An Efficient Nonlinear Regression Approach for Genome-wide Detection of Marginal and Interacting Genetic Variations

Seunghak Lee[1], Aurélie Lozano[2], Prabhanjan Kambadur[3], and Eric P. Xing[1,⋆]

[1]School of Computer Science, Carnegie Mellon University, USA
[2]IBM T. J. Watson Research Center, USA
[3]Bloomberg L.P., USA
epxing@cs.cmu.edu

Abstract. Genome-wide association studies have revealed individual genetic variants associated with phenotypic traits such as disease risk and gene expressions. However, detecting pairwise interaction effects of genetic variants on traits still remains a challenge due to a large number of combinations of variants ($\sim 10^{11}$ SNP pairs in the human genome), and relatively small sample sizes (typically $< 10^4$). Despite recent breakthroughs in detecting interaction effects, there are still several open problems, including: (1) how to quickly process a large number of SNP pairs, (2) how to distinguish between true signals and SNPs/SNP pairs merely correlated with true signals, (3) how to detect non-linear associations between SNP pairs and traits given small sample sizes, and (4) how to control false positives? In this paper, we present a unified framework, called SPHINX, which addresses the aforementioned challenges. We first propose a piecewise linear model for interaction detection because it is simple enough to estimate model parameters given small sample sizes but complex enough to capture non-linear interaction effects. Then, based on the piecewise linear model, we introduce randomized group lasso under stability selection, and a screening algorithm to address the statistical and computational challenges mentioned above. In our experiments, we first demonstrate that SPHINX achieves better power than existing methods for interaction detection under false positive control. We further applied SPHINX to late-onset Alzheimer's disease dataset, and report 16 SNPs and 17 SNP pairs associated with gene traits. We also present a highly scalable implementation of our screening algorithm which can screen ~ 118 billion candidates of associations on a 60-node cluster in < 5.5 hours.

1 Introduction

A fundamental problem in genetics is to understand the interaction (or epistatic) effects from pairs of or multiple single-nucleotide polymorphisms (SNPs) on phenotypic traits [33]. Existing methods for detecting causal SNP pairs include hypothesis-testing-based methods [36, 38, 42] and penalized multivariate regression (PMR) based methods [3, 21, 34]. Arguably PMR-based methods are more powerful than hypothesis-testing-based methods because PMR can in principle jointly estimate all marginal and interaction effects simultaneously [17, 21]. However, statistical and computational bottlenecks have prevented PMR from being widely used for detecting interaction effects on traits. Firstly, it is difficult to control false positives. One can use a "screen and clean" procedure to compute p-values [30, 39], but this strategy substantially downgrades the power in genome wide association mapping because only half of the samples can be used for each step of screening and cleaning. Secondly, the high correlations between pairs of SNPs also lead to decrease of the power of PMR, because PMR can only detect true associations accurately under conditions with little correlation between different SNPs/SNP pairs [6]. Lastly, there is a substantial computational challenge to overcome. If we were to consider millions of SNPs as candidates in studying a particular phenotypic trait, the number of potential pairwise interactions between pairs of SNPs to be considered is $> 10^{11}$. Such a massive pool of candidates of SNP pairs makes it infeasible to solve the mathematical optimization program underlying PMR with currently available tools.

The past several years have seen the emergence of several statistical methods that can potentially be employed to address the problems mentioned above. For the first problem of error control, Meinshausen and Bühlmann proposed a procedure known as *stability selection* [29]. The insight behind this technique is that, given randomly chosen multiple subsamples, true associations of covariates (e.g., SNPs or SNP

⋆ To whom correspondence should be addressed.

pairs) to responses (e.g., a trait) will be selected at high frequency because true association signals are likely to be insensitive to the random selection of subsamples. Second, to address the non-identifiability problem in regression due to inter-covariate correlation, a *randomized lasso* technique has been proposed that randomly perturbs the scale of covariates in the framework of stability selection, thereby relaxes the original requirements on small correlation for recovery of true association signals from all covariates [29]. Naturally, such a scheme is expected to help distinguishing between true and false associations of SNPs/SNP pairs because only true ones are likely to be selected under the perturbations. Finally, to combat the computational challenge due to a massive number of covariates, a *sure independence screening* (SIS) procedure [10] has been proposed to contain the operational size of the regression problem under provable guarantee of retaining true signals. It is possible to use the idea behind SIS to effectively perform simple independent tests on each pair of SNPs (or individual SNPs) and discard the large fraction of candidates with no associations, such that one can end up with only $O(NC)$ candidates (where N is sample size and C is a data dependent constant) of which no true associations will be missed with high probability. These theoretical development notwithstanding, their promised power remains largely unleashed for practical genome wide association mapping, especially in nontrivial scenarios such as non-additive epistatic effects, due to several remaining hurdles, including proper models for association, algorithms for screening with such models and on a computer cluster, and proper integration of techniques for error control, identifiability, screening, etc., in such a new paradigm.

In this paper, we present SPHINX (which stems from Sparse Piecewise linear model with High throughput screening for INteraction detection(X)), a new PMR-based approach built on the advancements in statistical methodologies mentioned above. It is an integrative platform that conjoins and extends the aforementioned three components, further enhanced with techniques allowing more realistic trait association patterns to be detected. In particular, SPHINX is designed to capture SNP pairs with non-linear interaction effects (synergistic/antagonistic epistasis) on traits using a piecewise linear model (PLM), which is better suited to model the complex interactions between a pair of SNPs and the traits. In short, SPHINX is designed as follows: using an extension of SIS based on PLM, it first selects a set of $O(NC)$ SNPs and SNP pairs with the smallest residual sum of squares. Then it runs the randomized group lasso based on PLM on the set of SNPs and SNP pairs selected in the previous step under stability selection. Finally, it reports SNPs and SNP pairs selected by stability selection, whose coefficients are non-zero given a majority of subsamples. In Fig. 1, we illustrate the overall framework of SPHINX. Note that in practical association analysis with all pairs of SNPs, we should address the three problems mentioned above simultaneously, which is a non-trivial task. To achieve this goal, we take the approach of unified framework, which requires statistically sound models and algorithms, and scalable system implementations.

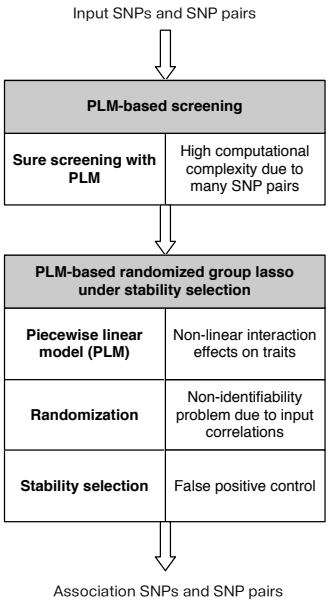


Fig. 1. Overall framework of SPHINX. Using a screening method, we first discard SNPs/SNP pairs without associations; given the SNPs/SNP pairs survived in the screening step, we run a method that incorporates three different techniques, each of which is introduced to address the problem on its right side.

In our experiments, we show the efficacy of SPHINX in controlling false positives, detecting true causal SNPs and SNP pairs, and using multiple cores/machines to deal with a large number of SNP pairs. Furthermore, with SPHINX, we analyzed late-onset Alzheimer's disease eQTL dataset [41], which contains $\sim$ 118 billion candidates of associations; the analysis took < 5.5 hours using 60-node cluster with 720 cores. As a result, we found 16 SNPs and 17 SNP pairs associated with gene traits. Among our findings, we report the analysis of 6 SNPs (*rs1619379*, *rs2734986*, *rs1611710*, *rs2395175*, *rs3135363*, *rs602875*) associated with immune system-related genes (i.e., *HLA* gene family) and a SNP pair (pair of *rs4272759* and *rs6081791*) associated with a dopamine-related gene (i.e., *DAT* gene); the role of dopamine and immune system in Alzheimer's disease have been studied in previous research [25, 28].

