The Annals of Applied Statistics
2014, Vol. 8, No. 4, 2292–2318
DOI: 10.1214/14-AOAS771

A FAST ALGORITHM FOR DETECTING GENE–GENE INTERACTIONS IN GENOME-WIDE ASSOCIATION STUDIES

BY JIAHAN LI*, WEI ZHONG†,1, RUNZE LI‡,2 AND RONGLING WU‡

University of Notre Dame, Xiamen University*†
and Pennsylvania State University‡

With the recent advent of high-throughput genotyping techniques, genetic data for genome-wide association studies (GWAS) have become increasingly available, which entails the development of efficient and effective statistical approaches. Although many such approaches have been developed and used to identify single-nucleotide polymorphisms (SNPs) that are associated with complex traits or diseases, few are able to detect gene–gene interactions among different SNPs. Genetic interactions, also known as epistasis, have been recognized to play a pivotal role in contributing to the genetic variation of phenotypic traits. However, because of an extremely large number of SNP–SNP combinations in GWAS, the model dimensionality can quickly become so overwhelming that no prevailing variable selection methods are capable of handling this problem. In this paper, we present a statistical framework for characterizing main genetic effects and epistatic interactions in a GWAS study. Specifically, we first propose a two-stage sure independence screening (TS-SIS) procedure and generate a pool of candidate SNPs and interactions, which serve as predictors to explain and predict the phenotypes of a complex trait. We also propose a rates adjusted thresholding estimation (RATE) approach to determine the size of the reduced model selected by an independence screening. Regularization regression methods, such as LASSO or SCAD, are then applied to further identify important genetic effects. Simulation studies show that the TS-SIS procedure is computationally efficient and has an outstanding finite sample performance in selecting potential SNPs as well as gene–gene interactions. We apply the proposed framework to analyze an ultrahigh-dimensional GWAS data set from the Framingham Heart Study, and select 23 active SNPs and 24 active epistatic interactions for the body mass index variation. It shows the capability of our procedure to resolve the complexity of genetic control.

1. Introduction. Genome-wide association studies (GWAS) have been a powerful tool for genetic and biomedical research. The past decade has witnessed the rapid development of GWAS and the substantial contributions it has made

Received July 2012; revised May 2014.

[1]Supported by NNSFC Grants 11301435, 71131008 and the Fundamental Research Funds for the Central Universities 20720140034. Wei Zhong is the corresponding author.

[2]Supported by NIDA Grant P50-DA10075 and NNSFC Grant 11028103.

Key words and phrases. Gene–gene interaction, GWAS, high-dimensional data, sure independence screening, variable selection.

[Altshuler, Daly and Lander (2008); Psychiatric GCCC (2009); Hirschhorn (2009); Das et al. (2011)]. With advances in high-throughput genotyping techniques and modern statistics, GWAS have been helping investigators understand the genetic basis of many complex traits or diseases, providing valuable clues to the genetic predisposition of common diseases and drug responses [Burton et al. (2007); Daly (2010)], among others.

In a typical GWAS, hundreds of thousands of single-nucleotide polymorphisms (SNPs) are usually genotyped on a cohort being studied to identify important genetic variants that are associated with the trait of interest. Although fast and inexpensive, the collection of genetic information is normally limited to a sample involving hundreds of subjects, which brings statistical challenges for estimating and identifying relevant genetic risk factors. With SNPs being predictors and phenotypes being the response, single-SNP analysis is mostly performed. However, such a single-SNP approach is neither efficient nor precise, since it fails to consider all SNPs and their possible interactions simultaneously, and to adjust the estimated effects accordingly. Therefore, many statistical procedures that consider all SNPs jointly have been proposed for analyzing the high-dimensional data sets generated by genome-wide association studies.

This is a feature selection problem for high-dimensional data, where the number of SNPs (p) is much larger than the number of observations (n). Penalized regressions, which are developed to overcome severe drawbacks of traditional variable selection techniques, are widely used to select a subset of important predictors from a large number of potential predictors. In the GWAS analysis of case-control studies, Wu et al. (2009) and Cho et al. (2009) applied LASSO penalized regression [Tibshirani (1996)] and elastic-net penalized regression [Zou and Hastie (2005)], respectively. Ayers and Cordell (2010) further conducted a comprehensive study to examine the performance of a variety of penalized regressions in case-control studies. They concluded that variable selection techniques based on penalized regressions outperform single-SNP analysis and stepwise selection. To further explore the potential of high-dimensional statistical models for identifying disease susceptibility genes, several two-stage approaches have been proposed for selecting significant main effects. Li et al. (2011) employed preconditioning and Bayesian LASSO on population cohorts to estimate genetic effects of SNPs on continuous traits. He and Lin (2011) developed a GWASelect procedure for case-control cohorts, where several steps of iterative sure independence screening (ISIS) and LASSO regression are involved.

These methods based on penalized regressions have demonstrated their statistical power and computational feasibility over the single-SNP analysis. However, since statistical methodologies and computations have already been challenged by the overwhelming number of whole-genome SNPs, these methods either do not consider gene–gene interactions or estimate interactions among only a small number of selected SNPs with significant main effects. However, without considering

the full picture of epistatic interactions in GWAS analysis, only a limited portion of phenotypic variation can be explained such that potential disease-associated pathways and risk factors can hardly be identified [Manolio et al. (2009); Cordell (2009)].

In light of recent developments in machine learning, many sophisticated approaches have been proposed to search for whole-genome interactions in genome-wide association studies, most of which are designed for case-control cohorts. These machine learning approaches include a Bayesian partitioning model [Zhang and Liu (2007)], a SNPRuler based on an association rule [Wan et al. (2010b)] and random forest approaches [Breiman (2001); Kim et al. (2009)]. However, these methods are computationally intensive and do not perform well in practice when the genome-wide SNP data are considered [Wan et al. (2010b); Wang et al. (2011); Szymczak et al. (2009)]. More recently, adaptive LASSO [Yang et al. (2010)] and Bayesian generalized linear models [Yi, Kaklamani and Pasche (2011)] are applied to detect epistatic interactions in case-control cohorts where all SNP pairs are exhaustively searched. Wang et al. (2011) present a comprehensive comparison of the prevailing epistatic interaction detection methods, including SNPRuler [Wan et al. (2010b)], SNPHarvester [Yang et al. (2009)], Screen and Clean [Wu et al. (2010)], BOOST [Wan et al. (2010a)] and TEAM [Zhang et al. (2010)]. They concluded that these methods perform differently in terms of statistical power, false positive rate and computational cost. However, methods other than Screen and Clean are specially designed for case-control studies where phenotypic values are binary, and cannot be applied to quantitative traits unless the phenotypical values are properly discretized.

In this paper, we propose a statistical framework for detecting whole-genome epistatic interactions in a population cohort where phenotype is continuous. The framework incorporates the well-developed penalized regressions, which have proved successful in detecting SNPs with significant main effects. Therefore, existing findings regarding penalized regression theories and their empirical performance in GWAS analysis can provide direct and valuable insights into our framework. Moreover, the proposed algorithm is suitable for parallel computing and does not involve computationally demanding techniques on the whole-genome SNP data that prevailing interaction models may involve, such as resampling strategies and Bayesian analysis. As a result, it is computationally efficient.

Specifically, we develop a two-stage sure independence screening (TS-SIS) procedure before variable selection. The screening step forms a pool of important SNPs, which may either have significant main effects or demonstrate no marginal effects but strong epistatic interactions. Since the two-stage screening is based on sure independence screening [Fan and Lv (2008)], the computational burden of selecting important interactions is greatly reduced. More importantly, this procedure guarantees the performance of the following variable selection procedure, in the sense that once important SNPs and interactions enter the pool, the probability of

identifying the correct ones is very high. We also propose a rates adjusted thresholding estimation (RATE) approach to determine the number of predictors retained by a variable screening procedure. This approach is based on soft-thresholding and bootstrapping, and relates the reduced model size to a false positive rate. Ueki and Tamiya (2012) proposed hard-thresholding-based sure independence screening (SIS) to select promising main genetic effects and interactions for penalized regressions. Motivated by the multifactor dimensionality reduction [MDR; Ritchie et al. (2001)] approach, they proposed dummy coding methods to effectively capture various patterns of interactions in case-control studies. Our approach, however, is more general and suitable for population-based GWAS and other variable screening problems.

We applied the newly developed statistical framework to analyze a GWAS data set from the Framingham Heart Study, aimed to identify genetic variants that are associated with obesity, blood pressure and heart disease. We find that, out of 349,985 SNPs, 23 SNPs and 24 epistatic interactions have notable effects on the body mass index (BMI). By applying gene-set enrichment analysis tools [Wang, Li and Bucan (2007); Holden et al. (2008)] in future studies, biological knowledge can be integrated to discover and prioritize signaling pathways implied by detected SNPs. Morever, SNP–SNP interactions will provide insight into functional related genes and the structure of genetic pathways, allowing better understanding of complex genetic architecture and cellular processes in a system level.

In Section 2 we introduce the TS-SIS procedure that reduces the model dimensionality and identifies potential gene–gene interactions. Section 3 proposes a rates adjusted thresholding estimation (RATE) approach to determine the number of predictors retained by a general variable screening procedure. Section 4 shows how penalized regression can fit in this framework and gives the estimation procedure for SCAD penalized regression [Fan and Li (2001)]. In Section 5 the statistical properties of this framework are investigated through simulation studies. Section 6 applies this framework to the Framingham Heart Study. Concluding remarks are given in Section 7.

2. Two-stage sure independence screening. In genome-wide association studies phenotypical measurements are explained by a handful of covariates and a great number of genetic factors represented by SNP genotypes. To select important SNPs and estimate their genetic effects precisely by adjusting for observed covariates, we employ a GWAS model that takes into account the effects of both genetic effects and covariate effects. Moreover, as we discussed, epistatic interactions play a central role in understanding metabolic pathways of complex diseases and traits. Therefore, a comprehensive GWAS model incorporating both main genetic effects and gene–gene interactions is more appropriate.

