BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 6. 1057–1068
Advance Access published on 21 November 2013
doi:10.1093/bib/bbt082

A model-free approach for detecting interactions in genetic association studies

Jiahan Li, Jun Dan, Chunlei Li and Rongling Wu

Submitted: 2nd August 2013; Received (in revised form): 6th October 2013

Abstract

Over the past few decades, genome-wide association studies analyzed by efficient statistical procedures have successfully identified single-nucleotide polymorphisms (SNPs) that are associated with complex traits or human diseases. However, due to the overwhelming number of SNPs, most approaches have focused on additive genetic model without genome-wide SNP–SNP interactions. In this study, we propose an efficient statistical procedure in a genetic model-free framework for detecting SNPs exhibiting main genetic effects as well as epistatic interactions. Specifically, the association between phenotype and genotype is characterized by an unknown function to be estimated using nonparametric techniques, and a two-stage non-parametric independence screening procedure is proposed to sequentially identify potentially important main genetic effects and interactions. Finally, the subset of genetic predictors implied by two-stage non-parametric independence screening is analyzed by penalized regressions such as LASSO, and a final model is identified. In this framework, specific genetic model is not assumed and interactions are not only among marginally important SNPs. Therefore, SNPs that are involved in genetic regulatory networks but missed by previous studies are expected to be recognized. In simulation studies, we show that the procedure is computationally efficient and has an outstanding finite sample performance in selecting potential SNPs as well as SNP–SNP interactions. A real data analysis further indicates the importance of epistatic interactions in explaining body mass index.

Keywords: *variable screening; nonparametric method; genome-wide association studies; variable selection*

INTRODUCTION

Analyzing genome-wide association studies (GWAS) is statistically challenging due to the overwhelming number of genetic markers, or single-nucleotide polymorphisms (SNPs), and a limited number of subjects in a study. The choice of genetic models that assume some specific association structures between the phenotype and all genotypes further complicates the analysis. Conventionally, single SNP analysis that tests the marginal effects of individual SNPs separately is implemented, and additive genetic model is widely used. This strategy has successfully identified many important genetic variants associated with human diseases and complex traits [1,2], but it is criticized for its biased genetic effect estimates, inflated false-positive rate and the lack of statistical power [3,4].

Identifying important genetic variants out of millions of observable genetic markers is a high-dimensional variable selection problem, where important

Corresponding author. Jiahan Li, Department of Applied and Computational Mathematics and Statistics, University of Notre dame, Notre Dame, IN 46556, USA. Tel.: +1 574 631 2741; Fax: +1 574 631 4822. E-mail: jiahan.li@nd.edu

Jiahan Li is an assistant professor of statistics in the Department of Applied and Computational Mathematics and Statistics (ACMS) at the University of Notre Dame. His research areas include statistical genetics, statistical genomics and high-dimensional modeling.

Jun Dan is a Ph.D. student in the Department of Applied and Computational Mathematics and Statistics (ACMS) at the University of Notre Dame. Her thesis work focuses on developing high-dimensional statistical models for analyzing genome-wide association studies.

Chunlei Li is a Ph.D. student in the Department of Applied and Computational Mathematics and Statistics (ACMS) at the University of Notre Dame. His research interests include stochastic dynamic systems and Monte Carlo simulations.

Rongling Wu is a professor of biostatistics and statistics and the Director of the Center for Statistical Genetics at the Pennsylvania State University. He is interested in statistical genetics and computational biology.

© The Author 2013. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

associations between phenotypic measurements and genetic variants are selected. Usually such selection is guided by the final model predictive performance. Because LASSO regression [5] is capable of producing sparse solutions by shrinking most of the regression coefficients to 0 through an L1 norm penalty term, LASSO-based variable selections have been proposed and widely used in analyzing whole-genome SNP data. For example, Wu and Lange [6] and Wu *et al*. [7] developed cyclic coordinate descent algorithm for analyzing GWAS data sets, where a fixed number of important SNPs could be specified. Cho *et al*. [8], Li *et al*. [9] and He and Lin [10] extended LASSO-based approached in different directions. All these findings suggest that variable selection methods have better statistical performance and computational feasibility than the single SNP analysis.

Despite important genes being identified by GWAS, the proportion of phenotypic variance explained by these genes is still very limited [4]. On the other hand, more and more evidence in systems biology and biomedicine indicates that gene–gene interactions play important roles in formulating genetic regulator networks and pathways. However, identifying interactions directly is usually computationally infeasible, and thus pairwise interactions are searched among marginally significant SNPs [7]. In case–control studies where both genotypes and phenotypes are discrete, machine learning methods have been developed to search for interactions by taking the advantage of special data structure (for review, see [11]). However, in population-based studies with quantitative traits, methods designed for analyzing case–control studies cannot be directly applied unless the continuous phenotypic values are properly discretized.

Testing additive genetic model alone, which is a common practice in analyzing GWAS data sets, may also reduce statistical power. Lettre *et al*. [12] showed that the maximal power is achieved only if the assumed genetic model is the actual underlying mode of inheritance of the causal allele. When the actual pattern of inheritance is unknown, however, testing the codominant model alone, or alternatively testing additive model, dominant model and recessive model together is recommended in population-based association studies. Although GWAS analysis tools such as Mendel [13] and SNPStats [14] provide options for specifying genetic models, the optimal method for determining genetic models is still not clear.

To this end, we propose a genetic model-free approach for detecting SNPs exhibiting main genetic effects and epistatic interactions in population-based association studies. Our nonparametric genetic model assumes that genetic predictors are associated with the phenotype through some unknown functions. These functions can be approximated nonparametrically by basis expansion, and are shrunk toward 0 in penalized regressions. This framework extends traditional genetic models, such as additive model, recessive model and dominant model, allowing non-linear effects of genetic controls. Without biological justifications for the choice of genetic models, this approach is robust against the risk of model misspecification.

In terms of selecting important genetic predictors including main effects and epistatic interactions, a hybrid of variable screening and variable selection is proposed, where variable screening is based on non-parametric independence screening (NIS) [15] and LASSO is used for variable selection. By using a two-stage NIS (TS-NIS) procedure, this framework identifies a subset of genetic predictors containing marginally important SNPs as well as SNPs that are not marginally significant but involved in interactions. We show that the sure independence screening (SIS) property [15,16] implies an overwhelming probability of retaining truly important SNPs.

