a few major disease-associated pathways (Xiong *et al.* 2012).

While `MegaLMM` addressed the statistical and computational challenges of applying MvLMMs to high-dimensional phenotypes, it permits a limited range of models for high-dimensional genotype data. Specifically, `MegaLMM` incorporates genomic data through one (or more) genomic relationship matrices, which imposes specific assumptions about the distribution and effect sizes of the underlying genetic variants, and does not allow direct inference on the identities of causal loci. Whole-genome regression methods such as the Bayesian Alphabet methods (Meuwissen *et al.* 2001; Park and Casella 2008; Kizilkaya *et al.* 2010; Habier *et al.* 2011; Cheng *et al.* 2015; Erbe *et al.* 2012; Cheng *et al.* 2018b), on the other hand, encode a wide range of different and more flexible distributions on the effect sizes of causal genomic loci and allow for inference of the causal loci themselves. However, fitting Bayesian Alphabet methods to very large numbers of markers can also be computationally demanding even for a single trait, and extensions of these methods to multivariate traits are very limited.

In this paper, we incorporate whole-genome regression approaches into a Bayesian sparse factor model named *MegaBayesianAlphabet* to incorporate thousands of traits for genome-wide prediction and association studies. The Bayesian Alphabet methods with mixture priors on marker effects (Kizilkaya *et al.* 2010; Habier *et al.* 2011; Moser *et al.* 2015; Wolc *et al.* 2016; Mehrban *et al.* 2017; Wang *et al.* 2020) are popular genetic models due to their incorporation of biologically meaningful assumptions and the variable selection procedure performed during model fitting. We focus on BayesC as an example of a Bayesian Alphabet method (Kizilkaya *et al.* 2010; Habier *et al.* 2011; Cheng *et al.* 2018b), but extensions of MegaBayesianAlphabet with other priors should be straightforward. We show that MegaBayesianAlphabet with BayesC prior (hereinafter referred to as MegaBayesC) can improve genomic prediction accuracy relative to multi-trait GBLUP and RR-BLUP methods by leveraging mixture priors on marker effects and information from thousands of traits. In association studies with millions of markers, MegaBayesianAlphabet is still computationally demanding, but we propose a two-step approach that can accurately estimate marker effects and improve power for association inference in both simulated and real data studies. MegaBayesianAlphabet is implemented in an R package called "MegaLMM".

Materials and Methods

In a conventional MvLMM, the genetic and non-genetic correlations among t traits are modeled through one or more $t \times t$ genetic covariance matrices ($\mathbf{G}_m$) and a $t \times t$ residual covariance matrix ($\mathbf{R}$), respectively. The computational cost of fitting a MvLMM can be prohibitive when t is large due to the difficulty in taking inverses of the covariance matrices (Gilmour *et al.* 1995; Yang *et al.* 2011; Zhou and Stephens 2014). To overcome the computational challenge and overfitting in conventional MvLMMs, we reparameterized the conventional MvLMM as a factor model (i.e., MegaLMM (Runcie *et al.* 2021) and MegaBayesianAlphabet), where K independent (unobserved) latent factors are introduced to account for the covariances among the t traits.

Model Description

In MegaBayesianAlphabet, the variation among t observed traits is decomposed into two parts: the variation caused by dependencies on K independent latent factors, which induces correlations among the t observed traits, and the variation that is unique, or idiosyncratic, to each trait. In MegaBayesianAlphabet, genetic values of latent factors are defined as a linear combination of all marker effects, and priors from the Bayesian Alphabet methods (Meuwissen *et al.* 2001; Park and Casella 2008; Kizilkaya *et al.* 2010; Habier *et al.* 2011; Cheng *et al.* 2015; Erbe *et al.* 2012; Cheng *et al.* 2018b) are assigned to the marker effects. The model specification of MegaBayesianAlphabet is described below.

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\Lambda} + \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_{2R}\mathbf{B}_{2R} + \mathbf{E}_R \quad (1)$$

with

$$\mathbf{F} = \mathbf{X}_{2F}\mathbf{B}_{2F} + \mathbf{E}_F \quad (2)$$

where $\mathbf{Y}$ is an $n \times t$ matrix of observations for n individuals on t traits, $\mathbf{F}$ is an $n \times K$ matrix of latent factors for n individuals across K latent factors, and $\boldsymbol{\Lambda}$ is a $K \times t$ factor loading matrix whose elements, such as λ_{kj}, describe the corresponding factor-trait relationships (e.g., the relationship between factor k and trait j). The K latent factors in $\mathbf{F}$ are further decomposed into genetic effects (i.e., $\mathbf{X}_{2F}\mathbf{B}_{2F}$) and residual effects (i.e., $\mathbf{E}_F$) as shown in Equation 2. The genetic effects of latent factors are expressed as multiple regressions on genotype covariates, where $\mathbf{X}_{2F}$ is an $n \times b_{2F}$ matrix of genotype covariates, and $\mathbf{B}_{2F}$ is a $b_{2F} \times K$ matrix of marker effects for the K latent factors at b_{2F} genotyped loci. $\mathbf{X}_1$ is an $n \times b_1$ incidence matrix allocating the observations on t traits to b_1 fixed effects with coefficient matrix $\mathbf{B}_1$. The residuals are similarly decomposed into trait-specific genetic effects (i.e., $\mathbf{X}_{2R}\mathbf{B}_{2R}$) and trait-specific residual effects (i.e., $\mathbf{E}_R$), with $\mathbf{B}_{2R}$ being a $b_{2R} \times t$ matrix of marker effects corresponding to the t traits at b_{2R} genotyped loci.

If all sources of correlation among observed traits are explained by the latent factors, the residuals conditional on these factors become uncorrelated between different traits. Since the sources of correlation among observed traits are explained by independent latent factors, samples at each iteration of Markov chain Monte Carlo (MCMC) can be obtained simultaneously in parallel across traits and factors, which leads to significant reduction in the computational cost of model fitting.

Prior Specification

Genetic Marker Effects Mixture priors are widely used for genetic marker effects in Bayesian regression methods in genome-enabled analysis. In this paper, the BayesC prior is used for the marker effects (e.g., coefficients in $\mathbf{B}_{2F}$) and we term this specific version of MegaBayesianAlphabet: MegaBayesC. The BayesC mixture prior assumes that marker effects are independently and identically distributed, each of which has a point mass at zero with a marker exclusion probability π, and follows a univariate normal distribution with a marker inclusion probability $1 - \pi$. For example, the prior distribution of the marker effect at locus i for the kth latent factor is shown as follows.

$$b_{2F_{k(i)}} = \begin{cases} N(0, \sigma^2_{B2F_k}) & \text{probability } (1 - \pi_{F_k}) \\ 0 & \text{probability } (\pi_{F_k}) \end{cases}$$

where $\sigma^2_{B2F_k}$ is the variance of marker effects corresponding to factor k. Due to the independence among latent factors, and the independence among traits conditional on $\mathbf{F}\boldsymbol{\Lambda}$, marker effects can be efficiently sampled from a set of univariate BayesC

models in parallel across traits and factors at each iteration of MCMC. We treat each marker exclusion probability for the K latent factors (e.g., π_{F_k}) and the t observed traits as an independent unknown parameter to be estimated. Note that if marker inclusion probabilities for all factors were set to 1.0 (i.e., all markers are included with equal variance), the model is equivalent to RR-BLUP, and we term this specific version of MegaBayesianAlphabet: MegaRRBLUP.

Factor Loading Matrix The factor loading matrix ($\mathbf{\Lambda}$) describes the relationship between latent factors and observed traits. Sparsity in this matrix implies that factors affect some, but not all traits, a key assumption in Bayesian sparse factor models (Carvalho *et al.* 2008). We use a BayesC mixture prior for the elements of $\mathbf{\Lambda}$. Because factor swaps do not change the likelihood, to improve the identifiablilty of the model, we introduce an additional parameter to the included-variable variance (τ_k^{-1}) that is stochastically decreasing across factors (Bhattacharya and Dunson 2011; Runcie and Mukherjee 2013; Runcie *et al.* 2021). For the factor loading that describes the relationship between factor k and trait j (i.e., λ_{kj}), its prior distribution is shown as follows.

$$\lambda_{kj} = \begin{Bmatrix} N(0, \tau_k^{-1}\sigma^2_{R_j}) & \text{probability } (1 - \pi_{\Lambda_k}) \\ 0 & \text{probability } (\pi_{\Lambda_k}) \end{Bmatrix} \quad (3)$$

$$\begin{aligned} \tau_k &= \prod_{h=1}^{k} \delta_h \\ \delta_1 &= 1 \\ \delta_h &\sim \text{Gamma}(a_\delta, b_\delta), h = 2...k \\ \sigma^2_{R_j} &\sim \text{Inv-Gamma}(a_\sigma, b_\sigma) \end{aligned}$$

Through the prior specification of $\mathbf{\Lambda}$, an appropriate level of truncation on the rows of $\mathbf{\Lambda}$ is able to ensure that the contribution from additional factors beyond the truncation point is negligible (Bhattacharya and Dunson 2011).

Other priors All other prior distributions are the same as used in Runcie *et al.* (2021).

Posterior Distributions for Gibbs Sampler

We use MCMC method to sample from the posterior distributions of all parameters. The full conditional posterior distributions for Gibbs sampler are derived for all the parameters in MegaBayesC in **Appendix**.

Estimation of Genetic Values for Genomic Prediction

We assessed the performance of MegaBayesC as a tool for genomic prediction using hyperspectral data as additional traits to assist wheat yield prediction.

Data Description Best linear unbiased estimators (BLUEs) of grain yield and reflectances from 62 wavelength bands collected with an areal hyperspectral camera on each of 10 time-points during the growing season for 1033 bread wheat lines were downloaded from CIMMYT Research Data (Krause *et al.* 2019). We analyzed results from the 2014-2015 breeding cycle under the Optimal Flat treatment. All lines were genotyped using the pipeline described in Poland *et al.* (2012). Markers with call rate $\leq$ 50% and minor allele frequency (MAF) $\leq$ 0.05 were removed. Missing genotypes were imputed by corresponding marker means. In our analysis, the 620 hyperspectral BLUEs were used as secondary traits (Runcie and Cheng 2019) to improve the prediction of the genetic value of grain yield, which is served as a focal trait in our prediction scenario. Both sets of traits were combined into a 1033×621 trait matrix $\mathbf{Y}$.

Models Four different models were used to predict the grain yield (GY): GBLUP, MegaGBLUP, MegaRRBLUP, and MegaBayesC. Posterior means were used as point estimates of parameters of interest. These four models are described below.

GBLUP A conventional single-trait GBLUP model (VanRaden 2008) with a variance-covariance matrix proportional to a genomic relationship matrix $\mathbf{K}$ fitted to the grain yield BLUEs, ignoring the hyperspectral data.

MegaGBLUP This model was described in Runcie *et al.* (2021). The fixed effects $\mathbf{B}_1$ included intercepts only. The random effects $\mathbf{B}_{2R}$ and $\mathbf{B}_{2F}$ were not included in the model. A random effect with covariance proportional to $\mathbf{K}$ was included in Equations 1 and 2 to model the genetic relationships among lines.

MegaBayesC The estimated individual genetic merits of grain yield in MegaBayesC were computed as: $\mathbf{u}_{GY} = \mathbf{X}_{2F}\hat{\mathbf{B}}_{2F}\hat{\boldsymbol{\lambda}}_1$, where $\hat{\boldsymbol{\lambda}}_1$ denotes the first column of $\hat{\mathbf{\Lambda}}$, which specifies the estimated relationship between all factors and grain yield. $\mathbf{B}_1$ included only an intercept, and $\mathbf{B}_{2R}$ was not included. We included one factor having non-zero effects only on grain yield, i.e., $\boldsymbol{\lambda}_{GY}^T = [1 \quad 0 \quad 0 \quad ... \quad 0]_{1\times t}$, to model direct genetic effects on grain yield. For the remaining factors, the probability of a element from $\mathbf{\Lambda}$ being zero was considered as known and set to be 0.9 to introduce sparsity to $\mathbf{\Lambda}$ and to shorten the time required for its convergence, while the probability of a marker having a null effect on a latent factor was considered as unknown and was estimated. $K = 100$ factors were fitted in MegaBayesC.

MegaRRBLUP This model mimics the priors for marker effects in RR-BLUP. The only difference between MegaRRBLUP and MegaBayesC lies in the prior distributions of marker effects on latent factors. Normal distributions instead of mixture priors are used for marker effects in MegaRRBLUP, indicating that all markers are included in the model. This model should be identical to **MegaGBLUP** except that the prior on elements of $\mathbf{\Lambda}$ is BayesC instead of the Bayesian Horseshoe.