For subject i in a population cohort consisting of a total of n subjects, we describe the observed phenotypic value y_i as

$$
\begin{aligned}
y_i = \mu &+ \sum_{k=1}^{q} x_{k,i}\alpha_k + \sum_{j=1}^{p} \xi_{j,i} a_j + \sum_{j=1}^{p} \zeta_{j,i}\, d_j + \sum_{j=1}^{p} \sum_{j'<j} \xi_{j,i}\xi_{j',i}\mathcal{I}^{aa}_{jj'} \\
&+ \sum_{j=1}^{p} \sum_{j'=1}^{p} \xi_{j,i}\zeta_{j',i}\mathcal{I}^{ad}_{jj'} + \sum_{j=1}^{p} \sum_{j'=1}^{p} \zeta_{j,i}\xi_{j',i}\mathcal{I}^{da}_{jj'} \\
&+ \sum_{j=1}^{p} \sum_{j'<j} \zeta_{j,i}\zeta_{j',i}\mathcal{I}^{dd}_{jj'} + \varepsilon_i,
\end{aligned} \tag{2.1}
$$

where μ is the overall mean, q is the number of nongenetic covariates, p is the number of SNPs, $x_{k,i}$ is the kth covariate for subject $i, k = 1, \ldots, q, i = 1, \ldots, n$, which could be either discrete or continuous, α_k is the effect of the kth covariate, a_j and d_j are the additive effect and dominant effect of the jth SNP, respectively, for $j = 1, \ldots, p$, $\mathcal{I}^{aa}_{jj'}$ is the additive $\times$ additive epistatic effect between the jth SNP and the j'th SNP, $\mathcal{I}^{ad}_{jj'}$, $\mathcal{I}^{da}_{jj}$ and $\mathcal{I}^{dd}_{jj'}$ are additive $\times$ dominant epistatic effect, dominant $\times$ additive epistatic effect and dominant $\times$ dominant epistatic effect, and ε_i is the residual error assumed to follow a $N(0, \sigma^2)$ distribution. If an effect is nonzero in the regression model (2.1), we say that the corresponding covariate or interaction is active. For subject i, $\xi_{j,i}$ and $\zeta_{j,i}$ are the indicators of the additive and dominant effects of the jth SNP, respectively, which are defined as

$$
\xi_{j,i} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } AA, \\ 0, & \text{if the genotype of SNP } j \text{ is } Aa, \\ -1, & \text{if the genotype of SNP } j \text{ is } aa, \end{cases}
$$

$$
\zeta_{j,i} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } Aa, \\ 0, & \text{if the genotype of SNP } j \text{ is } AA \text{ or } aa. \end{cases}
$$

Therefore, the additive effect a_j in model (2.1) measures the change of the average phenotypic value by substituting allele A with allele a in a population. Dominant effect d_j, on the other hand, represents how the effect of allele A is modified by the presence of allele a, allowing a more general nonadditive genetic model.

Given observed phenotypic traits, genetic information and covariates such as gender or age, our goal is to characterize the genetic control of the phenotype, by selecting active SNPs and gene–gene interactions and estimating their genetic effects. However, since in GWAS data sets, the number of SNPs usually far exceeds the number of subjects, it is almost impossible to directly estimate all genetic effects, as even epistatic interactions are not considered in the regression model. Recently, penalized regressions that regularize the size of regression coefficients are applied to GWAS models without interactions, and appropriate algorithms are

designed for high-dimensional inference, such as cyclical coordinate descent methods. But in many clinical trials where the number of SNPs is extremely large compared with the sample size, the empirical performance of penalized regression is not guaranteed. Moreover, if four interaction terms for each SNP pair are considered in the GWAS analysis, the estimation of all genetic effects in the ultrahigh-dimensional setting is infeasible from the perspective of both statistical theories and computational cost.

To identify this ultrahigh-dimensional model in practice, and to make the best use of GWAS data for better explanation and predictions, we need to put assumptions on the heredity structures of epistatic effects, although we want to make the restrictions as weak as possible. Two versions of the effect heredity principle are the following: strong heredity and weak heredity [Chipman (1996)]. Under strong heredity, if the interaction between two predictors is significant, both predictors should be marginally significant. Under weak heredity, only one needs to be marginally significant.

Obviously, in prevailing penalized regression models for GWAS, where interaction effects are tested after a subset of significant SNPs are selected, strong heredity assumption is implicitly imposed. However, throughout this paper we will assume only weak heredity, since, in practice, many important SNPs are marginally uncorrelated with the response, but interact with other SNPs in an epistasis network. With this biologically meaningful assumption in the epistatic GWAS model as well as large data sets collected in genome-wide studies, the potential of GWAS could be fully explored, and a detailed picture of genetic control and regulation could be unveiled.

Two SNPs involved in a two-way interaction will be denoted as "two roots." We will employ a two-stage sure independence screening (TS-SIS) procedure to identify SNPs which may have active main effects or may act as roots. Sure independence screening is a statistical learning technique for ultrahigh-dimensional data proposed by Fan and Lv (2008). In the context of GWAS analysis, it ranks the importance of SNPs according to their marginal correlations with the response and retains those SNPs whose marginal correlations are strong enough. It can be shown that under some technical conditions, sure independence screening enjoys the sure screening property. That is, the reduced model is capable of retaining all the active SNPs with asymptotic probability one.

Let $\mathcal{D}_a$ and $\mathcal{D}_d$ be two sets of indices of truly important additive effects and truly important dominant effects, respectively. The first SIS round will be performed between each SNP and the response to select active main effects. Since it is common practice to include covariates as linear predictors of the response in GWAS analysis, covariates are not subject to SIS and will later be added to the reduced model after TS-SIS. After the first stage of SIS, two subsets of SNPs with potential nonzero additive effects $\widehat{\mathcal{D}}_a$ and potential nonzero dominant effects $\widehat{\mathcal{D}}_d$ are selected. Sure screening property [Fan and Lv (2008)] implies that truly important main effects are retained in $\widehat{\mathcal{D}}_a$ and $\widehat{\mathcal{D}}_d$ with high probabilities.

Next, we formulate pairwise epistatic interactions between all SNPs in $\widehat{\mathcal{D}}_a$ or $\widehat{\mathcal{D}}_d$ and all genome-wide SNPs. In particular, an additive×additive interaction term is formulated by taking one SNP from $\widehat{\mathcal{D}}_a$ and taking any additive effect from all SNPs. The set of additive × additive interactions are denoted by $\mathcal{D}_{aa}^{(0)} = \{(j, j') : \xi_j \xi_{j'}, j \in \widehat{\mathcal{D}}_a, j' = 1, 2, \ldots, p\}$. Similarly, additive × dominant interactions $\mathcal{D}_{ad}^{(0)}$, dominant × additive interactions $\mathcal{D}_{da}^{(0)}$, and dominant × dominant interactions $\mathcal{D}_{dd}^{(0)}$ are formulated, and the GWAS model becomes

$$
\begin{aligned}
y_i = \mu &+ \sum_{k=1}^{q} x_{k,i}\alpha_k + \sum_{j\in\widehat{\mathcal{D}}_a} \xi_{j,i} a_j + \sum_{j\in\widehat{\mathcal{D}}_d} \zeta_{j,i} d_j + \sum_{(j,j')\in\mathcal{D}_{aa}^{(0)}} \xi_{j,i}\xi_{j',i}\mathcal{I}_{jj'}^{aa} \\
&+ \sum_{(j,j')\in\mathcal{D}_{ad}^{(0)}} \xi_{j,i}\zeta_{j',i}\mathcal{I}_{jj'}^{ad} + \sum_{(j,j')\in\mathcal{D}_{da}^{(0)}} \zeta_{j,i}\xi_{j',i}\mathcal{I}_{jj'}^{da} \\
&+ \sum_{(j,j')\in\mathcal{D}_{dd}^{(0)}} \zeta_{j,i}\zeta_{j',i}\mathcal{I}_{jj'}^{dd} + \varepsilon_i.
\end{aligned} \tag{2.2}
$$

After adding interaction terms in the model (2.2), the model dimensionality becomes extremely high compared with GWAS model without epistatic interactions. To test whether these interactions contribute to the observed variation in phenotypes, we again apply SIS to all interaction terms and select epistatic effects that are highly correlated with the response. Let $\widehat{\mathcal{D}}_{aa}$ be the index set for the selected additive × additive interactions between a SNP in $\widehat{\mathcal{D}}_a$ and another genome-wide SNP. Similarly, we define three other sets, $\widehat{\mathcal{D}}_{ad}$, $\widehat{\mathcal{D}}_{da}$ and $\widehat{\mathcal{D}}_{dd}$, which contain selected additive × dominant, dominant × additive and dominant×dominant interactions, respectively. Then the GWAS model after TS-SIS becomes

$$
\begin{aligned}
y_i = \mu &+ \sum_{k=1}^{q} x_{k,i}\alpha_k + \sum_{j\in\widehat{\mathcal{D}}_a} \xi_{j,i} a_j + \sum_{j\in\widehat{\mathcal{D}}_d} \zeta_{j,i} d_j + \sum_{(j,j')\in\widehat{\mathcal{D}}_{aa}} \xi_{j,i}\xi_{j',i}\mathcal{I}_{jj'}^{aa} \\
&+ \sum_{(j,j')\in\widehat{\mathcal{D}}_{ad}} \xi_{j,i}\zeta_{j',i}\mathcal{I}_{jj'}^{ad} + \sum_{(j,j')\in\widehat{\mathcal{D}}_{da}} \zeta_{j,i}\xi_{j',i}\mathcal{I}_{jj'}^{da} \\
&+ \sum_{(j,j')\in\widehat{\mathcal{D}}_{dd}} \zeta_{j,i}\zeta_{j',i}\mathcal{I}_{jj'}^{dd} + \varepsilon_i.
\end{aligned} \tag{2.3}
$$

Algorithm 1 summarizes the TS-SIS procedure, where the sizes of the reduced models in steps 1 and 3 will be determined by the RATE approach proposed in Section 3.

REMARK. If some nongenetic covariates (such as age) are known as truly significant predictors in model (2.1), the following modified independence screening procedure can be implemented to improve the performance in steps 1 and 3 of

Algorithm 1 Two-stage sure independence screening

Step 1. Apply the SIS approach to all additive and dominate main effects SNPs and estimate the reduced models $\widehat{\mathcal{D}}_a$ and $\widehat{\mathcal{D}}_d$.

