SPADIS: An Algorithm for Selecting Predictive and Diverse SNPs in GWAS

Serhan Yilmaz*, Oznur Tastan† and A. Ercument Cicek*,‡
*Computer Engineering Department, Bilkent University, Ankara, 06800, Turkey
†Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, 34956, Turkey
‡Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
Correspondence: †otastan@sabanciuniv.edu, ‡cicek@cs.bilkent.edu.tr

Abstract*—Phenotypic heritability of complex traits and diseases is seldom explained by individual genetic variants identified in genome-wide association studies (GWAS). Many methods have been developed to select a subset of variant loci, which are associated with or predictive of the phenotype. Selecting SNPs that are *close* on a biological network such as SNP-SNP networks have been proven successful in finding biologically interpretable and predictive SNPs. However, we argue that the closeness constraint favors selecting redundant features that affect similar biological processes and therefore does not necessarily yield better predictive performance. An approach, which awards diversity of the selected SNPs and affected functional processes, would boost the predictive power without compromising biological interpretability. In this paper, we propose a novel method called SPADIS that selects a set of loci such that diverse regions in the underlying SNP-SNP network are *covered*. Instead of enforcing selections based on closeness in the network, SPADIS favors the selection of remotely located SNPs in order to account for the complementary additive effects of SNPs that are associated with the phenotype. This is achieved by maximizing a submodular set function with a greedy algorithm that ensures a constant factor $(1 - 1/e)$ approximation to the optimal solution. We compare SPADIS to the state-of-the-art method SConES, on a dataset of *Arabidopsis Thaliana* genotype and continuous flowering time phenotypes. SPADIS has better regression performance in 12 out of 17 phenotypes on average, it identifies more candidate genes and runs faster. We also investigate the use of Hi-C data to construct SNP-SNP network in the context of SNP selection problem for the first time, which yields slight improvements in regression performance. SPADIS is available at *http://ciceklab.cs.bilkent.edu.tr/spadis

***Index Terms*—SNP Selection, Phenotype Prediction, Submodular Function, SNP-SNP Networks, Hi-C, GWAS.**

I. Introduction

Genome-Wide Association Studies (GWAS) have led to a wide range of discoveries over the last decade where individual variations in DNA sequences, usually single nucleotide polymorphisms (SNPs), have been associated with phenotypic differences (Visscher *et al.*, 2017). However, individual variants often fail to explain the heritability of complex traits and diseases (Manolio *et al.*, 2009; Goldstein *et al.*, 2009) as a large number of variants contribute to these phenotypes and each variant has a small overall effect (Kraft and Hunter, 2009; Christensen and Murray, 2007). Thus, research efforts have focused on evaluating and associating multiple loci with a given phenotype (Moore *et al.*, 2010; Cordell, 2009). Indeed, detecting genetic interactions (epistasis) among pairs of loci has proven to be a powerful approach (Phillips, 2008; Cordell, 2009; Wang *et al.*, 2010a; Wei *et al.*, 2014).

Detecting higher-order combinations of genetic variations is computationally challenging. For this reason, exhaustive search approaches have been limited to small SNP counts (up to few hundreds) (Nelson *et al.*, 2001; Ritchie *et al.*, 2001; Lou *et al.*, 2007; Lehár *et al.*, 2008; Hua *et al.*, 2010; Fang *et al.*, 2012) and greedy search algorithms have been limited to searching for small groups of SNPs –mostly around 3 (Storey *et al.*, 2005; Evans *et al.*, 2006; Yosef *et al.*, 2007; Varadan and Anastassiou, 2006; Varadan *et al.*, 2006; Zhang and Liu, 2007; Herold *et al.*, 2009; Tang *et al.*, 2009; Jiang *et al.*, 2009; Zhang *et al.*, 2010; Wang *et al.*, 2010b; Wan *et al.*, 2010; Guo *et al.*, 2014; Ding *et al.*, 2015; Ayati and Koyutürk, 2016; Tuo *et al.*, 2017). Multivariate regression-based approaches have been used (Shi *et al.*, 2008; Wu *et al.*, 2009; Cho *et al.*, 2010; Wang *et al.*, 2011a; Rakitsch *et al.*, 2012). However, (i) their predictive power is limited, (ii) incorporation of biological information in the models is not straightforward, and finally (iii) selected SNP set is often not biologically interpretable (Azencott *et al.*, 2013).

Assessing the significance of loci by grouping them based on functionally related genes, such as pathways, reduces the search space for testing associations and leads to discovery of more interpretable sets (Wang *et al.*, 2011b; de Leeuw *et al.*, 2015). Unfortunately, using gene sets and exonic regions for association restricts the search space to coding and nearby-coding regions. However, most of the genetic variation fall into non-coding genome (Hindorff *et al.*, 2009) and our knowledge of pathways are incomplete.

An alternative strategy to avoid literature bias is to select features on the SNP-SNP networks by applying regression based methods with sparsity and connectivity constraints (Jacob *et al.*, 2009; Huang *et al.*, 2011). These regularized methods jointly consider all predictors in the model as opposed to univariate test of associations. Nevertheless, using a SNP-SNP interaction network on GWAS yields intractable number of interactions. Azencott *et al.* presents an efficient method called SConES which uses a minimum graph cut-based approach to select predictive SNPs over a network of hundreds of thousands of SNPs (Azencott *et al.*, 2013;

Sugiyama *et al.*, 2014). In their network, edges denote either (i) spatial proximity on the genomic sequence or (ii) functional proximity as encoded with PPI closeness of loci. The method selects a connected set of SNPs that are individually related to the phenotype under additive effect model and has been shown to perform better than graph-regularized regression-based methods.

We argue that enforcing the selected features to be in close proximity encourages the algorithm to pick features that are in linkage disequilibrium or that have similar functional consequences. One extreme choice of this approach would be to choose all SNPs that fall into the same gene if they are individually found to be significantly associated with the phenotype. When there is an upper limit on the number of SNPs to be selected, this will lead to selecting functionally redundant SNPs and will fail to recall variants that hit diverse processes. Genetic complementation on the other hand, is a well-known phenomenon where multiple loci in multiple genes need to be mutated in order to observe the phenotype (Fincham, 1968). While there are numerous examples of long-range (trans) genetic interactions for transcription control (Miele and Dekker, 2008) and long-range epistasis is evident in complex genetic diseases such as type 2 diabetes (Wiltshire *et al.*, 2006), such complementary effects may not be treated with this approach. For disorders with complex phenotypes like Autism Spectrum Disorder (ASD), this would be even more problematic since multiple functionalities (thus gene modules in the network) are required to be disrupted for an ASD diagnosis, whereas damage in only one leads to a more restricted phenotype (Geschwind, 2008).

