Submitted to the Annals of Applied Statistics

SIMULTANEOUS SELECTION OF MULTIPLE IMPORTANT SINGLE NUCLEOTIDE POLYMORPHISMS IN FAMILIAL GENOME WIDE ASSOCIATION STUDIES DATA

By Subhabrata Majumdar, Saonli Basu and Snigdhansu Chatterjee

We propose a resampling-based fast variable selection technique for selecting important Single Nucleotide Polymorphisms (SNP) in multi-marker mixed effect models used in twin studies. Due to computational complexity, current practice includes testing the effect of one SNP at a time, commonly termed as 'single SNP association analysis'. Joint modeling of genetic variants within a gene or pathway may have better power to detect the relevant genetic variants, hence we adapt our recently proposed framework of e-values to address this. In this paper, we propose a computationally efficient approach for single SNP detection in families while utilizing information on multiple SNPs simultaneously. We achieve this through improvements in two aspects. First, unlike other model selection techniques, our method only requires training a model with all possible predictors. Second, we utilize a fast and scalable bootstrap procedure that only requires Monte-Carlo sampling to obtain bootstrapped copies of the estimated vector of coefficients. Using this bootstrap sample, we obtain the e-value for each SNP, and select SNPs having e-values below a threshold. We illustrate through numerical studies that our method is more effective in detecting SNPs associated with a trait than either single-marker analysis using family data or model selection methods that ignore the familial dependency structure. We also use the e-values to perform gene-level analysis in nuclear families and detect several SNPs that have been implicated to be associated with alcohol consumption.

1. Introduction. Genome Wide Association Studies (GWAS) have identified a large number of genetic variants associated with complex diseases [4, 41]. The advent of economical high-throughput genotyping technology enables researchers to scan the genome with millions of SNPs, and improvements in computational efficiency in analysis techniques has facilitated parsing through this huge amount of data to detect significant associations [36]. Detecting small effects of individual SNPs requires large sample size [27], which is a major challenge of these studies. For quantitative behavioral

Keywords and phrases: Family data, Twin studies, ACE model, Model selection, Resampling, Generalized bootstrap

traits such as alcohol consumption, drug abuse, anorexia and depression, variation in genetic effects due to environmental heterogeneity brings in additional noise, further amplifying the issue. This is one of the motivations of performing GWAS on families instead of unrelated individuals, through which the environmental variation can be reduced [2, 30, 37]. However, association analysis using families can be computationally very challenging and thus single SNP association analysis is the standard tool for detecting SNPs. The Minnesota Twin Family Study [30] with genome-wide data on identical twins, non-identical twins, biological offsprings, adoptees serve as the motivation for our methodology development in this paper.

Single-SNP analysis tests for association between the trait and a single SNP at a time. The SNPs with p-values lower than a particular threshold are considered to be associated with the trait. The GRAMMAR method of [1] and the association test of [7] are examples of such techniques applied to familial data. While they can efficiently analyze GWAS data, they assume that phenotypic similarity within families is entirely due to their genetic similarity and ignore the effect of shared environment. As a result, they tend to lose power when analyzing data where shared environmental effects explain a substantial proportion of the total phenotypic variation (see [29] and [11] for example). In contrast, the RFGLS method proposed by [24] takes into account genetic and environmental sources of familial similarity and still provides fast inference through a rapid approximation of the single-SNP mixed effect model.

Single-SNP methods are prone to be less effective in detecting SNPs with weak signals [27]. This includes instances where multiple SNPs are jointly associated with the phenotype [19, 34, 42]. Several methods of multi-SNP analysis have been proposed as alternatives. The kernel based association tests [6, 16, 33, 34] are prominent among such techniques. However, all of them test for whether a *group* of SNPs is associated with the phenotype being analyzed, and do not generally prioritize within the group and detect the individual SNPs primarily associated with the trait.

One way to solve this problem is to perform model selection. The methods of [12] and [43] take this approach, and perform SNP selection from a multi-SNP model on GWAS data from *unrelated individuals*. However, they rely on fitting models corresponding to multiple predictor sets, hence are computationally very intensive to implement in a linear mixed-effect framework for modeling familial data.

In this paper we propose a fast and scalable model selection technique thats fits a single model to a family data, and aims to identify important genetic variants with weak signals through joint modelling of multiple variants. We consider only main effects of the variants, but this can be extended to include higher-order interactions. We achieve this by extending the recently proposed framework of e-values [26]. For any estimation method that provides consistent estimates (at a certain rate relative to the sample size) 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 in a regression-like setup. A variable selection algorithm using the e-values has the following simple and generic steps:

1. Fit the full model, i.e. where all predictor effects are being estimated from the data, and use resampling to estimate its e-value;
2. Set an element of the full model coefficient estimate to 0 and get an e-value for that predictor using resampling distribution of previously estimated parameters- repeat this for all predictors;
3. Select predictors that have e-values below a pre-determined threshold.