Because the model dimensionality is dramatically reduced by TS-NIS, variable selection methods are expected to be more efficient in identifying causal SNPs and SNP–SNP interactions and formulating a final model. We, in particular, apply LASSO regression to further eliminate irrelevant genetic predictors. Simulation studies suggest LASSO could greatly reduce the false-positive rate while retaining most of the causal SNPs and SNP–SNP interactions. As a result, the whole statistical framework identifies important genetic risk factors and their interactive patterns out of a huge number of SNPs with high statistical power and low false-positive rate.

The rest of this article is organized as follows. In Section 2, we discuss the proposed non-parametric genetic model and the TS-NIS followed by variable selections. In Section 3, we provide simulation studies and comparisons with other approaches in analyzing GWAS with quantitative traits. Section 4 illustrates this framework by analyzing a data set from Framingham study, where main genetic effects and epistatic interactions associated with body mass

index (BMI) are identified. We give the concluding remarks and discussions in Section 5.

METHODS

Genetic model

Consider a GWAS data set consisting of n randomly selected subjects from a population. Let γ_i be the continuous phenotypic measurement for the ith subject and $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ be a p-dimensional vector of genotypes from p SNPs. Given two alleles, A and a, predictor x_{ij} from SNP j of subject i is defined as -1 for genotype aa, 0 for genotype Aa and 1 otherwise. The ultrahigh dimensionality of this problem implies that p could be $\exp(\mathcal{O}(n^a))$for some $a > 0$ [16].

A general non-parametric genetic model with both main genetic effects and pairwise interactions assumes

$$\gamma_i = \mu + \sum_{j=1}^{p} g_j(x_{ij}) + \sum_{j=k+1}^{p} \sum_{k=1}^{p} h_{jk}(x_{ij}x_{ik}) + e_i, \quad (1)$$

where μ is the population mean, $g_j(.)$ is an unknown function specifying the association between the phenotype and the jth SNP, $h_{jk}(.)$ is an unknown function specifying the association between the phenotype and the epistatic interaction between the jth SNP and the kth SNP and e_i is a random error with mean 0 and variance σ^2.

This model is an extension of the additive non-parametric model [17], and nests other genetic models that are commonly used. For example, an additive model without epistatic interactions assumes $g_j(x) = \beta_j x$ and $h_{jk}(x) = 0$, and a recessive model assumes a step function for $g_j(x)$ and $h_{jk}(x) = 0$. In practice, however, a particular genetic model may not hold for all SNPs, especially when the number of SNPs is extremely large, and the interaction patterns are even more complicated. Therefore, assuming such a general model is desirable for ultra-high-dimensional genetic association studies.

Because the phenotypic measurement γ_i is only determined by a small subset of genetic predictors, our goal is to identify important SNPs with non-zero functional coefficients $g_j(.)$ and $h_{jk}(.)$, where $g_j(.)$ is interpreted as main genetic effect and $h_{jk}(.)$ is the epistatic effect. In ultrahigh-dimensional settings, testing all pairwise interactions is neither theoretically valid nor computationally affordable. Moreover, even if two-way interactions can be tested in this way, the detection of higher-order interactions through exhaustive search is still challenging. Alternatively, prevailing methods usually consider interactions among significant main effects.

There are two versions of the heredity structures of interactions: strong heredity principle and weak heredity principle [18]. Under strong heredity assumption, interactions occur between two marginally significant predictors, whereas under weak heredity assumption, one of these predictors could have a zero marginal effect. Our approach only assumes weak heredity in the genetic model.

Basis expansion

We approximate unknown functional coefficients $g_j(.)$ and $h_{jk}(.)$ in (1) through a series of basis functions. Under mild regularity conditions,

$$g_j(x) = \sum_{l=0}^{\infty} c_{j,l}\varphi_l(x),\ j = 1, \ldots, p, \quad (2)$$

where $\varphi_l(.)$ is the lth basis function, and $c_{j,l}$ is the corresponding coefficient. Similar basis expansions exist for $h_{jk}(.)$:

$$h_{jk}(x) = \sum_{l=0}^{\infty} c_{jk,l}\varphi_l(x),\ k = 1, \ldots, p,\ j = k+1, \ldots, p. \quad (3)$$

The choices of sets of basis functions include Legendre orthogonal polynomials, B-splines and Fourier basis functions, all of which have been applied in modeling developmental trajectories in quantitative trait loci mapping [19–21].

According to the basis expansion of $g_j(.)$ and $h_{jk}(.)$, genetic model (1) becomes

$$\begin{aligned} \gamma_i = \mu &+ \sum_{j=1}^{p} \sum_{l} c_{j,l}\varphi_l(x_{ij}) \\ &+ \sum_{j=k}^{p} \sum_{k=1}^{p} \sum_{l} c_{jk,l}\varphi_l(x_{ij}x_{ik}) + e_i. \end{aligned} \quad (4)$$

Model (4) is a linear one. If model dimensionality is small to moderate, $c_{j,l}$ and $c_{jk,l}$ can be estimated by ordinary linear regression or penalized regressions. However, with ultrahigh-dimensional parameter space spanned by both main genetic effects and interactions, these methods cannot be directly applied for both computational purpose and variable selection purpose.

In what follows, we propose a two-stage variable screening approach followed by a step of variable selection to identify important genetic predictors, and a family of Legendre orthogonal polynomials is

used as the basis functions. The general form of a Legendre polynomial of order l is given by

$$\varphi_l(x) = \sum_{k=0}^{\lfloor l/2 \rfloor} (-1)^k \frac{(2l-2k)!}{2^l k!(l-k)!(l-2k)!} x^{l-2k},$$

where $\lfloor . \rfloor$ denotes the integer part of a positive real number. The first five Legendre polynomials are defined as $\varphi_0 = 1$, $\varphi_1 = x$, $\varphi_2 = \frac{1}{2}(3x^2 - 1)$, $\varphi_3 = \frac{1}{2}(5x^3 - 3x)$ and $\varphi_4 = \frac{1}{8}(35x^4 - 30x^2 + 3)$. In analyzing real data sets, we could determine the appropriate polynomial degree by implementing the whole procedure with different polynomial degrees, and selecting the polynomial degree that gives the lowest value of Bayesian Information Criterion. In simulations, however, this is computationally expensive. Therefore, the polynomial degree is fixed at 3 in simulation studies. As a robustness check, we also re-analyze all simulated data sets with polynomial degree $\nu = 4$.

