Supplementary Material to "Variable Prioritization in Nonlinear Black Box Methods: A Genetic Association Case Study"

Lorin Crawford[1-3,†], Seth R. Flaxman[4,5], Daniel E. Runcie[6], and Mike West[7]

1 Department of Biostatistics, Brown University, Providence, RI, USA
2 Center for Statistical Sciences, Brown University, Providence, RI, USA
3 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
4 Department of Mathematics, Imperial College London, London, UK
5 Data Science Institute, Imperial College London, London, UK
6 Department of Plant Sciences, University of California, Davis, CA, USA
7 Department of Statistical Science, Duke University, Durham, NC, USA

Contents

1 Supplementary Text

1.1 Motivating Simulation Study: Genomic Selection

In this particular subsection, our goal is to empirically motivate the desire to have a principled variable selection procedure for nonparametric and/or black box methods. We also want to empirically illustrate the behavior of the effect size analog. To do this, we consider a similar statistical genetics-inspired simulation design that we utilize in the main text. Briefly, we will assume that all of the observed genetic effects explain a fixed proportion of the total phenotypic variance. This proportion is referred to as the broad-sense heritability of the trait, which we set to be $\mathrm{H}^2 = 0.3$. From the more conventional statistics perspective, the parameter H^2 can alternatively be described as a factor controlling the signal-to-noise ratio. Next, we make use of $n = 2000$ synthetic genotypes that are independently generated to have $p = 500$ single nucleotide polymorphisms (SNPs) with allele frequencies randomly sampled from a uniform distribution over values ranging from $[0.05, 0.5]$. The resulting $n \times p$ simulated genotype matrix $\mathbf{X}$ is then used to generate continuous phenotypes that mirror genetic architectures affected by a combination of linear (additive) and interaction (epistatic) effects. Specifically, we randomly choose $j^* = 30$ "causal" (or truly associated) markers that we classify into two distinct groups: (i) a small set of 5 select variants, and (ii) a larger set of 25 causal variants. All causal markers have additive effects and, when applicable, the group 1 causal SNPs interact with group 2 causal SNPs, but never with each other (the same rule applies to the second group).

The linear effect sizes for all j^* associated genetic variants are assumed to come from a standard normal distribution or $\beta_{j^*} \sim \mathcal{N}(0, 1)$. Next, we create a separate matrix $\mathbf{W}$ which holds all pairwise interactions between the group 1 and 2 causal markers. These corresponding interaction effect sizes are also drawn as $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We scale both the additive and interaction effects so that collectively they explain a fixed proportion of H^2. Namely, the additive effects make up $\rho\%$, while the pairwise interactions make up the remaining $(1-\rho)\%$. Alternatively, the proportion of the heritability explained by additivity is said to be $\mathbb{V}(\mathbf{X}\boldsymbol{\beta}) = \rho\mathrm{H}^2$, while the proportion detailed by nonlinearity is given as $\mathbb{V}(\mathbf{W}\boldsymbol{\gamma}) = (1-\rho)\mathrm{H}^2$. Once we obtain the final effect sizes for all causal variants, we draw normally distributed random errors as $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to make up the remaining $(1 - \mathrm{H}^2)\%$ of the total phenotypic variance $\mathbb{V}(\mathbf{y})$. Finally, continuous phenotypes are then created by summing over all observed effects using the following two simulation models:

(i) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$

(ii) $\mathbf{y} = \mathbf{Z}\boldsymbol{\omega} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$

where $\mathbf{Z}$ contains covariates representing population structure, and $\boldsymbol{\omega}$ are the corresponding fixed effects which are also assumed to follow a standard multivariate normal distribution. We will consider three simulation scenarios. Scenario I involves phenotypes generated by model (i). Scenarios II and III consider model (ii) where we introduce population stratification effects by allowing the top 5 and 10 genotype principal components (PCs) $\mathbf{Z}$ to make up 30% of the overall variation in the simulated traits, respectively.

Within these three scenarios, we also consider two choices for the parameter $\rho = \{0.5, 1\}$. Intuitively, $\rho = 1$ represents the limiting case where the variation of a trait is driven by solely additive effects. For $\rho = 0.5$, the additive and interaction effects are assumed to equally contribute to the total phenotypic variance. Once again, in these simulations, we are interested in demonstrating the power of the nonparametric methods and their ability to facilitate out-of-sample prediction. We evaluate the predictive accuracy of two methods. The first is a standard GP regression model with a zero mean prior and a Gaussian covariance function. Posterior estimates of the function $\mathbf{f}$ are obtained by using a Gibbs sampler with 10,000 MCMC iterations and hyper-parameters set to $a = 5$ and $b = 2/5$ (see Algorithmic Overview below). The second method we consider is a standard linear model, which is fit by using ordinary least squares (OLS). The linear model is used to serve as a baseline. Mean squared error (MSE) and predictive

correlation (R) are used to compare out-of-sample predictive accuracy. We also record the tabulated frequency for which a given method exhibits the lowest MSE and greatest predictive R, which we denote as $\text{Opt}\%_{\text{MSE}}$ and $\text{Opt}\%_{R}$, respectively. We analyze 100 different simulated datasets for all simulation scenarios I-III and each case ρ. For each iterative run, we randomly split the data into training data with 80% of the samples and a test set with the remaining 20%.

Overall numerical results for each case of ρ are presented in Table S1, and then further illustrated as boxplots in Figure S1 to show how the two methods perform while taking into account variability across simulations. The GP regression outperforms the standard linear model OLS estimates in all of the simulation scenarios. As expected, this discrepancy is obvious when there is population stratification (i.e. Scenarios II and III), particularly when there are underlying interactions affecting the generation of phenotypes (i.e. $\rho = 0.5$). This is unsurprising given that the GP determines function estimates in a nonlinear space. Altogether, these results are consistent with past genomic selection and phenotypic prediction studies regarding nonparametric models [1].

1.2 Identifiability of the Effect Size Analog

In the main text of this paper, we consider a generalized projection operator between an infinite dimensional function space, called a reproducing kernel Hilbert space (RKHS), and the original genotype space. An RKHS may be defined based on a nonlinear transformation of data using a positive definite covariance function (or kernel). Here, we conduct inference by specifying a Gaussian process (GP) to describe a prior distribution over the elements in this space

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{S1}$$

where $f(\bullet)$ is completely specified by its mean function and positive definite covariance (kernel) function, $m(\bullet)$ and $k(\bullet, \bullet)$, respectively. In practice, we condition on a finite set of locations (i.e. a set of observed samples n), and jointly rewrite the Gaussian process prior as a multivariate normal [2]

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}, \quad \mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}). \tag{S2}$$

Altogether, we refer to the above as taking a "weight-space" view on Gaussian process regression [3]. Briefly, $\mathbf{y}$ is an n-dimensional vector of phenotypes, the residual noise $\boldsymbol{\varepsilon}$ is assumed to follow a multivariate normal distribution with mean zero and variance τ^2, and $\mathbf{I}$ is an identity matrix. The vector $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\intercal$ is assumed to come from a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{K} = \boldsymbol{\Psi}^\intercal\boldsymbol{\Psi}$ with each $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Additionally, the matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}(\mathbf{x}_1), \ldots, \boldsymbol{\psi}(\mathbf{x}_n)]^\intercal$ is a corresponding matrix of concatenated vector spaces $\boldsymbol{\psi}(\mathbf{x}) = \{\sqrt{\delta_\ell}\phi_\ell(\mathbf{x})\}_{\ell=1}^{\infty}$ detailing a subspace of the RKHS, $\mathcal{H}_{\mathbf{x}}$, that is realized by the span of the data. Namely,

$$\mathcal{H}_{\mathbf{x}} = \left\{ f \,|\, f(\mathbf{x}) = \boldsymbol{\Psi}_{\mathbf{x}}^\intercal\mathbf{c} \text{ and } \|f\|_{\text{K}}^2 < \infty \right\}$$

where $\| \bullet \|_{\text{K}}$ is the RKHS norm, and the coefficients $\mathbf{c}$ determine the nonlinear function.

Our goal is to specify an identifiability requirement for the effect size analog. Similar results have been previously presented for random Fourier feature maps [4]. The results in this section are, effectively, a generalization of these claims. A reasonable identifiability requirement for the effect size analog is that two different functions in $\mathcal{H}_{\mathbf{x}}$ will result in two different vectors for $\widetilde{\boldsymbol{\beta}}$. This requirement can be restated as the projection $\mathbf{P} = \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal = \mathbf{X}^\dagger\mathbf{f}$ should be an injective map from $\mathbf{c}$ to $\widetilde{\boldsymbol{\beta}}$. First we consider the classic linear regression setting

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{y}, \tag{S3}$$

where $\mathbf{X}^\dagger$ is the Moore-Penrose pseudoinverse — which, in the case of a full rank design matrix, equates to

$\mathbf{X}^\dagger = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ and leads to the standard ordinary least-squares (OLS) regression coefficient estimates. Observe that two vectors $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, which only differ in the null space of $\mathbf{X}$, will give rise to the same model estimate $\widehat{\mathbf{f}}$. This same issue will arise for our nonlinear effect size analog. Hence, the statement we will make about the injectivity of the map $\mathbf{P}$ will hold modulo the null space of $\mathbf{X}$.

Claim S1 (Generalizing Results from [4])**.** *Consider a strictly positive definite covariance matrix* $\mathbf{K}$ *with feature map* $\psi : \mathbb{R}^p \to \mathbb{R}^p$*. The projection* $\mathbf{P}$ *is injective for any coefficient vector for which the projection* $\mathbf{P}$ *is in the span of the design matrix* $\mathbf{X}$*. Alternatively, the projection* $\mathbf{P}$ *is injective for the span of the design matrix, span*$(\mathbf{X})$*.*

Proof. Consider positive definite covariance matrices $\mathbf{K}$. The assumption that the covariance function is positive definite is key as it implies that the resulting $\boldsymbol{\Psi}$ spans the entire p-dimensional predictor space. In the case that $\mathbf{K}$ is positive semi-definite, we have to understand the composition of the null space of $\mathbf{K}$ with the null space of the design matrix $\mathbf{X}$.

Let $\mathbf{c}_1$ and $\mathbf{c}_2$ be two different coefficient vectors corresponding to functions $\mathbf{f}_1$ and $\mathbf{f}_2$ in the restricted RKHS subspace, respectively. There exists $\boldsymbol{\delta}$ such that $\mathbf{c}_2 = \mathbf{c}_1 + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \neq \mathbf{0}$ and

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_1 &= \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal\mathbf{c}_1 \\ \widetilde{\boldsymbol{\beta}}_2 &= \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal\mathbf{c}_2 = \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal(\mathbf{c}_1 + \boldsymbol{\delta}) = \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal\mathbf{c}_1 + \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal\boldsymbol{\delta}. \end{aligned}$$

Since $\mathbf{K}$ is a positive definite matrix,

$$\mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal\boldsymbol{\delta} = \boldsymbol{\delta}_\parallel + \boldsymbol{\delta}_\perp,$$

where $\boldsymbol{\delta}_\parallel$ is the projection onto the span of $\mathbf{X}$, and $\boldsymbol{\delta}_\perp$ is the projection onto the null space of $\mathbf{X}$. Note that $\mathbf{X}\boldsymbol{\delta}_\perp = \mathbf{0}$, so we cannot separate $\widetilde{\boldsymbol{\beta}}_1 \neq \widetilde{\boldsymbol{\beta}}_2$ if the difference between $\mathbf{c}_1$ and $\mathbf{c}_2$ projects onto the null space of $\mathbf{X}$. By definition, if part of the vector $\boldsymbol{\delta}$ projects onto the span of $\mathbf{X}$, then $\mathbf{P}\boldsymbol{\delta} \neq \mathbf{0}$ and $\widetilde{\boldsymbol{\beta}}_1 \neq \widetilde{\boldsymbol{\beta}}_2$. □

1.3 Practical Computation of Distributional Centrality Measures

In the main text, we formally define the *effect size analog* as the result of projecting the design matrix $\mathbf{X}$ onto the vector $\mathbf{f} = \boldsymbol{\Psi}^\intercal\mathbf{c}$ via the linear map,

$$\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\boldsymbol{\Psi}^\intercal\mathbf{c} = \mathbf{X}^\dagger\mathbf{f}. \tag{S4}$$

We also assume that the posterior for $\widetilde{\boldsymbol{\beta}}$ is (approximately) multivariate normal with an empirical mean vector $\boldsymbol{\mu}$ and positive semi-definite covariance/precision matrices $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$ estimated via sampling methods. Under these assumptions, we may partition conformably as follows

$$\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \widetilde{\beta}_j \\ \widetilde{\boldsymbol{\beta}}_{-j} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_j \\ \boldsymbol{\mu}_{-j} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_j & \boldsymbol{\sigma}_{-1}^\intercal \\ \boldsymbol{\sigma}_{-j} & \boldsymbol{\Sigma}_{-j} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_j & \boldsymbol{\lambda}_{-j}^\intercal \\ \boldsymbol{\lambda}_{-j} & \boldsymbol{\Lambda}_{-j} \end{pmatrix},$$

where $\widetilde{\beta}_j$, μ_j, σ_j and λ_j are scalars; $\widetilde{\boldsymbol{\beta}}_{-j}$, $\boldsymbol{\mu}_{-j}$, $\boldsymbol{\sigma}_{-j}$, and $\boldsymbol{\lambda}_{-j}$ are $(p-1)$-dimensional vectors; and $\boldsymbol{\Sigma}_{-j}$ and $\boldsymbol{\Lambda}_{-j}$ are $(p-1)\times(p-1)$ positive definite, symmetric matrices. With this partitioning, the Kullback-Leibler divergence (KLD) — summarizing the influence/importance of the j-th variant and measuring the difference between $p(\widetilde{\boldsymbol{\beta}}_{-j} \,|\, \widetilde{\beta}_j)$ and $p(\widetilde{\boldsymbol{\beta}}_{-j})$ — simplifies to the following closed form solution

$$\mathrm{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + \mathrm{tr}(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right], \tag{S5}$$

where $\alpha_j = \boldsymbol{\lambda}_{-j}^{\intercal}\boldsymbol{\Lambda}_{-j}^{-1}\boldsymbol{\lambda}_{-j}$, $\log|\bullet|$ represents the log determinant function for a matrix, and $\text{tr}(\bullet)$ is the matrix trace function. Notice there are a few computationally expensive steps within this derivation. The first involves computing the log determinant and trace of a matrix product. With a reasonably sized data set (i.e. $p \gg 0$), both of these terms with $(1-p)$ remain relatively equal for each marker j and make a negligible contribution to the entire sum. Thus, we begin by simplifying computation to the following

$$\text{KLD}(\widetilde{\beta}_j) \approx \alpha_j(\widetilde{\beta}_j - \mu_j)^2/2. \tag{S6}$$

Next, notice that the KLD still relies on the full precision matrix $\boldsymbol{\Lambda}$. For large p, this is an expensive calculation; however, it only has to be done once and is used for all markers. Lastly, the rate of change parameter α_j depends on the partitioned matrix $\boldsymbol{\Lambda}_{-j}^{-1}$. This requires inverting a $p-1 \times p-1$ matrix separately for each marker j. Based on the assumed projection in Equation (S4), we implement the following procedures to reduce burden and complexity.