Step 2. Formulate pairwise epistatic interactions between all SNPs selected in $\widehat{\mathcal{D}}_a$ or $\widehat{\mathcal{D}}_d$ and all genome-wide SNPs $\mathcal{D} = \{1, 2, \ldots, p\}$. That is, $\mathcal{D}_{aa}^{(0)} = \{(j, j') : \xi_j \xi_{j'}, j \in \widehat{\mathcal{D}}_a, j' \in \mathcal{D}\}$, $\mathcal{D}_{ad}^{(0)} = \{(j, j') : \xi_j \zeta_{j'}, j \in \widehat{\mathcal{D}}_a, j' \in \mathcal{D}\}$, $\mathcal{D}_{da}^{(0)} = \{(j, j') : \zeta_j \xi_{j'}, j \in \widehat{\mathcal{D}}_d, j' \in \mathcal{D}\}$, and $\mathcal{D}_{dd}^{(0)} = \{(j, j') : \zeta_j \zeta_{j'}, j \in \widehat{\mathcal{D}}_d, j' \in \mathcal{D}\}$.

Step 3. Apply the SIS approach again to all epistatic interactions in step 2, that is, $\mathcal{D}_{aa}^{(0)}$, $\mathcal{D}_{ad}^{(0)}$, $\mathcal{D}_{da}^{(0)}$ and $\mathcal{D}_{dd}^{(0)}$, and obtain the reduced models $\widehat{\mathcal{D}}_{aa}$, $\widehat{\mathcal{D}}_{ad}$, $\widehat{\mathcal{D}}_{da}$ and $\widehat{\mathcal{D}}_{dd}$.

Step 4. Combine all reduced models in steps 1 and 3 to obtain the final selected model by the TS-SIS procedure: $\{\widehat{\mathcal{D}}_a, \widehat{\mathcal{D}}_d, \widehat{\mathcal{D}}_{aa}, \widehat{\mathcal{D}}_{ad}, \widehat{\mathcal{D}}_{da}, \widehat{\mathcal{D}}_{dd}\}$.

Algorithm 1. We run a linear regression of the response on each SNP and the significant nongenetic covariates, and utilize the magnitude of the SNP's estimated coefficient as a marginal screening utility.[3]

3. Rates adjusted thresholding estimation. In this section we propose a general rule to determine the size of the reduced model selected by an independence screening procedure. This rule can be applied to other independence screening methods. In its application to the proposed TS-SIS, it is equivalent to determining the cardinalities of sets $\widehat{\mathcal{D}}_a$, $\widehat{\mathcal{D}}_d$, $\widehat{\mathcal{D}}_{aa}$, $\widehat{\mathcal{D}}_{ad}$, $\widehat{\mathcal{D}}_{da}$ and $\widehat{\mathcal{D}}_{dd}$.

In general, the choice of the reduced model size is critical for any independence screening approach. If the model size is too large, the following penalized regression would be less efficient due to the presence of too many noise variables. If the model size is too small, on the other hand, it is likely to miss important predictors in the screening stage. Fan and Lv (2008) suggested the reduced model size being proportional to $[n/\log n]$ for the SIS procedure, where n is the sample size and $[\cdot]$ denotes the integer of a real number. Although this hard thresholding is easy to implement in practice, little theoretical evidence is provided to guarantee its performance in different data sets. Zhu et al. (2011) proposed a soft-thresholding rule by adding auxiliary variables in their Sure Independent Ranking and Screening (SIRS) procedure for multi-index models with ultrahigh-dimensional covariates. In what follows we propose a general data-driven procedure to determine the reduced model size that extends the soft-thresholding procedure.

Denote the p_n-dimensional vector of predictors by $\mathbf{x} = (X_1, \ldots, X_{p_n})$, and denote the vector of regression coefficients by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_n})$ in a linear regression model. Let $\mathcal{M}$ be the set of active predictors and $\mathcal{M}^c$ be its complement. That is, $\mathcal{M} = \{1 \le j \le p_n : \gamma_j \ne 0\}$ and $\mathcal{M}^c = \{1 \le j \le p_n : \gamma_j = 0\}$. The idea of the

[3]We thank the Associate Editor for suggesting this modified independence screening.

soft-thresholding rule in Zhu et al. (2011) is as follows. First, d auxiliary variables are generated independently and randomly $\mathbf{z} = (Z_1, \ldots, Z_d) \sim N_d(\mathbf{0}, \mathbf{I}_d)$. Next, an independence screening procedure is applied to the combined predictors set $(\mathbf{x}^{\mathrm{T}}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$. Let ρ_k be the marginal screening utility between each predictor and the response, where $k = 1, 2, \ldots, (p_n + d)$. Because $\mathbf{z}$ is known to be independent of the response, the marginal utility ρ_k between any Z_k and the response is exactly zero and the associated sample version $\hat{\rho}_k$ should be less than any marginal utility between the active predictors and the response. Zhu et al. (2011) suggested the maximal sample marginal utility of all auxiliary variables, $C_d = \max_{1 \le m \le d} \hat{\rho}_{p_n+m}$, as a natural cutoff to separate two sets of active and inactive predictors in $\mathbf{x}$. Thus, the selected model is determined by $\widehat{\mathcal{M}} = \{1 \le j \le p_n : \hat{\rho}_j > C_d\}$.

Although the soft-thresholding procedure may be useful, there are two major concerns of practical interest. The first is the choice of the number of auxiliary variables d. The larger the d value, the sparser the selected model, and thus the higher the probability of missing some active predictors. Besides, a larger d value implies more computation cost. On the other hand, a smaller d value gives a smaller cutoff, thus the reduced model dimensionality could still be very high. The second concern is how to generate independent auxiliary variables $\mathbf{z}$. The performance of the soft-thresholding rule depends on an exchangeability assumption between inactive predictors and auxiliary variables assumed in Theorem 3 of Zhu et al. (2011). But its validity is difficult to check in practice.[4] To address these concerns, we propose a rates adjusted thresholding estimation (RATE) approach to determine the number of auxiliary variables d by bootstrapping auxiliary variables from the original data.

In particular, we propose to relate the number of auxiliary variables d to the false positive rate of an independence screening procedure

$$\frac{|\widehat{\mathcal{M}} \cap \mathcal{M}^c|}{|\mathcal{M}^c|},$$

which is the proportion of inactive predictors that are incorrectly included in the selected model $\widehat{\mathcal{M}}$. In data mining and bioinformatics, statistical power is also known as sensitivity, and false positive rate is one minus specificity. Both sensitivity and specificity are performance measures of interest in genetic association studies [see, e.g., Duggal et al. (2008); Gorlov et al. (2008); Harley et al. (2008); Jacobs et al. (2009)].

The next theorem provides a lower bound on the probability that the false positive rate is controlled under a pre-specified level α.

THEOREM 1. *Suppose that the inactive variables* $\{X_j : j \in \mathcal{M}^c\}$ *and auxiliary variables* $\{Z_k : k = 1, \ldots, d\}$ *are exchangeable in the sense that the inac-*

[4]For instance, both the Editor and Associate Editor mentioned that the distribution of "noisy" SNPs is quite different from a normal distribution, so the exchangeability assumption may be violated. We thank the Editor and Associate Editor for pointing this out.

Algorithm 2 Rates adjusted thresholding estimation

Step 1. Solve the equation $\{1 - \frac{\alpha(p_n-n)}{p_n+d}\}^d = \beta$ to obtain d, for given p_n, n, α and β.

Step 2. Bootstrap the original data $\mathbf{x} = (X_1, \ldots, X_{p_n})$ to obtain d independent auxiliary variables $\mathbf{z} = (Z_1, \ldots, Z_d)$ in the following way. For each $i = 1, \ldots, n$, randomly assign one value of $(X_{1k}, \ldots, X_{(i-1)k}, X_{(i+1)k}, \ldots, X_{nk})$ to Z_{ik}, and then get the vector $Z_k = (Z_{1k}, \ldots, Z_{nk})$ for $k = 1, \ldots, d$. If $d > p_n$, one may iterate the above procedure until getting enough auxiliary variables.

Step 3. Compute the marginal screening utility $\hat{\rho}_k^*$ between each auxiliary variable Z_k and the response, $k = 1, \ldots, d$, and set the cutoff $C_d = \max_{1 \le k \le d} \hat{\rho}_k^*$.

Step 4. Compute the marginal screening utility $\hat{\rho}_j$ between each predictor X_j and the response, $j = 1, \ldots, p_n$, and select the reduced model as $\widehat{\mathcal{M}} = \{1 \le j \le p_n : \hat{\rho}_j > C_d\}$.

tive and auxiliary variables are equally likely to be selected by the independence screening procedure. Under the sparsity condition that $s_n < n$, the probability that the false positive rate can be controlled under a pre-specified level α is bounded from below. That is,

$$P\left(\frac{|\widehat{\mathcal{M}} \cap \mathcal{M}^c|}{|\mathcal{M}^c|} < \alpha\right) \ge 1 - \left\{1 - \frac{\alpha(p_n - n)}{p_n + d}\right\}^d. \tag{3.1}$$

The theorem implies that the probability of the false positive rate being controlled below a given level α is greater than $1 - \{1 - \frac{\alpha(p_n-n)}{p_n+d}\}^d$. Given a fixed confidence level $1 - \beta = 1 - \{1 - \frac{\alpha(p_n-n)}{p_n+d}\}^d$, the number of auxiliary variables d can be determined. According to Theorem 1, we propose the RATE procedure in Algorithm 2 for a general independence screening method.

We remark that the modified bootstrapping procedure in step 2 in Algorithm 2 is to guarantee the independence between the response and auxiliary variables $\mathbf{z}$. We obtain independent auxiliary variables by bootstrapping the original data instead of simulating them from a normal distribution. Consequently, the bootstrapped auxiliary variables have the same data structure as the original predictors, approximating the exchangeability condition in the soft-thresholding rule. Note that with given p_n and n, two rates α and β together determine the number of auxiliary variables d. Therefore, we call this approach the rates adjusted thresholding estimation (RATE). It will be shown later that the RATE approach has excellent performance in the simulation studies and the real data analysis.

