Prioritizing Individual Genetic Variants After Kernel Machine Testing Using Variable Selection

Qianchuan He[1*], Tianxi Cai[2], Yang Liu[1], Ni Zhao[1], Quaker E. Harmon[3], Lynn M. Almli[4], Elisabeth B. Binder[5], Stephanie M. Engel[6], Kerry J. Ressler[7], Karen N. Conneely[8], Xihong Lin[2], and Michael C. Wu[1*]

1. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.

2. Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A

3. Epidemiology Branch, NIEHS, Research Triangle Park, NC 27709, U.S.A

4. Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta GA 30322, U.S.A.

5. Max-Planck Institute of Psychiatry, Munich, Germany 80804

6. Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599, U.S.A.

7. Division of Depression & Anxiety Disorders, McLean Hospital, Belmont, MA 02478, U.S.A.

8. Department of Human Genetics, Emory University School of Medicine, Atlanta GA 30322, U.S.A.

Address for Correspondence: Qianchuan He, Ph.D and Michael C. Wu, Ph.D,

Public Health Sciences Division,

Fred Hutchinson Cancer Research Center,

Seattle, Washington 98109

Phone: (206) 667-7068, (206) 667-6603

Email: qhe@fredhutch.org, mcwu@fredhutch.org

ABSTRACT

Kernel machine learning methods, such as the SNP-set kernel association test (SKAT), have been widely used to test associations between traits and genetic polymorphisms. In contrast to traditional single-SNP analysis methods, these methods are designed to examine the joint effect of a set of related SNPs (such as a group of SNPs within a gene or a pathway) and are able to identify sets of SNPs that are associated with the trait of interest. However, as with many multi-SNP testing approaches, kernel machine testing can draw conclusion only at the SNP-set level, and do not directly inform on which one(s) of the identified SNP set is actually driving the associations. A recently proposed procedure, KerNel Iterative Feature Extraction (KNIFE), provides a general framework for incorporating variable selection into kernel machine methods. In this article, we focus on quantitative traits and relatively common SNPs, and adapt the KNIFE procedure to genetic association studies and propose an approach to identify driver SNPs after the application of SKAT to gene set analysis. Our approach accommodates several kernels that are widely used in SNP analysis, such as the linear kernel and the Identity By State (IBS) kernel. The proposed approach provides practically useful utilities to prioritize SNPs, and fills the gap between SNP set analysis and biological functional studies. Both simulation studies and real data application are used to demonstrate the proposed approach.

KEYWORDS: Genetic association studies; Kernel machine methods; KNIFE; Set-based; Variable selection.

INTRODUCTION

Gene, region, and pathway-based analyses have emerged as powerful strategies for analyzing genetic association studies (Wang et al., 2007; Yan et al., 2015). Under these strategies (collectively called set-based analysis), multiple, related genetic variants are grouped together into a set of variants (called a SNP-set) and then jointly tested for association with a complex trait or disease of interest. Set-based analysis can often offer improved power over standard analysis of genetic association studies which focuses on assessing the effect of each individual SNP, one-by-one. In particular, set-based analysis can improve power by reducing multiple testing burden, by enabling capture of multi-SNP effects, by harnessing linkage disequilibrium (LD) between SNPs, and even by possibly capturing epistatic or nonlinear effects (Wu et al., 2010).

Kernel machine testing approaches, such as the SNP-set or Sequence Kernel Association Test (SKAT) (Wu et al., 2011), are a particular class of approaches for conducting set-based analysis of both common and rare variants. The kernel machine testing framework operates by modeling the effect of a SNP-set on the outcome through a generally specified, possibly non-parametric function, which is defined based on a kernel function. Testing then proceeds by exploiting the connection between kernel machines and mixed models which enables utilization of a variance component score test (Lin, 1997). Operationally, the kernel function is a measure of similarity between two subjects based on the SNPs in the SNP-set, and the kernel machine test operates by comparing pair-wise similarity between subjects based on the SNP-set to pair-wise similarity between subjects based on the trait. If similarity in SNP-set profiles corresponds to similarity in the trait, then this suggests association between the SNPs and the trait. This class of approaches have been successfully applied to identify associations between genetic variants and a wide range of complex traits and diseases, such as fasting insulin (Cornes et al., 2013), hematological traits (Auer et al., 2014), and others. The approach has been extended to accommodate a wide range of types of traits and study designs (Lin et al., 2011; Ionita-Laza

et al., 2013).

Despite the popularity and successful application of kernel methods across a wide range of settings, a key limitation of the approach lies in the interpretation of significant results. More specifically, as a global test, kernel machine testing only provides an overall p-value for the association between a group of variants and the trait. Thus, significance indicates that one or more variants are associated with the outcome, but there is no indication of which variant(s) are driving the apparent association. Fine mapping and identification of individual SNPs that are driving associations is of prime importance in order to hypothesize mechanisms by which inherited variability influences complex traits (Edwards et al., 2013). Practically, for functional studies, experimental investigations require focusing on a modest number of candidate SNPs. However, despite the importance, it is currently unclear as to how to identify individual genetic variants driving significant associations for a number of reasons. First, by using a score test, the kernel machine test operates by estimating parameters under the null (which does not contain any genetic effects). Second, even if one does choose to do estimation under the kernel machine framework, as a non-parametric approach, the kernel machine framework only estimates the overall function of all of the SNPs. In other words, one can estimate the cumulative effect of all of the SNPs in the SNP set, but does not provide any information on the effect of any particular variant.

To overcome this difficulty and to facilitate the ongoing research efforts on functional studies of SNPs, we propose to apply variable selection, post-hoc, to identify individual variants that are driving the observed genetic associations when kernel machine methods are applied. This is closely related to fine mapping. In particular, for a SNP-set that has been found to be associated with a quantitative trait of interest, we propose to subsequently adapt the KerNel Iterative Feature Extraction (KNIFE) (Allen, 2013) method to select the individual SNPs that are driving the association. KNIFE is a recently developed approach that conducts variable selection within the kernel machine framework by imposing weights on different features while constructing the kernel. By shrinking some of these weights to be exactly zero, the corre-

sponding features are no longer used to estimate similarity and are therefore dropped from the model, enabling variable selection. We tailor the KNIFE method to conduct selection of genetic variables by applying KNIFE within the context of genetically relevant kernels and also making algorithmic adjustments to allow for covariate adjustment and reduce computational burden. Specifically, we (1) consider the linear, identity-by-state (IBS) and quadratic kernels which are powerful kernels for genetic association testing, (2) incorporate individual SNP specific weights, and (3) finally, design a two-step procedure for implementing the KNIFE approach for genetic data, which can sometimes offer improved behavior over multi-iteration procedures. We focus on quantitative traits and relatively common SNPs. When applied to a set of SNPs within a gene or a pathway, our approach removes noise SNPs from the gene set and yields a small subset of candidate SNPs that can serve as candidate SNPs for functional studies. Extensive simulation studies and a real data illustration are used to evaluate the performance of the proposed approach.