2 Methods

SPHINX is a framework for genome-wide association mapping, which consists of PLM-based screening technique, PLM-based randomized group lasso under stability selection. Among the SPHINX components, the effectiveness of the randomization technique and stability selection are demonstrated in [10, 29] with theory and experiments; the screening approach is extensively studied in both parametric and nonparametric settings [9,10]. In this section, we focus on describing our proposed novel model PLM-based group lasso with the randomization technique and stability selection. We then present the PLM-based screening method, followed by our system implementation of the screening method. Note that SPHINX runs the screening method prior to the PLM-based randomized group lasso, as shown in Fig. 1.

2.1 Piecewise Linear Model-Based Group Lasso

The relationships between genetic variations and phenotypic traits are complex, for example, non-linear. However, due to the highly under-determined nature of the mathematical problem — too many features (SNPs and SNP pairs) but too few samples — it is difficult to employ models that have a high degree of freedom. Traditionally, linear models have been used extensively in genome wide association studies despite the fact that these models are not flexible enough to capture the complexity of the trait-associated epistatic interactions between SNPs.

We introduce a multivariate piecewise linear model (PLM), which is better suited to model the complex interactions between a pair of SNPs and traits. Note that we employ PLM for adding additional degrees of freedom into a linear model in a high-dimensional multivariate regression setting. Therefore, it is different from the cases, where we change the degrees of freedom in statistical tests such as F-test. We denote the j-th SNP for the i-th individual by $x_j^i \in \{0, 1, 2\}$ with the number of minor alleles at the locus. Let us start converting a linear model into a piecewise linear model with two knots denoted by $\Delta = \{\eta_1, \eta_2\}$, where $\eta_1 = 1$ and $\eta_2 = 2$ for our SNP encoding. It uses three degrees of freedom, flexible enough to capture the change of gene expression with a change in the genotype. Specifically, let m_{jk}^i denote the genotype encoding for the interaction between j-th SNP and k-th SNP for the i-th individual, i.e., $m_{jk}^i \equiv x_j^i x_k^i$. Then, we have a piecewise linear model as follows:

$$\hat{\mathbf{y}} = \mathbf{1}C + \sum_{j=1}^{P} \mathbf{x}_j \beta_j + \sum_{j<k} \Psi(\boldsymbol{m}_{jk}, \{u_{jk}, t_{jk}, w_{jk}\}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{m}_{jk} = [m_{jk}^1, \ldots, m_{jk}^N]^T$, $\hat{\mathbf{y}}$ is an output trait based on the model, β_j is the regression coefficient for the j-th SNP, and $\boldsymbol{\epsilon}$ is Gaussian noise. Here, $\Psi(\cdot)$ is a piecewise linear function given by

$$\Psi(m_{jk}^i, \{u_{jk}, t_{jk}, w_{jk}\}) = \begin{cases} m_{jk}^i u_{jk} & \text{if } m_{jk}^i \le \eta_1, \\ m_{jk}^i u_{jk} + (m_{jk}^i - \eta_1) t_{jk} & \text{if } \eta_1 < m_{jk}^i \le \eta_2, \\ m_{jk}^i u_{jk} + (m_{jk}^i - \eta_1) t_{jk} + (m_{jk}^i - \eta_2) w_{jk} & \text{if } m_{jk}^i > \eta_2, \end{cases} \tag{2}$$

where u_{jk}, t_{jk} and w_{jk} represent the regression coefficients for the first, second and third line segment, respectively. Given the model in Eq. (1), to select significant SNPs/SNP pairs, we propose the following penalized multivariate piecewise-linear regression, referred to as PLM-based group lasso:

$$\min_{C, \{\beta_j\}, \{u_{jk}, t_{jk}, w_{jk}\}} \left\| \mathbf{y} - \left\{ \mathbf{1}C + \sum_{j=1}^{P} \mathbf{x}_j \beta_j + \sum_{j<k} \Psi(\boldsymbol{m}_{jk}, \{u_{jk}, t_{jk}, w_{jk}\}) \right\} \right\|_2^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| + \lambda_2 \sum_{j<k} \sqrt{|\Delta|} \sqrt{u_{jk}^2 + t_{jk}^2 + w_{jk}^2}, \tag{3}$$

where λ_1 and λ_2 are regularization parameters, determining the sparsity of the solutions. Here the first ℓ_1 and second ℓ_1/ℓ_2 norm are introduced to set the coefficients of individual SNPs and SNP pairs to exactly zero respectively if they are irrelevant to the observed trait $\mathbf{y}$. It is equivalent to group lasso penalty [40], and has been shown that it allows us to select true non-zero β_js and $\{u_{jk}, t_{jk}, w_{jk}\}$s under certain conditions [1]. We can optimize Eq. (3) using standard optimization techniques for group lasso such as a block coordinate descent [12], or a proximal gradient method [27] (we used a proximal gradient method to optimize Eq. (3) [26]) because the loss function is differentiable and $\Psi(\cdot)$ is linear. Further, the penalty is separable because there is no overlap between different groups of coefficients. Here, we considered the squared-loss for eQTL (expression quantitative trait loci) mapping with continuous traits; however, our methodology can be extended to other loss functions (e.g logistic loss in case/control studies).

Algorithm 1: PLM-based Randomized Group Lasso under Stability Selection

Input: $\mathbf{X}$: SNP matrix, $\mathbf{y}$: Expression of a single gene, Ω_{mar}: selected individual SNPs by screening, Ω_{int}: selected pairs of SNPs by screening, κ: maximum expected number of false positives, $\{\Lambda_t\}_{t=1}^T$: a set of regularization parameters of (λ_1, λ_2), π_{thr}: threshold for stability selection ($0.5 < \pi_{thr} \leq 1$).

Output: $\hat{S}_{mar}$: selected individual SNPs, $\hat{S}_{int}$: selected pairs of SNPs.

$\Pi_j^t = 0, j \in \Omega_{mar}$; $\Pi_{jk}^t = 0, (j,k) \in \Omega_{int}$ and $t = 1, \ldots, T$.
for $t = 1$ *to* T **do**
  $(\lambda_1, \lambda_2) \leftarrow \Lambda_t$.
  $s_j = 0, \forall j \in \Omega_{mar}; s_{jk} = 0, \forall (j,k) \in \Omega_{int}$
  for $k = 1$ *to* α **do**
    Randomly select $\lfloor N/2 \rfloor$ samples from N samples without replacement.
    Given the $\lfloor N/2 \rfloor$ subsamples, solve the PLM-based randomized group Lasso in Eq. (4).
    $s_j = s_j + 1$ for all selected individual term j
    $s_{jk} = s_{jk} + 1$ for all selected pairwise terms (j,k)
  Compute the maximum expected number of false positives ($E(V)$) via Eq. (5).
  if $E(V) \leq \kappa$ **then**
    $\Pi_j^t \leftarrow \frac{s_j}{\alpha}$ for all $j \in \Omega_{mar}$ $\Pi_{jk}^t \leftarrow \frac{s_{jk}}{\alpha}$ for all $(j,k) \in \Omega_{int}$
$\hat{S}_{mar} = \{j : \max_t \Pi_j^t \geq \pi_{thr}\}$; $\hat{S}_{int} = \{(j,k) : \max_t \Pi_{jk}^t \geq \pi_{thr}\}$

