Supplementary Material: Text

Simulation Study: Genomic Selection with GP Regression

In this subsection, our goal is to empirically motivate the desire to have a principled variable selection procedure for nonparametric (black box) methods. To do this, we consider the same statistical genetics-inspired simulation design that we consider 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 H^2. Two values $\mathrm{H}^2 = \{0.3, 0.6\}$ are examined — each corresponding to simulation scenarios I and II, respectively. 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 use a simulated genotype matrix $\mathbf{X}$ with $n = 500$ samples and $p = 250$ single nucleotide polymorphisms (SNPs) 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 set of 10 solely additive variants, and (ii) a set of 20 nonlinearly behaving variants. All causal markers in the second group have additive effects and are also involved in pairwise interactions.

The linear effect sizes for all j^* associated genetic variants are assumed to come from a standard uniform 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 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 $\mathrm{V}(\mathbf{X}\boldsymbol{\beta}) = \rho\mathrm{H}^2$, while the proportion detailed by nonlinearity is given as $\mathrm{V}(\mathbf{W}\boldsymbol{\gamma}) = (1-\rho)\mathrm{H}^2$. 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 $\mathrm{V}(\mathbf{y})$. Finally, continuous phenotypes are then created by summing over all observed effects using the simulation model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$.

We 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. 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). The second method we consider is a standard linear model, fit using the standard ordinary least squares (OLS), 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 $\mathrm{Opt\%}_{\mathrm{MSE}}$ and $\mathrm{Opt\%}_{r}$, respectively. We analyze 100 different simulated datasets for each scenario H^2 and 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 a majority of the simulation scenarios. This discrepancy is particularly obvious when there is a low signal-to-noise ratio (i.e. $\mathrm{H}^2 = 0.3$ in Scenario I), as well as when there are underlying interactions affecting the model (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 (e.g. Howard et al., 2014).

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 is completely specified by its mean function and positive definite covariance (kernel) function, $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$, 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 (Kolmogorov and Rozanov, 1960)

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

Altogether, we refer to the above as taking a "weight-space" view on Gaussian process regression (Rasmussen and Williams, 2006). Here, $\mathbf{y}$ 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,

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

where $\|f\|_{\mathrm{K}}$ 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 (Crawford et al., 2017). The results in this section will be 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 $\widehat{\boldsymbol{\beta}}_2$ that 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 (Crawford et al. (2017))**.** *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,

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

where $\boldsymbol{\delta}_{\|}$ 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 $\widetilde{\boldsymbol{\beta}}_1 \neq \widetilde{\boldsymbol{\beta}}_2$. □

Efficient 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 $\mathbf{f} = \mathbf{\Psi}^{\intercal}\mathbf{c}$ via the linear map,

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

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 $\mathbf{\Sigma} = \mathbf{\Lambda}^{-1}$ estimated via simulation methods. Under these assumptions, we may partition conformably as follows

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

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 $\mathbf{\Sigma}_{-j}$ and $\mathbf{\Lambda}_{-j}$ 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 $p(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j)$ and $p(\widetilde{\boldsymbol{\beta}}_{-j})$ — simplifies to the following closed form solution

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

where $\alpha_j = \boldsymbol{\lambda}^{\intercal}_{-j}\mathbf{\Lambda}^{-1}_{-j}\boldsymbol{\lambda}_{-j}$. 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, both of these terms remain relatively equal for each marker j and, when added to $(1-p)$, make a negligible contribution to the entire sum. Thus, we begin by simplifying our computation to the following

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

Next, notice that the KLD still relies on the full precision matrix $\mathbf{\Lambda}$. 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 $\mathbf{\Lambda}^{-1}_{-j}$. This requires inverting a $p-1\times p-1$ matrix separately for each marker j. Based on the assumed projection in (S4), we implement the following procedures to reduce burden and complexity.

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

Case #2: Calculating the Matrices $\mathbf{\Lambda}^{-1}_{-j}$ with $\boldsymbol{p > n}$. In this case, the calculation of $\mathbf{\Lambda}^{-1}_{-j}$ can be reduced to the inverse of an $n\times n$ matrix. Let $\mathbf{\Omega} = \sigma(\mathbf{f})$ be the empirical posterior covariance matrix of the estimated functions $\mathbf{f}$. Then from the projection in (S4), we estimate

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

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

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

as long as $\mathbf{X}$ has linearly independent rows.

Next, for this case, we need to compute the rate of change $\alpha_j = \boldsymbol{\lambda}^{\intercal}_{-j}\mathbf{\Lambda}^{-1}_{-j}\boldsymbol{\lambda}_{-j}$ for every marker j. Recall that the precision matrix $\mathbf{\Lambda} = \mathbf{X}^{\intercal}\mathbf{\Omega}^{\dagger}\mathbf{X} = \mathbf{\Sigma}^{\dagger}$. We will now find $\mathbf{L} = \mathbf{\Lambda}^{1/2}$ such that $\mathbf{\Lambda} = \mathbf{L}\mathbf{L}^{\intercal}$. 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 $(\mathbf{\Omega}^{\dagger})^{1/2}$ using a factorization such as SVD. Now, we may define $\mathbf{L} = \mathbf{X}^{\intercal}(\mathbf{\Omega}^{\dagger})^{1/2}$. 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

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

Using these quantities, we calculate $\alpha_j = \boldsymbol{\lambda}^{\intercal}_{-j}\mathbf{L}^{\intercal\dagger}_{-j}\mathbf{L}^{\dagger}_{-j}\boldsymbol{\lambda}_{-j}$. To compute $\mathbf{L}^{\dagger}_{-j}\boldsymbol{\lambda}_{-j}$, 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 (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.

Software Implementation

This implementation is used when computing the RATE measures, which is freely available in R code at `https://github.com/lorinanthony/RATE`. A complete algorithmic overview (without these efficiency modifications) are given later in the Supplementary Material. See Table S2 for an analysis of method scalability and empirical computational complexity.

Supporting Information: Algorithmic Overview

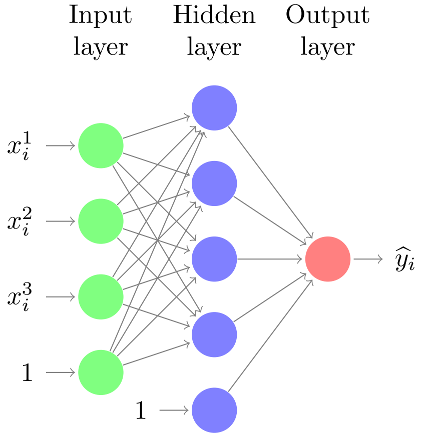
Algorithm 1 Gaussian Process Regression (GPR)

```
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 $\tau^2 \sim$ 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 $\mathbf{V}^* = \tau^2(\tau^2\mathbf{K} + \mathbf{I})^{-1}$;
    $\tau^2 \,|\, \mathbf{f}, \mathbf{y} \sim$ 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
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
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)

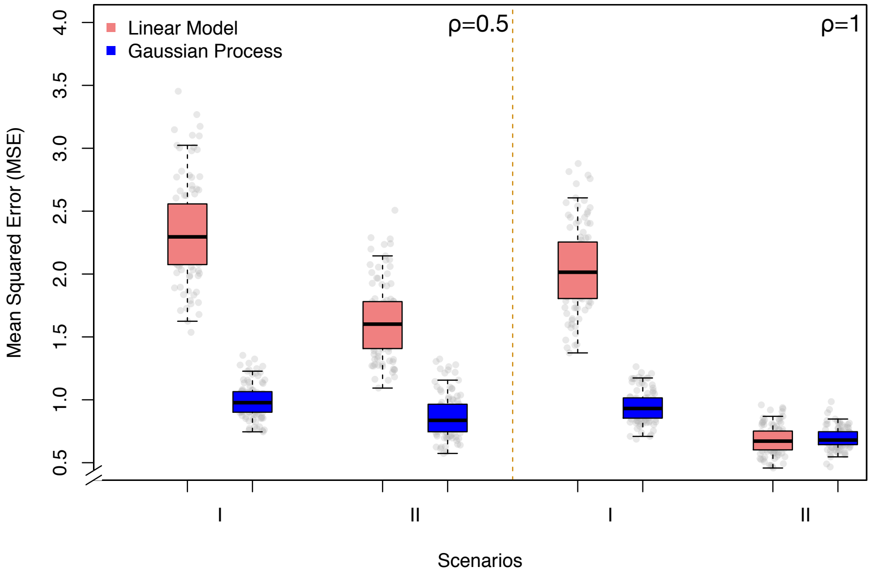
```
Select a positive definite covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are $n$-dimensional vectors from the design matrix.
Construct the $n \times n$ covariance matrix $\mathbf{K}$.
Define the full model where $\mathbf{y} = \mathbf{K}\boldsymbol{\vartheta} + \boldsymbol{\varepsilon}$ 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 $p(\widetilde{\boldsymbol{\beta}} \,|\, \mathbf{y})$ 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
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
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 3 Bayesian Neural Network (BNN)

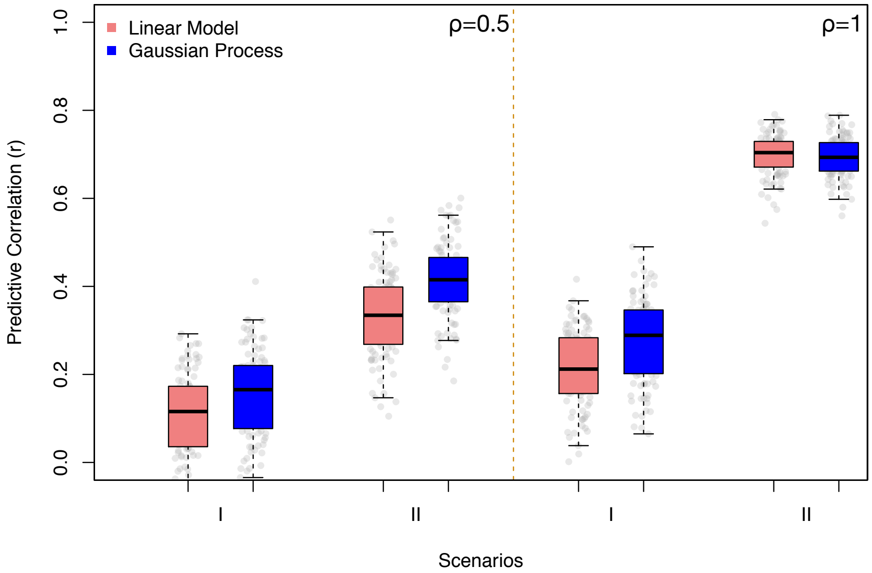


```
Specify the architecture of the neural network (e.g. see above), operating over input/output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Denote the output of the network for a given $x_i$ as $\widehat{y}_i$.
Specify a prior distribution $\pi(\cdot)$ over all parameters (e.g. weights and 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 $p(y_1^*, \ldots, y_n^* \mid \mathcal{D})$.
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 $p(\widetilde{\boldsymbol{\beta}} \mid \mathcal{D})$ 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
    $\text{KLD}(\widetilde{\beta}_j) = \frac{1}{2}\left[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + tr(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2\right]$.
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
```

