Published in final edited form as:

Ann Appl Stat. 2019 June ; 13(2): 958–989. doi:10.1214/18-aoas1222.

VARIABLE PRIORITIZATION IN NONLINEAR BLACK BOX METHODS: A GENETIC ASSOCIATION CASE STUDY[1]

Lorin Crawford[*], **Seth R. Flaxman**[†], **Daniel E. Runcie**[‡], **Mike West**[§]

[*]Brown University

[†]Imperial College London

[‡]University of California, Davis

[§]Duke University

Abstract

The central aim in this paper is to address variable selection questions in nonlinear and nonparametric regression. Motivated by statistical genetics, where nonlinear interactions are of particular interest, we introduce a novel and interpretable way to summarize the relative importance of predictor variables. Methodologically, we develop the "RelATive cEntrality" (RATE) measure to prioritize candidate genetic variants that are not just marginally important, but whose associations also stem from significant covarying relationships with other variants in the data. We illustrate RATE through Bayesian Gaussian process regression, but the methodological innovations apply to other "black box" methods. It is known that nonlinear models often exhibit greater predictive accuracy than linear models, particularly for phenotypes generated by complex genetic architectures. With detailed simulations and two real data association mapping studies, we show that applying RATE enables an explanation for this improved performance.

Keywords

Nonlinear regression; Gaussian processes; centrality measures; variable prioritization; genome-wide association studies; statistical genetics

lorin_crawford@brown.edu.

[1]Supported by start-up funds from Brown University.

Author contributions statement. LC conceived the study. LC and MW developed the methods. LC, SRF, and DER developed the algorithms. LC implemented the software and performed the analyses. All authors wrote and revised the manuscript.

L. CRAWFORD, DEPARTMENT OF BIOSTATISTICS AND CENTER FOR STATISTICAL SCIENCES AND CENTER FOR COMPUTATIONAL MOLECULAR BIOLOGY, BROWN UNIVERSITY, PROVIDENCE, RHODE ISLAND 02903, USA

S. R. FLAXMAN, DEPARTMENT OF MATHEMATICS AND DATA SCIENCE INSTITUTE, IMPERIAL COLLEGE LONDON, LONDON SW7 2AZ, UNITED KINGDOM

D. E. RUNCIE, DEPARTMENT OF PLANT SCIENCES, UNIVERSITY OF CALIFORNIA, DAVIS, DAVIS, CALIFORNIA 95616, USA

M. WEST, DEPARTMENT OF STATISTICAL SCIENCE, DUKE UNIVERSITY, DURHAM, NORTH CAROLINA 27708, USA

Competing financial interests. No competing interests exist.

SUPPLEMENTARY MATERIAL

Supplement to "Variable prioritization in nonlinear black box methods: a enetic association case study." (DOI: 10.1214/18-AOAS1222SUPPA; .pdf). This file contains supplementary derivations, figures, and tables referenced in the main text.

1. Introduction.

Classical statistical models and modern machine learning methodology have recently been dichotomized into two separate groups. The former are often characterized as interpretable modeling approaches and include conventional methods such as linear and logistic regressions. The latter, however, have sparked a greater debate as they have been frequently criticized as “black box” techniques with opaque implementations and uncertain internal workings. Whenever support vector machines or neural networks give meaningful performance gains over more conventional regression models, a challenge of interpretability arises. In these situations it is often questioned what characteristics of the input data are being most used by the black box. One of the key features leading to these performance gains is the automatic inclusion of higher order interactions between variables [Cotter, Keshet and Srebro (2011)]. Popular machine learning kernel functions and fully connected neural network layers implicitly enumerate all possible nonlinear effects [Wahba (1990)]. While this fact is in itself a partial explanation for improvement gains, we often wish to know precisely which variables are the most important—with the ultimate goals of furthering scientific understanding and performing model/feature selection [Barbieri and Berger (2004)].

As our main contribution we propose a “RelATive cEntrality” (RATE) measure for investigating variable importance in Bayesian nonlinear models, particularly those considered to be black box. Here, RATE identifies variables which are not just marginally important, but also those whose data associations stem from a significant covarying relationship with other variables. Our method is entirely general with respect to the modeling approach taken; the only requirement being that a method can produce uncertainty intervals for predictions. As an illustration we focus on Gaussian process modeling with Markov chain Monte Carlo (MCMC) inference. In addition we note that this variable selection approach immediately applies to other methodologies such as Bayesian neural networks [Richard and Lippmann (1991)], Bayesian additive regression trees [Chipman, George and McCulloch (2010)] and approximate inference methods like variational Bayes [Rasmussen and Williams (2006)].

While variable selection is the main utility for our method, we are motivated by the approach of continuous model expansion [Gelman, Hwang and Vehtari (2014)]. The goal is to build the best fitting or optimally predictive model while searching over many variables and the interactions between them but without explicitly worrying about sparsity. Indeed, this has become a recent focus of statistical methods research, especially in terms of understanding the relative importance of subsets of candidate predictors with respect to specific predictive goals [Lin, Chan and West (2016)]. While we believe strongly in regularization as a key ingredient in developing good statistical models, our choice of Gaussian process priors achieves robust inference without explicitly imposing a sparsity penalty. The reason to avoid sparsity constraints like the lasso is not just philosophical—as typically applied L1-regularization suffers from a lack of stability [Lim and Yu (2016), Piironen and Vehtari (2017)], and the use of Laplacian priors too has been criticized [Carvalho, Polson and Scott (2010)]. Simultaneously, we are also motivated by the rise of deep neural networks, which

are typically wildly overparameterized, and yet, when combined with large datasets, can give quite impressive improvements to model performance.

We assess our proposed approach in the context of association mapping (i.e., inference of significant variants or loci) in statistical genetics as a way to highlight data science applications that are driven by many covarying and interacting predictors. For example, understanding how statistical epistasis between genes (i.e., the polynomial terms of the variables in the genotype matrix) influence the architecture of traits and variation in phenotypes is of great interest in genetics applications [Crawford and Zhou (2018), Crawford et al. (2017), Mackay (2014), Phillips (2008), Prabhu and Pe'er (2012), Wan et al. (2010), Zhang and Liu (2007), Zhang et al. (2010)]. However, despite studies that have detected "pervasive epistasis" in many model organisms [Horn et al. (2011)] and improved genomic selection (i.e., phenotypic prediction) using nonlinear regression models [Howard, Carriquiry and Beavis (2014)], substantial controversies remain [Hill, Goddard and Visscher (2008)]. For example, in some settings, association mapping studies have identified many candidates of statistical epistasis or interactions that contribute to quantitative traits [Hemani et al. (2014)], but some of these results can be explained by additive effects of other unsequenced variants [Wood et al. (2014)]. To date, we have a limited understanding of this important biological question because it is often difficult to pinpoint how nonlinearities influence complex prioritization of associated genetic markers. Indeed, it has been suggested that if one aims to infer biological interactions, statistically modeled interactions and main effect terms should not be interpreted separately [Wang, Elston and Zhu (2011a, 2011b)]. Our contribution in this paper is therefore of direct scientific relevance in that RATE will enable scientists to consider embracing machine learning-type approaches by allowing them to open up the black box.

The remainder of this paper is organized as follows. In Section 2 we briefly detail the Gaussian process regression model and motivate the need for an effect size (regression coefficient) analog that serves to characterize the importance of the original input variables in nonparametric methods. In Section 3 we specify how to conduct association mapping using distributional centrality measures. Here, we also define the concept of relative centrality (RATE) which provides evidence for the relative importance of each variable. In Section 4 we show the utility of our methodology on real and simulated data. Finally, we close with a discussion in Section 5.

2. Motivating Bayesian nonparametric framework.

In this paper we propose a relative centrality measure as an interpretable way to summarize the importance of input variables for nonparametric methodologies. We will do this within the context of association mapping in statistical genetics. This effort will require the utilization of three components: (i) a motivating probabilistic model, (ii) a notion of an effect size (or regression coefficient) for each genetic variant and (iii) a statistical metric that determines marker significance. Each of these components are naturally given in linear regression. Our goal is to provide a computationally tractable way to derive the same necessary components for nonlinear methods.

In this section we focus on formulating components (i) and (ii), while component (iii) is developed later in Section 3. First, we begin by detailing Bayesian Gaussian process regression as our motivating probabilistic model. Next, we generalize a previous result which defines an effect size (regression coefficient) analog for the input data in nonparametric methods [Crawford et al. (2018)]. Extensions to other methodologies (e.g., Bayesian kernel ridge regression, neural networks) can be found in Supplementary Material [Crawford et al. (2019)]. For simplicity we make the assumption that the phenotypic response is continuous; although the frameworks discussed can be altered for dichotomous traits (e.g., case-control studies). This expansion would include steps similar to those outlined in previous works [Zhang, Dai and Jordan (2011)]. We leave these specific details to the reader.

2.1. Gaussian process regression.

We now state a Bayesian modeling framework, which we use to construct a generalized projection operator between an infinite dimensional function space, called a reproducing kernel Hilbert space (RKHS), and the original genotype space. This projection will allow us to define an effect size analog for Bayesian nonparametric analyses. We begin by considering standard linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \tag{1}$$

where $\mathbf{y}$ is an n-dimensional vector of phenotypes from n individuals, $\mathbf{X}$ is an $n \times p$ matrix of genotypes for p genetic variants encoded as {0, 1, 2} copies of a reference allele at each marker, $\boldsymbol{\beta}$ is the corresponding additive effect size, $\boldsymbol{\varepsilon}$ is assumed to follow a multivariate normal distribution with mean zero and variance τ^2 and $\mathbf{I}$ is an identity matrix. For convenience we will also assume that the genotype vector has been centered and standardized to have mean 0 and standard deviation 1.

In genetic applications the assumption that phenotypic variation can be fully explained by additive effects is often too restrictive [Mackay (2014), Phillips (2008)]. One natural way to overcome this problem is to conduct model inference within a high dimensional function space. Indeed, an RKHS may be defined based on a nonlinear transformation of the data using a positive definite covariance function (or kernel) that is assumed to have a finite integral operator with eigenfunctions $\{\phi_\ell\}_{\ell=1}^{\infty}$ and eigenvalues $\{\delta_\ell\}_{\ell=1}^{\infty}$. Namely,

$$\int k(\mathbf{x}, \mathbf{x}')\, \mathrm{d}(\mathbf{x}, \mathbf{x}') < \infty, \qquad \delta_\ell \phi_\ell(\mathbf{x}) = \int_{\chi} k(\mathbf{x}, \mathbf{x}') \phi_\ell(\mathbf{x}') \mathrm{d}\mathbf{x}'.$$

For these classes of covariance functions, the following infinite expansion holds $k(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{\infty} \delta_\ell \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}')$ [Mercer (1909)], and an RKHS function space may be formally defined via the closure of a linear combination of basis functions [Pillai et al. (2007)]. As a direct result we rewrite equation (1) as the following RKHS regression model [Zhang, Dai and Jordan (2011)]:

$$\mathbf{y} = \mathbf{\Psi}^{\mathsf{T}}\mathbf{c} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \tag{2}$$

where $\boldsymbol{\Psi}(\mathbf{x}) = \{\sqrt{\delta_\ell}\phi_\ell(\mathbf{x})\}_{\ell=1}^{\infty}$ is a vector space spanned by the bases, $\mathbf{\Psi} = [\boldsymbol{\psi}(\mathbf{x}_1), \ldots, \boldsymbol{\psi}(\mathbf{x}_n)]^{\mathsf{T}}$ is a corresponding matrix of concatenated basis functions and $\mathbf{c} = \{c_\ell\}_{\ell=1}^{\infty}$ are the corresponding basis coefficients. The above specification in equation (2) closely resembles the linear model in equation (1)—except now the bases are the feature vectors $\boldsymbol{\psi}(\mathbf{x})$ (rather than the unit basis), and the transformed space can be infinite dimensional. Theoretically, this is an important property because the inclusion of nonlinear interactions and covarying relationships are implicitly captured in the RKHS.

