Simultaneous Selection of Multiple Important Single Nucleotide Polymorphisms in Familial Genome Wide Association Studies data

Subhabrata Majumdar

Abstract: Genome Wide Association Studies (GWAS) on samples obtained from families instead of unrelated individuals are an established way of assessing gene-environment interactions and effects of Single Nucleotide Polymorphisms (SNP) on the development of behavioral disorders. Due to multiple reasons like weak signals of causal SNPs or multiple correlated SNPs, traditional methods of SNP detection based on single-marker analysis suffer from loss of power in such studies. Although statistical model selection methods are able to remedy this, owing to the dependent structure of the familial data they are computationally demanding. Here we propose a fast variable selection algorithm to address this problem. Working on a linear mixed model with three variance components and fixed effects for multiple SNPs, we obtain a quantity called the e-value for each SNP by only training the model with all covariates, and select SNPs having e-values below a threshold. To compute the e-values, we utilize a fast and scalable bootstrap procedure that relies on Monte-Carlo sampling to obtain bootstrapped copies of estimated fixed effect vectors. Numerical studies reveal our method to be more effective in detecting causal SNPs than either single-marker analysis on mixed models or model selection methods that ignore the familial dependency structure. We also use the e-values to perform gene-level analysis in a familial GWAS dataset and detect several SNPs that have potential effect on alcohol consumption in individuals.

Keywords: Model selection, bootstrap, data depth, family data, twin studies, ACE model, alcoholism

1 Introduction

Genome Wide Association Studies (GWAS), where genetic variants across the full human genome are analyzed, are becoming more and more relevant in recent years for the purpose of determining which of the variants are associated behind the expression of complex traits. The advent of efficient and economical genotyping technology enables researchers to scan the genome at hundreds of thousands of Single Nucleotide Polymorphisms (SNPs), and improvements in computational speed in the past few decades have helped in feasible analysis of the huge amount of data collected in order to detect significant associations (Visscher et al., 2012). One major challenge in such studies is the small effects individual SNPs have: detecting which requires large sample sizes (Manolio et al., 2009). For quantitative behavioral traits, for example alcohol dependence, drug abuse, Anorexia and depression, this problem is amplified because of the additional noise introduced by variation due to the environment the subject grew up in. This is one of the motivations of performing GWAS on families (GWAF) instead of unrelated individuals, through which the environmental variation can be reduced: so as to require smaller samples to detect the same magnitude of SNP effect. Another major reason, of course, of performing GWAS on familial data is to detect gene-environment interactions associated with development of behavioral traits. The data analyzed in such families typically consist of trait information and genotypes of from parents and their children, who can be either identical twins, non-identical twins or adpoted.

Single-marker tests, i.e. analyzing the effect of multiple SNPs separately on the quantitative phenotype and then selecting a group of SNPs by setting suitable thresholds on the resulting p-values, often after correction for multiple testing, is the most commonly used method to detect SNPs associated with the phenotype being studied. Although simultaneously estimating the fixed effect of a single SNP as well as the stratified population variance covariance matrix reflecting the familial structure, and repeating this for a large number of SNPs is a computationally prohibitive task, several fast approximation methods exist in the literature that tackle this while maintaining moderately high power. The GRAMMAR method of Aulchenko et al. (2007) and the association test of Chen and Abecasis (2007) are examples of this. While these two methods are able to efficiently analyze GWAF data, they assume that phenotypic similarity within families is entirely due to their genetic similarity and ignore the effect of shared environment. In GWAF data from nuclear (i.e. unrelated) families, the proportion of phenotypic variation explained by the shared environmental effects is often substantial, sometimes as

high as 51% (McGue et al., 2013) or 74% (De Neve et al., 2013): in which case such methods shall not be able to account for this added variation. To remedy this, Li et al. (2011) proposed a rapid method (RFGLS) that computes p-values corresponding to each SNP through a rapid approximation of the single-SNP generalized least squares model taking into account genetic and environmental sources of familial similarity.

A major issue with all such methods of single-marker analysis is that they are not always effective for detecting functionally relevant SNPs or regions in the genome. A single SNP is sometimes not enough to capture the extent of association (Ke, 2012; Yang et al., 2012). This includes cases when there are multiple causal SNPs closely located inside a gene in high Linkage Disequilibrium (LD) with one another. The causal SNP may even not be genotyped if its variants are unlikely to be present in the sample population (e.g. the variant of the SNP rs671 responsible for low alcohol tolerance in asians is rare in caucasians), and other SNPs highly correlated with it are genotyped instead.

Here we propose to tackle this through fitting mixed effect models with the behavioral trait phenotype as respone and a group of SNPs (e.g. SNPs inside a single gene) as fixed effect predictors, and selecting important SNPs through a model selection approach. Although the major impediment of applying model selection techniques in GWAS setup is the high computational cost, some fast methods have been proposed that are able to perform SNP seletion from a multi-SNP model on GWAS data from *unrelated individuals* (Frommelet et al., 2012; Zhang et al., 2014). However, these methods still rely on fitting models corresponding to multiple predictor sets. This makes them unsuitable to be adapted to the GWAF setup because of the much higher computational costs associated with training multiple mixed effect models that take into account within-family correlation between individuals, as compared to models that assume independent observations.

We remedy this by adapting the recently proposed framework of e-values (Majumdar and Chatterjee, 2017) to perform fast variable selection. In a regression-like setup, for any estimation method that provides consistent estimates (at a certain rate) of the vector of parameters, e-values quantify the proximity of the sampling distribution for a restricted parameter estimate to that of the full model estimate. A variable selection algorithm using the e-values has the following simple and generic steps:

1. Obtain coefficient estimates for the full model, i.e. where all covariates are being estimated from the data, and use resampling to estimate its sampling distribution;

2. Set an element of the coefficient vector to 0, obtain resampling approximation of this marginal

model. Compute *e*-value of that covariate by comparing this distribution with the full model distribution;

3. Go back to the full model and repeat step 2 for all other covariates;

4. Select covariates that have *e*-values below a pre-determined threshold.

The above one-step algorithm offers several benefits in the SNP selection scenario. It allows for fitting multi-SNP models, thereby accommodating the cases of modelling highly correlated SNPs or multiple causal SNPs situated close to one another. Since only the full model needs to be computed in the algorithm, it also offers the user more flexibility in utilizing a suitable method of estimation. Finally, we use the Generalized Bootstrap (Chatterjee and Bose, 2005) as our chosen resampling technique. Instead of fitting a separate model for each bootstrap sample, it computes bootstrap estimates using Monte-Carlo samples from the resampling distribution and reusing model objects obtained while fitting the full model. Consequently, the resampling step becomes very fast and parallelizable.

The rest of the paper is organized as follows. Section 2 provides background information on the GWAF dataset we shall be using in our case study, as well as introduce the statistical model we shall use to model this data. We start Section 3 by providing a technical introduction to the *e*-values framework, then elaborate on the necessary modifications for adapting it to our modelling scenario. Here we also present details of the generalized bootstrap procedure. We illustrate the performance of this method on synthetic datasets in 4. In Section 5 we analyze our GWAF dataset using the *e*-values technique to select SNPs from multiple genes that have demonstrated effect on alcohol consumption in individuals. Finally in Section 6 we provide a review of the work and outline future research directions. We include the proofs of all new results stated, specifically, 3.3, 3.4, 3.5, 3.13 and 3.14, in the supplementary material.

2   Data and model

2.1   The MCTFR data

The familial GWAS dataset collected and studied by Minnesota Center for Twin and Family Research (MCTFR)(Li et al., 2011; McGue et al., 2013; Miller et al., 2012) consists of samples from three longitudinal studies conducted by the MCTFR: (1) the Minnesota Twin Family Study (MTFS: Iacono

et al. (1999)) that covers twins and their parent, (2) the Sibling Interaction and Behavior Study (SIBS: McGue et al. (2007)) that includes adopted and biological sibling pairs and their parents, and (3) the enrichment study (ES: Keyes et al. (2009)) that extended the MTFS by oversampling 11 year old twins who are highly likely to develop substance abuse. While 9827 individuals completed the initial assessments for participation in the study, after several steps of screening the final sample consisted of 7188 caucasian individuals clustered in 2300 nuclear families.

DNA samples collected from the subjects were analyzed using Illuminas Human660W-Quad Array for 561,490 non-intensity SNP markers. After several data cleaning steps for quality control, 527,829 SNPs were retained. Covariates for each sample included age, sex, birth year, generation (parent or offspring), as well as two-way interactions between generation and other three covariates each. As for the quantitative phenotypes, five of them were studied in this GWAS: (1) Nicotine dependence, (2) Alcohol consumption, (3) Alcohol dependence, (4) Illegal drug usage, and (5) Behavioral disinhibition. The response variables corresponding to these phenotypes were derived from questionnaires using a hierarchical approach based on factor analysis (Hicks et al., 2011).

A more detailed description of the data is available in Miller et al. (2012). Several studies have been performed that focus on different aspects of this dataset. Li et al. (2011) used RFGLS to single out causal SNPs behind the height of participants, while McGue et al. (2013) used the same method to study SNPs behind the development of all five indicators of behavioral disinhibition mentioned above. Irons (2012) focused on the effect of several factors affecting alcohol use in the study population, namely the effects of polymorphisms in the ALDH2 gene and the GABA system genes, as well as the effect of early exposure to alcohols as adolescents to adult outcomes. Finally Coombes (2016) used a bootstrap-based combination test and a sequential score test to evaluate gene-environment interactions behind phenotypic outcomes in the data.