Supporting Information: Figures

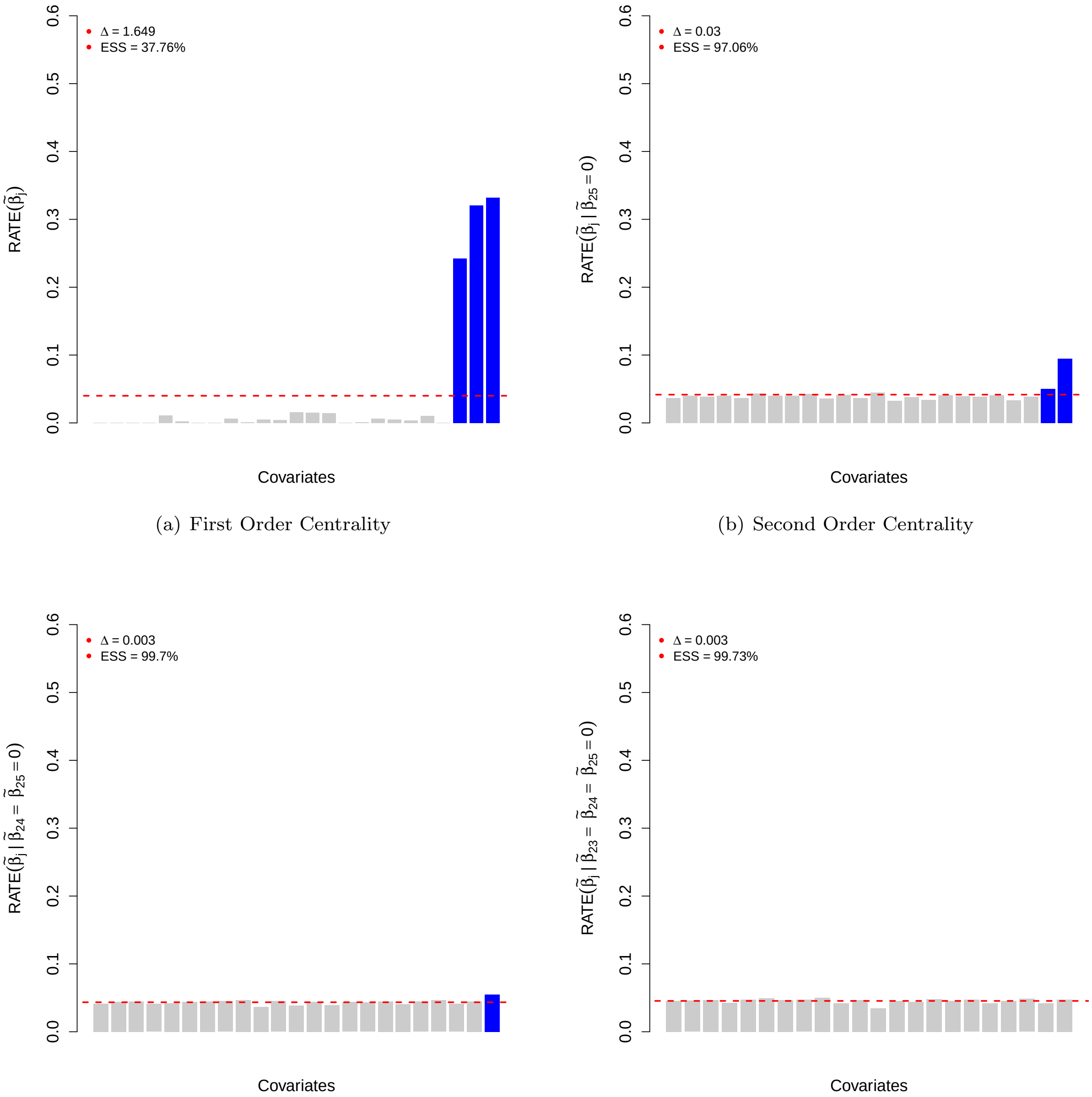


(a) Mean Squared Error (MSE)



(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. Scenarios I and II correspond to response variables being generated according to broad-sense heritability level $\mathrm{H}^2 = \{0.3, 0.6\}$ with control parameter $\rho = \{0.5, 1\}$. Here, $(1 - \rho)$ 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.



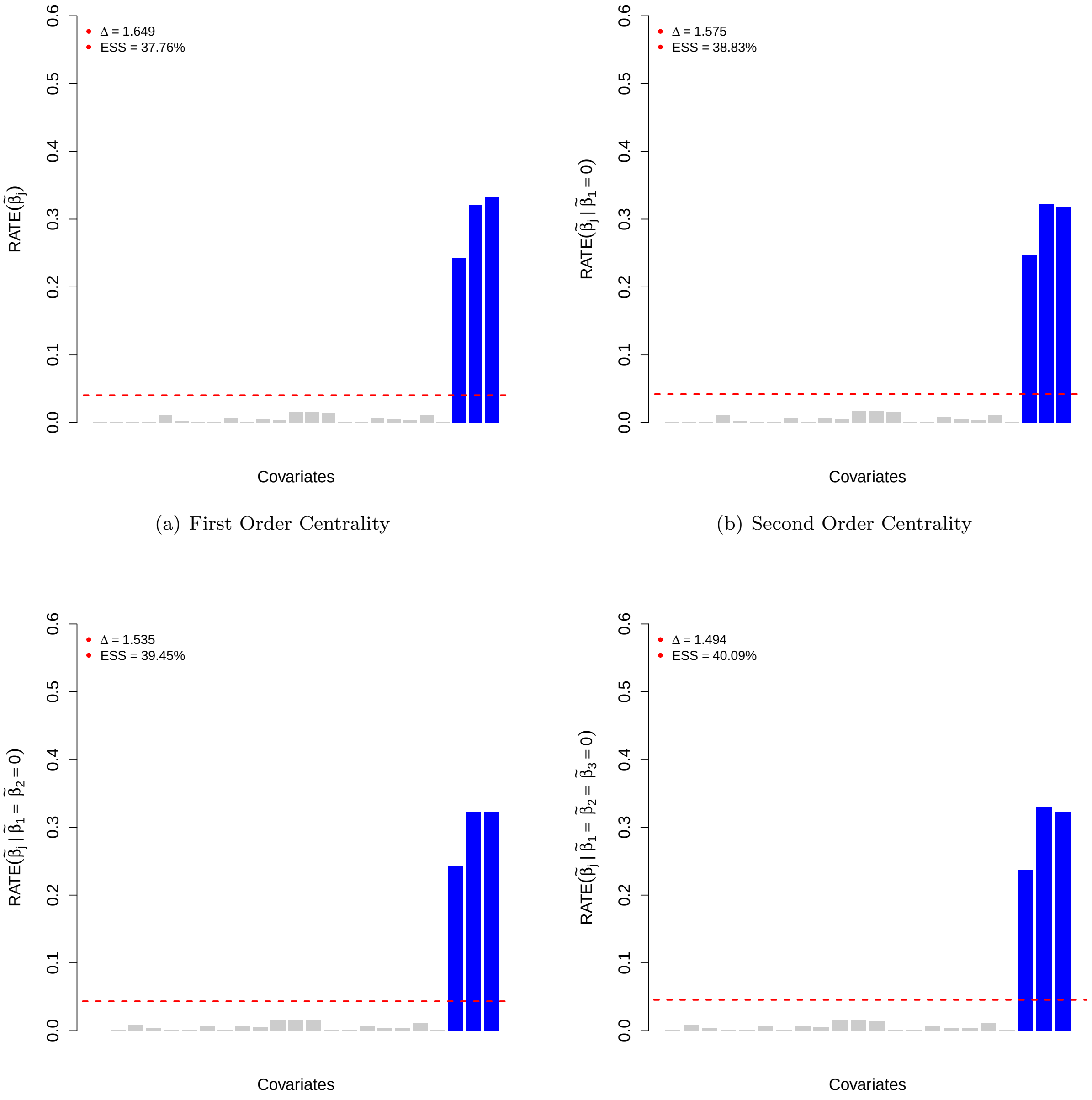
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S2: Orders of distributional centrality via RATE measures 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 $p^* = \{23, 24, 25\}$ (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.