Case #1: Calculating the Matrices $\boldsymbol{\Lambda}_{-j}^{-1}$ with $\boldsymbol{p \leq n}$. In this case, the calculation of $\boldsymbol{\Lambda}_{-j}^{-1}$ is not very expensive, and so can be done directly to calculate $\alpha_j = \boldsymbol{\lambda}_{-j}^{\intercal}\boldsymbol{\Lambda}_{-j}^{-1}\boldsymbol{\lambda}_{-j}$. The overall time complexity for this operation is $\mathcal{O}(p^4)$.

Case #2: Calculating the Matrices $\boldsymbol{\Lambda}_{-j}^{-1}$ with $\boldsymbol{p > n}$. This is case is more appropriate for the genetic association studies that we consider in the main text. Here, the calculation of $\boldsymbol{\Lambda}_{-j}^{-1}$ can be reduced to the inverse of an $n \times n$ matrix. Let $\boldsymbol{\Omega} = \sigma(\mathbf{f})$ be the empirical posterior covariance matrix of the estimated functions $\mathbf{f}$. Then from the projection in Equation (S4), we estimate

$$\boldsymbol{\Sigma} = \sigma(\widetilde{\boldsymbol{\beta}}) = \mathbf{X}^{\dagger}\boldsymbol{\Omega}\mathbf{X}^{\dagger\intercal}. \tag{S7}$$

Rather than calculate the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{\dagger}$ directly (which is $p \times p$ in computation), we can alternatively let $\boldsymbol{\Lambda} = \mathbf{X}^{\intercal}\boldsymbol{\Omega}^{\dagger}\mathbf{X}$ (which reduces computation to $n \times n$). This $\boldsymbol{\Lambda}$ satisfies the Moore-Penrose psuedoinverse conditions:

$$\begin{aligned}(i)\quad \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda} &= \mathbf{X}^{\intercal}\boldsymbol{\Omega}^{\dagger}\mathbf{X}\mathbf{X}^{\dagger}\boldsymbol{\Omega}\mathbf{X}^{\dagger\intercal}\mathbf{X}^{T}\boldsymbol{\Omega}^{\dagger}\mathbf{X} \\ &= \mathbf{X}^{\intercal}\boldsymbol{\Omega}^{\dagger}\boldsymbol{\Omega}\boldsymbol{\Omega}^{\dagger}\mathbf{X} \\ &= \mathbf{X}^{\intercal}\boldsymbol{\Omega}^{\dagger}\mathbf{X} \\ &= \boldsymbol{\Lambda}\end{aligned} \qquad \begin{aligned}(ii)\quad \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma} &= \mathbf{X}^{\dagger}\boldsymbol{\Omega}\mathbf{X}^{\dagger\intercal}\mathbf{X}^{T}\boldsymbol{\Omega}^{\dagger}\mathbf{X}\mathbf{X}^{\dagger}\boldsymbol{\Omega}\mathbf{X}^{\dagger\intercal} \\ &= \mathbf{X}^{\dagger}\boldsymbol{\Omega}\boldsymbol{\Omega}^{\dagger}\boldsymbol{\Omega}\mathbf{X}^{\dagger\intercal} \\ &= \mathbf{X}^{\dagger}\boldsymbol{\Omega}\mathbf{X}^{\dagger\intercal} \\ &= \boldsymbol{\Sigma}\end{aligned}$$

as long as $\mathbf{X}$ has linearly independent rows. Next, for this case, we need to compute the rate of change $\alpha_j = \boldsymbol{\lambda}_{-j}^{\intercal}\boldsymbol{\Lambda}_{-j}^{-1}\boldsymbol{\lambda}_{-j}$ for every marker j. Recall that the precision matrix $\boldsymbol{\Lambda} = \mathbf{X}^{\intercal}\boldsymbol{\Omega}^{\dagger}\mathbf{X} = \boldsymbol{\Sigma}^{\dagger}$. We will now find $\mathbf{L} = \boldsymbol{\Lambda}^{1/2}$ such that $\boldsymbol{\Lambda} = \mathbf{L}\mathbf{L}^{\intercal}$. This will allow us to reduce the complexity of repeatedly finding the pseudo-inverse of subsets of this matrix for every marker j. We first find the matrix $(\boldsymbol{\Omega}^{\dagger})^{1/2}$ using a factorization such as SVD. Now, we may define $\mathbf{L} = \mathbf{X}^{\intercal}(\boldsymbol{\Omega}^{\dagger})^{1/2}$. This matrix has dimension $p \times n$. Now that we have computed $\mathbf{L}$, we may partition for every marker j as $\mathbf{L} = [\mathbf{l}_j; \mathbf{L}_{-j}]$. This leads to

$$\boldsymbol{\Lambda}_{-j} = \mathbf{L}_{-j}\mathbf{L}_{-j}^{\intercal}, \quad \boldsymbol{\Lambda}_{-j}^{\dagger} = \mathbf{L}_{-j}^{\intercal\dagger}\mathbf{L}_{-j}^{\dagger} \tag{S8}$$

Using these quantities, we calculate $\alpha_j = \boldsymbol{\lambda}_{-j}^{\intercal}\mathbf{L}_{-j}^{\intercal\dagger}\mathbf{L}_{-j}^{\dagger}\boldsymbol{\lambda}_{-j}$. To compute $\mathbf{L}_{-j}^{\dagger}\boldsymbol{\lambda}_{-j}$, we use QR decomposition and solve the linear equation $\mathbf{L}_{-j}\mathbf{b} = \boldsymbol{\lambda}_{-j}$. This is a fast calculation because the dimension of the matrix L_{-j} is $(p-1) \times (n-1)$. Overall, the time complexity for these computations is $\mathcal{O}(p^3)$.

Note on Low Rank Design Matrices. If $n < p$, and $\text{rank}(\mathbf{X}) = r < n$, we need to modify the above calculations. In these such cases, assume the following matrix decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal,$$

where $\mathbf{U}$ is an $n \times r$ matrix, $\mathbf{D}$ is a $r \times r$ diagonal matrix, and $\mathbf{V}$ is a $p \times r$ matrix. Now by definition via Equation (S7)

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Omega}^*\mathbf{V}^\intercal, \quad \mathbf{\Lambda} = \mathbf{V}^\dagger\mathbf{\Omega}^{*\dagger}\mathbf{V}^{\dagger\intercal} \tag{S9}$$

where $\mathbf{\Omega}^* = \mathbf{D}^{-1}\mathbf{U}^\intercal\mathbf{\Omega}\mathbf{U}\mathbf{D}^{-1}$ is an $r \times r$ positive-definite matrix. The above computational reducing steps will now work for $n < p$ cases with $\mathbf{V}$ replacing $\mathbf{X}^\dagger$, $\mathbf{\Omega}^*$ replacing $\mathbf{\Omega}$, and r replacing n.

1.4 Preprocessing of Real Datasets

We use three real genetic datasets in the present study. The first dataset comes from the Wellcome Trust Case Control Consortium (WTCCC) 1 [5] (`http://www.wtccc.org.uk/`), which initially consisted of 2,938 shared controls with 458,868 SNPs after following the quality control procedures of previous studies [4, 6, 7]. Missing genotypes were further imputed by using the BIMBAM software [8] (`http://www.haplotype.org/bimbam.html`). In the main text, we utilize the real data from chromosome 22 and simulate continuous phenotypes to assess the power of RATE and other commonly used association mapping methods. All polymorphic SNPs with minor allele frequencies (MAFs) above 1% were used in the simulation studies. Exclusively considering this group of individuals and SNPs resulted in a final dataset consisting of n = 2,938 samples and p = 5,747 markers.

The second dataset we consider is small quantitative trait loci (QTL) association mapping study from the Versailles Arabidopsis Stock Center [9] (`http://publiclines.versailles.inra.fr/page/33`). Specifically, this study consists of $n = 403$ F6 plants from a Bay-0 $\times$ Shahdara recombinant inbred lines (RILs) population that were genotyped for $p = 1028$ genetic markers and phenotyped for sixty-three different metabolic traits [10] After pruning the genotypes of variants with near perfect correlation ($r^2 \geq 0.99$), we obtained a final set of $p = 524$ markers. In the main text, we limit the scope of our analysis to six biochemical content measurements including: allyl, Indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 8-methylthiooctyl (MT8), and 3-hydroxypropyl (OHP3).

The third dataset is from the Wellcome Trust Centre for Human Genetics (`http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml`). This study originally contains $n = 1,904$ heterogenous stock of mice from 85 families (all descending from eight inbred progenitor strains) [11], and 131 quantitative traits that are classified into 6 broad categories including behavior, diabetes, asthma, immunology, haematology, and biochemistry (`http://mtweb.cs.ucl.ac.uk/mus/www/GSCAN/index.shtml/index.old.shtml`). In the main text, we focused on three specific phenotypes: body weight (`Glucose.BodyWeight`), percentage of CD8+ cells (`Imm.PctCD8`), and high-density lipoprotein content (`Biochem.HDL`). All phenotypes were previously corrected for sex, age, body weight, season, and year effects [11]. A total of 12,226 autosomal SNPs were available for all mice. For individuals with missing genotypes, we imputed values by the mean genotype of that SNP in their corresponding family. All polymorphic SNPs with minor allele frequency above 5% were used for association mapping.

Other Resources. Significant loci in the RIL QTL study were mapped to the nearest gene(s) using the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov/gene/`). In the heterogenous stock of mice data applications, SNPs are mapped to the closest neighboring gene(s) using the Mouse Genome Informatics database (`http://www.informatics.jax.org`).

1.5 Dissection of Phenotypic Variance in Real Datasets

In the two real data applications presented in the main text, we detail the proportions of phenotypic variances explained by different orders of genetic effects. Namely, investigate whether the analyzed traits can be best be described by additivity, pairwise interactions, third order interactions, or (in the case of the heterogenous stock of mice) common environmental effects. To do so, we consider following linear mixed model (LMM) for both the *Arabidopsis* QTL study and the mice GWAS, respectively:

$$\mathbf{y} = \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \tau^2\mathbf{I}) \tag{S10}$$

where $\mathbf{g}_1 \sim \text{MVN}(\mathbf{0}, \sigma_1^2\mathbf{K})$ is the linear effects component; $\mathbf{g}_2 \sim \text{MVN}(\mathbf{0}, \sigma_2^2\mathbf{K}^2)$ is the pairwise interaction component; and $\mathbf{g}_3 \sim \text{MVN}(\mathbf{0}, \sigma_3^2\mathbf{K}^3)$ is the third order interaction component. Here, we let $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \sigma_2^2, \sigma_3^2\}$ be the corresponding random effect variance terms. The matrix $\mathbf{I}$ is an identity matrix. The covariance matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^\intercal/p$ is the conventional (linear) genetic relatedness matrix [6, 12–16]. The covariance matrix $\mathbf{K}^2 = \mathbf{K} \circ \mathbf{K}$ represents a pairwise interaction relationship matrix and is obtained by using the Hadamard product (i.e. the squaring of each element) of the linear kernel matrix with itself [17–19]. Similarly, the matrix $\mathbf{K}^3 = \mathbf{K} \circ \mathbf{K} \circ \mathbf{K}$ represents a third order interaction relationship matrix (i.e. the cubing of each element). The key purpose of these analyses is to directly estimate the contribution of nonlinear genetic effects across the different phenotypes and traits that we consider. We quantify these contributions by examining the proportion of phenotypic variance explained (pPVE) using the following equation [4, 6, 7, 20, 21]:

$$\text{pPVE}_r \propto \frac{\widehat{\sigma}_j^2}{n}\text{tr}(\boldsymbol{\Sigma}_j), \qquad \sum_j \text{pPVE}_j = 1, \qquad j = 1, \ldots, 3$$

where $\boldsymbol{\Sigma} = [\mathbf{K}, \mathbf{K}^2, \mathbf{K}^3]$. In the Tables S4 and S6, we detail the estimates of the pPVEs (and their standard errors) corresponding to the random effect variance terms $\widehat{\boldsymbol{\sigma}}^2 = \{\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\sigma}_3^2\}$. Intuitively, the variance component that explains the greatest proportion of the overall PVE then represents the most influential effect onto that particular phenotypic response. The LMM defined in Equation (S10) is implemented by using the `-vc 1` argument within the GEMMA open source software [15] (`http://www.xzlab.org/software.html`). Briefly, this configuration fits variance component models by using an MQS algorithm [21], which is based on a combination between a method of moments (MoM) [22] and minimal norm quadratic unbiased estimation criteria (MINQUE) [23].