Unfortunately, properly representing any given basis function in an empirically amenable form is a difficult task [Schölkopf, Herbrich and Smola (2001)]. To circumvent this analytical issue, one may alternatively conduct inference in an RKHS by specifying a Gaussian process (GP) as the prior distribution over the function space directly

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

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, if we condition on a finite set of locations (i.e., the set of observed samples n), the Gaussian process prior then becomes a multivariate normal [Kolmogorov and Rozanov (1960)]. By specifying a joint version of the nonparametric regression model above, we consider taking a "weight-space" view on Gaussian processes [Rasmussen and Williams (2006)],

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}, \qquad \mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \tag{3}$$

where, in addition to previous notation, $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^{\mathsf{T}}$ is assumed to come from a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{K} = \mathbf{\Psi}^{\mathsf{T}}\mathbf{\Psi}$ with each element $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Altogether, we refer to the family of models taking on this form as GP regression. The formulation of the weight space GP is similar to the linear mixed model (LMM) [Lippert et al. (2011), Zhou and Stephens (2012)] that is frequently used in genetics but with one key difference; the GP model utilizes a nonlinear covariance matrix $\mathbf{K}$ instead of the usual gram matrix $\mathbf{X}\mathbf{X}^{\mathsf{T}}/p$. From this perspective an RKHS model can be viewed as an extension of the LMM for modeling nonlinear effects such as statistical interactions. Indeed, the GP model still presents the same modeling benefits as an LMM, such as controlling for structured random effects. For example, notice that the Gaussian covariance function can be written as a product of three terms [Cotter, Keshet and Srebro (2011)

$$\exp\left\{-\frac{1}{2\theta^2}\|\mathbf{x}-\mathbf{x}'\|^2\right\} = \exp\left\{-\frac{1}{2\theta^2}(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2)\right\}\exp\left\{-\frac{1}{2\theta^2}\mathbf{x}^{\mathsf{T}}\mathbf{x}'\right\}.$$

The last term includes (nonlinear transformed) elements of the LMM relatedness matrix that has been well known to effectively control for population stratification in genetic studies [Kang et al. (2010), Wu et al. (2011), Yang et al. (2014), Zhou and Stephens (2014)].

Because of these properties, RKHS-based models have become powerful tools for predictive problems in many research areas and have been widely used for genomic selection in animal breeding programs [de los Campos et al. (2009, 2010)]. We replicate some of these sentiments via a small simulation study (see the Supplementary Material and Table S1).

Lastly, we want to point out that (although not explicitly considered here) the formulation of the GP regression model in equation (3) can also be easily extended to accommodate other fixed effects (e.g., age, sex or genotype principal components) [de los Campos et al. (2009), Shi et al. (2012)] as well as be adapted to account for interactions between variants and nongenetic risk factors [Cuevas et al. (2017), Weissbrod, Geiger and Rosset (2016)].

Note on bandwidth parameters.—In many cases the covariance function is indexed by a bandwidth parameter θ (also known as a smoothing parameter or lengthscale), which we expansively write as $k_\theta(\bullet, \bullet)$. For example, the previously mentioned Gaussian kernel can be specified as $k_\theta(\mathbf{x},\mathbf{x}') = \exp\{-\|\mathbf{x} - \mathbf{x}'\|^2/2\theta^2\}$. Within a fully Bayesian model this bandwidth parameter can be assigned a prior distribution, and its posterior distribution may be inferred [Zhang, Dai and Jordan (2011)]. However, for simplicity we follow recent studies using the "median heuristic" and work with a fixed bandwidth that we choose as $\theta = \text{median}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$ [Chaudhuri et al. (2017)].

Posterior inference and sampling.—We now briefly detail a simple MCMC sampling procedure for estimating the parameters in GP regression. Assume now that we have a completely specified hierarchical model

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}, \qquad \mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \qquad \tau^2 \sim \text{Scale-Inv} - \chi^2(a, b),$$

where, in addition to previous notation, we further assume that the residual variance parameter τ^2 follows a scaled-inverse chi-square distribution with degrees of freedom a and scale b as hyper parameters. Given the conjugacy of this model specification, we may use a Gibbs sampler to estimate the joint posterior distribution $\mathcal{P}(\mathbf{f}, \tau^2 \mid \mathbf{y})$. This consists of iterating between the following two conditional densities:

1. $\mathbf{f} \mid \tau^2, \mathbf{y} \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*)$ where $\mathbf{m}^* = \mathbf{K}(\mathbf{K} + \tau^2\mathbf{I})^{-1}\mathbf{y}$ and $\mathbf{V}^* = \mathbf{K} - \mathbf{K}(\mathbf{K} + \tau^2\mathbf{I})^{-1}\mathbf{K}$;

2. $\tau^2 \mid \mathbf{f}, \mathbf{y} \sim$ Scale-Inv-$\chi^2(a^*, b^*)$ where $a^* = a + n$ and $b^* = a^{*-1}[ab + (\mathbf{y} - \mathbf{f})^\intercal(\mathbf{y} - \mathbf{f})]$.

Iterating the above procedure T times results in a set of sampled draws from the target joint posterior distribution. Taking the mean over these draws yields posterior estimates for the model parameters (see the Supplementary Material for a detailed algorithmic overview).

2.2. Effect size analog for nonparametric methods.

A noteworthy downside to the GP regression model is the inability to find an effect size for causal variants. From a prediction and genomic selection perspective this loss is fine, but from the perspective of finding genetic markers that give rise to this improved predictive performance (i.e., association mapping) the interpretability of the model is lost. We now define the effect size analog for general nonparametric methods as a solution to this

limitation [Crawford et al. (2018)]. We first briefly outline the conventional wisdom for coefficients in linear regression. In linear models a natural interpretation of a regression coefficient is the projection of the genotypes **X** onto the phenotypic vector **y**,

$$\widehat{\boldsymbol{\beta}} = \text{Proj}(\mathbf{X}, \mathbf{y}), \tag{4}$$

with the choice of loss function, noise model as well as prior distributions or regularization penalties specifying the exact form of the projection. One standard projection operation is $\text{Proj}(\mathbf{X}, \mathbf{y}) = \mathbf{X}^{\dagger}\mathbf{y}$, where $\mathbf{X}^{\dagger}$ is the Moore–Penrose generalized inverse. For Bayesian procedures priors over the parameters $\boldsymbol{\beta}$ induce a distribution on the resulting projection procedure Proj (**X**, **y**) [Carvalho, Polson and Scott (2010), Liang et al. (2008)].

The general definition for the effect size analog is based on the similar idea of projecting a nonlinear function onto the design matrix. Specifically, consider a nonlinear function evaluated on n-observed samples such that $\mathbb{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{f}$. We formally define the *effect size analog* as the result of projecting the genotypic matrix **X** onto the nonlinearly estimated function vector **f**,

$$\widetilde{\beta} = \text{Proj}(\mathbf{X}, \mathbf{f}). \tag{5}$$

This projection operation and its practical calculation effectively requires two sets of coefficients: (i) the theoretical coefficients **c** on the basis functions; and (ii) the coefficients that determine the effect size analog $\widetilde{\beta}$. Following the formulation in equation (5), we use equations (2) and (3) to specify the joint projection of design matrix **X** onto the vector $\mathbf{f} = \boldsymbol{\Phi}^{\intercal}\mathbf{c}$ as the linear map,

$$\widetilde{\boldsymbol{\beta}} = \mathbf{X}^{\dagger}\boldsymbol{\Psi}^{\intercal}\mathbf{c} = \mathbf{X}^{\dagger}\mathbf{f}. \tag{6}$$

The argument for why the p-dimensional vector $\widetilde{\boldsymbol{\beta}}$ is an effect size analog for nonparametric regression models is that, on the n-observations, $\mathbf{f} \approx \mathbf{X}\widetilde{\boldsymbol{\beta}}$. In the Supplementary Material we rederive previous results to formally show that the map from **f** to $\widetilde{\boldsymbol{\beta}}$ is injective modulo the null space of the genotypic matrix [Crawford et al. (2018)]. This is similar to the classical linear regression case where two different coefficient vectors will result in the same estimated value if the difference between the vectors is in the null space of **X**. Additionally, the only requirement for equation (6) is a well-defined feature map $\psi(\bullet)$. This includes taking the Cholesky decomposition of the covariance matrix as a feature map, or even employing low-rank approximations such as the Nystrom approximation [Drineas and Mahoney (2005)], random Fourier features [Rahimi and Recht (2007)] or explicit Mercer expansions [Fasshauer and McCourt (2016)]. We should be clear that a variety of projection procedures (corresponding to various priors and loss functions) can be specified, and a systematic study elucidating which projections are efficient and robust is of interest for future research.

A key motivation for the effect size analog is to conduct nonlinear association mapping in the original genotype space while also accounting for population structure and significant covarying relationships between variants. When a phenotype or trait is solely driven by

additive effects, the projections (4) and 5) with the same genotypes $\mathbf{X}$ are equivalent, and the resulting effect size analog from equation (6) is the same as the OLS estimate derived by a standard linear model. Alternatively, it has been shown (via Taylor series expansions) that certain covariance functions enumerate nonlinear effects among observed markers [Jiang and Reif (2015)]. The Gaussian kernel, in particular, includes all higher-order interaction components, where the contribution of the terms decays polynomially with the order of nonlinearity [Cotter, Keshet and Srebro (2011)]. Therefore, when a given phenotype is driven by an arbitrary combination of additivity and interactions, a properly chosen nonlinear map $\psi(\bullet)$ will lead to an inversion in equation (6) that represents each $\widetilde{\beta}_j$ as a weighted sum of higher order interactions between marker j and all other markers (see text in the Supplementary Material).

3. Genetic association mapping using centrality measures.

The effect size analog serves as a nonlinear summary coefficient for each genetic variant in the original modeling space. However, since the explicit projection in equation (6) does not always guarantee a preserved mapping of sparse solutions [Crawford et al. (2018)], we cannot directly use standard Bayesian quantities such as posterior inclusion probabilities (PIPs) or Bayes factors (BFs) to rank markers in order of their significance. Indeed, there are many approaches to compute marginal association statistics based on corresponding effect size estimates [Barbieri and Berger (2004), Stephens and Balding (2009)], but many of these techniques rely on arbitrary thresholding. More importantly, they also fail to take advantage of significant underlying dependencies and covarying relationships between variants or sets of genomic loci.

We now develop our main methodological innovation. We introduce an analogy to traditional Bayesian hypothesis testing for nonparametric regression methods, a post-hoc approach for association mapping via a series of “distributional centrality measures” using Kullback–Leibler divergence (KLD) [Goutis and Robert (1998), Smith, Naik and Tsai (2006), Tan et al. (2017), Woo et al. (2015), Piironen and Vehtari (2016,2017), Alaa and van der Schaar (2017)]. Our strategy will be to use the posterior samples of the effect size analogs to infer the relative covariance between genetic variants. This underlying correlation structure will then be systematically searched over to posit significant individual associations. We refer to this approach as computing the RATE of genetic markers.

3.1. Kullback–Leibler divergence.

Typical questions in network studies simplify to the general issue of determining the “centrality” of nodes—the potential importance of individual components in relation to the other nodes in the entire network. When network relationships are modeled via multivariate distributions, this can be explored in various statistical ways. Assume here that we have a collection of deterministically computed samples from the implied posterior distribution of the effect size analog $\widetilde{\boldsymbol{\beta}}$ (via the projection in equation (6)). One interpretable way to summarize (in a single measure) the influence/importance of the jth variant in $\mathbf{x}_j$, on the rest of the variants in $\mathbf{X}_{-j}$, is via the computation of the KLD measuring the difference between $\mathscr{P}(\widetilde{\boldsymbol{\beta}}_{-j} \mid \widetilde{\beta}_j)$ and $\mathscr{P}(\widetilde{\boldsymbol{\beta}}_{-j})$. Specifically, this is defined by solving the following integral:

$$\mathrm{KLD}(\tilde{\beta}_j) = \int_{\tilde{\boldsymbol{\beta}}_{-j}} \log\left(\frac{\mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j})}{\mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j)}\right) \mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j}) \mathrm{d}\tilde{\boldsymbol{\beta}}_{-j}, \qquad j = 1, \ldots, p, \tag{7}$$

where we use the shorthand $\mathrm{KLD}(\tilde{\beta}_j) = \mathrm{KLD}(\mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j}) \parallel \mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j))$. Here, the KLD is a nonnegative quantity and in this context takes the value of zero if and only if $\mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j) = \mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j})$. Equivalently, this means that the KLD is zero if and only if the posterior distribution of $\tilde{\boldsymbol{\beta}}_{-j}$ is independent of the effect $\tilde{\beta}_j$. Therefore, the case for which $\mathrm{KLD}(\tilde{\beta}_j) = 0$ may simply be interpreted as meaning that variant j is not a key explanatory variable relative to others. Otherwise, for any given conditioning value $\tilde{\beta}_j$ the divergence in equation (7) represents the information (i.e., entropy) change induced on the distribution of $\tilde{\boldsymbol{\beta}}_{-j}$—naturally varying as the conditioning value $\tilde{\beta}_j$ varies.