Randomization As previously mentioned, high correlations between SNPs or SNP pairs make it hard to distinguish between true association SNPs/SNP pairs and the correlated ones. To address the problem, we randomly perturb the scale of covariates in Eq. (3) [29], called PLM-based randomized group Lasso:

$$\min_{C,\{\beta_j\},\{u_{jk},t_{jk},w_{jk}\}} \left\| \mathbf{y} - \left\{ \mathbf{1}C + \sum_{j=1}^{P} W_j \mathbf{x}_j \beta_j + \sum_{j<k} \Psi(W_{jk} \boldsymbol{m}_{jk}, \{u_{jk}, t_{jk}, w_{jk}\}) \right\} \right\|_2^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{j<k} \sqrt{|\Delta|} \sqrt{u_{jk}^2 + t_{jk}^2 + w_{jk}^2}. \quad (4)$$

Here $P(W_j = 1) = P(W_j = \delta) = P(W_{jk} = 1) = P(W_{jk} = \delta) = 0.5$, and $\delta \in (0, 1]$ determines the degree of perturbations (the smaller δ, the larger perturbations). It has been shown that this randomization with stability selection weakens the condition for the recovery of true non-zero coefficients [29]. Furthermore, Meinshausen and Bühlmann empirically showed that the randomization is very useful to distinguish between true causal signals and the false ones merely correlated with the true signals [29]. However, there is trade-off for the degree of random perturbations: as we increase the degree of perturbations, false positives will be reduced, but true negatives can be increased.

Stability Selection Next, to control false positives, we adopt stability selection [29], which takes the bootstrapping approach. Suppose we have a set of (λ_1, λ_2) parameters denoted by $\{\Lambda_t\}_{t=1}^T$, where $\Lambda_t = (\lambda_1, \lambda_2)_t$. For each Λ_t, we solve Eq. (4) based on randomly chosen samples size of $\lfloor N/2 \rfloor$ for α times. Then we select SNPs or SNP pairs if their coefficients are set to non-zero more than $\pi_{thr}\alpha$ ($0.5 < \pi_{thr} \leq 1$) times for any regularization parameters. Under certain assumptions, it has been shown that the expected number of false positives $E(V)$ is bounded by

$$E(V) \leq \frac{1}{2\pi_{thr} - 1} \frac{q_{\boldsymbol{\Lambda}}^2}{k_{mar} + k_{int}}, \quad (5)$$

where $q_{\boldsymbol{\Lambda}}$ is the expected number of non-zero coefficients in a solution of Eq. (4) [29]. Note that for whole genome-wide association studies, stability selection based on Eq. (4) is computationally challenging due to all SNP pairs considered. Specifically, optimizing Eq. (4) is non-trivial because it requires us to use an iterative algorithm such as a proximal gradient method [26], which sweeps over such a large number of SNP pairs multiple times. To address the problem, we introduce PLM-based screening algorithm, which efficiently gives us small candidate sets of association SNPs and SNP pairs, denoted by Ω_{mar} and Ω_{int}. We describe PLM-based randomized group lasso under stability selection in Algorithm 1.

2.2 Piecewise Linear Model-Based Screening

In Eq. (4), we include all P SNPs and $\binom{P}{2}$ ($= \frac{P(P-1)}{2}$) SNP pairs, which makes it impractical solve the problem at a whole-genome scale (e.g., millions of SNPs). To handle the quadratic explosion of the number

Algorithm 2: Piecewise Linear Model-Based Screening

Input: $\mathbf{X}$: SNP matrix, $\mathbf{y}$: Expression of a single gene, N: number of samples, P: number of SNPs, b_{mar}: number of SNPs to select per-iteration, b_{int}: number of SNP pairs to select per-iteration, k_{mar}: total number of SNPs to select, k_{int}: total number of SNP pairs to select.

Output: Ω_{mar}: selected individual SNPs, Ω_{int}: selected pairs of SNPs.

$\Omega_{mar} = \Omega_{int} = \emptyset, \mathbf{r} = \mathbf{y}$;
while $(k_{mar} > |\Omega_{mar}|)$ || $(k_{int} > |\Omega_{int}|)$ **do**
  $n_{mar} = \text{Min}((k_{mar} - |\Omega_{mar}|), b_{mar})$;
  $n_{int} = \text{Min}((k_{int} - |\Omega_{int}|), b_{int})$;
  for *SNP* $j \notin \Omega_{mar}$ **do**
    $RSS_j = \min_{C,\beta_j} ||\mathbf{r} - \mathbf{1}C - \mathbf{x}_j\beta_j||_2^2$;
  Sort(RSS_j,*ascending*);
  $\Omega_{mar} = \Omega_{mar} \cup \text{Top}(RSS_j, n_{mar})$;
  for *SNP* $(j,k) \notin \Omega_{int}$ **do**
    $RSS_{jk} = \min_{C,\{u_{jk},t_{jk},w_{jk}\}} ||\mathbf{r} - \mathbf{1}C - \Psi(\boldsymbol{m}_{jk}, \{u_{jk}, t_{jk}, w_{jk}\})||_2^2$;
  Sort(RSS_{jk},*ascending*);
  $\Omega_{int} = \Omega_{int} \cup \text{Top}(RSS_{jk}, n_{int})$;
  $(\hat{C}, \hat{\boldsymbol{\beta}}, \{\hat{\boldsymbol{u}}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{w}}\}) = \underset{C,\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{t},\boldsymbol{w}}{\text{argmin}} ||\mathbf{y} - \mathbf{1}C - \sum_{j\in\Omega_{mar}} \mathbf{x}_j\beta_j - \sum_{(j,k)\in\Omega_{int}} \Psi(\boldsymbol{m}_{jk}, \{u_{jk}, t_{jk}, w_{jk}\})||_2^2$;
  $\mathbf{r} = \mathbf{y} - \mathbf{1}\hat{C} - \sum_{j\in\Omega_{mar}} \mathbf{x}_j\hat{\beta}_j - \sum_{(j,k)\in\Omega_{int}} \Psi(\boldsymbol{m}_{jk}, \{\hat{u}_{jk}, \hat{t}_{jk}, \hat{w}_{jk}\})$;

of SNP pairs, we propose a scalable screening method based on PLM. Our screening method is designed to sequentially select potentially relevant SNPs and SNP pairs using a simple test scheme. Note that this screening step focuses on avoiding missing true positives supported by sure-screening theory [9].

Our screening algorithm is a variant of iterative sure independence screening based on Eq. (1). It greedily selects SNPs and SNP pairs based on the contribution of each candidate SNP or SNP pair to the decrease of residual sum of squares. The PLM-based screening is described in Algorithm 2. Note that there are two pairs of parameters (k_{mar}, k_{int}) and (b_{mar}, b_{int}). The pair (k_{mar}, k_{int}) determines the total number of SNPs and SNP pairs selected, and (b_{mar}, b_{int}) determines the number of candidates selected *per-iteration*. In our experiments, we used $(k_{mar}, k_{int}) = (N, N)$ and $(b_{mar}, b_{int}) = (10, 10)$ to limit the number of selected correlated SNPs/SNP pairs at each iteration by 10. When SNPs are highly correlated, we recommend large values of (k_{mar}, k_{int}) and small values of (b_{mar}, b_{int}) because it will allow us to select more independent SNPs/SNP pairs. After the screening step, we obtain small candidate sets of SNPs and SNP pairs, and thus it is computationally tractable to solve the high-dimensional problem in Eq. (4).