We hypothesize that diversifying the SNPs in terms of location would result in *covering* complementary modules in the underlying network that cause the phenotype. Based on this rationale, here, we present SPADIS, a novel SNP selection algorithm over a SNP-SNP interaction network that favors (i) loci with high univariate associations to the phenotype and (ii) that are diverse in the sense that they are far apart on a loci interaction network. In order to incorporate these principles, we design a submodular set scoring function. To maximize this set function, we use a greedy algorithm that is guaranteed to return a solution which is a constant factor $(1-1/e)$ approximate to the optimal solution. We compare our algorithm to the state-of-the-art, on a GWAS of *Arabidopsis Thaliana (AT)* with 17 continuous phenotypes related to flowering time (Atwell *et al.*, 2010). Replicating the experimental setting of Azencott *et al.* 2013 on the same dataset, we show that SPADIS has better regression performances on average in 12 out of 17 phenotypes with better runtime performance. We show that our method doubles the number of candidate genes identified and hits 23% more Gene Ontology (GO) terms proving that selection of SPADIS is more diverse.

Finally, we also employ Hi-C data in the context of SNP selection problem for the first time. Emerging evidence suggests that the spatial organization of the genome plays an important role in gene regulation Bickmore (2013) and contacts in 3D have been shown to affect the phenotype (Martin *et al.*, 2015; Jäger *et al.*, 2015). Hi-C technology can detect the 3D conformation genome-wide and yield contact maps which show loci that reside nearby in 3D (van Berkum *et al.*, 2010). We construct a SNP-SNP network based on genomic contacts in 3D as captured by HI-C and use this network to guide SNP selection. Our results show that use of Hi-C based network provides a slight overall increase in the prediction performance for all methods tested.

II. Methods

The problem is formalized as a feature selection problem over a network of SNPs. Let n be the number of SNPs. The problem is to find a SNP subset S with cardinality at most $k \ll n$ that explains the phenotype, given a background biological network $G(V, E)$. In G, vertices represent SNPs and edges link loci which are related based on spatial or functional proximity as explained in sections below. G can be a directed or an undirected graph.

We utilize a two-step approach. In the first step, we assess the relation of each SNP to the phenotype individually using the Sequence Kernel Association Test (SKAT) (Wu *et al.*, 2011). Let $\mathbf{c}$ be a scoring vector such that $c_i \in \mathbb{R}_{\geq 0}$ indicates the degree of i-th SNP's association with the phenotype. Our goal is to maximize the total score of SNP set while ensuring the selected set consists of SNPs that are diversely located on the network. Under the additive effect model, we define the set function shown in Equation (1) to encode this intuition.

$$F(S) = \sum_{i \in S} \left(c_i + \beta \left(1 - \sum_{j \in S} \frac{K(i,j)}{2k} \right) \right) \tag{1}$$

$$K(i,j) = \begin{cases} 1 - d(i,j)/D & d(i,j) \leq D, \quad i \neq j \\ 0 & otherwise \end{cases}$$

Here, $D \in \mathbb{R}_{>0}$ is a distance limit parameter and $d(i,j)$ is the shortest path between vertices $i, j \in V$. $K(i,j)$ is a function that penalizes vertices that are in *close* proximity. That is, the vertices i and j are considered *close* if and only if $d(i,j) \leq D$. The second parameter, $\beta \in \mathbb{R}_{\geq 0}$ controls the penalty to be applied when two close vertices are jointly included in S. Note that, $K(i,j) \in [0,1], \forall i,j \in V$ and c_i is non-negative.

Our aim is to find a subset of SNPs S^* of size k that maximizes F:

$$S^* = \underset{S \subseteq V, |S| \leq k}{\operatorname{argmax}} \; F(S) \tag{2}$$

Subset selection problem with cardinality constraint is NP-hard. Thus, exhaustive search is infeasible when k or V is not small. Hence, heuristic algorithms are required. We make use of the fact that the function defined in Equation (1) is submodular. Although submodular optimization itself is NP-hard as well (Krause and Guestrin, 2012), the greedy algorithm given in **Algorithm 1**, proposed by Nemhauser *et al.* 1978, guarantees a $\left(1 - \frac{1}{e}\right)$-factor approximation to the optimal solution under cardinality constraint for monotonically non-decreasing and non-negative submodular functions. The

Algorithm 1 Greedy Algorithm

Input: Set function F, ground set V, cardinality constraint $k \leq |V|$.
Output: Set $S \subset V$ such that $|S| = k$.
1: $S \leftarrow \emptyset$
2: **while** $|S| < k$ **do**
3: $\quad S \leftarrow S \cup \underset{x \in V \setminus S}{\operatorname{argmax}} F(S \cup x)$
4: **end while**

greedy algorithm starts with an empty set and at each step, adds an element that maximizes the set function. Note that, this is equivalent to adding elements with the largest marginal gain.

For each of the k iterations in the algorithm, where k is the size of S^*, a single source shortest path problem needs to be solved. Hence, the worst-case time complexity of the algorithm is $O(k(V+E))$ assuming that all edge weights are positive. For undirected graphs, $K(i,j) = K(j,i)$ and computations can be reduced by half even though the time complexity remains the same.

A submodular function is a set function for which the gain in the value of the function after adding a single item decreases as the set size grows (diminishing returns). Next, we prove that F is a submodular set function.

Definition 1. V is the ground set, $F : 2^V \rightarrow \mathbb{R}$ and $S \subseteq V$. The marginal gain of adding one element to the set S is : $G(S,x) = F(S \cup \{x\}) - F(S)$ where $x \in V \setminus S$.

By plugging the definition of F in Equation (1), we can rewrite G.

$$\begin{aligned} G(S,x) = &\left(\sum_{i \in S \cup \{x\}} c_i \right. \\ &\left. + \beta \sum_{i \in S \cup \{x\}} \left(1 - \sum_{j \in S \cup \{x\}} \left(\frac{K(i,j)}{2k} \right) \right) \right) \\ &- \left(\sum_{i \in S} c_i + \beta \sum_{i \in S} \left(1 - \sum_{j \in S} \left(\frac{K(i,j)}{2k} \right) \right) \right) \\ &= c_x + \beta - \frac{\beta}{2k} \sum_{i \in S} (K(i,x) + K(x,i)) \end{aligned} \tag{3}$$

Definition 2. A function F that is defined on sets, is *submodular* if and only if $G(A,x) \geq G(B,x)$ or equivalently $F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B)$ for all sets A, B where $A \subset B \subset V$ and $x \in V \setminus B$.

Lemma 1. $F(S)$ given in Equation (1) is *submodular*.