The above algorithm offers multiple important benefits in the SNP selection scenario. Unlike other model selection methods, only the full model needs to be computed here. It thus offers the user more flexibility in utilizing a suitable method of estimation for the full model. Our method allows for fitting multi-SNP models, thereby accommodating cases of modelling multiple correlated SNPs or closely located multiple causal SNPs simultaneously. Finally, we use the Generalized Bootstrap [5] 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 as weights, and reusing model objects obtained from 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 GWAS Family dataset we use in our case study, as well as introduces the statistical framework we 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, and present details of the bootstrap procedure. We illustrate the performance of this method on synthetic datasets in Section 4. In Section 5 we analyze our GWAS dataset using the e-values technique to identify novel SNPs from multiple genes that have been reported to influence alcohol consumption in individuals. Finally in Section 6 we outline directions of future research. We include the proofs of all new results stated, specifically, theorems 3.2 and 3.3, 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)[24, 29, 30] consists of samples from three longitudinal studies conducted by the MCTFR: (1) the Minnesota Twin Family Study (MTFS: [15]) that covers twins and their parents, (2) the Sibling Interaction and Behavior Study (SIBS: [28]) that includes adopted and biological sibling pairs and their parents, and (3) the enrichment study (ES: [20]) 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 [30] the final sample consisted of 7605 Caucasian individuals clustered in 2151 nuclear families. This consisted of 1109 families where the children are identical twins, 577 families with non-identical twins, 210 familes with adopted children, 162 families with non-twin biological siblings, and 93 families wher one child is adopted while the other is the biological child of the parents.

DNA samples collected from the subjects were analyzed using Illumina's Human660W-Quad Array, and after standard quality control steps [30], 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. Five quantitative phenotypes measuring substance use disorders 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 are derived from questionnaires using a hierarchical approach based on factor analysis [13].

A detailed description of the data is available in [30]. Several studies reported SNPs associated with phenotypes collected in MCTFR study [8, 24, 29]. [24] used RFGLS to detect association between height and genetic variants through single-SNP analysis, while [29] used the same method to study SNPs influencing the development of all five indicators of behavioral disinhibition mentioned above. [17] 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 [8] used a bootstrap-based combination test and a sequential score test to evaluate gene-environment interactions for alcohol consumption.

2.2. *Statistical model.* We 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 [21]. While the-state-of-the-art focuses on detection of a *single variant at a time*, we will incorporate *all* SNPs genotyped within a gene (or group of genes in some cases) as set of fixed effects in a *single model*.

Our model fitting process is invariant to pedigree sizes. In the present context we assume nuclear pedigrees, as previously implemented by [7, 24, 29]. Suppose there are m families in total, with the i^{th} pedigree containing n_i individuals. Denote by $y_i = (y_{i1}, \ldots, y_{in_i})^T$ 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.

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i \beta_g + \mathbf{C}_i \beta_c + \epsilon_i \tag{2.1}$$

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

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

The first component above is a within-family random effect term to account for polygenic effects. The matrix $\mathbf{\Phi}_i$ is the relationship matrix within the i^{th} pedigree. Its $(s,t)^{\text{th}}$ 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 [21]. The second variance component accounts for shared environmental effect within each family, while the third term 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 $\mathbf{\Phi}_i$ matrices for different types of families:

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

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

3. Variable selection using e-values. We present the details of our methodology in this section. Sections 3.1 and 3.2 summarize the existing method of e-values that performs best subset variable selection in a wide range of statistical models [26]. We build on this framework in Section 3.3, where we present new results for better detection of weak SNP signals. Section 3.4 elaborates on the bootstrap implementation of this methodology using the model in (2.1).

3.1. *Models and evaluation maps.* In a general modelling situation where one needs to estimate a set of parameters $\theta \in \mathbb{R}^p$ from a triangular array of samples $\mathcal{B}_n = \{B_{n1}, \ldots, B_{nk_n}\}$ at stage n, any hypothesis or statistical model corresponds to a subset of the full parameter space. Here we consider the model spaces $\mathbf{\Theta}_{mn} \subseteq \mathbb{R}^p$ in which some elements of the parameter vector have fixed values, while others are estimated from the data. Formally, a generic parameter vector $\theta_{mn} \in \mathbf{\Theta}_{mn}$ consists 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 $\mathcal{S}_n \subseteq \{1, \ldots, p\}$. Thus the estimable index set $\mathcal{S}_n$ and fixed elements $c_n = (c_{nj} : j \notin \mathcal{S}_n)$ fully specify any model in this setup.