Beyond the KNIFE approach, a wide range of other penalized variable selection procedures have been developed in recent years, such as the LASSO (Tibshirani, 1996) and elastic net methods (Zou, 2005). With an eye towards fine mapping, other penalized approaches have also been developed within the context of genetic association studies to identify genetic variants related to complex traits (Ayers and Cordell, 2010; Zhou et al., 2010; He and Lin, 2011). However, a commonality of these approaches is that they are all generally designed for selecting variables within classical parametric linear or generalized linear regression models, and are not applicable to the kernel machine settings, where the effect of each individual covariate is not directly specified except under simple linear kernels. The Component Selection and Smoothing (COSSO) method (Lin and Zhang, 2006) is designed for variable selection in non-parametric kernel models, but was proposed in the context of smoothing spline ANOVA and requires the use of univariate kernels which does not allow sufficient flexibility in terms of accommodating some of the most popular kernels that are used in genetic analysis.

METHODS

In this section, we first review the kernel machine testing framework with emphasis on both testing as well as estimation of the effects of a group of common variants on a quantitative trait. We then present the proposed variable selection procedure which is an adaptation of the KNIFE approach specifically targeted towards analysis of genetic variants. For simplicity, throughout this article, we restrict attention to quantitative traits and to common genetic variants.

Kernel Machine Testing and Modeling Framework

Focusing on just a single SNP-set, let y_i denote the trait value for the i^{th} person in the sample, $\boldsymbol{X}_i$ be a set of covariates for which we would like to control (including the intercept), and $\boldsymbol{Z}_i = [Z_{i1}, Z_{i2}, \ldots, Z_{ip}]'$ be the genotypes for the SNPs in a SNP-set. Specifically, each Z_{ij} is a trinary variable equal to 0, 1, or 2 for non-carriers, heterozygotes, and homozygous carriers of the rarer allele. Under the kernel machine regression framework, quantitative (continuous) outcomes can be related to the genotypes and any additional covariates through the semiparametric model:

$$y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + h(\boldsymbol{Z}_i) + \varepsilon_i,$$

where ε_i is an error term with mean zero and variance σ^2, and $\boldsymbol{\beta}$ are the regression coefficients for the covariates. In this model $h(\cdot)$ is a generally specified function that lies within a functional space $\mathcal{H}_K$ generated by a positive semi-definite kernel function $K(\cdot, \cdot)$. $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'})$ is a measure of the similarity between subjects i and i' based on the values of the SNPs in the gene set, and importantly, the kernel function fully specifies the relationship between the trait and the SNPs in the gene set, and vice versa. For example, it can be shown that if $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}) = \boldsymbol{Z}_i'\boldsymbol{Z}_{i'}$, called the linear kernel, then this implies that $h(\boldsymbol{Z}_i) = \boldsymbol{\alpha}'\boldsymbol{Z}_i$ for some vector of constants $\boldsymbol{\alpha}$, i.e. $h(\boldsymbol{Z}_i)$ is a linear function of the SNPs in the gene set. The converse is also true: setting $h(\boldsymbol{Z}_i) = \boldsymbol{\alpha}'\boldsymbol{Z}_i$ also implies that the kernel function is equal to the linear kernel. Some examples of commonly used kernel functions for genotype data include:

- *Linear Kernel*: $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}) = \boldsymbol{Z}_i'\boldsymbol{Z}_{i'}/2p$
- *Weighted Linear Kernel*: $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}) = \boldsymbol{Z}_i'\boldsymbol{W}\boldsymbol{Z}_{i'}$
- *IBS Kernel*: $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}) = \sum_{j=1}^{p} IBS(Z_{ij}, Z_{i'j}) = (2p)^{-1}\sum_{j=1}^{p}(2 - |Z_{ij} - Z_{i'j}|)$
- *Quadratic Kernel*: $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}) = (\boldsymbol{Z}_i'\boldsymbol{Z}_{i'} + 1)^2$

Other kernels are possible with the sole condition that they need to satisfy Mercer's theorem. Typically, under the testing framework, estimation of the function $h(\boldsymbol{Z}_i)$ is unnecessary since the test is score test. However, in contrast to the testing framework, in order to do variable selection, we are now conducting estimation instead of testing. Standard estimation of the nonparametric $h(\boldsymbol{Z}_i)$ proceeds by minimizing of the empirical loss function

$$\sum_{i=1}^{n}(y_i - h(\boldsymbol{Z}_i))^2 + \lambda||h||^2_{\mathcal{H}}. \tag{1}$$

Note that for simplicity of notation we omit the covariates $\boldsymbol{X}_i$, but will include them when we discuss the algorithm later. Let $\boldsymbol{Z}$ be the $n \times p$ genotype matrix. By the representer theorem, the solution to equation (1) can be expressed as $h(\boldsymbol{Z}) = \sum_{i=1}^{n}\gamma_i K(\boldsymbol{Z}, \boldsymbol{Z}_i) = \boldsymbol{K}\boldsymbol{\gamma}$ for some constants $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_n]'$ and a kernel matrix $\boldsymbol{K}$. This leads to the alternative dual objective function:

$$\sum_{i=1}^{n}\left(y_i - \sum_{i'=1}^{n}\gamma_{i'}K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'})\right)^2 + \lambda\boldsymbol{\gamma}^T\boldsymbol{K}\boldsymbol{\gamma} \tag{2}$$

which is minimized at $\hat{\boldsymbol{\gamma}} = (\lambda\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{y}$ such that $\hat{h}(\boldsymbol{Z}) = \boldsymbol{K}\hat{\boldsymbol{\gamma}}$, where $\boldsymbol{y}$ is the vector of the trait.

Modified KNIFE Procedure for Selecting Variants Driving Significance

Kernel machine tests are based on score tests which requires estimation under only the null. While this leads to improved computational efficiency and offers some attractive statistical properties, when a particular group of variants are called significant, it is difficult to identify the individual variants that are driving the significant result. Therefore, by adapting to the KNIFE

approach, we propose to apply variable selection methods to identify the variants driving the association. In this section, as with the original KNIFE procedure, we will first introduce weighting terms for individual genetic variants, but we specifically focus on genetically relevant kernels. We then describe modest departure from the original KNIFE and present a 2-step algorithm for estimating some of the weights as exactly zero (enabling variable selection).