Proof. F is *submodular* if and only if the following is true:

$$G(A,x) - G(B,x) \geq 0 \tag{4}$$

Let $H(A,B,x)$ be,

$$\begin{aligned} H(A,B,x) &= G(A,x) - G(B,x) \\ &= \left(c_x + \beta - \frac{\beta}{2k} \left(\sum_{i \in A} (K(i,x) + K(x,i)) \right) \right) \\ &\quad - \left(c_x + \beta - \frac{\beta}{2k} \left(\sum_{i \in B} (K(i,x) + K(x,i)) \right) \right) \\ &= \frac{\beta}{2k} \left(\sum_{i \in B \setminus A} (K(i,x) + K(x,i)) \right) \end{aligned} \tag{5}$$

Since $K(i,j) \geq 0 \ \forall i,j \in V$, $H(A,B,x) \geq 0$. Hence, F is *submodular*. □ □

To be able to use the greedy algorithm, F must be a monotonically non-decreasing and non-negative function. Below, we prove that F satisfies these properties.

Definition 3. $F(S)$ is *monotonically non-decreasing* function for sets if and only if the corresponding gain function is always non-negative i.e. $G(S,x) \geq 0$ for all sets $S \subset V$ and $x \in V$.

Lemma 2. $F(S)$ given in Equation (1) is *monotonically non-decreasing* for sets for which $|S| \leq k$.

Proof. Since $K(i,j) \leq 1 \ \forall ij$, $G(S,x)$ is bounded such that;

$$\begin{aligned} G(S,x) &\geq c_x + \beta - \frac{\beta}{2k} \sum_{i \in S} (1+1) \\ &\geq c_x + \beta - \frac{\beta}{2k} 2|S| \\ &\geq c_x + \beta(1 - |S|/k) \\ &\geq (1 - |S|/k) \\ &\geq 0 \end{aligned} \tag{6}$$

Since $|S| \leq k$, $F(S)$ is *monotonically non-decreasing*. □ □

Lemma 3. $F(S)$ given in Equation (1) is non-negative for sets $|S| \leq k$.

Proof. For any set $S = \{v_1, v_2, ..., v_n\}$ with cardinality n, let S^i denote the subset of S that contains elements up to the i-th element, i.e. $S^i = \{v_1, v_2, ..., v_i\}$. $S^i = \emptyset$ for $i = 0$. $F(S)$ can be decomposed as the summation of marginal gain functions:

$$F(S) = F(\emptyset) + \sum_{i=1}^{n} G(S^{i-1}, v_i) \tag{7}$$

$F(\emptyset) = 0$ by the definition of $F(S)$. **Lemma 2** states that $G(S,x) \geq 0$ for all sets $S \subset V$ and $x \in V \setminus S$ when $|S| \leq k$. Hence, $F(S) \geq 0$ for all sets $S \subset V$ where $|S| \leq k$. □ □

III. Results

A. Dataset

We use *AT* genotype and phenotype data from Atwell *et al.* 2010. The dataset includes 17 phenotypes related to flowering times (up to $m = 180$ samples and $n = 214,051$ SNPs). Gene-gene interaction network is constructed based on TAIR protein-protein interaction data[1]. SNPs with a minor allele frequency (MAF) $< 10\%$ are disregarded ($n = 173,219$ SNPs remained) and population stratification is corrected using the principal components of the genotype data (Price *et al.*, 2006). Genes pertaining to each phenotype is retrieved from Segura *et al.* (2012) and used for validating the models. Gene Ontology (GO) annotations are obtained from TAIR (Berardini *et al.*, 2004). We obtain the Hi-C data for *AT* from Wang *et al.* 2015 and process the intra-chromosomal contact matrices using the Fit-Hi-C method (Ay *et al.*, 2014). An edge is added on top of the GS network for loci pairs that are significantly close in 3D (FDR adjusted p-value ≤ 0.05).

B. Networks

We construct four undirected SNP-SNP networks. To be able to compare the performances of SPADIS and SConES in a controlled setting, we use three networks defined in Azencott *et al.* 2013: The *GS network* links loci that are adjacent on the DNA. The *GM (gene membership) network* additionally links two loci if both loci fall into the same gene or they are both close to the same gene below a threshold of 20,000 bp. The *GI (gene interaction) network* also links any two loci if their nearby genes are interacting in the protein interaction network. Note that, $GS \subset GM \subset GI$. To investigate the usefulness of the 3D conformation of the genome in this setting, we introduce a new network, *GS-HICN* which connects loci that are close in 3D in addition to 2D (GS). All networks contain 173,219 vertices. The number of edges are as follows: *GS*: 346,428, *GM*: 23,322,332, *GI*: 36,269,032, *GS-HICN*: 5,839,214.

C. Methods compared

We compare SPADIS with the following methods using the networks described in Section III-B:

SConES: It is a network-constrained SNP selection method that provides a max-flow based solution (Azencott *et al.*, 2013).

Univariate: We run a univariate linear regression and select SNPs that are found to be significantly associated with the phenotype (FDR-adjusted p-value ≤ 0.05) (Yekutieli and Benjamini, 1999). If more than k SNPs are found to be associated, the most significant k SNPs are picked.

Lasso: The Lasso regression (Tibshirani, 1996) that minimizes the prediction error with an the $\ell 1$-regularizer of the coefficient vectors. We use the SLEP implementation (Liu *et al.*, 2009).

GraphLasso and GroupLasso: We also compare our method to GraphLasso and GroupLasso (Jacob *et al.*, 2009) through simulations, using the implementation in the SLEP package. Due to the prohibitive runtimes of these algorithms, we can not perform a comparison on *AT* dataset (see Section III-E). For GraphLasso, SNP pairs connected with an edge constitute a separate group, i.e. one such group is constructed for every edge in the network. For GroupLasso, the groups are defined as follows. For *GS*: every consecutive SNP pair on the genome constitute a single group. This is equivalent to setting a group for an edge. For *GM*: the SNPs *near* (< 20 kbp) of a gene are considered as a group, and a separate group is constructed for every gene. For *GI*: the SNPs that are *near* interacting genes in the PPI network are combined and formed a single group. The SNPs that are *near* genes but do not participate in the interaction network are assigned to groups based on their gene membership as in GM. For *GS-HICN*: SNP pairs connected with an edge is considered as a separate group similar to the groups in GraphLasso.

[1] ftp://ftp.arabidopsis.org/home/tair/Proteins/

D. Experimental Setup Details

1) Experimental Settings:

Experimental Setting 1 (ES1). The first one is SConES' setting explained in (Azencott *et al.*, 2013). In a 10-fold cross validation setting, a parameter search is conducted for each fold separately. The parameters that maximize the desired objective (e.g., stability, regression performance) are selected. With the best parameter set, the SNPs are selected with the corresponding method (please see Section III-D2 for details on parameter search). Then, for evaluation, ridge regression is performed on the complete dataset using a 10-fold cross validation setting using this SNP set. Although this strategy is adopted due to the limited dataset size in Azencott *et al.* 2013, it also implicates that the test data affects the feature selection step and might lead to memorization.