Closed form derivation under approximate normal posteriors.—For our case study and immediate applications we are interested in straightforward computation of KLD measures in order to address problems with increasingly large numbers of genotypes and possible interactions. For these purposes and for the rest of the paper, we therefore restrict attention to contexts in which we can assume an adequate normal approximation to the full joint posterior distribution of the p-dimensional effect size analog $\tilde{\boldsymbol{\beta}}$. Ongoing and future work is concerned with computational and numerical aspects of the more general context, while the methodological and applied advances enabled by our approach are well-highlighted under the normal posterior assumption.

Thus, we take the posterior for $\tilde{\boldsymbol{\beta}}$ as (approximately) multivariate normal with an empirical mean vector $\boldsymbol{\mu}$ and positive semidefinite covariance/precision matrices $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$ estimated via simulation methods. Consider the association mapping case where we want to investigate the centrality or marginal importance of marker j. We may partition conformably as follows:

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

where $\tilde{\beta}_j$, μ_j, σ_j and λ_j are scalars; $\tilde{\boldsymbol{\beta}}_{-j}$, $\boldsymbol{\mu}_{-j}$, $\boldsymbol{\sigma}_{-j}$ 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. Under this partitioning we know that the marginally $\tilde{\boldsymbol{\beta}}_{-j} \sim \mathscr{N}(\boldsymbol{\mu}_{-j}, \boldsymbol{\Sigma}_{-j})$. Furthermore, we also know that, when conditioned on the jth variant, $\mathscr{P}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j)$ is a multivariate distribution with expectation and covariance

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j) = \boldsymbol{\mu}_{-j} + \boldsymbol{\theta}_j(\tilde{\beta}_j - \mu_j), \qquad \mathbb{V}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j) = \boldsymbol{\Lambda}_{-j}^{-1},$$

where $\boldsymbol{\theta}_j = -\boldsymbol{\Lambda}_{-j}^{-1}\boldsymbol{\lambda}_{-j}$ is a $(p-1)$-dimensional vector. Inserting these probability density forms into equation (7) with some algebraic rearrangement yields the following:

$$\mathrm{KLD}(\widetilde{\beta}_j) = \frac{1}{2}[-\log|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}| + \mathbb{E}(\mathbf{e}_{-j}^{\intercal}\boldsymbol{\Lambda}_{-j}\mathbf{e}_{-j}) - 2\mathbb{E}(\mathbf{e}_{-j}^{\intercal})\boldsymbol{\Lambda}_{-j}\boldsymbol{\theta}_j e_j - \mathbb{E}(\mathbf{e}_{-j}^{\intercal}\boldsymbol{\Sigma}_{-j}^{-1}\mathbf{e}_{-j}) + e_j^2\boldsymbol{\theta}_j^{\intercal}\boldsymbol{\Lambda}_{-j}\boldsymbol{\theta}_j], \tag{8}$$

where $\log|\bullet|$ represents the log determinant function of a matrix, $\mathbf{e}_{-j} = \widetilde{\boldsymbol{\beta}}_{-j} - \boldsymbol{\mu}_{-j}$ is a vector, $e_j = \widetilde{\beta}_j - \mu_j$ is a scalar and the expectations are taken with respect to the marginal posterior distribution of $\widetilde{\boldsymbol{\beta}}_{-j}$. Next, denote the following definition of an expectation of quadratic forms [Mathai and Provost (1992)],

$$\mathbb{E}(\mathbf{u}^{\intercal}\mathbf{Q}\mathbf{u}) = \mathbb{E}(\mathbf{u}^{\intercal})\mathbf{Q}\mathbb{E}(\mathbf{u}) + \mathrm{tr}(\mathbb{V}(\mathbf{u})\mathbf{Q}),$$

for any vector $\mathbf{u}$ and positive semidefinite covariance matrix $\mathbf{Q}$, where $\mathrm{tr}(\bullet)$ is the matrix trace function. Using this equality, the computation of the KLD in equation (8) simplifies to the following closed form

$$\mathrm{KLD}(\widetilde{\beta}_j) = \frac{1}{2}[-\log(|\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}|) + \mathrm{tr}(\boldsymbol{\Sigma}_{-j}\boldsymbol{\Lambda}_{-j}) + 1 - p + \alpha_j(\widetilde{\beta}_j - \mu_j)^2], \tag{9}$$

where $\alpha_j = \boldsymbol{\theta}_j^{\intercal}\boldsymbol{\Lambda}_{-j}\boldsymbol{\theta}_j = \boldsymbol{\lambda}_{-j}^{\intercal}\boldsymbol{\Lambda}_{-j}^{-1}\boldsymbol{\lambda}_{-j}$ and $\mathrm{tr}(\mathbf{I}) = p-1$. By symmetry in the notation for elements of subvectors and submatrices, it trivially follows that we may simply permute the order of the variables in $\widetilde{\boldsymbol{\beta}}$ and iteratively compute the KLD to measure the centrality of any variant *j*.

3.2. Prioritization and relative significance.

In the nonlinear regression context values $\widetilde{\beta}_j$ close to zero may be interpreted as "null hypotheses" with little to no relevance to the modeled outcome. Therefore, searching for the most central (i.e., influential) genetic markers simply reduces to looking for the greatest KLD when setting each $\widetilde{\beta}_j = 0$. More contextually specific questions arise when deciding if a given centrality measure is significant. Indeed, in practice a threshold may be chosen in order to determine if any given KLD represents a significant shift in entropy. Previous studies have done this through *k*-fold permutation to find an effective genome-wide threshold [Woo et al. (2015)]. This approach can be costly for datasets with many markers.

A more computationally efficient option for determining a natural ranked cutoff is to explore the relevance of variables recursively and to judge their significance via a scaled version of the KLD. We call this "RelATive cEntrality" or RATE,

$$\mathrm{RATE}(\widetilde{\beta}_j) = \mathrm{KLD}(\widetilde{\beta}_j) \,/\, \sum \mathrm{KLD}(\widetilde{\beta}_\ell), \qquad \sum \mathrm{RATE}(\widetilde{\beta}_j) = 1\,. \tag{10}$$

Here, the RATE measure is bounded within the range [0, 1] with the natural interpretation of measuring a variable's relative importance. Suppose that *j* identifies the genetic marker with the largest RATE value. Conditioning on a reduced margin and then repeating the

computation outlined in equations (9) and (10) will identify the relatively second most explanatory marker. We can repeat this procedure until each of the remaining variants appear to be equal in their relative importance. This would indicate that all significant variants had been identified, and all that remain are variants for which their influences on the posterior distribution are indistinguishable. This recursive process can be simplified to defining an initial set of candidate associated markers with first order centrality measures satisfying

$$\{j : \mathrm{RATE}(\widetilde{\beta}_j) > 1 / p\}.$$

The value $1/p$ represents the null assumption that there is relatively equal importance across all variants; hence, there are no central nodes that exist within the posterior distribution. We may quantify this behavior by checking the entropic difference between a uniform distribution and the observed RATE measures. Namely,

$$\Delta = \log(p) - H, \qquad H = -\sum \mathrm{RATE}(\widetilde{\beta}_j)\log(\mathrm{RATE}(\widetilde{\beta}_j)), \tag{11}$$

where H represents the intrinsic entropy of the relative centrality measures, and the case of no significantly associated markers yields an entropy of $\log(p)$. One way to calibrate Δ is linked to effective sample size (ESS) measures from importance sampling [Gruber and West (2016, 2017)]. In a very different applied context authors have exploited the use of an approximate ESS measure defined by

$$\mathrm{ESS} = 1 / (1 + \Delta) \times 100\,\%. \tag{12}$$

This ESS measure is a calibration metric that provides a notion of “loss in uniformity”. For example, 50% loss in terms of (1 − ESS) translates to a larger Δ value of 1. This equates to the presence of at least one variant that is significantly associated with the observed phenotypic trait. On the other hand a minor 5% loss corresponds to a more uniform case with Δ value of about 0.05. Again, this latter scenario would occur when there are hardly any influential markers within the data.

For any given set of significant variables, according to their estimated RATE measure, further analyses may be carried out involving the relative costs of false positives and negatives to make an explicitly reasoned decision about which specific variants to pursue [Stephens and Balding (2009)]. Unless stated otherwise, the results we present throughout the rest of the paper will be based on using RATE. We explore the power of this alternative approach for association mapping in Section 4.

3.3. Relationship to graphical models and precision analysis.

In conventional statistics the proposed variable selection procedure is very much related to precision analysis. It follows that the rate of change for the KLD (i.e., the first derivative of equation (9) with respect to a given effect size analog) is found via the term $\alpha_j = \boldsymbol{\lambda}_{-j}^{\intercal}\boldsymbol{\Lambda}_{-j}^{-1}\boldsymbol{\lambda}_{-j}$. This means that the closed form computation of the KLD is directly impacted by the deviations between the approximation of a given predictor’s posterior mean and the assumption that its true effect is zero. Therefore, α_j characterizes the implied linear rate of

change of information when the effect of any predictor is absent—thus, providing a natural (nonnegative) numerical summary of the role of $\tilde{\beta}_j$ in the multivariate distribution. In terms of weightings from the precision matrix, we see the following equivalent representation for the rate of change of the KLD,

$$\alpha_j = \sum_{k \neq j} \sum_{\ell \neq j} c_{k\ell} \lambda_{jk} \lambda_{j\ell},$$

where $c_{k\ell}$ is the corresponding k-ℓth element of the matrix $\mathbf{\Lambda}_{-j}^{-1}$. As derived in the previous subsection, we may alternatively denote $\alpha_j = \theta_{-j}^{\mathsf{T}} \mathbf{\Lambda}_{-j} \theta_{-j}$, where again $\theta_{-j} = -\mathbf{\Lambda}_{-j}^{-1} \lambda_{-j}$ is a $(p-1)$-dimensional vector and $\mathbf{\Lambda}_{-j}$ is the precision matrix of the conditional distribution $\mathcal{P}(\tilde{\boldsymbol{\beta}}_{-j} \mid \tilde{\beta}_j)$. These representations help show that, in the context of normal statistical regression, α_j computes the "variance explained" (i.e., the fitted sum-of-squares) by each covariate j.

The idea of variable selection via entropic shifts also has a key connection to graphical models. Often the goal of graphical models is to investigate if the precision matrix has some off-diagonal series corresponding to an underlying conditional independence structure between predictor [Carvalho and West (2007)]. RATE—a relative distributional centrality measure that assesses importance (or influence) of each variable on the network of relationships reflected in the graph—is greatly affected by the graphical structure resulting from the implied zeros in $\mathbf{\Lambda}$. A missing edge between two predictors j and ℓ means that $\lambda_{j\ell} = 0$; hence, limiting the contribution of node ℓ to the overall "network impact factor" of α_j. From the sum defining α_j above, we see that a term related to variables k and ℓ is nonzero only when both λ_{jk} and $\lambda_{j\ell}$ are nonzero. Therefore, the k-ℓth summation term is nonzero only for pairs of predictors that are direct neighbors of j in an undirected graph.

3.4. Software implementation.

Software for computing the RATE measure is carried out in R code which is freely available at https://github.com/lorinanthony/RATE. Detailed derivations of the algorithm, which utilizes low-rank matrix factorizations for a more practical implementation, are derived in the Supplementary Material [Crawford et al. (2019)].