4. SCAD penalized regression. After two-stage sure independence screening, the dimensionality of the GWAS model is greatly reduced. In order to precisely select important SNPs and epistatic interactions from a pool of candidate effects, penalized regressions widely used in main-effect analysis could be incor-

porated here. Specifically, we put penalties on the sizes of additive effects, dominant effects and all epistatic effects and minimize the following penalized least squares:

$$\begin{aligned}\frac{1}{2n}\|\mathbf{y}-E\mathbf{y}\|^2 &+ \sum_{j\in\widehat{\mathcal{D}}_a} p_\lambda(|a_j|) + \sum_{j\in\widehat{\mathcal{D}}_d} p_\lambda(|d_j|) + \sum_{(j,j')\in\widehat{\mathcal{D}}_{aa}} p_\lambda(|\mathcal{I}^{aa}_{jj'}|)\\ &+ \sum_{(j,j')\in\widehat{\mathcal{D}}_{ad}} p_\lambda(|\mathcal{I}^{ad}_{jj'}|) + \sum_{(j,j')\in\widehat{\mathcal{D}}_{da}} p_\lambda(|\mathcal{I}^{da}_{jj'}|) + \sum_{(j,j')\in\widehat{\mathcal{D}}_{dd}} p_\lambda(|\mathcal{I}^{dd}_{jj'}|),\end{aligned} \tag{4.1}$$

where the penalty function $p_\lambda(\cdot)$ is implemented to shrink sufficiently small effects to zero and thus exclude the inactive predictors.

We consider the smoothly clipped absolute deviation (SCAD) penalty function due to its unbiasedness, continuity and sparsity properties [Fan and Li (2001)]. The SCAD penalty is a nonconvex function and defined as follows:

$$\begin{aligned}p_\lambda(b) = \lambda|b|I\big(0\le|b|<\lambda\big) &+ \frac{a\lambda|b|-(b^2+\lambda^2)/2}{a-1}I\big(\lambda\le|b|\le a\lambda\big)\\ &+ \frac{(a+1)\lambda^2}{2}I\big(|b|>a\lambda\big),\end{aligned}$$

where $I(\cdot)$ is an indicator function and $a=3.7$ as suggested in Fan and Li (2001). λ is the tuning parameter which balances the model complexity and forecasting performance. We follow the idea of Wang, Li and Tsai (2007) and choose λ by a BIC tuning parameter selector.

Commonly-used algorithms for the SCAD penalized least squares include the local quadratic approximation (LQA) algorithm [Fan and Li (2001)], the perturbed LQA [Hunter and Li (2005)] and the local linear approximation (LLA) [Zou and Li (2008)] algorithm. With the aid of LLA, one may employ the LARS algorithm to obtain the SCAD estimate. Thus, we will use the LLA algorithm in this paper. Specifically, for a given initial value $\beta^{(0)}$, the penalty function $p_\lambda(\cdot)$ can be locally approximated by a linear function as

$$p_\lambda(|\beta|)\approx p_\lambda\big(|\beta^{(0)}|\big)+p'_\lambda\big(|\beta^{(0)}|\big)\big(|\beta|-|\beta^{(0)}|\big)\qquad\text{for } |\beta|\approx|\beta^{(0)}|. \tag{4.2}$$

With the aid of LLA, the estimates of regression coefficients in SCAD penalized least squares (4.1) can be obtained by minimizing

$$\begin{aligned}\frac{1}{2n}\|\mathbf{y}-E\mathbf{y}\|^2 &+ \sum_{j\in\widehat{\mathcal{D}}_a} p'_\lambda\big(|a_j^{(0)}|\big)|a_j| + \sum_{j\in\widehat{\mathcal{D}}_d} p'_\lambda\big(|d_j^{(0)}|\big)|d_j|\\ &+ \sum_{(j,j')\in\widehat{\mathcal{D}}_{aa}} p'_\lambda\big(|\mathcal{I}^{aa(0)}_{jj'}|\big)\big(|\mathcal{I}^{aa}_{jj'}|\big) + \sum_{(j,j')\in\widehat{\mathcal{D}}_{ad}} p'_\lambda\big(|\mathcal{I}^{ad(0)}_{jj'}|\big)\big(|\mathcal{I}^{ad}_{jj'}|\big)\\ &+ \sum_{(j,j')\in\widehat{\mathcal{D}}_{da}} p'_\lambda\big(|\mathcal{I}^{da(0)}_{jj'}|\big)\big(|\mathcal{I}^{da}_{jj'}|\big) + \sum_{(j,j')\in\widehat{\mathcal{D}}_{dd}} p'_\lambda\big(|\mathcal{I}^{dd(0)}_{jj'}|\big)\big(|\mathcal{I}^{dd}_{jj'}|\big),\end{aligned} \tag{4.3}$$

after constants are discarded. Note that this penalized least squares can be easily minimized based on L_1 penalized regression.

5. Simulated studies. In this section we investigate the GWAS analysis framework consisting of TS-SIS and variable selection through simulation studies. We simulate large data sets where SNPs may have either (a) main effects or (b) interaction effects. Our goal is to identify these active SNPs with high accuracy and low computational cost.

Specifically, genotypes of p SNPs across 23 chromosomes are generated for $n = 500$ subjects. For SNP j of subject i, $j = 1, \ldots, p$, $i = 1, \ldots, n$, its genotype $\xi_{j,i}$ is derived from $u_{j,i}$, where the vector $(u_{1,i}, \ldots, u_{p,i})$ is generated from multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{j,k})_{p\times p}$, $\sigma_{j,k} = \rho^{|j-k|}$ for $\rho = 0.2$, 0.5 or 0.8. Then, we set

$$\xi_{j,i} = \begin{cases} 1, & u_{ij} > c_{1j}, \\ 0, & c_{2j} \le u_{ij} \le c_{1j}, \\ -1, & u_{ij} < c_{2j}, \end{cases}$$

where c_{1j} and c_{2j} determine the minor allele frequency (MAFs). We consider two cases: homogeneous case, MAF = 0.5 for each j, and heterogeneous case, in which the MAF of each SNP is randomly set to 0.5, 0.35 or 0.2 with equal likelihood. Finally, the dominant effect indicator $\zeta_{j,i}$ is derived from $\xi_{j,i}$ by setting $\zeta_{j,i} = 1 - |\xi_{j,i}|$. In total, there are $p = 3948$ SNPs across 23 chromosomes, with the number of SNPs in each chromosome being one percent of that in a real data set we are going to work on.

We put 3 active main effects and 3 active epistatic interactions across the whole genome, whose positions and effect sizes are given in Table 1. Column "Interact with" in Table 1 indicates, out of 3 active main effects, which one the SNP interacts with. When simulating the response variable, we standardize the design matrix columnwisely, such that all columns of the design matrix have the same variance.

TABLE 1
Information of 6 *assumed genetic effects for data simulation*

Chr.	Position	Additive/dominant	Interact with	Effect size
		Main effects		
1	1	Additive	–	1
2	1	Dominant	–	1
3	1	Additive	–	1
		Epistatic interactions		
11	1	Additive	1	1
2	2	Dominant	2	1
12	1	Dominant	2	1

This step makes the comparison of detecting active main effects and active interactions fair. From Table 1, it can be seen that one SNP could interact with two other SNPs without marginal effects (three SNPs on chromosomes 2 and 12), and two SNPs involved in a two-way interaction may also be correlated (two SNPs on chromosomes 2). These interaction patterns add further complexity in the simulation studies.

For each simulated data set, we first implement SIS with the RATE procedure to select s_1 SNPs, which may exhibit notable main effects or epistatic effects. We determine s_1 in each simulation according to Theorem 1 with $\alpha = 0.01$ and $\beta = 0.0001$ and, on average, there are 11 SNPs selected in the first stage of TS-SIS. According to the sure screening property, this subset of s_1 SNPs should include the first SNPs on the first 3 chromosomes with high probability, which demonstrate active main effects and may serve as roots in two-way interactions.

To select those SNPs that have no marginal effects, but modify the genetic effects of other SNPs, two-way interactions are formed between each selected SNP in the first stage and any SNP across the genome according to model (2). SIS is carried out again, and, in total, s_2 pairs of SNPs are selected. We set $\alpha = 0.005$ and $\beta = 0.0001$. These SNP pairs should contain all epistatic interactions, although they may rank low in terms of the absolute value of marginal correlations. Finally, s_1 SNPs with potential main effects and s_2 SNP pairs enter model (2.3), and variable selections are implemented to select important SNPs and estimate their main effects and epistatic effects. We consider both LASSO regression and SCAD regression following TS-SIS.

Table 2 reports the statistical power, false positive rates and computational time using R code. The result is the average over 100 simulations with standard error in parenthesis. In columns labeled "TS-SIS" under "Power (%)," we present the statistical power of TS-SIS, or the proportion of active SNPs and interactions that are successfully included in the candidate pool of s_1 SNPs and s_2 interactions. In adjacent columns "TS-SIS-SCAD" and "TS-SIS-LASSO," we report statistical powers of two-stage SIS paired with SCAD regression or LASSO regression, or the percent of 6 active SNPs that are correctly identified by the whole procedure. Note that the statistical power under "TS-SIS-SCAD" or "TS-SIS-LASSO" cannot be greater than that of TS-SIS, since a SNP or an epistatic interaction is considered by the variable selection procedure only if it is correctly identified by TS-SIS. In each column under "False Positive Rate ($\times 10^{-4}$)," we report the false positive rate defined as the proportion of unimportant SNPs that are incorrectly identified. We also report the median computing time for TS-SIS with penalized regression over all replications. The simulation is conducted on a 32-bit windows 7 system, with an Intel (R) i5-2400 processor, 3.10 GHz, 4G memory.