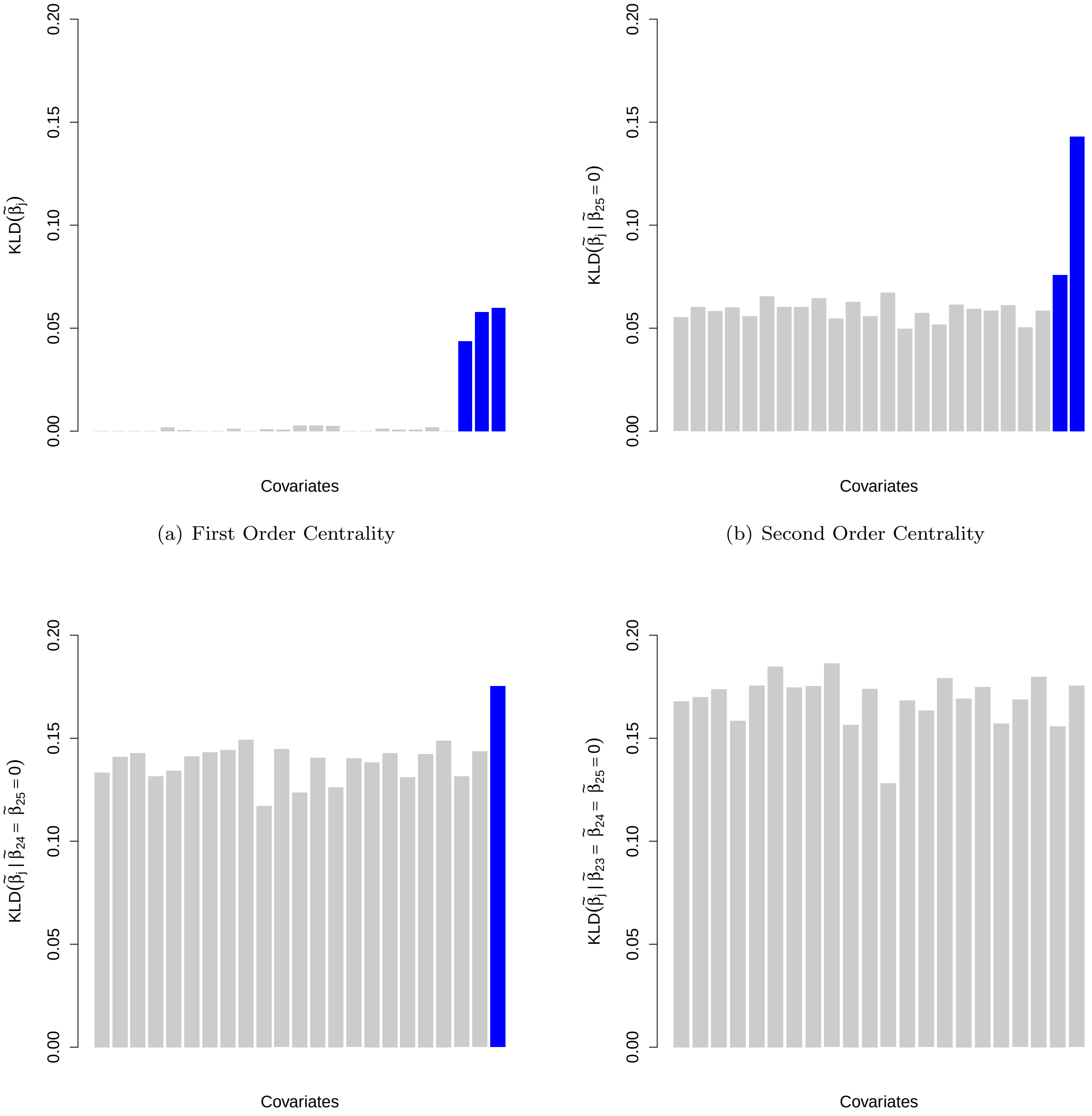
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S3: Orders of distributional centrality via RATE measures 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 $p^* = \{23, 24, 25\}$ (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.



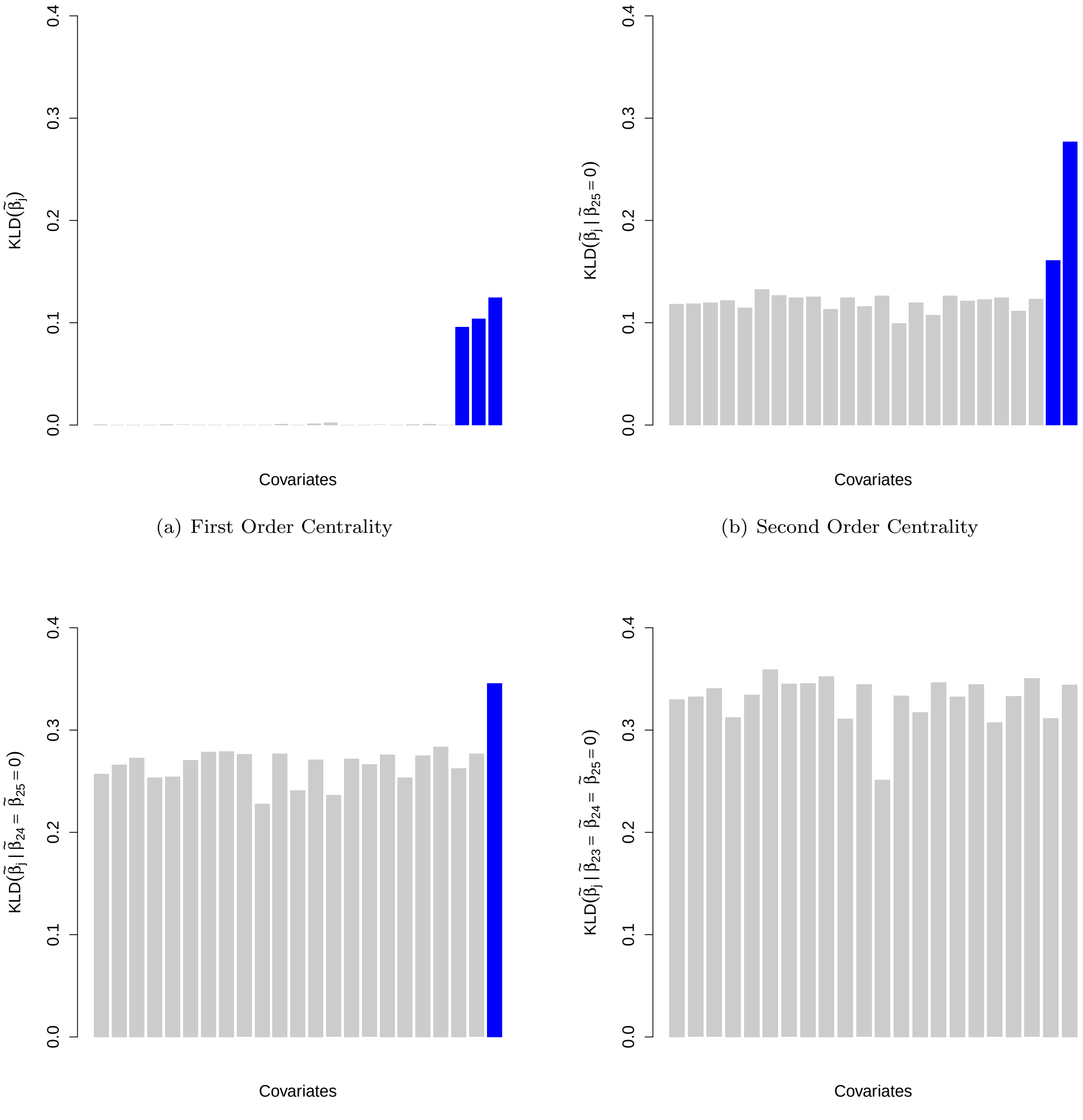
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S4: Orders of distributional centrality via raw and unscaled Kullback-Leibler divergence (KLD) measures 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 $p^* = \{23, 24, 25\}$ (blue) are nonzero. Figure (a) shows the first order centrality across all markers; (b)-(d) show results when the most significantly associated variants are iteratively nullified.



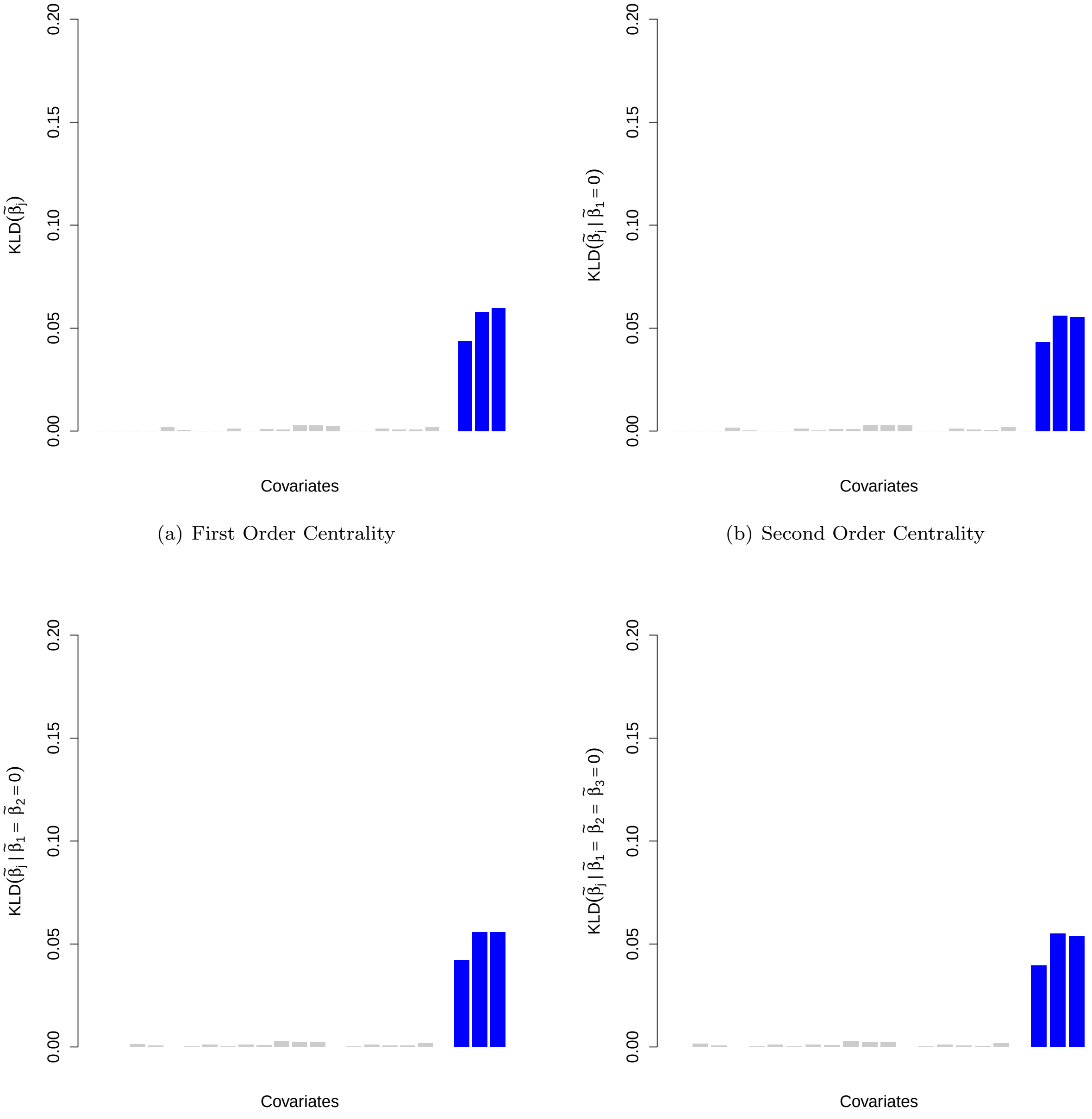
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S5: Orders of distributional centrality via raw and unscaled Kullback-Leibler divergence (KLD) measures 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 $p^* = \{23, 24, 25\}$ (blue) are nonzero. Figure (a) shows the first order centrality across all markers; (b)-(d) show results when the most significantly associated variants are iteratively nullified.