Cross Validation We used cross-validation to evaluate the predictive performance of different models by masking the grain yield of 516 randomly selected lines (around 50% of the population) of the population during model fitting and comparing the masked values to model predictions. Since we did not mask the hyperspectral data from these 516 model validation individuals, but used those data to enhance our genetic value predictions, using the Pearson's correlation between predicted and observed GY values could lead to biased and sub-optimal choices of models (Runcie and Cheng 2019). Instead, we used the estimated genetic correlation corrected by grain yield heritability to estimate the prediction accuracy (Runcie and Cheng 2019; Daetwyler *et al.* 2013) as implemented in Runcie *et al.* (2021). The cross-validation process was repeated 20 times with different masked lines.

Estimation of Marker Effects for Association Inference

In MegaBayesianAlphabet, covariances among high-dimensional phenotypic data are decomposed into K sources

of variation, each of which controls the correlation among a subset of observed traits through the factor loading matrix. In this way, information of correlated traits is used jointly to estimate their underlying pathways (i.e., latent factors), while the computational burden to analyze large-scale phenotypic data is significantly decreased. With the assistance of large-scale genetically correlated traits, MegaBayesianAlphabet is expected to boost the discovery of genetic variants associated with a trait of most interest (i.e., focal trait) and precisely quantify their effect sizes.

In this section, two simulation studies and one real data analysis were conducted to investigate the accuracy of the estimation of marker effects by MegaBayesC. First, a population with independent and uncorrelated SNPs was simulated to demonstrate the ability of MegaBayesC to distinguish the genetic and non-genetic sources of variation in a focal trait, utilizing the information of correlated traits. Second, a simulation study based on a real Arabidopsis population was conducted to study the effects of population structure and linkage among markers. Finally, the real phenotypes of flowering time from this Arabidopsis population was studied, utilizing expression data from 20,843 genes.

Simulated Study in A Population without structure or Linkage Disequilibrium We created a simulated population of $n = 5000$ individuals and $p = 2,000$ SNPs. An $n \times p$ matrix of genotypic covariates was generated by random sampling from $\{0, 1, 2\}$. We then created simulated phenotypic data for a single focal trait and many correlated "secondary" traits. The performance of MegaBayesC was compared to a single-trait BayesC model (ST-BayesC) based on the accuracy of estimated marker effects for the focal trait. We induced genetic and non-genetic covariation among the traits through latent factors. The majority of variance in the focal trait was attributed to the latent factors. In Scenario 1, we created latent factors whose variation was primarily determined by the genetic markers (i.e. high-heritability latent factors), and in Scenario 2, the latent factors were predominantly non-genetic (i.e. low-heritability factors).

We studied four parameters that we expected to influence the relative performance of MegaBayesC and ST-BayesC. They are 1) the number of latent factors (n_{factor}), 2) the number of correlated traits ($n_{trait/factor}$) controlled by each factor, 3) the number of QTL (i.e., causal variants) that control each factor ($n_{qtl/factor}$), and 4) the heritability of the factors. In this simulation study, $n_{factor} = \{2, 6, 9\}$, $n_{trait/factor} = \{2, 20\}$, $n_{qtl/factor} = \{10, 20, 30\}$, and two heritable patterns of latent factors were considered.

To generate the simulated phenotype data, we first used n_{factor} and $n_{trait/factor}$ to construct a factor loading matrix ($\mathbf{\Lambda}$). For example, when $n_{trait/factor} = 2$ and $n_{factor} = 2$, 4 (i.e., $n_{factor} \times n_{trait/factor}$) observed traits were simulated, with two different observed traits linked to each factor. Since the first observed trait was treated as the focal trait, and all factors were assumed to contribute to its variation, factor loadings in the first column of $\mathbf{\Lambda}$ were set to 1. To minimize the complexity of this simulation, non-zero factor loadings in $\mathbf{\Lambda}$ were set to be equal to 1. Therefore, the simulated $\mathbf{\Lambda}$ given $n_{trait/factor} = 2$ and $n_{factor} = 2$ was expressed as:

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Based on the constructed $\mathbf{\Lambda}$, all factors except the first factor were linked to $n_{trait/factor} + 1 = 3$ observed traits, while the first factor was linked to $n_{trait/factor} = 2$ observed traits. A similarly structured $\mathbf{\Lambda}$ was constructed for other combinations of $n_{trait/factor}$ and n_{factor}.

After defining $\mathbf{\Lambda}$, genetic variation controlled by selected QTL and non-genetic variation in each factor were simulated. $n_{qtl/factor}$ QTL were selected for each factor, and variation was simulated such that the variance explained by these QTL was a defined percentage of the total variation in the factor. In Scenario 1, the QTL accounted for 95% of the variance of each factor (i.e., $\sigma^2_{FG}/(\sigma^2_{FG} + \sigma^2_{FE}) = 0.95$ with σ^2_{FG} being the genetic variance of factors and σ^2_{FE} being the residual variance of factors). In Scenario 2, only the first factor was associated with QTL (again with 95% of its variance explained by the QTL), and the remaining factors had independent variation. Finally, additional trait-specific variance was added to each trait, accounting for approximately 10% of its total variance.

As a consequence of these simulation choices, the two scenarios differed in several key aspects of the genetic architecture and correlation structures between the focal trait and the secondary traits. In Scenario 1, all factors were controlled predominately by genetic variation and all QTL for every factor was therefore a QTL for the focal trait. Therefore, all secondary traits had strong genetic correlations with the focal trait. In Scenario 2, most factors were controlled by non-genetic variation; only the first factor was controlled by QTL. Therefore while all secondary traits were phenotypically correlated with the focal trait, most of these correlations were non-genetic.

In both scenarios, as n_factor and/or $n_{qtl/factor}$ increased, the magnitude of variation explained by each QTL decreased to hold the total percentage of variation in the focal trait explained by QTL constant. In Scenario 1, when $n_{factor} = 9$ and $n_{qtl/factor} = 10$, the 90 QTL each explained $\approx 0.97\%$ of the total variance (Figure 3). As $n_{qtl/factor}$ increased to 30, the number of QTL for the focal trait increased to 270 and each accounted for around 0.29% of the total variance of focal trait. In Scenario 2, when $n_{factor} = 9$, the variance explained by each marker decreased from 0.90% to 0.31% as the number of QTL increased from 10 to 30. For a given n_{factor} and $n_{qtl/factor}$ the per-QTL effect sizes were comparable, but since there were more factors with QTL in Scenario 1, the total variance in the focal trait controlled by all QTL was larger.

In Scenario 2, as n_{factor} increased, the proportion of variance explained by QTL decreased. For example, when $n_{factor} = 6$, the genetic variance accounted for 14% of the total variance of focal trait. With $n_{factor} = 9$, the percent of variance explained by genetic markers decreased to 9%. In Scenario 1, the percentage of variance explained by QTL was constant across values of n_{factor}. In this scenario, all QTL for all factors contributed to the variation in the focal trait. For example, when $n_{factor} = 6$ and $n_{qtl/factor} = 10$, each factor was influenced by 10 QTL, which were randomly selected from all SNPs, leading to a total of 60 QTL selected. During this process, some SNPs may be stochastically selected more than once, and thus, some QTL may have effects on more than one factor.

Based on the combination of $n_{trait/factor}$, n_{factor}, $n_{qtl/factor}$, and the heritable patterns, a total of $3 \times 3 \times 2 \times 2$ conditions were studied in this simulation study. 10 replicates were conducted for each of the 36 conditions.

When fitting models to these simulated data, we included the intercept for each trait as the only fixed effect. The model specifi-

cation of MegaBayesC was similar to that used in the **Genomic Prediction** application, except: 1) no fixed factor loadings were included in $\mathbf{\Lambda}$, 2) the probability of a element from $\mathbf{\Lambda}$ being zero was considered as unknown and was estimated, 3) the number of factors fitted in the model was $K = 10$. We estimated the total marker effects on the focal trait obtained by MegaBayesC as: $\boldsymbol{\alpha}_f = \mathbf{B}_{2F}\boldsymbol{\lambda}_1$, where $\mathbf{B}_{2F}$ is the matrix of marker effects of latent factors and $\boldsymbol{\lambda}_1$ denotes the first column of $\mathbf{\Lambda}$ specifying the relationship between factors and focal trait, i.e., summing up the QTL effects on the latent factors weighted by the relationships between each factor and the focal trait. We measured the performance of each method (i.e., MegaBayesC and ST-BayesC) by calculating the square root of the mean square error (RMSE) of estimated marker effects.

Simulation Study in a Real Arabidopsis Population We created a second set of simulated datasets based on real genotypes from 1003 Arabidopsis thaliana accessions. Genotype data were downloaded from the 1001 genomes project (Alonso-Blanco *et al.* 2016). In a real population, the presence of linkage disequilibrium (LD) between loci and variable allele frequencies among markers increase the complexity of genetic association analyses. We removed SNPs with MAF ≤ 0.05 and missing genotype rate ≥ 0.1 using PLINK 1.9 (Purcell *et al.* 2007), leaving 802,427 variants used for downstream analysis.

To ensure the QTL were independent, we pruned SNPs with an LD threshold of 0.8 in windows of 500 SNPs, using a sliding window of 100 SNPs. We randomly selected 20 QTL from these SNPs, and generated 10 latent factors, each was affected by 2 different QTL. In this simulation, the structure of the variance of the focal trait was simplified. All genetic variance in all traits was driven by the QTL effects on the latent factors, while all non-genetic variance was trait-specific. In this way, the observed traits ($\mathbf{Y}$) was expressed as: $\mathbf{Y} = \mathbf{X}_2\mathbf{B}_{2F}\mathbf{\Lambda} + \mathbf{E}_R$.

We set each element of the first column of $\mathbf{\Lambda}$ to 0.5 so that all 10 of the factors contributed equally to the focal trait. Each factor was additionally linked to 20 different secondary traits with factor loadings equal to 1. Other elements in $\mathbf{\Lambda}$ were set to be 0. Therefore, a total of 201 traits were simulated.

The proportion of genetic variance in the focal trait was set to be around 60% (i.e., $h^2_{focal} = 0.6$). To ensure that the variance explained by each QTL was consistent ($\approx 1 - 5\%$ of the total variance), QTL effects were sampled from a uniform distribution $U(3, 5)$, and a randomly chosen half of those effects were multiplied by -1. In addition, since the heritability of secondary traits such as gene expression is often higher than that of focal trait in real data applications, the heritabilities of the 200 secondary traits were each set to be 0.8.

Finally, to parallel our real data analysis below, secondary trait data was simulated for only 649 of the 1003 Arabidopsis accessions. The 354 remaining accessions had the records for only the focal trait.

After creating the simulated data, we applied three methods to identify QTL and estimate their effects on the focal trait: 1) single-trait Genome-Wide Association Studies (GWAS) using GCTA (Yang *et al.* 2011) (ST-GCTA); 2) single-trait BayesC implemented in JWAS (Cheng *et al.* 2018a) (ST-BayesC); and 3) MegaBayesC implemented in MegaLMM. Since whole-genome regression models with hundreds of thousands of candidate markers are computationally prohibitive, a two-stage analysis was implemented for ST-BayesC and MegaBayesC. In the first stage (i.e., the pre-selection stage), we selected a small proportion of SNPs to take forward into a full BayesC analysis by running a single-trait GWAS using GCTA on only the 354 individuals without records on secondary traits. After running the GWAS, we used LD-based clumping to select ≈ 2000 potentially important SNPs. First, we sorted SNPs by p-value, removed SNPs with p-values larger than 0.01, then used a greedy algorithm to select the most-significant SNPs and mask all nearby SNPs (within 250Kb) with $r^2 > 0.5$ (Purcell *et al.* 2007).

After the pre-selection stage, the records of focal and secondary traits from the remaining 649 individuals, which are considered as an independent population, were analysed in MegaBayesC using only the pre-selected potentially important SNPs. The model specification of MegaBayesC was similar to that used in the previous simulation study for independent population, except we set $K = 30$. In MegaBayesC, the total marker effects of the focal trait were computed as: $\boldsymbol{\alpha}_f = \mathbf{B}_{2F}\boldsymbol{\lambda}_f$, where $\mathbf{B}_{2F}$ is a $b_{2F} \times K$ matrix of marker effects for latent factors, with b_{2F} being the number of SNPs selected at the pre-selection stage, and $\boldsymbol{\lambda}_f$ denotes the column of $\mathbf{\Lambda}$ that specifies the relationship between factors and focal trait. Furthermore, to demonstrate that the improved performance of MegaBayesC is attributed to not only the use of the BayesC prior on the marker effects but also the utilization of information from correlated secondary traits, a ST-BayesC was also performed for the 649 individuals at the second stage. MCMC chains of 50,000 iterations were run for the BayesC-based methods with the first 10,000 iterations discarded as burn-in.