At this point, the original framework in [26] introduces a few concepts to provide a detailed treatment considering a general scenario. For our specific problem, i.e. variable selection, we only require vastly simplified versions of them. We consider sample size $k_n = n$ for all n, and assume constant sequences of candidate models: $\mathcal{M}_n = \mathcal{M}$ for all n. We also drop the subscripts in $\mathcal{S}_n$ and c_n. Thus the 'full model', i.e. the model with all covariates is denoted by $\mathcal{M}_* = (\{1, \ldots, p\}, \emptyset)$.

Given data of size n, we obtain the full model estimates as minimizers of an estimating equation:

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

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

(P1) The population version of (3.1) has a unique minimizer θ_0, i.e.

$$\theta_0 = \arg\min_{\theta} \mathbb{E} \sum_{i=1}^{n} \Psi_i(\theta, B_i)$$

(P2) There exist a sequence of positive numbers $a_n \uparrow \infty$ and a p-dimensional probability distribution $\mathbb{T}_0$ such that $a_n(\hat{\theta} - \theta_0) \rightsquigarrow \mathbb{T}_0$.

We designate θ_0 as the *true parameter vector*, some elements of which are potentially set to 0. We can now classify any candidate model $\mathcal{M}$ into one of the two classes: the ones that satisfy $\theta_0 \in \mathbf{\Theta}_m$, and the ones that do not. We denote these two types of models by *adequate* and *inadequate models*, respectively. Given the data and unknown θ_0, we want to determine if a candidate model is adequate or inadequate.

For this we need coefficient estimates $\hat{\theta}_m$ corresponding to a model. We do so by just replacing elements of $\hat{\theta}$ not in $\mathcal{S}$ by corresponding elements of c. This means that for the j^{th} element, $j = 1, \ldots, p$, we have

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

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

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

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

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

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

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

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

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

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

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

Data depths [35, 44, 45] also constitute a broad class of point-to-distribution proximity functions that satisfy the above regularity conditions for evaluation maps. Indeed, [26] used halfspace depth [35] as evaluation function to perform model selection. However, the conditions (E1) and (E4) are weaker than those imposed on a traditional depth function [45]. Conditions (E2) and (E3) are not required of depth functions in general, but they arise implicitly in several implementations of data depth [31]. 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 we use the non-depth evaluation functions E_1 and E_2 as in (3.2) above.

3.2. *Model selection using e-values.* Depending on the choice of the data sequence $\mathcal{B}_n$, $E(\hat{\theta}_m, [\hat{\theta}])$ can take different values. For any candidate model $\mathcal{M}$, we denote the distribution of the corresponding random evaluation map by $\mathbb{E}_{mn}$. For simplicity we drop the n in its subscript, i.e. $\mathbb{E}_{mn} \equiv \mathbb{E}_m$. These distributions are informative of how model estimates behave, and we use them as a tool to distinguish between inadequate and adequate models. Given a single set of samples, we use resampling schemes that satisfy standard regularity conditions [26] to get consistent approximations of $\mathbb{E}_m$.

We now define a quantity called the ***e*-value** 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, [26] 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 θ_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 $\hat{\theta}$ 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 $\hat{\theta}$.

As $n \to \infty$, the above algorithm provides consistent model selection, i.e. $\mathbb{P}(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \to 1$, 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 $n \to \infty$ [26].

3.3. *Quantiles of* $\mathbb{E}_m$ *as e-values.* When 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 in a sample setting. 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 3.1 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 use tail quantiles as e-values.

We denote the q^{th} population quantile of $\mathbb{E}_m$ by $c_q(\mathbb{E}_m)$. Then we have equivalent results to Theorem 3.1 as $n \to \infty$:

THEOREM 3.2. *Given that the estimator* $\hat{\theta}$ *satisfies conditions (P1) and*