Experimental Setting 2 (ES2). We use nested cross-validation, the outer 10-fold cross-validation splits the data into training and test sets, and the inner loop is used to select the parameters via 10-fold cross-validation on the training set. In the inner loop at each fold, the scheme described in ES1 is run on the training data. Note that, parameter selection is repeated 10 times in ES2 and the test data is never seen by the algorithms.

To check if we could replicate the values reported in the paper, we use ES1. Under this setting, we show that we can successfully reproduce the reported precision of the selected SNPs and R^2 values of SConES in Supplementary Figures 2 and 3, respectively. In all other experiments, we use ES2.

2) Parameter Selection: A fair comparison among such a diverse range of methods is challenging. SPADIS operates with a cardinality constraint, whereas other methods have parameters that affect the number of selected SNPs. To account for these differences, we consider two parameter selection settings.

Parameter Selection Setting 1 (PSS1). We measure the predictive power of methods by constraining them to select a fixed number of SNPs (k). We achieve this by applying binary search over a range of sparsity parameter values that yields close to k chosen SNPs.

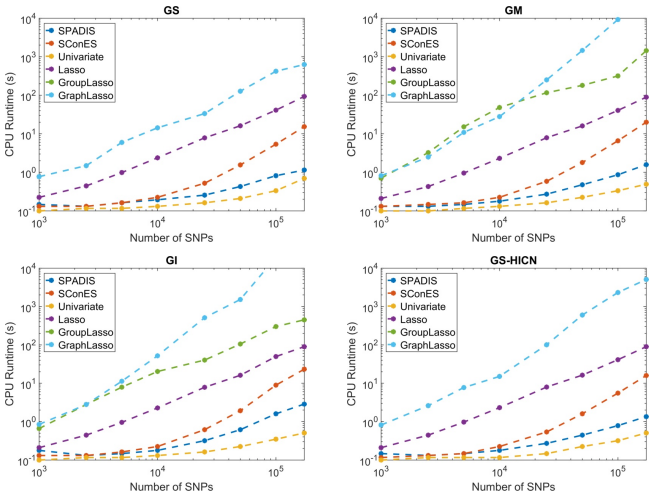


Fig. 1: CPU time measurements of SPADIS, SConES, Univariate, Lasso, GroupLasso and GraphLasso from 1.000 to 173.219 SNPs on four different networks. GS, GM, GI, GS-HICN respectively from left to right. Note that, runtimes of GroupLasso and GraphLasso are the same for GS and GS-HICN networks by construction.

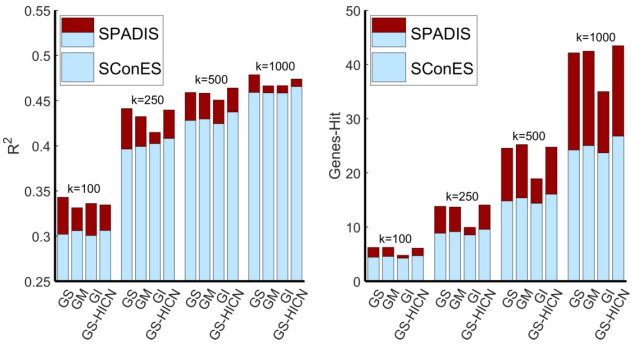


Fig. 2: Figure shows the improvement of SPADIS over SConES in terms of number of distinct candidate genes-hit (left) and Pearson's squared correlation coefficient (right) for different number of SNPs selected, k. Values shown are averages over 17 phenotypes. Blue bar indicates the maximum of SConES(S) and SConES(R) for the corresponding network and k value. The red bar indicates the amount of improvement of SPADIS over SConES.

Parameter Selection Setting 2 (PSS2). We allow the methods to select SNP sets of different sizes as long as the set size is smaller than an upper bound. This upper bound is set as the 1% of the total number of SNPs which is 1733.

In each setting, we select parameters either based on stability denoted with (S) (via consistency index) or regression performance denoted with (R). For consistency index based parameter selection (S), the common set of SNPs consistently selected across all training folds are chosen. In regression based parameter selection (R), SNP set is selected by a single run with the best parameter set on the training data. For SPADIS, we use only the regression performance (R) as SPADIS performs better with this strategy, for other methods we experiment with both of them. More details on parameter selection for each method are in Supplementary Text 3.1.

E. Time Performance

We report the CPU runtime of all methods, across a range of number of SNPs (from 1.000 to 173.219) and four networks. The measurements are taken on a single dedicated core of Intel I7-6700HQ processor. The runtime tests are conducted for one cross-validation fold with preset parameters on a single phenotype FT Field, which has the most number of samples available ($m = 180$).

We consider a method to time-out if it takes more than 10^3 seconds for a single run because the runtime of the complete test (10 folds with parameter selection) would take more than 1 CPU week (10^3 seconds x 10 evaluation folds x 10 training folds x at least 7 parameters).

Results show that SPADIS is more efficient than all other methods except the Univariate method (Fig. 1). GroupLasso and GraphLasso do not scale to SNP selection problem in GWAS. Hence, they are not included in the performance experiments described in Section III-G.