2.3 System Implementation of Piecewise Linear Model-Based Screening

We implemented a highly efficient shared- and distributed-memory parallel PLM-based screening algorithm in C++. Our implementation can exploit parallelism when running on multi-core machines, or on clusters of multi-core machines. To exploit shared-memory parallelism, we used PFunc [18], a lightweight and portable library that provides C and C++ APIs to express task parallelism. For distributed-memory parallelism, we used MPI [31, 32], a popular library specification for message-passing that is used extensively in high-performance computing. In this section, we briefly describe some salient features of our implementation that optimize memory and computational efficiency.

First of all, we optimize the memory footprint of SPHINX by storing each SNP using 2 bits (to represent $0, 1, 2$), thereby giving us 4 SNPs per-byte of data. This way, the entire SNP dataset is compressed and most of the operations, such as tests for SNP-SNP interactions, are performed as bit-wise operations. For example, using this scheme, a 200 patient, 500,000 SNP dataset only occupies 250MB of storage that can be entirely cached in-memory on most modern machines. The SNP-SNP interaction pairs are constructed on-the-fly in order to save space instead of explicitly storing $\binom{P}{2}$ additional columns.

We also optimize our implementation for computational efficiency. Note that the significant SNPs and SNP-SNP interactions are selected by solving millions of linear systems, followed by computation of the 2-norm of the resulting residual. To quickly solve the linear systems, we use Cholesky factorization [5] because Cholesky decomposition is faster (although less numerically stable in some cases) than other alternatives such as QR decomposition [5] and singular value decomposition (SVD) [14]. Furthermore, we use BLAS and LAPACK kernels to optimize all the linear operations. After the selection of the first SNP and SNP pair,

incremental linear models are built; that is, given a set of selected SNPs and SNP pairs, the best SNP or SNP pair to add to our model has to be determined. As the number of selected candidates increases, the linear system becomes more expensive to solve, thereby making successive later iterations expensive. In order to offset these costs, we resort to using a hand-coded incremental version of Cholesky factorization, which keeps the per-iteration costs near constant.

3 Simulation Study

In this section, we validate the effectiveness of SPHINX in terms of false positive control, statistical power, and the benefits of using a piecewise linear model over a linear model via simulations because ground-truth associations are unknown in real datasets. Furthermore, we show the scalability of our screening implementation on multi-node, multi-core, and hybrid settings. We first conduct an extensive simulation study to demonstrate and statistically validate the efficacy of SPHINX, in comparison to two popular existing approaches: the *two-locus test* by PLINK [36] with the `--epistasis` option and the maximum likelihood method with fully parametrized two-locus model (saturated two-locus test) [8] with Bonferroni correction at significance level 0.01. We set $|\Omega_{mar}| = N$, $|\Omega_{int}| = N$, and $\kappa = 3$ for SPHINX, allowing for three false positives on average, and used the following sequence of regularization parameters: $\lambda_1 = \lambda_2 \in \{0.5, 0.1, 0.05, 0.01, 0.005\}$. We simulated chromosome 1 with 22834 SNPs and 2000 individuals using GWAsimulator [24], and generated traits under additive and non-additive scenarios: association SNP pairs have (1) additive, and (2) non-additive interaction effects. We ran the methods on 50 different data sets generated by randomly choosing 200 samples and 300 consecutive SNPs from the simulated genome for each simulation setting. In our plots, we report the average performance with error bars of 1/2 standard deviation.

3.1 Generation of Simulation Data

Let us denote $\mathbf{S}_1$ by a set of SNPs with marginal effects, and $\mathbf{S}_2$ by a set of SNP pairs with interaction effects. For the additive scenario, we generate simulation data as follows:

$$y_i = \sum_{j \in \mathbf{S}_1} x_j^i \beta_j + \sum_{(j,k) \in \mathbf{S}_2} x_j^i x_k^i \beta_{jk} + \epsilon_i, \tag{6}$$

where y_i is the continuous response (e.g., gene expression level) for the i-th individual, $x_j^i \in \{0, 1, 2\}$ represents the encoding of the j-th SNP for the i-th individual (i.e., the number of minor alleles), ϵ_i represents Gaussian noise with zero mean and unit variance for the i-th individual, and β_j and β_{jk} are constants that represent the size of marginal and interaction effects, respectively.

For the non-additive scenario, we generate simulation data as follows:

$$y_i = \sum_{j \in \mathbf{S}_1} x_j^i \beta_j + \sum_{(j,k) \in \mathbf{S}_2} f(x_j^i x_k^i) + \epsilon_i, \tag{7}$$

where $f(x_j^i x_k^i)$ is given by

$$f(x_j^i x_k^i) = \begin{cases} r_1 : \text{ if } x_j^i x_k^i = 0, \\ r_2 : \text{ if } x_j^i x_k^i = 1, \\ r_3 : \text{ if } x_j^i x_k^i = 2, \\ r_4 : \text{ if } x_j^i x_k^i = 4, \end{cases}$$

where $r_q \sim \text{Unif}(-\beta_{jk}, \beta_{jk})$ for all $q = 1, \ldots, 4$. Note that in this non-additive scenario, the relationship between $\mathbf{y}$ and a pair of SNPs $\mathbf{x}_j$ and $\mathbf{x}_k$ is non-additive due to the function $f(x_j^i x_k^i)$, which randomly assigns the size of interaction effects according to the input genotype $x_j^i x_k^i$.

In our experiments below, we denote N by the sample size, P by the number of SNPs, $\boldsymbol{\nu}$ by the association strength of marginal and interaction effects (i.e., $\boldsymbol{\nu} = \{\beta_j, \beta_{jk}\}$), and $\boldsymbol{\xi}$ by the number of true association SNPs and SNP pairs (i.e., $\boldsymbol{\xi} = \{|\mathbf{S}_1|, |\mathbf{S}_2|\}$). Furthermore, we randomly choose $\mathbf{S}_2$ such that each SNP pair in $\mathbf{S}_2$ has the minor allele frequency less than MAF1 and MAF2. For the set $\mathbf{S}_1$, we randomly choose SNPs with marginal effects among the SNPs with minor allele frequency less than 0.1.

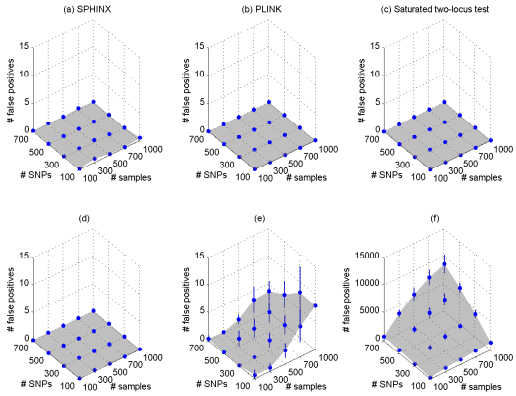


Fig. 2. Number of false positives of SNP pairs found by SPHINX (a,d), PLINK (b,e) and saturated two-locus test (c,f) with different sample sizes and the number of SNPs under two null hypotheses (see text for details).

3.2 False Positive Control

We first confirm that SPHINX effectively controls the number of false positives of SNP pairs under two null hypotheses: (1) there exist no marginal and no interaction effects, (2) there exist only marginal effects but no interaction effects. As shown in Fig. 2, for both null hypotheses, false positives were well controlled with different sample sizes from 100 to 1000 and different numbers of SNPs from 100 to 700 (less than one false positive under both null hypotheses). However, PLINK and the saturated two-locus test did not effectively control the number of false positives under the second scenario (up to 7.94 and 8470, respectively) because SNP pairs correlated with SNPs having some marginal effects were falsely detected.