4. Results.

We now illustrate the utility of using centrality measures for genetic association mapping through extensive simulation studies and real data analyses. The motivation for each set of examples is to better understand the performance and behavior of RATE under different types of genetic architectures. First, we use a small simulation study to help the reader build a stronger intuition about how RATE prioritizes influential variables in a dataset. It is during this demonstration where we also explore what happens to the concepts of "centrality" and "uniformity," when the effects of all known significant markers are assumed to be absent from the model. Next, we use more realistic simulations to assess the mapping power of our approach in genetic-based applications. Here, the goal is to show that RATE performs association mapping as well as the most commonly used Bayesian and regularization

modeling techniques. Finally, we assess the potential of the our approach in two real datasets. The first is an *Arabidopsis thaliana* QTL mapping study consisting of six different metabolic traits from an F6 Bay-0 × Shahdara recombinant inbred lines (RILs) population. The second is a genome-wide association study (GWAS) in a heterogeneous stock of mice from the Wellcome Trust Centre for Human Genetics.

4.1. Simulation studies.

For all synthetic demonstrations and assessments we consider a simulation design that is often used to explore the utility of statistical methods across different genetic architectures underlying complex phenotypic traits [Crawford and Zhou (2018), Crawford et al. (2017), Zeng and Zhou (2017)]. First, we 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 denote as H^2. From the more conventional statistics perspective the parameter H^2 can alternatively be described as a factor controlling the signal-to-noise ratio in the simulations. Next, we use a genotypic matrix $\mathbf{X}$ with n samples and p single nucleotide polymorphisms (SNPs) to generate synthetic real-valued phenotypes that mirror genetic architectures affected by a combination of linear (additive) and interaction (epistatic) effects.

We randomly choose a select subset of j^* "causal" (or truly associated) SNPs as the determining factors of the data generating process. The linear effect sizes for all j^* associated genetic variants are assumed to come from a standard normal distribution, $\beta_{j^*} \sim \mathcal{N}(0, 1)$. When applicable, we also create a separate matrix $\mathbf{W}$ which holds all pairwise interactions between the causal SNPs. These corresponding interaction effect sizes are 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 ρ%, while the pairwise interactions make up the remaining $(1 - \rho)$%. Alternatively, the proportion of the heritability explained by additivity is said to be $\mathbb{V}(\mathbf{X}\boldsymbol{\beta}) = \rho H^2$, while the proportion detailed by nonlinearity is given as $\mathbb{V}(\mathbf{W}\boldsymbol{\gamma}) = (1 - \rho)H^2$. 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. 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 $(1 - H^2)$% of the total $\mathbb{V}(\mathbf{y})$. Finally, continuous phenotypes are then created by summing over all observed effects using two simulation models:

- **i.** Standard model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$.
- **ii.** Population stratification model: $\mathbf{y} = \mathbf{Z}\boldsymbol{\omega} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$,

where $\mathbf{Z}$ contains covariates representing additional population structure, and $\boldsymbol{\omega}$ are the corresponding fixed effects which are also assumed to follow a standard multivariate normal distribution. Alternatively, one can think of the combined effect of $\mathbf{Z}\boldsymbol{\omega}$ as structured noise. To this end, simulations under model (ii) will make the appropriate assumption that $\mathbb{V}(\mathbf{Z}\boldsymbol{\omega}) + \mathbb{V}(\boldsymbol{\varepsilon}) = (1 - H^2)$. For any simulations conducted under model (ii), genotype PCs are

not included in any of the model fitting procedures, and no other preprocessing normalizations were carried out to account for the added population structure.

It is helpful to point out here that the main purpose of the following simulations is to demonstrate the utility of RATE in providing an explicit ranking of variable importance, so as to uncover the implicit ranking assigned by nonparametric regression methods. Our simulation comparisons are thus targeted to illustrate how RATE can be used in this task, and how its overall variable selection performance differs from standard parametric mapping procedures in different scenarios.

4.1.1. Proof of concept simulations: Demonstrating centrality.—In this subsection we show how distributional centrality measures may be used and interpreted when prioritizing genetic markers in an association mapping study. Our main concern is to familiarize the reader with the behavior and concepts underlying RATE. To do this, we make use of $n = 2000$ synthetic genotypes that are independently generated to have $p = 25$ 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 create continuous phenotypes using the standard generative model (i). Here, we assume that only the last three variants $j^* = \{23, 24, 25\}$ are causal, and that their combined genetic effects make up $H^2 = 60\%$ of the total phenotypic variation. We then examine the full two cases for the parameter $\rho = \{0.5, 1\}$. As a brief reminder ρ represents the proportion of broad-sense heritability that is contributed by additivity versus interaction effects. Indeed, these simulation assumptions are not realistic in terms of the qualities observed in real data applications; however, we stress that this section merely serves as a simple demonstration of "centrality" and "uniformity." The small number of variants allows us to clearly illustrate and visualize these proofs of concepts.

Throughout the rest of this subsection, we detail the behavior RATE in the simple linear case with $\rho = 1$. Similar results for $\rho = 0.5$ can be found in the Supplementary Material. For each simulation we fit a standard GP regression model under a zero mean prior and a Gaussian covariance function using a Gibbs sampler with 10,000 MCMC iterations and hyper-parameters set to $a = 5$ and $b = 2/5$. During each iterate a corresponding nonlinear projection is computed as in equation (5). This results in an approximation of the implied posterior distribution for the effect size analog. With these conditional draws we calculate the distribution's empirical posterior mean, covariance, and precision. Next, we use the closed form solutions in equations (9) and (10) to derive a RATE measure for each genetic marker.

Figure 1(a) depicts an illustration of first order centrality across the 25 variants. Here, the three known causal SNPs are colored in blue. As a reference we also display a red dashed line that is drawn at the level of relative equivalence (i.e., $1/p$). This represents the value for which all variants are approximately uniform in their centrality or significance. To put this into better context, we provide uniformity checks: (i) the entropic difference Δ according to equation (11) and (ii) the corresponding empirical ESS estimate as computed in equation (12). In this first panel figure we see that RATE accurately determines variants #23–25 as being the most central to the posterior distribution.

To demonstrate what it conceptually means to be central to a distribution, we next consider a series of follow-up analyses. Here, we iteratively assume that the genetic effect of the most significantly associated SNP has been nullified from the dataset. We then condition on a reduced margin for the posterior distribution and recompute the RATE measures. The key takeaway is that, without the effect of the data's most influential SNPs, the relative importance of the remaining variants will continue to increase until each of them are approximately equal in weight—hence, resembling a uniform distribution. Consider the ongoing example and assume that we nullify the effect of variant # 24. After recomputing $\text{RATE}(\widetilde{\beta}_j \mid \widetilde{\beta}_{24}) = 0$ for every $j \neq 24$th variant, we see that while markers #25 and #23 are still the most significant according to their second order centrality; the importance levels of the other markers have shifted closer to becoming relatively equivalent (see Figure 1(b)). This shift continues when the effects of the remaining causal variants are also removed successively (see Figures 1(c) and (d) respectively). Also notice during this transition, $\Delta \rightarrow 0$ and $\text{ESS} \rightarrow 100\%$. This same trend happens both in the presence of interaction effects (Figure S2), as well as when the causal variants are in nearly perfect collinearity (or "linkage disequilibrium" (LD)) with noncausal markers (Figures S3 and S4). In the latter case we force variants # 23–25 to have a correlation coefficient $R = 0.9$ with variants #1–3.

It is also important to demonstrate what happens to the proposed centrality measures if one mistakenly removes the effect of a genetic marker that is not central to explaining the observed phenotypic variation. Reconsider the ongoing example where, instead of iteratively removing the effect of the most central variant, we simply nullify the effect of markers #1–3, which we know to be nonsignificant (see Figure 2(a)). Figures 2(b)–(d) (and Figure S5) illustrate that the three true causal variants (i.e., markers #23–25) are continuously identified as the most associated or central to the overall posterior distribution. Noticeably, with each passing removal of a noncentral variant, the degree to which the RATE measures begin to look uniform has slowed substantially.

As a final demonstration we show what happens when the null assumptions of relative centrality are met. Recall that under the null hypothesis, RATE assumes that every variant equally contributes to the broad-sense heritability of a trait— that is, no one SNP is more important or more central than the others. To illustrate this, we generate synthetic phenotypes such that the effect sizes of all 25 SNPs in the data are set to 1. Figure S6 shows results from four different datasets. The key takeaway here is that in these cases RATE produces much more uniformly distributed first order centrality measures as indicated by the entropic statistics Δ and ESS. For completeness, in Figure S7 we also show what happens to the raw and unscaled KLDs when phenotypes have been permuted.

4.1.2. Power assessment and method comparisons.—

We now assess the power of RATE and its ability to effectively prioritize truly associated variants under different genetic architectures. To do this, we now consider simulations that mirror more realistic genetic applications. Here, we utilize real genotypes from chromosome 22 of the control samples in the Wellcome Trust Case Control Consortium (WTCCC) 1 study [The Wellcome Trust Case Control Consortium (2007)] (http://www.wtccc.org.uk/) to generate continuous phenotypes (see the Supplementary Material for details). Exclusively considering this group

of individuals and SNPs leaves us with an initial dataset consisting of $n = 2938$ samples and $p = 5747$ markers. During each simulation run we randomly choose $j^* = 30$ SNPs, which we classify into the two distinct causal groups: (1) a small set of five variants, and (2) a larger set of 25 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). We will consider three simulation scenarios. Scenario I involves phenotypes generated by standard model (i); while scenarios II and III consider model (ii) where we introduce population stratification effects by allowing the top five 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 set the broad-sense heritability to be $H^2 = 0.3$ and consider two choices for the parameter $\rho = \{0.5, 1\}$.

We compare the GP regression model and our proposed distributional centrality measures to a list of standard Bayesian and regularization modeling techniques. Specifically, these methods include: (a) a genome scan with a single-SNP univariate linear model that is typically used in GWAS applications (SCANONE) [Yandell et al. (2007)], (b) L1-regularized lasso regression; (c) the combined regularization utilized by the elastic net [Waldmann et al. (2013)]; and (d) a commonly used spike and slab prior model, also commonly known as Bayesian variable selection regression [Guan and Stephens (2011)] which places a prior distribution on each SNP as a mixture of a point mass at zero and a diffuse normal centered around zero. For each Bayesian method we run a Gibbs sampler for 10,000 MCMC iterations. Regularization approaches were fit by first learning tuning parameter values via 10-fold cross validation.

All results described in the main text are based on scenarios I and II, while results for scenario III can be found in the Supplementary Material. We evaluate each method's ability to effectively prioritize the causal SNPs in 100 different simulated datasets. The criteria we use compares the false positive rate (FPR) with the rate at which true variants are identified by each model (TPR). This is further quantified by assessing the area under the curve (AUC). Note that SCANONE produces p-values, lasso and the elastic net give magnitude of regression coefficients and the Bayesian variable selection model computes posterior inclusion probabilities (PIPs). Method performance varies depending on the two factors: (a) the presence of interaction effects, and (b) additional structure due to population stratification. For example, in the first simulation scenario all methods exhibit lower power when a proportion of the broad-sense heritability is made up of interaction effects (e.g., Figure 3(a)). This power increases when additive effects dominate the heritability (e.g., Figure 3(b)). Overall, the lasso is the worst performing method. In the cases where there are no additional population stratification effects, the SCANONE approach proved to be better method. These results are unsurprising since this scenario best suites the assumptions of this approach.

While the performance of our distributional centrality measures are comparable in the first setting, its true advantage becomes apparent when there is some underlying population structure between genotypes (i.e., scenarios II and III). Importantly, under this type of data the power of RATE is consistently better than its counterparts (e.g., Figures 3(c), (d) and S8). From a significance threshold perspective RATE also proves to have the best "optimal"

selection metric. Solely considering SNPs with RATEs > $1/p$ consistently yielded more associative mapping power than observing both (a) the equivalence of the Bayesian "median probability model" (i.e., PIPs > 0.5) [Barbieri and Berger (2004)], and (b) SCANONE p-values below the Bonferroni-corrected significance threshold (i.e., $P < 8.7 \times 10^{-6}$) (see Figure S9). For example, in simulation scenario II the "optimal" RATE model identified 72% and 78% of the casual variables for $\rho = 0.5$ and 1 respectively. This compared to 24% and 37% for the median probability model, and 32% and 46% for the multiple testing corrected SCANONE model (see Figure S9). This trend is consistent across all of the simulation settings that we consider.