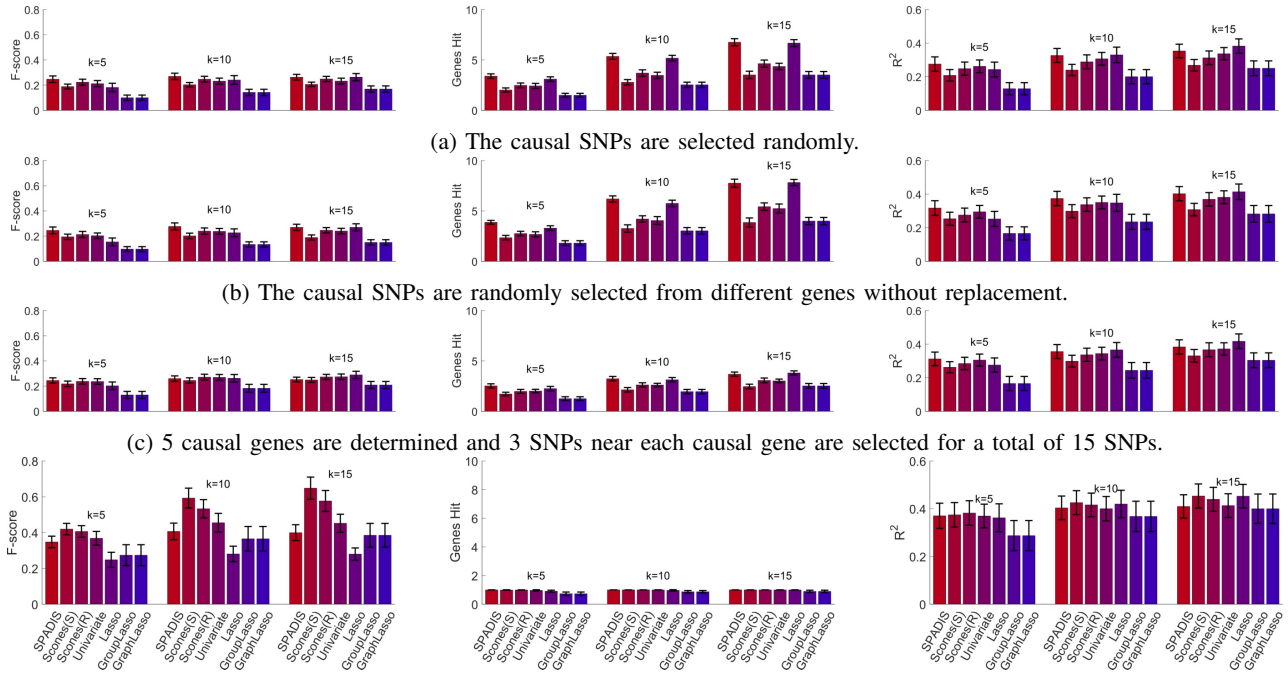
F. Simulated Experiments

To assess the performance of the methods in a controlled setting, we conduct simulated experiments. We randomly choose 200 samples out of 1307 samples in *AT* data. We select 500 random SNPs with MAF ¿ 10% as follows.

Then, we select 20 random SNPs near ($<$ 20 kbp) each gene. In each experiment, we designate 15 SNPs to be causal. We generate phenotypes using the regression model: $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, where $\mathbf{y} \in \mathbb{R}^{m\times 1}$ is the phenotype vector, $\mathbf{X} \in \mathbb{R}^{m\times n}$ is the genotype matrix, $\mathbf{w} \in \mathbb{R}^{n\times 1}$ is thrweight vector for each SNP, and ϵ is the error term. Both $\mathbf{w}$ and ϵ are normally distributed. We sample the weights of the non-causal SNPs from a normal distribution with zero mean and 0.1 standard deviation.

We compare the methods under four different simulation settings: (a) the causal SNPs are randomly selected, (b) the causal SNPs are selected randomly such that they are near different genes, (c) 5 causal genes are determined and 3 SNPs near each causal gene are selected for a total of 15 SNPs, and (d) the causal SNPs are selected near a single random gene.

For each method, we adopt the parameter selection strategy PSS1 and experimental setting ES2 and select the parameters such that the number of selected SNPs is k. We test with $k = 5$, 10 and 15. For evaluation, we consider three metrics : (1) F-score based on the number of causal SNPs that are selected, (2) Number of causal genes hit (a gene is hit if a SNP near that gene is selected), (3) Pearson's squared correlation coefficient. We perform 10-fold cross-validation 50 times and report averages over all folds. The 95% confidence interval for the means of the specified statistics are calculated assuming a t-distribution on the error.



(a) The causal SNPs are selected randomly.

(b) The causal SNPs are randomly selected from different genes without replacement.

(c) 5 causal genes are determined and 3 SNPs near each causal gene are selected for a total of 15 SNPs.

(d) The causal SNPs are selected such that they are near the same gene.

Fig. 3: The simulation results of SPADIS, SConES(R), SConES(S), Lasso, Univariate, GroupLasso and GraphLasso on GS network for $k = 5$, $k = 10$ and $k = 15$. (a) Causal SNPs are picked randomly. (b) All causal SNPs are from different genes. (c) Causal SNPs are from 5 different genes (d) All causal SNPs are from the same gene. (Left) F-score calculated for number of causal SNPs hit, (Middle) Number of causal genes hit, (Right) Pearson's squared correlation coefficient. Black bars indicate the 95% confidence intervals.

In all experimental settings except the fourth on the GS network, SPADIS outperforms other methods when k is less than the number of causal SNPs (Fig. 3). Otherwise, SPADIS performs comparably with Lasso and outperforms other methods. In the fourth setting, SPADIS underperforms compared to others in term of F-score. However, its regression performance is still comparable. This is the setting where methods with graph connectivity assumption should perform well (casual SNPs are close). However, we argue that all associated SNPs are rarely that close for complex traits.

G. Phenotype Prediction Performance

First, we compare the regression performances of SConES(S), SConES(R) and SPADIS using the Pearson's squared correlation coefficient R^2 under ES2 and PSS1 settings. Here, we report results wherein all approaches select 500 SNPs (Fig. 4). The results for k = 100, 250, and 1000 are provided in Supplementary Figures 4-6, respectively.

Out of 68 tests that is performed for $k = 500$ over 17 phenotypes using 4 different networks as input, SPADIS outperforms SConES(S) in 46 tests and SConES(R) in 47 tests. The improvement in R^2 is up to 0.15 in a single phenotype and 0.03 on average. Overall, this corresponds to an improvement in 12 out of 17 phenotypes, when they are averaged over all networks. We test whether the differences in R^2 are statistically significant (FDR adjusted p-value ≤ 0.05) using the method described in Hittner *et al.*. The multiple hypothesis correction is conducted as in Yekutieli and Benjamini. 3 results of SPADIS are found to be significantly better than SConES, whereas none of the results of SConES is found to be significantly better than SPADIS. The same type of comparison (k=500) using ES1, can be found in Supplementary Figure 1.

The improvement of SPADIS over SConES in regression performance for a varying number of selected SNPs ($k = 100, 250, 500, 1000$), is summarized in Fig. 2. We observe that performance of both methods increase as the set size grows. Therefore, for a fair comparison, we believe that it is important to compare the methods when they select the same number of SNPs. For all k values tested, SPADIS provides a consistent improvement in regression performance over SConES on average. In addition, the improvement of SPADIS is particularly prevalent when the number of selected SNPs is smaller. This is because, when the SNP set size is smaller, the additional information leveraged by the methods becomes more evident because there is more room for improvement in the regression performance. We argue that this property of SPADIS may provide an advantage in complex phenotypes with large number of causal SNPs.