3.3 Comparison of Different Methods for the Detection of SNP Pairs with Interaction Effects

We present our comparison results among SPHINX, the two-locus test by PLINK [36] with the `--epistasis` option, and the saturated two-locus test [8] with various experimental settings. We evaluate the performance of SPHINX, PLINK [36] and the saturated two-locus test [8] on simulation datasets with different number of true association SNP pairs, different MAFs of true association SNP pairs, and different association strengths.

Comparison with Different Number of True association SNP pairs We performed experiments to show that SPHINX exhibits high power even when false positives are well-suppressed. In the simulation, we randomly chose three SNPs for marginal effects (out of 300 SNPs), and set the number of SNP pairs from 1 to 5 (out of 44850 possible pairs) for interaction effects (SNPs with minor allele frequency between 0 and 0.2 were randomly chosen). Compared to PLINK and the saturated two-locus test, as shown in Fig. 3, SPHINX showed significantly larger true positive rates (up to $\sim 40\%$) while generating fewer number of false positives (< 0.18) under both scenarios of additive and non-additive interaction effects. PLINK found a smaller fraction of SNP pairs with true interaction effects (up to $\sim 10\%$), and the number of false positives was less than 1.04. The saturated two-locus test found more true positives than PLINK but the number of false positives was very large (> 1000).

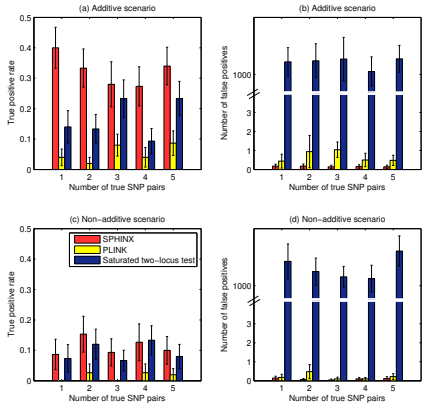
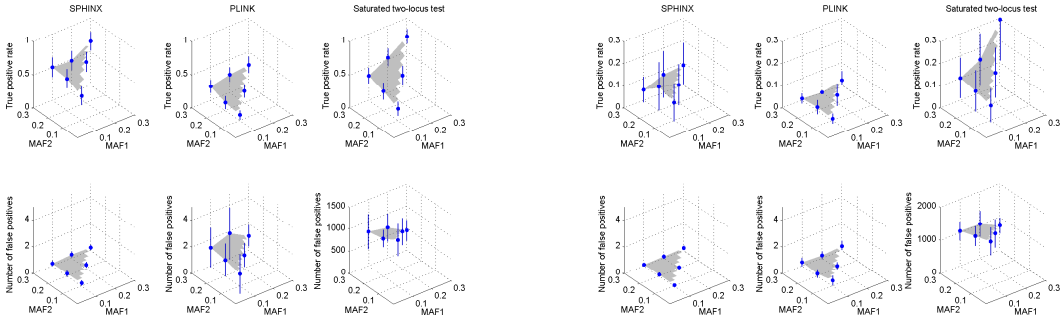


Fig. 3. Comparison of true positive rate and the number of false positives among SPHINX, PLINK and saturated two-locus test with different number of true association SNP pairs under the additive scenario (a,b) and the non-additive scenario (c,d).



(a) Additive scenario (b) Non-additive scenario

Fig. 4. Comparison of true positive rate and the number of false positives for SPHINX (first column), two-locus test by PLINK (second column), and saturated two-locus test (third column) under the linear scenario (a) and the non-linear scenario (b). In each panel, x-axis and y-axis show MAFs of true association SNP pairs (MAF1, MAF2), and z-axis represents the true positive rate or the number of false positives.

Comparison with Different Minor Allele Frequencies We evaluated the three different methods on simulation datasets with $N = 200$ (sample size), $P = 300$ (the number of SNPs), $\boldsymbol{\nu} = \{3, 3\}$ (association strength of marginal and interaction effects), and $\boldsymbol{\xi} = \{3, 3\}$ (the number of true association SNPs and SNP pairs). Fig. 4 shows true positive rate and the number of false positives of the three different methods (columns) with different MAFs of true association SNP pairs (i.e., MAF1=MAF2 $\in \{0.1, 0.2, 0.3\}$) under the linear scenario (Fig. 4(a)) and the non-linear scenario (Fig. 4(b)). Overall, SPHINX achieved the best performance considering both true positive rate and the number of false positives. When we compare between PLINK and SPHINX, under both the additive and non-additive scenarios, SPHINX showed significantly better true positive rate than PLINK while producing fewer number of false positives. Furthermore, SPHINX effectively controlled the number of false positives over all regions of MAFs, showing that the theory of stability selection [29] is in agreement with the empirical results (e.g., under the additive scenario, SPHINX had 0.12 false positives on average). When we compare between the saturated two-locus test and SPHINX, for both scenarios, the saturated two-locus test found slightly more true positives but much larger number of false positives than SPHINX, which makes the saturated two-locus test impractical. It can be explained by the fact that many parameters in the saturated two-locus test led to over-fitting of the model.

Comparison with Different Association Strengths We also tested the three methods with different association strengths $\nu_1 = \nu_2 = 1, \ldots, 5$ ($N = 200, P = 300$, MAF1 = MAF2 = 0.1, $\boldsymbol{\xi} = \{3, 3\}$), and show true positive rate and the number of false positives under the additive scenario in Fig. 5(a,b) and under the non-additive scenario in Fig. 5(c,d). Overall, SPHINX showed the best performance among the three methods as it found a relatively large number of true positives while effectively suppressing false positives over all association strengths. Furthermore, under the non-additive scenario, only SPHINX effectively increased true positive rate as association strength increased under the control of false positives. PLINK showed very low true positive rate (true positive rate was < 0.05 for all association strengths), and the saturated two-locus test produced many false positives (> 200 in most cases).

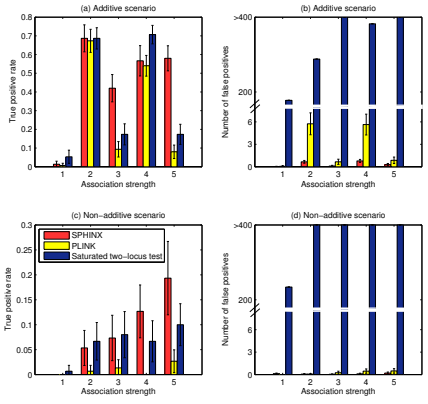


Fig. 5. Comparison of true positive rate and the number of false positives among SPHINX, PLINK and saturated two-locus test with different association strengths under the additive scenario (a,b) and the non-additive scenario (c,d).

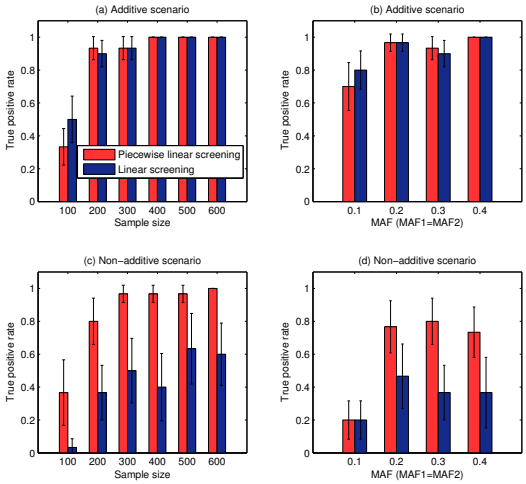


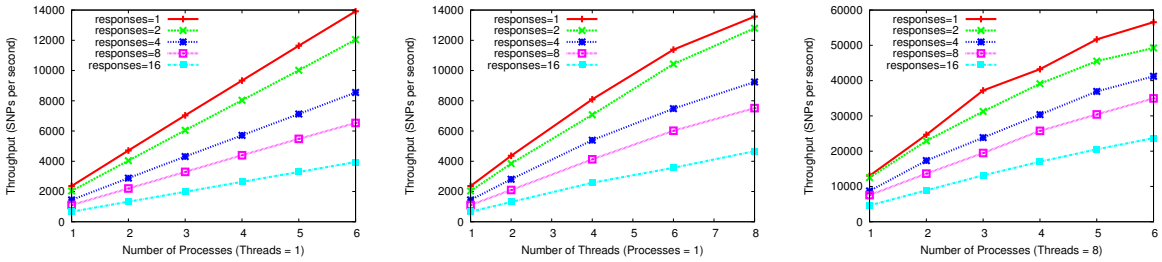
Fig. 6. Comparison of true positive rate between piecewise linear screening and linear screening under different sample sizes and MAFs of true association SNP pairs under additive scenario (a,b) and non-additive scenario (c,d).