Introduction of Individual SNP Weighting Terms

The fundamental idea underlying the KNIFE method is the introduction of a variable specific weight which can be shrunken to zero. Following this idea, we introduce the weighting term, c_j, for each SNP j which we can then shrink to zero in some instances. However, whereas the KNIFE work focused on generic kernels, we restrict attention to some of the kernels that are most genetically relevant. Specifically, we can define the following new kernels:

- Linear: $K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; \boldsymbol{c}) = \sum_{j=1}^{p} c_j Z_{ij} Z_{i'j}$
- Weighted Linear: $K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; \boldsymbol{c}) = \sum_{j=1}^{p} c_j w_j Z_{ij} Z_{i'j}$
- IBS: $K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; \boldsymbol{c}) = \sum_{j=1}^{p} c_j IBS(Z_{ij} Z_{i'j})/(2p) = (2p)^{-1} \sum_{j=1}^{p} c_j \left(2 - |Z_{ij} - Z_{i'j}|\right)$
- d^{th} degree polynomial: $K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; d, \boldsymbol{c}) = (\sum_{j=1}^{p} c_j Z_{ij} Z_{i'j} + 1)^d$

Note that the relationship between the variants and the trait is fully defined based on kernel function. Consequently, if some c_j is exactly zero such that the j^{th} SNP is not used to calculate the similarity between individuals, then the relationship between the trait and the genetic variants does not at all depend on the j^{th} SNP. In this way, SNPs can be dropped from the model allowing for variable selection.

Two-Step KNIFE Estimation Procedure

Although the general KNIFE procedure could be used, here, we propose to use a simplified two-step procedure to do variable selection. We further allow for covariate adjustment which

is imperative for genetic studies. In particular, letting $\boldsymbol{K}_G$ be the kernel matrix induced by $K_G(\cdot,\cdot)$, we propose to use the following procedure:

Step 1: Initialize $\hat{c}_j = 1$ for $j = 1, \ldots, p$. Fix $\boldsymbol{c} = \hat{\boldsymbol{c}}$, then minimize

$$\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \sum_{i'=1}^{n}\gamma_{i'}K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; \hat{\boldsymbol{c}}))^2 + \lambda\boldsymbol{\gamma}^T\boldsymbol{K}_G(\hat{\boldsymbol{c}})\boldsymbol{\gamma}.$$

The solution is known to be $\hat{\boldsymbol{\beta}} = (X'(I + \lambda^{-1}\boldsymbol{K}_G)^{-1}X)^{-1}X'(I + \lambda^{-1}\boldsymbol{K}_G)^{-1}\boldsymbol{y}$ and $\hat{\boldsymbol{\gamma}} = (\lambda\boldsymbol{I} + \boldsymbol{K}_G(\hat{c}))^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$ (Liu et al., 2007), where X is the covariate matrix (including the column of 1 for intercept).

Step 2: Fix $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$, and solve

$$\min_{\boldsymbol{c}}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} - \sum_{i'=1}^{n}\hat{\gamma}_{i'}K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; \boldsymbol{c}))^2, \quad \text{s.t. } \sum_{j=1}^{p}c_j \leq s, c_j \geq 0.$$

Here, s is used to encourage sparsity on c_j. When s is small, then some of the c_j are estimated as exactly zero. We note that for fixed λ and s there are closed form solutions for all of the parameters in step 1. For step 2, some constrained optimization needs to be done and this requires some tailoring towards the particular kernel being used; we will describe computation algorithm for conducting the optimization via cyclic coordinate descent. In principle, λ and s can be selected by performing a 2-dimensional grid search and minimizing a generalized cross validation (GCV) or k-fold CV prediction error. However, the searching of two tuning parameters can be extremely time-consuming and results are often relatively robust to particular values of λ. Thus, in line with Wu et al. (2009), we suggest fixing $\lambda = \sqrt{p/n}$ and using CV to choose s.

This procedure is similar to the original KNIFE approach, but while the original KNIFE procedure essentially iterates between the two steps until convergence, we choose to stop after the second step. In addition to reducing computational expense, the two-step procedure can often offer improved performance over multi-iteration procedures. This is due to the fact that the model is slightly over-parameterized and is in line with other two-step variable selection procedures. By using just two-steps, our work becomes closely related to the well

established non-negative garrote procedure (Breiman, 1995) (and by extension the adaptive LASSO) which we demonstrate in the next section. Further note that the original KNIFE procedure does not explicitly consider covariate adjustment which is a requisite to control for potential confounders and population stratification.

Computational Procedure

As noted the constrained optimization in step 2 requires some tailoring depending on the particular kernel under consideration. In this section, we describe the details of the computational algorithm for estimating some of the weights as exactly zero, focusing on several kernels that are widely used in SNP analysis, i.e., the linear (and weighted linear) kernel, the IBS kernel, and the polynomial kernel.

Linear and Weighted Linear Kernels

By definition $\boldsymbol{K}_G = \boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}'$ where $\boldsymbol{C} = diag\{c_1, \ldots, c_p\}$. Then in the first step, by initializing all $\hat{c}_j = 1$, we estimate

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(X'(I + \lambda^{-1}\boldsymbol{K}_l)^{-1}X\right)^{-1} X'(I + \lambda^{-1}\boldsymbol{K}_l)^{-1}\boldsymbol{y}, \\ \hat{\boldsymbol{\gamma}} &= (\lambda I + \boldsymbol{K}_l)^{-1}\tilde{\boldsymbol{y}}, \end{aligned}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ and $\boldsymbol{K}_l = \boldsymbol{Z}\boldsymbol{Z}'$.

In step 2: to find $\hat{\boldsymbol{c}}$, we minimize:

$$L(\boldsymbol{c}; \hat{\boldsymbol{\gamma}}) = (\tilde{\boldsymbol{y}} - \boldsymbol{K}_G\hat{\boldsymbol{\gamma}})'(\tilde{\boldsymbol{y}} - \boldsymbol{K}_G\hat{\boldsymbol{\gamma}}), \text{subject to: } \sum_{j=1}^{p} c_j \leq s \text{ and } c_j \geq 0.$$

If we substitute in $\boldsymbol{K}_G$ and $\hat{\boldsymbol{\gamma}}$, then

$$L(\boldsymbol{c}; \hat{\boldsymbol{\gamma}}) = (\tilde{\boldsymbol{y}} - \boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}'(\lambda I + \boldsymbol{Z}\boldsymbol{Z}')^{-1}\tilde{\boldsymbol{y}})'(\tilde{\boldsymbol{y}} - \boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}'(\lambda I + \boldsymbol{Z}\boldsymbol{Z}')^{-1}\tilde{\boldsymbol{y}}).$$

We solve the above objective function (with the linear inequality constraints) by implementing the cyclic coordinate descent algorithm (Friedman et al., 2007).

Now we show that our objective function is closely connected with the non-negative garrote objective function (Breiman, 1995). Assume that there are no adjusting covariates. Note that if $\hat{\boldsymbol{\alpha}}$ are the linear ridge regression estimates, then $\hat{\boldsymbol{\alpha}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda I)^{-1}\boldsymbol{Z}'\boldsymbol{y}$ and $\boldsymbol{Z}\hat{\boldsymbol{\alpha}} = (\boldsymbol{Z}\boldsymbol{Z}' + \lambda I)^{-1}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda I)^{-1}\boldsymbol{Z}'\boldsymbol{y}$. Then we see that the dual objective function for the second stage is given as:

$$\begin{aligned} L(\boldsymbol{c}; \hat{\boldsymbol{\gamma}}) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}'(\lambda I + \boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{y})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}'(\lambda I + \boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{y}) \\ &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{C}\hat{\boldsymbol{\alpha}})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{C}\hat{\boldsymbol{\alpha}}). \end{aligned}$$

The estimate for h is given as $\hat{h} = \boldsymbol{K}_G(\hat{\boldsymbol{c}})\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}\hat{\boldsymbol{C}}\hat{\boldsymbol{\alpha}}$ in this case.