2.2 Statistical model

We shall use a Linear Mixed Model (LMM) with three variance components accounting for several potential sources of variation to model effect of SNPs behind a quantitative phenotype. This is known as *ACE model* in the literature (Kohler et al., 2011). While the-state-of-the-art focuses on using a single SNP and other covariates as fixed effects, we shall incorporate all SNPs that are genotyped within a gene (or group of genes in some cases) in the set of fixed effects.

We assume that the families modeled are unrelated to one another, i.e. they are nuclear pedigrees. We stick to this structure for ease of representation, although the model fitting process remains unchanged for larger pedigrees. Suppose there are m families in total, with the i-th pedigree containing n_i individuals. Denote by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ the quantitative trait values for individuals in that pedigree, while the matrix $\mathbf{G}_i \in \mathbb{R}^{n_i \times p_g}$ contains their genotypes for a number of SNPs. Let $\mathbf{C}_i \in \mathbb{R}^{n_i \times p}$ denote the data on p covariates for individuals in the pedigree i. Given these, we consider the following model.

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i \boldsymbol{\beta}_g + \mathbf{C}_i \boldsymbol{\beta}_c + \boldsymbol{\epsilon}_i \tag{2.1}$$

with α the intercept term, $\boldsymbol{\beta}_g$ and $\boldsymbol{\beta}_c$ fixed coefficient terms corresponding to the multiple SNPs and covariates, respectively, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{V}_i)$ the random error term. To account for the within-family dependency structure, we break up the random error variance into three independent components:

$$\mathbf{V}_i = \sigma_a^2 \boldsymbol{\Phi}_i + \sigma_c^2 \mathbf{1}\mathbf{1}^T + \sigma_e^2 \mathbf{I}_{n_i} \tag{2.2}$$

The first component above represents a within-family random effect term to account for effects of other SNPs. The matrix $\boldsymbol{\Phi}_i$ is the relationship matrix within the i-th pedigree. Its (s, t)-th element represents two times the kinship coefficient, which is the probability that two alleles, one randomly chosen from individual s in pedigree i and the other from individual t, are 'identical by descent', i.e. come from same common ancestor. The second variance component accounts for shared environmental effect within the family, while the third term finally quantifies other sources of variation unique to an individual.

Following basic probability, the kinship coefficient of a parent-child pair is 1/4, a full sibling pair or non-identical (or dizygous = DZ) twins is 1/4, and for identical (or monozygous = MZ) twins is 1/2 in a nuclear pedigree. Following this, we can construct the $\boldsymbol{\Phi}_i$ matrices for different types of families:

$$\boldsymbol{\Phi}_{MZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1 \\ 1/2 & 1/2 & 1 & 1 \end{bmatrix}, \boldsymbol{\Phi}_{DZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{bmatrix}, \boldsymbol{\Phi}_{Adopted} = \mathbf{I}_4$$

for families with parents (indices 1 and 2) and MZ twins, DZ twins, or two adapted children (indices 3 and 4), respectively.

We use the R package `regress` to fit the above model with additive error structure. The package requires specifying the dependency structure of all samples in the data. For ease of representation, we only consider nuclear pedigrees with MZ twins in our simulation study and data analysis, which simplifies the overall relationship matrix $\mathbf{\Phi} = \text{diag}(\mathbf{\Phi}_1, \dots, \mathbf{\Phi}_m)$ as $\mathbf{I}_m \otimes \mathbf{\Phi}_{MZ}$. Note that, situations in which the pedigree structure is not nuclear can be readily handled in this situation by supplying the overall $\mathbf{\Phi}$ matrix. The second overall structural component will be $\mathbf{I}_m \otimes \mathbf{1}\mathbf{1}^T$. The `regress` procedure includes the third structure in (2.2) by default.

3   Variable selection using e-values

3.1   Models and evaluation maps

In a general modelling situation where one needs to estimate a set of parameters $\boldsymbol{\theta} \in \mathbb{R}^{p_n}$ from a triangular array of samples $\mathcal{B}_n = \{B_{n1}, \dots, B_{nk_n}\}$ at stage n, any hypothesis or statistical model corresponds to a subset of the full parameter space. Here we consider the model spaces $\mathbf{\Theta}_{mn} \subseteq \mathbb{R}^{p_n}$ in which some elements of the parameter vector have fixed values, while others are estimated from the data. Thus a generic parameter vector $\boldsymbol{\theta}_{mn} \in \mathbf{\Theta}_{mn}$ shall consist of entries

$$\theta_{mnj} = \begin{cases} \text{Unknown } \theta_{mnj} & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases}$$

for some $\mathcal{S}_n \subseteq \{1, \dots, p\}$. Thus the estimable index set $\mathcal{S}_n$ and fixed elements $\mathbf{c}_n = (c_{nj} : j \notin \mathcal{S}_n)$ fully specify any model in this setup.

At this point, the original framework in Majumdar and Chatterjee (2017) introduces a few concepts in order to provide a detailed treatment considering a general scenario. For our specific problem, i.e. variable selection, we shall only require vastly simplified versions of them.

To start with, we consider a fixed covariate dimension $p_n = p$ and sample size $k_n = n$ for all n. Consequently we also drop the subscripts in $\mathcal{S}_n$ and $\mathbf{c}_n$. We denote such a candidate model by $\mathcal{M} := (\mathcal{S}, \mathbf{c})$. Thus the 'full model', i.e. the model with all covariates is denoted by $\mathcal{M}_* = (\{1, \dots, p\}, \emptyset)$. We also assume that there is a 'true' data-generating parameter vector $\boldsymbol{\theta}_0$, some elements of which

are potentially set to 0. We can now classify any of the candidate models into two classes: those that contain $\boldsymbol{\theta}_0$, and those do not. We denote these two types of models by *adequate* and *inadequate models*, respectively.

Given data of size n and an unknown $\boldsymbol{\theta}_0$, we want to determine if a candidate model belongs which of the above two categories. For this we need to obtain coefficient estimates $\hat{\boldsymbol{\theta}}_m$ corresponding to a model. We obtain the full model estimates as minimizers of an estimating equation:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Psi_i(\boldsymbol{\theta}, B_i) \tag{3.1}$$

The only conditions we impose on these generic estimating functionals $\Psi_i(.)$ are:

(P1) The true parameter vector $\boldsymbol{\theta}_0$ is the unique minimizer of the population version of (3.1), i.e.

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \mathbb{E} \sum_{i=1}^{n} \Psi_i(\boldsymbol{\theta}, B_i)$$

(P2) There exist a sequence of positive numbers $a_n \uparrow \infty$ and a p-dimensional probability distribution $\mathbb{T}_0$ such that $a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightsquigarrow \mathbb{T}_0$.

We define coefficient estimates $\hat{\boldsymbol{\theta}}_m$ corresponding to any other candidate model $\mathcal{M} = (\mathcal{S}, \mathbf{c})$ by replacing the elements of $\hat{\boldsymbol{\theta}}$ not in $\mathcal{S}$ by corresponding elements of $\mathbf{c}$. This means that for the j-the element, $j = 1, \ldots, p$, we have

$$\hat{\theta}_{mj} = \begin{cases} \text{Unknown } \hat{\theta}_j & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_j & \text{for } j \notin \mathcal{S}_n \end{cases}$$

From now on we denote the probability distribution of a random variable $\mathbf{T}$ by $[\mathbf{T}]$. With this notation, we want to be able to compare the above model estimate distributions with the full model distribution, i.e. $[\hat{\boldsymbol{\theta}}_m]$ with $[\hat{\boldsymbol{\theta}}]$. For this we define an *evaluation map* function $E : \mathbb{R}^p \times \tilde{\mathbb{R}}^p \to [0, \infty)$ that measures the relative position of $\hat{\boldsymbol{\theta}}_m$ with respect to $[\hat{\boldsymbol{\theta}}]$. Here $\tilde{\mathbb{R}}^p$ is the set of probability measures on $\mathbb{R}^p$. We assume that this function satisfies the following conditions:

(E1) For any probability distribution $\mathbb{G} \in \tilde{\mathbb{R}}^p$ and $\mathbf{x} \in \mathbb{R}^p$, E is invariant under location and scale transformations:

$$E(\mathbf{x}, \mathbb{G}) = E(a\mathbf{x} + \mathbf{b}, [a\mathbb{G} + \mathbf{b}]); \quad a \in \mathbb{R} \neq 0, \mathbf{b} \in \mathbb{R}^p$$

where the random variable $\mathbf{G}$ has distribution $\mathbb{G}$.

(E2) The evaluation map E is lipschitz continuous under the first argument:

$$|E(\mathbf{x}, \mathbb{G}) - E(\mathbf{y}, \mathbb{G})| < \|\mathbf{x} - \mathbf{y}\|^{\alpha}; \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^p, \alpha > 0$$

(E3) Suppose $\{\mathbb{Y}_n\}$ is a tight sequence of probability measures in $\tilde{\mathbb{R}}^p$ with weak limit $\mathbb{Y}_\infty$. Then $E(\mathbf{x}, \mathbb{Y}_n)$ converges uniformly to $E(\mathbf{x}, \mathbb{Y}_\infty)$.