2 Supplementary Algorithmic Overview

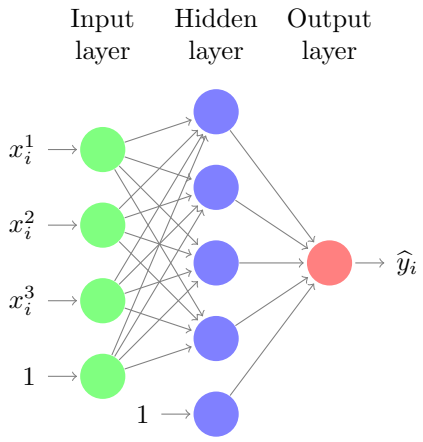
Algorithm 1 Gaussian Process Regression (GPR)

Select a positive definite covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are n-dimensional vectors from the design matrix.
Construct the $n \times n$ covariance matrix $\mathbf{K}$.
Define the full model where $\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$.
Specify the prior distributions $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ and $\tau^2 \sim \text{Scale-Inv-}\chi(a, b)$.
Run the Gibbs Sampler (T Iterations).
for $t = 1 \to T$ **do**
    $\mathbf{f} \,|\, \tau^2, \mathbf{y} \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*)$ where $\mathbf{m}^* = \tau^{-2}\mathbf{V}^*\mathbf{y}$ and $\mathbf{V}^* = \tau^2(\tau^2\mathbf{K} + \mathbf{I})^{-1}$;
    $\tau^2 \,|\, \mathbf{f}, \mathbf{y} \sim \text{Scale-Inv-}\chi^2(a^*, b^*)$ where $a^* = a + n$ and $b^* = a^{*-1}[ab + (\mathbf{y} - \mathbf{f})^\intercal(\mathbf{y} - \mathbf{f})]$;
    $\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{f}$.
end for
Calculate the empirical mean, covariance, and precision of the posterior distribution $p(\widetilde{\boldsymbol{\beta}} \,|\, \mathbf{y})$ as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$, respectively.
Compute the centrality of every p predictor via Kullback-Leibler Divergence (KLD).
for $j = 1 \to p$ **do**
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + \text{tr}(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
end for
Scale each centrality measure for the p predictors to determine their relative importance.
for $j = 1 \to p$ **do**
    $\text{RATE}(\widetilde{\beta}_j) = \text{KLD}(\widetilde{\beta}_j) / \sum \text{KLD}(\widetilde{\beta}_\ell)$.
end for

Algorithm 2 Bayesian Kernel Ridge Regression (BKRR)

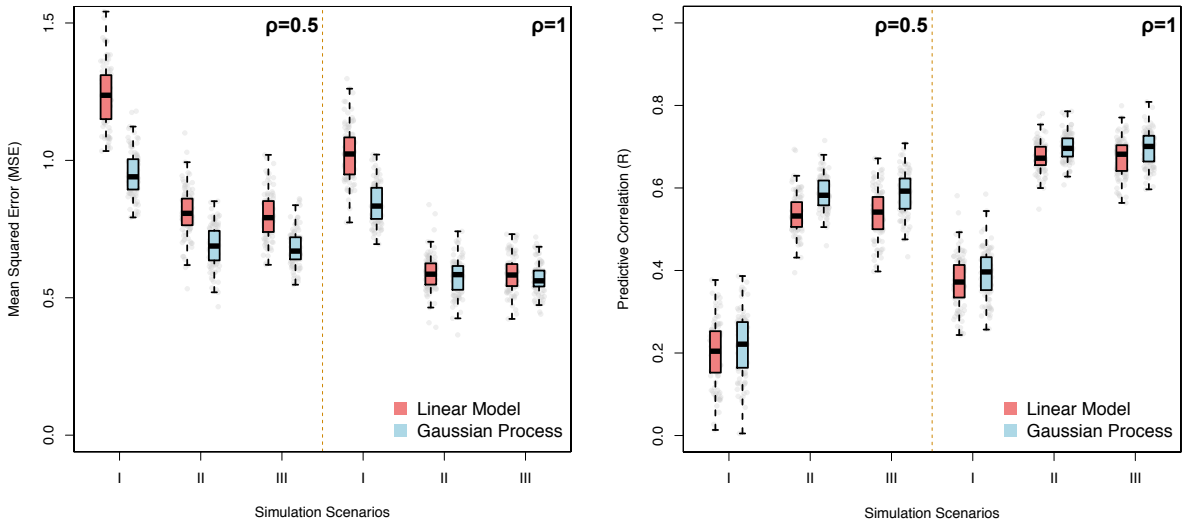
Select a positive definite covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are n-dimensional vectors from the design matrix.
Construct the $n \times n$ covariance matrix $\mathbf{K}$.
Define the full model where $\mathbf{y} = \mathbf{K}\boldsymbol{\vartheta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$.
Specify the prior distributions $\boldsymbol{\vartheta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{K}^{-1})$ and $\sigma^2, \tau^2 \sim$ Scale-Inv-$\chi(a, b)$.
Run the Gibbs Sampler (T Iterations).
for $t = 1 \to T$ **do**
    $\boldsymbol{\vartheta} \mid \sigma^2, \tau^2, \mathbf{y} \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*)$ with $\mathbf{m}^* = \tau^{-2}\mathbf{V}^*\mathbf{K}^\intercal\mathbf{y}$ and $\mathbf{V}^* = \tau^2\sigma^2(\tau^2\mathbf{K}^{-1} + \sigma^2\mathbf{I})^{-1}$;
    $\sigma^2 \mid \boldsymbol{\vartheta}, \tau^2, \mathbf{y} \sim$ Scale-inv-$\chi^2(a_\sigma^*, b_\sigma^*)$ where $a_\sigma^* = a + q$ and $b_\sigma^* = a_\sigma^{*-1}(ab + \boldsymbol{\vartheta}^\intercal\mathbf{K}^{-1}\boldsymbol{\vartheta})$;
    $\tau^2 \mid \boldsymbol{\vartheta}, \sigma^2, \mathbf{y} \sim$ Scale-inv-$\chi^2(a_\tau^*, b_\tau^*)$ where $a_\tau^* = a+n$ and $b_\tau^* = a_\tau^{*-1}(ab+\boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon} = \mathbf{y}-\mathbf{K}\boldsymbol{\vartheta}$;
    $\widetilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{K}\boldsymbol{\vartheta}$.
end for
Calculate the empirical mean, covariance, and precision of the posterior distribution $p(\widetilde{\boldsymbol{\beta}} \mid \mathbf{y})$ as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$, respectively.
Compute the centrality of every p predictor via Kullback-Leibler Divergence (KLD).
for $j = 1 \to p$ **do**
    $\mathrm{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + \mathrm{tr}(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
end for
Scale each centrality measure for the p predictors to determine their relative importance.
for $j = 1 \to p$ **do**
    $\mathrm{RATE}(\widetilde{\beta}_j) = \mathrm{KLD}(\widetilde{\beta}_j) / \sum \mathrm{KLD}(\widetilde{\beta}_\ell)$.
end for

Algorithm 3 Bayesian Neural Network (BNN)



Specify the architecture of the neural network (e.g. see above), operating over input/output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Denote the output of the network for a given x_i as $\widehat{y}_i$.
Specify a prior distribution $\pi(\cdot)$ over all parameters (e.g. weights and biases) in the network, summarized in a vector $\boldsymbol{\theta}$.
Use an MCMC sampler or any approximate Bayesian method to obtain a set of T samples $\{\widehat{\mathbf{y}}^{(t)}\}_{t=1}^T$ from the posterior predictive distribution $p(y_1^*, \ldots, y_n^* \,|\, \mathcal{D})$.
for $t = 1 \to T$ **do**
    $\widetilde{\boldsymbol{\beta}}^{(t)} = \mathbf{X}^\dagger \widehat{\mathbf{y}}^{(t)}$.
end for
Using the samples for $\widetilde{\boldsymbol{\beta}}$, calculate the empirical mean, covariance, and precision of the posterior distribution $p(\widetilde{\boldsymbol{\beta}} \,|\, \mathcal{D})$ as $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$, respectively.
Compute the centrality of every j predictor via Kullback-Leibler Divergence (KLD).
for $j = 1 \to p$ **do**
    $\mathrm{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + \mathrm{tr}(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
end for
Scale each centrality measure for the p predictors to determine their relative importance.
for $j = 1 \to p$ **do**
    $\mathrm{RATE}(\widetilde{\beta}_j) = \mathrm{KLD}(\widetilde{\beta}_j) / \sum \mathrm{KLD}(\widetilde{\beta}_\ell)$.
end for

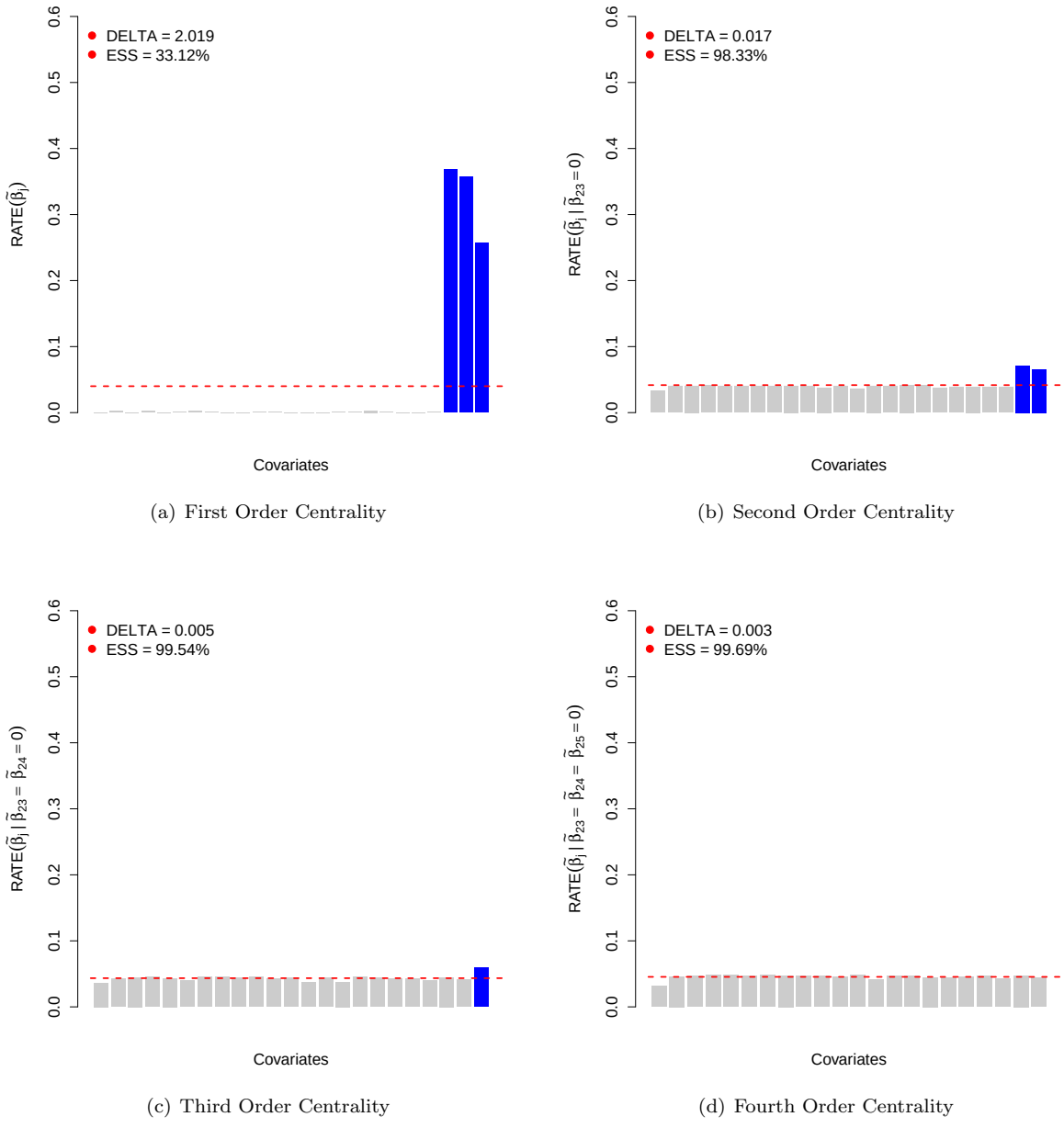
3 Supplementary Figures



(a) Mean Squared Error (MSE)

(b) Predictive Correlation (R)

Figure S1. Comparisons of the out-of-sample predictive mean squared errors (MSE) and predictive correlations (R) for the linear regression model using the standard OLS estimates and the GP regression method using the effect size analogue. Scenario I corresponds to phenotypic outcomes being generated via a standard linear model without population structure. Scenarios II and III introduce population stratification effects by allowing the top 5 and 10 genotype PCs to make up 30% of the phenotypic variance, respectively. Here, the broad-sense heritability is set to $\mathrm{H}^2 = 0.3$ with control parameter $\rho = \{0.5, 1\}$, which is used to determine the proportion of signal that is contributed by interaction effects. Figure (a) corresponds to MSE results, while Figure (b) depicts results for predictive correlation. Results are based on 100 replicates in each case.



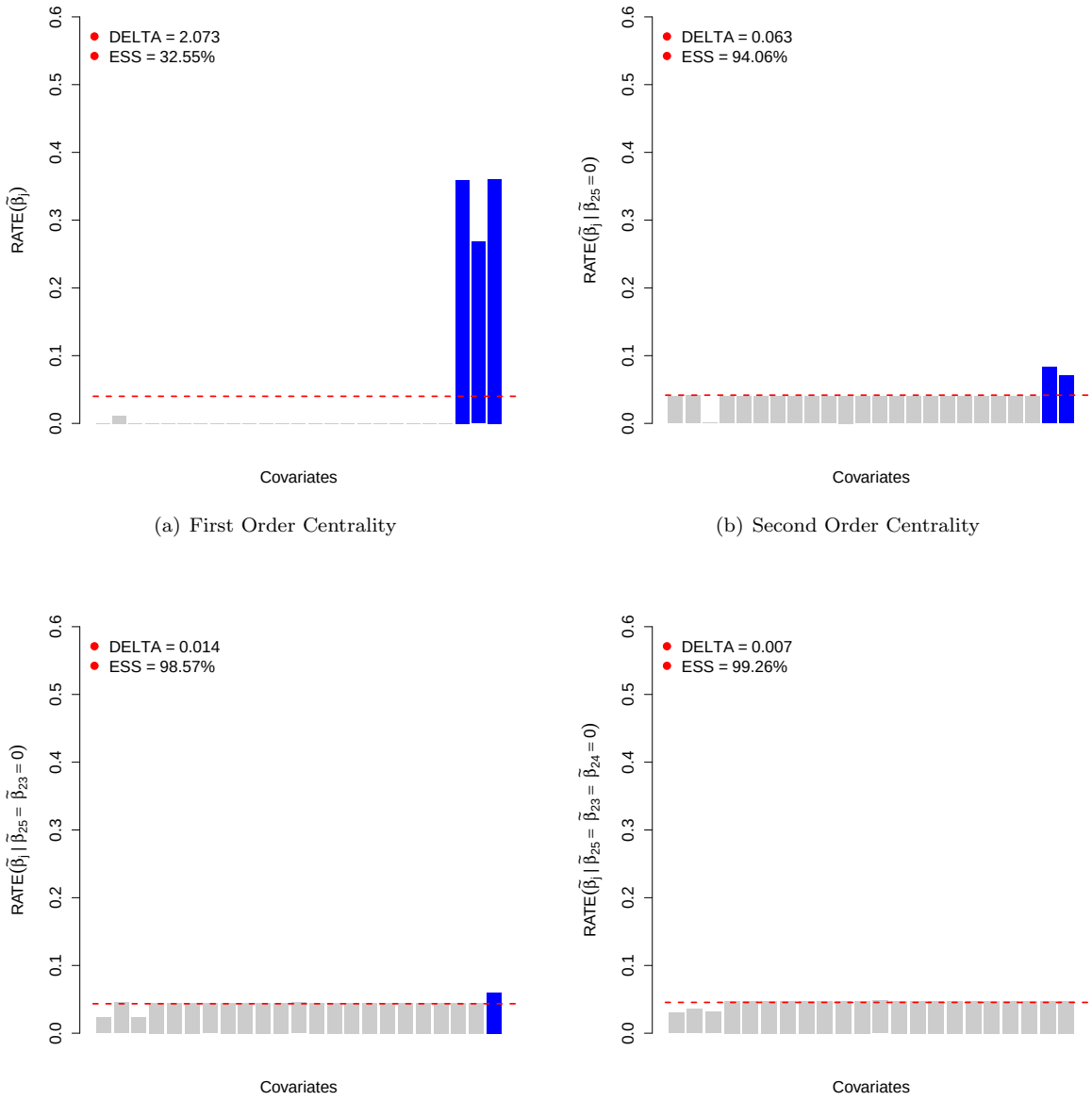
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S2. Orders of distributional centrality via RATE measures in the presence of interaction effects. These are simple proof of concept simulations with broad-sense heritability level $H^2 = 0.6$ and $\rho = 0.5$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Data are simulated such that the effects of only the last three genetic variants $j^* = \{23, 24, 25\}$ (blue) are nonzero. The dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) shows the first order centrality across all markers; (b)-(d) show results when the most significantly associated variants are iteratively nullified. Uniformity check values are also reported: (i) the entropic difference Δ, and (ii) the corresponding empirical effective sample size (ESS) estimates.