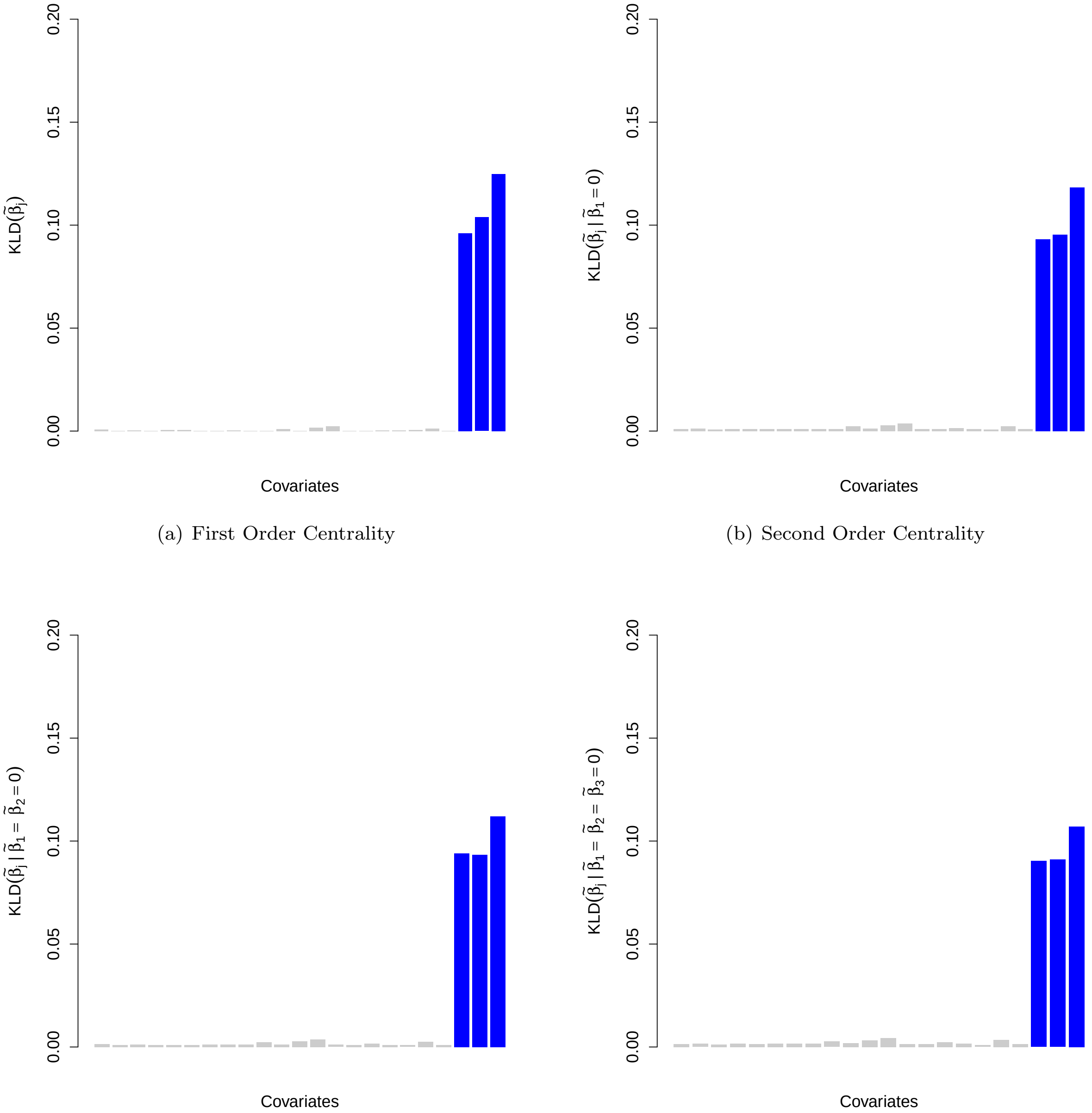
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S6: Orders of distributional centrality via raw and unscaled Kullback-Leibler divergence (KLD) measures 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 $p^* = \{23, 24, 25\}$ (blue) are nonzero. Figure (a) shows the first order centrality across all markers; (b)-(d) show the results when nonsignificant markers #1-3 are iteratively nullified.



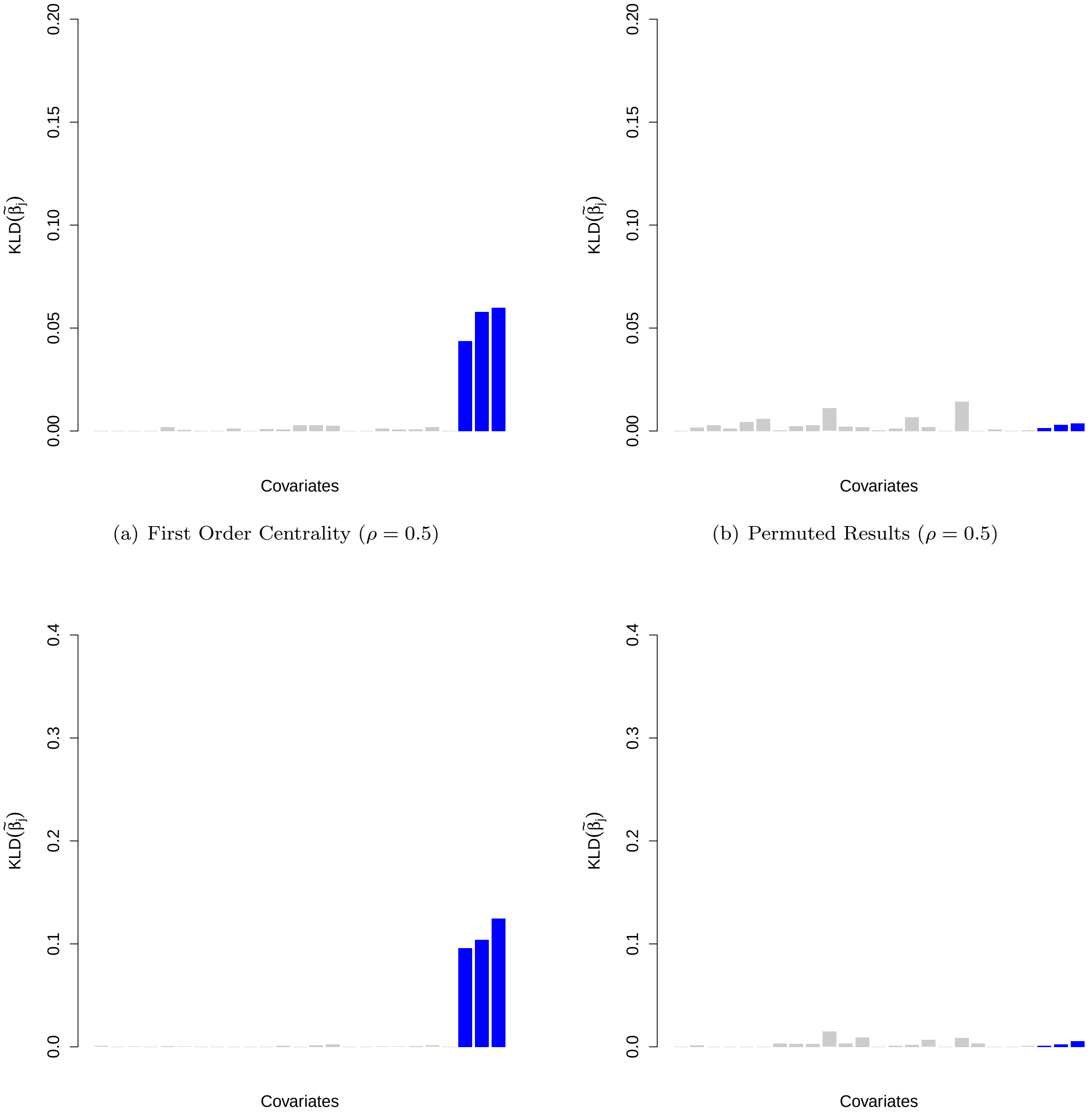
(a) First Order Centrality

(b) Second Order Centrality

(c) Third Order Centrality

(d) Fourth Order Centrality

Figure S7: Orders of distributional centrality via raw and unscaled Kullback-Leibler divergence (KLD) measures with broad-sense heritability level $\text{H}^2 = 0.6$ 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 $p^* = \{23, 24, 25\}$ (blue) are nonzero. Figure (a) shows the first order centrality across all markers; (b)-(d) show the results when nonsignificant markers #1-3 are iteratively nullified.