At the same time, the nonnegative garrote estimates $\boldsymbol{c}$ (Breiman, 1995) are found by minimizing an objective function:

$$L(\boldsymbol{c}; \tilde{\boldsymbol{\alpha}}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{C}\tilde{\boldsymbol{\alpha}})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{C}\tilde{\boldsymbol{\alpha}})$$

subject to the constraints on $\boldsymbol{c}$, where $\tilde{\boldsymbol{\alpha}}$ are some regression coefficient estimates for genotype matrix $\boldsymbol{Z}$. If $\tilde{\boldsymbol{\alpha}}$ are taken to be the ridge estimates, then we can see that the nonnegative garrote estimates are the same as the estimate for h in our proposed model. The equivalence between the nonnegative garrote and our two-step procedure provides some additional justification (beyond the simulations presented later) that our proposed modifications to the original KNIFE procedure are reasonable.

IBS Kernel

The IBS kernel is generally used to model complicated effects among SNPs. The 'similarity' between two subjects induced by the IBS kernel lies in the absolute value of the difference for a set of SNPs. Similar to the linear kernel, we first obtain the $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ which have closed form solutions. For the second step, we need to minimize

$$L(\boldsymbol{c}; \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \sum_{i'=1}^{n}\hat{\gamma}_{i'}K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; \boldsymbol{c}))^2$$

subject the constraint on $\boldsymbol{c}$. The complicated form of IBS kernel creates challenges for optimization. However, we show in the Appendix that this objective function can be transformed into a nonnegative garrote problem with a new design matrix. Then, the newly formed objective function can be solved by an algorithm similar to the linear kernel.

Quadratic Kernel

The quadratic kernel involves interaction terms between SNPs, and the corresponding objective function can not be directly cast as a non-negative garrote problem. However, as shown in the KNIFE method, polynomial kernels can be linearized by a first-order Taylor expansion. Let $\boldsymbol{w}_{ii'} = (Z_{i1}Z_{i'1}, \ldots, Z_{ip}Z_{i'p})'$. The quadratic kernel $K(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}) = (\sum_{j=1}^{p} c_j Z_{ij} Z_{i'j} + 1)^2$ can be approximated by

$$\begin{aligned} & (\sum_{j=1}^{p} \tilde{c}_j Z_{ij} Z_{i'j} + 1)^2 + 2(\sum_{j=1}^{p} \tilde{c}_j Z_{ij} Z_{i'j} + 1)\boldsymbol{w}'_{ii'}(\boldsymbol{c} - \tilde{\boldsymbol{c}}) \\ = & \left(1 - (\sum_{j=1}^{p} \tilde{c}_j Z_{ij} Z_{i'j})^2\right) + 2(\sum_{j=1}^{p} \tilde{c}_j Z_{ij} Z_{i'j} + 1)\boldsymbol{w}'_{ii'}\boldsymbol{c}, \end{aligned}$$

where $\tilde{\boldsymbol{c}} = [\tilde{c}_1, \ldots, \tilde{c}_p]'$ is some initial estimate of $\boldsymbol{c}$. Thus, given $\hat{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{c}}$, $L(\boldsymbol{c}; \hat{\boldsymbol{\gamma}})$ can be approximated by $L(\boldsymbol{c}; \hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{c}})$, which is a linear kernel problem. Then, we can iteratively solve the objective function by updating $\tilde{\boldsymbol{c}}$ in each iteration.

RESULTS

Simulation Studies

We conducted simulation studies to examine the performance of the proposed approach. We first simulated 10,000 sequence haplotypes using cosi (Schaffner et al., 2005) on a 1 megabase region, with parameters set to mimic sequence data consistent with a population with European ancestry. We then excluded SNPs with minor allele frequency (MAF) less than 0.05, and pruned off highly correlated SNPs (with correlation coefficient $|\rho| > 0.95$). We considered

a SNP-set with 10 SNPs. We randomly picked 10 consecutive SNPs from the simulated haplotypes, and then fixed these 10 SNPs for the following simulation studies. Haplotypes were randomly drawn from the pool of 10,000 haplotypes to form genotypes.

We first simulated the trait under the linear model, $y_i = 1+0.5\times X_i+0.5\times Z_{i1}+0.5\times Z_{i3}+\epsilon_i$, where $X_i \sim N(0,1)$, Z_{i1} and Z_{i3} represent the driver SNPs, and $\epsilon_i \sim N(0,1)$. That is, only the first and the third SNPs contribute to the trait (i.e., driver SNPs), while all the other 8 SNPs are noise SNPs. We name this model set-up as Model Structure I. We are interested to know whether the proposed approach can identify the driver SNPs out of the noise SNPs. We tested three kernels that are widely used in SNP studies, the linear kernel, the quadratic kernel, and the IBS kernel. We considered sample size of 500 and 1000, and the number of Monte Carlo experiments is 100. The LASSO, Elastic Net, and MCP (Zhang, 2010), which are penalized regression methods that can be used for fine mapping, were included for comparison as potential competitors. We calculated several quantities to measure the performance of the compared approaches, described as follows. Let $(\hat{c}_1,\ldots,\hat{c}_p)$ be the final estimates of $(c_1,\ldots,c_p)$, and $I(\cdot)$ be the indicator function. We calculate (1) the number of SNPs being selected, i.e., $\sum_{j=1}^{p} I(\hat{c}_j \neq 0)$, (2) the proportion of driver SNPs being selected (Capture rate), which is defined as $\sum_{j=1}^{p} I(\hat{c}_j \neq 0)I(c_j \neq 0)/\sum_{j=1}^{p} I(c_j \neq 0)$, (3) the false positive rate (FPR), i.e., $\sum_{j=1}^{p} I(\hat{c}_j \neq 0)I(c_j = 0)/\sum_{j=1}^{p} I(\hat{c}_j \neq 0)$, and (4) the proportion of experiments in which the selected SNPs cover all the driver SNPs (Coverage probability), $\sum_{m=1}^{M} I\big(\{j : c_j \neq 0\} \subseteq \{j : \hat{c}_j^{(m)} \neq 0\}\big)/M$, where $\hat{c}_j^{(m)}$ is the estimate of c_j in the mth experiment, and M is the total number of experiments. We also calculated the rank-sum of the estimated coefficients for the driver SNPs with respect to the noise SNPs, $\sum_{j=1}^{p} \Re(\hat{c}_j)I(c_j \neq 0)$, where $\Re(\cdot)$ is the rank function defined on $\{\hat{c}_1,\ldots,\hat{c}_p\}$; this metric measures how often an approach yields higher ranks for the driver SNPs than the noise SNPs. The results are shown in Table I. As can be seen, the linear kernel has a higher capture rate and lower FPR than the quadratic and IBS kernels. In each experiment, the linear kernel also tends to cover all the driver SNPs, as shown by its high coverage probability. The LASSO, Elastic Net and

MCP also have high coverage probability, but their empirical FPRs appear to be higher than the linear kernel. The linear kernel approach tends to assign higher ranks to the driver SNPs compared to the other two kernel approaches. This indicates that when the true model is a linear model, the linear kernel outperforms the other two kernels in prioritizing SNPs that are of importance.