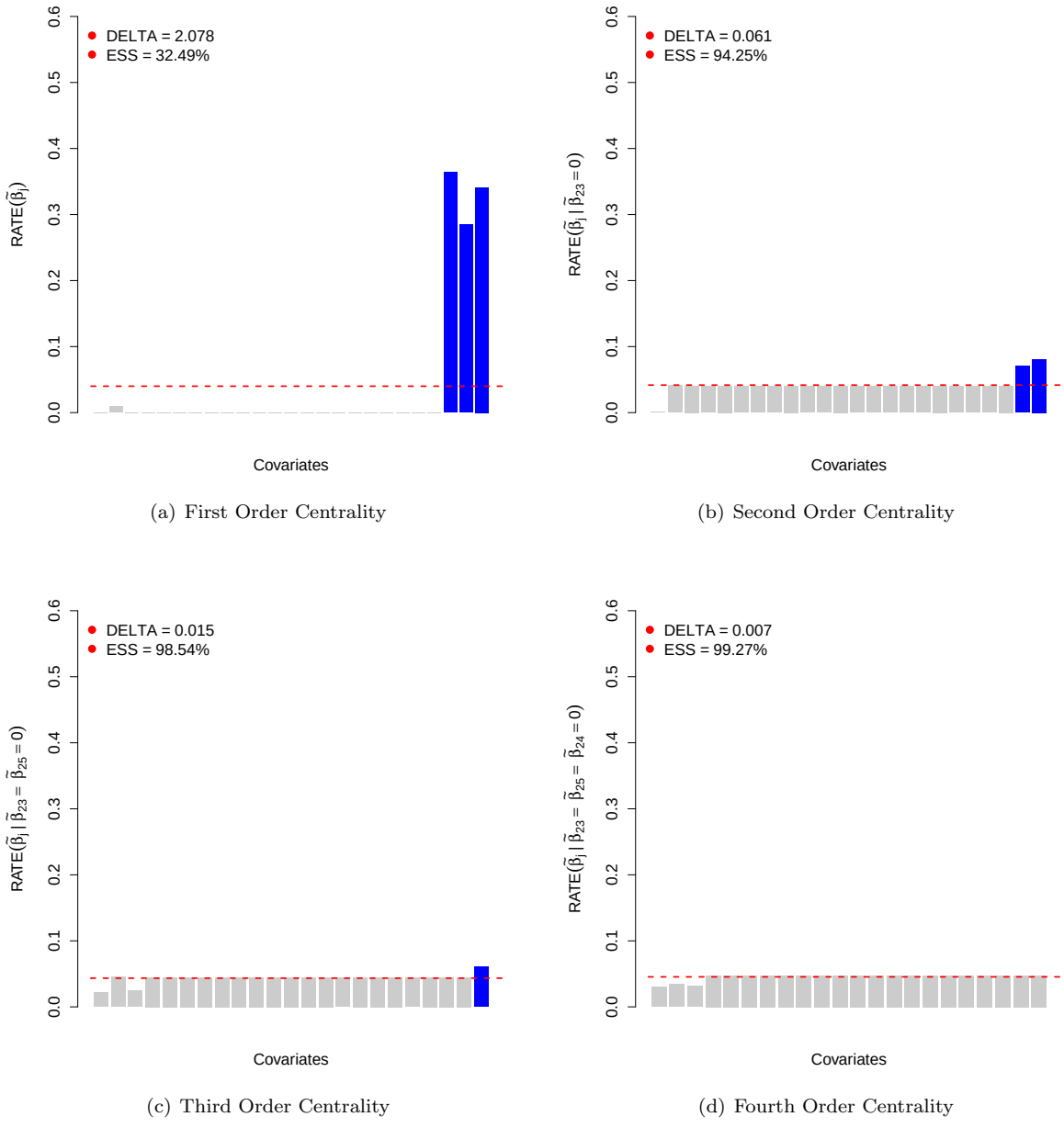
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S3. Orders of distributional centrality via RATE measures when associated variants are strongly correlated with non-associated markers. These are simple proof of concept simulations with broad-sense heritability level $\mathrm{H}^2 = 0.6$ and $\rho = 0.5$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Data are simulated such that the effects of only the last three genetic variants $j^* = \{23, 24, 25\}$ (blue) are nonzero; yet, they are positively correlated ($R = 0.9$) with nonsignificant markers #1-3, respectively. The dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) shows the first order centrality across all markers; (b)-(d) show results when the most significantly associated variants are iteratively nullified. Uniformity check values are also reported: (i) the entropic difference Δ, and (ii) the corresponding empirical effective sample size (ESS) estimates.



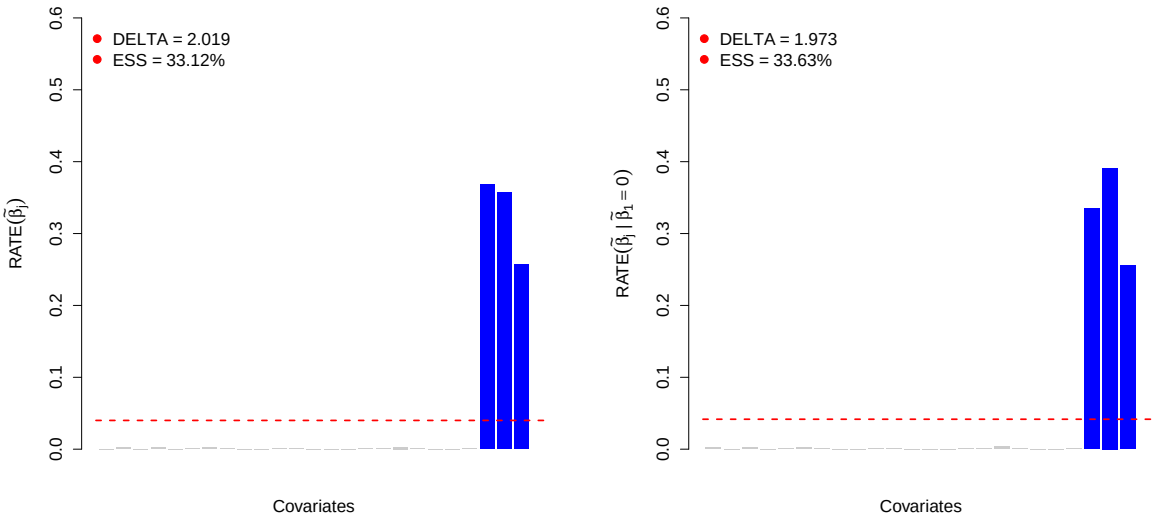
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

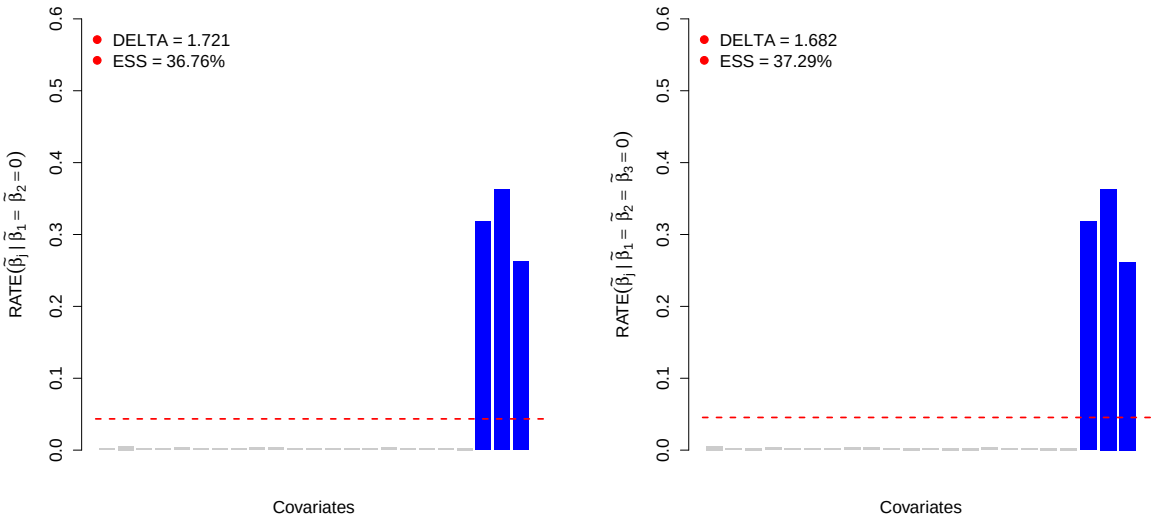
(d) Fourth Order Centrality

Figure S4. Orders of distributional centrality via RATE measures when associated variants are strongly correlated with non-associated markers. These are simple proof of concept simulations with broad-sense heritability level $\mathrm{H}^2 = 0.6$ and $\rho = 1$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Data are simulated such that the effects of only the last three genetic variants $j^* = \{23, 24, 25\}$ (blue) are nonzero; yet, they are positively correlated ($R = 0.9$) with nonsignificant markers #1-3, respectively. The dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) shows the first order centrality across all markers; (b)-(d) show results when the most significantly associated variants are iteratively nullified. Uniformity check values are also reported: (i) the entropic difference Δ, and (ii) the corresponding empirical effective sample size (ESS) estimates.



(a) First Order Centrality

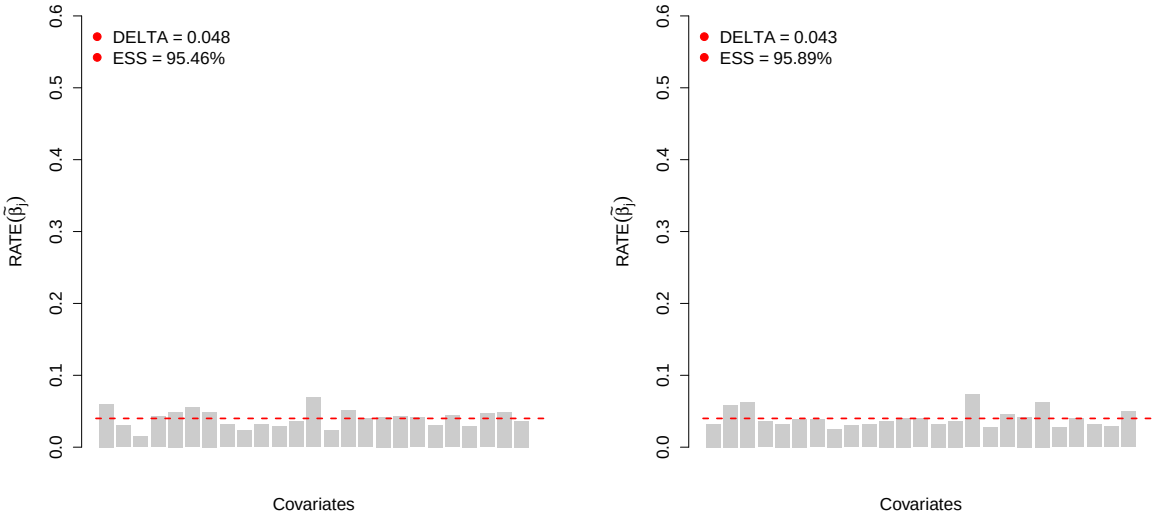
(b) Second Order Centrality



(c) Third Order Centrality

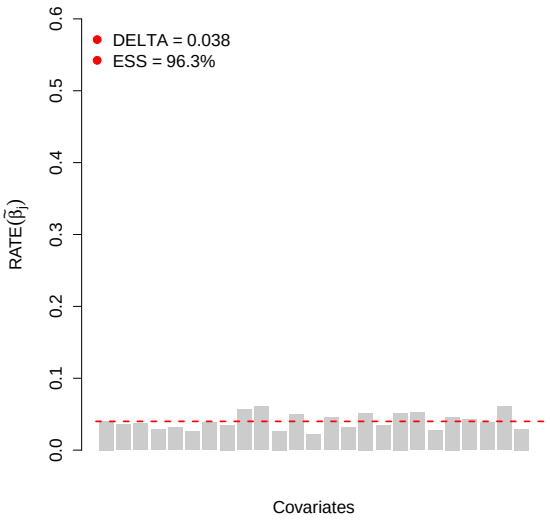
(d) Fourth Order Centrality

Figure S5. Orders of distributional centrality via RATE measures when non-associated variants are deemed significant. These are simple proof of concept simulations with broad-sense heritability level $\mathrm{H}^2 = 0.6$ and $\rho = 0.5$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Data are simulated such that the effects of only the last three genetic variants $j^* = \{23, 24, 25\}$ (blue) are nonzero. The dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figure (a) shows the first order centrality across all markers; (b)-(d) show the results when nonsignificant markers #1-3 are iteratively nullified. Uniformity check values are also reported: (i) the entropic difference Δ, and (ii) the corresponding empirical effective sample size (ESS) estimates.