(E4) Suppose $\mathbf{Z}_n$ is a sequence of random variables such that $\|\mathbf{Z}_n\| \overset{P}{\to} \infty$. Then $E(\mathbf{Z}_n, \mathbb{Y}_n) \overset{P}{\to} 0$.

For any $\mathbf{x} \in \mathbb{R}^p$ and $[\mathbf{X}] \in \tilde{\mathbb{R}}^p$ such that $\mathbb{V}\mathbf{X}$ is positive definite, following are examples of the evaluations functions covered by the above set of conditions:

$$E_1(\mathbf{x}, [\mathbf{X}]) = \left[1 + \left\|\frac{\mathbf{x} - \mathbb{E}\mathbf{X}}{\sqrt{\mathrm{diag}(\mathbb{V}\mathbf{X})}}\right\|^2\right]^{-1}; \quad E_2(\mathbf{x}, [\mathbf{X}]) = \exp\left[-\left\|\frac{\mathbf{x} - \mathbb{E}\mathbf{X}}{\sqrt{\mathrm{diag}(\mathbb{V}\mathbf{X})}}\right\|\right] \tag{3.2}$$

Data depths (Tukey, 1975; Zuo, 2003; Zuo and Serfling, 2000) also constitute a very broad class of point-to-distribution proximity functions that satisfy the above regularity conditions for evaluation maps. Indeed, Majumdar and Chatterjee (2017) had used halfspace depth (Tukey, 1975) as evaluation function to perform model selection. However, the conditions (E1) and (E4) are weaker than those imposed on a traditional depth function (Zuo and Serfling, 2000). Lipschitz continuity and uniform convergence are not required of depth functions in general, but they arise implicitly in several implementations of data depth (Mosler, 2013). The theoretical results we state here are based on a general evaluation map and not depth functions *per se*. To emphasize this point, in the numerical sections that follow we shall use the non-depth evaluation functions E_1 and E_2 given in (3.2) above.

3.2 Model selection using e-values

Depending on the choice of the data sequence $\mathcal{B}_n$, $E(\hat{\boldsymbol{\theta}}_m, [\hat{\boldsymbol{\theta}}])$ can take different values. For any candidate model $\mathcal{M}$, we shall denote the distribution of the corresponding random evaluation map by $\mathbb{E}_{mn}$. For simplicity we shall drop the n in the subscripts for such sampling distributions, i.e. $\mathbb{E}_{mn} \equiv \mathbb{E}_m$. These sampling distributions are informative of how model estimates behave, and we shall use them as a tool to distinguish between inadequate and adequate models. Given a single set of samples, we shall use resampling schemes that satisfy standard regularity conditions (Majumdar and

Chatterjee, 2017) to get consistent approximations of $\mathbb{E}_m$.

We now define a quantity called the *e-***value**, through which we shall be able to compare the different model estimates and eventually perform selection of important SNPs from a multi-SNP model. Loosely construed, any functional of the evaluation map distribution $\mathbb{E}_m$ that can act as model evidence is an *e*-value. For example, under a much general setup Majumdar and Chatterjee (2017) took the mean functional of $\mathbb{E}_m$ (say $\mu(\mathbb{E}_m)$) as *e*-value, and proved a result that, when adapted to our setting, states as:

Theorem 3.1. *Consider estimators satisfying conditions (P1) and (P2), and an evaluation map E satisfying the conditions (E1), (E2) and (E4). Also suppose that*

$$\lim_{n \to \infty} \mu(\mathbb{Y}_n) = \mu(\mathbb{Y}_\infty) < \infty$$

for any tight sequence of probability measures $\{\mathbb{Y}_n\}$ in $\tilde{\mathbb{R}}^p$ with weak limit $\mathbb{Y}_\infty$. Then as $n \to \infty$,

1. *For the full model, $\mu(\mathbb{E}_*) \to \mu_\infty$ for some $0 < \mu_\infty < \infty$;*

2. *For any adequate model, $|\mu(\mathbb{E}_m) - \mu(\mathbb{E}_*)| \to 0$;*

3. *For any inadequate model, $\mu(\mathbb{E}_m) \to 0$.*

Taking data depths as evaluation functions leads to a further result that $\mu(\mathbb{E}_*) < \mu(\mathbb{E}_m)$ for any adequate model $\mathcal{M}$ and large enough n. Following this, non-zero indices of $\boldsymbol{\theta}_0$ (say $\mathcal{S}_0$) can be recovered through a fast algorithm that has these generic steps:

1. Estimate the *e*-value of the full model, i.e. $\hat{\mu}(\mathbb{E}_*)$, through bootstrap approximation of $\mathbb{E}_*$;

2. For the j-th predictor, $j = 1, \ldots, p$, consider the model with the j-th coefficient of $\hat{\boldsymbol{\theta}}$ replaced by 0, and get its *e*-value. Suppose this is $\hat{\mu}(\mathbb{E}_{-j})$;

3. Collect the predictors for which $\hat{\mu}(\mathbb{E}_{-j}) < \hat{\mu}(\mathbb{E}_*)$. Name this index set $\hat{\mathcal{S}}_0$: this is the estimated set of non-zero coefficients in $\hat{\boldsymbol{\theta}}$.

As $n \to \infty$, the above algorithm provides consistent model selection, i.e. $\mathbb{P}(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \xrightarrow{P} 1$, with the underlying resampling distribution having mean 1 and variance τ_n^2 such that $\tau_n \to \infty, \tau_n/\sqrt{n} \to 0$ as $n \to \infty$ (Majumdar and Chatterjee, 2017).

3.3 Quantiles of $\mathbb{E}_m$ as e-values

The above formulation of e-values leads to favorable finite sample results compared to several existing covariate selection techniques in linear and linear mixed models (Majumdar and Chatterjee, 2017). However, its performance is dependent on the relative magnitude of the non-zero coefficients to the random error term, i.e. the signal-to-noise ratio (SNR). When the true signals are weak, the above method of variable selection leads to very conservative estimates of non-zero coefficient indices, i.e. a large number of false positives. This happens because even though at the population level $\mu(\mathbb{E}_*)$ separates the population means of inadequate model sampling distributions and those of adequate models, for weak signals bootstrap estimates of adequate model distributions almost overlap with those of the full model.

Figure 1 demonstrates this phenomenon in our simulation setup. Here we analyze data on 250 families with monozygotic twins, each individual being genotyped for 50 SNPs. Four of these 50 SNPs are causal: each having a heritability of $h/6\%$ with respect to the total error variation present. The four panels show density plots of $\hat{\mathbb{E}}_{-j}$ for $j = 1, \ldots, p$, as well as $\hat{\mathbb{E}}_*$: based on resampling schemes with four different values of the standard deviation parameter $s \equiv s_n = \tau_n/\sqrt{n}$. While smaller values of s are able to separate out the bootstrap estimates of $\mathbb{E}_{-j}$ for inadequate and adequate models, all the density plots are to the left of the curve corresponding to the full model.

However, notice that the inadequate and adequate model distributions have different tail behaviors for smaller values of s, and setting an appropriate upper threshold to tail probabilities for a suitable fixed quantile of these distributions with respect to the full model distribution can possibly provide a better separation of the two types of distributions. For this reason we shall use tail quantiles as e-values.

For $q \in (0, 1)$, denote the corresponding q-th quantile of $\mathbb{E}_m$ by $c_q(\mathbb{E}_m)$. Given that the estimator $\hat{\boldsymbol{\theta}}$ satisfies conditions (P1) and (P2), and the evaluation map satisfies conditions (E1)-(E4), we have equivalent results to Theorem 3.1 for the population quantiles as $n \to \infty$:

$$c_q(\mathbb{E}_*) \to c_{q,\infty} < \infty \tag{3.3}$$

$$|c_q(\mathbb{E}_m) - c_q(\mathbb{E}_*)| \to 0 \text{ when } \mathcal{M} \text{ is adequate} \tag{3.4}$$