Next, we introduced interaction terms into the model to simulate the trait. We let $y_i = 1 + 0.5 \times X_i + 0.8 \times Z_{i2} + 0.8 \times Z_{i7} - 1.0 \times Z_{i2} \times Z_{i7} + \epsilon_i$. That is, the trait is influenced by the interaction effect between SNPs 2 and 7. Under this set up, because the interaction term has opposite sign with respect to the main effects, $cor(y, Z_2)$ and $cor(y, Z_7)$ tend to be small in magnitude, and this makes it challenging to tease out driver SNPs from noise SNPs. We wish to test whether the three approaches can still identify SNPs 2 and 7 as the driver SNPs. As shown in Table II, the quadratic kernel has high probability to pick up the driver SNPs. The other approaches show high FPR, and tend to have low power to capture the driver SNPs. This example shows that the quadratic kernel can perform much better than the other two kernels when the true model contains interaction effects.

Finally, we simulated the trait under a non-linear model: $y_i = 0.5 \times X_i + 0.3 \times I(Z_{i2} = 0) + 1.0 \times I(Z_{i2} = 1) + 0.1 \times I(Z_{i2} = 2) + \epsilon_i$. In other words, the heterozygote has higher effect on the trait than the two types of homozygotes. In the biology literature, this type of model is known as the Heterozygote Advantage model, and an example can be seen in Penn et al. (2002). Through basic calculations, we show in the Appendix that (1) when there are no adjusting covariates, the covariance between y and Z_2 is solely dependent upon the MAF of Z_2 and the effect sizes of the three genotypes of Z_2, and (2) the correlation between y and Z_2 tends to be small under this Heterozygote Advantage model. In fact, the marginal association between y and Z_2 can be nearly zero. Under such a nonlinear situation, the linear kernel is expected to have low power to detect Z_2, the driver SNP. On the other hand, it is straightforward to show that the Heterozygote Advantage model considered herein can be characterized by a model that contains both linear and quadratic effects for Z_2. Thus, we

anticipate that the quadratic kernel should perform well in identifying the driver SNP. As shown in Table III, the linear kernel, LASSO, Elastic Net and MCP tend to miss the driver SNP, while the quadratic kernel captures the driver SNP with high probability. The IBS kernel also seems to have good performance under this model. This is likely due to the fact that the IBS kernel has complicated basis functions and can accommodate certain non-linear effects.

In addition to these simulations, we further considered scenarios in which effect sizes were smaller with larger number of variants. Results (see Supplement) were qualitatively similar and also support our method, though when signal is too weak, no method can perform well. Further simulations considering rare variants and alternative implementations of our procedure (under multiple iterations and with two-dimensional grid search to select tuning parameters) are also presented as Supplemental Material.

Application to Birth-weight Studies

We illustrate our approach via application to a real dataset, examining the association between birth outcomes and genetic variants at a candidate gene. In particular, we considered a study in which 20 SNPs within the EDN1 (Endothelin 1) gene were genotyped in a sample of 853 singleton, live births from women of European Ancestry in the Pregnancy, Infection and Nutrition Cohort (Savitz et al., 2001). Our overall objective in this analysis was to examine the association between the SNPs in EDN1 and birth-weight, which is an important determinant of many subsequent health conditions (Hack et al. 2002).

The particular objectives of our analysis here were to, first, assess the overall association between the EDN1 SNPs and birth-weight, and second, to identify any SNPs which may be driving potential associations. Of the 20 SNPs in EDN1, two SNPs have correlation coefficient equal to 1, and we removed one of them from our analysis. We first applied the SKAT test with the linear kernel to EDN1 while adjusting for gender, preterm birth status, maternal smoking status, and parity. The resulting SKAT p-value is 0.028, indicating that there is potential association between EDN1 and birth weight. However, SKAT does not allow for

identification of individual driving variants. Thus, it is unclear whether the result is due to one very strongly associated SNP or whether there are multiple, modestly associated SNPs.

To identify SNPs that are driving the observed association, we applied our approach with the linear kernel to EDN1. Among the 19 SNPs, only the $\hat{c}_j$ for rs6931867 is nonzero. The LASSO, Elastic Net, and MCP selected 2, 5, and 3 SNPs, respectively, and they all included rs6931867. We then applied SKAT to the post-selection SNPs for each model, and the p-values for linear kernel, LASSO, Elastic Net and MCP are 0.004, 0.009, 0.007 and 0.009, respectively. To quantify the effect size of rs6931867 on birth weight, we then fitted an unregularized linear regression model for rs6931867, along with other adjusting covariates such as the gender and preterm birth. The results are shown in Table IV. Perhaps not surprising, the preterm birth status has the largest effect (-492.78) on birth weight among all the considered covariates. On the other hand, rs6931867 also shows a strong effect (111.79) on birth weight, even stronger than the 'gender' (-86.41) and 'smoking' (-86.06). We next examined rs6931867 using the UCSC Genome Brower. We plotted rs6931867 along with its neighbor SNPs using the SNAP software (Broad Institute), and it can be seen that this SNP is located in the 5'UTR of the EDN1 gene (Figure 1). According to the UCSC genome brower, rs6931867 falls into a DNase I Hypersensitivity Cluster, indicating that this SNP is possibly engaged in gene regulation. These findings suggest that rs6931867 is an intriguing SNP for further study; the evaluation of its function role may shed light on the regulation mechanism of EDN1 expression.

Application to Grady Trauma Project Data

We also applied our method to analyze the genetic regulation of gene expression using data collected from the Grady Trauma Project (Gillespie et al., 2009), a study investigating the genetic factors in response to stressful life events. 337 study subjects were recruited from the waiting rooms of primary care and obstetrics-gynecology clinics of Grady Memorial Hospital in Atlanta, Georgia. Gene expression and genotypes were both measured using the whole blood samples. The expression data are available at GEO (Gene Expression Omnibus) under the

accession GSE58137. In this manuscript, we are interested in the cis-regulation, i.e., whether the genotype of the gene can influence the expression of the same gene.