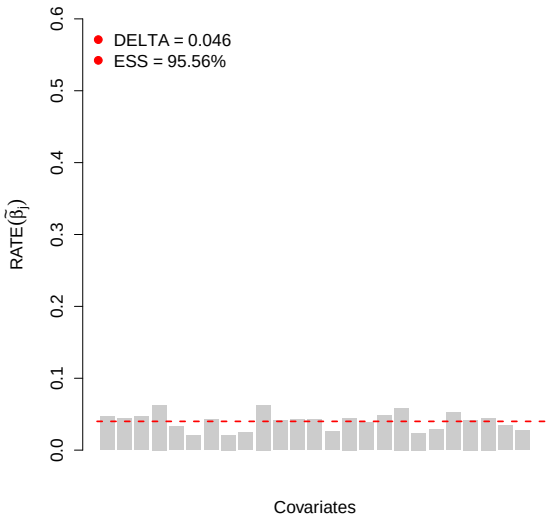


(a) First Order Centrality

(b) Second Order Centrality

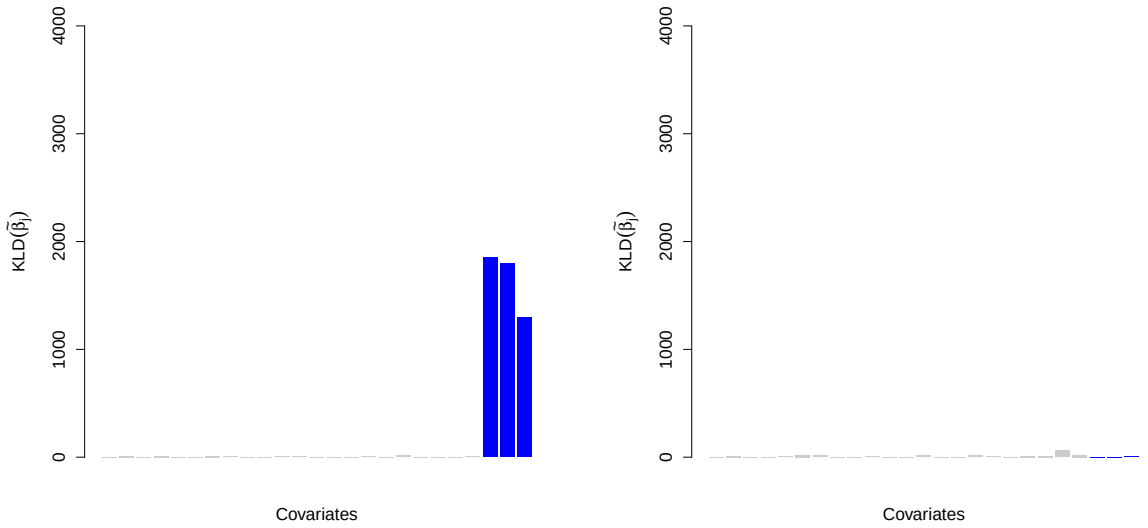


(c) Third Order Centrality



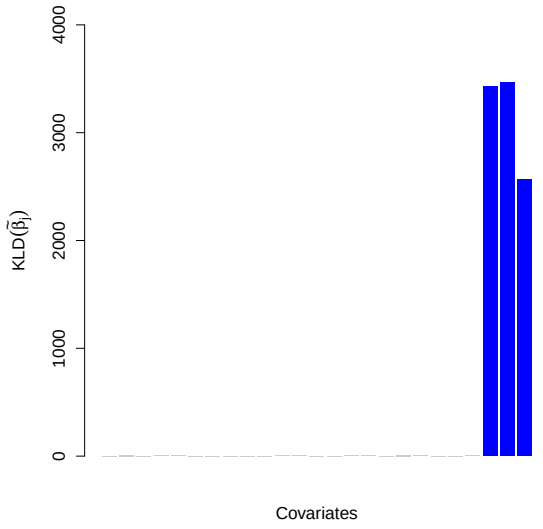
(d) Fourth Order Centrality

Figure S6. Orders of distributional centrality via RATE measures under the null hypothesis that all variants in the data contribute equally to the heritability of a trait. These are simple proof of concept simulations with broad-sense heritability level $H^2 = 0.6$. Data are simulated such that the effects of all genetic variants are the same. The dashed line is drawn at the level of relative equivalence (i.e. $1/p$). Figures (a)-(d) show results for four different randomly generated datasets. Uniformity check values are also reported: (i) the entropic difference Δ, and (ii) the corresponding empirical effective sample size (ESS) estimates.

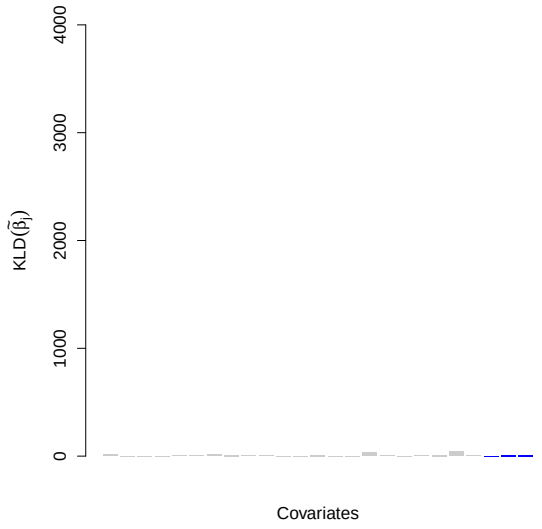


(a) First Order Centrality ($\rho = 0.5$)

(b) Permuted Results ($\rho = 0.5$)

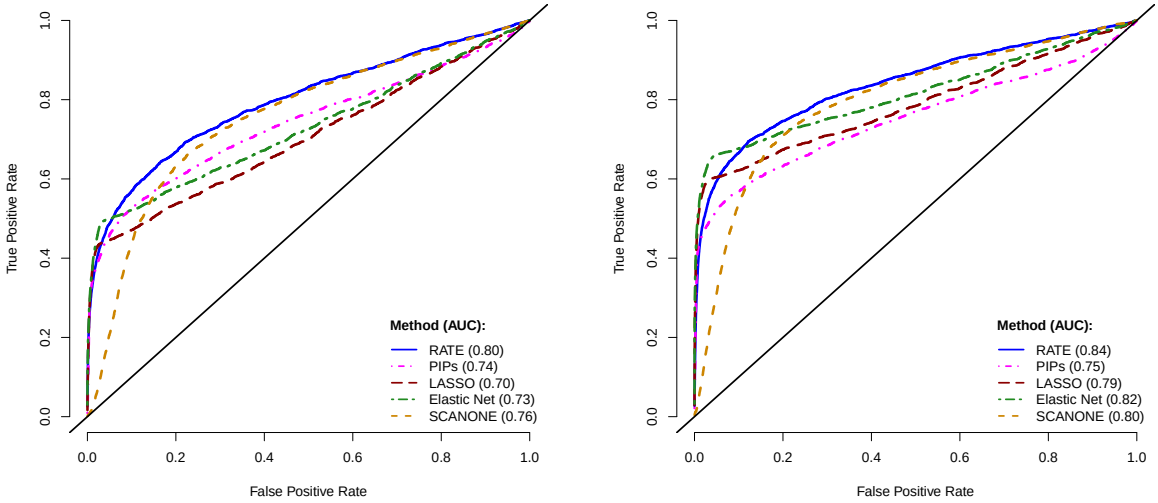


(c) First Order Centrality ($\rho = 1$)



(d) Permuted Results ($\rho = 1$)

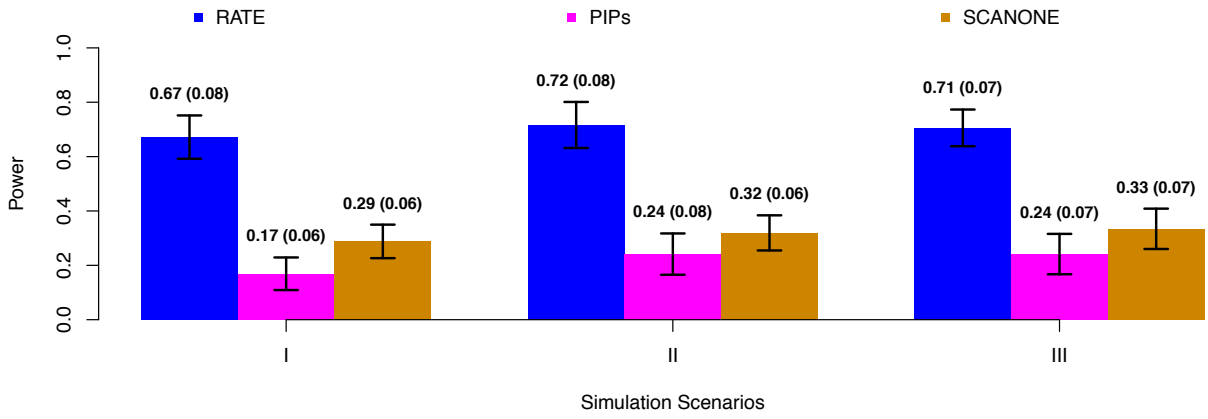
Figure S7. Orders of distributional centrality via raw and unscaled Kullback-Leibler divergence (KLD) measures after phenotypes have been permuted. These are simple proof of concept simulations with broad-sense heritability level $\mathrm{H}^2 = 0.6$ and $\rho = \{0.5, 1\}$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Data are simulated such that the effects of only the last three genetic variants $j^* = \{23, 24, 25\}$ (blue) are nonzero. Figures (a) and (c) show the first order centrality across all markers; (b) and (d) show comparative results when the phenotypes have been permuted once.



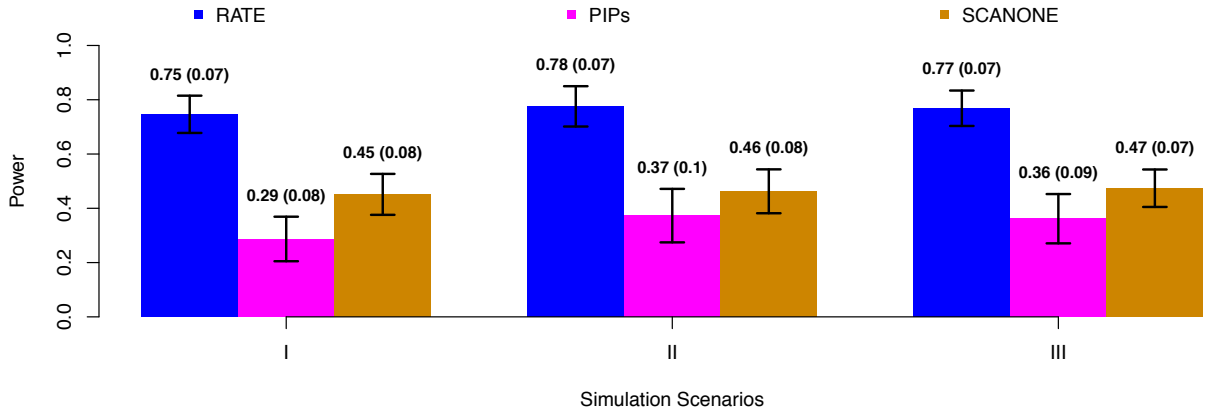
(a) Scenario III ($\rho = 0.5$)

(b) Scenario III ($\rho = 1$)

Figure S8. Power analysis for prioritizing genetic variants in the presence of population stratification effects (Top 10 PCs). Phenotypes are simulated with broad-sense heritability level $H^2 = 0.3$ with control parameter $\rho = \{0.5, 1\}$ in Figures (a) and (b), respectively. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Compared approaches include Gaussian process regression with RATE (blue), Bayesian variable selection with a spike and slab prior (PIPs) (pink), lasso regression (red), the elastic net (green), and the SCANONE method (orange). Area under the curve (AUC) is reported to facilitate comparisons. Results are based on 100 replicates in each case, where data is created under simulation model (ii) with the top 10 genotype PCs.

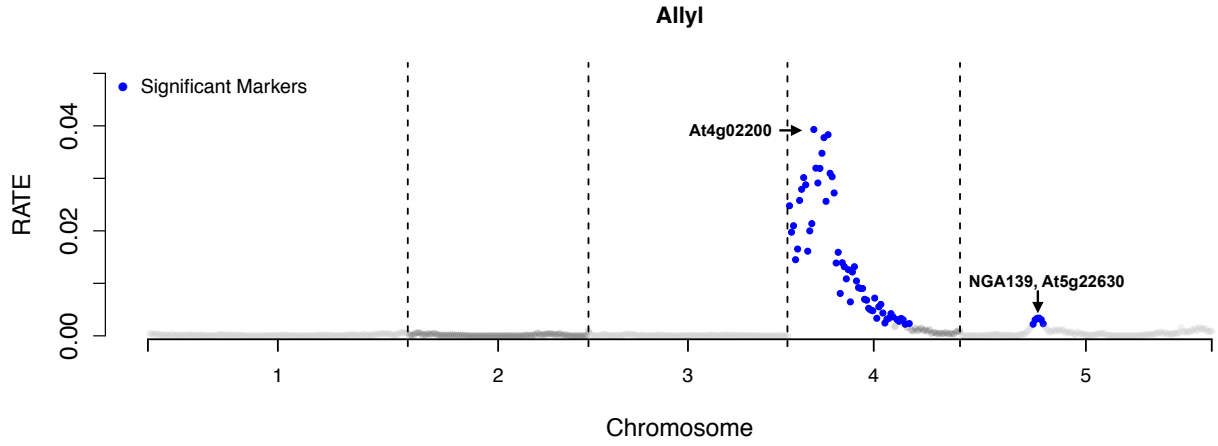


(a) Interaction Case ($\rho = 0.5$)

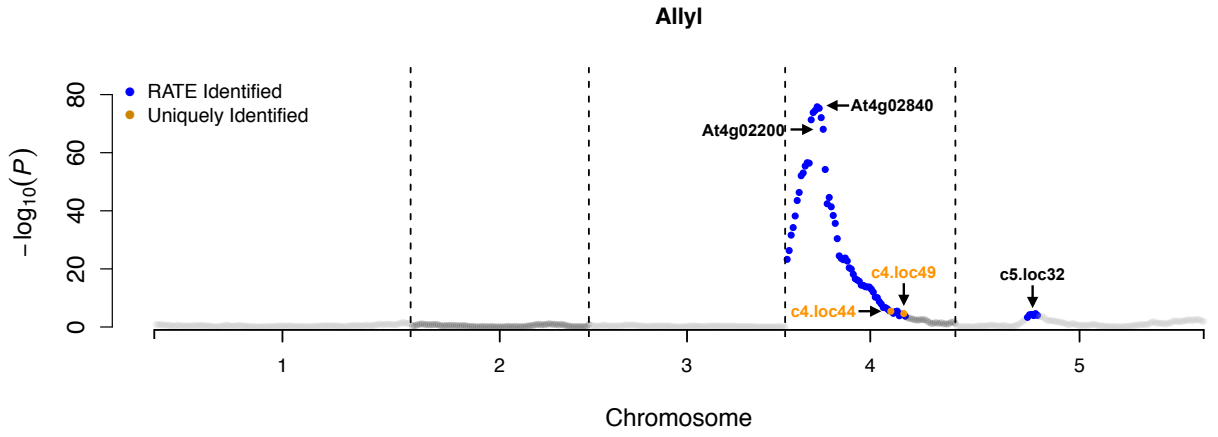


(b) Additive Case ($\rho = 1$)

Figure S9. Power analysis for prioritizing causal variants under the "optimal" model criterion. Phenotypes are simulated with broad-sense heritability level $H^2 = 0.3$ with control parameter $\rho = \{0.5, 1\}$ in Figures (a) and (b), respectively. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Criterions considered include RATEs $> 1/p$ (blue), the Bayesian "median probability model" (pink) (i.e. PIPs > 0.5), and the multiple testing corrected SCANONE method $P < 8.33 \times 10^{-6}$ (orange). Scenario I corresponds to phenotypic outcomes being generated via simulation model (i) without population structure. Scenarios II and III introduce population stratification effects with simulation model (ii) by allowing the top 5 and 10 genotype PCs to make up 30% of the phenotypic variance, respectively. Results are based on 100 replicates in each case.