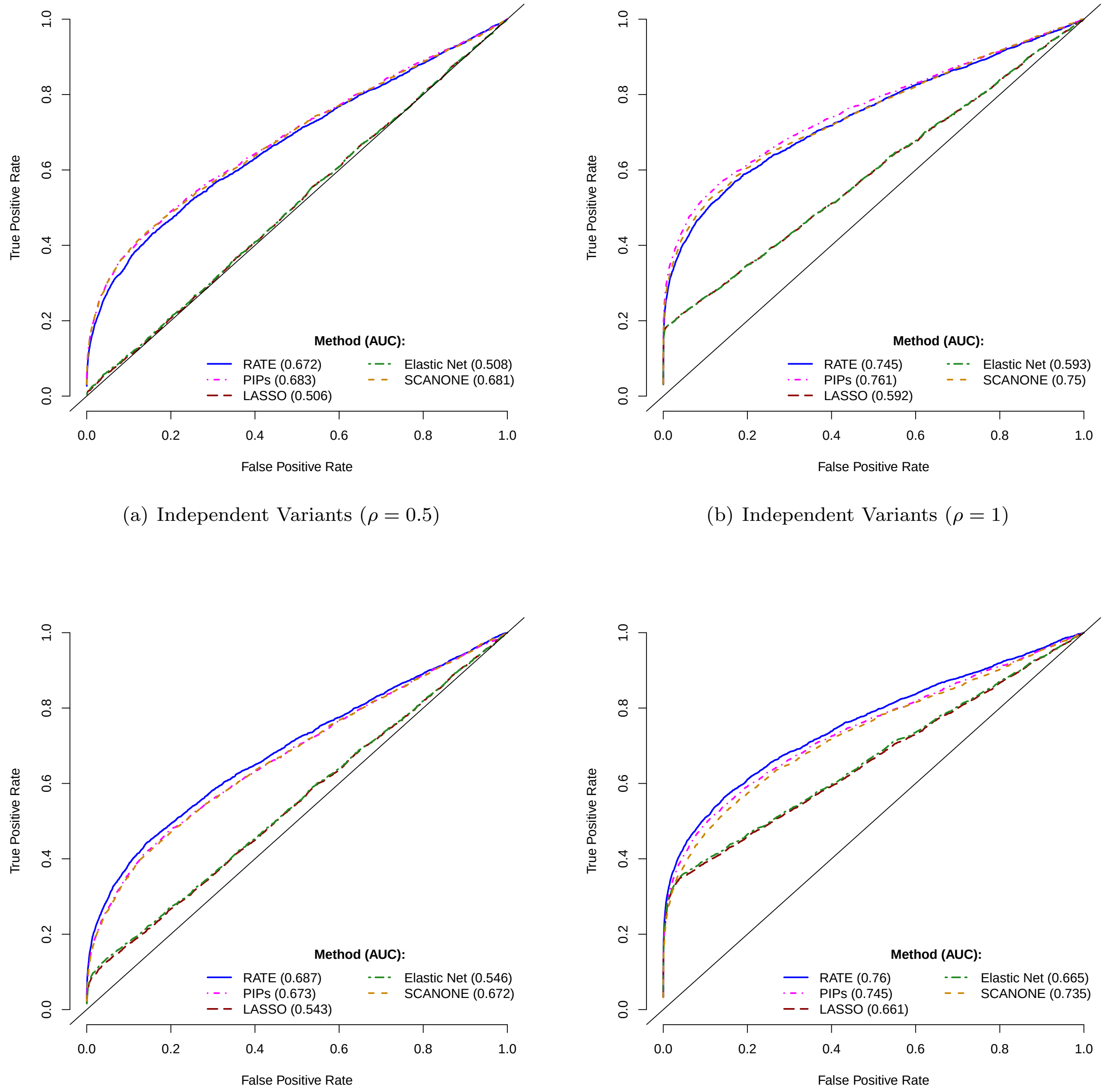
(a) First Order Centrality ($\rho = 0.5$)

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

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

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

Figure S8: Orders of distributional centrality via raw and unscaled Kullback-Leibler divergence (KLD) measures 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 $p^* = \{23, 24, 25\}$ (blue) are nonzero. Figure (a) and (c) show the first order centrality across all markers; (b) and (d) show comparative results when the phenotypes have been permuted once.



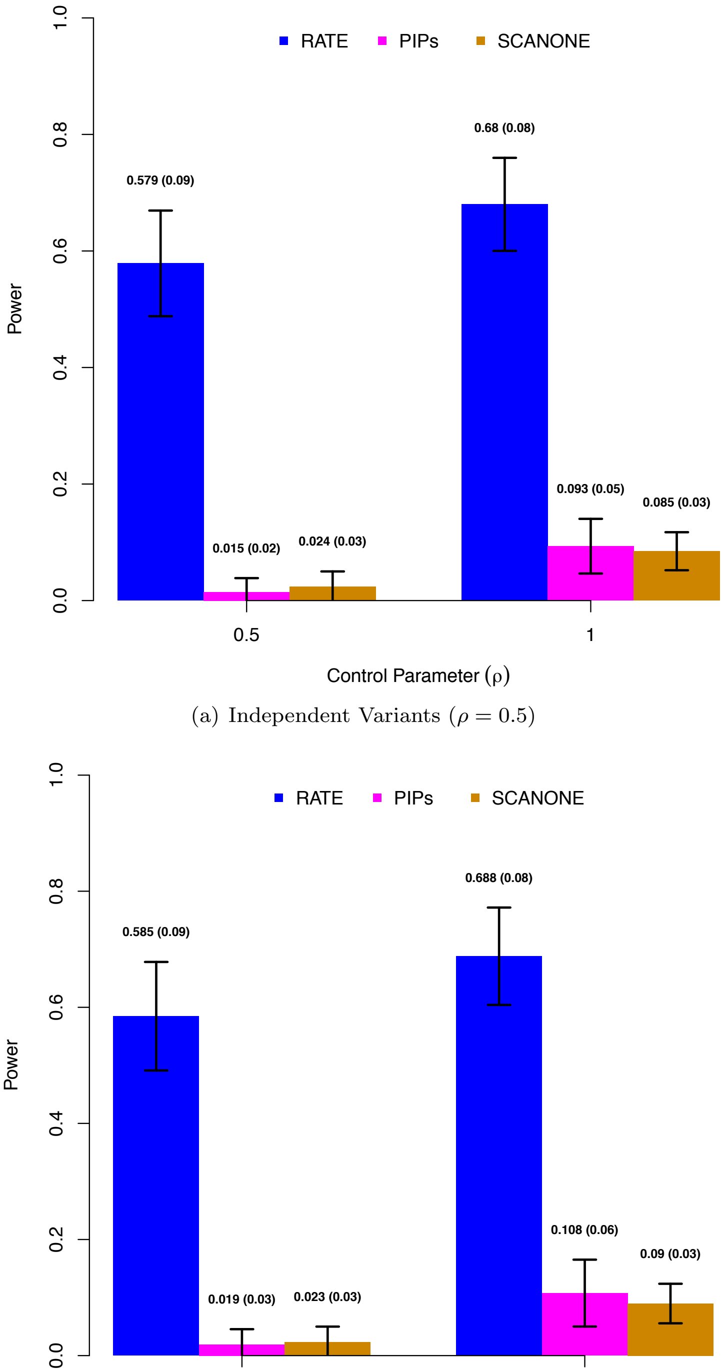
(a) Independent Variants ($\rho = 0.5$)

(b) Independent Variants ($\rho = 1$)

(c) Structured Variants ($\rho = 0.5$)

(d) Structured Variants ($\rho = 1$)

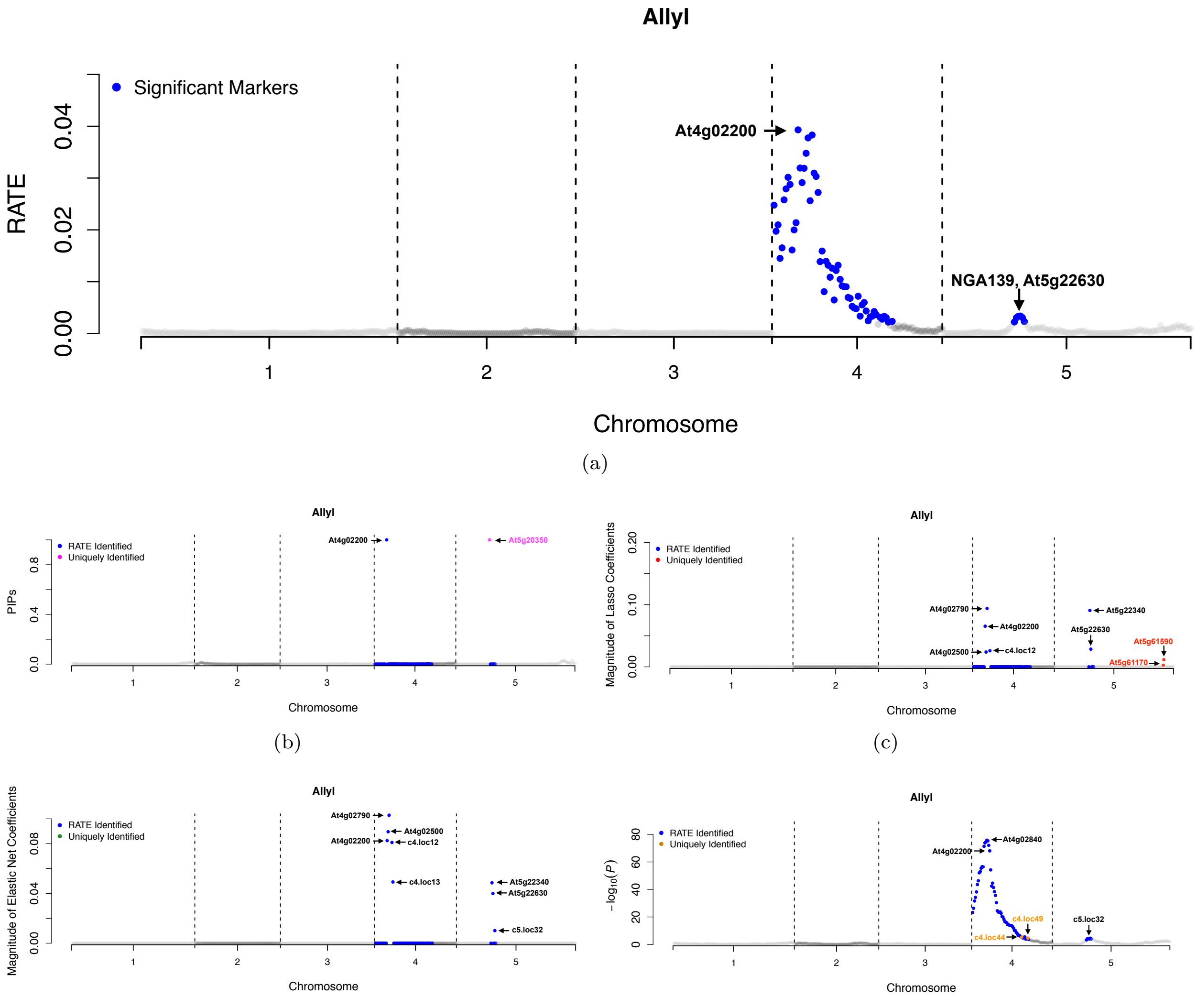
Figure S9: Power analysis for prioritizing genetic variants. Data were simulated with $n = 500$ individuals and $p = 2500$ SNPs. Phenotypes are simulated with broad-sense heritability level $\mathrm{H}^2 = 0.3$ with control parameter $\rho = \{0.5, 1\}$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Compared are RATE (blue), lasso regression (red), the elastic net (green), the SCANONE method (orange), and a Bayesian spike and slab prior (PIPs) (pink). Figures (a)-(b) show results using the standard model; while (c)-(d) are results with population stratification. Area under the curve (AUC) is also reported to facilitate comparisons.