A more natural setting for SConES and other methods is to let each method decide the number of SNPs based on their parameter search. We perform a second set of experiments

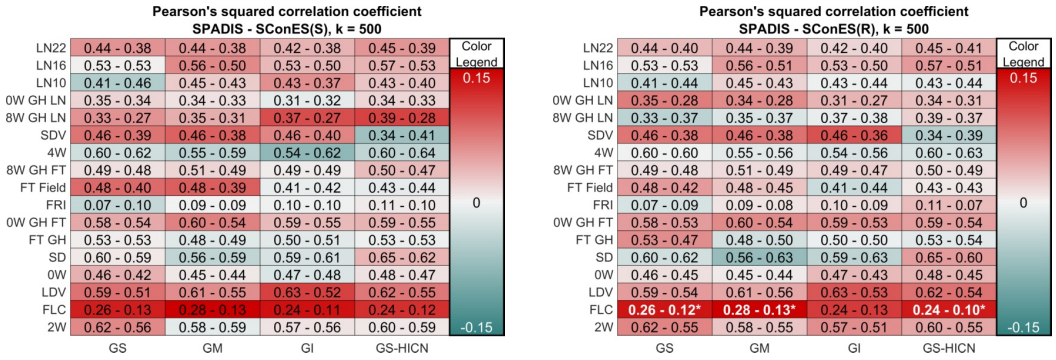


Fig. 4: The regression performances of SPADIS, SConES(R) and SConES(S) on AT data for $k = 500$. The rows denote phenotypes and the columns denote networks. The numbers in each cell show Pearson's squared correlation coefficients attained by SPADIS and SConES respectively. The background color encodes the difference in correlation coefficients between SPADIS and SConES. Red indicates SPADIS performs better than SConES while blue indicates otherwise. Differences that are found to be statistically significant are shown in bold and white and marked with star (*).

(PSS2) in which we allow methods to pick the SNP set size as long as the set sizes are bounded from above by 1% of all SNPs, which is 1733. We compare SPADIS with SConES(S), SConES(R), Univariate, Lasso(S) and Lasso(R). The test is performed on the following two phenotypes: 4W (worst relative performance of SPADIS compared to SConES in Fig. 4) and FLC (best performance of SPADIS compared to SConES). We let SPADIS select the maximum number of SNPs permitted (1733). In 4W, SPADIS's performance improves as it selects more SNPs compared to the previous experiment and it performs better than SConES(S), but its performance is worse than SConES(R). Lasso(R) on the other hand outperforms all methods for this phenotype. On the FLC phenotype, SPADIS outperforms all other methods.

H. Diverse Selection of SNPs

The core idea in this paper is to select a diverse set of SNPs over the SNP-SNP network. Here, we verify that the diversification is achieved with SPADIS and argue that the prediction performance increase in Section III-G is due to this effect.

We compare the average number of candidate genes hit by each method (out of 165 candidates). A gene is considered *hit* if the method selects a SNP *near* the gene ($\leq$ 20 kbp). We find that SPADIS hits [18.9 - 25.2] distinct candidate genes, which is an improvement between 31% and 66% compared to the maximum amount SConES(S) or SConES(R) could detect over different networks (see I).

Moreover, to test our intuition that SPADIS will discover SNPs that are related to diverse processes, we check how many distinct GO biological processes are hit by the SNPs discovered in each method. As shown in Table I, SNPs discovered by SPADIS covers 490 GO-terms on average (by hitting genes annotated with those GO terms). This is an increase 21% and 33% over the best SConES(S) and SConES(R) results on different networks.

Finally, we compare the methods with respect to the ratio of the number of selected SNPs that are near a candidate gene

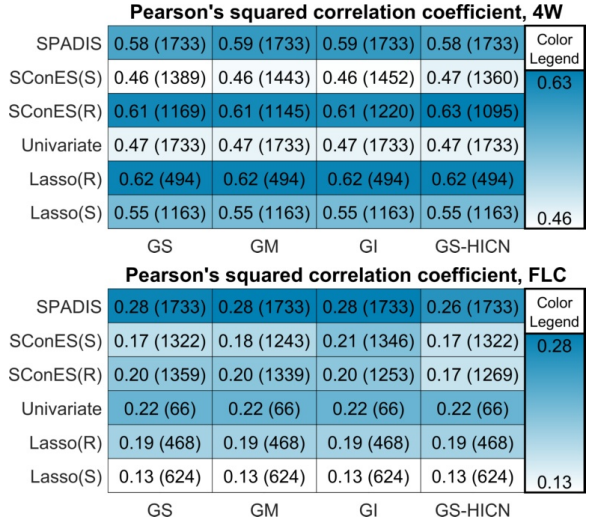


Fig. 5: The regression performances of SPADIS, SConES(R), SConES(S), lasso(S) and lasso(R) for PSS2 for two phenotypes: 4W (Left - worst performance of SPADIS in PSS1), FLC (Right - best performance in PSS1). The rows denote methods and the columns denote networks. The numbers in each cell show Pearson's squared correlation coefficients (R^2) achieved by the respective method. The number of selected SNPs are given in parentheses. The tone of the background color reflects the magnitude of R^2. Darker blue indicates higher values.

and the total number of selected SNPs, as done in Azencott *et al.* 2013 for the sake of completeness. This metric measures the *precision* of the selected SNPs, hence we denote it as such. As shown in Table I, SPADIS consistently underperforms in this metric. Nevertheless, we argue that it is not a good measure of how well the methods perform. *Precision* considers all SNPs near a candidate gene as true positives. That is why the methods that favor connectivity of SNPs on the network perform well with respect to this metric. Consider

the following extreme case: a method that selects solely a set of SNPs near a single candidate gene can easily achieve *precision* value of 1.0. Hence, *precision* indirectly rewards the selection of SNPs that fall into a smaller number of genes. On the other hand, the diversification of the SNPs in terms of genes and biological processes help explain the phenotype better. This metric is in clear contrast with the number of genes hit, hence the regression performance - see Table I.

I. Contribution of the Hi-C Data

We evaluate the information leveraged by using the Hi-C data, (i) when using other networks (GS, GM, and GI) versus (ii) when using GS-HICN, in all 17 phenotypes for $k = 500$. As shown in Fig. 6, Hi-C data provides slight improvements in regression performance over all networks: 0.7% higher than GS and GM, 1.3% higher than GI, on average. Also as it can be seen in Table I, Hi-C data using SPADIS covers the largest number of biological processes hit: 2.2% more than GS, 4.5% more than GM and 10.6% higher than GI on average. Moreover, the improvement is consistent over phenotypes compared to GM and GI: in 15 out of 17 phenotypes, GS-HICN covers more biological processes compared to GM, and this number is 16 when compared to GI, and 10 when compared to GS. We argue that similar results of GS-HICN and GS is because GS is a subset of GS-HICN. We observe similar trends when $k = 100$, $k = 250$ and $k = 1000$ as well. See corresponding results in Supplementary Figures 7-9 and Supplementary Tables 1-3.

TABLE I: Table shows statistics about the genes and biological processes hit by the selected SNPs for each method when the number of selected SNPs is 500. The values are averages over all 17 phenotypes. Genes-Hit is the number of distinct candidate flowering time genes identified. GO-Hit is the number of distinct biological processes hit by the selected SNPs. A process is considered hit if the method chooses a SNP near a gene which is annotated with that biological term. Precision is the ratio of the number of selected SNPs near candidate genes and the total number of selected SNPs. R^2 is the Pearson's squared correlation coefficient.