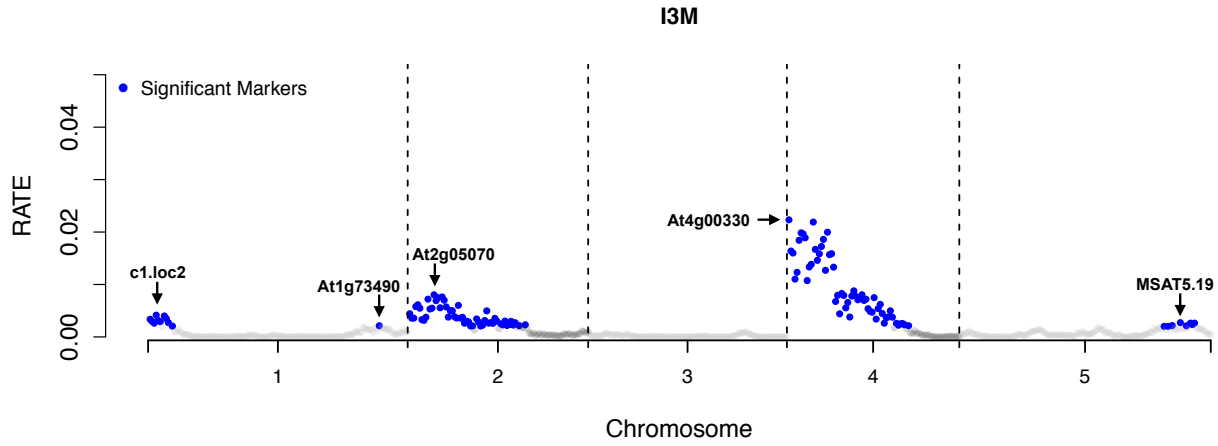


(a) Gaussian Process Regression

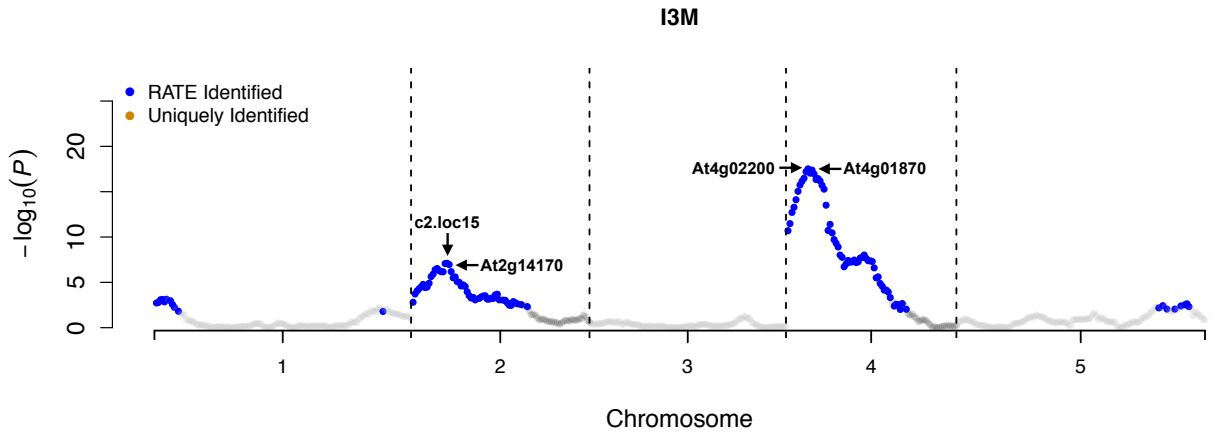


(b) SCANONE

Figure S10. Genetic map wide scan for the allyl content metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) Gaussian process regression with RATE and (b) SCANONE (orange). Significant markers are determined by RATE($\widetilde{\beta}$) $> 1/p$ and $P < 9 \times 10^{-5}$, respectively. The latter represents the genome-wide Bonferroni-corrected significance threshold. To ease the comparisons, points in blue represent genetic markers with significant distributional centrality measures. Markers labeled in color were not found by RATE.

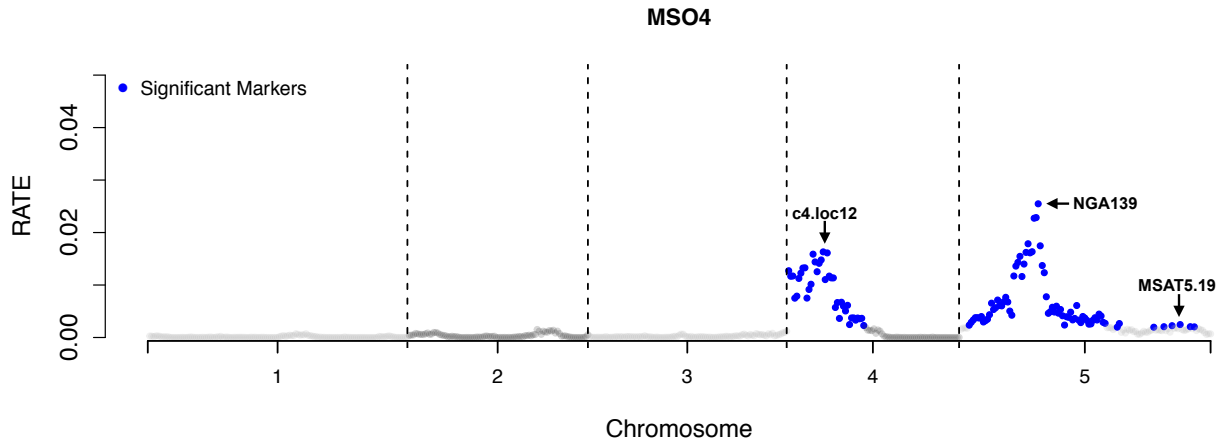


(a) Gaussian Process Regression

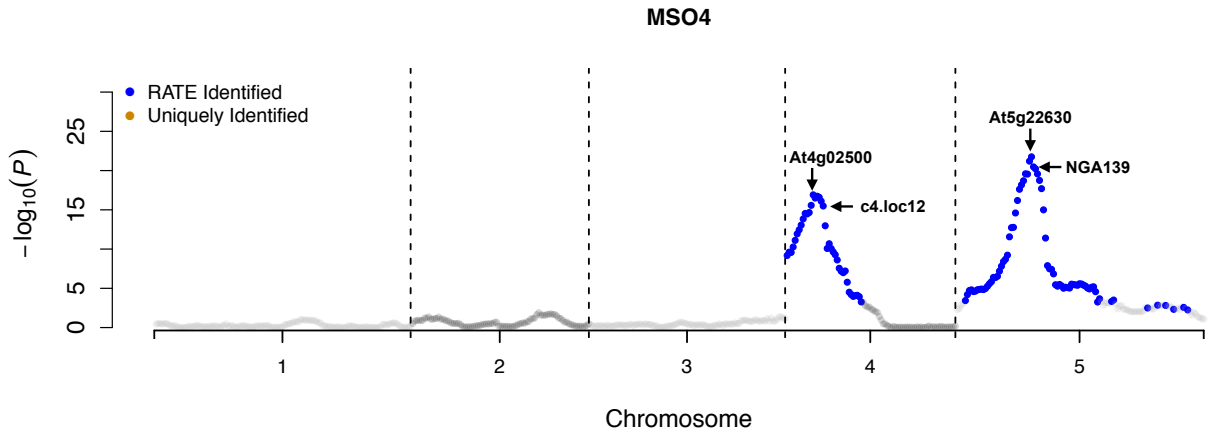


(b) SCANONE

Figure S11. Genetic map wide scan for the indol-3-ylmethyl (I3M) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) Gaussian process regression with RATE and (b) SCANONE (orange). Significant markers are determined by $\text{RATE}(\widetilde{\beta}) > 1/p$ and $P < 9 \times 10^{-5}$, respectively. The latter represents the genome-wide Bonferroni-corrected significance threshold. To ease the comparisons, points in blue represent genetic markers with significant distributional centrality measures. Markers labeled in color were not found by RATE.

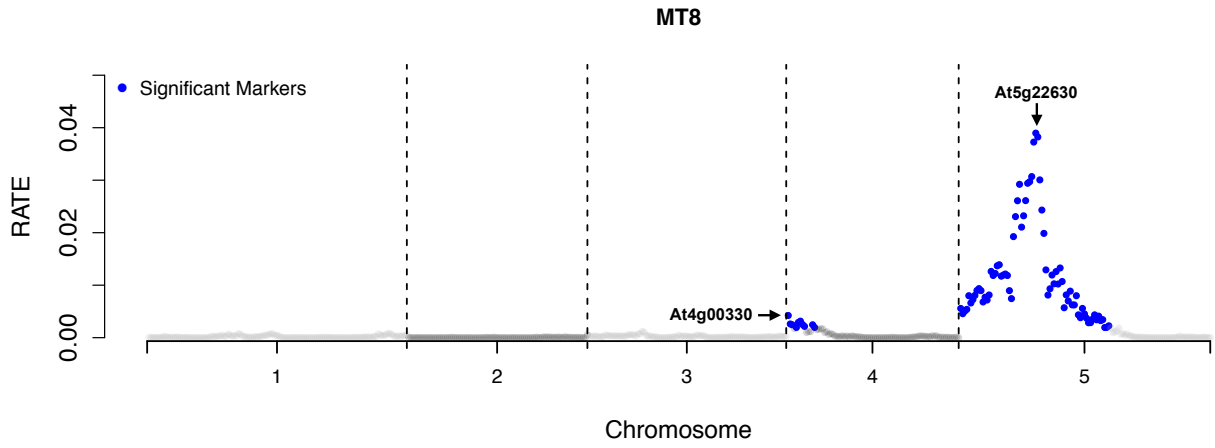


(a) Gaussian Process Regression

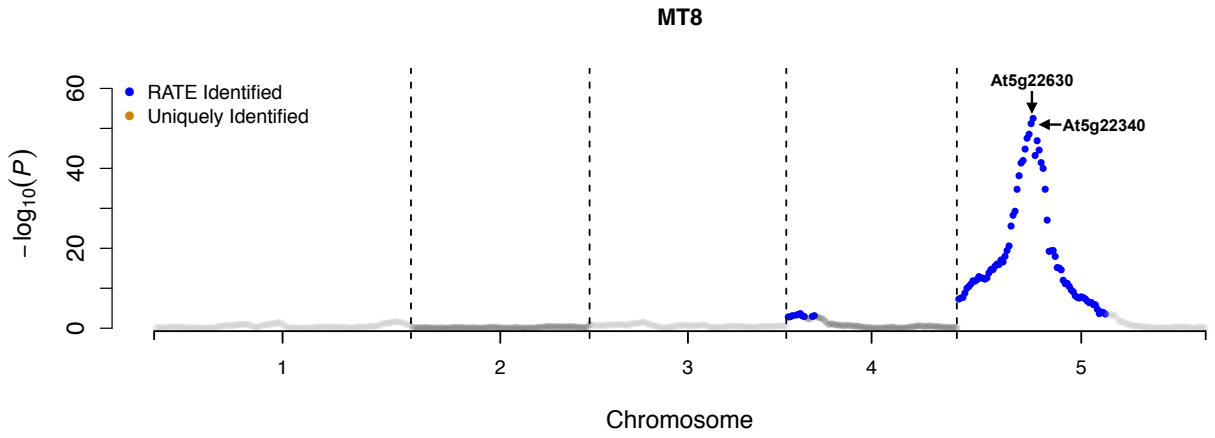


(b) SCANONE

Figure S12. Genetic map wide scan for the 4-methylsulfinylbutyl (MSO4) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) Gaussian process regression with RATE and (b) SCANONE (orange). Significant markers are determined by RATE($\widetilde{\beta}$) $> 1/p$ and $P < 9 \times 10^{-5}$, respectively. The latter represents the genome-wide Bonferroni-corrected significance threshold. To ease the comparisons, points in blue represent genetic markers with significant distributional centrality measures. Markers labeled in color were not found by RATE.