We considered gene MTHFR (methylenetetrahydrofolate reductase), a key regulator in folate, thiol, homocysteine, methylation and thymidine metabolism. MTHFR has been shown to play an important role in inflammation and oxidative stress (Faraci and Lentz, 2004), as well as in the development of many diseases, including heart diseases, cancers, and mental disorders (Odin et al., 2006). We first applied the SKAT method to evaluate the association between the 22 SNPs in the gene and the expression of MTHFR. Under the linear kernel, the p-value for association is 2.61e-07; and under the IBS kernel, the p-value is 5.10e-11, indicating a possible non-linear relationship between the genotype and the expression. Using the IBS kernel, the proposed variable selection method identified six important SNPs (SNP number: 2, 6, 13, 17, 18, 19), which overlap largely with the 12 SNPs (SNP number: 2, 5, 6, 7, 10, 11, 13, 16, 17, 19, 21, 22) that were selected using the linear kernel, with only one exception (SNP 18). The SKAT model using the six selected SNPs generated a p-value of 1.08e-14, which is considerably more significant than using the 12 SNPs that were selected using the linear kernel (p-value = 4.98E-08).

To further examine the effects of the selected SNPs, we fitted an unregularized linear regression model using the six SNPs that were selected from the IBS model. In order to assess the potential nonlinear effect, we coded each SNP (except SNPs 2 and 13 which have only values 0 and 1) by two dummy variables using the genotype of 0 as the reference, i.e. using a co-dominant coding. This allows every genotype to have a different and nonadditive effect. Table V shows the effect size and p-values obtained from this unregularized linear regression model. Noteworthy, SNP 18 (rs2066470) showed a strong non-linear effect in regulating the gene expression. The effect estimates for a heterozygous change and a homozygous change in this SNP are in the opposite direction, which is different from the additive assumption that the linear kernel assumes. This example shows that the variable selection using nonlinear kernels can be more effective in identifying important SNPs in a SNP-set.

DISCUSSION

Set-based approaches have become a powerful approaches for genetic association studies. However, the major limitation of set-based approaches is that they provide little information on which SNPs may be (or closely related to) the driver SNPs. Yet fine mapping of the individual driver variants is imperative for development of further functional studies and facilitating interpretation of identified signals. The proposed approach conducts post-SKAT variable selection to identify important SNPs, and hence well complements the SKAT for SNP-set analysis. The selected SNPs will help to narrow down candidate regions for biological functional studies, which have recently attracted considerable attention from the biomedical research community (Wang et al., 2015).

In this article, we have focused on relatively common SNPs, with the understanding that the kernel machine testing is often used for analysis of common genetic variants. That said, SKAT is perhaps even more popular for the analysis of rare genetic variants. We have conducted some initial simulations examining the possibility of applying our approach for rare variants with initially promising results. We emphasize, however, that these results are not meant to serve as a comprehensive examination of the topic and merely demonstrate that our approach is potentially applicable under the important setting of rare variants. Rare variant analysis is made challenging by a range of unique features. Because of the low MAF, the data effectively become binary such that issues of non-linearity are less visible and while interactive effects are still important, when individual MAFs are low then the interaction will become exceedingly uncommon. Further consideration of this and related issues, such as the need to accommodate extrinsic information (e.g. functionality) and limited ability to observe the causal variants, deserves dedicated attention which is beyond the scope of this article.

Although our approach is powerful for enabling prioritization of individual variants, a limitation of the approach is that when the SNP density of the studied gene is not very high, the driver SNP is likely to be a tagging SNP, and more refined mapping will be necessary

to track down the likely functional SNPs (Yao et al., 2014). With the rise of sequencing technology and improved imputation, however, it is increasingly likely that the true causal variant will be genotyped. Related to this point is the fact that many SNPs are often in high LD. Even in our data illustration, two SNPs were perfectly correlated. In this scenario, it is impossible for any computational technique to identify the causal variant without external information and/or additional experiments. Nonetheless, the proposed method can allow for identification of a restricted set of putative SNPs that drive the associations and aid in the design of down-stream experiments.

While our approach can be used to prioritize individual variants, a limitation is that it is difficult to conduct formal inference on the individual selected variants. Due to the selection procedure, subsequently obtaining p-values for the individual selected variants (without consideration of unselected variants) will yield optimistic p-values. Similarly, as observed in the real data analysis, re-testing just the selected variants tends to yield more significant results. Accordingly, we recommend caution in conducting or interpreting any post-hoc inference.

An assumption underlying our approach is that a particular kernel has already been chosen. In general, our approach is primarily designed as a follow-up to testing, and we suggest directly using the same kernel that was used to obtain the significant testing results. However, we acknowledge that it is not always the case that a single kernel is obvious and the best kernel may actually be a weighted average of multiple kernels (Wu et al., 2013). We can extend our approach to simultaneously consider the problem of kernel choice by jointly considering multiple kernels together as a composite kernel. Then the weights for the composite kernel can also be shrunken such that we are conducting joint kernel and individual SNP selection. This approach would not only allow for selection of driving variants but also provide clues as to how the variants are influencing the outcome.

Currently, the proposed approach does not use any external information, yet there is considerable interest in the field in accommodating prior knowledge into analyses, both to improve power and to improve interpretation. SNP annotation tools, such as the PolyPhen-2

(Adzhubei, 2013), can also be used to assign a functional score to each SNP, which can then be transformed into weights representing prior expectation that each SNP influences the trait. A simple modification can be made to allow for incorporation of prior biological information on SNP function or likely effects by adjusting the threshold s to be different for each variant (this would be equivalent to simply re-scaling the SNP values based on prior knowledge). How to best translate prior knowledge into weights remains a topic of future research.

ACKNOWLEDGMENTS

This research was supported in part by NIH R21HD060207, R01HG007508, R01HG006292, R01MH071537, R01MH096764, the Fred Hutchinson Cancer Research Center Institutional Research Support, and the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

REFERENCES

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet, Chap.7:Unit7.20.

Allen GI. 2013. Automatic feature selection via weighted kernels and regularization. J Comp Graph Stat 22: 284-299.

Auer PL, Teumer A, Schick U, O'Shaughnessy A, Lo KS, Chami N, Carlson C, de Denus S, Dube MP, Haessler J and others. 2014. Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat Genet 46: 629-634.

Ayers KL and Cordell HJ. 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34:879-891.

Breiman L. 1995. Better subset regression using the nonnegative garrote. Technometrics 37: 373-384.

Cornes B, Brody J, Morrison A, Siscovick D, Meigs J, CHARGE-S Diabetes WG. 2013. Abstract P054: rare variants in and near IRS1 are associated with fasting insulin in CHARGE-S. Circulation 127: AP054.

Edwards SL, Beesley J, French JD, Dunning AM. 2013. Beyond GWASs: Illuminating the dark road from associationn to function. Am J Hum Genet 93: 779-797.

Faraci F and Lentz S. 2004. Hyperhomocysteinemia, oxidative stress, and cerebral vascular dysfunction. Stroke 35: 345-347.