According to Table 2, the TS-SIS captures most of the SNPs with active main effects, as well as SNPs without main effects but demonstrating active interactions. As a result, important SNPs are selected in the reduced model, and the majority of irrelevant SNPs are eliminated before variable selection. This critical step greatly

TABLE 2
Statistical power, false positive rate and running time of the proposed TS-SIS approach

	Power (%)			False positive rate ($\times 10^{-4}$)			
(ρ, σ^2)	TS-SIS	TS-SIS-SCAD	TS-SIS-LASSO	TS-SIS	TS-SIS-SCAD	TS-SIS-LASSO	Time (seconds)
			Homogeneous case ($MAF = 0.50$)				
(0.8, 6)	97.0	92.7	86.5	4.22	2.18	2.92	12.29
	(8.3)	(11.7)	(13.3)	(2.35)	(1.09)	(1.47)	
(0.8, 8)	95.5	86.2	82.3	4.52	2.42	3.19	13.64
	(9.1)	(13.4)	(15.9)	(2.51)	(1.23)	(1.68)	
(0.5, 6)	97.8	94.5	88.2	2.93	1.74	2.29	10.05
	(6.6)	(9.5)	(13.2)	(1.58)	(0.88)	(1.16)	
(0.5, 8)	95.8	90.7	88.5	2.76	1.72	2.21	11.13
	(8.3)	(13.0)	(11.6)	(1.69)	(1.01)	(1.33)	
(0.2, 6)	95.9	91.8	87.7	2.89	1.75	2.27	7.88
	(9.0)	(11.7)	(13.4)	(1.88)	(1.05)	(1.43)	
(0.2, 8)	95.5	87.8	87.1	2.70	1.81	2.28	9.12
	(9.7)	(13.6)	(13.1)	(1.71)	(1.08)	(1.42)	
			Heterogeneous case (*mixed MAFs*)				
(0.8, 6)	98.17	89.83	89.50	4.71	2.30	3.18	12.41
	(5.24)	(11.58)	(11.28)	(3.51)	(1.30)	(1.99)	
(0.8, 8)	96.50	85.50	88.33	5.02	2.46	3.42	13.15
	(7.22)	(12.46)	(11.73)	(4.17)	(1.49)	(2.11)	
(0.5, 6)	98.17	91.50	90.67	3.47	1.87	2.45	11.17
	(5.24)	(11.48)	(11.68)	(2.93)	(1.26)	(1.78)	
(0.5, 8)	95.00	87.67	88.33	3.08	1.82	2.42	10.38
	(9.91)	(13.94)	(14.11)	(2.36)	(1.12)	(1.75)	
(0.2, 6)	97.67	90.17	90.83	3.08	1.80	2.37	10.59
	(6.28)	(12.10)	(11.93)	(2.15)	(1.10)	(1.54)	
(0.2, 8)	94.33	87.50	89.33	2.58	1.53	2.05	9.25
	(9.54)	(13.06)	(11.24)	(2.33)	(0.99)	(1.61)	

improves the probability of effectively identifying important SNPs and interactions in GWAS analysis. After TS-SIS, SNPs and interactions in the reduced model are selected by either SCAD or LASSO. As expected, variable selection further reduces the false positive rate and increases the interpretability of the final model. Compared with LASSO, SCAD can identify truly important SNPs with higher probability for the homogeneous case. As more SNPs have lower MAFs in the heterogeneous case, two penalized regressions have comparable statistical powers. In addition, SCAD delivers smaller false positive rates consistently in all simulation scenarios. Table 2 further suggests that, as σ^2 decreases, the statistical powers increase, but the linkage disequilibrium of two SNPs measured by ρ plays a lim-

TABLE 3
Statistical power of detecting interactions of the proposed TS-SIS approach

	Homogeneous case			Heterogeneous case		
(ρ, σ^2)	**TS-SIS**	**TS-SIS-SCAD**	**TS-SIS-LASSO**	**TS-SIS**	**TS-SIS-SCAD**	**TS-SIS-LASSO**
(0.8, 6)	93.3	93.3	92.7	97.0	97.0	97.0
	(13.4)	(13.4)	(14.7)	(9.6)	(9.6)	(9.6)
(0.8, 8)	96.7	96.3	96.3	94.7	94.3	94.7
	(10.1)	(10.5)	(10.5)	(13.2)	(14.3)	(13.2)
(0.5, 6)	96.7	96.7	96.7	96.7	96.7	96.7
	(10.1)	(10.1)	(10.1)	(10.1)	(10.1)	(10.1)
(0.5, 8)	92.6	92.6	92.6	91.0	91.0	91.0
	(15.1)	(15.1)	(15.1)	(18.3)	(18.3)	(18.3)
(0.2, 6)	93.5	93.5	93.5	96.0	96.0	96.0
	(13.3)	(13.3)	(13.3)	(10.9)	(10.9)	(10.9)
(0.2, 8)	92.7	92.7	92.7	91.0	91.0	91.0
	(16.1)	(16.1)	(16.1)	(17.0)	(17.0)	(17.0)

ited role in this setting. Besides, this variable screening procedure is very fast even though millions of potential pairwise interactions are present in each simulation.

Table 3 gives the statistical power of detecting interactions, or the average proportion of interactions that are selected over 100 simulations. By comparing Table 3 with Table 2, it can be seen that interactions are relatively more difficult to capture by variable screenings than main effects. This is understandable since the number of interaction terms is huge compared with the number of main effect terms. Once important interactions are identified by TS-SIS, however, they are unlikely to be missed by the following penalized regression. As a result, the statistical power of the entire procedure is very close to that of TS-SIS. In Table 4 we report the results when the number of SNPs is doubled ($p = 6996$) for MAF $= 0.50$, with all other specifications unchanged. Interestingly, although the statistical power of TS-SIS increases, the power of SCAD and LASSO regressions slightly decreases, because the same α and β in Theorem 1 imply a larger reduced model from TS-SIS.[5] However, given that the number of interactions increases from about 24.5 million to about 98 million, the performance of TS-SIS is excellent, as can be seen from the increased statistical power and decreased false positive rates. In Table 4 we do not change α and β for comparison purposes; we consider in future research the effects of user-specified rates.

We also compare this framework with other methods for detecting SNP–SNP interactions in simulation studies with MAF $= 0.5$. Although most of the avail-

[5]On average, the total number of main effects and interactions selected by TS-SIS increases from 46.9 to 66.2.

TABLE 4
Statistical power, false positive rate and running time of the proposed TS-SIS approach when the number of SNPs is doubled ($p = 6996$) for MAF = 0.5

	Power (%)			False positive rate ($\times 10^{-4}$)			
(ρ, σ^2)	TS-SIS	TS-SIS-SCAD	TS-SIS-LASSO	TS-SIS	TS-SIS-SCAD	TS-SIS-LASSO	Time (seconds)
(0.8, 6)	99.5	90.0	83.7	1.39	0.74	0.97	39.37
	(3.7)	(12.9)	(14.2)	(0.64)	(0.29)	(0.40)	
(0.8, 8)	97.3	87.8	84.1	1.32	0.78	1.01	44.68
	(7.8)	(14.3)	(12.7)	(0.60)	(0.29)	(0.39)	
(0.5, 6)	97.3	89.8	85.0	1.15	0.68	0.89	29.90
	(7.0)	(11.8)	(14.3)	(0.58)	(0.32)	(0.41)	
(0.5, 8)	97.1	86.3	83.7	1.17	0.74	0.95	36.56
	(8.2)	(14.8)	(14.9)	(0.53)	(0.33)	(0.42)	
(0.2, 6)	98.8	93.8	87.2	1.06	0.67	0.86	25.97
	(4.3)	(10.2)	(12.5)	(0.47)	(0.28)	(0.36)	
(0.2, 8)	94.7	80.7	84.0	1.18	0.77	0.99	33.52
	(9.7)	(15.1)	(13.8)	(0.52)	(0.29)	(0.38)	

able interaction detection methods are designed for binary phenotypes, the Mendel software program [Lange et al. (2001, 2013)] and the Screen and Clean (SC) method [Wu et al. (2010)] can identify important SNPs as well as interactions in GWAS analysis for the quantitative phenotype. Moreover, they are scalable and computationally efficient. Specifically, Analysis Option 24 in Mendel software is very convenient to test for main genetic effects and interaction effects based on marginal p-values or LASSO type analysis [Wu and Lange (2008); Wu et al. (2009); Zhou et al. (2010)]. Table 5 reports the results from four major analysis options of Mendel: (1) marginal analysis for main effects followed by testing important marginal effects against all SNPs for interactions (Mendel 1), (2) marginal analysis for main effects followed by testing all pairwise interactions among top SNPs (Mendel 2), (3) LASSO analysis for main effects followed by testing important marginal effects against all SNPs for interactions (Mendel 3), and (4) LASSO analysis for main effects followed by testing all pairwise interactions among top SNPs (Mendel 4). Since these four analysis options generate final models with pre-determined sizes, we use the default model size of 10 for main effects and then determine the number of selected interactions in a way that the final model size is the same as our method (TS-SIS-SCAD). Table 5 also reports the performance of the Screen and Clean (SC) method (column "SC") and hard-thresholding-based TS-SIS (column "Hard-SCAD" and column "Hard-LASSO"), where the first $[n/\log n]$ SNPs are selected in TS-SIS. Since the final model size of Mendel is user specified, the false positive rate is not reported.

TABLE 5
Statistical power and false positive rate of alternative methods

	Power (%)							FPR ($\times 10^{-4}$)		
(ρ, σ^2)	Hard-SCAD	Hard-LASSO	SC	Mendel 1	Mendel 2	Mendel 3	Mendel 4	Hard-SCAD	Hard-LASSO	SC
(0.8, 6)	89.3	81.0	69.9	85.3	49.0	89.3	49.3	4.0	4.4	4.8
	(9.3)	(8.5)	(10.5)	(8.7)	(4.0)	(10)	(3.3)	(0.4)	(0.4)	(2.7)
(0.8, 8)	81.5	80.5	61.9	85.7	49.7	88.0	49.0	4.2	4.6	5.3
	(11.3)	(11.9)	(12.0)	(9.5)	(2.4)	(10.1)	(4.0)	(0.3)	(0.4)	(3.2)
(0.5, 6)	87.5	81.7	68.8	83.3	50.0	86.7	49.7	4.5	4.8	5.1
	(10.7)	(9.6)	(13.5)	(6.7)	(0)	(9.5)	(2.4)	(0.3)	(0.3)	(2.8)
(0.5, 8)	81	81.3	60.5	83.0	49.0	86.0	49.7	4.6	4.9	5.3
	(11.1)	(12.8)	(11.6)	(7.9)	(4.0)	(10.3)	(2.4)	(0.3)	(0.4)	(3.5)
(0.2, 6)	86.2	81.8	66.8	86.0	49.3	89.3	49.3	4.5	4.9	4.8
	(10.1)	(9.2)	(12.1)	(10.3)	(3.3)	(10.5)	(3.3)	(0.3)	(0.4)	(2.2)
(0.2, 8)	79.8	80.2	63.1	83.3	49.3	85.0	49.7	4.6	5.0	4.2
	(12.6)	(14.9)	(9.2)	(8.9)	(3.3)	(9.7)	(2.4)	(0.3)	(0.3)	(2.3)

Among all alternative approaches, Mendel 3 has the best performance followed by Mendel 1 and hard-thresholding-based approaches. Both Mendels 3 and 1 test the interactions between marginally important SNPs and all SNPs, but Mendel 3 selects marginally important SNPs by LASSO regressions and Mendel 1 is based on the conventional marginal analysis. Mendels 2 and 4 cannot give statistical power greater than 50% since only interactions among top SNPs are considered. In terms of hard-thresholding-based TS-SIS procedures ("Hard-SCAD" and "Hard-LASSO"), their performance is less satisfactory since too many variables retained after variable screening lead to a lower statistical power and an inflated false positive rate. But similar to Table 2, SCAD regression tends to be associated with a higher statistical power and a smaller false positive rate. Last, the Screen and Clean method has a low statistical power and a large and unstable false positive rate.