TS-NIS

Let us call two SNPs involved in a two-way interaction two roots, and let $M_0 = \{1,2,\ldots,p\}$ be a set of indices including all SNPs. This is the initial model before variable screening. To estimate a subset of important SNPs with non-zero main effects $M_1 = \{1 \le j \le p : g_j(x) \ne 0 \text{for some} x\}$, we implement NIS on M_0. NIS technique was developed by [15], where predictors are ranked by their non-parametric marginal correlations with the response, and top-ranked predictors are retained for the following variable selection. Although important predictors may have relatively low marginal correlations, as long as they are retained in the reduced model, they could be correctly identified by variable selections that jointly analyze all predictors.

In particular, we approximate function $g_j(x), j = 1,\ldots,p$ with ν-dimensional basis functions $\varphi = (\varphi_1,\ldots,\varphi_\nu)^T$ and estimate $c_{j,l}, l = 1,\ldots,\nu$ in marginal non-parametric regressions

$$y_i = \mu + \sum_{l=1}^{\nu} c_{j,l}\varphi_l(x_{ij}) + e_i, \text{ for } j = 1,\ldots,p. \tag{5}$$

Then the squared L2 norm $\sum_{l=1}^{\nu} \hat{c}_{j,l}^2$ serves as a marginal utility measure of the jth SNP, and those SNPs whose marginal utility measures are greater than a threshold C are selected. In other words, we estimate $M_1 = \{1 \le j \le p : g_j(x) \ne 0\}$ by $\widehat{M}_1 = \left\{1 \le j \le p : \sum_{l=1}^{\nu} \hat{c}_{j,l}^2 > C\right\}$. The sure screening property of NIS guarantees that $\widehat{M}_1$ contains all SNPs with non-zero marginal effects with a probability tending to 1. These selected SNPs, however, could exhibit additive effect, dominant effect, recessive effect or codominant effect.

Then according to weak heredity principle, any SNPs in M_0 may interact with marginally important SNPs in $\widehat{M}_1$, and the model becomes

$$\begin{aligned} y_i = \mu &+ \sum_{j \in \widehat{M}_1} \sum_l c_{j,l}\varphi_l(x_{ij}) \\ &+ \sum_{j \in M_0} \sum_{k \in \widehat{M}_1} \sum_l c_{jk,l}\varphi_l(x_{ij}x_{ik}) + e_i, \end{aligned} \tag{6}$$

To screen out covariates that are involved in interactions but missed by the first round NIS due to their negligible marginal effects, a second-stage NIS is carried out between the phenotype and each pairwise interaction term $x_{ij}x_{ik}, j \in M_0$, $k \in \widehat{M}_1$. Similar approximations and screening procedures are applied in the second stage, and truly important interactions $M_2 = \{1 \le j \le p, 1 \le k \le p : h_{jk}(x_{ij}x_{ik}) \ne 0\}$ are estimated by $\widehat{M}_2 = \left\{j \in M_0, k \in \widehat{M}_1 : \sum_{l=1}^{\nu} \hat{c}_{jk,l}^2 > C'\right\}$ for some $C' > 0$.

After TS-NIS, the genetic model becomes

$$y_i = \mu + \sum_{j \in \widehat{M}_1} \sum_l c_{j,l}\varphi_l(x_{ij}) + \sum_{j,k \in \widehat{M}_2} \sum_l c_{jk,l}\varphi_l(x_{ij}x_{ik}) + e_i. \tag{7}$$

As suggested by [15] and [16], two thresholds C and C' are determined so that the top-ranked $\lfloor n/\log(n) \rfloor$ genetic predictors in each stage are retained.

The advantage of this two-stage SIS (TS-SIS) procedure is its efficiency and accuracy in identifying notable phenotype–genotype associations of any form. Owing to the weak heredity principle, covariate in $\widehat{M}_2$ may either have non-zero marginal genetic effects or only modify the effect of other SNPs by involving in epistatic interactions.

Sure screening property

Fan, Feng and Song (2011) showed the sure screening property of a NIS procedure for ultrahigh-dimensional additive models. This property states that the probability of including all important predictors in the final model goes to 1. To show the sure screening property for the proposed two-stage approach, we assume that conditions A–F in Fan,

Feng and Song (2011) are satisfied for both stages of screening.

To be specific, conditions A–C in Fan, Feng and Song (2011) imply that if we approximate non-linear genetic effects by some non-parametric techniques and the approximation error is bounded above, the signal from the basis expansion of any important SNP is at the same level as its true non-linear effect. Moreover, the minimum signal of all truly important predictors is bounded below by some positive number, indicating that the separation of important predictors and unimportant predictors is possible through a variable screening procedure.

Conditions D–F ensure that only the number of genetic predictors with non-zero effects matter for the purpose of sure screening. In other words, the probability of selecting all important SNPs and SNP–SNP interactions is high, and this probability only depends on the number of truly important genetic predictors, not the total number of SNPs.

Given these conditions, Theorem 1 in Fan, Feng and Song (2011) implies that for some positive real numbers a_1, b_1 and κ_1,

$$P\left(M_1 \subset \widehat{M}_1\right) \geq 1 - v|M_1| \left\{(8+2\upsilon)e^{-a_1 n^{1-4k_1}\upsilon^{-3}} + 6\upsilon e^{-b_1 n\upsilon^{-3}}\right\}, \tag{8}$$

where $|M_1|$ is the number of truly important main effects. Similarly, for some positive real numbers a_2, b_2 and κ_2,

$$P\left(M_2 \subset \widehat{M}_2\right) \geq 1 - v|M_2| \left\{(8+2\upsilon)e^{-a_2 n^{1-4k_2}\upsilon^{-3}} + 6\upsilon e^{-b_2 n\upsilon^{-3}}\right\}, \tag{9}$$

where $|M_2|$ is the number of truly important epistatic effects. By combining (8) and (9), it is straightforward to show

$$P\left(M_1 \subset \widehat{M}_1 \text{and } M_2 \subset \widehat{M}_2\right) \geq 1 - v\sum_{s=1}^{2}|M_s| \left\{(8+2\upsilon)e^{-a_s n^{1-4k_s}\upsilon^{-3}} + 6\upsilon e^{-b_s n\upsilon^{-3}}\right\}. \tag{10}$$

Therefore,

$$P\left(M_1 \subset \widehat{M}_1 \text{ and } M_2 \subset \widehat{M}_2\right) \to 1. \tag{11}$$

In other words, the proposed two-stage non-parametric screening procedure has the sure screening property. In detecting SNPs with non-zero main effects and epistasis, the probability of including truly important SNPs in the reduced model goes to 1.