3.4 Benefits of Using a Piecewise Linear Model for Screening

We tested the benefits of using a piecewise linear model instead of a simple linear model during the screening procedure. Throughout this section, we use PLS to indicate using a piecewise linear model for screening and LS to indicate using a simple linear model for screening. For this experiment, we simulated data with $P = 500$ (that generates candidates of 124750 SNP pairs), three SNPs having marginal effects with association strength of 1, and three SNP pairs having interaction effects with association strength of 3. Given the simulation data, for both PLS and LS, N candidates of SNP pairs were selected. We then evaluated true positive rate of PLS and LS under different minor allele frequencies (MAFs) of true association SNP pairs from 0.1 to 0.4 (fixing $N = 200$) and different sample sizes from 100 to 600 (fixing MAF1=MAF2= 0.1). Fig. 6 represents the average true positive rate of PLS and LS with error bars of 1/2 standard deviation when the underlying true interaction effect was additive (See Fig. 6(a,b)), and non-additive (See Fig. 6(c,d)). In general, our results show that PLS is very useful under various simulation settings. When true model was linear, true positive rates of PLS and LS were comparable in most of our settings as shown in Fig. 6(a,b), which was not expected because a simple linear model would be ideal given finite data under the additive scenario. It seems that the model complexity of PLS was small enough not to lose much power. When true model was non-linear, PLS showed clear benefits over LS. As seen in Fig. 6(c,d), true positive rate of PLS substantially increased as sample size and minor allele frequency increased but true positive rate of LS marginally improved. It seems that true positive rate of PLS significantly increased due to the fact that additional degrees of freedom allowed PLS to fit well into the data under the non-additive scenario.

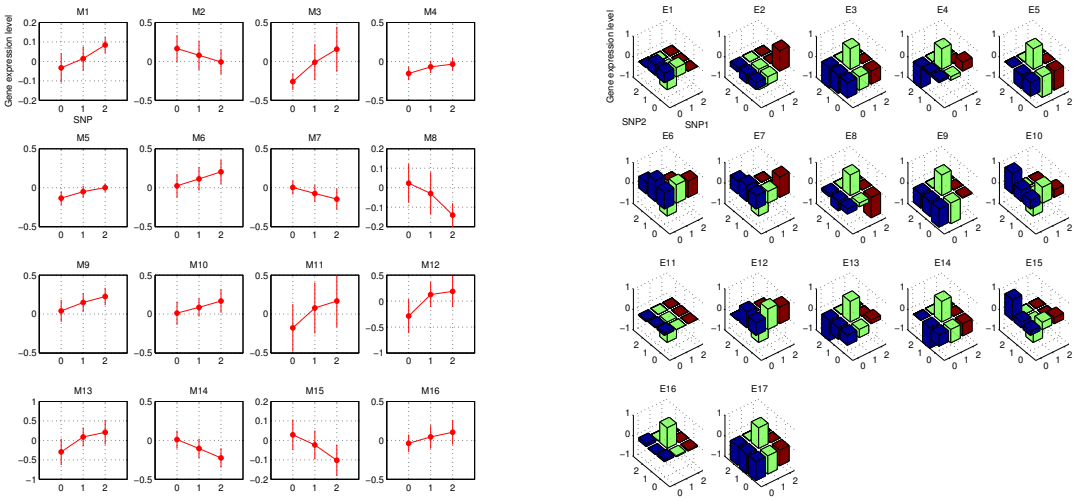
3.5 Scalability of Screening Implementation

We carried out scalability experiments for our screening implementation on oxygen, a six-node cluster of dual-socket, quad-core Intel Xeon E5410 machine with 32GB of RAM per-node running Linux Kernel 2.6.31-23 (total 48 cores). Fig. 7 shows the throughput (SNPs processed per second) for various scenarios when running on a simulated dataset that had 200 SNPs, 500 samples, and 20 true association SNPs. Each experiment was run for 50 iterations, where each iteration considered 200 marginal candidates and $\binom{200}{2}$ interaction candidates. To test the effect of simultaneously evaluating multiple responses (e.g., eQTL mapping on many gene traits), we ran experiments with the number of responses varying from 1 to 16. The left panel depicts the multi-node (cluster) performance of our implementation on oxygen; as can be seen, our implementation is able to process up to 14000 SNPs per second on 6 machines and shows near-linear speedup. Furthermore, our implementation handles an increasing number of responses (e.g., gene traits) gracefully; processing 16



(a) Multi-node (b) Multi-core (c) Hybrid

Fig. 7. Performance of the parallel implementation of our screening algorithm on `oxygen` cluster.



(a) SNPs with marginal effects (b) SNP pairs with interaction effects

Fig. 8. Gene expression levels according to the genotypes of (a) 16 SNPs, and (b) 17 SNP pairs found by SPHINX. In (a), x-axis represents genotypes and y-axis shows the average gene expression levels of individuals who possess the corresponding genotype with error bars of 1/2 standard deviation. In (b), x and y-axis represent genotypes of a SNP pair and z-axis shows the average gene expression levels.

responses in a multi-task fashion only results in a 3.5x slowdown when compared to processing just one response (4000 SNPs per second as opposed 14000 SNPs per second). The middle panel shows the near-linear speedup achieved when we use pure multi-threading on a single node of oxygen. The scalability is slightly less than the multi-node case because of memory bandwidth issues that result from BLAS-2 operations such as matrix-vector products. Finally, the right panel demonstrates our algorithm's capability to exploit both multi-core and cluster architectures together. In this experiment, we ran 8 threads per-node and increased the number of nodes from 1 to 6 to achieve a near-linear speedup. To conclude, our implementation is able to efficiently process large datasets while scaling near-linearly.

4 Association Analysis of Late-Onset Alzheimer's Disease Data

We applied SPHINX to late-onset Alzheimer's disease (AD) data from Harvard Brain Tissue Resource Center and Merck Research Laboratories [41], in an attempt to detect causal SNPs associated either marginally or epistatically to molecular traits of interest. This data concerns 206 AD cases with 555091 SNPs in total and expression levels of 37585 DNA probes including known and predicted genes, miRNAs and non-coding

Table 1. Significant trait-associated SNPs in Alzheimer's disease dataset [41] found by SPHINX. For each SNP, we represent GENE which is located within 50kb from the corresponding SNP. The stability score represents the proportion for which the SNP was selected in stability selection.