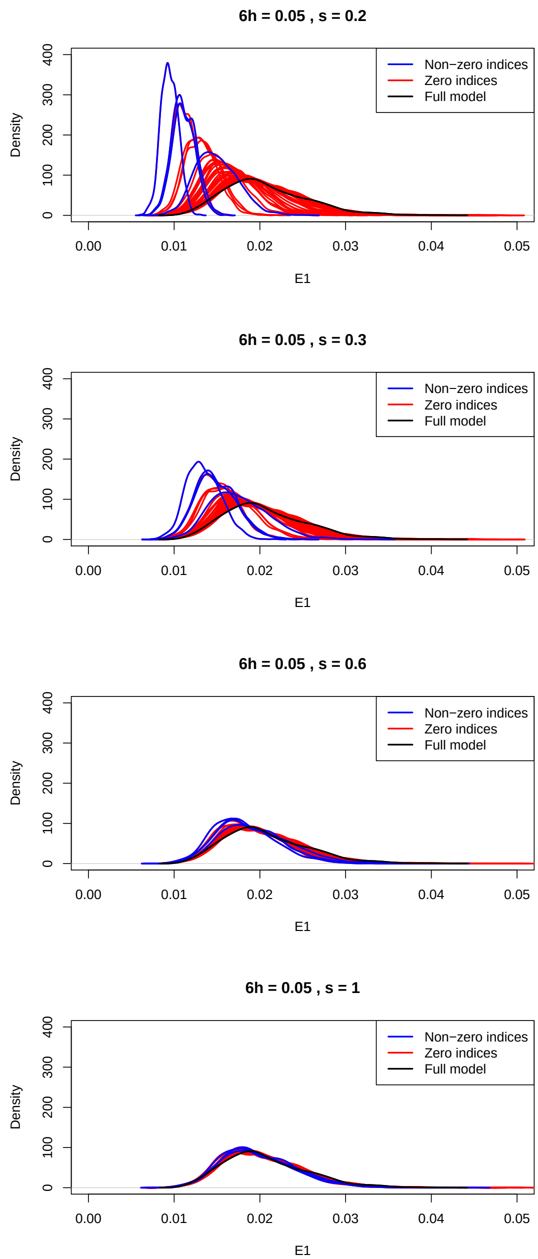


Fig 3.1: Density plots of bootstrap approximations for $\mathbb{E}_*$ and $\mathbb{E}_{-j}$ for all j in simulation setup, with $s = 0.2, 0.3, 0.6, 1$

(P2), and the evaluation map satisfies conditions (E1)-(E4), we have

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

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

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

When the q^{th} quantile is taken as the e-value instead of the mean, 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 [5] to obtain approximations of the sampling distributions $\mathbb{E}_{-j}$ and $\mathbb{E}_*$. It calculates bootstrap equivalents of the parameter estimate $\hat{\theta}$ by minimizing a version of the estimating equation in (3.1) with random weights:

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

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:

$$\mathbb{E}\mathbb{W}_1 = 1; \quad \mathbb{V}\mathbb{W}_1 = \tau_n^2 \uparrow \infty; \quad \tau_n^2 = o(a_n^2) \tag{3.7}$$

$$\mathbb{E}W_1 W_2 = O(n^{-1}); \quad \mathbb{E}W_1^2 W_2^2 \to 1; \quad \mathbb{E}W_1^4 < \infty \tag{3.8}$$

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 $\Psi(.)$ [5, 26] and conditional on the data, $(a_n/\tau_n)(\hat{\theta}_w - \hat{\theta})$ converges to the same asymptotic distribution as $a_n(\hat{\theta} - \theta_0)$, 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 $[\hat{\theta}_r]$ and $[\hat{\theta}_{r_1}]$;

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

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

Conditions (3.7) and (3.8) on the resampling weights ensure bootstrap-consistent approximation of the evaluation map quantiles:

Theorem 3.3. *Given the estimator $\hat{\theta}$ and evaluation map E in Theorem 3.2, and a generalized bootstrap scheme satisfying (3.7) and (3.8), we get*

$$|c_q(\hat{\mathbb{E}}_m) - c_q(\hat{\mathbb{E}}_*)| \stackrel{P_n}{\to} o_P(1) \text{ when } \mathcal{M} \text{ is adequate} \tag{3.9}$$

$$c_q(\hat{\mathbb{E}}_m) \stackrel{P_n}{\to} o_P(1) \text{ when } \mathcal{M} \text{ is inadequate} \tag{3.10}$$

where P_n is probability conditional on the data.

Generalized bootstrap 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 θ_0 and some other conditions in [5], there is an approximate representation of $\hat{\theta}_w$:

$$\hat{\theta}_w = \hat{\theta} - \frac{\tau_n}{a_n} \left[\sum_{i=1}^n W_i \Psi_i''(\hat{\theta}, B_i) \right]^{-1} \sum_{i=1}^n W_i \Psi_i'(\hat{\theta}, B_i) + \mathbf{R}_{wn} \tag{3.11}$$

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

Given the full model estimate $\hat{\theta}$, and the score vectors $\Psi_i'(\hat{\theta}, B_i)$ and hessian matrices $\Psi_i''(\hat{\theta}, B_i)$, (3.11) allows us to obtain multiple copies of $\hat{\theta}_w$ 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 (3.11) to the LMM in (2.1). We first obtain the maximum likelihood estimates $\hat{\beta}_g, \hat{\sigma}_a^2, \hat{\sigma}_c^2, \hat{\sigma}_e^2$ through fitting the LMM. 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:

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

We take m random draws from Gamma$(1,1)-1$, say $\{w_{r1},\dots,w_{rm}\}$, as resampling weights in (3.11), using the same weight for all members of a pedigree. Consequently, the bootstrapped coefficient estimate $\hat{\beta}_{rg}$ has the following representation:

$$\hat{\beta}_{rg} \simeq \hat{\beta}_g + \frac{\tau_n}{\sqrt{n}}(\mathbf{G}^T\hat{\mathbf{V}}^{-1}\mathbf{G})^{-1}\mathbf{W}_r\mathbf{G}^T\hat{\mathbf{V}}^{-1}(y-\mathbf{G}\hat{\beta}_g) \tag{3.12}$$

with $\mathbf{G} = (\mathbf{G}_1^T,\dots,\mathbf{G}_m^T)^T$ and $\mathbf{W}_r = \mathrm{diag}(w_{r1}\mathbf{I}_4,\dots,w_{rm}\mathbf{I}_4)$. 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 $\{\hat{\beta}_{1g},\dots,\hat{\beta}_{Rg}\}$.

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 generate them in correlated blocks of 6, 4 ,6, 4 and 30 to simulate correlation among SNPs in the genome. 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 generally each associated SNP explains only a small proportion of the overall variability of the trait. 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. The value of the non-zero coefficient in k-th block: $k = 1,...,4$, say β_k is calculated using the formula:

$$\beta_k = \sqrt{\frac{h}{100(\sigma_a^2+\sigma_c^2+\sigma_e^2).2\mathrm{MAF}_k(1-\mathrm{MAF}_k)}} \tag{4.1}$$

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 $y_1,\dots,y_{250}$ 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 [12] and provides approximate False Discovery Rate (FDR) control at level 0.05 [3].

(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: [24]), 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 $\mathbb{E}_*$ 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.

Since the above procedure depends on the bootstrap standard deviation parameter s, 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 $\hat{\mathcal{S}}_t(s)$ that minimizes fixed effect prediction error (PE) on an independently generated test dataset $\{(y_{test,i}, \mathbf{G}_{test,i}), i = 1, \ldots, 250\}$ from the same setup above:

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

$$\hat{\mathcal{S}}_t = \arg\min_s \text{PE}_t(s)$$

We use the following metrics to evaluate each method we implement: (1) True Positive (TP), which is the proportion of causal SNPs detected; (2) True Negative (TN), which is the proportion of non-causal SNPs undetected; (3) Relaxed True Positive (RTP), which is the: proportion of detecting any

SNP in each of the 4 blocks with causal SNPs, i.e. for the selected index set by some method m, say $\hat{\mathcal{S}}_m$,

$$\mathrm{RTP}(\hat{\mathcal{S}}_m) = \frac{1}{4}\sum_{i=1}^{4} \mathbb{I}(\text{Block } i \cap \hat{\mathcal{S}}_m \neq \emptyset)$$

and finally (4) Relaxed True Negative (RTN), which is the 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 [12]. 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 4.1. For all heritability values, applying mBIC2 on linear models performs poorly compared to applying RFGLS and then correcting for multiple testing. This is expected because the linear model ignores within-family error components.

Our method works better than the two competing 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 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, indicating that the right tails of evaluation map distributions are more useful for this purpose. Finally for $h = 0$, we report only TN and RTN 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 RTN performance as RFGLS+BH for all choices of t.

RTP 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 RTN 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 RTP) while having the same TN (or RTN) rates.

5. Analysis of the MCTFR data. We now apply the above methods on SNPs from the MCTFR dataset. We assume a nuclear pedigree structure, and for simplicity only analyze pedigrees with MZ and DZ twins. After setting aside samples with missing response variables, we end up with 1019 such 4-member families. We look at the effect of genetic factors behind the

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 4.1: (Top) Average True Positive (TP)/ True Negative (TN) rates for mBIC2, RFGLS+BH and the e-values method with E_1 and E_2 as evaluation maps and different values of t over 1000 replications, and (Bottom) Average Relaxed True Positive (RTP) and Relaxed True Negative (RTN) rates

Gene	Total no. of SNPs	No. of SNPs detected by		
		e-value	RFGLS+BH	mBIC2
GABRA2	11	5	0	0
ADH	44	3	1	0
OPRM1	47	25	1	0
CYP2E1	9	5	0	0
ALDH2	6	5	0	1
COMT	15	14	0	0
SLC6A3	18	4	0	0
SLC6A4	5	0	0	0
DRD2	17	0	0	1

TABLE 5.1
Table of analyzed genes and number of detected SNPs in them by the three methods

response variable pertaining to the amount of alcohol consumption, which is highly heritable in this dataset according to previous studies [29]. We analyze SNPs inside some of the most-studied genes with respect to alcohol abuse: GABRA2, ADH1A, ADH1B, ADH1C, ADH4-ADH7, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT [9] through separate gene-level models. The ADH genes did not contain many SNPs individually, so we club all seven of them together. We include sex, birth year, age and generation (parent or offspring) of individuals as covariates to control for their potential effect.