Friedman J, Hastie T, Hofling H, Tibshirani R. 2007. Pathwise coordinate optimization. Ann Appl Stat 1: 302-332.

Gillespie C, Bradley B, Mercer K, Smith A, Conneely K, Gapen M, Weiss T, Schwartz A, Cubells J, Ressler K. 2009. Trauma exposure and stress-related disorders in inner city primary care patients. Gen Hosp Psychiatry 31: 505-514.

Hack M, Flannery D, Schluchter M, Cartar L, Borawski E, Klein N. 2002. Outcomes in young adulthood for very-low-birth-weight infants. N Eng J Med 346: 149-157.

He Q and Lin DY. 2011. A variable selection method for genome-wide association studies. Bioinformatics 27: 1-8.

Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. 2013. Family-based association tests for sequence data, and comparisons with population-based association tests. Eur J Hum Genet 21: 1158-62.

Lin X. 1997. Variance component testing in generalized linear models with random effects. Biometrika 84: 309-326.

Lin X, Cai T, Wu M, Zhou Q, Liu G, Christiani D, Lin X. 2011. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet Epidemiol 35: 620-631.

Lin Y and Zhang HH. 2006. Component selection and smoothing in multivariate nonparametric regression. Ann Statist 34: 2272-2297.

Liu D, Lin X, and Ghosh D. 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63: 1079-88.

Odin E, Wettergren Y, Carlsson G, Danenberg P, Termini A, Willen R, Gustavsson B. Expression and clinical significance of methylenetetrahydrofolate reductase in patients with colorectal cancer. Clin colorectal Cancer 5: 344-349.

Penn DJ, Damjanovich K, Potts WK. 2002. MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc Natl Acad Sci 99: 11260-4.

Savitz D, Dole N, Terry J, Zhou H, Thorp J. 2001. Smoking and pregnancy outcome among African-American and white women in central North Carolina. Epidemiology 12: 636-642.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15, 1576-1583.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B 58: 267-288.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81: 1278-1283.

Wang R, Li M, Zhou S, Zeng D, Xu X, Xu R, Sun G. 2015. Effect of a single nucleotide polymorphism in miR-I46a on COX-2 protein expression and lung function in smokers with chronic obstructive pulmonary disease. Int J COPD 10: 463-473.

Wu MC, Zhang L, Wang Z, Christian DC, Lin X. 2009. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. Bioinformatics 25: 1145-51.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP set analysis for case-control genome wide association studies. Am J Hum Genet 86: 929-942.

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

Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Molldrem JJ., Armistead PM. 2013. Kernel machine SNP-set testing under multiple candidate kernels. Genet Epidemiol 37: 267-75.

Yan B, Wang S, Jia H, Liu X, Wang X. 2015. An efficient weighted tag SNP-set analytical method in genome-wide association studies. BMC Genetics 16: 25.

Yao L, Tak Y, Berman B, Farnham P. 2014. Functional annotation of colon cancer risk SNPs. Nat Communic 5:5114.

Zhang C. 2010. Nearly unbiased variable selection under minimax concave penalty. Ann Statist 2: 894-942.

Zhou H, Sehl M, Sinsheimer J, and Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26: 2357-2382.

Zou H and Hastie T. 2005. Regularization and variable selection via the elastic net. J Roy Stat Soc B 67: 301-320.

APPENDIX I: TRANSFORM THE IBS KERNEL OBJECTIVE FUNCTION INTO A NONNEGATIVE GARROTE PROBLEM

For the IBS kernel,

$$\begin{aligned} L(\boldsymbol{c};\hat{\gamma}) &= \sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \sum_{i'=1}^{n}\hat{\gamma}_{i'}K_G(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'}; c))^2 \\ &= \sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \sum_{i'=1}^{n}\hat{\gamma}_{i'} \times (2p)^{-1}\sum_{j=1}^{p}c_j(2 - |Z_{ij} - Z_{i'j}|)\right)^2 \\ &= \sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \sum_{j=1}^{p}c_j\sum_{i'=1}^{n}\hat{\gamma}_{i'}(2 - |Z_{ij} - Z_{i'j}|)/(2p)\right)^2 \end{aligned}$$

Now, define $\xi_{ij} = \sum_{i'=1}^{n}\hat{\gamma}_{i'}(2 - |Z_{ij} - Z_{i'j}|)/(2p)$, then the objective function becomes

$$\sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i'\boldsymbol{\beta} - \sum_{j=1}^{p}c_j\xi_{ij}\right)^2,$$

subject to $\sum_{j=1}^{p}c_j \leq s, c_j \geq 0$. Given fixed $\hat{\gamma}_{i'}$, this is equivalent to a nonnegative garrote problem and can be solved accordingly.

APPENDIX II: Quantify the covariance between y and Z_2 under the Heterozygote Advantage Model

The Heterozygote Advantage model specifies that $y_i = \beta_1 \times I(Z_{i2} = 0) + \beta_2 \times I(Z_{i2} = 1) + \beta_3 \times I(Z_{i2} = 2) + \epsilon_i$, where $\beta_2 > \beta_1$ and $\beta_2 > \beta_3$. Assume that the MAF of Z_2 is p and that Z_2 is under the Hardy-Weinberg Equilibrium. Let $q = 1 - p$. We wish to evaluate $Cov(y, Z_2) = E(yZ_2) - E(y)E(Z_2)$.

First, it can be shown that $E(Z_2) = 2p$. Next, we notice that $E(y) = E(E(y|Z_2)) = \sum_{j\in(0,1,2)} E(y|Z_2 = j) \times P(Z_2 = j) = \beta_1 q^2 + 2\beta_2 pq + \beta_3 p^2$. We also note that $E(yZ_2) = E(E(yZ_2|Z_2)) = E(Z_2 E(y|Z_2)) = \sum_{j\in(0,1,2)} j \times E(y|Z_2) \times P(Z_2 = j) = 2\beta_2 pq + 2\beta_3 p^2$.

It follows that $Cov(y, Z_2) = 2pq\,(q(\beta_2 - \beta_1) + p(\beta_3 - \beta_2))$, which is solely dependent upon the MAF of Z_2 and the three genetic effects. Hence, when $\beta_2 > \beta_1$ and $\beta_2 > \beta_3$, the term $(q(\beta_2 - \beta_1) + p(\beta_3 - \beta_2))$ tends to be small due to the opposite effect of $(\beta_2 - \beta_1)$ and $(\beta_3 - \beta_2)$. In particular, when $q(\beta_2 - \beta_1) + p(\beta_3 - \beta_2) = 0$, we have $Cov(y, Z_2) = 0$.