In summary, TS-SIS guided by the RATE approach is effective and efficient in selecting truly important genetic effects and eliminating false positives for the following penalized regressions. In the context of the ultrahigh-dimensional GWAS model where a huge number of potential predictors are considered, they are recommended in the real data analysis.

6. Framingham data analysis. We use the newly developed framework to analyze a real GWAS data set from the Framingham Heart Study, a cardiovascular study based in Framingham, Massachusetts, supported by the National Heart, Lung, and Blood Institute and Boston University [Dawber, Meadors and Moore (1951)]. Recently, 550,000 SNPs have been genotyped for the entire Framingham

cohort [Jaquish (2007)], from which 977 unrelated subjects including 418 males and 559 females were randomly chosen for our data analysis, conforming to the assumption of population-based GWAS. For each subject, body mass index (BMI) is measured at multiple time points between age 29 and age 61. We take the first measurement for each individual, although the age of receiving the first measurement varies across individuals.

As a common practice in GWAS analysis, SNPs with rare allele frequency $<10\%$ were excluded from data analysis, which leaves 349,985 SNPs across 23 chromosomes of the whole genome. 5.16% of the remaining SNPs, however, contain missing genotypes for some subjects. Since we are interested in detecting active genetic effects rather than handling missing data in this study, for each missing genotype of each subject, we randomly draw a genotype according to the SNP's genotypic frequencies across all subjects whose genotypes are known. Then, by including gender and age as two covariates, we follow the procedure described in previous sections to select SNPs with active main effects and construct an epistatic network explaining the observed BMI variations. In the RATE assisted TS-SIS procedure, in particular, the confidence level is the same as that in simulation studies ($\beta = 0.0001$), but α is set to 0.0005 in screening for main effects and to 0.00001 in detecting interactions.

Out of 349,985 SNPs and numerous two-way interaction terms, 23 active main effects and 24 active epistatic interactions are detected by the TS-SIS procedure followed by SCAD penalized regression. Then, we refit a linear regression model with these selected SNPs and two covariates being predictors, and obtain the estimated regression coefficient and heritability for each selected SNP. Tables 6 and 7

TABLE 6
Information of SNPs with active main effects in the Framingham Heart Study

Additive effects					Dominant effects				
Chr.	**Name**	**MAF**	**Effect**	**Heritability (%)**	**Chr.**	**Name**	**MAF**	**Effect**	**Heritability (%)**
1	ss66041272	0.49	−0.82	1.92	3	ss66173500	0.29	−0.26	0.35
1	ss66276746	0.13	−0.42	0.23	3	ss66142093	0.30	−0.63	2.06
4	ss66346559	0.28	−0.51	0.60	4	ss66354801	0.27	0.30	0.45
4	ss66159949	0.29	0.12	0.03	6	ss66166806	0.34	−0.45	1.09
5	ss66316662	0.38	0.37	0.37	6	ss66299053	0.34	0.41	0.91
5	ss66118377	0.50	0.06	0.01	6	ss66090554	0.27	0.24	0.29
7	ss66083530	0.19	−0.47	0.39	7	ss66083530	0.35	−0.28	0.43
8	ss66177628	0.23	−0.01	0.00	7	ss66249128	0.33	−0.64	2.19
9	ss66095597	0.28	−0.60	0.83	7	ss66314446	0.21	0.62	1.70
12	ss66086159	0.36	−0.45	0.53	8	ss66381612	0.27	−0.32	0.51
21	ss66511535	0.16	−0.44	0.30	11	ss66369823	0.25	0.21	0.21
					13	ss66487154	0.30	0.38	0.75

TABLE 7
Information of SNPs with significant interactions in the Framingham Heart Study

Root 1			Root 2			Effect	Heritability (%)
Chr.	**Name**	**MAF**	**Chr.**	**Name**	**MAF**		
			Additive × additive interactions				
1	ss66041272	0.49	6	ss66061582	0.21	−0.95	2.08
9	ss66095597	0.28	4	ss66151090	0.11	−0.80	0.89
			Additive × dominant interactions				
1	ss66137441	0.49	17	ss66248774	0.49	1.06	1.70
3	ss66081331	0.28	3	ss66142093	0.35	0.56	0.30
3	ss66375852	0.38	23	ss66107600	0.33	−1.13	0.83
4	ss66159949	0.29	11	ss66132273	0.38	0.74	0.71
8	ss66177628	0.23	13	ss66487154	0.34	−1.06	1.32
			Dominant × additive interactions				
3	ss66142093	0.3	2	ss66430035	0.25	−1.00	0.97
3	ss66142093	0.3	3	ss66081331	0.28	−0.74	0.61
3	ss66142093	0.3	3	ss66483001	0.30	−0.33	0.14
4	ss66354801	0.27	7	ss66416257	0.21	−0.72	0.53
6	ss66316737	0.29	21	ss66113670	0.10	1.42	2.75
7	ss66468842	0.33	7	ss66083530	0.19	−0.14	0.03
7	ss66249128	0.33	8	ss66047672	0.23	−0.88	0.79
15	ss66058021	0.38	1	ss66325411	0.17	1.01	0.90
			Dominant × dominant interactions				
3	ss66142093	0.3	4	ss66444506	0.26	1.09	0.89
3	ss66142093	0.3	8	ss66468875	0.39	0.97	0.71
3	ss66142093	0.3	11	ss66152909	0.35	1.08	0.78
7	ss66468842	0.33	11	ss66318229	0.29	1.02	0.68
7	ss66249128	0.33	12	ss66451087	0.14	2.25	2.29
7	ss66249128	0.33	12	ss66109005	0.16	−1.08	0.61
11	ss66369823	0.25	10	ss66482189	0.42	1.23	1.74
11	ss66369823	0.25	3	ss66142093	0.35	−0.40	0.15
18	ss66306728	0.3	16	ss66394113	0.13	1.19	0.55

tabulate the information of selected SNPs with nonzero main and epistatic interaction effects, respectively, including chromosomes, names, minor allele frequencies (MAF), estimated genetic effects and heritabilities. Specifically, heritability is the proportion of the phenotypic variance explained by the genetic variance of a particular effect. For an additive or dominant effect, it is calculated as

$$h^2 = \frac{2p_A p_a(\hat{a}_j + (p_A - p_a)\hat{d}_j)^2 + (2p_A p_a \hat{d}_j)^2}{\mathrm{var}(y)},$$

where p_A is the allele frequency for A and p_a is the allele frequency for a. For the epistatic interactions, the heritability calculation under our general genetic model is more involved. Suppose SNP j has alleles A and a, and SNP j' has alleles B

and b. Then for genotypes $AABB$, $AABb$, $AAbb$, $AaBB$, $AaBb$, $Aabb$, $aaBB$, $aaBb$ and $aabb$, the vector of genotype frequencies is

$$\boldsymbol{\omega} = (p_A^2 p_B^2, 2p_A^2 p_B p_b, p_A^2 p_b^2, 2p_A p_a p_B^2, 4p_A p_a p_B p_b, \\ 2p_A p_a p_b^2, p_a^2 p_B^2, 2p_a^2 p_B p_b, p_a^2 p_b^2)^T,$$

and the associated genetic values are

$$\mathbf{g} = (\hat{a}_j + \hat{a}_{j'} + \widehat{\mathcal{I}}_{jj'}^{aa}, \hat{d}_j + \hat{a}_{j'} + \widehat{\mathcal{I}}_{jj'}^{da}, -\hat{a}_j + \hat{a}_{j'} - \widehat{\mathcal{I}}_{jj'}^{aa} \\ \hat{a}_j + \hat{d}_{j'} + \widehat{\mathcal{I}}_{jj'}^{ad}, \hat{d}_j + \hat{d}_{j'} + \widehat{\mathcal{I}}_{jj'}^{dd}, -\hat{a}_j + \hat{d}_{j'} - \widehat{\mathcal{I}}_{jj'}^{ad} \\ \hat{a}_j + \hat{a}_{j'} - \widehat{\mathcal{I}}_{jj'}^{aa}, \hat{d}_j - \hat{a}_{j'} - \widehat{\mathcal{I}}_{jj'}^{da}, -\hat{a}_j - \hat{a}_{j'} + \widehat{\mathcal{I}}_{jj'}^{aa})^T.$$

Therefore, the genetic variance is $\boldsymbol{\omega}^T \mathbf{g}^2 - (\boldsymbol{\omega}^T \mathbf{g})^2$, and the epistatic variance is this genetic variance minus the genetic variances of two main effects. Finally, the associated epistatic heritability is the epistatic variance divided by the phenotypic variance. If dominant effects are not modeled, this formula gives exactly the same result as the one proposed in Wu and Zhao (2009), where two SNPs' additive effects and their additive × additive interaction are considered.

Generally speaking, main genetic effects contribute to 16.1% of the phenotypic BMI variation, among which 5.2% is due to the additive genetic effects and 10.9% is due to the dominant genetic effects. Epistasis, on the other hand, explains 23.0% of the phenotypic variation. It is worth noting that a few SNPs and interactions demonstrate stronger genetic effects than others. In other words, although the expression of the BMI trait is determined by many SNPs, there exist some SNPs that may be more influential. For example, out of the 23 SNPs exhibiting significant additive or dominant effects, five have heritabilities greater than 1%. This number increases to six for epistatic interactions.