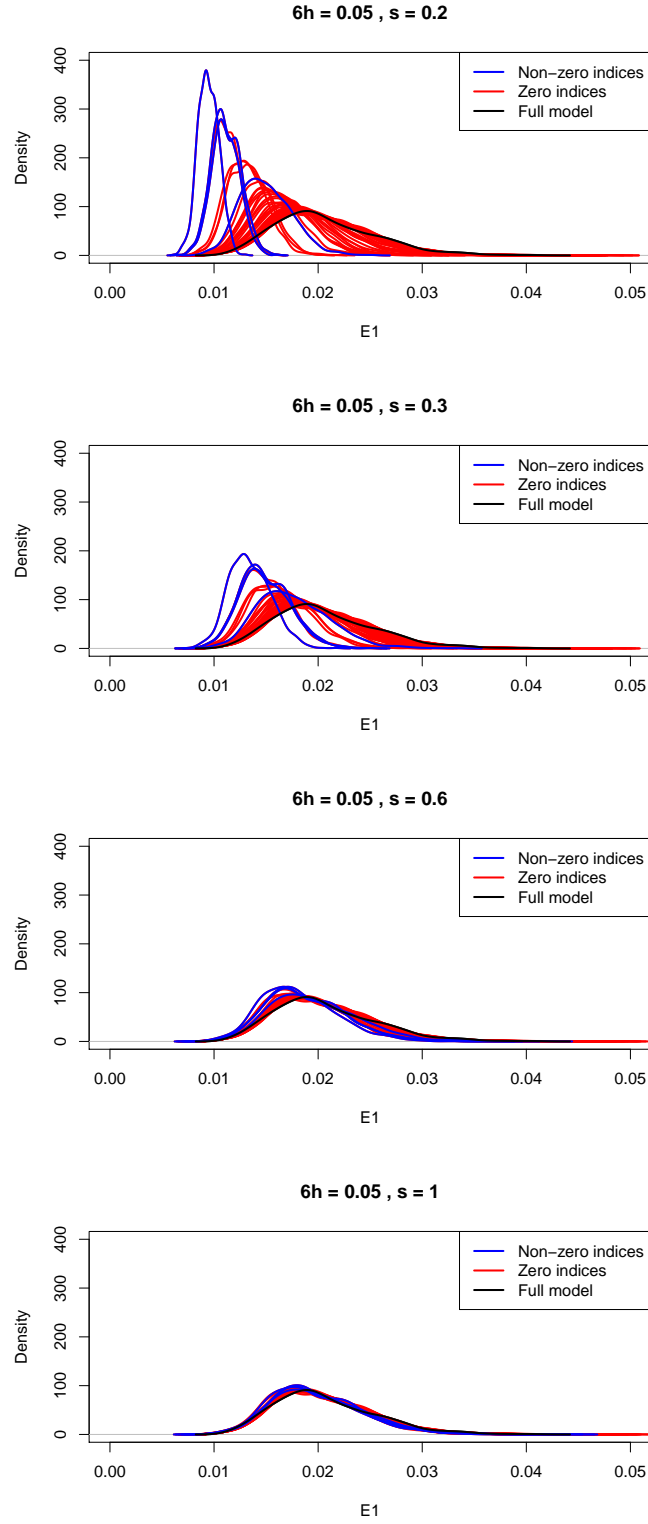$$c_q(\mathbb{E}_m) \to 0 \text{ when } \mathcal{M} \text{ is inadequate} \tag{3.5}$$

Figure 1: Density plots of bootstrap approximations for $\mathbb{E}_*$ and $\mathbb{E}_{-j,n}$ for all j in simulation setup, with $s = 0.2, 0.3, 0.6, 1$

These results are proved by taking a similar approach as that of proving Theorem 3.1. We do this in the supplementary material (**?**).

When the q-th quantile is taken as the e-value, we set a lower detection threshold than the same functional on the full model, i.e. choose all j such that $c_q(\mathbb{E}_{-j}) < c_{qt}(\mathbb{E}_*), 0 < t < 1$ to be included in the model. The choice of t potentially depends on several factors such as the value of quantile evaluated, the statistical model used, sample size and degree of sparsity of parameters in the data generating process. We illustrate this point on simulated data in Section 4.

3.4 Bootstrap procedure

We use generalized bootstrap (Chatterjee and Bose, 2005) to obtain approximations of the sampling distributions $\mathbb{E}_{-j}$ and $\mathbb{E}_*$. It calculates bootstrap equivalents of the parameter estimate $\hat{\boldsymbol{\theta}}$ by minimizing a version of the estimating equation in (3.1) with random weights:

$$\hat{\boldsymbol{\theta}}_w = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathbb{W}_i \Psi_i(\boldsymbol{\theta}, B_i) \tag{3.6}$$

The resampling weights $(\mathbb{W}_1, \ldots, \mathbb{W}_n)$ are non-negative exchangeable random variables chosen independent of the data, and satisfy the following conditions:

$$\mathbb{E}\mathbb{W}_1 = 1 \tag{3.7}$$

$$\mathbb{V}\mathbb{W}_1 = \tau_n^2 \uparrow \infty \tag{3.8}$$

$$\tau_n^2 = o(a_n^2) \tag{3.9}$$

$$\mathbb{E}W_1 W_2 = O(n^{-1}) \tag{3.10}$$

$$\mathbb{E}W_1^2 W_2^2 \to 1 \tag{3.11}$$

$$\mathbb{E}W_1^4 < \infty \tag{3.12}$$

with $W_i := (\mathbb{W}_i - 1)/\tau_n; i = 1, \ldots, n$ being the centered and scaled resampling weights. Under standard regularity conditions on the estimating functional $\Psi(.)$ (Chatterjee and Bose, 2005; Majumdar and Chatterjee, 2017) and conditional on the data, $(a_n/\tau_n)(\hat{\boldsymbol{\theta}}_w - \hat{\boldsymbol{\theta}})$ converges to the same asymptotic distribution as $a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, i.e. $\mathbb{T}_0$.

We use empirical quantiles of the full model bootstrap samples as the quantile e-value estimates.

Specifically, we go through the following steps:

1. Fix $q, t \in (0, 1)$;

2. Generate two independent set of bootstrap weights, of size R and R_1, and obtain the corresponding approximations to the full model sampling distribution, say $[\hat{\boldsymbol{\theta}}_r]$ and $[\hat{\boldsymbol{\theta}}_{r_1}]$;

3. For $j = 1, 2, \ldots p$ and estimate the e-value of the j-th predictor as the empirical q-th quantile of $\hat{\mathbb{E}}_{-j} := [E(\hat{\boldsymbol{\theta}}_{r,-j}, [\hat{\boldsymbol{\theta}}_{r_1}])]$, with $\hat{\boldsymbol{\theta}}_{r,-j}$ obtained from $\hat{\boldsymbol{\theta}}_r$ by replacing the j-th coordinate with 0;

4. Estimate the set of non-zero covariates as

$$\hat{\mathcal{S}}_0 = \{j : c_q(\hat{\mathbb{E}}_{-j}) < c_{qt}(\hat{\mathbb{E}}_*)\};$$

The conditions (3.7)-(3.12) on the resampling weights ensure bootstrap-consistent approximation of the evaluation map quantiles:

$$|c_q(\hat{\mathbb{E}}_m) - c_q(\hat{\mathbb{E}}_*)| \overset{P_n}{\to} o_P(1) \text{ when } \mathcal{M} \text{ is adequate} \tag{3.13}$$

$$c_q(\hat{\mathbb{E}}_m) \overset{P_n}{\to} o_P(1) \text{ when } \mathcal{M} \text{ is inadequate} \tag{3.14}$$

The generalized bootstrap method covers a large array of resampling procedures, for example the m-out-of-n bootstrap and a scale-enhanced version of the bayesian bootstrap. Furthermore, given that $\Psi_i(.)$ are twice differentiable in a neighborhood of $\boldsymbol{\theta}_0$ and some other conditions in Chatterjee and Bose (2005), there is an approximate representation of $\hat{\boldsymbol{\theta}}_w$:

$$\hat{\boldsymbol{\theta}}_w = \hat{\boldsymbol{\theta}} - \frac{\tau_n}{a_n} \left[\sum_{i=1}^n W_i \Psi_i''(\hat{\boldsymbol{\theta}}, B_i) \right]^{-1/2} \sum_{i=1}^n W_i \Psi_i'(\hat{\boldsymbol{\theta}}, B_i) + \mathbf{R}_{wn} \tag{3.15}$$

with $\mathbb{E}_w \|\mathbf{R}_{wn}\|^2 = o_P(1)$.

Given the full model estimate $\hat{\boldsymbol{\theta}}$, and the score vectors $\Psi_i'(\hat{\boldsymbol{\theta}}, B_i)$ and hessian matrices $\Psi_i''(\hat{\boldsymbol{\theta}}, B_i)$, (3.15) allows us to obtain multiple copies of $\hat{\boldsymbol{\theta}}_w$ through Monte-Carlo simulation of several arrays of bootstrap weights. This bypasses the need to fit the full model for each bootstrap sample, resulting in extremely fast computation of e-values.

We adapt the approximation of (3.15) to the LMM in (2.1). We first obtain the maximum likeli-

hood estimates $\hat{\boldsymbol{\beta}}_g, \hat{\sigma}_a^2, \hat{\sigma}_c^2, \hat{\sigma}_e^2$ through fitting the LMM with `regress`. Then we replace the variance components in (2.2) with corresponding estimates to get $\hat{\mathbf{V}}_i$ for i-th pedigree, and aggregate them to get the covariance matrix estimate for all samples:

$$\hat{\mathbf{V}} = \mathrm{diag}(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_m)$$

We take m random draws from $\mathrm{Gamma}(1, 1) - 1$, say $\{w_{r1}, \ldots, w_{rm}\}$, as resampling weights in (3.15), using the same weight for all members of a pedigree. It is now straightforward that the bootstrapped coefficient estimate $\hat{\boldsymbol{\beta}}_{rg}$ has the following representation:

$$\hat{\boldsymbol{\beta}}_{rg} \simeq \hat{\boldsymbol{\beta}}_g + \frac{\tau_n}{\sqrt{n}}(\mathbf{G}^T\hat{\mathbf{V}}^{-1}\mathbf{G})^{-1}\mathbf{W}_r\mathbf{G}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{G}\hat{\boldsymbol{\beta}}_g) \qquad (3.16)$$

with $\mathbf{G} = (\mathbf{G}_1^T, \ldots, \mathbf{G}_m^T)^T$ and $\mathbf{W}_r = \mathrm{diag}(w_{r1}\mathbf{I}_4, \ldots, w_{rm}\mathbf{I}_4)$. Finally we repeat the procedure for two independent sets of resampling weights, say of sizes R and R_1, to obtain two collections of bootstrapped estimates $\{\hat{\boldsymbol{\beta}}_{1g}, \ldots, \hat{\boldsymbol{\beta}}_{Rg}\}$.