In addition to the two-stage analysis, a one-stage ST-GCTA was performed using the whole-genome SNP information and the phenotypes of the focal trait from all 1003 individuals. To compare with the two-stage analysis, the selection of potentially important SNPs was done based on the one-stage ST-GCTA result in the same manner as that in the pre-selection stage.

Figure 1 shows the procedures performed to estimate marker effects in the three different methods. Simulations were repeated 100 times.

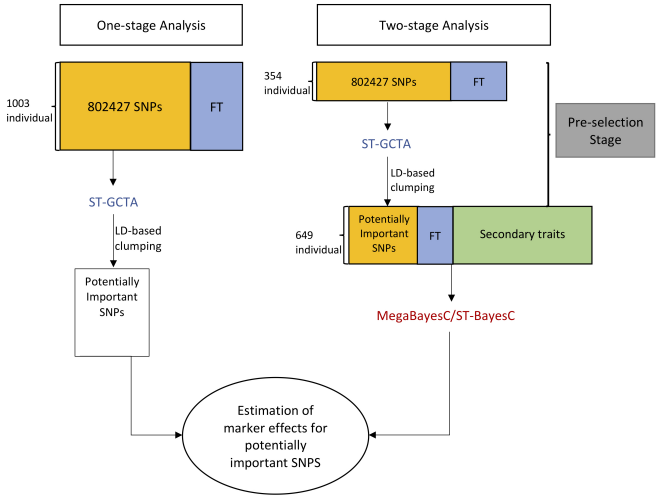

Figure 1 Graphic representation for the procedure of one-stage and two-stage analyses performed for the estimation of marker effects. FT represents the focal trait, ST-GCTA represents single-trait GWAS implemented in GCTA, and ST-BayesC represents single-trait BayesC method. In ST-BayesC, only phenotypes of FT and genotypes of the pre-selected potentially important SNPs were used.

RMSEs of the QTL effect sizes and the percentage of variance in the focal trait explained by the QTL were used to evaluate the accuracy of estimation of QTL effects by different methods. The variance explained by marker l was computed as: $var(\alpha_{lf}\mathbf{x}_2)$, where α_{lf} is the marker effect of SNP l on focal trait, and $\mathbf{x}_2$ is the vector of genotypic covariates for SNP l. To score QTL accuracy, we parsed the detected QTL in three ways: 1) If the true QTL were selected in the set of potentially important SNPs (e.g., Stage 1), the estimated effects were compared directly to the true effects. 2) If a SNP in imperfect LD with the true QTL was selected instead of the true QTL, we flagged its estimated effect size in the accuracy comparison because the incomplete linkage and different allele frequencies of the two SNPs mean that the estimated effect size will not be directly comparable to that of the true QTL. However, the percentage of variance attributed to the marker should be similar to the true QTL as long as r^2 is high; 3) If neither the true QTL nor any of its linked SNPs was selected in the potentially important SNPs, we set the estimated marker effect and variance explained by this QTL to 0. For the purpose of unit consistency, RMSE of estimated marker effects and estimated marker-explained standard deviation (i.e., square root of marker-explained variance) across different methods were compared. In this study, a SNP was considered to be linked to a QTL if the squared correlation between its genotypic covariate and the QTL genotype was greater than 0.4.

Genetic Association Analysis of Arabidopsis thaliana Flowering Time and Gene Expression Phenotypes of flowering time from 1003 accessions and expression data of 20843 genes from 649 accessions were used, with flowering time selected as our focal trait. Gene expression data was downloaded from NCBI GEO (Barrett *et al.* 2012). Genes with average counts smaller than 10 were removed and the remaining gene counts were normalized and variance stabilized as per Runcie *et al.* (2021) using DESeq2 (Love *et al.* 2014). The two-stage MegaBayesC and one-stage ST-GCTA analyses described above were performed again on this dataset. LD-based clumping was done to select potentially important SNPs for both methods. The model specification of MegaBayesC was similar to that used in the **Genomic Prediction** section above. A MCMC chain of 80,000 was run with the first 20,000 iterations discarded as burn-in. In the two-stage MegaBayesC analysis, potentially important SNPs with explained proportion of variance $> 0.1\%$ were classified as significant SNPs, while in the one-stage ST-GCTA analysis, potentially important SNPs with p-value $< 1 \times 10^{-5}$ were classified as significant SNPs. We compared each significant SNP to a list of genes previously known to influence flowering time in Arabidopsis (Bouché *et al.* 2016), and counted as a match (i.e., a true positive hit) if a SNP was within +/- 100 Kb distance from at least one of the reported genes. Otherwise the significant SNP was conservatively considered as a false positive association.

Data availability

Scripts for running all analyses are archived at GitHub: https://github.com/Jiayi-Qu/Mega-BayesC. The Bayesian Alphabet implementation is available on the "BayesAlphabet" branch of the MegaLMM GitHub repository: https://github.com/deruncie/MegaLMM/tree/BayesAlphabet. Data from the wheat breeding trial were downloaded from CIMMYT Research Data (Krause *et al.* 2019). Arabidopsis flowering time data was downloaded from Arapheno: https://arapheno.1001genomes.org/phenotype/261/. Gene expression data was downloaded from NCBI GEO (Barrett *et al.* 2012). Genotype data were downloaded from the 1001 genomes project (Alonso-Blanco *et al.* 2016).

Results

MegaBayesC Improves Estimation of Genetic Values

We tested if MegaBayesianAlphabet models could match or exceed the performance of MegaLMM in trait-assisted genomic prediction using data from a breeding trial of bread wheat. We compared the genomic value prediction accuracy of MegaBayesC and MegaRRBLUP to MegaGBLUP in this dataset, where we leveraged 620 hyperspectral phenotypes measured on 1033 bread wheat lines to supplement genotype-based predictions of genomic value for grain yield. As a baseline, we performed conventional univariate GBLUP-based genomic value prediction as well. Prediction accuracy was assessed by cross-validation where for each of 20 replicates, grain yield values of 50% of the lines were masked and used as an independent testing set. Estimated genetic correlations between predicted and observed yields in the testing set were used as the cross-validation statistic.

As shown in Figure 2, univariate GBLUP achieved a prediction accuracy of 0.43 in this dataset. MegaGBLUP fitted to all traits in MegaLMM with a single random effect based on the genomic relationship matrix $\mathbf{K}$ achieved an average prediction accuracy of 0.69. MegaRRBLUP fitted in MegaBayesianAlphabet achieved an average prediction accuracy of 0.68. No significant difference was observed between MegaGBLUP and MegaRRBLUP. RR-BLUP and GBLUP are mathematically equivalent (Whittaker *et al.* 2000; Meuwissen *et al.* 2001; Habier *et al.* 2007) models that account for the contributions of the genetic markers, but MegaGBLUP uses a horseshoe prior for the elements of $\mathbf{\Lambda}$ while MegaRRBLUP uses the BayesC prior for these parameters with fixed $\pi = 0.9$. MegaBayesC, with its BayesC prior on the marker effects, achieved an average accuracy of 0.75, significantly higher than the other methods. These results show that the use of biologically meaningful prior on marker effects can further improve the genomic selection in breeding programs.

MegaBayesC improves Estimation of Marker Effects in Simulated Populations with Independent Markers

Next, we ran a set of simulations to evaluate the ability of MegaBayesC to identify and accurately estimate the effect sizes of genetic variants for a set of correlated traits under different genetic architectures. Specifically, we tested whether MegaBayesC improved the estimation of variant effect sizes of a single focal trait when phenotypes of other correlated traits (i.e., secondary traits) were provided.

Since the magnitude and causes (genetic vs. non-genetic) of the covariance structures among traits determine the usefulness of the secondary traits, we considered two covariance structures. In both cases, we began by simulating a set of latent factors partially controlled by genetic variation. In Scenario 1, the majority of variation in the focal trait was controlled by latent factors that were dominated by genetic variation. In Scenario 2, the majority of variation in the focal trait was controlled by latent factors dominated by non-genetic sources of variation. We compared the estimation of marker effects between ST-BayesC (which ignored all secondary traits) and MegaBayesC (which used all trait data at once). We scored the accuracy of each method by the RMSE of estimated marker effects. In both scenarios, as the genetic architecture increased in complexity (i.e., the number of QTL increased and the average size of each QTL decreased to

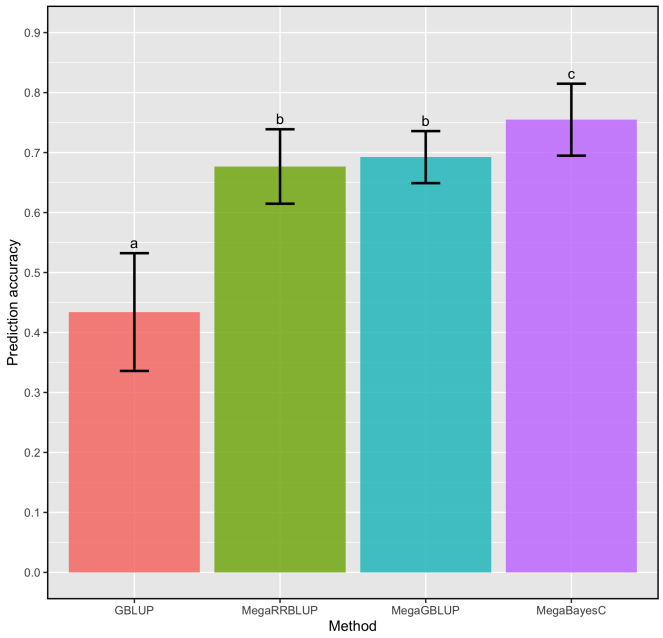

Figure 2 Genomic value prediction performance of 4 models for wheat yield. Records of yield, 620 hyperspectral phenotypes, and genotype data for 1033 lines were available. 20 replicate validations were used. Bars show the mean prediction accuracy (± standard error) for each model, and letters show the statistical significance of mean difference between methods based on a paired t-test.

keep the total percentage of variation attributable to the QTL constant), the performance of ST-BayesC decreased (RMSE increased) much more dramatically than MegaBayesC. Figure 3 shows RMSE of estimated effects for QTL and SNP, respectively, (*i.e.* QTL are markers with a non-zero effect and SNPs are markers with a true effect size of zero) under the two scenarios for the simulation setting where the largest difference of RMSE was observed between MegaBayesC and ST-BayesC, with $n_{trait/factor}$ = 2 and n_{factor} = 9. Results for other combinations of n_{factor}, $n_{trait/factor}$, and $n_{qtl/factor}$ are shown in **Appendix** (Figure 9). In Scenario 1, the number of latent factors had no direct effect on the performance of ST-BayesC beyond its effect on the number of QTL. Also, the number of traits linked to each factor (i.e. $n_{trait/factor}$) did not significantly affect the performance of MegaBayesC in both Scenario 1 and Scenario 2. This shows the ability of MegaBayesC to capture the underlying sources of correlations among traits by optimizing the utilization of secondary traits, even when each factor only has one linked secondary trait included in the model.

For ST-BayesC, the RMSE of estimated marker effects increased significantly as marker-explained variances decreased in both scenarios. Compared to Scenario 1, the increase of RMSE for estimated effects of QTL was greater in Scenario 2, while the increase of RMSE for estimated effects of SNPs were similar between the two scenarios. This indicates that the performance of ST-BayesC to identify QTL was affected by the marker-explained variance as well as the variance structure of the focal trait.