Altogether, we want to stress that these simulation results are important from a model interpretation perspective. Even though methods like SCANONE effectively prioritize SNPs in certain scenarios, their significance metrics struggle to create separation between selected and nonselected markers. Therefore, if a practitioner were to choose variants satisfying some "optimal" genome-wide threshold, the more conservative methods will simply miss the majority of the true causal variables (i.e., a higher count of false negatives). RATE, on the other hand, is consistently able to distinguish among the SNPs in a given set. Even in the scenarios where phenotypes are simulated without population stratification effects, RATE is more likely to deem associated variants as significant genome-wide—just at the possible cost of slightly more false positives.

4.2. Real data analysis: Arabidopsis QTL study.

We now apply our approach to a quantitative trait loci (QTL) association mapping study focused on the characterization of complex phenotypes in *Arabidopsis thaliana,* a small flowering plant native to Eurasia. The specific dataset that we consider comes from the Versailles Arabidopsis Stock Center [Loudet et al. (2002)] (http://publiclines.versailles.inra.fr/page/33) and has been previously used for evaluating the mapping power of other statistical methods [Demetrashvili, den Heuvel and Wit (2013)]. More descriptively, it consists of $n = 403$ F6 plants from a Bay-0 × Shahdara recombinant inbred lines (RILs) population that were genotyped for $p = 1028$ genetic markers and phenotyped for 63 different metabolic traits [Wentzell et al. (2007)]. After pruning the genotypes of variants with near perfect correlation ($R \geq 0.99$), we obtained a final set of 524 markers (see the Supplementary Material for details). 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) (see Table S2). Importantly, the goal of the original study was to highlight complex connections between gene expression and metabolite (glucosinolate) variation [Wentzell et al. (2007)]. Here, we consider this particular case study not only because it presents a variety of quantitative traits, but also because the data contains a mixture of additive and some epistatic effects. Indeed, this dataset presents a realistic mix between the cases we previously examined for simulation scenario I.

For each metabolic trait we provide a summary table which lists centrality measures for all gene expression polymorphisms as detected via GP regression and RATE (see Table S3). To contrast the associations identified by our nonparametric method, we also directly compared

results from implementing the SCANONE approach, since it proved to be the most powered of the competing methods in simulations (again, see Table S3). Figures 4 and S10–S14 display plots of enrichment for a genome-wide scan on the six traits according to the RATE enrichment metric. These figures also show the comparative results for the standard single-variant testing approach. Referenced in all images are blue points which represent genetic markers with significant distributional centrality measures above the line of relative equivalence (i.e., RATEs > $1/p$). In Table 1 we report the number of significant markers that are identified by both methods. Once again, these are determined by markers with $\text{RATE}(\widetilde{\beta}) > 1 \,/\, p$ and $P < 9 \times 10^{-5}$ respectively. Again, the latter represents the genome-wide Bonferroni-corrected significance threshold. In the second part of Table 1, we take the significant markers identified by each model and refit simple linear regressions with them. Here, we report R^2 as a way to assess which method was able to select markers that explain the greatest proportion of variance in all six traits.

Overall, RATE consistently identified genomic locations that correspond to known members of biosynthetic pathways in *Arabidopsis thaliana*. Most of these, as in the original study, were small networks of QTLs known to control biosynthetic pathways. For example, in MO4I3M, the most central loci appeared on the second chromosome and were headlined by the marker tagged At2g14170 (see Figure 4(a)). This variant is associated with *ALDH6B2*—a gene within the *Arabidopsis* genome known to catalyze enzymatic reactions in valine and pyrimidine catabolism (i.e., destructive metabolism) [Hou and Bartels (2015), Kirch et al. (2004)]. Similarly, on the first chromosome RATE featured a small group of central loci lead by At1g78370—which encodes a core glucosinolate biosynthesis gene *GSTU20* and plays a key role in glutathione transferase activity and metabolism [Wu et al. (2016)]. For the trait MT8 content RATE deemed the most important region of the genome to be on the fifth chromosome (see Figure S13). Here, the marker At5g22630 had the greatest relative centrality measure. This polymorphism represents *ADT5* which has recently been suggested to moonlight proteins that play an enzymatic role in biosynthesis [Bross et al. (2017)]. This same marker is also highlighted as being moderately influential in explaining the variability inallyl content across the plants (see Figure S10). This makes sense because of the strong positive correlation between the content of these two traits.

These validated findings from previous experimentally based studies lead us to believe that our results contain true positives. Lastly, in order to bolster confidence in the relative centrality measures identified by our nonparametric approach, we also display the correlation structure across the genotypes and phenotypes for the 403 Bay-0 × Shahdara RILs (see Figures S15 and S16). Consistent with our results, there appeared to be strong *cis*-type covariances between groups of genetic markers located on the same chromosome. This underlying genetic architecture resembles data analytic situations where our approach is most powered.

In order to better explain why our nonparametric approach and the SCANONE method performed similarly in each of the six phenotypes, we use a variance component analysis to evaluate how different types of genetic effects (i.e., linear vs. nonlinear) contribute to the overall broad-sense heritability [Zhou (2017)] (see text in the Supplementary Material for details). Briefly, we use a linear mixed model with multiple random effects to partition the

phenotypic variance into three different categories: (a) an additive component, (b) a pairwise interaction component and (c) a third order interaction component. Disregarding the contribution of random noise, we quantify the contribution of these genetic effects by estimating the proportion of heritability that is explained via their corresponding variance components. Table S4 displays these results which effectively highlights that each of the six traits are primarily dominated by additivity.

4.3. Real data analysis: Heterogenous stock of mice GWAS.

We lastly assess RATE's association mapping ability in a more traditional GWAS setting by analyzing three quantitative traits in a heterogeneous stock of mice dataset [Valdar et al. (2006)] from the Wellcome Trust Centre for Human Genetics (http://mtweb.cs.ucl.ac.uk/mus/www/mouse/index.shtml). This data contains $n \approx 2000$ individuals and $p \approx 10{,}000$ SNPs with minor allele frequencies above 5%—with exact numbers varying slightly depending on the phenotype (see the Supplementary Material for details). The three quantitative traits we consider include body weight, percentage of CD8+ cells and high-density lipoprotein (HDL) content. We consider this particular dataset not only because it contains a wide variety of quantitative traits but also because the data contains related samples. Relatedness has been shown to manifest different orders of interaction effects [Hemani, Knott and Haley (2013), Crawford et al. (2018,2017)], and thus this dataset also presents a realistic mix between the cases we examined in simulation scenarios II and III.

Once again, we compare the GP regression model to the single-SNP approach via SCANONE which serves as a baseline. For each trait we provide a summary table which lists the corresponding RATEs and p-values for all SNPs (see Table S5). Figures 5, S17 and S18 then visually display this information via Manhattan plots. In these figures chromosomes are shown in alternating colors for clarity with the top five most enriched regions (according to RATE measures) being highlighted as a way to facilitate comparisons between the mapping approaches.

As in the previous real data application our nonparametric approach was able to detect multiple loci that have been previously validated as having functional associations with the traits of interests. Many of these findings were also indicated in the original study that produced this dataset [Valdar et al. (2006)]. For example, the X chromosome is well known to majorly influence adiposity and metabolism in mice [Rance, Hill and Keightley (1997), Chen et al. (2012, 2013), Cox, Bonthuis and Rissman (2014)]. As expected, in the body weight and HDL content traits, our approach identified significant enrichment in this genomic region—headlined by the chromatin remodeling complex gene *Smarca1* in both cases. Additionally, for the body weight phenotype, RATE also prioritized markers on chromosomes 7 and 10 as having notable associations. Previous computational studies have shown variants on both of these chromosomes to have additive effects and statistical epistatic interactions that influence mice body composition [Ankra-Badu et al. (2009), Brockmann et al. (1998), Diament and Warden (2003), Kleyn et al. (1996)]. In this particular analysis we attribute the selection of these loci to the nonlinear properties of the Gaussian covariance function and the nonparametric nature of the GP regression model. Similarly, for HDL content RATE found many significant SNPs on the first, eleventh and twelfth

chromosomes. The corresponding spike on chromosome 1 is a genomic location that most notably harbors the HDL driver gene *Ath-1* [Paigen et al. (1987)] (see Figure 5(a)). Finally, for the phenotype detailing the percentage of CD8+ cells, our method identified the majority of significant SNPs to be on the seventeenth chromosome—including those within boundary of *Myof1*, a gene that has been suggested to modulate cell adhesion and motility in the immune system [Kim et al. (2006)]. Overall, this general genomic location that has been validated to greatly determine the ratio of T-cells [Yalcin et al. (2010)].

Once again, we use variance component analysis to now dissect the broad-sense heritability of these three mice traits and help better explain why there could be differences in the loci discovered by RATE and SCANONE (see Table S6). As in the previous subsection we implement a linear mixed model to partition the overall broad-sense heritability into the same additive, second order (pairwise) interaction and third order interaction genetic effect types. Note that, unlike in the *Arabidopsis* QTL study, additive effects do not dominate the genotypic contribution in any of the three mice phenotypes that we consider—this is particularly obvious for the trait detailing the HDL content (Figure 5 and Table S6). Instead, the variance components corresponding to the second and third order interactions make up the majority of the broad-sense heritability. We believe that accounting for these nonlinear relationships, as well as controlling for the relatedness between samples, allows RATE to identify loci that SCANONE misses.

5. Discussion.

In this paper, we proposed a new general measure for conducting variable selection in “black box” Bayesian methodologies. While many of these black box approaches often give notable predictive performance gains, the reasoning behind these results can be difficult to explain and interpret. Within a statistical genetics context we discussed how the previously proposed effect size analog for nonparametric regression enables the prioritization of variants based on their marginal associations. Recognizing that one of the main sources of performance gains in black box modeling is through underlying interactions and nonlinear effects between predictor variables, we introduced our new distributional centrality measure RATE —meant to rank genetic markers based on their influence on the joint distribution with other markers. As we demonstrated with simulation studies, our new measure can be used for feature selection, giving state-of-the-art performance even in the presence of population structure. In real QTL and GWAS data applications, RATE allowed us to uncover biologically relevant markers by simultaneously taking into account significant interactions when ranking variants based on their relative importance.

In its current form we have focused on demonstrating RATE with a Gaussian process regression model. Although our entire illustration of the method is based on the manipulation of approximate posterior distributions in Bayesian applications, each of the innovations that we present can be applied in a frequentist setting. The effect size analog is merely a summary statistic which can be derived after fitting any model. Therefore, one could envision a frequentist setting in which parameter estimation and uncertainty is done using bootstrap, for example. In particular this would lead to a multivariate normal-like estimator for the mean and covariance of the effect size analog. One could then proceed to

compute the relative centrality measures with this distribution. The utility of our approach, from this alternative point of view, remains an open question.

RATE is not without its limitations. One particular limitation of RATE is that while it provides a measure of general association for nonparametric methods, it cannot be used to directly identify the component (i.e., linear vs. nonlinear) that drives individual variable associations. Thus, despite being able to detect significant variants that are associated to a response in a nonlinear fashion, the RATE measure is unable to directly identify the detailed orders of interaction effects. A key part of our future work is learning how to disentangle this information. A second, and perhaps the most noticeable, limitation of RATE is that the computation of the centrality measures scales at least cubically with the number of features in the input data (see Table S7 in the Supplementary Material). This is opposed to the other methods we compare in this study (e.g., single-SNP tests) which take a fraction of the time to compute. In future work we would like to consider the challenges of analyzing large scale studies. An example of this would be consortium-sized efforts in human-based genome-wide association studies with millions of markers and thousands of genotyped individuals [Sudlow et al. (2015), The 1000 Genomes Project Consortium (2010), The Wellcome Trust Case Control Consortium (2007)]. In these settings one possible immediate fix would be to use a two step procedure. In the first step we implement a more scalable mapping method [Lippert et al. (2011), Purcell et al. (2007), Zhou and Stephens (2012)] as a screen to select the top marginally associated markers. Then, in the second step we test for more detailed nonlinear prioritization using centrality measures. Nonetheless, new algorithms and alternative code implementations are likely needed to scale RATE up to datasets that are orders of magnitude larger in size.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

We would like to the Editor, Associate Editor and two anonymous referees for their constructive comments. We would also like to thank Andrew Gelman, Elizabeth R. Hauser, Steve Oudot, Sohini Ramachandran and Xiang Zhou for helpful conversations and suggestions. This study makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113, 085475, and 090355.