4 Simulation

We now evaluate the performance of the above formulation of quantile e-values in a simulation setup. For this, consider the model in (2.1) with no environmental covariates. We consider familes with MZ twins and first generate the SNP matrices $\mathbf{G}_i$. We take a total of $p_g = 50$ SNPs, and to simulate correlation among SNPs in the genome generate them in correlated blocks of 6, 4 ,6, 4 and 30. We set the correlation between two SNPs inside a block at 0.7, and consider the blocks to be uncorrelated. For each parent we generate two independent vectors of length 50 with the above correlation structure, and entries within each block being 0 or 1 following Bernoulli distributions with probabilities 0.2, 0.4, 0.4, 0.25 and 0.25 (Minor Allele Frequency or MAF) for SNPs in the 5 blocks, respectively. The genotype of a person is then determined by taking the sum of these two vectors: thus entries in $\mathbf{G}_i$ can take the values 0, 1 or 2. Finally we set the common genotype of the twins by randomly choosing one allele vector from each of the parents and taking their sum.

We repeat the above process for $m = 250$ families. In GWAS there are generally a small number of causal SNPs, each explaining small proportions of the overall variability in response variable. To reflect this in our simulation setup, we assume that the first entries in each of the first four blocks

above are causal, and each of them explains $h/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)\%$ of the overall variability. The term h is known as the *heritability* of the corresponding SNP (and can of course vary across SNPs). The value of the non-zero coefficient in k-th group: $k = 1, ..., 4$, say β_k is calculated using the formula:

$$\beta_k = \sqrt{\frac{h}{(\sigma_a^2 + \sigma_c^2 + \sigma_e^2).2\mathrm{MAF}_k(1 - \mathrm{MAF}_k)}} \tag{4.1}$$

We fix the following values for the error variance components: $\sigma_a^2 = 4, \sigma_c^2 = 1, \sigma_e^2 = 1$, and generate pedigree-wise response vectors $\mathbf{y}_1, \dots, \mathbf{y}_{250}$ using the above setup. To consider different SNP effect sizes, we repeat the above setup for $h \in \{10, 7, 5, 3, 2, 1, 0\}$, generating 1000 datasets for each value of h.

4.1 Methods and metrics

For this simulated data, we compare our e-value based approach using the evaluation maps E_1 and E_2 in 3.2 with two other methods:

(1) Model selection on linear model: Here we ignore the dependency structure within families by training linear models on the simulated data and selecting SNPs with non-zero effects by backward deletion using a modification of the BIC called mBIC2. This has been showed to give better results than single-marker analysis in GWAS for unrelated individuals (Frommelet et al., 2012) and provides approximate False Discovery Rate (FDR) control at level 0.05 (Bogdan et al., 2011).

(2) Single-marker mixed model: We train single-SNP versions of (2.1) using a fast approximation of the Generalized Least Squares procedure (named Rapid Feasible Generalized Least Squares or RFGLS: Li et al. (2011)), obtain marginal p-values from corresponding t-tests and use the Benjamini-Hochberg (BH) procedure to select significant SNPs at FDR $= 0.05$.

With the e-value being the q-th quantile of the evaluation map distribution, we set the detection threshold value at the t-th multiple of q for some $0 < t < 1$. This means all indices j such that q-th quantile of the bootstrap approximation of $\mathbb{E}_{-j}$ is less than the tq-th quantile of the bootstrap approximation of $\mathbb{E}_*$ will get selected as the set of active predictors. To enforce stricter control on the selected set of SNPs we repeat this for $q \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, and take the SNPs that get selected for *all* values of q as the final set of selected SNPs.

Finally the above procedure depends on the bootstrap standard deviation parameter s. Consequently, we repeat the process for $s \in \{0.3, 0.15, \ldots, 0.95, 2\}$, and take as the final estimated set of SNPs the SNP set $\hat{\mathcal{S}}(s, t)$ that minimizes fixed effect prediction error (PE) on an independently generated test dataset $\{(\mathbf{y}_{test,i}, \mathbf{G}_{test,i}), i = 1, \ldots, 250\}$ from the same setup above:

$$\text{PE}(s, t) = \sum_{i=1}^{250} \sum_{j=1}^{4} \left(y_{test,ij} - \mathbf{g}_{test,ij}^T \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}(s,t)} \right)^2;$$

$$\hat{\mathcal{S}}_0(t) = \arg\min_s \text{PE}(s, t)$$

The metrics to evaluate each method we implement are:

1. True Positive (TP): proportion of causal SNPs detected;

2. True Negative (TN): proportion of non-causal SNPs undetected;

3. Relaxed True Positive (TPR): proportion of detecting any SNP in each of the 4 blocks with causal SNPs, i.e. for the selected index set $\hat{\mathcal{S}}_0(q, t)$,

$$\text{TPR}(\hat{\mathcal{S}}_0(q, t)) = \frac{1}{4} \sum_{i=1}^{4} \mathbb{I}(\text{Block } i \cap \hat{\mathcal{S}}_0(q, t) \neq \emptyset)$$

4. Relaxed True Negative (TNR): proportion of SNPs in block 5 undetected.

We consider the third and fourth metrics to cover situations in which the causal SNP is not detected itself, but highly correlated SNPs with the causal SNP are. This is common in GWAS. Finally, we average all the above proportions over 1000 replications, and repeat the process for two different ranges of t for E_1 and E_2.

4.2   Results

We present the simulation results in table 1. Applying mBIC2 on linear models performs poorly compared RFGLS and then correction for multiple testing on marginal LMMs for all heritability values. This is expected because the linear model ignores the within-family error components.

Our proposed e-values work better than the two methods for detecting true signals across different values of h: the average TP rate going down slowly than other methods across the majority of choices for t. Both mBIC2 and RFGLS+BH have very high true negative detection rates, which is matched

	Method	$h=10$	$h=7$	$h=5$	$h=3$	$h=2$	$h=1$	$h=0$
	mBIC2	0.79/0.99	0.59/0.99	0.41/0.99	0.2/0.99	0.11/0.99	0.05/0.99	-/0.99
	RFGLS+BH	0.95/0.92	0.82/0.95	0.62/0.97	0.29/0.98	0.14/0.99	0.04/1	-/1
E_1	$t=\exp(-1)$	0.95/0.98	0.87/0.97	0.74/0.97	0.47/0.97	0.28/0.97	0.12/0.98	-/0.99
	$t=\exp(-2)$	0.94/0.98	0.85/0.98	0.69/0.98	0.43/0.98	0.25/0.98	0.09/0.99	-/0.99
	$t=\exp(-3)$	0.94/0.99	0.82/0.98	0.65/0.98	0.37/0.99	0.2/0.99	0.07/0.99	-/1
	$t=\exp(-4)$	0.92/0.99	0.79/0.99	0.61/0.99	0.32/0.99	0.17/0.99	0.06/1	-/1
	$t=\exp(-5)$	0.9/0.99	0.75/0.99	0.55/0.99	0.26/1	0.13/1	0.04/1	-/1
E_2	$t=0.8$	0.97/0.98	0.9/0.97	0.79/0.96	0.54/0.96	0.34/0.97	0.15/0.98	-/0.99
	$t=0.74$	0.96/0.98	0.88/0.97	0.75/0.97	0.48/0.97	0.29/0.98	0.12/0.98	-/0.99
	$t=0.68$	0.95/0.99	0.87/0.98	0.72/0.98	0.45/0.98	0.26/0.98	0.1/0.99	-/0.99
	$t=0.62$	0.95/0.99	0.84/0.98	0.68/0.98	0.4/0.99	0.22/0.99	0.09/0.99	-/0.99
	$t=0.56$	0.94/0.99	0.82/0.99	0.65/0.99	0.36/0.99	0.19/0.99	0.07/1	-/1
	$t=0.5$	0.92/0.99	0.79/0.99	0.6/0.99	0.31/0.99	0.16/1	0.05/1	-/1

	Method	$h=10$	$h=7$	$h=5$	$h=3$	$h=2$	$h=1$	$h=0$
	mBIC2	0.84/0.99	0.66/0.99	0.48/0.99	0.26/0.99	0.16/0.99	0.08/0.99	-/0.98
	RFGLS+BH	0.96/0.99	0.83/0.99	0.64/0.99	0.32/0.99	0.16/1	0.05/1	-/1
E_1	$t=\exp(-1)$	0.95/0.98	0.87/0.97	0.75/0.97	0.5/0.97	0.32/0.98	0.15/0.98	-/0.98
	$t=\exp(-2)$	0.94/0.99	0.85/0.98	0.71/0.98	0.45/0.98	0.28/0.98	0.12/0.99	-/0.98
	$t=\exp(-3)$	0.94/0.99	0.83/0.99	0.67/0.99	0.39/0.99	0.22/0.99	0.09/0.99	-/0.99
	$t=\exp(-4)$	0.92/0.99	0.8/0.99	0.62/0.99	0.33/0.99	0.18/0.99	0.07/1	-/1
	$t=\exp(-5)$	0.9/0.99	0.75/0.99	0.56/0.99	0.27/1	0.14/1	0.05/1	-/1
E_2	$t=0.8$	0.97/0.98	0.91/0.97	0.8/0.96	0.57/0.96	0.38/0.97	0.2/0.98	-/0.97
	$t=0.74$	0.96/0.98	0.89/0.98	0.76/0.97	0.51/0.97	0.33/0.98	0.15/0.98	-/0.98
	$t=0.68$	0.95/0.99	0.87/0.98	0.73/0.98	0.48/0.98	0.29/0.98	0.12/0.99	-/0.98
	$t=0.62$	0.95/0.99	0.85/0.99	0.69/0.98	0.42/0.99	0.24/0.99	0.11/0.99	-/0.99
	$t=0.56$	0.94/0.99	0.83/0.99	0.66/0.99	0.38/0.99	0.2/0.99	0.08/0.99	-/0.99
	$t=0.5$	0.92/0.99	0.79/0.99	0.61/0.99	0.32/0.99	0.17/1	0.06/1	-/1

Table 1: (Top) Average True Positive (TP), True Negative (TN) and (Bottom) Average Relaxed True Positive (TPR) and Relaxed True Negative (TNR) proportions over 1000 replications for the e-values method with E_1 and E_2 as evaluation maps, and the other two methods

by our method for higher values of q. Since all reduced model distributions reside on the left of the full model distribution, we expect the variable selection process to turn more conservative at lower values of t.This effect is more noticeable for lower q. This indicates that the right tails of evaluation map distributions are more useful for this purpose. Finally for $h = 0$, we report only TN or TNR values since no signals should ideally be detected: in terms of this a value of $q = 0.9$ or $q = 0.5$ leads to the same TN and TNR performance as RFGLS+BH for all choices of t.