In contrast, the performance of MegaBayesC was relatively constant across scenarios as measured by RMSE. In terms of the estimation of effect sizes of QTL, the influence of the variance

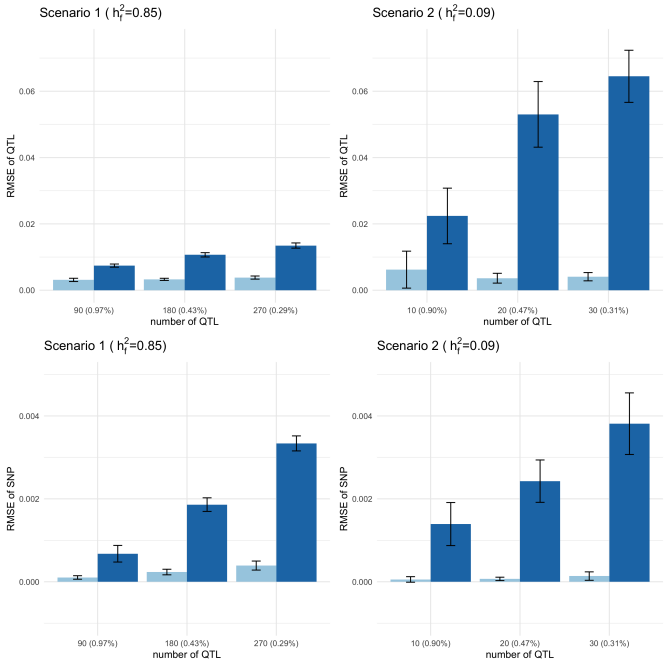

Figure 3 Root mean square error (RMSE) of estimated QTL effects and SNP effects, respectively, under two scenarios.The upper panels show RMSE of estimated QTL effects under two scenarios. The lower panels show RMSE of estimated SNP effects under two scenarios. The left panels show RMSE for Scenario 1, where all latent factors had high heritability ($h^2 = 0.95$). The right panels show RMSE for Scenario 2, where only one of the factors had high heritability (i.e., factor 1 had $h^2 = 0.95$ and the remainder factors had $h^2 = 0$). Results are shown for the simulation setting with $n_{trait/factor}$ = 2 and n_{factor} = 9. The average proportion of total variance explained by one QTL was shown in the parenthesis.

structure and the marker-explained variance was negligible, which lead to a relatively constant RMSE across the simulation settings. At the same time, MegaBayesC was able to shrink most SNPs more effectively towards zero, especially in Scenario 2, when the ratio of number of QTL to number of SNPs was smaller.

To further explore the differences in the performance of ST-BayesC and MegaBayesC in Scenario 2, we plotted the estimated marker effects under one example simulation with $n_{factor} = 9$, $n_{qtl/factor} = 30$, and $n_{trait} = 2$ (Figure 4).

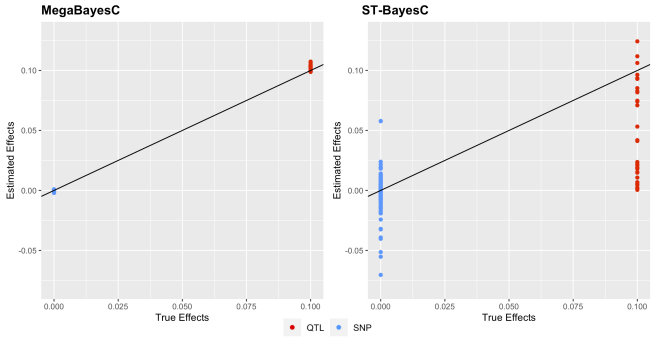

Figure 4 Scatter plot of estimated marker effects versus true marker effects for the simulation setting with $n_{factor} = 9$, $n_{qtl/factor} = 30$, and $n_{trait} = 2$ in Scenario 2, where all factors have effects on the focal trait but only one of them is a genetic factor (i.e., $h^2 > 0$). Red and blue colors specify QTL (effect size $\neq$ 0) and SNP (effect size = 0), respectively. The solid black line represents the line y = x.

For ST-BayesC, some QTL were successfully select by the model and their effect sizes were accurately estimated close to the true value of 0.1. However, for the majority of QTL, the estimated marker effects were shrunk toward 0s. On the other hand, ST-BayesC erroneously estimated effect sizes of SNPs with true effect sizes of 0 from -0.07 to 0.06. In contrast, the marker effects of QTL and null-effect SNPs were accurately estimated by MegaBayesC (Figure 4).

Estimation of Explained Variance of Markers in a Population Simulated Using Real Genotype Data

To explore the ability of MegaBayesC to accurately identify QTL and estimate their effect sizes in the presence of LD, we generated simulated phenotypes based on real genotypes from an Arabidopsis population. We then ran association analyses using three methods: The direct (*i.e.* one-stage) method, ST-GCTA, that only uses the focal trait, and two two-stage methods: ST-BayesC and MegaBayesC, which both rely on a pre-selection stage to select a set of candidate SNPs using one partition of the population, and then an assay stage where the effects of those SNPs on the focal trait are modeled in the second partition of the population. We compared the performance of the models by the RMSE of estimated marker effects and marker-explained variances.

Figure 5 shows the RMSE of estimated marker effects and estimated marker-explained standard deviations from the simulated phenotype data. The two-stage MegaBayesC method achieved the lowest RMSE for both marker effects and marker-explained standard deviations, followed by the two-stage analysis incorporating ST-BayesC, and then the one-stage single-trait GWAS (ST-GCTA). The RMSE of the one-stage single-trait GWAS

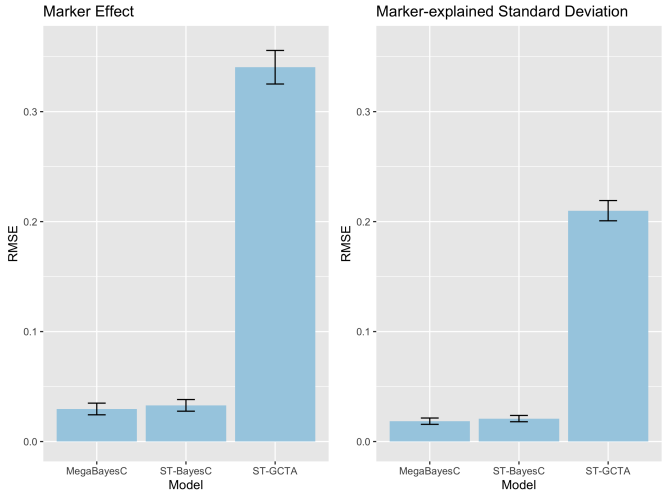

Figure 5 Root mean square error (RMSE) of estimated marker effects and estimated marker-explained standard deviations (i.e., square root of marker-explained variances) across different methods. The performance of two-stage (ST-BayesC and MegaBayesC) methods and one-stage (ST-GCTA) method were compared.

was around ten times larger than that of the two-stage BayesC-based analyses, while the difference between ST-BayesC and MegaBayesC was much smaller. Furthermore, the RMSE of estimated marker-explained standard deviations was generally lower than that of estimated marker effects. The larger RMSE of estimated marker effects is likely due to the selection of linked SNPs rather than the true causal QTL in the pre-selection stage.

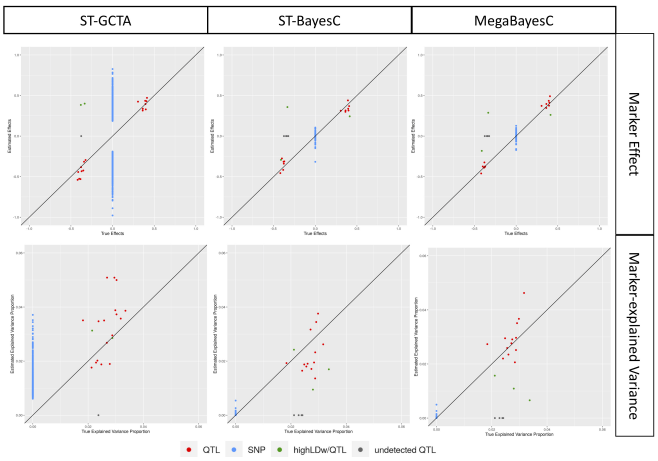

Figure 6 The relationship between estimated and true values of marker effects and marker-explained proportion of variance for focal trait. Three different methods (ST-GCTA, ST-BayesC, and MegaBayesC) were compared. Details of each method are presented in **Materials and Methods**.

To further explore the difference in the performance of ST-BayesC and MegaBayesC in this simulation scenario, we present the relationship between true and estimated marker effects for one replicate in Figure 6. In this simulation, 19/20 true causal QTL were selected by ST-GCTA, and only 16 were selected in the pre-selection stage for the two-stage methods, ST-BayesC

and MegaBayesC. In all these three cases, the effect sizes of these selected QTL were accurately estimated. However, the effect sizes of many null-effect SNPs were dramatically overestimated by ST-GCTA, leading to an overall high false positive rate. In contrast, although a few true causal QTL were missed in the pre-selection stage, SNPs with null effects that were moved forward into stage two were estimated to have very small effects by both ST-BayesC and MegaBayesC.

Note that in some cases, SNPs that are in LD with true QTL were selected instead of the causal QTL. When the linkage phase was negative, the estimated effect sizes for linked SNPs have the opposite sign, which increases the reported RMSE. However, even in these cases, the proportion of variance explained by these linked markers is close to the proportion that would have been explained by the true QTL, so the effect of LD on the RMSE of marker-explained variances is minimized.

Identifying Candidate Genes for Flowering Time in Arabidopsis using Gene Expression Data as Secondary Traits

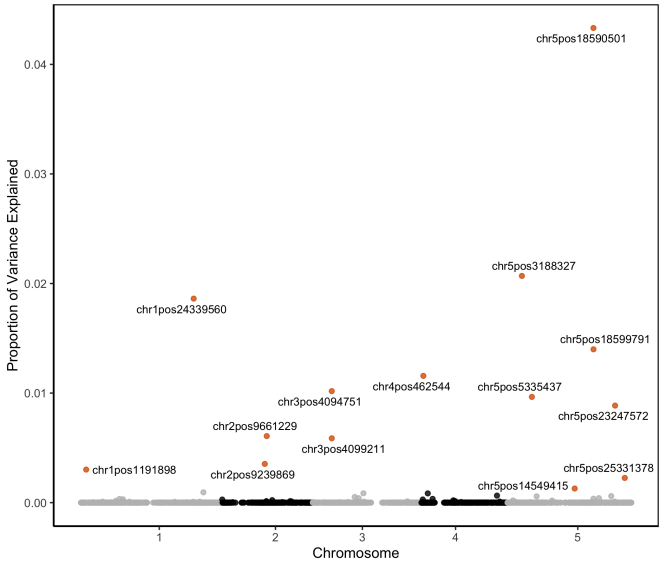

Figure 7 Marker-explained proportion of variance for potentially important SNPs by the two-stage analysis using MegaBayesC. The top 14 SNPs that explained the greatest proportions of variance in flowering time are highlighted.

We applied the two-stage MegaBayesC and the one-stage single-trait GWAS (ST-GCTA) to the task of identifying candidate genes that regulate flowering time in *Arabidopsis thaliana* using actual flowering time measurements and genotype data from 1003 *A. thaliana* accessions. In MegaBayesC, we included the expression of 20843 genes measured on 649 of the accessions as secondary traits.

Potentially important SNPs with marker-explained variance greater than 0.1% in MegaBayesC and potentially important SNPs with p-value smaller than 10^{-5} in ST-GCTA were selected as significant SNPs. MegaBayesC was better able to select a limited number of candidate SNPs based on per-marker variance explained (Figure 7) then ST-GCTA (Figure 8) by shrinking the vast majority of SNP effects close to zero.

We assessed the accuracy of these associations by checking whether known flowering time-related genes are located near to the SNPs selected by each model. Using MegaBayesC, we

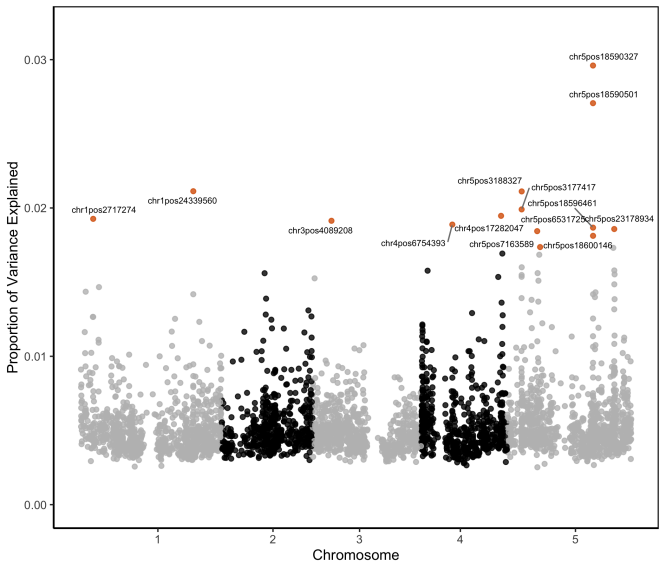

Figure 8 Marker-explained proportion of variance for potentially important SNPs by the one-stage ST-GCTA analysis. The top 14 SNPs that explained the greatest proportions of variance in flowering time are highlighted.

selected 14 significant SNPs and 13 of these were located within 100Kb of known flowering time-related genes. Note that these known genes were generally not the nearest gene to the significant SNPs, but associations at this distance are not uncommon in Arabidopsis (Sasaki *et al.* 2021). For ST-GCTA, we selected 34 significant SNPs, among which 26 SNPs were located within 100Kb of known flowering time-related genes. In total, based on our prior knowledge, 14 and 15 genes were detected by MegaBayesC and ST-GCTA, respectively. Detailed comparison on detected genes between MegaBayesC and ST-GCTA is shown in Table 1.