For model selection we use E_2 as the evaluation function because of its slighty better performance in the simulations. For each gene-level model, 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 set of SNPs that minimizes fixed effect prediction error on the data from the other 25% of families over this grid of (s, t).

As seen in Table 5.1, our e-value based technique detects a much higher number of SNPs than the two competing methods. Our method selects all but one SNP in the genes ALDH2 and COMT. These are small genes of size 50kb and 30kb, respectively, thus SNPs within them have more chance of being in high Linkage Disequilibrium (LD). On the other hand, it does not select any SNPs in SLC6A4 and DRD2. Variants of these genes are known to interact with each other and are jointly associated with multiple behavioral disorders [18, 39].

A number of SNPs we detect (or SNPs situated close to them) have known associations with alcohol-related behavioral disorders. We summarize this in Table 5.2. Prominent among them are rs1808851 and rs279856 in the GABRA2 gene, which are at perfect LD with rs279858 in the larger, 7188-individual version of our twin studies dataset [17]. This SNP is the marker

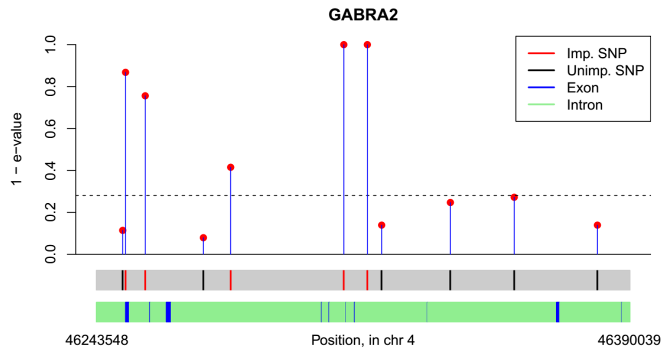
Gene	Detected SNPs with known associations	Reference for associated SNP
GABRA2	rs1808851, rs279856: close to rs279858	[10]
ADH genes	rs17027523: 20kb upstream of rs1229984	Multiple studies (https://www.snpedia.com/index.php/Rs1229984)
OPRM1	rs12662873: 1 kb upstream of rs1799971	Multiple studies (https://www.snpedia.com/index.php/Rs1799971)
CYP2E1	rs9419624: 600b downstream of rs4646976; rs9419702: 10kb upstream of rs4838767	[25]
ALDH2	rs16941437: 10kb upstream of rs671	Multiple studies (https://www.snpedia.com/index.php/Rs671)
COMT	rs4680, rs165774	[38]
SLC6A3	rs464049	[14]

TABLE 5.2
Table of detected SNPs with known references

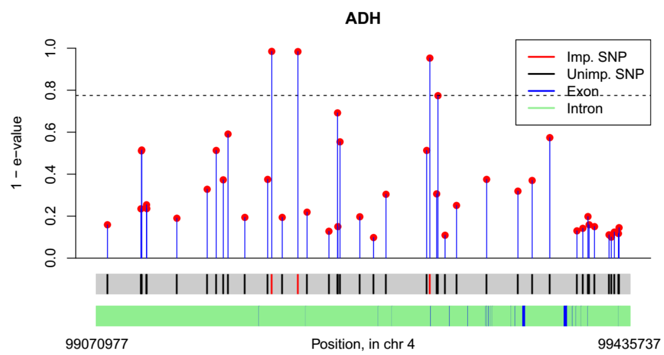
in GABRA2 that is most frequently associated in the literature with alcohol abuse [10], but was not genotyped in our sample. A single SNP RFGLS analysis of the same twin studies data that used Bonferroni correction on marginal p-values missed the SNPs we detect [17]: highlighting the advantage of our approach. We give a gene-wise discussion of associated SNPs, as well as information on all SNPs, in the supplementary material.

We plot the 90^{th} quantile e-value estimates in Figures 5.1, 5.2 and 5.3. We obtained gene locations, as well as the locations of coding regions of genes, i.e. exons, inside 6 of these 9 genes from annotation data extracted from the UCSC Genome Browser database [32]. Exon locations were not available for OPRM1, CYP2E1 and DRD2. In general, SNPs tend to get selected in groups with neighboring SNPs, which suggests high LD. Also most of the selected SNPs either overlap or in close proximity to the exons, which underline their functional relevance.