Network	Method	Genes-Hit	GO-Hit	Precision(%)	R^2
GS	SPADIS	24.6	500	6.5%	0.459
GS	SConES(S)	13.3	347	9.0%	0.426
GS	SConES(R)	14.8	375	8.7%	0.428
GM	SPADIS	25.2	489	6.4%	0.458
GM	SConES(S)	15.3	383	8.6%	0.425
GM	SConES(R)	15.4	385	8.6%	0.430
GI	SPADIS	18.9	462	5.5%	0.451
GI	SConES(S)	12.7	373	7.0%	0.424
GI	SConES(R)	14.4	381	8.0%	0.422
GS-HICN	SPADIS	24.8	511	6.6%	0.464
GS-HICN	SConES(S)	16.1	391	8.7%	0.437
GS-HICN	SConES(R)	15.6	387	8.6%	0.434

IV. Discussion and Conclusions

SPADIS seeks for a subset of SNPs on a network derived from biological knowledge, such that the selected SNP set is associated with the phenotype. Even though there are methods for tackling the same problem with a similar formulation, they rest on the assumption that causal SNPs tend to be connected on the network. Thus, they incorporate constraints that favor the connectivity of selected SNPs. However, we argue that SNPs affecting diverse biological processes would be complementary and explain the phenotype better. The SNPs that are nearby might not provide additional predictive power as they can be in haplotype blocks and bring redundant information. To derive meaningful insights, a method that can highlight different parts of the networks, thus, different potentially hit biological processes, will be useful. To address this issue, we propose a new formulation: As opposed to enforcing graph connectivity over the set of selected features, we set out to discover SNPs that are far apart in terms of their location on the genome, which translate into diversity in function. To the best of our knowledge, none of the current approaches operate with this principle. Our results indicate that selecting SNPs remotely located on the network indeed hit genes that are related to a larger number of distinct biological processes. This property can help in gaining more biological insights into the genetic basis of the complex traits and diseases.

The technical contribution of this paper involves formulating this principle through a submodular function. We empirically show that SPADIS can recover SNPs known to be associated with the phenotype and the optimization is efficient. Another alternative would be to formulate an optimization function that directly rewards the number of distinct process hits. However, given the incomplete knowledge of the annotations, this could lead to literature bias. Therefore, we refrain from incorporating such a term directly in the model, instead, we let the diversity on the 2D and 3D distances lead the selection.

To score each SNPs relevance to the phenotype, we use sequence kernel association test (SKAT) based on its success and for drawing a fair comparison to the existing literature. There are other alternatives such as Pearsons correlation coefficient, or maximal information coefficient Reshef *et al.* (2011), which can be used as long as the computed scores are non-negative or are transformed to a non-negative range.

For the first time, we investigate the utility of Hi-C data for selecting a SNP set. Our results show that Hi-C data provides consistently slight improvements in regression performance. We think it is a promising source of information for SNP association. We currently limit the use of data to intra-chromosomal contacts due to much better higher resolution compared to inter-chromosomal contact maps (2 kbp vs. 20 kbp). We also discard contacts that fall outside of the significance range. These choices are likely to over-constrain the method, and further research is needed to fully utilize such information, which we leave as future work.

SPADIS can be used for discovering associated SNP sets for complex genetic disorders. For instance in autism, the former research efforts have mostly focused on identifying risk genes through whole exome sequencing studies (De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014). However, close to 90% of the point mutations fall outside of the coding regions (Hindorff

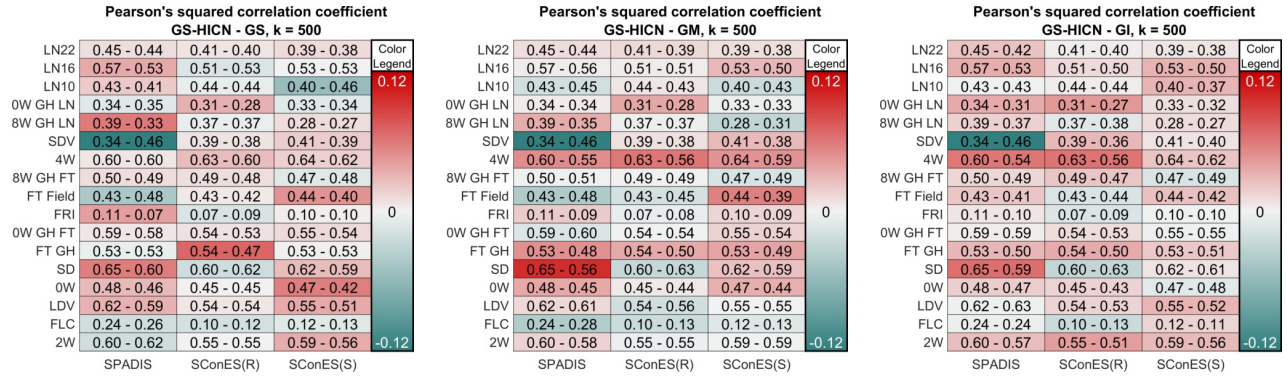


Fig. 6: The regression performances of the methods for GS, GM, GI, GS-HICN networks on AT data for $k = 500$. The rows denote phenotypes and columns denote methods. The figure focuses on the Pearson's squared correlation coefficients differences between GS-HICN and other networks. Left : GS-HICN vs GS , Middle : GS-HICN vs GM, Right : GS-HICN vs GI. The background is colored in accordance with their differences. Red indicates GS-HICN performs better than the other network while blue indicates vice versa.

et al., 2009). Discovering a set of non-coding risk mutations will certainly help to uncover the genetic architecture. Very recently, a large-scale effort to collect GWAS data of autism families along with clinical information of patients is reported (Yuen *et al.*, 2017). In this article, we introduce SPADIS and benchmark its performance on *AT* genotype and phenotypes. In future work, we will apply SPADIS on autism, which should help explain the heterogeneity in wide spectrum of phenotypes.

Acknowledgements

We thank Chlo-Agathe Azencott and Dominik Grimm for their help on running SConES, Mehmet Koyuturk and Utku Norman for feedback on SPADIS, and TUBITAK for supporting this work via TUBITAK Career Grant #116E148 to AEC.

References

Atwell, S. *et al.* (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, **465**(7298), 627–631.

Ay, F. *et al.* (2014). Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, **24**(6), 999–1011.

Ayati, M. and Koyutürk, M. (2016). Pocos: Population covering locus sets for risk assessment in complex diseases. *PLoS computational biology*, **12**(11), e1005195.