Finally, TPR performances for all methods are better than the corresponding TP/TN performances. However, for mBIC2 this seems to be due to detecting SNPs in the first four blocks by chance since for $h = 0$ its TNR is less than TN. Also E_2 seems to perform slightly better than E_1, in the sense that it yields a higher TP (or TPR) while having the same TN (or TNR) rates.

Gene	Total/detected SNP	Non-zero SNPs ordered per position in genome
GABRA2	11/5	rs572227(-), rs534459(+), rs502038(+), rs1808851(+), rs279856(-)
ADH	44/8	rs10006414(+), rs2034677(+), rs12508445(-), rs10005811(+), rs10516428(-), rs10022047(+), rs12503056(+), rs2004316(-)
OPRM1	46/25	rs9371718(-), rs1937600(-), rs9397637(+), rs1316368(+), rs1937587(+), rs6921403(-), rs1937580(+), rs1937645(+), rs1892361(-), rs1937633(-), rs1937631(-), rs12527197(+), rs1892356(+), rs1937619(-), rs9285539(+), rs9322439(+), rs4869813(+), rs4870241(-), rs9384156(+), rs689219(-), rs12199858(+), rs9371762(-), rs612450(+), rs9384159(+), rs6938958(-)
CYP2E1	9/6	rs10872828(-), rs9419702(-), rs9419624(+), rs7906770(-), rs9419569(+), rs9419629(+)
ALDH2	5/5	rs7398343(+), rs7297186(-), rs3803167(+), rs10219736(-), rs3742004(-)
COMT	15/7	rs4646312(+), rs165656(-), rs165722(+), rs2239393(-), rs174699(+), rs165728(-), rs2518824(+)
SLC6A3	18/6	rs11133767(+), rs2042449(+), rs464049(-), rs460700(-), rs460000(+), rs2617605(+)
SLC6A4	5/1	rs11658318(-)
DRD2	17/3	rs7937641(-), rs10750025(+), rs6589386(+)

Table 2: Table of analyzed genes and detected SNPs in them. Positive/ negative sign indicates type of association found, as indicated by the sign of the corresponding coefficient

5 Analysis of the MCTFR data

We now apply the above technique on SNPs from the MCTFR dataset. We assume a nuclear pedigree structure, and for simplicity only analyze pedigrees with MZ twins. After adjusting for missing response variables in a part of the main dataset, we end up with 682 such 4-member families. We look at the effect of genetic factors behind the response variable pertaining to the amount of alcohol consumption, which has previously been found to be highly heritable in this dataset (McGue et al., 2013). We decide to analyze SNPs inside some of the most-studied genes with respect to alcohol abuse: GABRA2, ADH1B, ADH1C, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT (Coombes, 2016) through separate gene-level models. The ADH genes did not contain many SNPs individually, so we decided to club all existing ADH genes (ADH1-ADH7) together in our analysis. We also include sex, birth year, age and generation (parent or offspring) of individuals to control for their potential effect.

We use E_2 as the evaluation function here because of its slighty better performance in the simulations. For each gene, We train the LMM in (2.1) on 75% of randomly selected families, perform our e-values procedure for $s = 0.2, 0.4, \ldots, 2.8, 3, t = 0.1, 0.15, \ldots, 0.75, 0.8$; and select the predictor set $\hat{\mathcal{S}}(s,t)$ that minimizes fixed effect prediction error on the data from the other 25% of families.

We plot the 90-th quantile e-value estimates from our gene-specific analyses in Figures 2, 3 and 4. We obtained gene locations, as well as the locations of exons inside 6 of these 9 genes from annotation

data extracted from the UCSC Genome Browser database (Rosenbloom et al., 2015). Exon locations were not available for OPRM1, CYP2E1 and DRD2. Also Table 2 summarizes the selected SNPs for each gene. In general, SNPs tend to get selected in groups with neighboring SNPs, which suggests high Linkage Disequilibrium (LD). Also most of the selected SNPs either overlap or in close proximity to the coding regions of genes, i.e. exons, which underline their functional relevance.

Finally, below are some gene-specific observations. We highlight some interesting findings here, and provide full tables of SNPs with more information from the model outputs in the supplementary material.

GABRA2: As seen in the plots, the first two SNPs detected are close to two separate exons. The 4th and 5th detected SNPs, rs1808851 and rs279856, are at perfect LD with rs279858 in the larger 7188-individual dataset (Irons, 2012). This SNP had not been genotyped in our sample, but is the marker in GABRA2 that is most frequently associated in the literature with alcohol abuse (Cui et al., 2012). Interestingly, a single SNP RFGLS analysis of the same twin studies data that used Bonferroni correction on marginal p-values to detect SNPs had missed these SNPs (Irons, 2012). This highlights the advantage of our approach.

ADH genes: Multiple studies have associated rs1229984 in the ADH1B gene (position 99318162 of chromosome 4) with alcohol dependence (`https://www.snpedia.com/index.php/Rs1229984`), which as seen in the plot of ADH2 is close to an exon region. Our data does not contain this marker, but detects rs10022047 and rs10516430 at position 99296818. Macgregor et al. (2008) found strong association between alcohol consumption and the SNP rs1042026 at position 99307309: this is also close to our detected SNP.