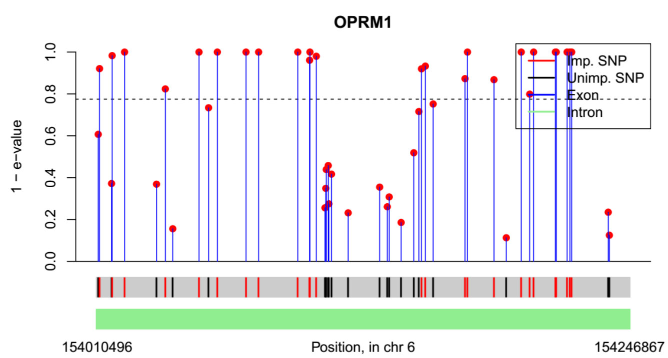
6. Discussion and conclusion. To expand the above approach to a genome-wide scale, we need to incorporate strategies for dealing with the hierarchical structure of causal SNPs: there are a few causal genes behind a quantitative phenotype, which can be further attributed to a proportion of SNPs inside each gene. 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 [23] and min-P test



(a)

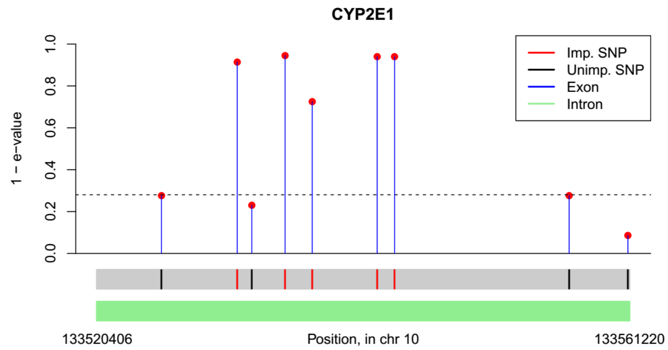


(b)

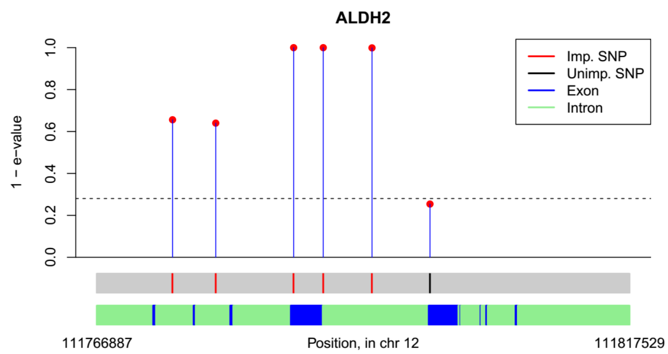


(c)

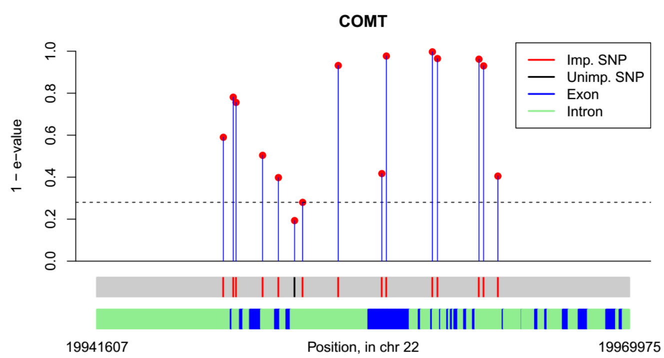
Fig 5.1: Plot of e-values for genes analyzed: (a) GABRA2, (b) ADH1 to ADH7, (c) OPRM1. For ease of visualization, $1 - e$-values are plotted in the y-axis.



(d)

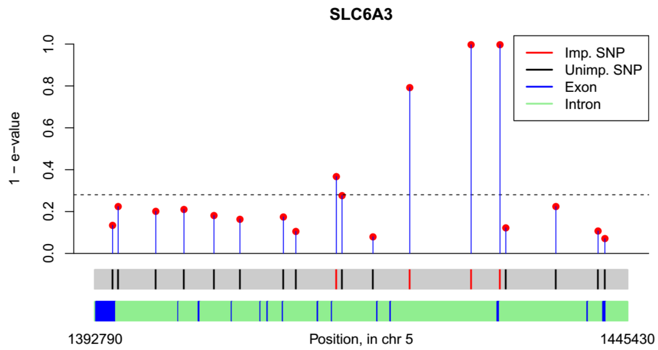


(e)

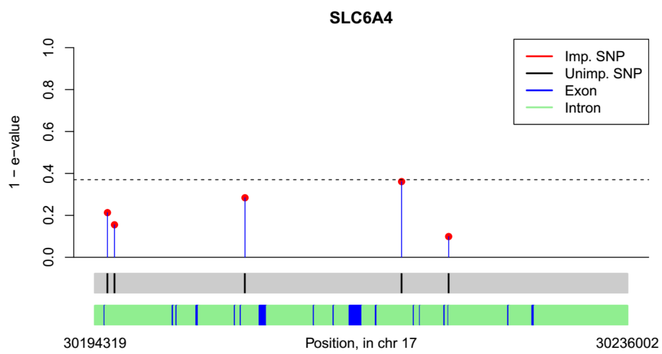


(f)