References

Alaa AM and van der Schaar M (2017). Bayesian nonparametric causal inference: Information rates and learning algorithms. Available at ArXiv:1712.08914.

Ankra-Badu GA, Pomp D, Shriner D, Allison DB and Yi N (2009). Genetic influences on growth and body composition in mice: Multilocus interactions. Int. J. Obes 33 89–95. DOI:10.1038/ijo.2008.215.

Barbieri MM and Berger JO (2004). Optimal predictive model selection. Ann. Statist 32 870–897. MR2065192

Brockmann GA, Haley CS, Renne U, Knott SA and Schwerin M (1998). Quantitative trait loci affecting body weight and fatness from a mouse line selected for extreme high growth. Genetics 150 369–381. [PubMed: 9725853]

Bross CD, Howes TR, Abolhassani Rad S, Kljakic O and Kohalmi SE (2017). Subcellular localization of *Arabidopsis* arogenate dehydratases suggests novel and non-enzymatic roles. J. Exp. Bot 68 1425–1440. [PubMed: 28338876]

Carvalho CM, Polson NG and Scott JG (2010). The horseshoe estimator for sparse signals. Biometrika 97 465–480. MR2650751

Carvalho CM and West M (2007). Dynamic matrix-variate graphical models. Bayesian Anal. 2 69–97. MR2289924

Chaudhuri A, Kakde D, Sadek C, Gonzalez L and Kong S (2017). The mean and median criterion for automatic kernel bandwidth selection for support vector data description. Available at arXiv:1708.05106.

Chen X, McClusky R, Chen J, Beaven SW, Tontonoz P, Arnold AP and Reue K (2012). The number of X chromosomes causes sex differences in adiposity in mice. PLoS Genet. 8 e1002709. [PubMed: 22589744]

Chen X, McClusky R, Itoh Y, Reue K and Arnold AP (2013). X and Y chromosome complement influence adiposity and metabolism in mice. Endocrinology 154 1092–1104. DOI:10.1210/en.2012-2098. [PubMed: 23397033]

Chipman HA, George EI and McCulloch RE (2010). BART: Bayesian additive regression trees. Ann. Appl. Stat 4 266–298. MR2758172

Cotter A, Keshet J and Srebro N (2011). Explicit approximations of the Gaussian kernel. Available at arXiv:1109.4603.

Cox KH, Bonthuis PJ and Rissman EF (2014). Mouse model systems to study sex chromosome genes and behavior: Relevance to humans. Front. Neuroendocrinol 35 405–419. DOI: 10.1016/j.yfrne.2013.12.004. [PubMed: 24388960]

Crawford L and Zhou X (2018). Genome-wide marginal epistatic association mapping in case-control studies. *BioRxiv* 374983.

Crawford L, Zeng P, Mukherjee S and Zhou X (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 13 e1006869. [PubMed: 28746338]

Crawford L, Wood KC, Zhou X and Mukherjee S (2018). Bayesian approximate kernel regression with variable selection. J. Amer. Statist. Assoc 113 1710–1721. MR3902240

Crawford L, Flaxman SR, Runcie DE and West M (2019). Supplement to "Variable prioritization in nonlinear black box methods: A genetic association case study". DOI:10.1214/18-AOAS1222SUPPA, DOI:10.1214/18-AOAS1222SUPPB, DOI:10.1214/18-AOAS1222SUPPC, DOI:10.1214/18-AOAS1222SUPPD.

Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Pérez-Rodríguez P and de Los Campos G (2017). Bayesian genomic prediction with genotype × environment interaction kernel models. G3 (Bethesda) 7 41–53. [PubMed: 27793970]

Demetrashvili N, den Heuvel ERV and Wit EC (2013). Probability genotype imputation method and integrated weighted lasso for QTL identification. BMC Genet. 14 125. [PubMed: 24378210]

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K and Cotes J (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182 375–385. [PubMed: 19293140]

de los Campos G, Gianola D, Rosa GJM, Weigel KA and Crossa J (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res 92 295–308.

Diament AL and Warden CH (2003). Multiple linked mouse chromosome 7 loci influence body fat mass. Int. J. Obes 28 199 EP.

Drineas P and Mahoney MW (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. J. Mach. Learn. Res 6 2153–2175. MR2249884

Fasshauer G and McCourt M (2016). Kernel-Based Approximation Methods Using MATLAB. World Scientific, Hackensack, NJ.

Gelman A, Hwang J and Vehtari A (2014). Understanding predictive information criteria for Bayesian models. Stat. Comput 24 997–1016. MR3253850

Goutis C and Robert CP (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. Biometrika 85 29–37. MR1627250

Gruber L and West M (2016). GPU-accelerated Bayesian learning and forecasting in simultaneous graphical dynamic linear models. Bayesian Anal. 11 125–149. MR3447094

Gruber LF and West M (2017). Bayesian online variable selection and scalable multivariate volatility forecasting in simultaneous graphical dynamic linear models. Econ. Stat 3 3–22. MR3666239

Guan Y and Stephens M (2011). Bayesian variable selection regression for Genome-wide association studies and other large-scale problems. Ann. Appl. Stat 5 1780–1815. MR2884922

Hemani G, Knott S and Haley C (2013). An evolutionary perspective on epistasis and the missing heritability. PLoS Genet. 9 e1003295. [PubMed: 23509438]

Hemani G, Shakhbazov K, Westra H-J, Esko T, Henders AK, McRae AF, Yang, Gibson, Martin NG, Metspalu A, Franke L, Montgomery GW, Visscher PM and Powell JE (2014). Detection and replication of epistasis influencing transcription in humans. Nature 508 249–253. [PubMed: 24572353]

Hill WG, Goddard ME and Visscher PM (2008). Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4 e1000008. [PubMed: 18454194]

Horn T, Sandmann T, Fischer B, Axelsson E, Huber W and Boutros M (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nat. Methods 8 341–346. [PubMed: 21378980]

Hou Q and Bartels D (2015). Comparative study of the aldehyde dehydrogenase (ALDH) gene superfamily in the glycophyte *Arabidopsis thaliana* and *Eutrema* halophytes. Ann. Bot 115 465–479. [PubMed: 25085467]

Howard R, Carriquiry AL and Beavis WD (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda) 4 1027–1046. [PubMed: 24727289]

Jiang Y and Reif JC (2015). Modeling epistasis in genomic selection. Genetics 201 759–768. [PubMed: 26219298]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C and Eskin E (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet 42 348–354. [PubMed: 20208533]

Kim SV, Mehal WZ, Dong X, Heinrich V, Pypaert M, Mellman I, Dembo M, Mooseker MS, Wu D and Flavell RA (2006). Modulation of cell adhesion and motility in the immune system by Myo1f. Science 314 136–139. [PubMed: 17023661]

Kirch H-H, Bartels D, Wei Y, Schnable PS and Wood AJ (2004). The ALDH gene superfamily of Arabidopsis. Trends Plant Sci. 9 371–377. [PubMed: 15358267]

Kleyn PW, Fan W, Kovats SG, Lee JJ, Pulido JC, Wu Y, Berkemeier LR, Misumi DJ, Holmgren L et al. (1996). Identification and characterization of the mouse obesity gene tubby: A member of a novel gene family. Cell 85 281–290. [PubMed: 8612280]

Kolmogorov AN and Rozanov Ju. A. (1960). On a strong mixing condition for stationary Gaussian processes. Theory Probab. Appl 5 222–227. MR0133175

Liang F, Paulo R, Molina G, Clyde MA and Berger JO (2008). Mixtures of g priors for Bayesian variable selection. J. Amer. Statist. Assoc 103 410–423. MR2420243

Lim C and Yu B (2016). Estimation stability with cross-validation (ESCV). J. Comput. Graph. Statist 25 464–492. MR3499690

Lin L, Chan C and West M (2016). Discriminative variable subsets in Bayesian classification with mixture models, with application in flow cytometry studies. Biostatistics 17 40–53. MR3449849 [PubMed: 26040910]

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI and Heckerman D (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods 8 833–835. [PubMed: 21892150]

Loudet O, Chaillou S, Camilleri C, Bouchez D and Daniel-Vedele F (2002). Bay-0 × Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in *Arabidopsis*. Theor. Appl. Genet 104 1173–1184. [PubMed: 12582628]

Mackay TFC (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. Nat. Rev. Genet 15 22–33. [PubMed: 24296533]

Mathai AM and Provost SB (1992). Quadratic Forms in Random Variables. Theory and Applications. Statistics: Textbooks and Monographs 126 Dekker, New York MR1192786

Mercer J (1909). Functions of positive and negative type and their connection with the theory of integral equations. Philos. Trans. R. Soc. Lond. Ser. A 209 415–446.

Paigen B, Mitchell D, Reue K, Morrow A, Lusis AJ and LeBoeuf RC (1987). Ath-1, a gene determining atherosclerosis susceptibility and high density lipoprotein levels in mice. Proc. Natl. Acad. Sci. USA 84 3763–3767. [PubMed: 3473481]

Phillips PC (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet 9 855–867. DOI:10.1038/nrg2452. [PubMed: 18852697]

Piironen J and Vehtari A (2016). Projection predictive model selection for Gaussian processes. in IEEE International Workshop on Machine Learning for Signal Processing 1–6. IEEE, New York.

Piironen J and Vehtari A (2017). Comparison of Bayesian predictive methods for model selection. Stat. Comput 27 711–735. MR3613594

Pillai NS, Wu Q, Liang F, Mukherjee S and Wolpert RL (2007). Characterizing the function space for Bayesian kernel models. J. Mach. Learn. Res 8 1769–1797. MR2332448

Prabhu S and Pe'er I (2012). Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Res. 22 2230–2240. [PubMed: 22767386]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81 559–575. DOI:10.1086/519795. [PubMed: 17701901]

Rahimi A and Recht B (2007). Random features for large-scale kernel machines. Adv. Neural Inf. Process. Syst 3 5.

Rance KA, Hill WG and Keightley PD (1997). Mapping quantitative trait loci for body weight on the X chromosome in mice. I. Analysis of a reciprocal F2 population. Genet. Res 70 117–124. [PubMed: 9449188]

Rasmussen CE and Williams CKI (2006). Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA MR2514435

Richard MD and Lippmann RP (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Comput. 3 461–483. [PubMed: 31167331]

Schölkopf B, Herbrich R and Smola AJ (2001). A generalized representer theorem. In Computational Learning Theory (Amsterdam, 2001) Lecture Notes in Computer Science 2111 416–426. Springer, Berlin MR2042050

Shi JQ, Wang B, Will EJ and West RM (2012). Mixed-effects Gaussian process functional regression models with application to dose-response curve prediction. Stat. Med 31 3165–3177. MR2993619 [PubMed: 22865484]

Smith A, Naik PA and Tsai C-L (2006). Markov-switching model selection using Kullback-Leibler divergence. J. Econometrics 134 553–577. MR2328419

Stephens M and Balding DJ (2009). Bayesian statistical methods for genetic association studies. Nat. Rev. Genet 10 681–690. [PubMed: 19763151]

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 e1001779. [PubMed: 25826379]

Tan S, Caruana R, Hooker G and Lou Y (2017). Detecting bias in black-box models using transparent model distillation. Available at arXiv:1710.06169.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature 467 1061–1073. [PubMed: 20981092]

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 661–678. [PubMed: 17554300]

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JNP, Mott R and Flint J (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet 38 879–887. [PubMed: 16832355]

Wahba G (1990). Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics 59 SIAM, Philadelphia, PA MR1045442

Waldmann P, Mészáros G, Gredler B, Fürst C and Sölkner J (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Front. Genet 4 270. [PubMed: 24363662]

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL and Yu W (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am. J. Hum. Genet 87 325–340. [PubMed: 20817139]

Wang X, Elston RC and Zhu X (2011a). Statistical interaction in human genetics: How should we model it if we are looking for biological interaction? Nat. Rev. Genet 12 74. [PubMed: 21102529]

Wang X, Elston RC and Zhu X (2011b). The meaning of interaction. Hum. Hered 70 269–277.