OPRM1: Many of the SNPs analyzed in this gene have very low e-values, and tend to cluster together. Although the minor allele of the SNP rs1799971 (chr 6, position 154039662) has been associated with stronger alcohol cravings (`https://www.snpedia.com/index.php/Rs1799971`), we do not detect any SNP that is close to this.
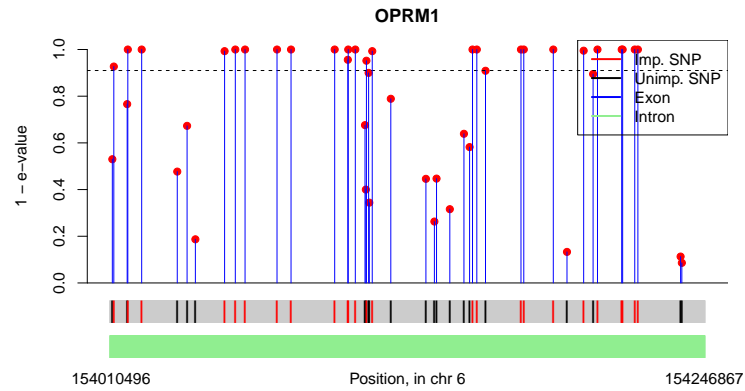
CYP2E1: Seven of the 9 SNPs studied are detected through our analysis. Five of them have very high 90-th quantile e-values, and are within 10 kb of one another (base pairs 133534822 to 133543210 in chr 10). In the analysis of Lind et al. (2012) rs4646976 at 133534223 position was most associated
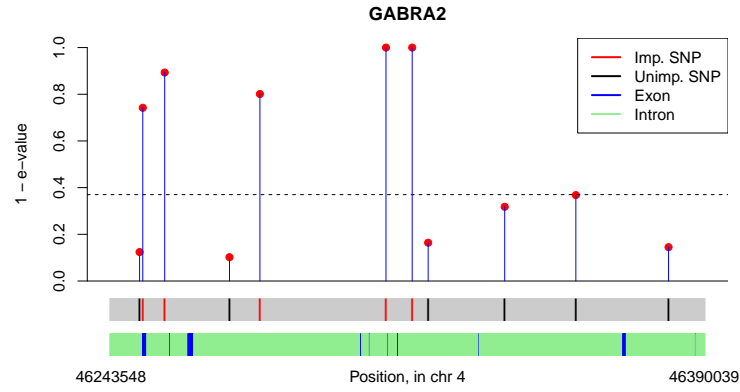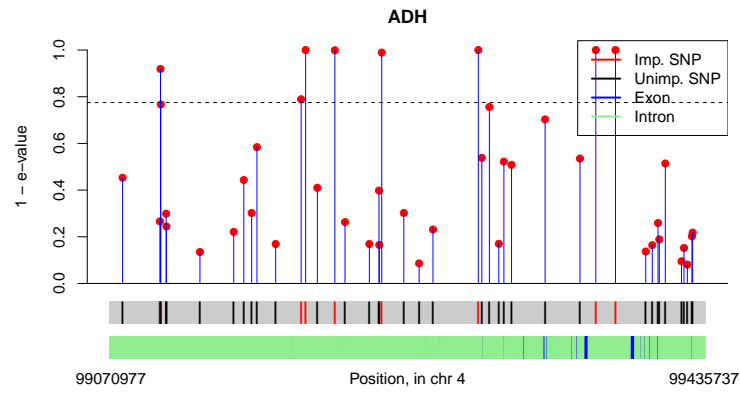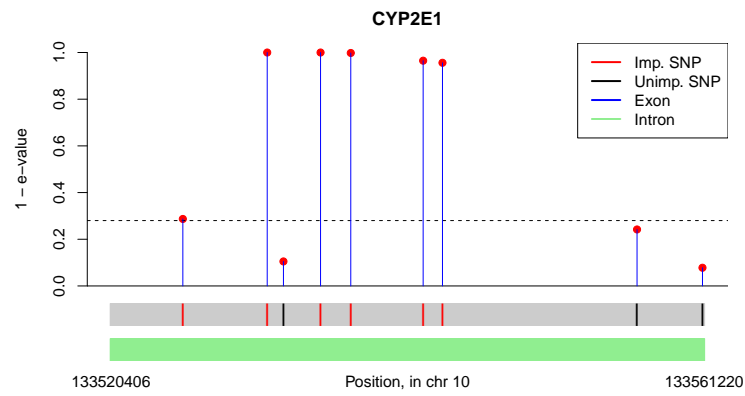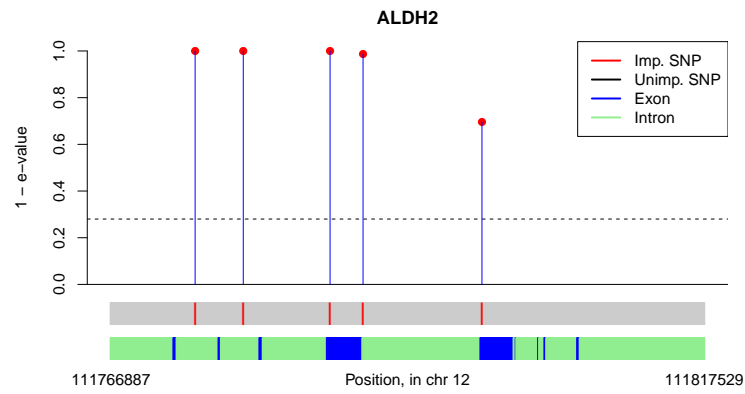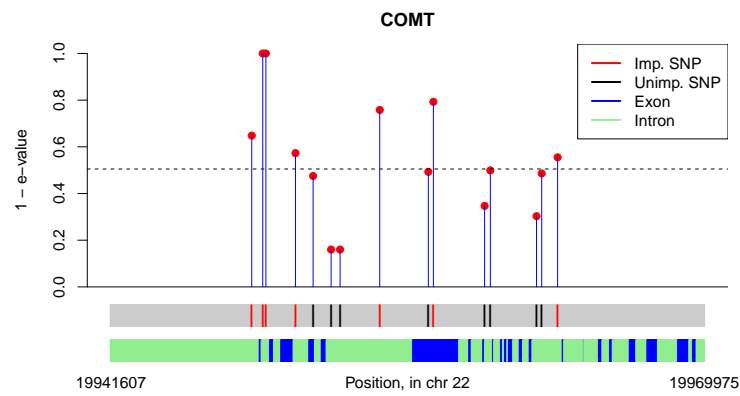
Figure 2: Plot of *e*-values for genes analyzed: (a) GABRA2, (b) ADH1 to ADH7, (c) OPRM1. For ease of visualization, $1 - e$-values are plotted in the y-axis.
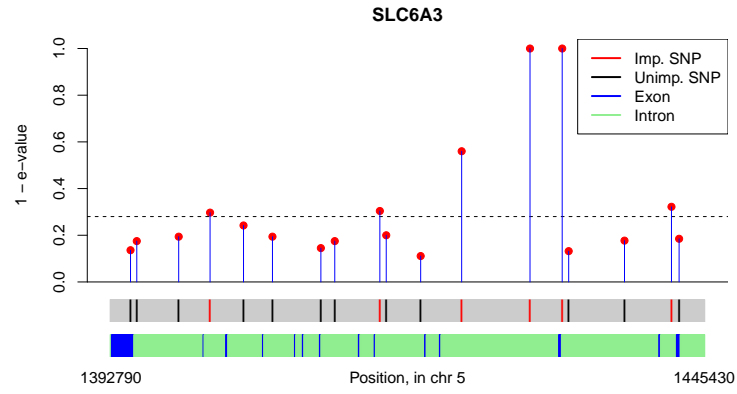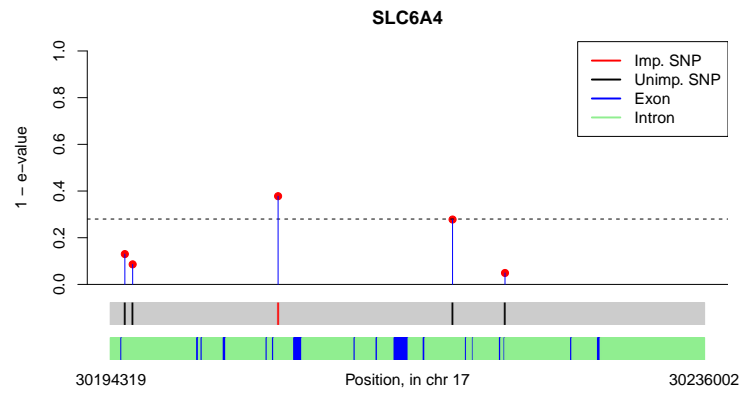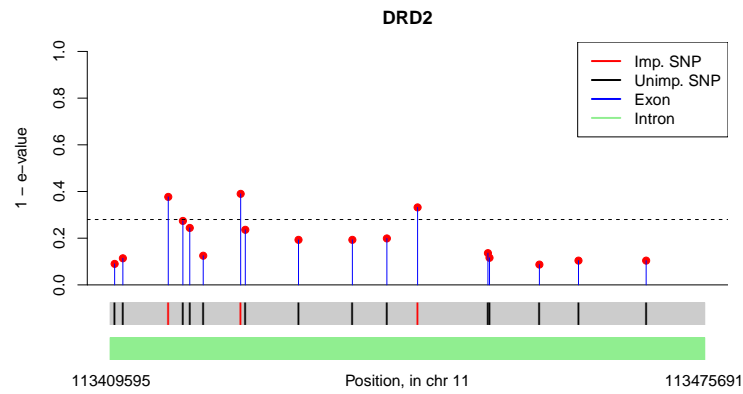
(d)

(e)

(f)

Figure 3: Plot of *e*-values for genes analyzed: (d) CYP2E1, (e) ALDH2, (f) COMT

22

(g)

(h)

(i)

Figure 4: Plot of *e*-values for genes analyzed: (g) SLC6A3, (h) SLC6A4, (i) DRD2

with a measure of breath alcohol concentration: this is within our detected region. This study had also detected rs4838767 in the promoter region of CYP2E1 (position 133520114) associated with multiple alcohol consumption measures, but we did not detect the closest SNP to this as having non-zero effect on our response.

ALDH2: All 5 SNPs we study are close to exons, and get picked up by the e-value procedure. While all five are at a lesser base pair position than the well-known SNP rs671 (`https://www.snpedia.com/index.php/Rs671`, position 111803962), one of the SNPs we analyze (rs3742004) is within 5 kb of this SNP.

COMT: The SNP rs4680 has long been associated with schizophrenia and substance abuse, including alcoholism. A case-control study (Voisey et al., 2011) associated rs4680 and rs165774 with alcohol dependence through a SNP-wise chi-squared test, and had these two SNPs in high LD in their study population. Compared to this, in our simultaneous model of all COMT polymorphisms, both SNPs have e-values above threshold. A previous case-control GWAS on unrelated individuals with European and African ancestry that also focused on analysis of well-studied genes also did not find any association between rs4680 and alcohol dependence (Olfson and Bierut, 2012).

SLC6A3: Our analysis does not detect rs27072, which has been associated with alcohol withdrawal symptoms (`https://www.snpedia.com/index.php/Rs27072`). However, all the SNPs we detect are very close to exons.

SLC6A4: The only SNP detected in this gene is within 10 kb of rs1042173 that has repeatedly been associated with alcohol consumption (`https://www.snpedia.com/index.php/Rs1042173`). Interestingly we do not detect two other SNPs that are located closer to this SNP than the one detected.

DRD2: Among the 3 detected SNPs, rs7937641 is within 3 kb of rs1076560, which has multiple references of association with alcoholism (`https://www.snpedia.com/index.php/Rs1076560`). Among the three DRD2 SNPs associated with alcoholism in a case-control study on an Eastern Indian study sample (Bhaskar et al., 2010), one of them (rs2734835) is also close to rs7937641, while the other two (rs1116313 and TaqID) are within 4 kb of another detected SNP in our analysis: rs10750025.