Method	Number of Significant SNPs	Number of False Positives	Detected Genes
MegaBayesC	14	1	AGL17, CRY2, **FLC**, **FT**, **FRL1**, GRP7, HDA6, NF-YC4, **PIE1**, SEF, VP, **VIN3**, **ZTL**, and **DOG1**
ST-GCTA	34	8	CIB2, **FLC**, **FT**, **FRL1**, JMJ14, LATE, LIF2, MRG1, AtNDX, **PIE1**, PRMT4A, TSF, **VIN3**, **ZTL**, and **DOG1**

Table 1 Detailed information on detected genes from ST-GCTA and MegaBayesC. Bold fonts are used to indicate genes that are detected in both methods.

Discussion

The emergence of new types of phenotype data, such as gene expression or spectral reflectances, has created a demand for the development of robust models that are able to analyze large numbers of phenotypes in genome-enabled analysis. Although Bayesian regression models with mixture priors allow for more biologically meaningful prior assumptions on the effect size distributions of causal variants, their corresponding multivariate models (Cheng *et al.* 2018b) suffer from a high computational burden. In this paper, we developed a Bayesian sparse factor model

with mixture priors on marker effects (MegaBayesianAlphabet) to implement both genome-wide prediction and association for analyses with hundreds to tens-of-thousands of phenotypes. MegaBayesianAlphabet uses a moderate number of latent factors (K) to account for the covariance among the observed traits. This substantially reduces the computational burden relative to either a multivariate Bayesian regression model or a multivariate linear mixed model with fully-parameterized trait covariance matrices when the number of traits (t) is large.

However, the sparse factor structure of MegaBayesianAlphabet does not reduce the model complexity enough to enable mixture priors over the millions of genetic markers that are available in many systems from high-density genotyping arrays or whole genome sequencing. When marker effects of the factors and the trait-specific residuals are both included in the model, the number of marker effects to be estimated is equivalent to $(t+K)\times p$, with t being the number of observed traits, K being the number of factors, and p being the number of total SNPs, which would require a tremendous amount of computational time and memory storage for whole-genome analysis.

We therefore developed two approximations to greatly reduce the time complexity of the full model. First, we forced the marker effects to affect the secondary traits through the K factors (although we do allow marker effects to independently control the focal trait). This reduces the number of marker effects to $(K+1)p$. Second, we developed a two-stage approach to prune the candidate markers before subjecting the pruned markers to the MegaBayesC analysis. For our MegaBayesC analysis of the *Arabidopsis* dataset with n = 649, t = 20844, and p = 2804, it took around 3 hours to sample a MCMC chain of 10,000 iterations on a computer with 1 node and 20 CPU.

While MegaBayesC, and MegaBayesianAlphabet more generally, shows promise in its ability to integrate thousands of traits in genome-wide prediction and association, the trade-off between the benefit of incorporating secondary traits and the computational cost brought from the increased model complexity must be considered. Based on our simulated study, MegaBayesC can effectively disentangle the genetic and non-genetic sources of covariation among observed traits. When there is an important environmental component in the variation of focal trait, and this environmental component is shared by many other highly correlated traits, we expect MegaBayesianAlphabet models to provide a large benefit by providing a tool to effectively control for this environmental variation. However, when the secondary traits are not highly correlated with the focal trait, or the heritability of the focal trait is already sufficiently high, MegaBayesianAlphabet may prove less useful.

In this paper, we have focused on two versions of MegaBayesianAlphabet: MegaBayesC with the BayesC prior on the marker effects, and MegaRRBLUP with a ridge prior on the marker effects. Implementing other mixture priors in the `MegaLMM` R package is relatively straightforward, and we anticipate that the BayesA, BayesB or BayesR priors may provide benefits in specific datasets.

Appendix

Gibbs Sampler Updates

Sample F given all other parameters

To sample $\mathbf{F}$, we transpose Eq. 1:

$$\mathbf{Y}^T = \mathbf{\Lambda}^T\mathbf{F}^T + \mathbf{M}_R^T + \mathbf{E}_R^T \tag{4}$$

where $\mathbf{M}_R = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_{2R}\mathbf{B}_{2R}$. Conditioning on $\mathbf{B}_{2F}, \mathbf{B}_{2R}$, columns of $\mathbf{F}^T$ and $\mathbf{M}_R^T$ are uncorrelated and we can represent Eq. 4 as a set of simple linear regressions:

$$(\widetilde{\mathbf{Y}}^T)_i = \widetilde{\mathbf{\Lambda}}^T(\mathbf{F}^T)_i + (\widetilde{\mathbf{M}}_R^T)_i + (\widetilde{\mathbf{E}}_R^T)_i \tag{5}$$

$$(\mathbf{F}^T)_i \sim N(\boldsymbol{\mu}_{(\mathbf{F}^T)_i}, \mathbf{D}_f) \tag{6}$$

$$(\widetilde{\mathbf{E}}_R^T)_i \sim N(\mathbf{0}, \mathbf{D}_{(\widetilde{\mathbf{Y}}^T)_i}) \tag{7}$$

where $\widetilde{\cdot}$ denotes the removal of missing trait data from the corresponding entity. For example, $(\widetilde{\mathbf{Y}}^T)_i$ is the sub-vector of non-missing traits in the ith row of $\mathbf{Y}$. $(\mathbf{F}^T)_i$ denotes the ith row of $\mathbf{F}$, which follows a multivariate normal distribution with mean $\boldsymbol{\mu}_{(\mathbf{F}^T)_i} = \mathbf{B}_{2F}^T(\mathbf{X}_{2F}^T)_i$ and (co)variance matrix $\mathbf{D}_f = \mathbf{\Psi}_{FE}$. $\mathbf{D}_{(\widetilde{\mathbf{Y}}^T)_i} = \widetilde{\mathbf{\Psi}}_{RE}$. $\mathbf{\Psi}_{FE}$ and $\widetilde{\mathbf{\Psi}}_{RE}$ are diagonal matrices.

Let $(\widetilde{\mathbf{Y}}_{cor}^T)_i = (\widetilde{\mathbf{Y}}^T)_i - (\widetilde{\mathbf{M}}_R^T)_i$, we have

$$(\widetilde{\mathbf{Y}}_{cor}^T)_i = \widetilde{\mathbf{\Lambda}}^T(\mathbf{F}^T)_i + (\widetilde{\mathbf{E}}_R^T)_i \tag{8}$$

For simplicity, let $(\widetilde{\mathbf{Y}}_{cor}^T)_i = \mathbf{y}_{cor_i}$, $\widetilde{\mathbf{\Lambda}}^T = \mathbf{\Lambda}^T$, $(\mathbf{F}^T)_i = \mathbf{f}_i$, $\boldsymbol{\mu}_{(\mathbf{F}^T)_i} = \boldsymbol{\mu}_{f_i}$ and $\mathbf{D}_{(\widetilde{\mathbf{Y}}^T)_i} = \mathbf{D}_Y$. The full conditional posterior distribution for $(\mathbf{F}^T)_i$ is derived as:

$$\begin{aligned}
f(\mathbf{f}_i|ELSE) &\propto f(\mathbf{y}_{cor_i}|\mathbf{\Lambda}^T, \mathbf{f}_i, \mathbf{D}_Y) f(\mathbf{f}_i|\boldsymbol{\mu}_{f_i}, \mathbf{D}_f) \\
&\propto \exp\{-\frac{1}{2}(\mathbf{y}_{cor_i} - \mathbf{\Lambda}^T\mathbf{f}_i)^T(\mathbf{D}_Y)^{-1}(\mathbf{y}_{cor_i} - \mathbf{\Lambda}^T\mathbf{f}_i)\} \\
&\times \exp\{-\frac{1}{2}(\mathbf{f}_i - \boldsymbol{\mu}_{f_i})^T(\mathbf{D}_f)^{-1}(\mathbf{f}_i - \boldsymbol{\mu}_{f_i})\} \\
&\propto \exp\{-\frac{1}{2}(\mathbf{f}_i^T(\mathbf{D}_f^{-1} + \mathbf{\Lambda}\mathbf{D}_Y^{-1}\mathbf{\Lambda}^T)\mathbf{f}_i - 2(\mathbf{y}_{cor_i}^T\mathbf{D}_Y^{-1}\mathbf{\Lambda}^T + \boldsymbol{\mu}_{f_i}^T\mathbf{D}_f^{-1})\mathbf{f}_i)\} \\
&\propto \exp\{-\frac{1}{2}(\mathbf{f}_i^T\mathbf{C}\mathbf{f}_i - 2\mathbf{r}^T\mathbf{f}_i)\} \\
&\propto N(\mathbf{C}^{-1}\mathbf{r}, \mathbf{C}^{-1})
\end{aligned}$$

Therefore, $(\mathbf{F}^T)_i|ELSE \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ with

$$\mathbf{\Sigma} = \left[\mathbf{D}_f^{-1} + \widetilde{\mathbf{\Lambda}}\mathbf{D}_{(\widetilde{\mathbf{Y}}^T)_i}^{-1}\widetilde{\mathbf{\Lambda}}^T\right]^{-1} \tag{9}$$

$$\boldsymbol{\mu} = \mathbf{\Sigma}\left[\widetilde{\mathbf{\Lambda}}\mathbf{D}_{(\widetilde{\mathbf{Y}}^T)_i}^{-1}(\widetilde{\mathbf{Y}}_{cor}^T)_i + \mathbf{D}_f^{-1}\boldsymbol{\mu}_{(\mathbf{F}^T)_i}\right] \tag{10}$$

Sample parameters in $\mathbf{\Lambda}$

Full conditional posterior distribution of $\mathbf{\Lambda}$ The prior for $\boldsymbol{\lambda}_j$ is specified as follows:

$$\boldsymbol{\lambda}_{kj} = \begin{cases} N(0, \tau_k^{-1}\sigma_{R_j}^2) & \text{probability } (1-\pi_k) \\ 0 & \text{probability } (\pi_k) \end{cases} \tag{11}$$

$$\sigma_{R_j}^2 \sim iG(a_\sigma, b_\sigma) \tag{12}$$

$$\tau_k = \prod_{h=1}^{k}\delta_h \tag{13}$$

$$\delta_1 = 1, \quad \delta_h \sim Ga(a_\delta, b_\delta) \quad h = 2...k \tag{14}$$

This mixture prior for $\boldsymbol{\lambda}_j$ can be parameterized as: $\mathbf{D}_{\gamma_j}\boldsymbol{\beta}_{\lambda_j}$, where $\mathbf{D}_{\gamma_j} = Diag(\boldsymbol{\gamma}_{\lambda_j})$ with

$$\gamma_{\lambda_j(k)} = \begin{cases} 1 & \text{probability } (1-\pi_{\Lambda_k}) \\ 0 & \text{probability } (\pi_{\Lambda_k}) \end{cases} \tag{15}$$

and $\boldsymbol{\beta}_{\lambda_j} \sim N(\mathbf{0}, \sigma^2_{R_j}\mathbf{D}_\lambda = \sigma^2_{R_j} Diag(\tau_k^{-1}))$ for $k = 1, 2, \ldots K$.