To depict an overall picture of genetic control for BMI by SNP–SNP epistasis, we draw a web of additive × additive, additive × dominant, dominant × additive and dominant × dominant interactions in Figure 1 which shows the genomic distribution of SNPs that interact with each other. From this figure, we obtain the following interesting results: (1) epistasis appears to be distributed randomly throughout the genome, although a few SNPs, such as ss66142093 on chromosome 3 and ss66249128 and ss66468842 on chromosome 7 tend to interact with many other SNPs. (2) Active epistasis may not be due to interactions between two SNPs, both of which display active marginal effects. Of the 24 selected pairs, there are two cases in which both SNPs have active marginal effects and there are 14 cases in which only one SNP has an active marginal effect, whereas the counterpart has none. There are as many as 8 pairs in which no SNP is active for its marginal effect. Notably, the dominant × dominant interaction between SNP ss66249128 on chromosome 7 and SNP ss66451087 on chromosome 12 can explain 2.29% of the BMI variation, although the latter is marginally uncorrelated with BMI. In the

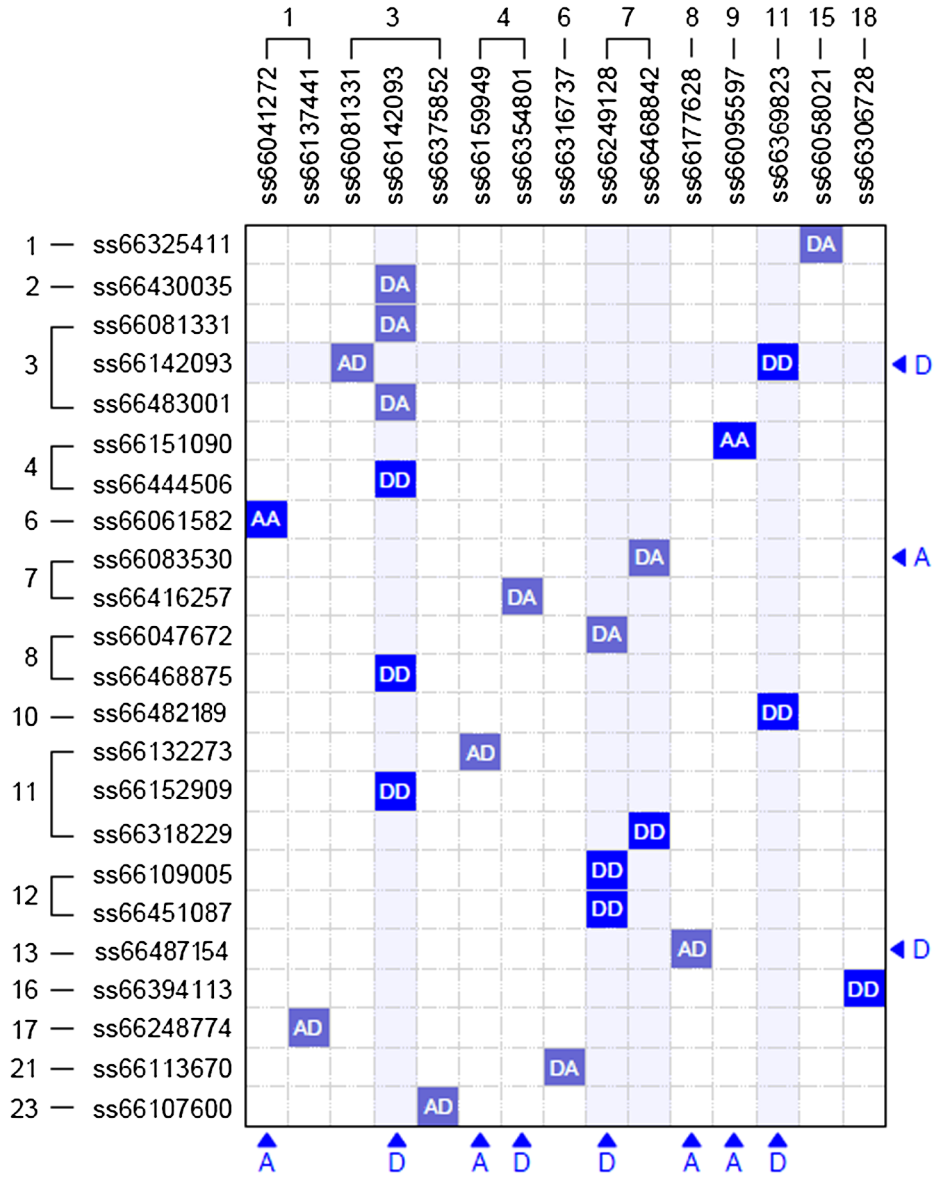

FIG. 1. *A picture of significant SNP–SNP interactions for BMI in the Framingham Heart Study. The numbers beside SNPs are chromosome numbers. The SNPs that display significant additive (A) and dominant (D) effects are indicated by arrows. The pairs of SNPs with significant additive × additive, additive × dominant, dominant × additive and dominant × dominant interactions are indicated as AA, AD, DA and DD, respectively.*

presence of SNP ss66451087, the dominant genetic effect of SNP ss66249128 is dramatically impacted (Table 7).

Since our model allows a large number of SNPs to be analyzed simultaneously, the resulting discoveries should be more biologically relevant and statistically robust than those from traditional single-SNP approaches. For example, SNP ss66142093 on chromosome 3 was detected to explain 2.97% heritability. This SNP is near a candidate gene ANAPC13 involved in pathways for bone and cartilage development that affects human height and stature through cell cycle regulation and mitosis [Weedon and Frayling (2008)].

In other GWAS for BMI [Frayling et al. (2007); Scuteri et al. (2007); Speliotes et al. (2010)], significant SNPs were repeatedly detected on chromosomes 1, 3, 4, 6, 7 and 11 in NEGR1, ETV5, GNPDA2. BDNF and MTCH2 loci. Our results of main genetic effects are in agreement with previous reports about the presence of common variants near these loci associated with biochemical pathways toward obesity. The result on epistatic interactions suggests large epistatic effects among chromosomes 3, 7 and 11 which have not been reported in previous studies, possibly showing the unique power of this new approach. A recent review on identifying genes responsible for type 2 diabetes confirms genomic regions harboring disease susceptibility loci [Frayling et al. (2007)]. These regions include two on chromosome 3 and one on each of chromosomes 4, 6, 9 and 12. We have noticed many SNPs identified in this study overlap with those detected by previous studies targeting type 2 diabetes, suggesting the underlying correlations between BMI and type 2 diabetes. Additionally, our analysis shows that the regression coefficients for gender and age are -0.12 and 0.01, respectively. That is, after adjusting for these genetic factors, the risk of obesity is higher for females, and the risk increases with age.

To further evaluate the significance and predictability of the proposed method, we randomly partition the original real data set into two parts: the training data set with 900 subjects and the validation data set with the remaining 77 individuals. We apply the proposed RATE assisted TS-SIS followed by the SCAD penalized regression to the training data set, and then use the validation data set to evaluate the estimated model. Denote by Y_i^* the response BMI value of the ith subject in the validation data set, and $\widehat{Y}_i^*$ the predicted BMI value by the estimated model using the training data set, where $i = 1, 2, \ldots, 77$. We compute the following two criteria to evaluate the prediction performance. First, we calculate the relative mean absolute prediction error (RMAPE), which is the difference between Y_i^* and $\widehat{Y}_i^*$ divided by the true value of Y_i^*:

$$\text{RMAPE} = \frac{1}{77}\sum_{i=1}^{77}\frac{|Y_i^* - \widehat{Y}_i^*|}{Y_i^*}.$$

Second, we note that a primary interest of predicting BMI is to predict whether the individual is obese or not, that is, BMI$>$ 30. Thus, we compute the classification accuracy (CA) of the validation data set using the estimated model:

$$\text{CA} = 1 - \frac{1}{77}\sum_{i=1}^{77}\big|I\big(Y_i^* > 30\big) - I\big(\widehat{Y}_i^* > 30\big)\big|,$$

where $I(\cdot)$ is an indicator function. Then, we repeat the above validation experiment 10 times. The average RMAPE is 14.10%, and the standard deviation of RMAPE is 0.85%. The average CA is 82.77%, with a standard deviation of 3.37%. These results suggest that our model predicts well in the out-of-sample validation data sets.

7. Discussion. Identifying genetic interaction network is an important task in genome-wide association studies, but is challenged by the sheer volume of genetic data. In this paper we present a comprehensive GWAS model and propose a statistical framework to identify important SNPs and interactions which jointly explain the observed phenotypes. Specifically, a two-stage sure independence screening procedure (TS-SIS) is proposed to formulate a candidate pool of SNPs, including those without weak main effects, but serving as a root in two-way interactions. This procedure expands the literature by relaxing the restrictive assumption that two roots in an interaction have to be marginally correlated with the response. A RATE approach is also proposed to determine the number of predictors retained in each stage of TS-SIS. This approach can also be applied to other variable screening problems.

Wu and Zhao (2009) derived an analytical approach to calculate the power of a model selection strategy in GWAS that is similar to the proposed TS-SIS. Their approach allows for random genotypes, correlation among test statistics as well as a false-positive control. It is straightforward to apply their power calculations to our framework. Since the TS-SIS procedure provided a relatively low-dimensional regression model containing important SNPs with high probability, existing penalized least squares estimations and their empirical performances in GWAS analysis provided valuable guidance for selecting important SNPs and constructing a gene–gene interaction network.

The new model has been used to analyze GWAS data from the Framingham Heart Study [Dawber, Meadors and Moore (1951)], aimed to identify genetic variants that affect cardiovascular diseases and their related traits such as blood pressure and BMI [Jaquish (2007)]. To the best of our knowledge, this is likely the first study that has detected genetic interactions for obesity-related traits in GWAS. Since the detected SNPs displaying important interactions may be harbored in genes of the BMI-associated metabolic pathways [Speliotes et al. (2010)], plus higher heritabilities collectively explained by them, our model should provide a powerful and useful tool for understanding the underlying genetic mechanisms and regulatory network of obesity. For example, dopamine, which is a neurotransmitter, modulates motivation and rewarding properties of eating. Wang et al. (2001) confirmed by biomedical experiments that brain dopamine levels are significantly lower in the obese individuals, suggesting strong correlations between BMI and genetic regulatory networks. The use of our model to detect dopamine-associated SNPs in a GWAS study should help to unravel the genetic architecture of obesity.