Finally, most e-values for the last 3 genes, i.e. SLC6A3, SLC6A4 and DRD2, are large: indicating weak SNP signals. We found this observation interesting, because variants of these genes have known interaction effects behind alcohol withdrawal-induced seizure (Karpyak et al., 2010) and bipolar disorder (Wang et al., 2014), as well as additive effect on the susceptibility to smoking addiction (Erblich et al., 2005).

6 Discussion and conclusion

In the above sections we have proposed a fast covariate selection method to detect SNP signals in multi-SNP mixed effect models. The speed advantage is achieved because of two reasons: calculation of only a single model for the full algorithm, and utilizing a parallelizable bootstrap technique that uses only Monte-Carlo samples and previously obtained model objects to get resampling estimates of the coefficient vector. Our method achieves this by using the recently proposed e-values framework and comparing sampling distributions of reduced model estimates with that of the full model through an evaluation map function.

To expand the above approach to a genome-wide scale, we need to incorporate strategies for dealing with the hierarchical structure of causal SNPs: there are a few causal genes behind the quantitative phenotype, and that is due to the effect of a small number of SNPs inside these genes. To apply the e-values method here, it is plausible to start with an initial screening step to eliminate evidently non-relevant genes. Methods like the grouped Sure Independent Screening (Li et al., 2012) and min-P test (Westfall and Young, 1993) can be useful here. Following this, in a multi-gene predictor set, there are several possible strategies to select important genes *and* important SNPs in them:

1. **Two-step e-values:** First construct multi-SNP models for each gene, trained on SNP data inside that gene and a common behavioral trait response and select SNPs in each model using e-values. Now train a model using selected SNPs from all genes, and run group selection procedure in this model using e-values. This means dropping *groups* of predictors from the full model and checking the reduced model e-values. The one-step e-values method described in this paper is applicable here, and setting the groups as the collection of SNPs corresponding to a gene should be able to achieve our objective.

2. **SNP-level e-values only:** First select important genes using an aggregation method of SNP-

trait associations (e.g. Lamparter et al. (2016)) and run e-value based SNP selection on the set of SNPs within these genes.

3. **Gene-level e-values only:** Train separate models for each gene (after initial screening), select SNPs within those models using a fast screening method (e.g. RFGLS) and run group-level or SNP-level e-value selection in that full set of SNPs.

We plan to study merits and demerits of these strategies and the computational issues associated with them in detail through synthetic studies as well as in the GWAF data from MCTFR. Finally, the present data depth based formulation requires the existence of an asymptotic distribution for the full model estimate. We plan to explore alternative formulation of evaluation maps under weaker conditions to bypass this, so as to be able to tackle potentially high-dimensional situations.

References

Aulchenko, Y. S., Koning, D. J. D., and Haley, C. (2007). Genome-wide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. *Nat. Genet.*, 177:577–585.

Bhaskar, L. V., Thangaraj, K., Non, L. V., et al. (2010). Population-based case-control study of drd2 gene polymorphisms and alcoholism. *J. Addict. Dis.*, 29:475–480.

Bogdan, M., Chakrabarti, A., Frommelet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, 39:1551–1579.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.*, 33:414–436.

Chen, W. M. and Abecasis, G. (2007). Family-based association tests for genome-wide association scans. *Am. J. Hum. Genet.*, 81:913–926.

Coombes, B. J. (2016). *Tests for detection of rare variants and gene-environment interaction in cohort and twin family studies.* PhD thesis, University of Minnesota.

Cui, W. Y., Seneviratne, C., Gu, J., and Li, M. D. (2012). Genetics of GABAergic signaling in nicotine and alcohol dependence. *Hum. Genet.*, 131:843–855. doi: 10.1007/s00439-011-1108-4.

De Neve, J.-E., Mikhaylov, S., Dawes, C. T., et al. (2013). Born to Lead? A Twin Design and Genetic Association Study of Leadership Role Occupancy. *Leadersh Q*, 24:45–60.

Erblich, J. A., Lerman, C., Self, D. W., et al. (2005). Effects of dopamine d2 receptor (drd2) and transporter (slc6a3) polymorphisms on smoking cue-induced cigarette craving among african-american smokers. *Mol. Psychiatry*, 10:407–414.

Frommelet, F., Ruhaltinger, F., Twaróg, P., and Bogdan, M. (2012). Modified versions of Bayesian Information Criterion for genome-wide association studies. *Comput. Stat. Data Anal.*, 56:1038–1051.

Hicks, B. M., Schalet, B. D., Malone, S., Iacono, W. G., and McGue, M. (2011). Psychometric and Genetic Architecture of Substance Use Disorder and Behavioral Disinhibition Measures for Gene Association Studies. *Behav Genet.*, 41:459–475. doi:10.1007/s10519-010-9417-2.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., and McGue, M. (1999). Behavioral disinhibition and the development of substance use disorders: Findings from the Minnesota Twin Family Study. *Dev. Psychopathol.*, 11:869–900.

Irons, D. E. (2012). *Characterizing specific genetic and environmental influences on alcohol use*. PhD thesis, University of Minnesota.

Karpyak, V. M., Biernacka, J. M., Weg, M. W., et al. (2010). Interaction of SLC6A4 and DRD2 polymorphisms is associated with a history of delirium tremens. *Addict. Biol.*, 15:23–34. doi:10.1111/j.1369-1600.2009.00183.x.

Ke, X. (2012). Presence of multiple independent effects in risk loci of common complex human diseases. *Am. J. Hum. Genet.*, 91:185–192.

Keyes, M. A., Malone, S. M., Elkins, I. J., Legrand, L., McGue, M., and Iacono, W. G. (2009). The Enrichment Study of the Minnesota Twin Family Study: Increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Res. Hum. Genet.*, 12:489–501.

Kohler, H. P., Behrman, J. R., and Schnittker, J. (2011). Social science methods for twins data: integrating causality, endowments, and heritability. *Biodemography Soc Biol.*, 57:88–141.

Lamparter, D., Marbach, D., Rueedi, R., et al. (2016). Fast and Rigorous Computation of Gene

and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput. Biol.*, 12:e1004714. doi:10.1371/journal.pcbi.1004714.

Li, R., Zhong, W., and Zhu, L. (2012). Feature Screening via Distance Correlation Learning. *J. Amer. Statist. Assoc.*, 107:1129–1139.

Li, X., Basu, S., Miller, M. B., Iacono, W. G., and McGue, M. (2011). A Rapid Generalized Least Squares Model for a Genome-Wide Quantitative Trait Association Analysis in Families. *Hum. Hered.*, 71:67–82. doi:10.1159/000324839.

Lind, P. A., Macgregor, S., Heath, A. C., and Madden, P. A. F. (2012). Association between *in vivo* alcohol metabolism and genetic variation in pathways that metabolize the carbon skeleton of ethanol and NADH reoxidation in the Alcohol Challenge Twin Study. *Alcohol Clin. Exp. Res.*, 36:2074–2085. doi:10.1111/j.1530-0277.2012.01829.x.

Macgregor, S., Lind, P. A., Bucholtz, K. K., et al. (2008). Associations of adh and aldh2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. *Hum. Mol. Genet.*, 18:580–593.

Majumdar, S. and Chatterjee, S. (2017). Generalized Model Discovery using Statistical Evaluation Maps. Final draft.

Manolio, T. A., Collins, F. S., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

McGue, M., Keyes, M., Sharma, A., Elkins, I. J., Legrand, L. N., Johnson, W., and Iacono, W. G. (2007). The environments of adopted and non-adopted youth: Evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). *Behav. Genet.*, 37:449–462.

McGue, M., Zhang, Y., Miller, M. B., et al. (2013). A Genome-Wide Association Study of Behavioral Disinhibition. *Behav Genet.*, 43. doi:10.1007/s10519-013-9606-x.

Miller, M. B., Basu, S., Cunningham, J., et al. (2012). The Minnesota Center for Twin and Family Research Genome-Wide Association Study. *Twin Res Hum Genet.*, 15:767–774. doi:10.1017/thg.2012.62.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and Complex Data Structures*, pages 17–34. Springer Berlin Heidelberg.

Olfson, E. and Bierut, L. J. (2012). Convergence of gwa and candidate gene studies for alcoholism. *Alcohol Clin. Exp. Res.*, 36:2086–2094.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, 43:D670–81. doi:10.1007/s10519-010-9417-2.

Tukey, J. (1975). Mathematics and picturing data. In James, R., editor, *Proceedings of the International Congress on Mathematics*, volume 2, pages 523–531.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five Years of GWAS Discovery. *Amer. J. Hum. Genet.*, 90:7–24. doi: 10.1016/j.ajhg.2011.11.029.

Voisey, J., Swagell, C. D., Hughes, I. P., et al. (2011). A novel SNP in COMT is associated with alcohol dependence but not opiate or nicotine dependence: a case control study. *Behav. Brain Funct.*, 7(51). doi: 10.1186/1744-9081-7-51.

Wang, T. Y., Lee, S. Y., Chen, S. L., et al. (2014). Gender-specific association of the SLC6A4 and DRD2 gene variants in bipolar disorder. *Int. J. Neuropsychopharmacol.*, 17:211–222. doi: 10.1017/S1461145713001296.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley; New York.

Yang, J., Ferreira, T., Morris, A. P., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, 44:369–375 S361S363.

Zhang, H., Shi, J., Liang, F., et al. (2014). A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies. *Eur. J. Hum. Genet.*, 22:696–701.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.*, 31:1460–1490.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482.