(a) Independent Variants ($\rho = 0.5$)

(b) Independent Variants ($\rho = 1$)

Figure S10: Power analysis for prioritizing causal variants under the "optimal" model criterion. Data were simulated with $n = 500$ individuals and $p = 2500$ SNPs. Phenotypes are simulated with broad-sense heritability level $\mathrm{H}^2 = 0.3$ with control parameter $\rho = \{0.5, 1\}$. Here, $(1 - \rho)$ is used to determine the proportion of signal that is contributed by interaction effects. Compared are RATEs $> 1/p$ (blue), the Bayesian "median probability model" (pink) (i.e. PIPs > 0.5), and the multiple testing corrected SCANONE method $P < 2 \times 10^{-5}$ (orange). Figure (a) show results using the standard model; while (b) are results with population stratification.



(a)

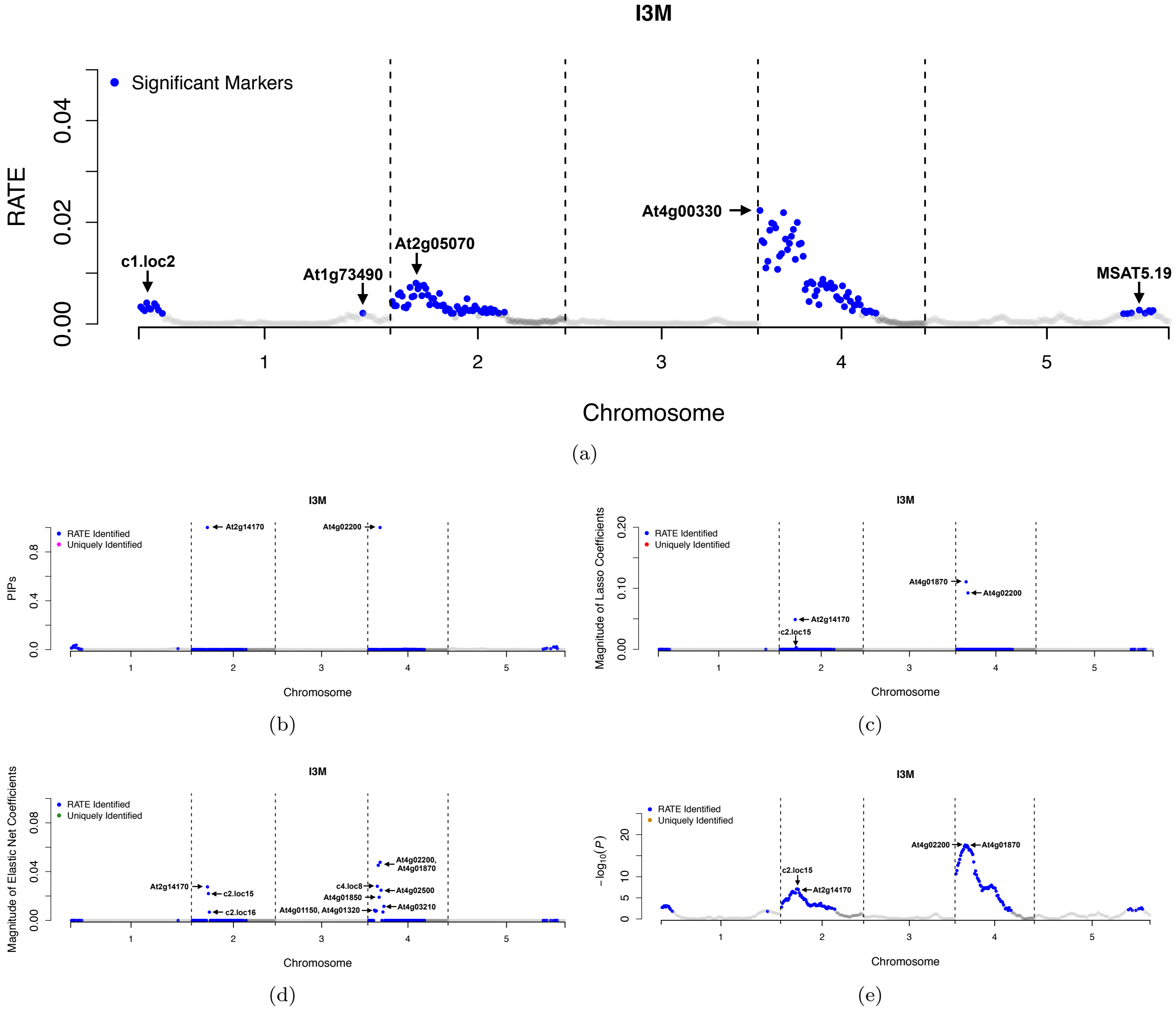
(b)

(c)

(d)

(e)

Figure S11: Genetic map wide scan for the allyl content metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) RATE, (b) the Bayesian spike and slab prior (pink), (c) lasso regression (red), (d) elastic net regularization (green), and (e) SCANONE (orange). Significant markers are determined by $\text{RATE}(\widetilde{\beta}) > 1/p$, $\text{PIP}(\beta) > 0.5$, $|\widehat{\beta}| > 0$, 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)

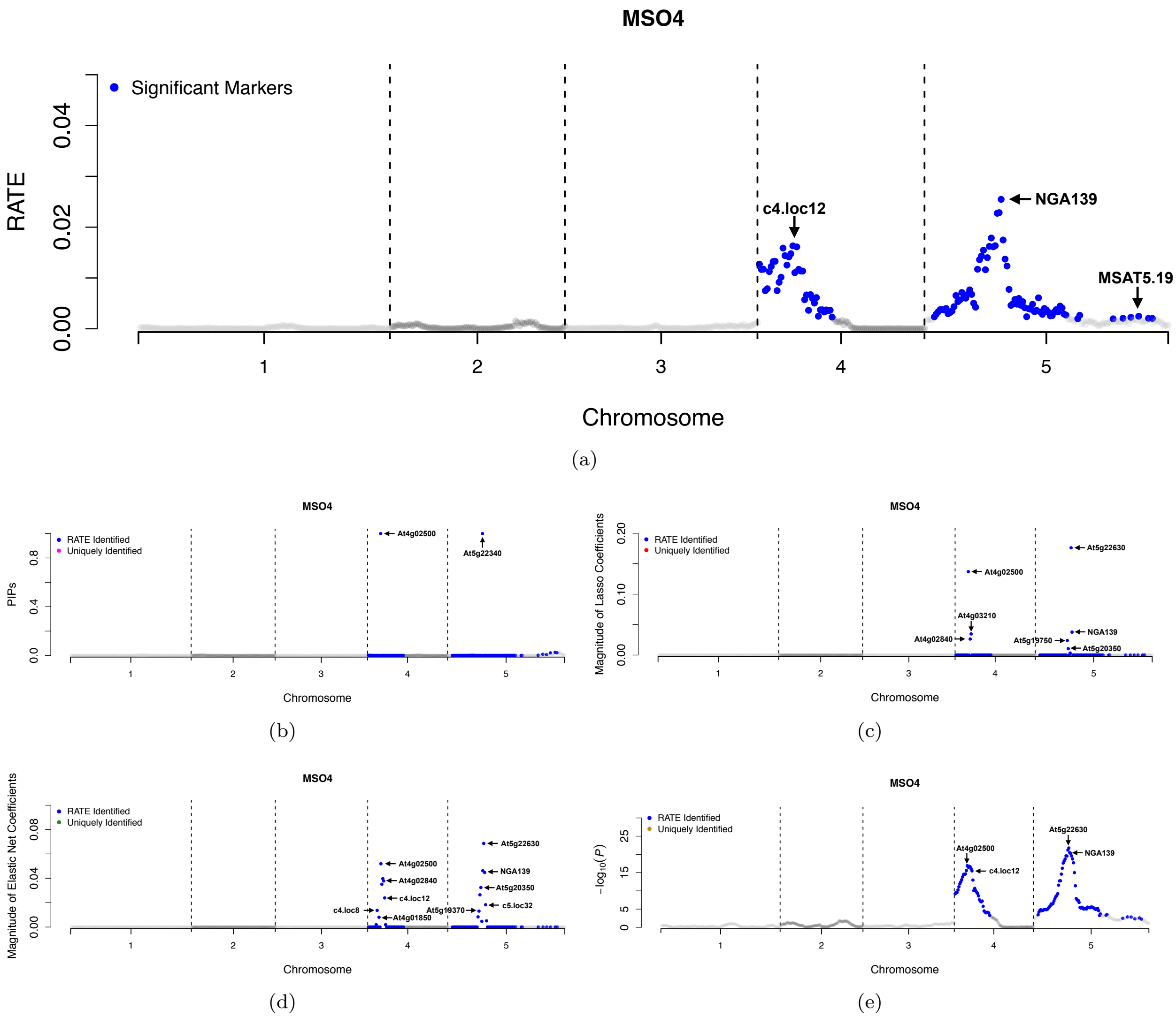
(b)