Conditional on $\mathbf{F}$, Eq. 1 can be simplified into t independent univariate linear mixed models for the columns of $\mathbf{Y}$. For the jth column of $\mathbf{Y}$:

$$\mathbf{y}_j = \mathbf{X}_1\mathbf{b}_{1j} + \mathbf{F}\mathbf{D}_{\gamma_j}\boldsymbol{\beta}_{\lambda_j} + \mathbf{X}_{2R}\mathbf{b}_{2R_j} + \mathbf{e}_{R_j} \tag{16}$$

where $\mathbf{e}_{R_j} \sim N(\mathbf{0}, \sigma^2_{R_j}\mathbf{I})$.

$$\begin{aligned}
f(\boldsymbol{\beta}_{\lambda_j}|ELSE) &\propto f(\mathbf{y}_j|\mathbf{b}_{1j}, \mathbf{F}, \mathbf{D}_{\gamma_j}, \boldsymbol{\beta}_{\lambda_j}, \mathbf{b}_{2R_j}, \sigma^2_{R_j}) f(\boldsymbol{\beta}_{\lambda_j}|\mathbf{D}_\lambda, \sigma^2_{R_j}) \\
&\propto \exp\{-\frac{1}{2\sigma^2_{R_j}}(\boldsymbol{\epsilon} - \mathbf{F}\mathbf{D}_{\gamma_j}\boldsymbol{\beta}_{\lambda_j})^T(\boldsymbol{\epsilon} - \mathbf{F}\mathbf{D}_{\gamma_j}\boldsymbol{\beta}_{\lambda_j})\} \times \exp\{-\frac{1}{2\sigma^2_{R_j}}\boldsymbol{\beta}^T_{\lambda_j}\mathbf{D}^{-1}_\lambda\boldsymbol{\beta}_{\lambda_j}\} \\
&\propto \exp\{-\frac{1}{2}[\boldsymbol{\beta}^T_{\lambda_j}(\frac{\mathbf{D}^T_{\gamma_j}\mathbf{F}^T\mathbf{F}\mathbf{D}_{\gamma_j}}{\sigma^2_{R_j}} + \frac{\mathbf{D}^{-1}_\lambda}{\sigma^2_{R_j}})\boldsymbol{\beta}_{\lambda_j} - 2\frac{\boldsymbol{\epsilon}^T\mathbf{F}\boldsymbol{D}_{\gamma_j}}{\sigma^2_{R_j}}\boldsymbol{\beta}_{\lambda_j}]\} \\
&\propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}^T_{\lambda_j}\mathbf{C}\boldsymbol{\beta}_{\lambda_j} - 2\mathbf{r}^T\boldsymbol{\beta}_{\lambda_j})\} \\
&\propto N(\mathbf{C}^{-1}\mathbf{r}, \mathbf{C}^{-1})
\end{aligned}$$

where $\boldsymbol{\epsilon} = \mathbf{y}_j - \mathbf{X}_1\mathbf{b}_{1j} - \mathbf{X}_{2R}\mathbf{b}_{2R_j}$, $\mathbf{C} = \frac{\mathbf{D}^T_{\gamma_j}\mathbf{F}^T\mathbf{F}\mathbf{D}_{\gamma_j} + \mathbf{D}^{-1}_\lambda}{\sigma^2_{R_j}}$, and $\mathbf{r} = \frac{\mathbf{D}^T_{\gamma_j}\mathbf{F}^T\boldsymbol{\epsilon}}{\sigma^2_{R_j}}$.

Besides the full conditional posterior distribution for the multivariate $\boldsymbol{\beta}_{\lambda_j}$ as derived above, a univariate version for the elements in $\boldsymbol{\beta}_{\lambda_j}$ is also derived as follows to prepare for the derivation of $\gamma_{\lambda_{kj}}$.

$$\begin{aligned}
f(\beta_{\lambda_{kj}}|ELSE) &\propto f(\mathbf{y}_j|\mathbf{b}_{1j}, \mathbf{F}, \mathbf{D}_{\gamma_j}, \boldsymbol{\beta}_{\lambda_j}, \mathbf{b}_{2R_j}, \sigma^2_{R_j}) f(\beta_{\lambda_{kj}}|\sigma^2_{R_j}, \tau_k) \\
&\propto \exp\{-\frac{1}{2\sigma^2_{R_j}}(\boldsymbol{\epsilon} - \sum_{i=1}^{K}\mathbf{F}_{\cdot i}\gamma_{\lambda_{ij}}\beta_{\lambda_{ij}})^T(\boldsymbol{\epsilon} - \sum_{i=1}^{K}\mathbf{F}_{\cdot i}\gamma_{\lambda_{ij}}\beta_{\lambda_{ij}})\} \times \exp\{-\frac{\tau_k\beta^2_{\lambda_{kj}}}{2\sigma^2_{R_j}}\} \\
&\propto \exp\{-\frac{1}{2\sigma^2_{R_j}}(\boldsymbol{\epsilon}^* - \mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}\beta_{\lambda_{kj}})^T(\boldsymbol{\epsilon}^* - \mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}\beta_{\lambda_{kj}})\} \times \exp\{-\frac{\tau_k\beta^2_{\lambda_{kj}}}{2\sigma^2_{R_j}}\} \\
&\propto \exp\{-\frac{1}{2}[(\frac{\mathbf{F}^T_{\cdot k}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}}{\sigma^2_{R_j}} + \frac{\tau_k}{\sigma^2_{R_j}})\beta^2_{\lambda_{kj}} - 2\frac{\boldsymbol{\epsilon}^{*T}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}}{\sigma^2_{Rj}}\beta_{\lambda_{kj}}]\} \\
&\propto \exp\{-\frac{1}{2}[A_{kj}\beta^2_{\lambda_{kj}} - 2r\beta_{\lambda_{kj}}]\} \\
&\propto N(A^{-1}_{kj}r, A^{-1}_{kj})
\end{aligned}$$

where $\boldsymbol{\epsilon}^* = \mathbf{y}_j - \mathbf{X}_1\mathbf{b}_{1j} - \mathbf{X}_{2R}\mathbf{b}_{2R_j} - \sum_{i=1, i\neq k}^{K}\mathbf{F}_{\cdot i}\gamma_{\lambda_{ij}}\beta_{\lambda_{ij}}$, $A_{kj} = \frac{\mathbf{F}^T_{\cdot k}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}} + \tau_k}{\sigma^2_{Rj}}$, and $r = \frac{\boldsymbol{\epsilon}^{*T}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}}{\sigma^2_{Rj}}$.

Full conditional posterior distribution of $\gamma_{\lambda_{kj}}$ From the model specification, γ variables can take either 0 or 1. Let $\boldsymbol{\theta}$ denote all other parameters except for $\beta_{\lambda_{kj}}$ and $\gamma_{\lambda_{kj}}$, the marginal full conditional distribution of $\gamma_{\lambda_{kj}}$ that integrates $\beta_{\lambda_{kj}}$ is shown as:

$$f(\gamma_{\lambda_{kj}}|\boldsymbol{\theta}, \mathbf{y}) = \frac{f(\gamma_{\lambda_{kj}}, \boldsymbol{\theta}, \mathbf{y})}{\sum_{\gamma_{\lambda_{kj}}} f(\gamma_{\lambda_{kj}}, \boldsymbol{\theta}, \mathbf{y})} \tag{17}$$

$$= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) f(\boldsymbol{\theta}) f(\gamma_{\lambda_{kj}}|\pi_{\Lambda_k})}{\sum_{\gamma_{\lambda_{kj}}} f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) f(\boldsymbol{\theta}) f(\gamma_{\lambda_{kj}}|\pi_{\Lambda_k})} \tag{18}$$

$$= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) f(\gamma_{\lambda_{kj}}|\pi_{\Lambda_k})}{\sum_{\gamma_{\lambda_{kj}}} f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) f(\gamma_{\lambda_{kj}}|\pi_{\Lambda_k})} \tag{19}$$

Since $f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) = \int f(\mathbf{y}, \beta_{\lambda_{kj}}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) d\beta_{\lambda_{kj}}$, the derivation for $f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}})$ is shown as follows.

$$\begin{aligned}
f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) &= \int f(\mathbf{y}, \beta_{\lambda_{kj}}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) d\beta_{\lambda_{kj}} \\
&= \int f(\mathbf{y}|\beta_{\lambda_{kj}}, \boldsymbol{\theta}, \gamma_{\lambda_{kj}}) f(\beta_{\lambda_{kj}}|\sigma^2_{R_j}, \tau_k) d\beta_{\lambda_{kj}} \\
&\propto \int \exp\{-\frac{1}{2}[(\frac{\mathbf{F}^T_{\cdot k}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}} + \tau_k}{\sigma^2_{R_j}})\beta^2_{\lambda_{kj}} - 2\frac{\boldsymbol{\epsilon}^{*T}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}}{\sigma^2_{R_j}}\beta_{\lambda_{kj}}]\} d\beta_{\lambda_{kj}} \\
&\quad \times \exp\{-\frac{\boldsymbol{\epsilon}^{*T}\boldsymbol{\epsilon}^*}{2\sigma^2_{R_j}}\} \\
&\propto \int \exp\{-\frac{1}{2}[A_{kj}\beta^2_{\lambda_{kj}} - 2r\beta_{\lambda_{kj}} + r^2A^{-1}_{kj}]\} d\beta_{\lambda_{kj}} \\
&\quad \times \exp\{-\frac{1}{2}(\frac{\boldsymbol{\epsilon}^{*T}\boldsymbol{\epsilon}^*}{\sigma^2_{R_j}} - r^2A^{-1}_{kj})\} \\
&\propto \exp\{-\frac{1}{2}(\frac{\boldsymbol{\epsilon}^{*T}\boldsymbol{\epsilon}^*}{\sigma^2_{R_j}} - r^2A^{-1}_{kj})\}
\end{aligned}$$

where $\boldsymbol{\epsilon}^* = \mathbf{y}_j - \mathbf{X}_1\mathbf{b}_{1j} - \mathbf{X}_{2B}\mathbf{b}_{2R_j} - \sum_{i=1, i\neq k}^{K}\mathbf{F}_{\cdot i}\gamma_{\lambda_{ij}}\beta_{\lambda_{ij}}$, $A_{kj} = \frac{\mathbf{F}^T_{\cdot k}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}} + \tau_k}{\sigma^2_{Rj}}$, and $r = \frac{\boldsymbol{\epsilon}^{*T}\mathbf{F}_{\cdot k}\gamma_{\lambda_{kj}}}{\sigma^2_{Rj}}$.

Given Eq. 19, we have

$$\begin{aligned}
f(\gamma_{\lambda_{kj}} = 0) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}} = 0) f(\gamma_{\lambda_{kj}} = 0|\pi_{\Lambda_k})}{\sum_{\gamma_{\lambda_{kj}}} f(\mathbf{y}|\boldsymbol{\theta}, \gamma_{\lambda_{kj}}) f(\gamma_{\lambda_{kj}}|\pi_{\Lambda_k})} \\
&= \frac{\pi_{\Lambda_k} \times \exp\{\frac{1}{2}r^2A^{-1}_{kj}\}}{\sum_{\gamma_{\lambda_{kj}}} \pi_{\Lambda_k} \times \exp\{\frac{1}{2}r^2A^{-1}_{kj}\}} \\
&= \frac{\pi_{\Lambda_k}}{\pi_{\Lambda_k} + (1 - \pi_{\Lambda_k})\exp\{\frac{1}{2}(\frac{\boldsymbol{\epsilon}^{*T}\mathbf{F}_{\cdot k}}{\sigma^2_{R_j}})^2(\frac{\mathbf{F}^T_{\cdot k}\mathbf{F}_{\cdot k} + \tau_k}{\sigma^2_{R_j}})^{-1}\}}
\end{aligned}$$

Full conditional posterior distribution of δ_l In order to sample τ_k, we need to firstly sample δ_l when $K > 1$. To derive the full conditional posterior distribution of δ_l, vectorize $\boldsymbol{\Lambda}$ as $\boldsymbol{\lambda}$. Then, we have

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \cdots \\ \boldsymbol{\lambda}_t \end{bmatrix}_{Kt\times 1} \sim N(\mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} \tau_1^{-1}\sigma^2_{R_1} & & & & & \\ & \ddots & & & & \\ & & \tau_K^{-1}\sigma^2_{R_1} & & & \\ & & & \tau_1^{-1}\sigma^2_{R_t} & & \\ & & & & \ddots & \\ & & & & & \tau_K^{-1}\sigma^2_{R_t} \end{bmatrix}_{Kt\times Kt})$$

Note that the determinant of a diagonal matrix is the product of elements of its diagonal.