Variable selection by LASSO regression

Once the model dimensionality is dramatically reduced, variable selection methods for non-parametric additive models could be used to further eliminate irrelevant covariates in $\widehat{M}_1$ and interactions $\widehat{M}_2$ [22–25]. To facilitate genetic interpretability, we fit a genetic model including both additive and dominant effects for SNPs in $\widehat{M}_1$, additive × additive interactions, additive × dominant interactions, dominant × additive interactions and dominant × dominant interactions for each pair in $\widehat{M}_2$:

$$\begin{aligned}\gamma_i = \mu &+ \sum_{j\in\widehat{M}_1} a_j x_{ij} + \sum_{j\in\widehat{M}_1} d_j z_{ij} + \sum_{j,k\in\widehat{M}_2} I^{aa}_{jk} x_{ij}x_{ik} \\ &+ \sum_{j,k\in\widehat{M}_2} I^{ad}_{jk} x_{ij}z_{ik} + \sum_{j,k\in\widehat{M}_2} I^{da}_{jk} z_{ij}x_{ik} \\ &+ \sum_{j,k\in\widehat{M}_2} I^{dd}_{jk} z_{ij}z_{ik} + e_i, i = 1,\ldots,n,\end{aligned} \tag{12}$$

where $z_{ij} = 1$ for genotypes AA or aa and 0 for Aa, a_j and d_j are the additive effect and dominant effect of SNP j, respectively, and I^{aa}_{jk}, I^{ad}_{jk}, I^{da}_{jk} and I^{dd}_{jk} denote four different modes of epistatic interactions between SNP j and SNP k. This comprehensive genetic model has been widely used in the literature [26–29].

With a reasonable number of genetic predictors, LASSO regression implemented by coordinate descent algorithms [6–7] is used to further refine the reduced genetic model and estimate genetic effects, which minimizes the following penalized least squares:

$$\sum_{i=1}^{n}(\gamma_i - EY_i)^2 + \lambda\sum_{j\in\widehat{M}_1}(|a_j| + |d_j|) + \lambda \sum_{j,k\in\widehat{M}_2}\left(\left|I^{aa}_{jk}\right| + \left|I^{da}_{jk}\right| + \left|I^{ad}_{jk}\right| + \left|I^{dd}_{jk}\right|\right).$$

We select the tuning parameter using 5-fold cross-validation. Of course, other variable selection methods proposed in the GWAS literature can also be applied. As suggested by Fan and Lv (2008), a variable selection step following variable screening could further remove non-significant predictors with increased interpretability. Finally, we use linear regressions to unbiasedly estimate the effect of SNPs and SNP–SNP interactions that are selected by the LASSO regression.

SIMULATION STUDY

In this section, we present simulation studies using the proposed TS-NIS scheme to select important SNPs and SNP–SNP interactions. In each replication, we simulate a data set according to a genetic model with both additive effects and dominant effects. This genetic model would be the optimal one in detecting causal SNPs. However, such a genetic model is assumed to be unknown to investigators.

Specifically, genotypic data are simulated as follows. For SNP j of subject i, $i = 1, \cdots, n = 1000$, $j = 1, \cdots, p$, the genotype x_{ij} is derived from a marginal standard normal random variable s_{ij}. If SNP j and SNP k are in the same chromosome, s_{ij} and s_{ik} have a correlation of $\rho^{|j-k|} = corr(s_{ij}, s_{ik})$, where ρ is specified as 0.2, 0.5 or 0.8 in different scenarios. Then genotypic value x_{ij} is derived from s_{ij} according to

$$x_{ij} = \begin{cases} 1, & s_{ij} < c, \\ 0, & c \le s_{ij} \le -c, \\ -1, & s_{ij} > -c, \end{cases}$$

where c is the first quartile of a standard normal distribution.

In simulating phenotypic values, we include 12 main effects and 8 epistatic interactions whose positions and effects are given in Table 1. The first two columns of Table 1 (column 'Chr' and column 'Position') list the position an SNP resides on. Column 'Additive/Dominant' represents whether SNP demonstrates additive effect or dominant effect. Column 'interact with' presents the row index of the SNP that is interacting with the current one. For the first epistatic interaction, for example, SNP 1 on chromosome 11 interacts with the first SNP with non-zero genetic effect, which is SNP 1 on chromosome 1. With this specification, some SNPs only have main effects (SNPs on chromosomes 9 and 10); some SNPs only involve in epistatic interactions but do not exhibit main effects (all SNPs under 'Epistatic Interactions' in Table 1). Moreover, two SNPs in an epistatic interaction could be highly correlated if they are on the same chromosome (SNPs on chromosomes 3, 4, 5 and 6).

Owing to the computational cost, we further simulate 3500 SNPs with 0 genetic effects in each replication, leading to 6.2 million pairwise interactions. All simulated SNPs are distributed into 23 chromosomes, and the proportion of SNPs on each chromosome is the same as that in our real data analysis. Then the phenotypical value of each subject is generated according to a genetic model consisting of all main effects and interactions in Table 1, where genetic effects equal to 1. Three noise levels are considered in simulating the phenotype $\sigma^2 = 6,8,10$.

Table 1: Main effect SNPs and SNP–SNP interactions in simulated data

Chromosome	Position	Additive/ Dominant	Interact with
Main effects			
1	1	Additive	–
2	1	Additive	–
3	1	Additive	–
4	1	Additive	–
5	1	Dominant	–
6	1	Dominant	–
7	1	Dominant	–
8	1	Dominant	–
9	1	Dominant	–
10	1	Additive	–
9	5	Dominant	–
10	5	Additive	–
Epistatic interactions			
11	1	Additive	1
12	1	Additive	2
3	2	Additive	3
4	3	Additive	4
5	2	Dominant	5
6	3	Dominant	6
17	1	Dominant	7
18	1	Dominant	8

For each simulated data set, we implement TS-NIS and then apply LASSO to the reduced model. Our goal is to select a small subset of SNPs and SNP–SNP interactions that includes all important ones. In Table 2, we report the average statistical power and false-positive rate over 30 simulations for each simulation scenario, where standard errors are in parentheses.

As can be seen from Table 2, the statistical power TS-NIS is high for all different (ρ, σ^2) combinations. In other words, TS-NIS is an efficient dimension reduction technique, which produces a reduced model that identifies most of the truly significant SNPs of >6 million genetic predictors. The reduced model contains $2n/\log(n) = 288$ main genetic effects and interactions, a reasonable number of predictors for the following LASSO regression. Also, the variances of statistical power and false-positive rate are very small, suggesting the consistency of this procedure. Furthermore, linkage disequilibrium ρ has a very limited impact, and increased σ^2 is associated with decreased statistical power.