Weissbrod O, Geiger D and Rosset S (2016). Multikernel linear mixed models for complex phenotype prediction. Genome Res. 26 969–979. [PubMed: 27302636]

Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA and Klieben-Stein DJ (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 3 e162.

Woo JH, Shimoni Y,Yang WS, Subramaniam P, Iyer A, Nicoletti P, Rodríguez Martínez M, López G, Mattioli M et al. (2015). Elucidating compound mechanism of action by network perturbation analysis. Cell 162 441–451. [PubMed: 26186195]

Wood AR, Tuke MA, Nalls MA, Hernandez DG, Bandinelli S, Singleton AB, Melzer D, Ferrucci L, Frayling TM and Weedon MN (2014). Another explanation for apparent epistasis. Nature 514 E3–E5. [PubMed: 25279928]

Wu MC, Lee S, Cai T, Li Y, Boehnke M and Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet 89 82–93. [PubMed: 21737059]

Wu J, Zhao Q,Yang Q,Liu H,Li Q,Yi X, Cheng Y,Guo L,Fan C and Zhou Y (2016). Comparative transcriptomic analysis uncovers the complex genetic network for resistance to Sclerotinia sclerotiorum in Brassica napus. Sci. Rep 6 19007 EP. [PubMed: 26743436]

Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, Østerås M, Whitley A, Yuan W. et al. (2010). Commercially available outbred mice for genome-wide association studies. PLoS Genet. 6 e1001085. [PubMed: 20838427]

Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, von smith R and Yi N (2007). R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. Bioinformatics 23 641–643. DOI:10.1093/bioinformatics/btm011. [PubMed: 17237038]

Yang J, Zaitlen NA, Goddard ME, Visscher PM and Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet 46 100–106. [PubMed: 24473328]

Zeng P and Zhou X (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun. 8 456. [PubMed: 28878256]

Zhang Z, Dai G and Jordan MI (2011). Bayesian generalized kernel mixed models. J. Mach. Learn. Res 12 111–139. MR2773550

Zhang Y and Liu JS (2007). Bayesian inference of epistatic interactions in case-control studies. Nat. Genet 39 1167–1173. [PubMed: 17721534]

Zhang X, Huang S, Zou F and Wang W (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics 26 i217–i227. DOI:10.1093/bioinformatics/btq186. [PubMed: 20529910]

Zhou X (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat 11 2027–2051. MR3743287 [PubMed: 29515717]

Zhou X and Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet 44 821–825. [PubMed: 22706312]

Zhou X and Stephens M (2014). Efficient multivariate linear mixed model algorithms for genomewide association studies. Nat. Methods 11 407–409. [PubMed: 24531419]

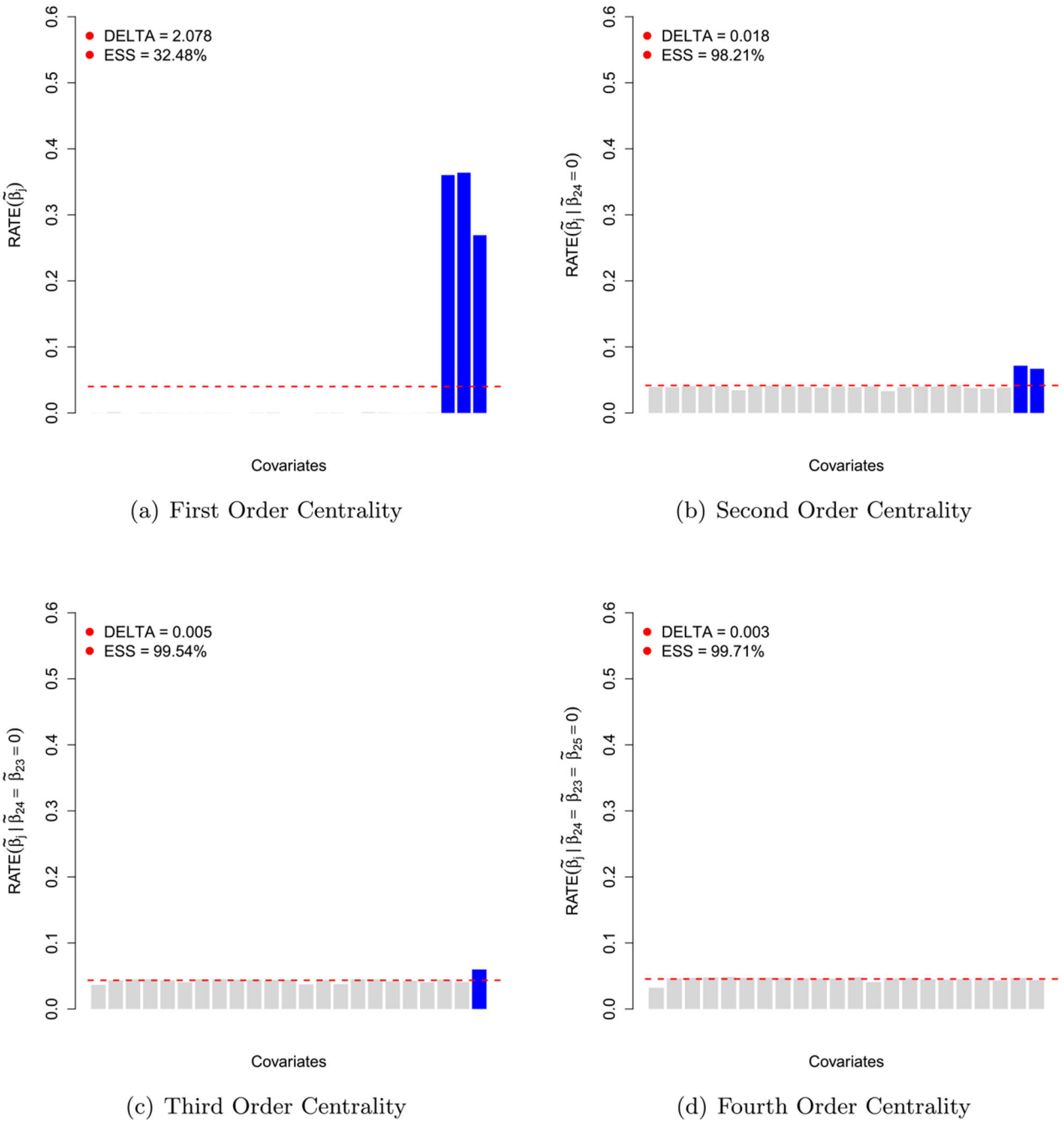

Fig. 1.

Orders of distributional centrality via RATE measures. These are simple proof of concept simulations with broad-sense heritability level $H^2 = 0.6$ *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. 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*.

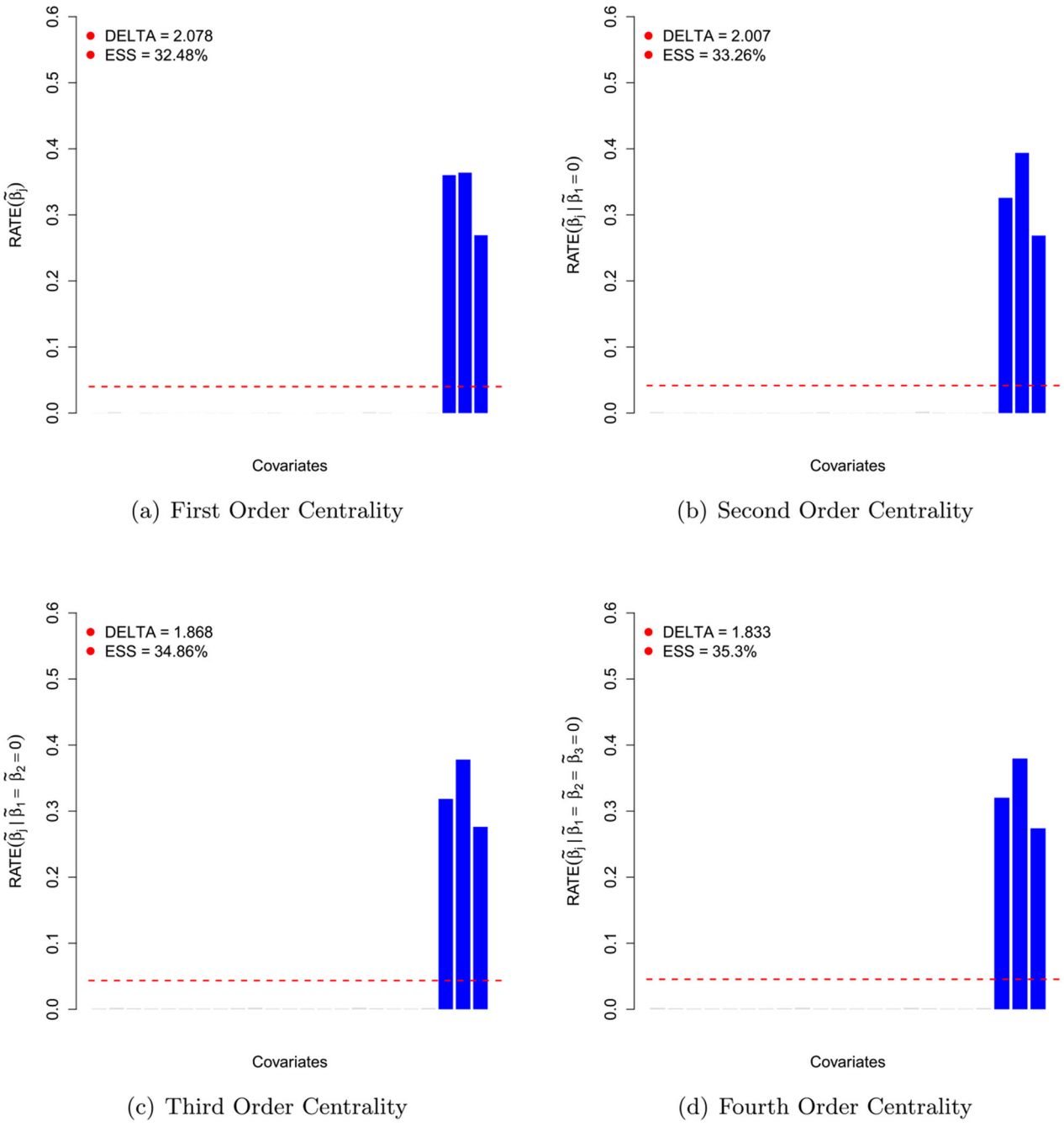

Fig. 2.

Orders of distributional centrality via RATE measures when nonassociated variants are deemed significant. These are simple proof of concept simulations with broad-sense heritability level $H^2 = 0.6$ *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. 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*.