(c)

(d)

(e)

Figure S12: Genetic map wide scan for the indol-3-ylmethyl (I3M) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) RATE, (b) the Bayesian spike and slab prior (pink), (c) lasso regression (red), (d) elastic net regularization (green), and (e) SCANONE (orange). Significant markers are determined by RATE($\widetilde{\beta}$) $> 1/p$, PIP(β) > 0.5, $|\widehat{\beta}| > 0$, 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)

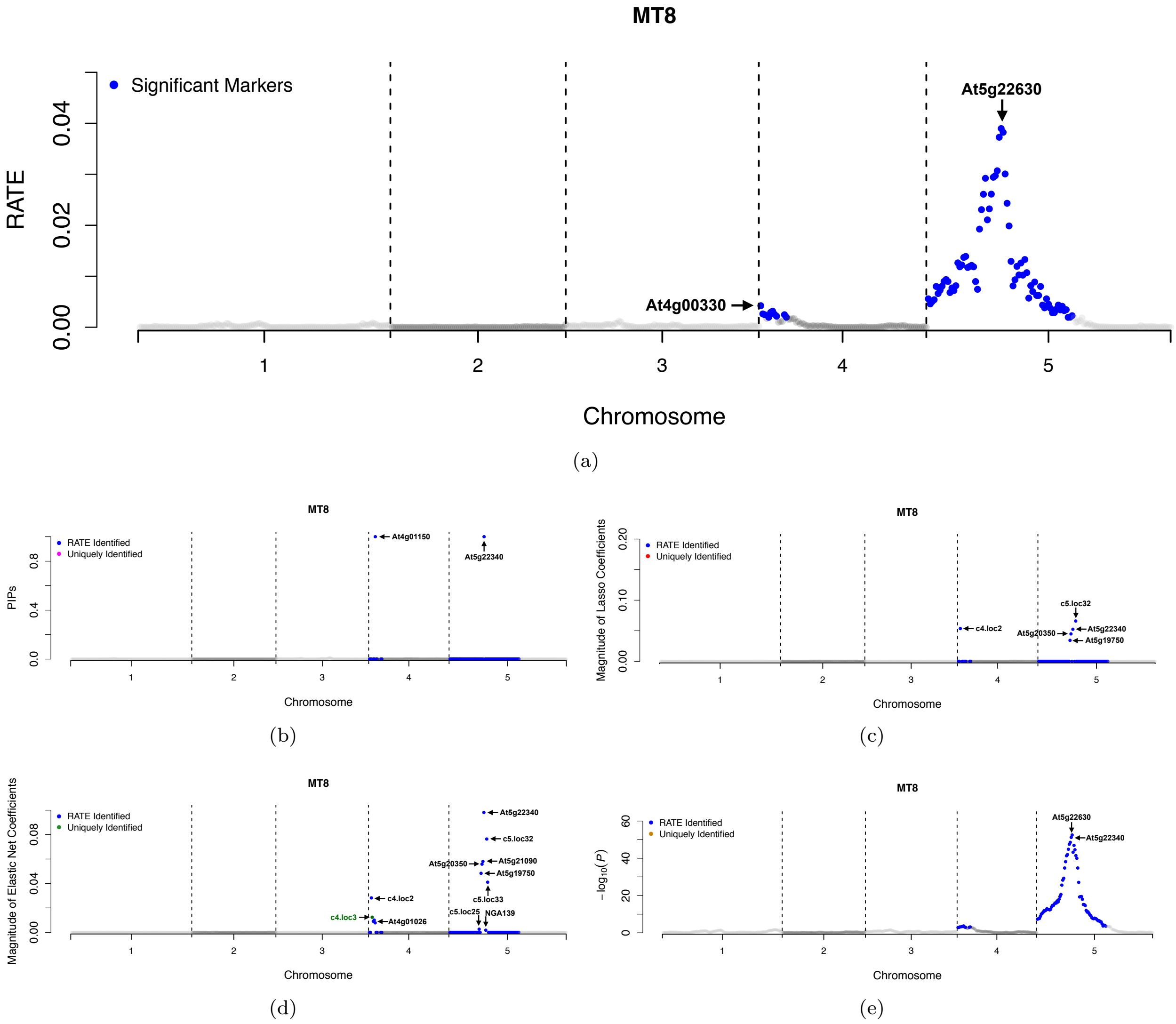
(b)

(c)

(d)

(e)

Figure S13: Genetic map wide scan for the 4-methylsulfinylbutyl (MSO4) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) RATE, (b) the Bayesian spike and slab prior (pink), (c) lasso regression (red), (d) elastic net regularization (green), and (e) SCANONE (orange). Significant markers are determined by RATE($\widetilde{\beta}$) $> 1/p$, PIP(β) > 0.5, $|\widehat{\beta}| > 0$, 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)

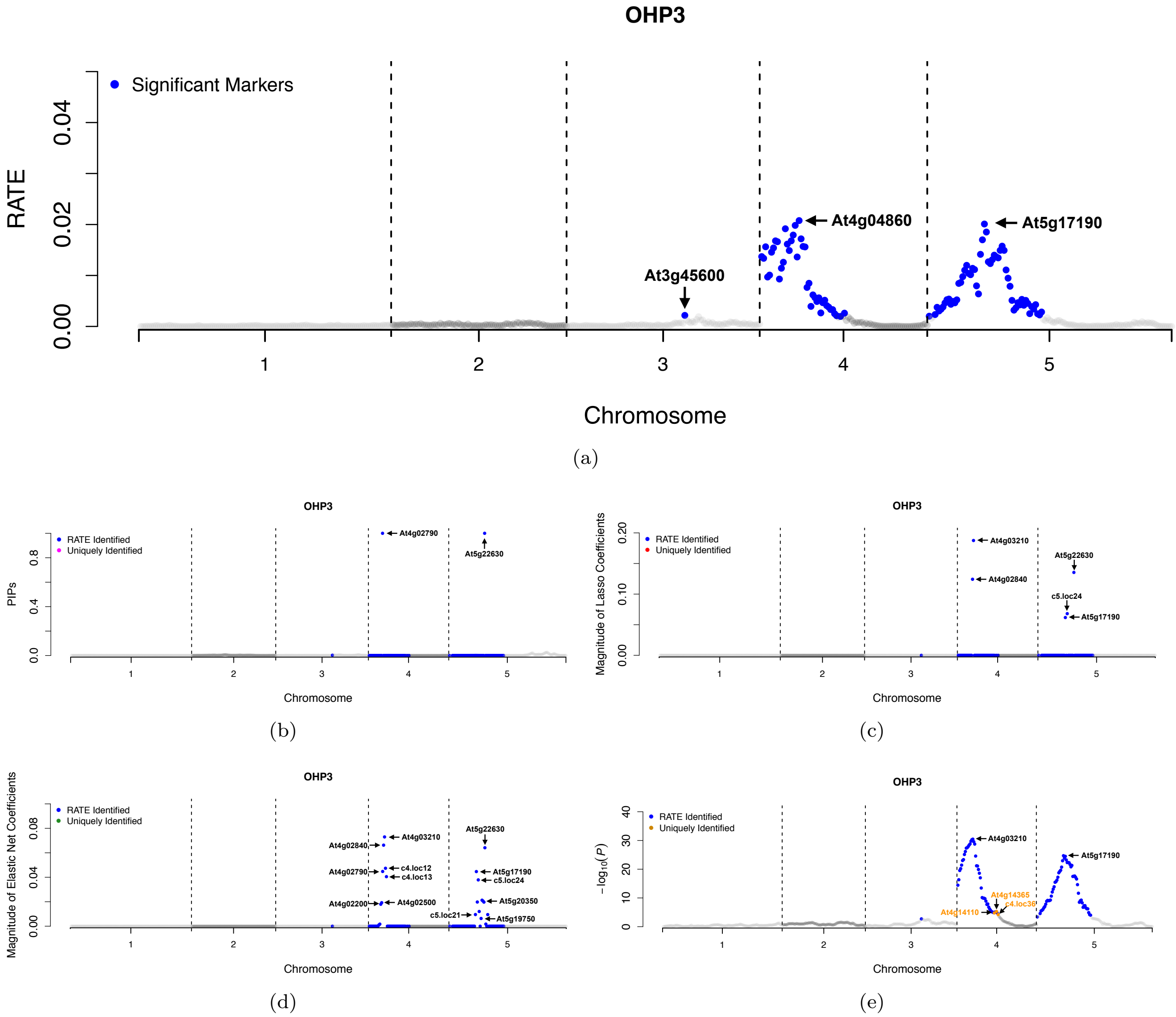
(b)

(c)

(d)

(e)

Figure S14: Genetic map wide scan for the 8-methylthiooctyl (MT8) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) RATE, (b) the Bayesian spike and slab prior (pink), (c) lasso regression (red), (d) elastic net regularization (green), and (e) SCANONE (orange). Significant markers are determined by RATE$(\widetilde{\beta}) > 1/p$, PIP$(\beta) > 0.5$, $|\widehat{\beta}| > 0$, 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)

(b)

(c)

(d)

(e)

Figure S15: Genetic map wide scan for the 3-hydroxypropyl (OHP3) metabolism trait analyzed in *Arabidopsis thaliana* QTL mapping study. Compared methods are (a) RATE, (b) the Bayesian spike and slab prior (pink), (c) lasso regression (red), (d) elastic net regularization (green), and (e) SCANONE (orange). Significant markers are determined by RATE($\widetilde{\beta}$) $> 1/p$, PIP(β) > 0.5, $|\widehat{\beta}| > 0$, 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.