TABLE I. Comparison of different methods under Model Structure I

	LASSO	E.Net	MCP	Linear	Polynom.	IBS
n=500						
#SNPs selected	4.5	4.3	5.1	3.0	3.3	5.5
Capt. rate	1.0	1.0	1.0	1.0	0.8	1.0
FPR	0.44	0.43	0.54	0.22	0.36	0.61
Rank-sum	18.9	19.0	19.0	18.8	16.4	14.7
Cover. Prob.	1.0	1.0	1.0	1.0	0.6	1.0
n=1000						
#SNPs selected	4.1	4.4	4.8	2.6	3.4	4.9
Capt. rate	1.0	1.0	1.0	1.0	0.9	1.0
FPR	0.42	0.45	0.51	0.13	0.32	0.54
Rank-sum	19.0	19.0	19.0	19.0	17.8	15.3
Cover. Prob.	1.0	1.0	1.0	1.0	0.8	1.0

TABLE II. Comparison of different methods under Model Structure II

	LASSO	E.Net	MCP	Linear	Polynom.	IBS
n=500						
#SNPs selected	2.1	2.1	3.1	2.1	3.0	2.3
Capt. rate	0.3	0.3	0.4	0.3	1.0	0.2
FPR	0.65	0.73	0.73	0.58	0.24	0.92
Rank-sum	11.7	11.6	11.6	12.3	18.8	10.1
Cover. Prob.	0.1	0.1	0.2	0.1	1.0	0.1
n=1000						
#SNPs selected	2.6	2.9	4.1	2.0	2.4	2.8
Capt. rate	0.4	0.4	0.5	0.4	1.0	0.3
FPR	0.67	0.66	0.71	0.50	0.11	0.90
Rank-sum	12.5	12.5	12.4	12.8	19.0	10.6
Cover. Prob.	0.2	0.2	0.3	0.1	1.0	0.2

TABLE III. Comparison of different methods under the Heterozygote Advantage model

	LASSO	E.Net	MCP	Linear	Polynom.	IBS
n=500						
#SNPs selected	1.6	1.7	2.5	1.9	2.0	1.9
Capt. rate	0.1	0.1	0.2	0.3	1.0	1.0
FPR	0.94	0.98	0.95	0.81	0.27	0.27
Rank-sum	5.2	5.0	5.0	5.9	9.8	9.9
Cover. Prob.	0.1	0.1	0.2	0.3	1.0	1.0
n=1000						
#SNPs selected	1.3	1.2	2.2	1.4	1.7	1.6
Capt. rate	0.2	0.1	0.3	0.3	1.0	1.0
FPR	0.87	0.87	0.86	0.71	0.20	0.21
Rank-sum	5.6	5.6	5.7	6.5	9.9	10.0
Cover. Prob.	0.2	0.1	0.3	0.3	1.0	1.0

TABLE IV. Effects of rs6931867 and other adjusting covariates on birth weight

	gender	preterm birth	parity	smoking	rs6931867
Effect	-86.41	-492.78	7.28	-86.06	111.79

TABLE V. Effects of the six SNPs on the gene expression of MTHFR (genotype values in parenthesis)

SNP	Effect	p-value	SNP	Effect	p-value
SNP 2 (1)	-0.118	0.009	SNP 17 (2)	0.338	0.128
SNP 6 (1)	0.070	0.035	SNP 18 (1)	0.141	0.0067
SNP 6 (2)	0.088	0.250	SNP 18 (2)	-0.275	0.282
SNP 13 (1)	-0.058	0.259	SNP 19 (1)	-0.056	0.061
SNP 17 (1)	0.008	0.866	SNP 19 (2)	-0.119	0.424

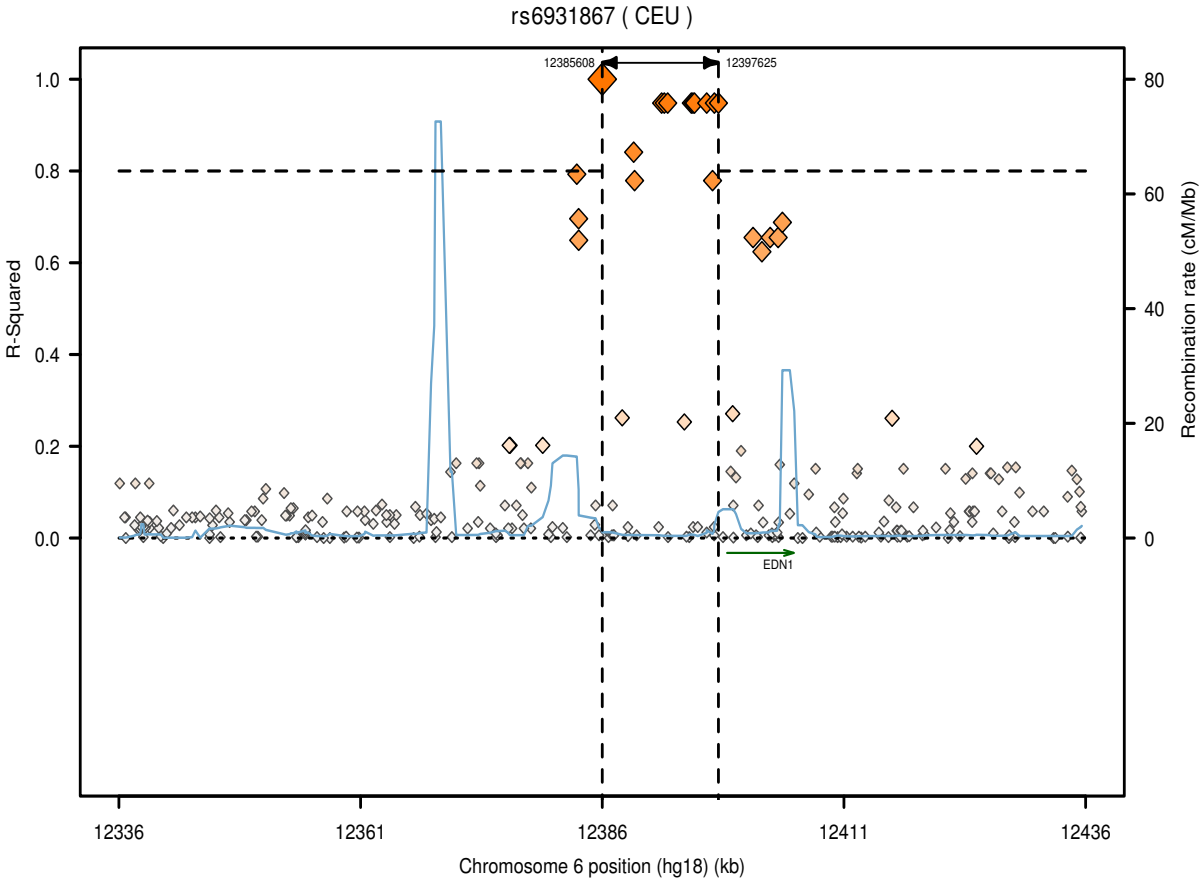

Figure 1: rs6931867 and other SNPs near the EDN1 region (Plot is based on the 1000 Genomes Pilot 1 CEU data; diamond represents SNP, and solid line shows the recombination rate.)