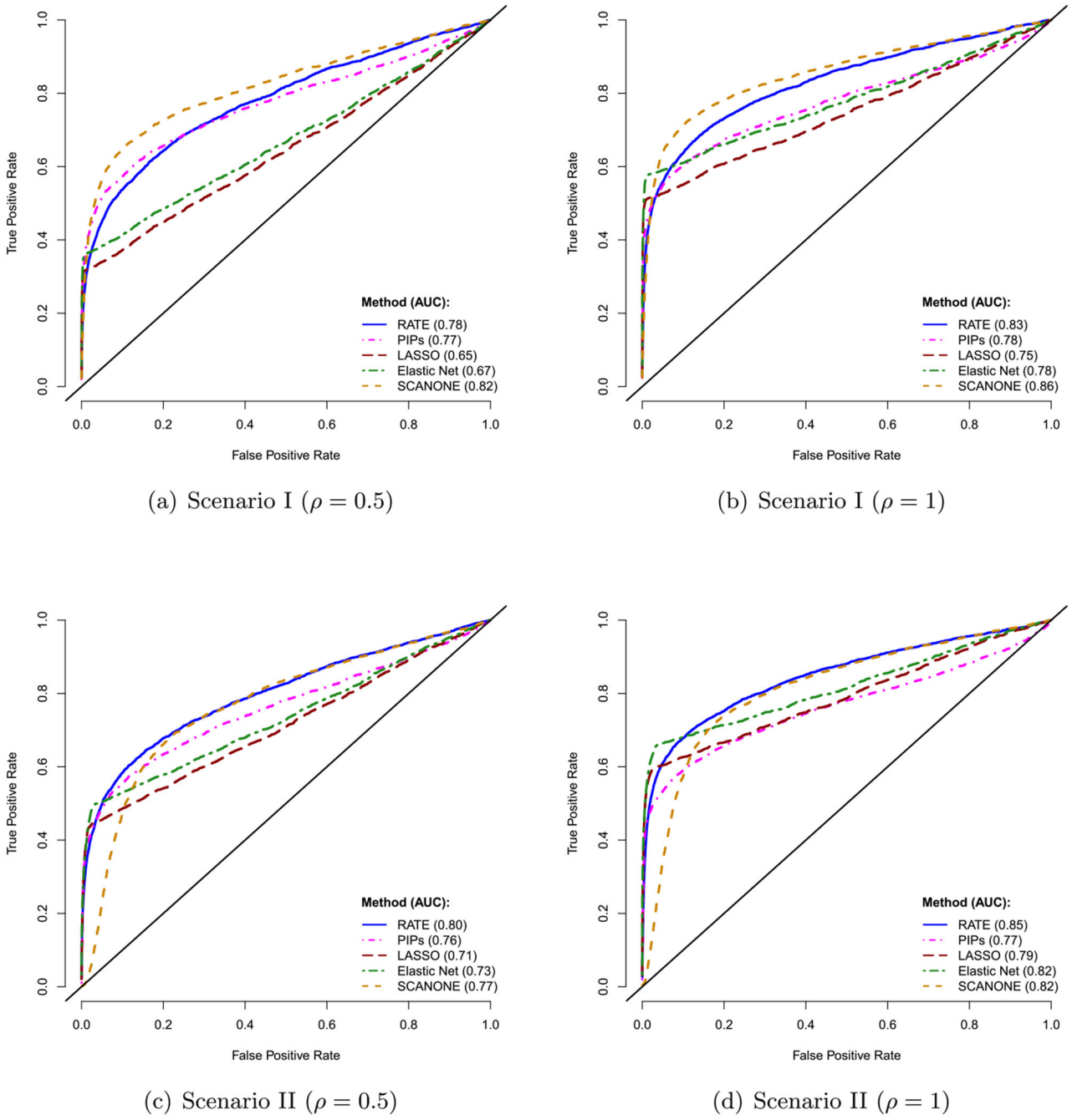

Fig. 3.

Power analysis for prioritizing genetic variants. Phenotypes are simulated with broad-sense heritability level $H^2 = 0.3$ *with control parameter* $\rho = \{0.5, 1\}$ *in Figures* (a) *and* (c) *and Figures* (b) *and* (d) *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. Scenario I corresponds to*

phenotypic outcomes being generated via simulation model (i). *Scenario II introduces population stratification effects with simulation model* (ii) *by allowing the top five genotype PCs to make up* 30% *of the phenotypic variance. Results are based on* 100 *replicates in each case.*

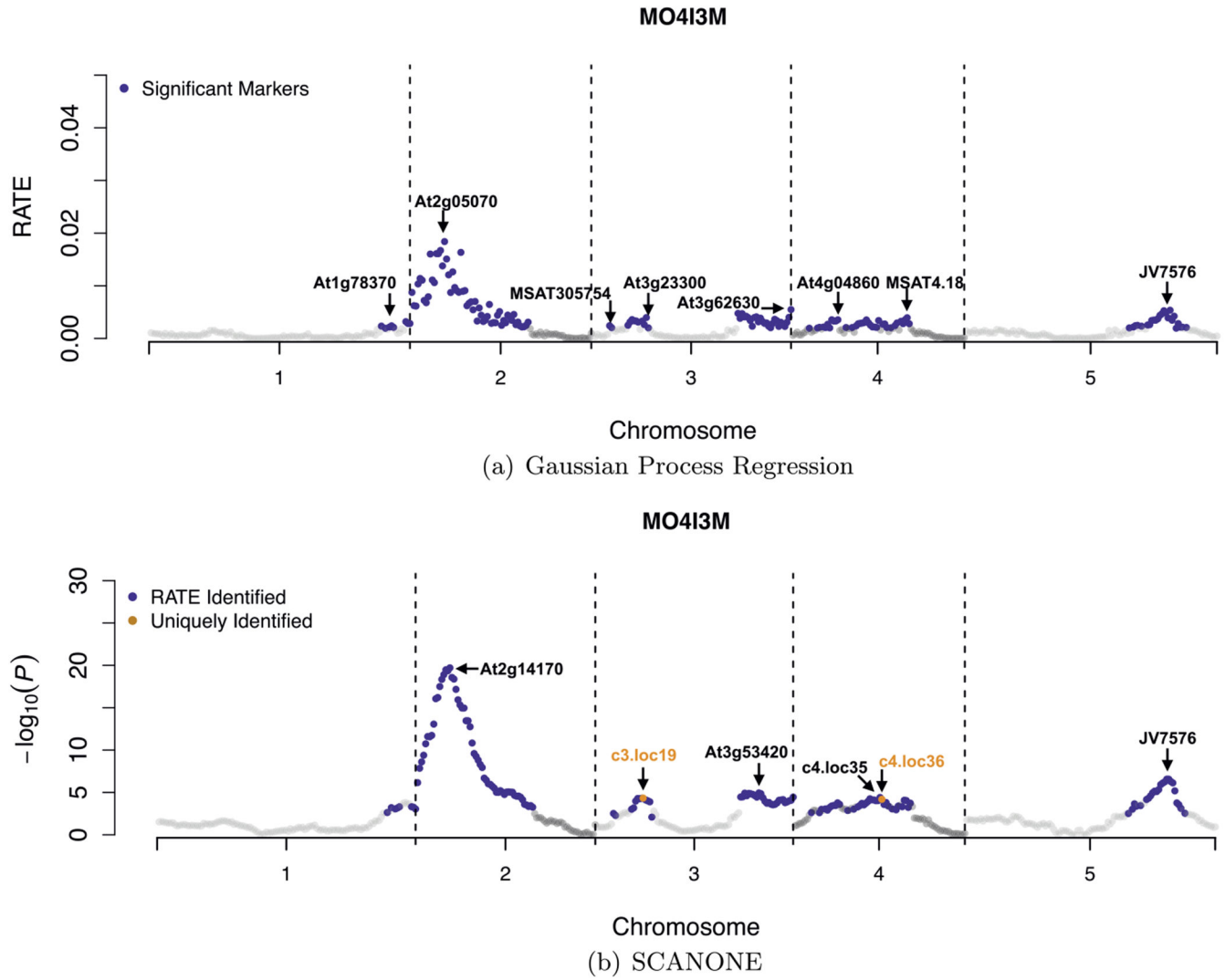

Fig. 4.

Genetic map wide scan for the 4-methoxy-indol-3-ylmethyl (*MO4I3M*) *glucosinolate 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* $\text{RATE}(\tilde{\beta}) > 1 \,/\, p$ 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.*

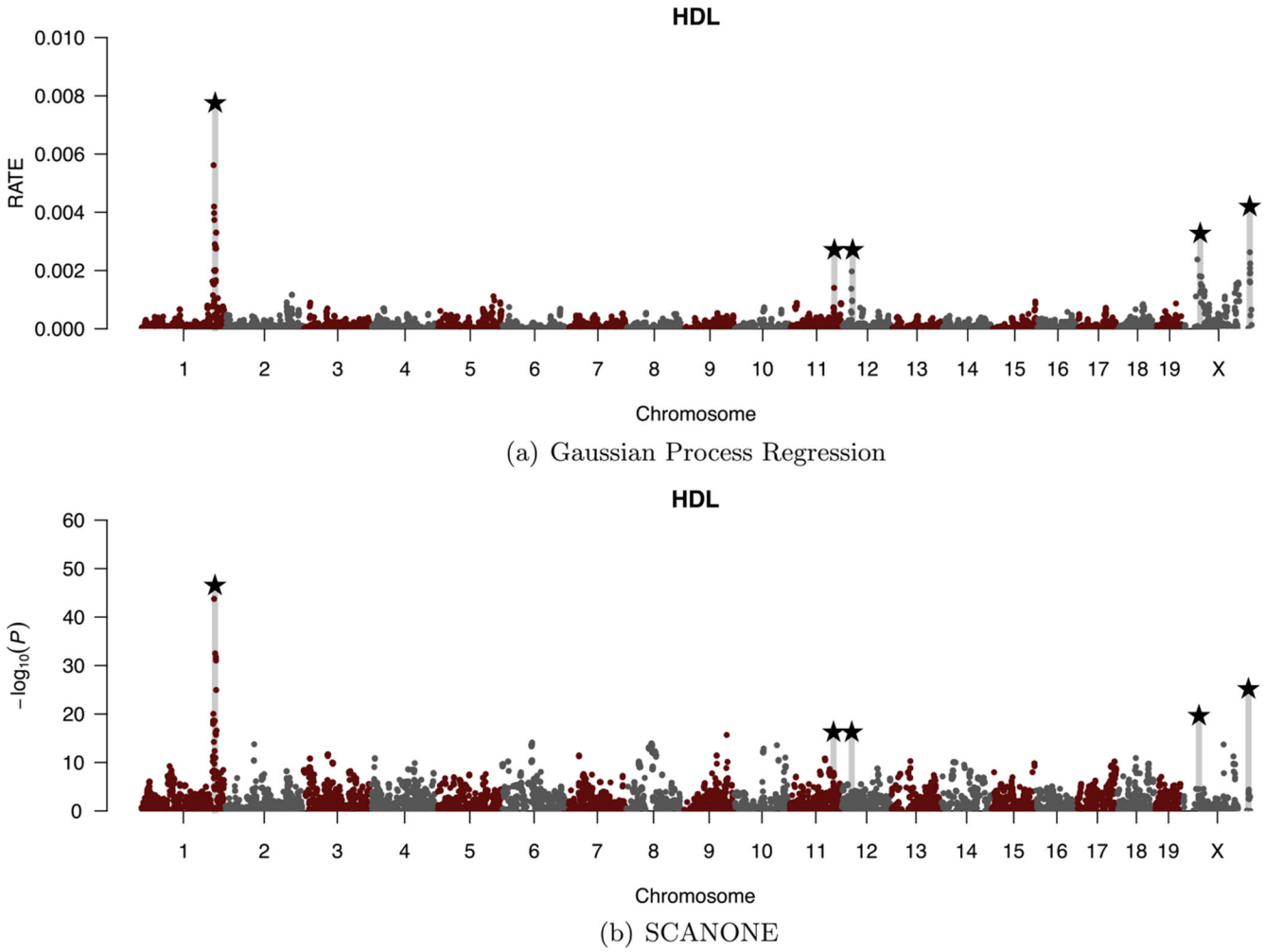

Fig. 5.

Genome-wide scan for high-density lipoprotein (*HDL*) *content in the heterogeneous 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 five most enriched regions* (*according to RATE*) *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.*

Table 1

Comparing RATE and the SCANONE mapping approach in the Arabidopsis QTL study. Glucosinolate content traits include allyl content, indol-3-ylmethyl (I3M), 4-methoxy-indol-3-ylmethyl (MO4I3M), 4-methylsulfinylbutyl (MSO4), 8-methylthiooctyl (MT8) and 3-hydroxypropyl (OHP3). Significant markers are determined by $\mathrm{RATE}(\tilde{\beta}) > 1 / p$ *and* $P < 9 \times 10^{-5}$ *respectively. The latter represents the genome-wide Bonferroni-corrected significance threshold. Values in bold denote the best according to* R^2 *when considering "optimal" model fit with the significant markers. The last section describes the percent overlap between the significant markers found using the two methods*

		Phenotypic traits					
Category	**Method**	**Allyl**	**I3M**	**MO4I3M**	**MSO4**	**MT8**	**OHP3**
# Sig. Markers	RATE	64	130	165	117	85	96
	SCANONE	61	75	99	100	71	98
R^2 of Sig. Model	RATE	**0.686**	**0.472**	**0.570**	**0.544**	**0.610**	**0.569**
	SCANONE	0.675	0.353	0.452	0.494	0.527	0.566
% Overlap	SCANONE ⊆ RATE	97%	100%	98%	100%	100%	97%