SNP	GENE	Affected Gene	Stability Score
rs1047631	*DTNBP1*	*DTNBP1*	0.705
rs536635	*C9orf72*	*SELL*	0.651
rs7483826	*WT1*	*WT1*	0.979
rs2699411	*LRPAP1*	*LRPAP1*	0.824
rs16844487	*LRPAP1*	*LRPAP1*	0.763
rs1323580	*PTPRD*	*HHEX*	0.631
rs4701834	*SEMA5A*	*SEMA5A*	0.631
rs7852952	*PTPRD*	*PTPRD*	0.724
rs2734986	*HLA-A*	*HLA-A*	0.628
rs1611710	*HLA-A*	*HLA-A*	0.617
rs2395175	*HLA-DRB1*	*HLA-DRB1*	0.692
rs602875	*HLA-DQB1*	*HLA-DQB1*	0.809
rs3135363	*HLA-DRB1*	*HLA-DQB1*	0.717
rs1619379	*HLA-A*	*HLA-A*	0.967
rs156697	*GSTO2*	*GSTO2*	0.943
rs7759273	*ABCB1*	*PARK2*	0.67

RNAs in three brain regions including cerebellum, visual cortex, and dorsolateral prefrontal cortex, profiled on a custom-made Agilent 44K microarray. Specifically, we are interested in the expression traits of all 718 genes in visual cortex related to neurological diseases according to GAD (genetic association database) [2], and we focused on the 18137 SNPs residing within 50kb from these genes, in an attempt to search for *cis*-acting causal SNPs or "restricted" *trans*-acting (i.e., acting on genes within the same functional group) SNPs related to neurological diseases. This results in a massive problem involving 18137 SNPs, $\sim$ 164 million SNP-pairs and 718 gene traits, that is, $\sim$ 118 billion candidates of associations between SNPs/SNP-pairs and traits. We employed a cluster with a total of 720 cores (see Methods for experimental details), which took 4.5 hours to perform screening and $<$ 1 hour for stability selection with PLM-based randomized group lasso. Using SPHINX, we found 16 SNPs and 17 SNP pairs significantly associated with the expression traits (see Tables 1 and 2 for the list of all SNPs and SNP pairs found by SPHINX). Note that most association studies on AD have focused on detecting SNPs with marginal effects, and SNP pairs associated with AD are largely unknown. The patterns of marginal and interaction effects are illustrated in Fig. 8.

4.1 Marginal Effects in Late-Onset Alzheimer's Disease Dataset

Among 16 SNPs identified with marginal effects, 13 SNPs were located nearby affected genes (12 SNPs are located within 50kb, and 1 SNP is located within 130kb from their associated genes), and 3 SNPs were associated with a gene trait in a different chromosome. As an example, here we investigate 6 SNPs (*rs1619379*, *rs2734986*, *rs1611710*, *rs2395175*, *rs3135363*, *rs602875*) associated with HLA (human leukocyte antigen) genes including *HLA-A*, *HLA-DRB1*, and *HLA-DQB1*, related to immune system. All the 6 SNPs were located nearby the affected HLA genes, which encode proteins for antigen presentation [4]. We observed that 5 SNPs out of the 6 SNPs had positive correlation with the expression levels of their associated genes, whereas one SNP (*rs1619379*) had negative correlation with the expression levels of its associated gene (*HLA-A*).

We found out that 5 SNPs (out of the 6 SNPs) had genome annotations in their locations. For associations between the three SNPs (*rs1619379*, *rs2734986*, *rs1611710*) and *HLA-A*, we observed that *rs1619379* and *rs1611710* coincide with H3K27Ac histone mark and transcription factor binding sites, respectively, and *rs2734986* aligns with spliced ESTs. It suggests that *rs1619379* and *rs1611710* may perturb regulatory elements of *HLA-A*, and *rs2734986* may be related to a mechanism for DNA transcription. In case of

Table 2. Significant trait-associated SNP pairs identified by SPHINX in Alzheimer's disease dataset [41]. For each SNP A(B), we represent GENE A(B), which is located within 50kb from the SNP. The stability score represents the proportion for which the pair was selected in stability selection.

SNP A	GENE A	SNP B	GENE B	Affected Gene	Stability Score
rs10501554	*DLG2*	rs7805834	*NOS3*	*NEFH*	0.684
rs4547324	*Intergenic*	rs7870939	*PTPRD*	*MEIS1*	0.602
rs1956993	*NUBPL*	rs6677129	*LOC199897*	*FARP1*	0.633
rs27744	*LTC4S*	rs13209308	*PARK2*	*CLCN2*	0.629
rs17150898	*MAGI2*	rs7798194	*CDK5*	*NINJ2*	0.605
rs2802247	*FLT1*	rs9533787	*DNAJC15*	*ADH1C*	0.629
rs10883782	*CYP17A1*	rs10786737	*CNNM2*	*SCN1B*	0.683
rs7139251	*ITPR2*	rs12915954	*IGF1R*	*IL6*	0.605
rs11207272	*PDE4D*	rs2274932	*ZBP1*	*ARSB*	0.635
rs2634507	*TOX*	rs11790283	*VLDLR*	*SFXN2*	0.622
rs17309944	*BDNF*	rs358523	*HTR1A*	*GRIK1*	0.665
rs10501554	*DLG2*	rs17318454	*RFX4*	*GNAS*	0.611
rs4900468	*CYP46A1*	rs10217447	*PTPRD*	*CAPN5*	0.64
rs17415066	*KCNJ10*	rs912666	*SUSD1*	*SEMA5A*	0.663
rs6578750	*CCKBR*	rs12340630	*TAL2*	*CTNNA3*	0.631
rs4272759	*PGR*	rs6081791	*PDYN*	*DAT*	0.71
rs2679822	*MYRIP*	rs4538793	*NXPH1*	*CPT7*	0.85

the association between *rs2395175* and *HLA-DRB1*, *rs2395175* was in an intron of a *HLA-DRB1* gene (chr6:32489683-32557613). Finally, for associations between the two SNPs (*rs3135363*, *rs602875*) and *HLA-DQB1*, we observed that *rs3135363* coincides with both transcription factor binding site and H3K27Ac histone mark, which hints that *rs3135363* may be related to regulatory mechanisms for *HLA-DQB1*. For *rs602875*, we did not find any specific genome annotations.

It should be noted that associations between *HLA* genes and late-onset AD [23, 28] have been found and these findings have been replicated in previous studies. It has been reported that there is association between *HLA-A* and late-onset of AD [15, 35], and Lehmann et al. replicated the association between *HLA-B7* and AD [22]. Furthermore, recently, HLA-DRB1 identified by SPHINX has been reported as a new susceptibility locus for AD [20]. Lambert et al. identified 11 new loci associated with AD that includes HLA-DRB1 from 17008 AD cases and 37154 controls. This dataset is independent from ours, which indicates that our findings can be reproducible. As we found associations between the 6 SNPs and *HLA* genes, and previous studies reported associations between *HLA* genes and AD, it would be interesting to further investigate whether these associations are related to regulatory mechanisms or transcription factor bindings.

4.2 Interaction Effects in Late-Onset Alzheimer's Disease Dataset

Among 17 SNP pairs identified with interaction effects, as an example, we investigate the biological underpinnings of one of our findings — the pair *rs4272759* (chr11:100899750) and *rs6081791* (chr20:1988298) that is jointly (but not marginally) associated with *DAT* (Dopamine Active Transporter, chr5:1392905-1445545) — to demonstrate the biological validity of our results. Specifically, the expression level of *DAT* is high only when both SNPs are heterozygous (i.e., both SNPs have only one minor allele). SNP *rs4272759* is located 605 base-pairs upstream of the start position of gene *PGR* (ProGesterone Receptor, chr11:100900355-101001255), whereas SNP *rs6081791* is 13407 base-pairs downstream of gene *PDYN* (ProDYNorphin, chr20:1959402-1974891). An extensive literature survey has yielded intriguing biological evidence to explain the association involving *DAT* with *PGR* and *PDYN*, suggesting that our finding is biologically plausible. It was reported that progesterone treatment could increase the dynorphin concentration and prodynorphin mRNA level (prodynorphin is the precursor protein of dynorphin) [11], suggesting that a disruption of the *PGR* function could alter the activity of *PDYN*, which supports our finding that the SNPs in *PGR* and *PDYN* are epistatic.