Azencott, C.-A. *et al.* (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, **29**(13), i171–i179.

Berardini, T. Z. *et al.* (2004). Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiology*, **135**(2), 745–755.

Bickmore, W. A. (2013). The spatial organization of the human genome. *Annual review of genomics and human genetics*, **14**, 67–84.

Cho, S. *et al.* (2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Annals of human genetics*, **74**(5), 416–428.

Christensen, K. and Murray, J. C. (2007). *N Engl J Med*, **356**(11), 1094–1097.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10**(6), 392–404.

de Leeuw, C. A. *et al.* (2015). Magma: generalized gene-set analysis of gwas data. *PLoS computational biology*, **11**(4), e1004219.

De Rubeis, S. *et al.* (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**(7526), 209–215.

Ding, X. *et al.* (2015). Searching high-order snp combinations for complex diseases based on energy distribution difference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **12**(3), 695–704.

Evans, D. M. *et al.* (2006). Two-stage two-locus models in genome-wide association. *PLoS Genetics*, **2**(9), e157.

Fang, G. *et al.* (2012). High-order snp combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. *PloS one*, **7**(4), e33531.

Fincham, J. R. S. (1968). Genetic complementation. *Science Progress (1933-)*, pages 165–177.

Geschwind, D. H. (2008). Autism: many genes, common pathways? *Cell*, **135**(3), 391–395.

Goldstein, D. B. *et al.* (2009). Common genetic variation and human traits. *New England Journal of Medicine*, **360**(17), 1696.

Guo, X. *et al.* (2014). Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. *BMC bioinformatics*, **15**(1), 102.

Herold, C. *et al.* (2009). Intersnp: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, **25**(24), 3275–3281.

Hindorff, L. A. *et al.* (2009). Potential etiologic and functional implications of genome-wide association loci for human

diseases and traits. *Proceedings of the National Academy of Sciences*, **106**(23), 9362–9367.

Hittner, J. B. *et al.* (2003). A monte carlo evaluation of tests for comparing dependent correlations. *The Journal of general psychology*, **130**(2), 149–168.

Hua, X. *et al.* (2010). Testing multiple gene interactions by the ordered combinatorial partitioning method in case–control studies. *Bioinformatics*, **26**(15), 1871–1878.

Huang, J. *et al.* (2011). Learning with structured sparsity. *Journal of Machine Learning Research*, **12**(Nov), 3371–3412.

Iossifov, I. *et al.* (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **515**(7526), 216–221.

Jacob, L. *et al.* (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM.

Jäger, R. *et al.* (2015). Capture hi-c identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications*, **6**.

Jiang, R. *et al.* (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC bioinformatics*, **10**(1), S65.

Kraft, P. and Hunter, D. J. (2009). Genetic risk predictionare we there yet? *New England Journal of Medicine*, **360**(17), 1701–1703.

Krause, A. and Guestrin, C. E. (2012). Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*.

Lehár, J. *et al.* (2008). High-order combination effects and biological robustness. *Molecular Systems Biology*, **4**(1), 215.

Liu, J. *et al.* (2009). Slep: Sparse learning with efficient projections. *Arizona State University*, **6**(491), 7.

Lou, X.-Y. *et al.* (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics*, **80**(6), 1125–1137.

Manolio, T. A. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

Martin, P. *et al.* (2015). Capture hi-c reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature communications*, **6**, 10069.

Miele, A. and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. *Molecular biosystems*, **4**(11), 1046–1057.

Moore, J. H. *et al.* (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**(4), 445–455.

Nelson, M. *et al.* (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome research*, **11**(3), 458–470.

Nemhauser, G. L. *et al.* (1978). An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, **14**(1), 265–294.

Phillips, P. C. (2008). Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**(11), 855–867.

Price, A. L. *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**(8), 904–909.

Rakitsch, B. *et al.* (2012). A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, **29**(2), 206–214.

Reshef, D. N. *et al.* (2011). Detecting novel associations in large data sets. *science*, **334**(6062), 1518–1524.

Ritchie, M. D. *et al.* (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, **69**(1), 138–147.

Segura, V. *et al.* (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, **44**(7), 825–830.

Shi, W. *et al.* (2008). Lasso-patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and its Interface*, **1**(1), 137.

Storey, J. D. *et al.* (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS biology*, **3**(8), e267.

Sugiyama, M. *et al.* (2014). Multi-task feature selection on multiple networks via maximum flows. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 199–207. SIAM.

Tang, W. *et al.* (2009). Epistatic module detection for case-control studies: a bayesian model with a gibbs sampling strategy. *PLoS genetics*, **5**(5), e1000464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tuo, S. *et al.* (2017). Niche harmony search algorithm for detecting complex disease associated high-order snp combinations. *Scientific Reports*, **7**(1), 11529.

van Berkum, N. L. *et al.* (2010). Hi-c: a method to study the three-dimensional architecture of genomes. *J Vis Exp*, (39).

Varadan, V. and Anastassiou, D. (2006). Inference of disease-related molecular logic from systems-based microarray analysis. *PLoS computational biology*, **2**(6), e68.

Varadan, V. *et al.* (2006). Computational inference of the molecular logic for synaptic connectivity in c. elegans. *Bioinformatics*, **22**(14), e497–e506.

Visscher, P. M. *et al.* (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**(1), 5–22.

Wan, X. *et al.* (2010). Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, **87**(3), 325–340.

Wang, C. *et al.* (2015). Genome-wide analysis of local chromatin packing in arabidopsis thaliana. *Genome research*, **25**(2), 246–256.

Wang, D. *et al.* (2011a). Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso. *Journal of agricultural, biological, and environmental statistics*, **16**(2), 170–184.

Wang, L. *et al.* (2011b). An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics*, **27**(5), 686–692.

Wang, X. *et al.* (2010a). The meaning of interaction. *Human heredity*, **70**(4), 269–277.

Wang, Z. *et al.* (2010b). A general model for multilocus epistatic interactions in case-control studies. *PLoS One*, **5**(8), e11384.

Wei, W.-H. *et al.* (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, **15**(11), 722–733.

Wiltshire, S. *et al.* (2006). Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21-25 and 10q23-26 in northern europeans. *Annals of human genetics*, **70**(6), 726–737.

Wu, M. C. *et al.* (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, **89**(1), 82–93.

Wu, T. T. *et al.* (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**(6), 714–721.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, **82**(1), 171–196.

Yosef, N. *et al.* (2007). A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data. *Bioinformatics*, **23**(2), e91–e98.

Yuen, R. K. *et al.* (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, **20**(4), 602–611.

Zhang, W. *et al.* (2010). A bayesian partition method for detecting pleiotropic and epistatic eqtl modules. *PLoS computational biology*, **6**(1), e1000642.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, **39**(9), 1167–1173.