Our statistical procedure is capable of identifying epistatic interactions and enables researchers to decipher a detailed picture of the genetic architecture of human diseases or complex traits. So far, we have concentrated on detecting interactions for a continuous trait in GWAS. The proposed TS-SIS assisted SCAD regression can be readily extended to case-control cohorts, family trios or survival data analysis in genome-wide association studies. The framework can also be applied to other statistical problems, where the accurate detection of interactions is desired in the presence of high-dimensional data sets or ultrahigh-dimensional data sets.

APPENDIX

PROOF OF THEOREM 1. Let any $r \in \mathcal{N}_+$, the set of positive integers. The event $\{|\widehat{\mathcal{M}} \cap \mathcal{M}^c| \geq r\}$ represents that at least r unimportant variables rank on the top of all auxiliary variables. Because the inactive variables $\{X_j : j \in \mathcal{M}^c\}$ and auxiliary variables $\{Z_k : k = 1, \ldots, d\}$ are exchangeable, we follow the idea of Zhu et al. (2011) and have that

$$
\begin{aligned}
P\big(|\widehat{\mathcal{M}} \cap \mathcal{M}^c| \geq r\big) &\leq \frac{(p_n - s_n)!}{(p_n - s_n - r)!r!} \Big/ \frac{(p_n - s_n + d)!}{(p_n - s_n + d - r)!r!} \\
&\leq \frac{(p_n - s_n + d - r) \times \cdots \times (p_n - s_n + 1 - r)}{(p_n - s_n + d) \times \cdots \times (p_n - s_n + 1)} \\
&\leq \left(1 - \frac{r}{p_n + d}\right)^d.
\end{aligned} \tag{A.1}
$$

$|\mathcal{M}^c| = p_n - s_n > p_n - n$ by the sparsity principle. If we can assume $|\mathcal{M}| < n$, it follows that

$$
\begin{aligned}
P\left(\frac{|\widehat{\mathcal{M}} \cap \mathcal{M}^c|}{|\mathcal{M}^c|} < \alpha\right) &= 1 - P\big(|\widehat{\mathcal{M}} \cap \mathcal{M}^c| \geq \alpha |\mathcal{M}^c|\big) \\
&\geq 1 - P\big(|\widehat{\mathcal{M}} \cap \mathcal{M}^c| \geq \alpha(p_n - n)\big) \\
&\geq 1 - \left\{1 - \frac{\alpha(p_n - n)}{p_n + d}\right\}^d,
\end{aligned} \tag{A.2}
$$

where the second inequality follows by (A.1). □

Acknowledgments. The Framingham Heart Study project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (N01 HC25195).

The authors acknowledge the investigators that contributed the phenotype, genotype and simulated data for this study. The manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University or the NHLBI. The authors are grateful to Dr. Zhong Wang for sharing the idea to create Figure 1 in this paper. The authors thank the Editor, the Associate

Editor and three anonymous referees for their constructive comments, which have led to a significant improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or NNSFC.

REFERENCES

ALTSHULER, D., DALY, M. J. and LANDER, E. S. (2008). Genetic mapping in human disease. *Science* **322** 881–888.

AYERS, K. L. and CORDELL, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* **34** 879–891.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BURTON, P. R., CLAYTON, D. G., CARDON, L. R., CRADDOCK, N., DELOUKAS, P., DUNCANSON, A. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

CHIPMAN, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36. MR1394738

CHO, S., KIM, H., OH, S., KIM, K. and PARK, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proceedings* **3** S25.

CORDELL, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10** 392–404.

DALY, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet.* **11** 241–246.

DAS, K., LI, J., WANG, Z., TONG, C., FU, G., LI, Y., XU, M., AHN, K., MAUGER, D., LI, R. and WU, R. (2011). A dynamic model for genome-wide association studies. *Hum. Genet.* **8** 1–8.

DAWBER, T. R., MEADORS, G. F. and MOORE, F. E. (1951). Epidemiological approaches to heart disease: The Framingham study. *Am. J. Publ. Health* **41** 279–286.

DUGGAL, P., GILLANDERS, E. M., HOLMES, T. N. and BAILEY-WILSON, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* **9** 516.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

FRAYLING, T. M., TIMPSON, N. J., WEEDON, M. N., ZEGGINI, E., FREATHY, R. M., LINDGREN, C. M. et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316** 889–894.

GORLOV, I. P., GORLOVA, O. Y., SUNYAEV, S. R., SPITZ, M. R. and AMOS, C. I. (2008). Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **82** 100–112.

HARLEY, J. B., ALARCN-RIQUELME, M. E., CRISWELL, L. A., JACOB, C. O., KIMBERLY, R. P., MOSER, K. L. et al. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.* **40** 204–210.

HE, Q. and LIN, D. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* **27** 1–8.

HIRSCHHORN, J. N. (2009). Genomewide association studies—Illuminating biologic pathways. *N. Engl. J. Med.* **360** 1699–1701.

HOLDEN, M., DENG, S., WOJNOWSKI, L. and KULLE, B. (2008). GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24** 2784–2785.

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

JACOBS, K. B., YEAGER, M., WACHOLDER, S., CRAIG, D., KRAFT, P., HUNTER, D. J. et al. (2009). A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat. Genet.* **41** 1253–1257.

JAQUISH, C. E. (2007). The Framingham Heart Study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.* **8** 63.

KIM, Y., WOJCIECHOWSKI, R., SUNG, H., MATHIAS, R., WANG, L., KLEIN, A., LENROOT, R., MALLEY, J. and BAILEY-WILSON, J. (2009). Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proceedings* **3** S64.

LANGE, K., CANTOR, R., HORVATH, S., PEROLA, M., SABATTI, C., SINSHEIMER, J. and SOBEL, E. (2001). Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am. J. Hum. Genet.* **69 (Suppl. 1)** A1886.

LANGE, K., PAPP, J. C., SINSHEIMER, J. S., SRIPRACHA, R., ZHOU, H. and SOBEL, E. M. (2013). Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics* **29** 1568–1570.

LI, J., DAS, K., FU, G., LI, R. and WU, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* **27** 516–523.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

PSYCHIATRIC GCCC (2009). Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *Am. J. Psychiatr.* **166** 540–556.

RITCHIE, M. D., HAHN, L. W., ROODI, N., BAILEY, L. R., DUPONT, W. D., PARL, F. F. and MOORE, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69** 138.

SCUTERI, A., SANNA, S., CHEN, W.-M., UDA, M., ALBAI, G., STRAIT, J. et al. (2007). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3** e115.

SPELIOTES, E. K., WILLER, C. J., BERNDT, S. I., MONDA, K. L., THORLEIFSSON, G. et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42** 937–948.

SZYMCZAK, S., BIERNACKA, J. M., CORDELL, H. J., GONZALEZ-RECIO, O., KONIG, I. R., ZHANG, H. and SUN, Y. V. (2009). Machine learning in genome-wide association studies. *Genet. Epidemiol.* **33** S51–S57.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

UEKI, M. and TAMIYA, G. (2012). Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis. *BMC Bioinformatics* **13** 72.

WAN, X., YANG, C., YANG, Q., XUE, H., FAN, X., TANG, N. and YU, W. (2010a). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **87** 325–340.

WAN, X., YANG, C., YANG, Q., XUE, H., TANG, N. and YU, W. (2010b). Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* **26** 30–37.

WANG, G., VOLKOW, N., LOGAN, J., PAPPAS, N., WONG, C., ZHU, W., NETUSLL, N. and FOWLER, J. (2001). Brain dopamine and obesity. *The Lancet* **357** 354–357.

WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

WANG, K., LI, M. and BUCAN, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81** 1278–1283.

WANG, Y., LIU, G., FENG, M. and WONG, L. (2011). An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* **27** 2936–2943.

WEEDON, M. N. and FRAYLING, T. M. (2008). Reaching new heights: Insights into the genetics of human stature. *Trends Genet*. **24** 595–603.

WU, J., DEVLIN, B., RINGQUIST, S., TRUCCO, M. and ROEDER, K. (2010). Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genet. Epidemiol*. **34** 275–285.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.

WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat*. **2** 224–244. MR2415601

WU, Z. and ZHAO, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet*. **5** e1000582.

YANG, C., HE, Z., WAN, X., YANG, Q., XUE, H. and YU, W. (2009). SNPHarvester: A filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* **25** 504–511.

YANG, C., WAN, X., YANG, Q., XUE, H. and YU, W. (2010). Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group lasso. *BMC Bioinformatics* **11** S18.

YI, N., KAKLAMANI, V. G. and PASCHE, B. (2011). Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Ann. Hum. Genet*. **75** 90–104.

ZHANG, X., HUANG, S., ZOU, F. and WANG, W. (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* **26** 217–227.

ZHANG, Y. and LIU, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet*. **39** 1167–1173.

ZHOU, H., SEHL, M. E., SINSHEIMER, J. S. and LANGE, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26** 2375–2382.

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc*. **106** 1464–1475. MR2896849

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol*. **67** 301–320. MR2137327

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist*. **36** 1509–1533. MR2435443

J. LI
DEPARTMENT OF APPLIED AND COMPUTATIONAL MATHEMATICS AND STATISTICS
UNIVERSITY OF NOTRE DAME
NOTRE DAME, INDIANA 46556
USA
E-MAIL: jli7@nd.edu

W. ZHONG
WANG YANAN INSTITUTE FOR STUDIES IN ECONOMICS
DEPARTMENT OF STATISTICS
SCHOOL OF ECONOMICS
FUJIAN KEY LABORATORY OF STATISTICAL SCIENCE
XIAMEN UNIVERSITY
XIAMEN, FUJIAN 361005
CHINA
E-MAIL: wzhong@xmu.edu.cn

R. LI
THE METHODOLOGY CENTER
DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: rli@stat.psu.edu

R. WU
CENTER FOR STATISTICAL GENETICS
PENNSYLVANIA STATE UNIVERSITY
HERSHEY, PENNSYLVANIA 17033
USA
E-MAIL: rwu@hes.hmc.psu.edu