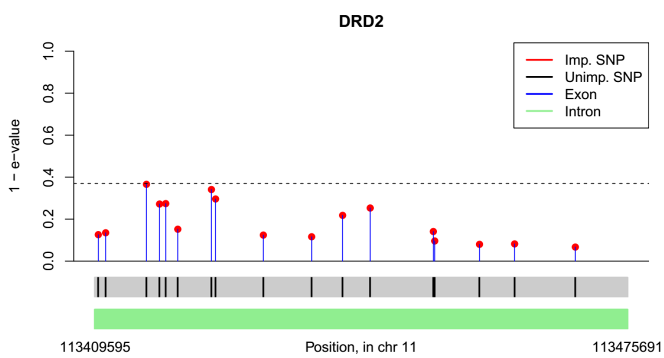
Fig 5.2: Plot of e-values for genes analyzed: (d) CYP2E1, (e) ALDH2, (f) COMT



(g)



(h)



(i)

Fig 5.3: Plot of e-values for genes analyzed: (g) SLC6A3, (h) SLC6A4, (i) DRD2

[40] 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. Firstly, one can use a two-stage e-value based procedure. The first stage is same as the method described in this paper, i.e. selecting important SNPs from each gene using multi-SNP models trained on SNPs in that gene. In the second stage, a model will be trained using the aggregated set of SNPs obtained in the first step, and a group selection procedure will be run on this model using e-values. This means dropping *groups* of predictors (instead of single predictors) from the full model, checking the reduced model e-values, and selecting a SNP group only if dropping it causes the e-value to go below a certain cutoff. Secondly, one can start by selecting important genes using an aggregation method of SNP-trait associations (e.g. [22]) and then run the e-value based SNP selection on the set of SNPs within these genes. Thirdly, one can also take the aggregated set of SNPs obtained from running the e-values procedure on gene-level models, then use a fast screening method (e.g. RFGLS) to select a subset of those 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 GWAS data from MCTFR. Finally, the current evaluation map 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, thus being able to tackle high-dimensional ($n < p$) situations.

References.

[1] 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.

[2] Benyamin, B., Visscher, P. M. and McRae, A. F. (2009). Family-based genome-wide association studies. *Pharmacogenomics* **10** 181–190.

[3] 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.

[4] Chang, C. Q., Yesupriya, A., Lowell, J. L., Pimentel, C. B. et al. (2014). A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur, J. Hum. Genet.* **22** 402–408.

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

[6] Chen, H., Meigs, J. B. and Dupuis, J. (2013). Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genet. Epidemiol.* **37** 196–204.

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

[8] Coombes, B., Basu, S. and McGue, M. (2017). A combination test for detection of gene-environment interaction in cohort studies. *Genet. Epidemiol.* **41** 396–412.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Hicks, B. M., Schalet, B. D., Malone, S. M., 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.

[14] Huang, C. C., Kuo, S. C., Weh, Y. W. et al. (2017). The SLC6A3 gene possibly affects susceptibility to late-onset alcohol dependence but not specific personality traits in a Han Chinese population. *PLoS ONE* **12** e0171170.

[15] 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.

[16] Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet.* **21** 1158–1162.

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

[18] 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.

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

[20] Keyes, M. A., Malone, S. M., Elkins, I. J., Legrand, L. N., 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] Majumdar, S. and Chatterjee, S. (2017). Fast and General Model Selection using Data Depth and Resampling. https://arxiv.org/abs/1706.02429.

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

[28] 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.

[29] McGue, M., Zhang, Y., Miller, M. B. et al. (2013). A Genome-Wide Association Study of Behavioral Disinhibition. *Behav Genet.* **43**.

[30] 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.

[31] Mosler, K. (2013). Depth Statistics. In *Robustness and Complex Data Structures* (C. Becker, R. Fried and S. Kuhnt, eds.) 17-34. Springer Berlin Heidelberg.

[32] Rosenbloom, K. R., Armstrong, J., Barber, G. P. et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43** D670–81.

[33] Schaid, D. J., McDonnell, S. K., Sinnwell, J. P. and Thibodeau, S. N. (2013). Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. *Genet. Epidemiol.* **37** 409–418.

[34] Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L., Peyser, P. A. and Lin, X. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* **36** 797–810.

[35] Tukey, J. W. (1975). Mathematics and picturing data. In *Proceedings of the International Congress on Mathematics* (R. D. James, ed.) **2** 523-531.

[36] 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.

[37] Visscher, P. M., Wray, N. R., Zhang, Q. et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101** 5–22.

[38] 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**.

[39] 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.

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

[41] Wheeler, E. and Barroso, I. (2011). Genome-wide association studies and type 2 diabetes. *Brief. Funct. Genet.* **10** 52–60.

[42] 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.

[43] 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.

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

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