$$
\begin{aligned}
&f(\delta_l|ELSE) \propto f(\boldsymbol{\lambda}|\boldsymbol{\Sigma}) f(\delta_l|a_\delta, b_\delta) \\
&\propto \prod_{k=1}^{K}\prod_{j=1}^{t}[(\tau_k)^{-1\times(-1/2)} \exp\{-\frac{1}{2}\frac{\lambda_{kj}^2}{\tau_k^{-1}\sigma_{R_j}^2}\}] \\
&\times (\delta_l)^{a_\delta-1}\exp\{-b_\delta\delta_l\} \\
&\propto [\prod_{k=l}^{K}\prod_{j=1}^{t}(\delta_l)^{1/2}] \times \delta_l^{a_\delta-1} \times \exp\{-\frac{1}{2}\sum_{k=l}^{K}\sum_{j=1}^{t}\frac{\lambda_{kj}^2\tau_k}{\sigma_{R_j}^2}\}\exp\{-b_\delta\delta_l\} \\
&\propto [\prod_{k=l}^{K}\prod_{j=1}^{t}(\delta_l)^{1/2}] \times \delta_l^{a_\delta-1} \times \exp\{-b_\delta\delta_l\} \\
&\times \exp\{-\frac{1}{2}\sum_{k=l}^{K}\sum_{j=1}^{t}\frac{\lambda_{kj}^2(\prod_{h=1,h\neq l}^{k}\delta_h)\delta_l}{\sigma_{R_j}^2}\} \\
&\propto (\delta_l)^{\frac{t(K-l+1)}{2}+a_\delta-1}\exp\{-b_\delta\delta_l\} \\
&\times \exp\{-\frac{1}{2}[\sum_{k=l}^{K}(\prod_{h=1,h\neq l}^{k}\delta_h)\sum_{j=1}^{t}\frac{\lambda_{kj}^2}{\sigma_{R_j}^2}]\delta_l\} \\
&\propto Ga(a_\delta+\frac{t(K-l+1)}{2}, b_\delta+\frac{1}{2}\sum_{k=l}^{K}(\prod_{h=1,h\neq l}^{k}\delta_h)\sum_{j=1}^{t}\frac{\lambda_{kj}^2}{\sigma_{R_j}^2})
\end{aligned}
$$

Parallel Model Setting

Given $\mathbf{F}$ and $\boldsymbol{\Lambda}$, although the design matrices may differ for columns of $\mathbf{Y}$ and $\mathbf{F}$, the form of both sets of conditional model can be similarly expressed as:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\alpha} + \mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta} + \mathbf{e} \tag{20}$$

where

$$
\begin{aligned}
\boldsymbol{\alpha} &\sim N(\mathbf{0}, \infty) && (21)\\
\boldsymbol{\beta} &\sim N(0, \sigma_\beta^2\mathbf{I}) && (22)\\
\mathbf{D}_\gamma &= Diag(\boldsymbol{\gamma}) && (23)\\
\gamma_i &= \begin{cases} 1 & \text{probability } (1-\pi) \\ 0 & \text{probability } (\pi) \end{cases} && (24)\\
\mathbf{e} &\sim N(\mathbf{0}, \sigma^2\mathbf{I}) && (25)\\
\sigma^2 &\sim iG(a_0, b_0) && (26)\\
\sigma_\beta^2 &\sim iG(a_\beta, b_\beta) && (27)
\end{aligned}
$$

Conditional on $\mathbf{F}$ and $\boldsymbol{\Lambda}$, Eq. 1 can be simplified into t independent univariate linear mixed models for the columns of $\mathbf{Y}_{cor} = \mathbf{Y} - \mathbf{F}\boldsymbol{\Lambda}$:

$$\mathbf{y}_{cor_j} = \mathbf{X}_1\mathbf{b}_{1_j} + \mathbf{X}_{2B}\boldsymbol{\beta}_{B2R_j} \circ \boldsymbol{\gamma}_{B2R_j} + \mathbf{e}_{R_j} \tag{28}$$

where

$$
\begin{aligned}
\mathbf{b}_{1_j} &\sim N(\mathbf{0}, \infty\mathbf{I}) && (29)\\
\boldsymbol{\beta}_{B2R_j} &\sim N(\mathbf{0}, \sigma_{B2R_j}^2\mathbf{I}) && (30)\\
\boldsymbol{\gamma}_{B2R_{j(i)}} &= \begin{cases} 1 & \text{probability } (1-\pi_j) \\ 0 & \text{probability } (\pi_j) \end{cases} && (31)\\
\mathbf{e}_{R_j} &\sim N(\mathbf{0}, \sigma_{R_j}^2\mathbf{I}_n) && (32)
\end{aligned}
$$

Besides the columns of $\mathbf{Y}$, the columns of $\mathbf{F}$ (Eq. 2) can be similarly expressed into K independent univariate linear mixed models:

$$\mathbf{f}_k = \mathbf{X}_{2F}\boldsymbol{\beta}_{B2F_k} \circ \boldsymbol{\gamma}_{B2F_k} + \mathbf{e}_{F_k} \tag{33}$$

where

$$
\begin{aligned}
\boldsymbol{\beta}_{B2F_k} &\sim N(\mathbf{0}, \sigma_{B2F_k}^2\mathbf{I}) && (34)\\
\boldsymbol{\gamma}_{B2F_{k(i)}} &= \begin{cases} 1 & \text{probability } (1-\pi_{F_k}) \\ 0 & \text{probability } (\pi_{F_k}) \end{cases} && (35)\\
\mathbf{e}_{F_k} &\sim N(\mathbf{0}, \sigma_{F_k}^2\mathbf{I}_n) && (36)\\
& && (37)
\end{aligned}
$$

Here, factor-specific and trait-specific prior on the marker exclusion probability (π_{F_k} and π_j) and the variance of marker effects ($\sigma_{B2F_k}^2$ and $\sigma_{B2R_j}^2$) are used for each latent factor and observed trait. We can see that the columns of $\mathbf{Y}$ and $\mathbf{F}$ can be generally expressed by Eq. 20. That is, for columns of $\boldsymbol{Y}$, $\mathbf{y} = \mathbf{y}_{cor_j}$, $\boldsymbol{\alpha} = \mathbf{b}_{1_j}$, $\mathbf{D}_\gamma = Diag(\boldsymbol{\gamma}_{B2R_j})$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{B2R_j}$, $\mathbf{e} = \mathbf{e}_{R_j}$, $\sigma^2 = \sigma_{R_j}^2$, $\sigma_\beta^2 = \sigma_{B2R_j}^2$. Similarly, for columns of $\mathbf{F}$, $\mathbf{y} = \mathbf{f}_k$, $\boldsymbol{\alpha}$ is empty, $\mathbf{D}_\gamma = Diag(\boldsymbol{\gamma}_{B2F_k})$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{B2F_k}$, $\mathbf{e} = \mathbf{e}_{F_k}$, $\sigma^2 = \sigma_{F_k}^2$, $\sigma_\beta^2 = \sigma_{B2F_k}^2$. Furthermore, we defined the following term based on the notation in Eq. 20:

$$\mathbf{V}_\beta = \mathbf{X}_2\mathbf{D}_\gamma\mathbf{X}_2^T\sigma_\beta^2 + \sigma^2\mathbf{I}$$

Full conditional posterior distribution of $\boldsymbol{\alpha}$ The conditional posterior distribution for $\boldsymbol{\alpha}(i.e., \mathbf{b}_{1j})$ is derived as (integrating out $\boldsymbol{\beta}$):

$$
\begin{aligned}
&f(\boldsymbol{\alpha}|\cdot) \propto f(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{V}_\beta) \\
&\propto \exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\alpha})^T\mathbf{V}_\beta^{-1}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\alpha})\} \\
&\propto \exp\{-\frac{1}{2}(\boldsymbol{\alpha}^T\mathbf{X}_1^T\mathbf{V}_\beta^{-1}\mathbf{X}_1\boldsymbol{\alpha} - 2\mathbf{y}^T\mathbf{V}_\beta^{-1}\mathbf{X}_1\boldsymbol{\alpha})\} \\
&\propto \exp\{-\frac{1}{2}(\boldsymbol{\alpha}^T\mathbf{A}_\alpha\boldsymbol{\alpha} - 2\mathbf{r}^T\boldsymbol{\alpha})\} \\
&\propto N(\mathbf{A}_\alpha^{-1}\mathbf{r}, \mathbf{A}_\alpha^{-1})
\end{aligned}
$$

where $\mathbf{A}_\alpha = \mathbf{X}_1^T\mathbf{V}_\beta^{-1}\mathbf{X}_1$, $\mathbf{r} = \mathbf{X}_1^T\mathbf{V}_\beta^{-1}\mathbf{y}$. The dimension of $\mathbf{A}_\alpha$ is $a(b_1) \times a(b_1)$, and the dimension of $\mathbf{V}_\beta$ is $n \times n$.

Full conditional posterior distribution of σ^2 The conditional posterior distribution for $\sigma^2(i.e., \sigma_{R_j}^2$ and $\sigma_{F_k}^2)$ is derived as:

$$
\begin{aligned}
&f(\sigma^2|\cdot) \propto f(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{D}_\gamma, \boldsymbol{\beta}, \sigma^2) f(\sigma^2|a_0, b_0) \\
&\propto (\sigma^2)^{-\frac{n}{2}}\exp\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\alpha}-\mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}_1\boldsymbol{\alpha}-\mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta})\} \\
&\times (\sigma^2)^{-a_0-1}\exp\{-\frac{b_0}{\sigma^2}\} \\
&\propto (\sigma^2)^{-(\frac{n}{2}+a_0)-1}\exp\{-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/2}{\sigma^2}\}\exp\{-\frac{b_0}{\sigma^2}\} \\
&\propto iG(\frac{n}{2}+a_0, \frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{2}+b_0)
\end{aligned}
$$

where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta}$.

Full conditional posterior distribution of $\boldsymbol{\beta}$ The conditional posterior distribution for $\boldsymbol{\beta}$ is derived as:

$$
\begin{aligned}
f(\boldsymbol{\beta}|\cdot) &\propto f(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{D}_\gamma, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}|\sigma^2_\beta) \\
&\propto \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta})\} \\
&\times \exp\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^T\boldsymbol{\beta}\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\epsilon} - \mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta})^T(\boldsymbol{\epsilon} - \mathbf{X}_2\mathbf{D}_\gamma\boldsymbol{\beta})\} \times \exp\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^T\boldsymbol{\beta}\} \\
&\propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}^T(\frac{\mathbf{D}_\gamma^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{D}_\gamma}{\sigma^2} + \frac{1}{\sigma^2_\beta}\mathbf{I})\boldsymbol{\beta} - 2\frac{\boldsymbol{\epsilon}^T\mathbf{X}_2\mathbf{D}_\gamma}{\sigma^2}\boldsymbol{\beta})\} \\
&\propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}^T\mathbf{A}_\beta\boldsymbol{\beta} - 2\mathbf{r}^T\boldsymbol{\beta})\} \\
&\propto N(\mathbf{A}_\beta^{-1}\mathbf{r}, \mathbf{A}_\beta^{-1})
\end{aligned}
$$

where $\mathbf{A}_\beta = \frac{\mathbf{D}_\gamma^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{D}_\gamma}{\sigma^2} + \frac{1}{\sigma^2_\beta}\mathbf{I}$, $\mathbf{r} = \frac{\mathbf{D}_\gamma^T\mathbf{X}_2^T\boldsymbol{\epsilon}}{\sigma^2}$. The dimension of $\mathbf{A}_\beta$ is $b \times b$. For columns of $\mathbf{Y}$, $b = b_{2R}$, $\boldsymbol{\epsilon} = \mathbf{y}_{cor_j} - \mathbf{X}_1\mathbf{b}_{1_j}$. For columns of $\mathbf{F}$, $b = b_{2F}$, $\boldsymbol{\epsilon} = \mathbf{f}_k$. Besides the full conditional posterior distribution of the multivariate $\boldsymbol{\beta}$ as derived above, a univariate version for the elements β_l in $\boldsymbol{\beta}$ is also written as follows.

$$
\begin{aligned}
f(\beta_l|\cdot) &\propto f(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{D}_\gamma, \boldsymbol{\beta}, \sigma^2) f(\beta_l|\sigma^2_\beta) \\
&\propto \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \sum_{i=1}^{b}\mathbf{X}_{2\cdot i}\gamma_i\beta_i)^T(\mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \sum_{i=1}^{b}\mathbf{X}_{2\cdot i}\gamma_i\beta_i)\} \\
&\times \exp\{-\frac{\beta_l^2}{2\sigma^2_\beta}\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\epsilon} - \mathbf{X}_{2\cdot l}\gamma_l\beta_l)^T(\boldsymbol{\epsilon} - \mathbf{X}_{2\cdot l}\gamma_l\beta_l)\}\exp\{-\frac{\beta_l^2}{2\sigma^2_\beta}\} \\
&\propto \exp\{-\frac{1}{2}[(\frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2} + \frac{1}{\sigma^2_\beta})\beta_l^2 - 2\frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2}\beta_l]\} \\
&\propto exp\{-\frac{1}{2}(A_\beta\beta_l^2 - 2r\beta_l)\} \\
&\propto N(A_\beta^{-1}r, A_\beta^{-1})
\end{aligned}
$$

where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \sum_{i=1, i\neq l}^{b}\mathbf{X}_{2\cdot i}\gamma_i\beta_i$, $A_\beta = \frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2} + \frac{1}{\sigma^2_\beta}$, $r = \frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2}$.