Table 2: Power and FPR of simulated data using TS-NIS and TS-NIS-LASSO ($v = 3$)

(ρ,σ^2)	Power (%)		False-positive rate ($\times 10^{-4}$)	
	TS-NIS	TS-NIS-LASSO	TS-NIS	TS-NIS-LASSO
(0.8, 6)	99.2	97.7	2.45	0.72
	(2.3)	(3.4)	(9.09×10^{-3})	(0.09)
(0.8, 8)	97.5	94.7	2.45	0.72
	(3.4)	(4.7)	(1.34×10^{-2})	(0.09)
(0.8, 10)	98.2	96.3	2.45	0.76
	(3.3)	(4.1)	(1.32×10^{-2})	(0.10)
(0.5, 6)	98.8	98.5	2.45	0.87
	(2.2)	(2.3)	(8.48×10^{-3})	(0.15)
(0.5, 8)	99.0	98.2	2.45	0.88
	(2.0)	(2.8)	(8.02×10^{-3})	(0.10)
(0.5, 10)	98.2	97.2	2.45	0.91
	(3.1)	(4.1)	(1.21×10^{-2})	(0.13)
(0.2, 6)	99.0	98.8	2.45	0.88
	(2.4)	(2.5)	(9.54×10^{-3})	(0.11)
(0.2, 8)	98.5	97.7	2.45	0.90
	(2.7)	(3.1)	(1.05×10^{-2})	(0.14)
(0.2, 10)	98.5	97.8	2.45	0.88
	(2.7)	(3.1)	(1.05×10^{-2})	(0.11)

After TS-NIS, LASSO regression is applied to the reduced model to identify genetic predictors with non-zero effects. Although, by construction, the number of truly significant SNPs selected by LASSO is less than or equal to that selected by TS-NIS, the decrease of statistical power is limited. In contrast, the decrease of false-positive rate is obvious, suggesting the effectiveness of LASSO regression in further eliminating irrelevant genetic predictors. Moreover, as linkage disequilibrium level increases, both statistical power and false-positive rate of the whole procedure (TS-NIS-LASSO) slightly decreases.

We also compare our TS-NIS procedure with four other methods. Screen and Clean (SC; [30]) is a computationally efficient approach to detect important SNPs and interactions in GWAS with quantitative phenotype. This approach applies LASSO regression in search of candidate main effects and interactions. Then a cleaning process is implemented to identify significant genetic predictors. Moreover, we compare our approach with a similar procedure where NIS in each stage is replaced with SIS (16), which uses marginal linear correlation to select SNPs. We call this approach TS-SIS. This allows a direct comparison between parametric and non-parametric approaches.

Furthermore, we compare our approach with Forward LASSO [31]. Being a multistage approach to select important high-order interactions, forward LASSO selects main effect SNPs in the first stage, selects SNP-SNP interactions in the second stage, selects SNP–SNP–SNP interactions in the third stage and so on. The last method our approach is compared with is the well-known multi-dimensionality reduction (MDR; [32]) approach. The comparison of our approach with these four methods will provide a comprehensive evaluation.

Table 3 summarizes the result of TS-SIS and SC. We see that TS-SIS has a power ~95%, which is lower than that of TS-NIS in Table 2. Because TS-SIS implements two-stage variable screening based on a linear model, non-linear interaction patterns tend to be missed. Then once LASSO regression is applied to the reduced model implied by TS-SIS, the power further decreases to ~90%. This is a notable decrease compared with the decrease when LASSO applied to the reduced model from TS-NIS. Because genetic predictors in the reduced model could be highly correlated, LASSO regression tends to randomly select genetic predictors and miss important ones in the final model. As a result, the performance of TS-SIS-LASSO is not stable over replications, as can be seen from their higher standard errors. On the other hand, unimportant SNPs retained by TS-NIS (Table 2) may not have strong linear correlations with important SNPs, leading to better variable selection performance. In implementing SC method, we set both the number of main effects and the number of epistatic interactions to $n/\log(n)$. The power of SC is ~80%.

Table 4 summarizes the result of Forward LASSO and MDR. The power of Forward LASSO is in the range of 33.3 and 63.8%, and the variances of statistical power are large in general. On the other hand, MDR method has statistical power from 61 to 66%, and slightly lower variances. Note that in Forward LASSO, an additive genetic model is assumed in all simulations. Higher statistical power could be expected if true inheritance patterns are assumed, for example, in a genetic model including dominant effects and interactions among additive effects and dominant effects. Among all alternative approaches, decreased σ^2 leads to increased statistical power. However, their false-positive rates are more than twice as much as that of TS-NIS-LASSO. Therefore, in analyzing ultrahigh-dimensional genetic data from GWAS, TS-NIS-LASSO is recommended in identifying important main effects as well as epistatic interactions.

Table 3: Power and FPR of simulated data using TS-SIS, TS-SIS-LASSO and SC

(ρ,σ^2)	Power (%)			False-positive rate ($\times 10^{-4}$)		
	TS-SIS	TS-SIS-LASSO	SC	TS-NIS	TS-SIS-LASSO	SC
(0.8, 6)	94.7	90.3	82.2	2.46	1.49	2.46
	(5.4)	(7.2)	(3.6)	(0.02)	(0.12)	(0.6)
(0.8, 8)	95.5	89.3	79.7	2.46	1.55	2.46
	(4.2)	(6.8)	(3.5)	(0.02)	(0.16)	(0.05)
(0.8, 10)	92.8	85.3	79.8	2.47	1.59	2.47
	(4.5)	(7.3)	(4.5)	(0.02)	(0.17)	(0.06)
(0.5, 6)	96.2	92.0	79.3	2.46	1.68	2.46
	(4.9)	(6.4)	(2.9)	(0.02)	(0.12)	(0.04)
(0.5, 8)	95.5	87.0	78.8	2.46	1.79	2.46
	(5.3)	(7.8)	(3.9)	(0.02)	(0.12)	(0.05)
(0.5, 10)	95.7	87.8	78.2	2.46	1.83	2.47
	(5.2)	(7.0)	(3.8)	(0.02)	(0.12)	(0.05)
(0.2, 6)	97.5	92.8	79.5	2.45	1.76	2.47
	(3.2)	(5.7)	(2.4)	(0.01)	(0.12)	(0.05)
(0.2, 8)	95.7	88.7	79.3	2.46	1.82	2.45
	(4.9)	(6.9)	(3.1)	(0.02)	(0.12)	(0.06)
(0.2, 10)	95.3	88.2	77.5	2.46	1.85	2.48
	(3.5)	(7.3)	(5.5)	(0.01)	(0.12)	(0.06)