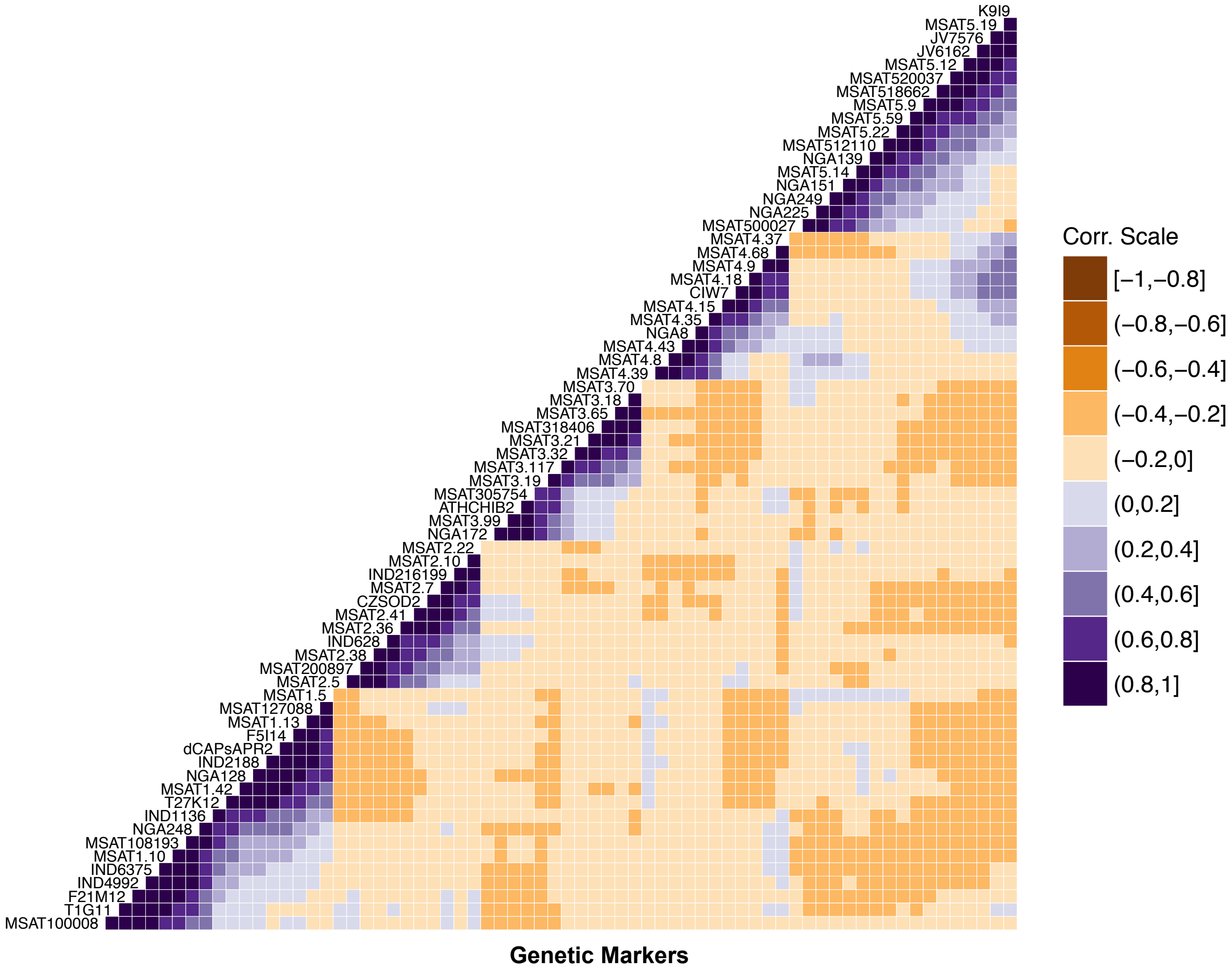


Figure S16: Lower-triangular heat map illustrating the correlation structure for a proportion of the genotyped markers in the *Arabidopsis thaliana* QTL mapping study. The legend represents a correlation scale on an [-1,1] interval that has been evenly divided into ten shorter subintervals. There appears to be an underlying covarying structure between groups of markers located on different chromosomes.

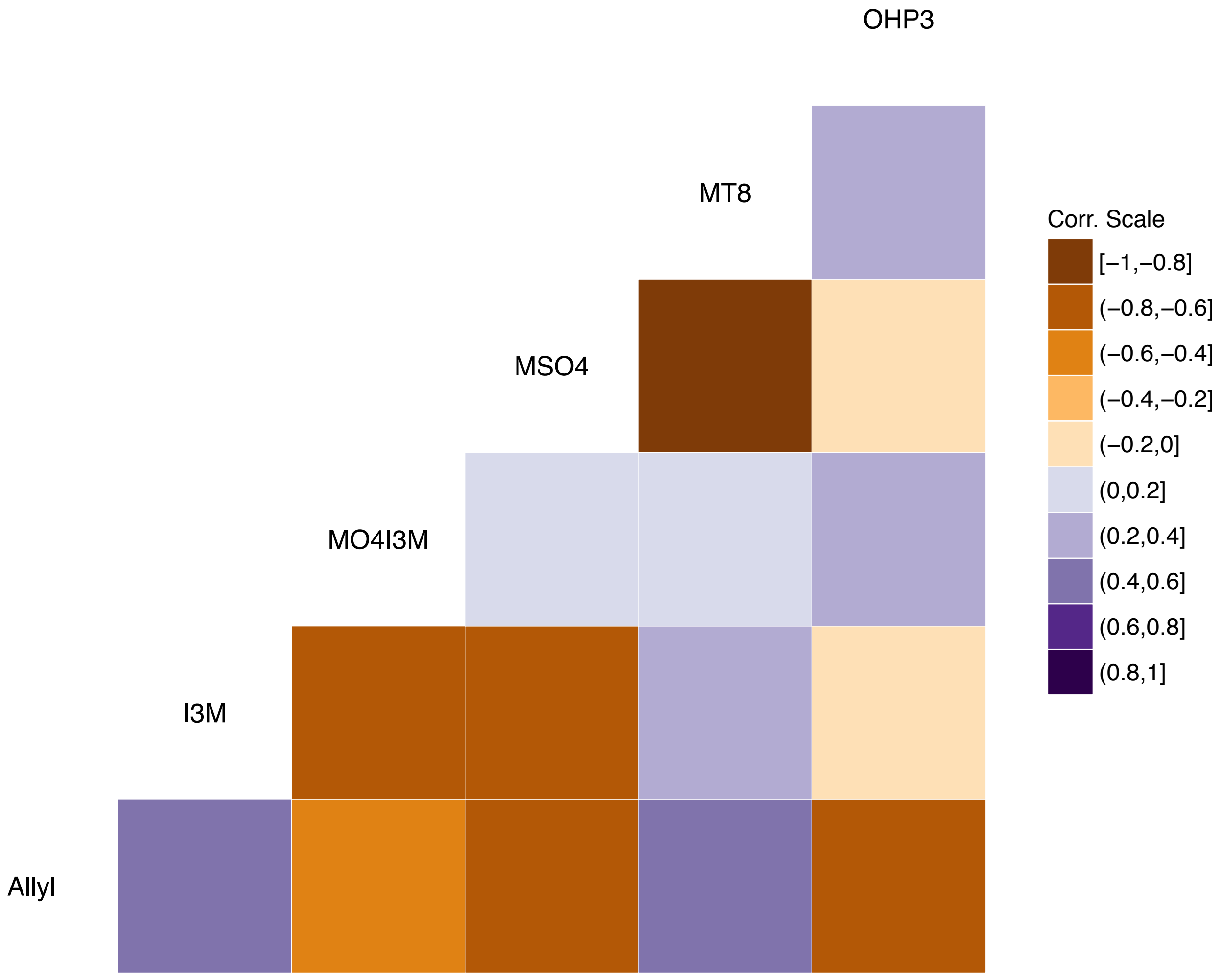


Figure S17: Lower-triangular heat map illustrating the correlation structure between the six metabolic phenotypes in the *Arabidopsis thaliana* QTL mapping study. The six traits analyzed include: allyl content, indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 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.

Supporting Information: Tables

Table S1: Mean squared errors (MSE) and predictive correlations (r) using the OLS estimates and the effect size analog. Scenarios I and II correspond to broad-sense heritability level $\mathrm{H}^2 = \{0.3, 0.6\}$ 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 Opt%$_{\mathrm{MSE}}$ and Opt%$_r$, respectively. Values in bold 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	2.33 (0.39)	**0.99 (0.14)**	2.04 (0.33)	**0.94 (0.13)**
	II	1.63 (0.30)	**0.88 (0.17)**	**0.68 (0.11)**	0.69 (0.09)
Opt%$_{\mathrm{MSE}}$	I	0.00	**1.00**	0.00	**1.00**
	II	0.00	**1.00**	**0.59**	0.41
r	I	0.11 (0.09)	**0.16 (0.10)**	0.21 (0.10)	**0.28 (0.10)**
	II	0.33 (0.10)	**0.41 (0.10)**	**0.69 (0.05)**	0.68 (0.05)
Opt%$_r$	I	0.24	**0.76**	0.18	**0.82**
	II	0.08	**0.92**	**0.55**	0.45

Table S2: Computational complexity for calculating RATE as a function of the number of variables that are present within the data and the number of available computing clusters for parallelization. Each entry represents the mean computation time (in seconds). Computations were performed using the Athena computing cluster at the Center for Statistical Sciences at Brown University. To create synthetic data for these simulations, we generated $p = \{100, 500, 1000, 2500, 5000\}$ genetic markers, respectively. Sample sizes were fixed at $n = 500$ individuals. Values in the parentheses are the standard deviations of the estimates.

	Average Time (sec)				
Computing Cores	$p = 100$	$p = 500$	$p = 1000$	$p = 2500$	$p = 5000$
$n = 1$	0.47 (0.06)	3.94 (0.49)	13.23 (3.55)	157.64 (4.98)	674.77 (24.77)
$n = 4$	0.15 (0.05)	2.13 (0.57)	8.40 (0.45)	98.05 (0.41)	439.11 (0.64)

Table S3: 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 S4: 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)

References

Crawford, L., K. C. Wood, X. Zhou, and S. Mukherjee (2017). Bayesian approximate kernel regression with variable selection. *Journal of the American Statistical Association.* Available online from https://doi.org/10.1080/01621459.2017.1361830.

Howard, R., A. L. Carriquiry, and W. D. Beavis (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics 4*(6), 1027–1046.

Kolmogorov, A. N. and Y. A. Rozanov (1960). On strong mixing conditions for stationary Gaussian processes. *Theory of Probability & Its Applications 5*(2), 204–208.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning.* Cambridge, MA: MIT Press.