Full conditional posterior distribution of γ_l Let $\boldsymbol{\theta}$ denote all other parameters except for β_l and γ_l, the marginal full conditional distribution of γ_l that integrates out β_l is shown as:

$$
\begin{aligned}
f(\gamma_l|\boldsymbol{\theta}, \mathbf{y}) &= \frac{f(\gamma_l, \boldsymbol{\theta}, \mathbf{y})}{\sum_{\gamma_l} f(\gamma_l, \boldsymbol{\theta}, \mathbf{y})} && (38) \\
&= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l) f(\boldsymbol{\theta}) f(\gamma_l|\pi)}{\sum_{\gamma_l} f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l) f(\boldsymbol{\theta}) f(\gamma_l|\pi)} && (39) \\
&= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l) f(\gamma_l|\pi)}{\sum_{\gamma_l} f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l) f(\gamma_l|\pi)} && (40)
\end{aligned}
$$

Since $f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l) = \int f(\mathbf{y}, \beta_l|\boldsymbol{\theta}, \gamma_l) d\beta_l$, the derivation for $f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l)$ is shown as follows.

$$
\begin{aligned}
f(\mathbf{y}|\boldsymbol{\theta}, \gamma_l) &= \int f(\mathbf{y}, \beta_l|\boldsymbol{\theta}, \gamma_l) d\beta_l \\
&= \int f(\mathbf{y}|\beta_l, \boldsymbol{\theta}, \gamma_l) f(\beta_l|\sigma^2_\beta) d\beta_l \\
&\propto \int \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\epsilon} - \mathbf{X}_{2\cdot l}\gamma_l\beta_l)^T(\boldsymbol{\epsilon} - \mathbf{X}_{2\cdot l}\gamma_l\beta_l)\}\exp\{-\frac{\beta_l^2}{2\sigma^2_\beta}\} d\beta_l \\
&\propto \int \exp\{-\frac{1}{2}[(\frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2} + \frac{1}{\sigma^2_\beta})\beta_l^2 - 2\frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2}\beta_l]\} d\beta_l \\
&\times \exp\{-\frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\} \\
&\propto \int \exp\{-\frac{1}{2}(A_\beta\beta_l^2 - 2r\beta_l + r^2A_\beta^{-1})\} d\beta_l \\
&\times \exp\{-\frac{1}{2}(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/\sigma^2 - r^2A_\beta^{-1})\} \\
&\propto \exp\{-\frac{1}{2}(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/\sigma^2 - r^2A_\beta^{-1})\}
\end{aligned}
$$

where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}_1\boldsymbol{\alpha} - \sum_{i=1, i\neq l}^{b}\mathbf{X}_{2\cdot i}\gamma_i\beta_i$, $A_\beta = \frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2} + \frac{1}{\sigma^2_\beta}$, $r = \frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}\gamma_l}{\sigma^2}$.

Given Eq. 40, we have

$$
\begin{aligned}
f(\gamma_l = 0) &= \frac{\pi \times \exp\{\frac{1}{2}r^2A_\beta^{-1}\}}{\sum_{\gamma_l}\pi_{\gamma_l} \times \exp\{\frac{1}{2}r^2A_\beta^{-1}\}} \\
&= \frac{\pi}{\pi + (1-\pi)\exp\{\frac{1}{2}(\frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}}{\sigma^2})^2(\frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}}{\sigma^2} + \frac{1}{\sigma^2_\beta})^{-1}\}}
\end{aligned}
$$

$$
\begin{aligned}
f(\gamma_l = 1) &= \frac{(1-\pi) \times \exp\{\frac{1}{2}r^2A_\beta^{-1}\}}{\sum_{\gamma_l}\pi_{\gamma_l} \times \exp\{\frac{1}{2}r^2A_\beta^{-1}\}} \\
&= \frac{(1-\pi) \times \exp\{\frac{1}{2}(\frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}}{\sigma^2})^2(\frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}}{\sigma^2} + \frac{1}{\sigma^2_\beta})^{-1}\}}{\pi + (1-\pi)\exp\{\frac{1}{2}(\frac{\boldsymbol{\epsilon}^T\mathbf{X}_{2\cdot l}}{\sigma^2})^2(\frac{\mathbf{X}_{2\cdot l}^T\mathbf{X}_{2\cdot l}}{\sigma^2} + \frac{1}{\sigma^2_\beta})^{-1}\}}
\end{aligned}
$$

Full conditional posterior distribution of σ^2_β

$$
\begin{aligned}
f(\sigma^2_\beta) &\propto f(\boldsymbol{\beta}|\sigma^2_\beta) f(\sigma^2_\beta|a_\beta, b_\beta) \\
&\propto (\sigma^2_\beta)^{-\frac{b}{2}}\exp\{-\frac{1}{2\sigma^2_\beta}\boldsymbol{\beta}^T\boldsymbol{\beta}\} \times (\sigma^2_\beta)^{-a_\beta-1}\exp\{-\frac{b_\beta}{\sigma^2_\beta}\} \\
&\propto (\sigma^2_\beta)^{-\frac{b}{2}-a_\beta-1}\exp\{-\frac{1}{\sigma^2_\beta}(\frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2} + b_\beta)\} \\
&\propto iG(\frac{b}{2} + a_\beta, \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2} + b_\beta)
\end{aligned}
$$

Specification of parameters for the real data analysis performed in the paper

Model (Analysis)	K	Chain Length	Burn-in
MegaBayesC (GP)	100	10K	2K
MegaGBLUP (GP)	100	10K	2K
MegaRRBLUP (GP)	100	10K	2K
MegaBayesC (GWAS)	100	80K	20K

Supplementary Plots

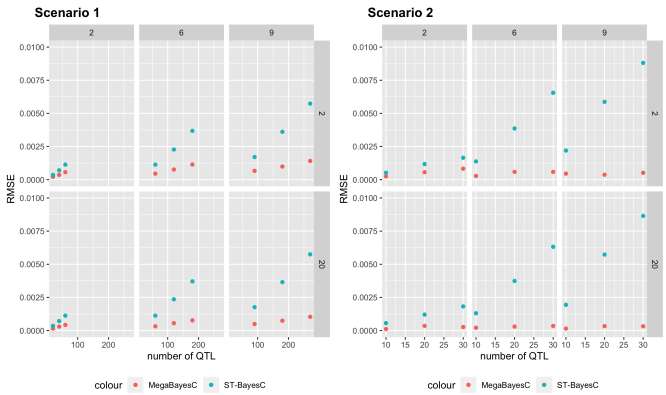

Figure 9 RMSE of estimated marker effects under two scenarios for the total 36 different simulation settings. The performance of single-trait BayesC and MegaBayesC were compared. The performance of models for the simulation setting with n_{trait} = 2 and 20 are presented at the first and second row, respectively. The performance of models for the simulation setting with n_{factor} = 2, 6, 9 are presented at the first, second, and third column, respectively.

Literature Cited

Alonso-Blanco, C., J. Andrade, C. Becker, F. Bemm, J. Bergelson, *et al.*, 2016 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell **166**: 481–491.

Araus, J. L., S. C. Kefauver, M. Zaman-Allah, M. S. Olsen, and J. E. Cairns, 2018 Translating high-throughput phenotyping into genetic gain. Trends in plant science **23**: 451–466.

Barrett, T., S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, *et al.*, 2012 Ncbi geo: archive for functional genomics data sets—update. Nucleic acids research **41**: D991–D995.

Bhattacharya, A. and D. B. Dunson, 2011 Sparse bayesian infinite factor models. Biometrika pp. 291–306.

Bouché, F., G. Lobet, P. Tocquin, and C. Périlleux, 2016 Florid: an interactive database of flowering-time gene networks in arabidopsis thaliana. Nucleic Acids Research **44**: D1167–D1171.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, *et al.*, 2008 High-dimensional sparse factor modeling: applications in gene expression genomics. Journal of the American Statistical Association **103**: 1438–1456.

Cheng, H., R. Fernando, and D. Garrick, 2018a Jwas: Julia implementation of whole-genome analysis software. In *Proceedings of the world congress on genetics applied to livestock production*, volume 11, p. 859.

Cheng, H., K. Kizilkaya, J. Zeng, D. Garrick, and R. Fernando, 2018b Genomic Prediction from Multiple-Trait Bayesian Regression Methods Using Mixture Priors. Genetics **209**: genetics.300650.2018.

Cheng, H., L. Qu, D. J. Garrick, and R. L. Fernando, 2015 A fast and efficient Gibbs sampler for BayesB in whole-genome analyses. Genet Sel Evol **47**: 80.

Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics **193**: 347–365.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of dairy science **95**: 4114–4129.

Gibson, G. and B. Weir, 2005 The quantitative genetics of transcription. TRENDS in Genetics **21**: 616–623.

Gilmour, A., R. Thompson, and B. Cullis, 1995 Linear mixed models algorithm for average information reml: an efficient in linear mixed models variance parameter estimation. Biometrics **51**: 1440–1450.

Habier, D., R. L. Fernando, and J. C. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**: 2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC bioinformatics **12**: 1–12.

Henderson, C. and R. Quaas, 1976 Multiple trait evaluation using relatives' records. Journal of Animal Science **43**: 1188–1197.

Kizilkaya, K., R. Fernando, and D. Garrick, 2010 Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. Journal of animal science **88**: 544–551.

Krause, M. R., L. González-Pérez, J. Crossa, P. Pérez-Rodríguez, O. Montesinos-López, *et al.*, 2019 Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. G3: Genes, Genomes, Genetics **9**: 1231–1247.

Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**: 1–21.

McGraw, E. A., Y. H. Ye, B. Foley, S. F. Chenoweth, M. Higgie, *et al.*, 2011 High-dimensional variance partitioning reveals the modular genetic basis of adaptive divergence in gene expression during reproductive character displacement. Evolution: International Journal of Organic Evolution **65**: 3126–3137.

Mehrban, H., D. H. Lee, M. H. Moradi, C. IlCho, M. Naserkheil, *et al.*, 2017 Predictive performance of genomic selection methods for carcass traits in hanwoo beef cattle: impacts of the genetic architecture. Genetics Selection Evolution **49**: 1–13.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Moser, G., S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, *et al.*, 2015 Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLOS Genetics **11**: e1004969.

Park, T. and G. Casella, 2008 The bayesian lasso. Journal of the American Statistical Association **103**: 681–686.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, *et al.*, 2012 Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome **5**: 103–113.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, *et al.*, 2007 Plink: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics **81**: 559–575.

Runcie, D. and H. Cheng, 2019 Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. G3: Genes, Genomes, Genetics **9**: 3727–3741.

Runcie, D. E. and S. Mukherjee, 2013 Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. Genetics **194**: 753–767.

Runcie, D. E., J. Qu, H. Cheng, and L. Crawford, 2021 Megalmm: Mega-scale linear mixed models for genomic predictions with thousands of traits. Genome Biology **22**: 1–25.

Rutkoski, J., J. Poland, S. Mondal, E. Autrique, L. G. Pérez, *et al.*, 2016 Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3: Genes, Genomes, Genetics **6**: 2799–2808.

Sasaki, E., T. Köcher, D. L. Filiault, and M. Nordborg, 2021 Revisiting a gwas peak in arabidopsis thaliana reveals possible confounding by genetic heterogeneity. Heredity **127**: 245–252.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of dairy science **91**: 4414–4423.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, *et al.*, 2017 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics **101**: 5–22.

Wang, Z., D. Chapman, G. Morota, and H. Cheng, 2020 A multiple-trait bayesian variable selection regression method for integrating phenotypic causal networks in genome-wide association studies. G3: Genes, Genomes, Genetics **10**: 4439–4448.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genetics Research **75**: 249–252.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, *et al.*, 2016 Mixture models detect large effect qtl better than gblup

and result in more accurate and persistent predictions. Journal of animal science and biotechnology **7**: 1–6.

Xiong, Q., N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey, 2012 Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome research **22**: 386–397.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 Gcta: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.

Zhou, X. and M. Stephens, 2014 Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods **11**: 407–409.