Table 4: Power and FPR of simulated data using Forward LASSO and MDR

(ρ,σ^2)	Power (%)		False-positive rate ($\times 10^{-4}$)	
	Forward LASSO	MDR	Forward LASSO	MDR
(0.8, 6)	38.8	61.3	2.82	2.12
	(6.3)	(2.2)	(7.32×10^{-3})	(0.24)
(0.8, 8)	45.2	61.3	2.82	2.25
	(6.8)	(2.2)	(8.76×10^{-3})	(0.31)
(0.8, 10)	33.3	61.0	2.82	2.24
	(8.4)	(1.8)	(9.64×10^{-3})	(0.23)
(0.5, 6)	63.8	62.0	2.82	2.30
	(5.7)	(3.0)	(8.92×10^{-3})	(0.20)
(0.5, 8)	44.2	63.5	2.82	2.31
	(6.8)	(2.9)	(9.59×10^{-3})	(0.24)
(0.5, 10)	37.8	60.7	2.82	2.37
	(6.5)	(1.8)	(8.64×10^{-3})	(0.29)
(0.2, 6)	62.8	65.0	2.80	2.29
	(7.2)	(4.9)	(1.28×10^{-2})	(0.25)
(0.2, 8)	59.8	64.0	2.82	2.30
	(7.5)	(3.9)	(1.02×10^{-2})	(0.28)
(0.2, 10)	47.8	66.0	2.82	2.31
	(6.1)	(3.5)	(9.16×10^{-3})	(0.22)

So far, we have used Legendre polynomials of degree $\nu = 3$. We recommend specifying ν large enough so that satisfactory approximations are provided for all non-linear effects. To demonstrate that the procedure works well as long as ν is not under-specified, we re-analyze all simulated data with $\nu = 4$ and report results in Table 5. It can be seen that all statistical power and false-positive rates are on the same level, suggesting that larger ν does not impact the performance of the TS-NIS procedure in an obvious way.

REAL DATA ANALYSIS

We apply the proposed procedure to a real data set from Framingham Heart Study [33]. In this study, 977 subjects including 418 males and 559 females are randomly selected from Framingham, Massachusetts, and are genotyped, and their BMI, sex and age are measured. After excluding SNPs with minor allele frequency <10%, a total of 349 985 SNPs is used for detecting significant main effects and epistatic effects that are associated with BMI. We implement the proposed procedure with sex and age being two non-genetic covariates.

After TS-NIS followed by LASSO regression, 41 main effect SNPs and 36 SNP–SNP interactions are identified. To enhance the model interpretability, we refit the final model by including two covariates, all selected SNPs and pairwise interactions in a genetic model with both additive and dominant effects. In Table 6, we summarize SNP information with non-zero main genetic effects. Table 7 provides similar information for epistatic interactions. In sum, 41 main genetic effects contribute to 14.40% of the phenotypic variation, whereas epistatic SNP–SNP interactions contribute to 24.38%. Of the contribution of main effect SNPs, 3.56% are from additive effects and 10.84% are from dominant effects. Out of the contribution of epistatic interactions, 8.74% are from additive × additive interactions, 2.55% are from additive × dominant interactions, 6.62% are from dominant × additive interactions and 6.47% are from dominant × dominant interactions. This result implies that gene–gene interactions may play a more important role in the genetic determinants of BMI.

In our genetic model, we assume weak heredity condition by allowing one marginally unimportant SNP in an interaction. If strong heredity condition is assumed instead, all interactions could only explain 11.06% of the phenotypic variation, less than half of the epistatic heritability implied by the proposed model. Therefore, in practice, epistatic models that only consider marginally important SNPs may underestimate the contribution of genetic risk factors.

In terms of effects from non-genetic factors, regression coefficient from age and gender are–0.0107 and 0.0598, respectively, and both are statistically significant. This implies that the risk of obesity increases with age, and females tend to be associated with higher risks of obesity. The identification of BMI susceptibility genes suggests novel metabolic pathways and provides new potential drug targets from the perspective of pharmacogenomics.

Table 5: Power and FPR of simulated data using TS-NIS and TS-NIS-LASSO ($v = 4$)

(ρ, σ^2)	Power (%)		False-positive rate ($\times 10^{-4}$)	
	TS-NIS	TS-NIS-LASSO	TS-NIS	TS-NIS-LASSO
(0.8, 6)	98.2	96.5	2.45	0.75
	(2.45)	(3.3)	(9.65×10^{-3})	(0.12)
(0.8, 8)	97.8	96.3	2.45	0.77
	(2.52)	(3.7)	(9.93×10^{-3})	(0.10)
(0.8, 10)	97.3	95.0	2.45	0.80
	(2.86)	(4.4)	(1.12×10^{-2})	(0.10)
(0.5, 6)	98.7	98.3	2.45	0.87
	(2.91)	(3.3)	(1.15×10^{-2})	(0.14)
(0.5, 8)	98.5	98.2	2.45	0.90
	(3.26)	(3.8)	(1.28×10^{-2})	(0.13)
(0.5, 10)	98.3	97.5	2.45	0.89
	(3.30)	(4.1)	(1.30×10^{-2})	(0.12)
(0.2, 6)	98.8	98.7	2.45	0.85
	(2.53)	(2.6)	(9.93×10^{-3})	(0.12)
(0.2, 8)	98.8	98.2	2.45	0.86
	(2.53)	(2.8)	(9.93×10^{-3})	(0.10)
(0.2, 10)	98.7	97.3	2.45	0.88
	(2.61)	(3.1)	(1.03×10^{-2})	(0.12)

DISCUSSION

Although statistical tools that analyze all SNPs simultaneously prove valuable to a deeper understanding of the genetic components of human diseases and traits, restrictive assumptions of genetic models may not capture the complex genetic architecture underlying regulatory pathways. Without a rule to determine the optimal genetic models, different genetic models have to be considered jointly [12]. On the other hand, models without epistatic interactions are usually associated with large fractions of the 'missing heritability' [4,10]. In the presence of ultrahigh-dimensional genetic data, addressing these two problems is challenging.