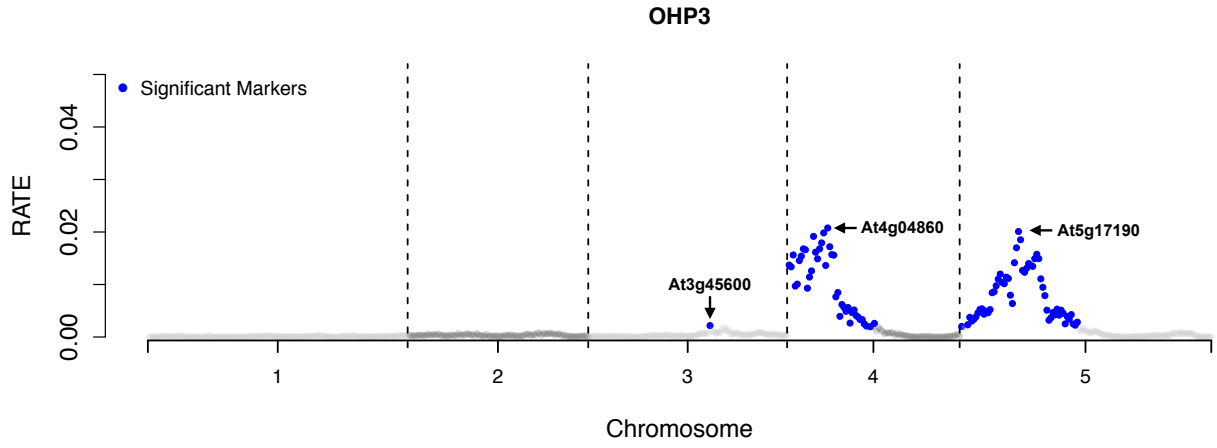


(a) Gaussian Process Regression

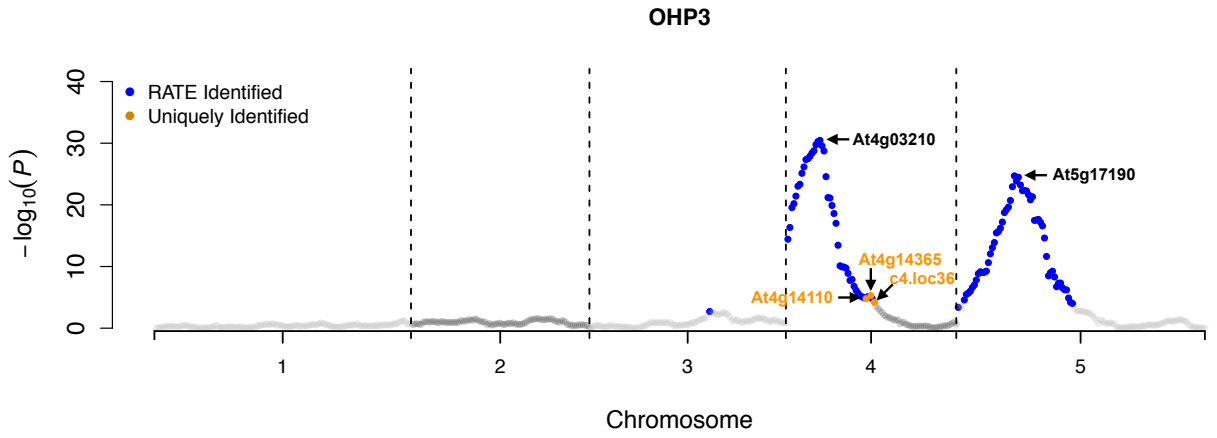


(b) SCANONE

Figure S13. Genetic map wide scan for the 8-methylthiooctyl (MT8) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) Gaussian process regression with RATE and (b) SCANONE (orange). Significant markers are determined by $\text{RATE}(\widetilde{\beta}) > 1/p$ and $P < 9 \times 10^{-5}$, respectively. The latter represents the genome-wide Bonferroni-corrected significance threshold. To ease the comparisons, points in blue represent genetic markers with significant distributional centrality measures. Markers labeled in color were not found by RATE.



(a) Gaussian Process Regression



(b) SCANONE

Figure S14. Genetic map wide scan for the 3-hydroxypropyl (OHP3) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) Gaussian process regression with RATE and (b) SCANONE (orange). Significant markers are determined by $\text{RATE}(\widetilde{\beta}) > 1/p$ and $P < 9 \times 10^{-5}$, respectively. The latter represents the genome-wide Bonferroni-corrected significance threshold. To ease the comparisons, points in blue represent genetic markers with significant distributional centrality measures. Markers labeled in color were not found by RATE.

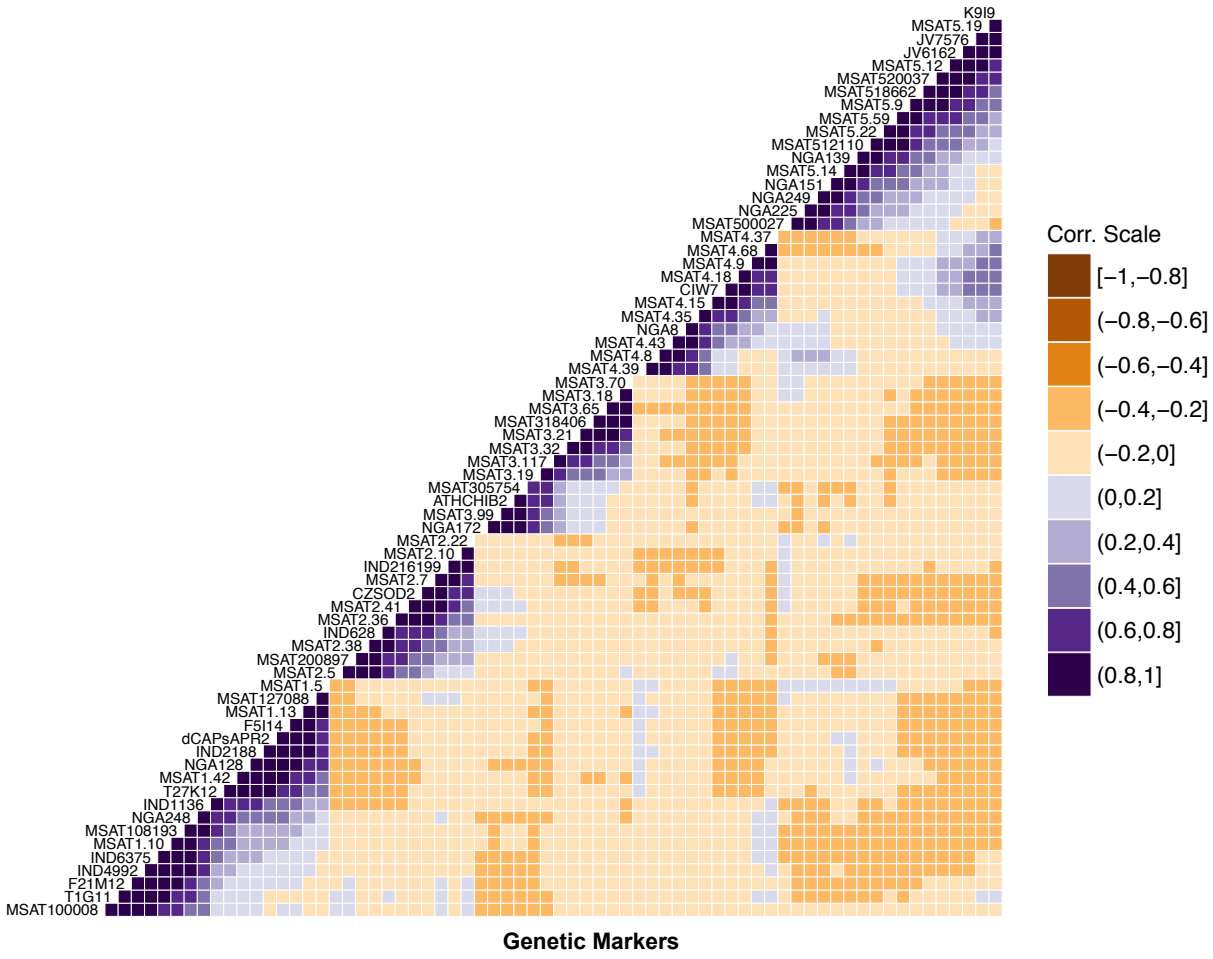


Figure S15. Lower-triangular heat map illustrating the correlation structure for a proportion of the genotyped markers in the *Arabidopsis thaliana* QTL mapping study. The legend represents a correlation scale on an [-1,1] interval that has been evenly divided into ten shorter subintervals. The main takeaway here is that there appears to be an underlying covarying structure between groups of markers located on different chromosomes.

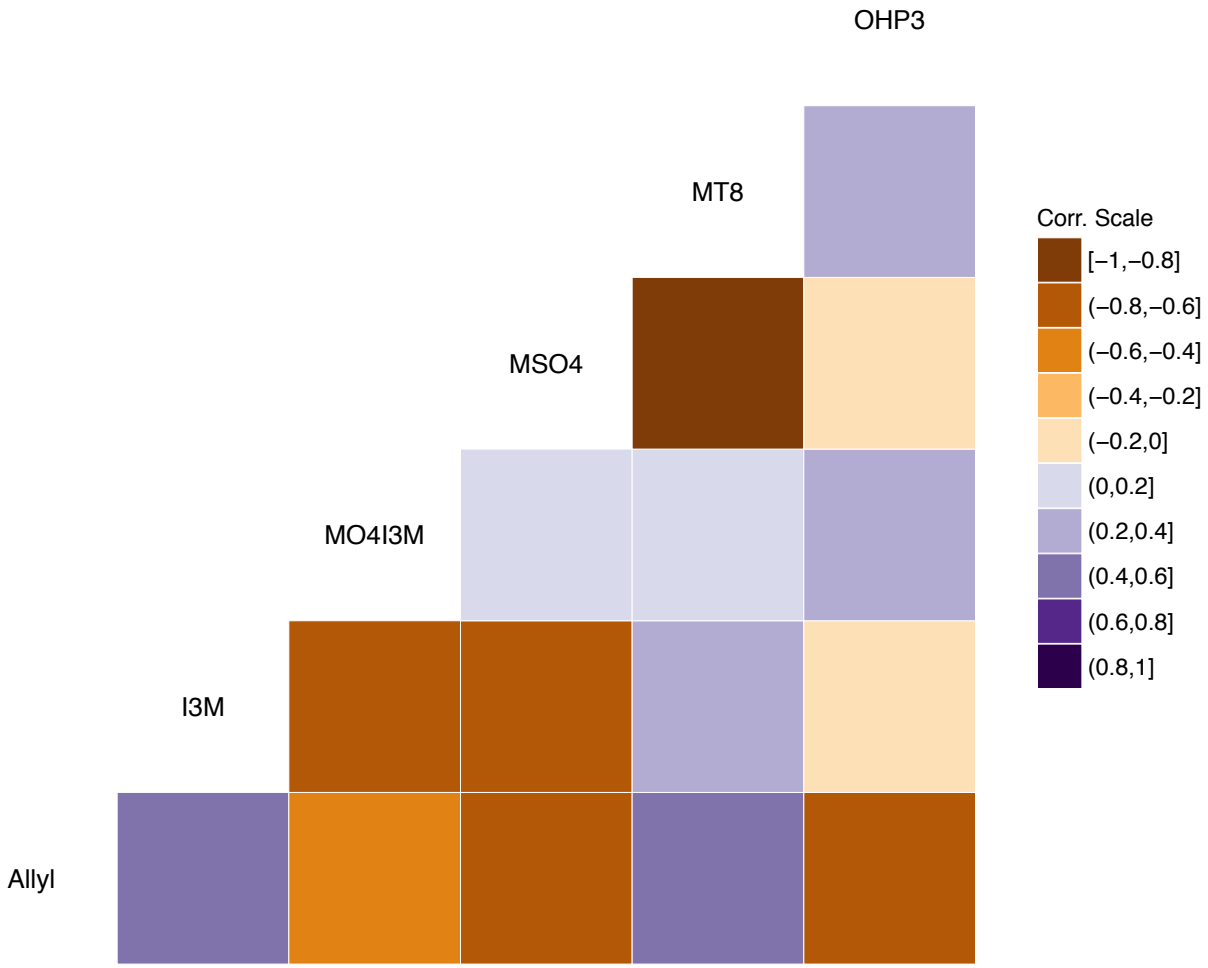
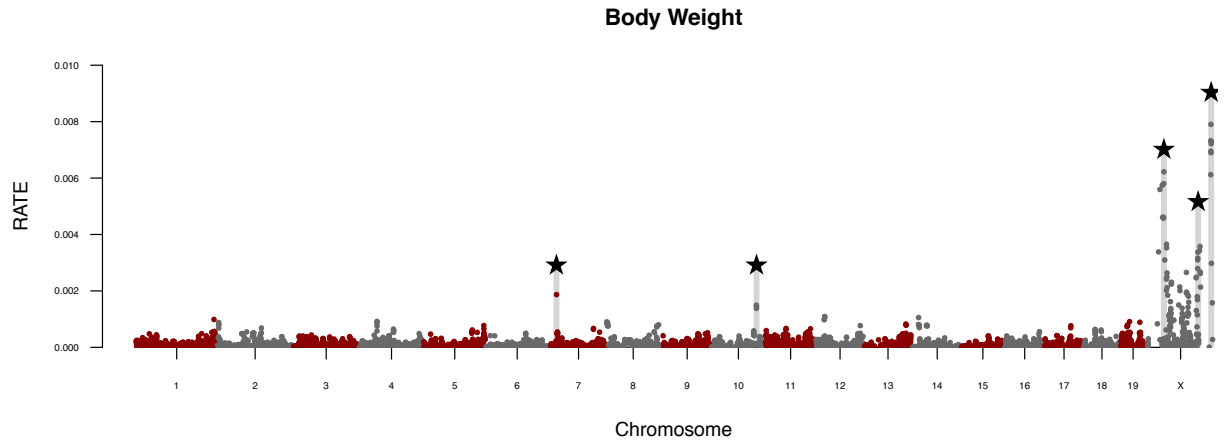
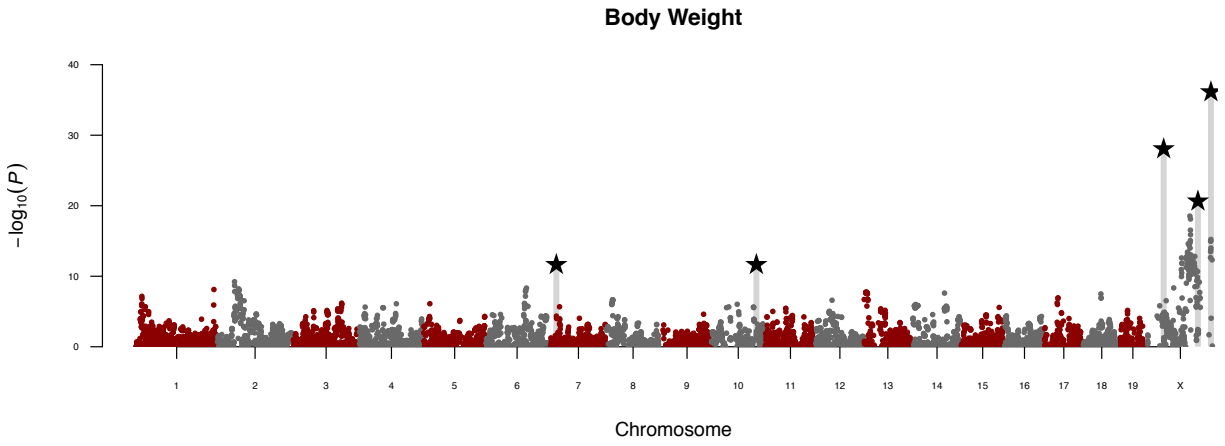


Figure S16. Lower-triangular heat map illustrating the correlation structure between the six metabolic phenotypes in the *Arabidopsis thaliana* QTL mapping study. The six traits analyzed include: allyl content, indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 8-methylthiooctyl (MT8), and 3-hydroxypropyl (OHP3). The legend represents a correlation scale on an [-1,1] interval that has been evenly divided into ten shorter subintervals.