A direct association between *PDYN* and *DAT* has also been reported. For example, it has been reported that prodynorphin expression in the striatum is associated with D1 dopamine receptor stimulation [13]; furthermore, in the experiments with DAT knock-down mice, Cagniard et al. [7] found that the increased level of dopamine is associated with the level of dynorphin expression. Overall, evidence from the literature seems to support a hypothesis drawn from our association analysis of interacting genetic variations that, a pair of SNPs affecting *PGR* and *PDYN* are likely to lead to an epistatic effect on *DAT*, and it would be interesting to further examine the status of *DAT* in the case studied in [11], and the status of *PGR* in cases studied in [13] and [7] to directly confirm and characterize such an epistatic effect.

5 Conclusions

We developed a unified framework for detecting marginal and pairwise interaction effects on traits, built on state-of-the-art techniques including screening, randomization, and stability selection. Furthermore, to facilitate the detection of SNPs and SNP pairs associated with traits at a whole genome scale, we implemented an efficient and scalable screening program. We validate the efficacy of SPHINX via simulations and the analysis of late-onset Alzheimer's disease dataset. Note that detecting pairwise interaction effects on traits requires us to address computational and statistical challenges simultaneously, which stem from a large number of SNP pairs to be tested, correlations between SNPs/SNP pairs, and non-linear patterns of marginal and interaction effects; to our knowledge, SPHINX is the first attempt to address these challenges within a single framework. We further note that by redefining m^i_{jk} in Eq. (1), it is possible to investigate different choices of interaction encodings (e.g., data-driven encoding [16]). In this paper, we adopted the widely used genotype encoding for the pairwise interaction (i.e., multiplication of two SNPs). For future work, we plan to (1) incorporate diverse prior knowledge into our model such as trait networks using graph-guided fused lasso [19] or grouping information on both genotypes (e.g., LD structures) and phenotypic traits (e.g., pathways) using structured input-output lasso [21], (2) use kernel techniques for detecting multi-way interactions among SNPs, (3) detect interaction effects under case-control settings via logistic regression, and (4) combine linear mixed model with SPHINX to correct for population structures [37].

Acknowledgments This work was done under a support from NIH 1 R01 GM087694-01; NIH 1RC2HL101487-01 (ARRA); NSF IIS-0713379; and P30 DA035778A1.

References

1. F. R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.
2. K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang. The genetic association database. *Nature Genetics*, 36(5):431–432, 2004.
3. J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.
4. W. F. Bodmer and J. G. Bodmer. Evolution and function of the hla system. *British Medical Bulletin*, 34(3):309–316, 1978.
5. O. Bretscher. *Linear algebra with applications*. Prentice Hall Eaglewood Cliffs, NJ, 1997.
6. P. Bühlmann, P. Rütimann, S. van de Geer, and C. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 2013.
7. B. Cagniard, P. D. Balsam, D. Brunner, and X. Zhuang. Mice with chronically elevated dopamine exhibit enhanced motivation, but not learning, for a food reward. *Neuropsychopharmacology*, 31(7):1362–1370, 2005.
8. D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genetics*, 2(9):e157, 2006.
9. J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.
10. J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.
11. C. D. Foradori, R. L. Goodman, V. L. Adams, M. Valent, and M. N. Lehman. Progesterone increases dynorphin a concentrations in cerebrospinal fluid and preprodynorphin messenger ribonucleic acid levels in a subset of dynorphin neurons in the sheep. *Endocrinology*, 146(4):1835–1842, 2005.

12. J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
13. C. R. Gerfen, T. M. Engber, L. C. Mahan, Z. Susel, T. N. Chase, FJ Monsma, and D. R. Sibley. D1 and d2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science*, 250(4986):1429–1432, 1990.
14. G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.
15. F. R. Guerini, C. Tinelli, E. Calabrese, C. Agliardi, M. Zanzottera, A. De Silvestri, M. Franceschi, L. M. Grimaldi, R. Nemni, and M. Clerici. HLA-A*01 is associated with late onset of Alzheimer's disease in italian patients. *International Journal of Immunopathology and Pharmacology*, 22:991–999, 2009.
16. D. He, Z. Wang, and L. Parida. Data-driven encoding for quantitative genetic trait prediction. *BMC bioinformatics*, 16(Suppl 1):S10, 2015.
17. G. E. Hoffman, B. A. Logsdon, and J. G. Mezey. PUMA: A unified framework for penalized multiple regression analysis of gwas data. *PLoS Computational Biology*, 9(6):e1003101, 2013.
18. P. Kambadur, A. Gupta, A. Ghoting, H. Avron, and A. Lumsdaine. PFunc: modern task parallelism for modern high performance computing. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, page 43. ACM, 2009.
19. S. Kim and E. P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5(8):e1000587, 2009.
20. J. Lambert et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 2013.
21. S. Lee and E. P. Xing. Leveraging input and output structures for joint mapping of epistatic and marginal eqtls. *Bioinformatics*, 28(12):i137–i146, 2012.
22. D. J. Lehmann, M. C. Barnardo, S. Fuggle, I. Quiroga, A. Sutherland, D. R. Warden, L. Barnetson, R. Horton, S. Beck, and A. D. Smith. Replication of the association of HLA-B7 with Alzheimer's disease: a role for homozygosity? *Journal of Neuroinflammation*, 3(1):33, 2006.
23. D. J. Lehmann et al. HLA class I, II & III genes in confirmed late-onset Alzheimer's disease. *Neurobiology of Aging*, 22(1):71–77, 2001.
24. C. Li and M. Li. GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, 24(1):140–142, 2008.
25. J. Li, M. Zhu, A. B. Manning-Bog, D. A. Di Monte, and A. L. Fink. Dopamine and l-dopa disaggregate amyloid fibrils: implications for parkinson's and Alzheimer's disease. *The FASEB journal*, 18(9):962–964, 2004.
26. J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.
27. J. Liu and J. Ye. Moreau-yosida regularization for grouped tree structure learning. *Advances in Neural Information Processing Systems*, 187:195–207, 2010.
28. E. Maggioli, C. Boiocchi, M. Zorzetto, E. Sinforiani, C. Cereda, G. Ricevuti, and M. Cuccia. The human leukocyte antigen class III haplotype approach: new insight in Alzheimer's disease inflammation hypothesis. *Current Alzheimer Research*, 10(10):1047–1056, 2013.
29. N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
30. N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.
31. Message Passing Interface Forum. MPI, June 1995. `http://www.mpi-forum.org/`.
32. Message Passing Interface Forum. MPI-2, July 1997. `http://www.mpi-forum.org/`.
33. J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.
34. M. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2008.
35. H. Payami et al. Evidence for association of HLA-A2 allele with onset age of Alzheimer's disease. *Neurology*, 49(2):512–518, 1997.
36. S. Purcell et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.
37. B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, 2013.
38. X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N.L.S. Tang, and W. Yu. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, 87(3):325, 2010.
39. L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.
40. M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.

41. B. Zhang et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimers disease. *Cell*, 153(3):707–720, 2013.
42. X. Zhang, F. Zou, and W. Wang. FastANOVA: an efficient algorithm for genome-wide association study. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 821–829. ACM, 2008.