Because a single genetic model can hardly hold for all genetic predictors jointly, in this article we propose an ultrahigh-dimensional non-parametric genetic model. This genetic model could

Table 6: Main effects detected in the real data study

Additive effects					Dominant effects				
Chr	Name	MAF	Effect	Heritability (%)	Chr	Name	MAF	Effect	Heritability (%)
1	rs2236817	0.13	−0.5371	0.0859	1	rs10801706	0.33	−0.6247	0.4885
1	rs17406104	0.13	0.6896	0.1370	1	rs10801713	0.35	−0.3869	0.1921
1	rs723015	0.10	0.4702	0.0421	3	rs698210	0.25	−0.5634	0.3171
1	rs2821309	0.48	−0.4892	0.3406	3	rs9862388	0.29	0.5425	0.3323
3	rs698210	0.38	−0.4104	0.2132	3	rs7637740	0.30	0.6891	0.5578
5	rs275442	0.38	0.4835	0.2994	4	rs12511093	0.27	0.5480	0.3227
5	rs4920897	0.38	0.4952	0.3088	5	rs158999	0.25	−0.7333	0.5443
6	rs4945604	0.36	−0.3785	0.1729	6	rs1535556	0.34	−0.4021	0.2055
7	rs2533449	0.33	−0.3992	0.1767	6	rs4945604	0.29	1.1102	1.4055
8	rs10951131	0.14	0.4304	0.0600	7	rs17282885	0.32	−0.5914	0.4319
8	rs10950919	0.37	0.4091	0.2098	7	rs2533449	0.30	0.7714	0.7024
8	rs2053313	0.23	−0.4530	0.1448	8	rs10950919	0.27	−0.8982	0.8584
8	rs16887751	0.28	0.3709	0.1266	8	rs13235166	0.27	0.5736	0.3541
9	rs803917	0.28	0.7125	0.4744	9	rs10756080	0.39	0.4195	0.2378
12	rs1607868	0.34	−0.6390	0.4665	10	rs303207	0.26	−0.4769	0.2408
12	rs1520779	0.24	0.5204	0.2039	12	rs10876943	0.29	−0.4743	0.2547
13	rs17632604	0.22	0.3750	0.0970	13	rs2858978	0.39	−0.5925	0.4756
					13	rs17632604	0.34	−0.9074	1.0430
					14	rs8007682	0.27	0.3966	0.1682
					18	rs1866854	0.30	0.6447	0.4865
					18	rs16940905	0.15	−0.8443	0.4036
					18	rs1605973	0.26	−0.5323	0.2906
					19	rs11670504	0.28	−0.4871	0.2661
					20	rs206815	0.29	0.4798	0.2604

Table 7: Epistatic interactions in the real data study

Root 1			Root 2			Effect	Heritability (%)
Chr	**Name**	**MAF**	**Chr**	**Name**	**MAF**		
Additive × additive interactions							
20	rs852027	0.10	17	rs10468553	0.14	−0.2430	0.2795
1	rs1158407	0.10	17	rs2586118	0.20	−0.3109	0.3393
1	rs1158407	0.10	17	rs11079178	0.37	−0.2472	0.3207
1	rs6668038	0.23	8	rs2527036	0.21	−0.3213	0.4645
18	rs16940905	0.38	8	rs10954339	0.14	−0.7363	2.9192
18	rs16940905	0.38	8	rs3095006	0.14	−0.2936	0.4624
1	rs2096148	0.32	14	rs1177586	0.19	−0.3438	0.6213
1	rs619311	0.12	14	rs17097936	0.11	−0.1655	0.0660
18	rs27411182	0.22	14	rs17671047	0.15	−0.9016	3.2180
1	rs1158407	0.10	14	rs4363775	0.14	0.1381	0.0485
Additive × dominant interactions							
1	rs1158407	0.10	1	rs1777258	0.18	0.4492	0.5280
20	rs6084164	0.16	7	rs12703874	0.27	0.1623	0.0836
1	rs2096148	0.32	10	rs2994665	0.26	−0.1777	0.1006
1	rs1158407	0.10	20	rs6011600	0.11	0.1532	0.0477
4	rs2725771	0.28	20	rs11697106	0.14	−0.4303	0.5873
8	rs10950583	0.30	12	rs758163	0.22	0.5362	0.9261
1	rs1158407	0.10	2	rs12714205	0.14	0.3533	0.2731
Dominant × additive Interactions							
1	rs2096148	0.32	11	rs10792757	0.17	−0.3411	0.3747
4	rs2725771	0.28	5	rs7443778	0.41	−0.6713	1.3961
4	rs2522474	0.29	5	rs4702271	0.20	−0.2084	0.1398
1	rs10801706	0.23	5	rs249721	0.17	−0.6133	1.1710
18	rs16940905	0.38	6	rs1343488	0.18	0.2908	0.2702
13	rs2858978	0.13	6	rs332562	0.14	−0.3926	0.3822
1	rs2096148	0.32	6	rs9344765	0.21	−0.1611	0.0835
19	rs11670504	0.33	6	rs1325476	0.28	−0.5477	0.9472
5	rs269696	0.24	3	rs9877175	0.36	0.5899	1.1060
1	rs1158407	0.10	3	rs13095765	0.15	−0.1466	0.0501
6	rs1535556	0.23	3	rs11711363	0.25	0.1595	0.0813
1	rs2096148	0.32	3	rs640039	0.15	−0.4386	0.6191
Dominant × dominant interactions							
12	rs10876943	0.36	2	rs10496319	0.36	0.7408	3.1063
1	rs1158407	0.10	2	rs1922289	0.29	0.2399	0.2849
20	rs852027	0.26	18	rs7237668	0.15	−0.3181	0.5093
18	rs16940905	0.38	18	rs17657594	0.13	0.1545	0.1315
1	rs2096148	0.32	21	rs718099	0.22	0.2204	0.2642
1	rs10801706	0.23	22	rs714026	0.33	−0.5387	1.5866
1	rs1158407	0.10	13	rs2390886	0.23	−0.3620	0.5905

characterize different modes of genotype–phenotype associations for different genetic variants, and thus offers unprecedented flexibility in analyzing genome-wide genetic data. Moreover, we propose a TS-NIS procedure to identify a reduced model for variable selections. With sure independence property, this procedure eliminates the majority of irrelevant genetic variants, and recent developments in variable selection-guided GWAS analysis could be incorporated.

We use the newly developed procedure to identify genetic variants that are associated with BMI. Surprisingly, the proportion of BMI variations explained by epistatic interactions is greater than that explained by the main genetic effects, and many epistatic interactions exhibit notable heritabilities. These findings indicate novel BMI-associated metabolic pathways, and provide evidence for better understanding the underlying genetic mechanisms of obesity. For example, dopamine in the brain is a neurotransmitter that surges a 'feel-good' hormone after eating. Wang *et al*. [11] found that the brain dopamine levels are significantly lower in the obese individuals, suggesting genetic regulatory networks underlying BMI.