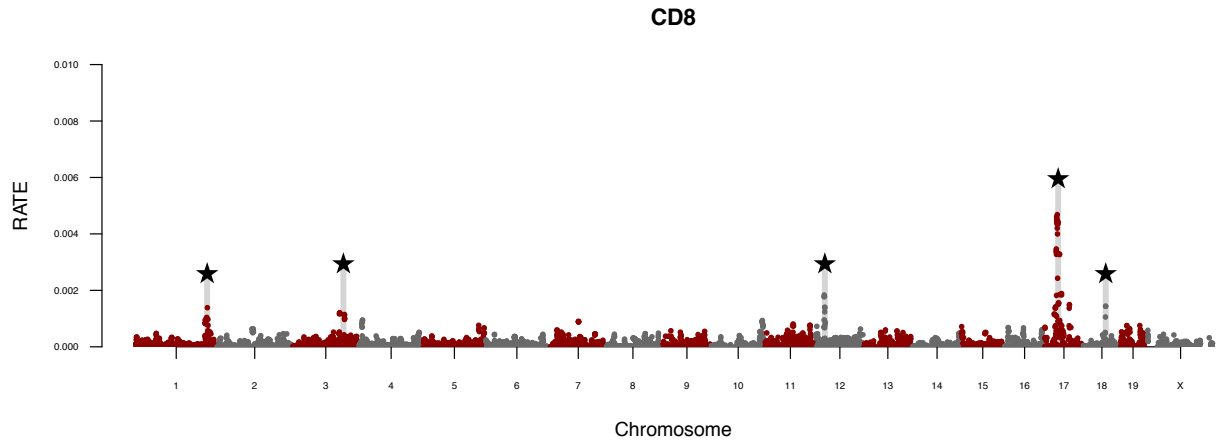


(a) Gaussian Process Regression

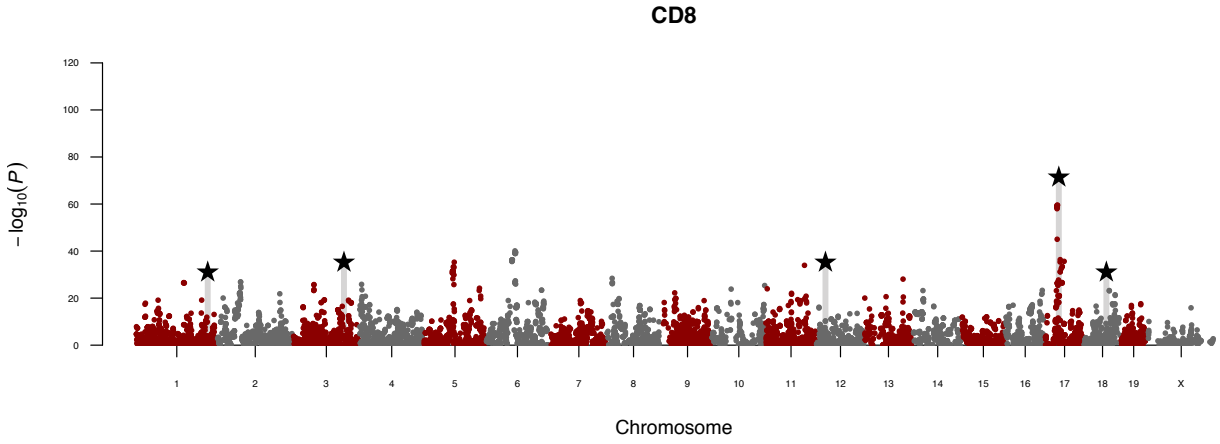


(b) SCANONE

Figure S17. Genome-wide scan for body weight in the heterogenous stock of mice dataset. Figure (a) depicts the relative distributional centrality measures (RATE) of quality-control-positive SNPs plotted against their genomic positions. Gaussian process regression was used to derive these measures. Chromosomes are shown in alternating colors for clarity, with the top 5 most enriched regions being highlighted by the star symbol. Figure (b) serves as a direct comparison and depicts results from a typical GWAS analysis using SCANONE. Here, we overlay the enriched regions detected by RATE to simplify the comparison.



(a) Gaussian Process Regression



(b) SCANONE

Figure S18. Genome-wide scan for percentage of CD8+ cells in the heterogenous stock of mice dataset. Figure (a) depicts the relative distributional centrality measures (RATE) of quality-control-positive SNPs plotted against their genomic positions. Gaussian process regression was used to derive these measures. Chromosomes are shown in alternating colors for clarity, with the top 5 most enriched regions being highlighted by the star symbol. Figure (b) serves as a direct comparison and depicts results from a typical GWAS analysis using SCANONE. Here, we overlay the enriched regions detected by RATE to simplify the comparison.

4 Supplementary Tables

Table S1. Mean squared errors (MSE) and predictive correlation coefficient (R) using the OLS estimates and the effect size analog. Scenarios I, II, and III correspond to phenotypes being generated under a linear model without population structure and one where stratification effects are introduced via the top 5 and 10 genotypic principle components, respectively. We assume that a broad-sense heritability level $H^2 = \{0.3\}$ with control parameter $\rho = \{0.5, 1\}$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. The proportion of times that a method exhibits the lowest MSE or greatest R is denoted as $\text{Opt}\%_{\text{MSE}}$ and $\text{Opt}\%_R$. Values in bold represent the approach with the best (and most robust) performance. Standard errors are given in parentheses.

		$\rho = 0.5$		$\rho = 1$	
	Scenario	LM	GP	LM	GP
MSE	I	1.24 (0.12)	**0.95 (0.09)**	1.03 (0.10)	**0.84 (0.07)**
	II	0.81 (0.09)	**0.69 (0.08)**	0.59 (0.07)	**0.58 (0.06)**
	III	0.80 (0.08)	**0.68 (0.07)**	0.59 (0.07)	**0.57 (0.06)**
$\text{Opt}\%_{\text{MSE}}$	I	0.00	**1.00**	0.02	**0.98**
	II	0.02	**0.98**	0.34	**0.66**
	III	0.04	**0.96**	0.42	**0.58**
R	I	0.20 (0.08)	**0.22 (0.08)**	0.37 (0.06)	**0.39 (0.06)**
	II	0.54 (0.05)	**0.59 (0.04)**	0.68 (0.04)	**0.70 (0.03)**
	III	0.54 (0.06)	**0.59 (0.05)**	0.67 (0.05)	**0.70 (0.04)**
$\text{Opt}\%_R$	I	0.21	**0.79**	0.23	**0.77**
	II	0.02	**0.98**	0.11	**0.89**
	III	0.00	**1.00**	0.13	**0.87**

Table S2. A table that lists a description of the six quantitative phenotypes that are analyzed in the *Arabidopsis thaliana* QTL mapping study. The six metabolic content traits analyzed include: allyl content, indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 8-methylthiooctyl (MT8), and 3-hydroxypropyl (OHP3). (XLSX)

Table S3. Table of all genetic markers and their distributional centrality measures for each of the six metabolic traits in the *Arabidopsis thaliana* QTL mapping study. Listed are the relative centrality (RATE) measures for each variant, along with their L1-regularized effect sizes as computed by lasso regression, the combined L1 and L2-penalized coefficients from the elastic net, the -$\log_{10}$ transformed p-values from SCANONE, and the posterior inclusion probabilities (PIPs) derived from the Bayesian variable selection model. All genetic markers are given in order of their positions along the genome. (XLSX)

Table S4. Proportions of the phenotypic variance explained (PVE) by different orders of genetic effects for the six quantitative phenotypes analyzed in the *Arabidopsis thaliana* QTL mapping study. The six metabolic content traits analyzed include: allyl content, indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 8-methylthiooctyl (MT8), and 3-hydroxypropyl (OHP3). The orders of the genetic effects include: (1) additive effects, (2) pairwise interactions, and (3) third order interactions. Values in bold represent the type of effect that explains the greatest proportion of the overall PVE. Standard errors are given in parentheses.

	Phenotypic Traits					
Genetic Effects	Allyl	I3M	MO4I3M	MSO4	MT8	OHP3
Additive	**0.66 (0.23)**	**0.50 (0.14)**	**0.49 (0.20)**	**0.50 (0.17)**	**0.92 (0.20)**	**0.50 (0.20)**
Pairwise	0.10 (0.05)	0.20 (0.07)	0.28 (0.09)	0.21 (0.07)	0.03 (0.03)	0.06 (0.05)
Third Order	0.24 (0.19)	0.30 (0.13)	0.23 (0.15)	0.29 (0.16)	0.05 (0.05)	0.44 (0.19)

Table S5. The distributional centrality measures for all SNPs in the heterogenous stock of mice dataset. The three traits analyzed include: body weight (page 1), percentage of CD8+ cells (page 2), and high-density lipoproteins (HDL) content (page 3), respectively. Listed are the RATE values for each variant as computed via Gaussian process regression, as well as their marginal p-values computed by using the standard GWAS analysis model SCANONE. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S6. Proportions of the phenotypic variance explained (PVE) by different orders of genetic effects for the three quantitative traits analyzed in the heterogenous stock of mice GWAS dataset. The three traits analyzed include: body weight, percentage (%) of CD8+ cells, and high-density lipoproteins (HDL) content (page 3) The order classes for these genetic effects include: (1) additive effects, (2) pairwise interactions, and (3) third order interactions. Values in bold represent the type of effect that explains the greatest proportion of the overall PVE. Standard errors are given in parentheses.

	Phenotypic Traits		
Genetic Effects	Body Weight	% of CD8+ Cells	HDL Content
Additive	0.20 (0.01)	0.23 (0.10)	0.13 (0.02)
Pairwise	0.14 (0.03)	0.27 (0.22)	0.37 (0.07)
Third Order	**0.66 (0.04)**	**0.50 (0.25)**	**0.50 (0.11)**

Table S7. Empirical computational times for running RATE as a function of sample size and the number of SNPs. Each entry represents the mean computation time (in minutes) it takes to run RATE under implementation strategies. These approaches are using the full rank computation, and the other utilizes the low rank matrix approximations derived in the Supplementary Text. Computations were performed using 28 core nodes on the Athena computing cluster at the Brown University Center for Statistical Sciences. To create genetic data for these simulations, we generated 5.0×10^3, 1×10^4, 1.5×10^4, and 2.0×10^4 genetic markers, respectively. Sample sizes were set to 500, 1000, and 2500. Values in the parentheses are the standard deviations of the estimates across 25 runs.

		Average Time (min) per SNP Size (p)			
Implementation	Total Sample Size	5.0×10^3	1×10^4	1.5×10^4	2.0×10^4
Full Rank	$n = 500$	1.30 (0.01)	5.34 (0.01)	12.24 (0.04)	22.18 (0.09)
	$n = 1{,}000$	7.79 (0.03)	31.43 (0.09)	70.75 (0.09)	126.34 (0.09)
	$n = 2{,}500$	33.8 (0.13)	67.6 (0.17)	352.1 (9.74)	688.5 (19.05)
Low Rank	$n = 500$	0.44 (0.01)	2.08 (0.01)	4.80 (0.01)	8.64 (0.05)
	$n = 1{,}000$	2.11 (0.01)	8.30 (0.10)	18.57 (0.10)	33.34 (0.20)
	$n = 2{,}500$	13.86 (0.02)	51.23 (0.03)	113.13 (0.12)	200.24 (0.40)

References

1. Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda). 2014;4(6):1027–1046.

2. Kolmogorov AN, Rozanov YA. On strong mixing conditions for stationary Gaussian processes. Theory Probab Its Appl. 1960;5(2):204–208.

3. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, MA: MIT Press; 2006.

4. Crawford L, Wood KC, Zhou X, Mukherjee S. Bayesian approximate kernel regression with variable selection. J Am Stat Assoc. 2018;Available from: `https://doi.org/10.1080/01621459.2017.1361830`.

5. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–678. Available from: `http://dx.doi.org/10.1038/nature05911`.

6. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.

7. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 2017;13(7):e1006869–. Available from: `https://doi.org/10.1371/journal.pgen.1006869`.

8. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. 2007;3(7):e114–. Available from: `https://doi.org/10.1371/journal.pgen.0030114`.

9. Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F. Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor Appl Genet. 2002;104(6):1173–1184.

10. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 2007;3(9):e162–. Available from: `https://doi.org/10.1371/journal.pgen.0030162`.

11. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet. 2006;38(8):879–887. Available from: `http://dx.doi.org/10.1038/ng1840`.

12. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–1723. Available from: `http://www.genetics.org/content/178/3/1709.abstract`.

13. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–354. Available from: `http://dx.doi.org/10.1038/ng.548`.

14. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Meth. 2011;8(10):833–835. Available from: `http://dx.doi.org/10.1038/nmeth.1681`.

15. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–825.

16. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24:1550–1557.

17. Rönnegård L, Pong-Wong R, Carlborg Ö. Defining the assumptions underlying modeling of epistatic QTL using variance component methods. J Hered. 2008;99(4):421–425. Available from: http://jhered.oxfordjournals.org/content/99/4/421.abstract.

18. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015;201:759–768.

19. Ning C, Wang D, Kang H, Mrode R, Zhou L, Xu S, et al. A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. Bioinformatics. 2018;34(11):1817–1825. Available from: http://dx.doi.org/10.1093/bioinformatics/bty017.

20. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genomewide association studies. Nat Meth. 2014;11(4):407–409.

21. Zhou X. A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann Appl Stat. 2017;11(4):2027–2051. Available from: https://projecteuclid.org:443/euclid.aoas/1514430276.

22. Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982;50(4):1029–1054.

23. Rao CR. Minimum variance quadratic unbiased estimation of variance components. J Multivar Anal. 1971;1(4):445–456. Available from: http://www.sciencedirect.com/science/article/pii/0047259X71900194.