This work explores a flexible genetic model with two-way interactions. In the future, this framework could be extended to higher-order interactions and GWAS with case–control cohorts. For example, we could add more stages to the current framework of TS-NIS. In particular, after selecting important epistatic SNP–SNP interactions, an SNP–SNP–SNP interaction triplet can be formed by combining one important SNP–SNP interaction (selected in the second stage) with any SNP from the genome. Then the same variable screening procedure can be extended to select important SNP–SNP–SNP interactions. This procedure is also implied by the weak heredity condition, and enjoys sure screening property.

In our software package, the reduced model suggested by TS-NIS is analyzed by LASSO regression. However, extensions of LASSO have been proposed to enhance the statistical power and reduce the false-positive rate of GWAS analysis [8–10,34,35]. These approaches could also be used in analyzing the reduced model suggested by TS-NIS, allowing more customized genetic analyses. In addition, the identified disease susceptibility variants and interactions could be better understood through gene-set enrichment analysis tools [14,36], which will provide valuable insight into functional related genes and the complex genetic architecture of regulatory pathways.

Key Points

- In analyzing ultrahigh-dimensional data sets from GWAS, a model-free framework is proposed to identify SNPs exhibiting main genetic effects and epistatic interactions. This framework nests various genetic models.
- A TS-NIS procedure is formulated to identify a subset of SNPs for the following variable selection, where SNPs that are marginally uncorrelated with the phenotype but are involved in SNP–SNP interactions could be identified.
- After TS-NIS, the LASSO regression with coordinate descent steps is applied to select truly important SNPs and estimate their genetic effects. Because two-stage variable screening greatly reduces the model dimensionality, truly important SNPs can be efficiently identified by LASSO.
- The proposed method is validated by using both extensive computer simulations and a real data set from Framingham Heart study, where epistatic interactions are found to be more important than individual main genetic effects in explaining BMI variations.

ACKNOWLEDGEMENTS

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

FUNDING

Startup Grant from The University of Notre Dame [341410] (to J.L.).

References

1. Burton PR, Clayton DG, Cardon LR, *et al*. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447**:661–78.
2. Daly AK. Genome-wide association studies in pharmacogenomics. *Nat Rev Genet* 2010;**11**:241–6.
3. McCarthy MI, Abecasis GR, Cardon LR, *et al*. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;**9**:356–69.
4. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
5. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 1996;**58**:267–88.
6. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Stat* 2008;**2**:224–44.
7. Wu TT, Chen YF, Hastie T, *et al*. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 2009;**25**:714–21.
8. Cho S, Kim H, Oh S, *et al*. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. In: *BMC Proceedings 2009*, Vol. 3. No. Suppl 7, p. S25. BioMed Central Ltd.
9. Li J, Das K, Fu G, *et al*. The Bayesian lasso for genome-wide association studies. *Bioinformatics* 2011;**27**:516–23.
10. He Q, Lin DY. A variable selection method for genome-wide association studies. *Bioinformatics* 2011;**27**:1–8.
11. Wang Y, Liu G, Feng M, *et al*. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* 2011;**27**:2936–43.
12. Lettre G, Lange C, Hirschhorn JN. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* 2007;**31**:358–62.
13. Lange K, Cantor R, Horvath S, *et al*. 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* 2001;**69(Suppl. 1)**:A1886.
14. Solé X, Guinó E, Valls J, *et al*. SNPStats: a web tool for the analysis of association studies. *Bioinformatics* 2006;**22**:1928–9.
15. Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J Am Stat Assoc* 2011;**106**:494.

16. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B* 2008;**70**:849–911.
17. Stone CJ. Additive regression and other nonparametric models. *Ann Stat* 1985;**13**:689–705.
18. Chipman H. Bayesian variable selection with related predictors. *Canadian J Stat* 1996;**24**:17–36.
19. Cui Y, Zhu J, Wu R. Functional mapping for genetic control of programmed cell death. *Physiol Genom* 2006;**25**: 458–69.
20. Zhao W, Wu R. Wavelet-based nonparametric functional mapping of longitudinal curves. *J Am Stat Assoc* 2008;**103**:482.
21. Xing J, Li J, Yang R, *et al*. Bayesian B-spline mapping for dynamic quantitative traits. *Genet Res* 2012;**94**:85–95.
22. Koltchinskii V, Yuan M. Sparse recovery in large ensembles of kernel machines. In: Servedio RA, Zhang T (eds). *21st Annual Conference on Learning Theory—COLT 2008, Helsinki, Finland, July 9–12*. Omnipress, 2008;229–38.
23. Ravikumar P, Liu H, Lafferty J, *et al*. Spam: sparse additive models. *J R Stat Soc Ser B* 2009;**71**:1009–30.
24. Meier L, Geer V, Bühlmann P. High-dimensional additive modeling. *Ann Stat* 2009;**37**:3779–821.
25. Huang J, Horowitz J, Wei F. Variable selection in nonparametric additive models. *Ann Stat* 2010;**38**:2282–313.
26. Wu R, Ma C, Lin M, *et al*. A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* 2004;**166**:1541–51.
27. Cui Y, Wu R. Mapping genome–genome epistasis: a high-dimensional model. *Bioinformatics* 2005;**21**:2447–55.
28. Wills D M, Burke J M. Quantitative trait locus analysis of the early domestication of sunflower. *Genetics* 2007;**176**: 2589–99.
29. Wang Y, Liu G, Feng M, *et al*. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* 2011;**27**:2936–43.
30. Wu J, Devlin B, Ringquist S, *et al*. Screen and Clean: a tool for identifying interactions in genome-wide association studies. *Genet Epidemiol* 2010;**34**:275–85.
31. Gao H, Wu Y, Li J, *et al*. Forward LASSO analysis for high-order interactions in genome-wide association study. *Brief Bioinformatics* 2013;**15**:552–61.
32. Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;**69**:138–47.
33. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the framingham study. *Am J Public Health* 1951;**41**:279–86.
34. Shi G, Boerwinkle E, Morrison AC, *et al*. Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genet Epidemiol* 2011;**35**:111–18.
35. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol* 2010;**34**:879–91.
36. Holden M, Deng S, Wojnowski L, *et al*. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 2008;**24**: 2784–5.