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

Serhan Yilmaz, Oznur Tastan, and A. Ercument Cicek

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 connected SNPs on SNP-SNP networks have been proven successful in finding biologically interpretable and predictive SNPs. However, we argue that the connectedness constraint favors selecting redundant features that affect similar biological processes and therefore does not necessarily yield better predictive performance. In this paper, we propose a novel method called SPADIS that favors the selection of remotely located SNPs in order to account for their complementary effects in explaining a phenotype. SPADIS selects a diverse set of loci on a SNP-SNP network. This is achieved by maximizing a submodular set function with a greedy algorithm that ensures a constant factor approximation to the optimal solution. We compare SPADIS to the state-of-the-art method SConES, on a dataset of *Arabidopsis Thaliana* with continuous flowering time phenotypes. SPADIS has better average phenotype prediction performance in 15 out of 17 phenotypes when the same number of SNPs are selected and provides consistent improvements across multiple networks and settings on average. Moreover, it identifies more candidate genes and runs faster.

Index Terms—Phenotype prediction, GWAS, SNP selection, SNP-SNP networks, Hi-C, submodular function

1 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 [1]. However, individual variants often fail to explain the heritability of complex traits and diseases [2], [3] as a large number of variants contribute to these phenotypes and each variant has a small overall effect [4], [5]. Thus, evaluating and associating multiple loci with a given phenotype is critical [6], [7]. Indeed, detecting genetic interactions (epistasis) among pairs of loci has proven to be a powerful approach as discussed in several reviews [7], [8], [9], [10].

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) [11], [12], [13], [14], [15] and greedy search algorithms have been limited to searching for small combinations of SNPs—mostly around 3 [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. Multivariate regression-based approaches have been used [30], [31], [32], [33], [34]. 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 [35].

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 the discovery of more interpretable sets [36], [37]. 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 [38] 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 [39], [40]. These regularized methods jointly consider all predictors in the model as opposed to univariate test of associations. Nevertheless, using a SNP-SNP interaction network with these regression based methods on GWAS yields intractable number of interactions. An efficient method called SConES uses a minimum graph cut-based approach to select predictive SNPs over a network of hundreds of thousands of SNPs [35], [41]. 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

- *S. Yilmaz is with the Computer Engineering Department, Bilkent University, Ankara 06800, Turkey. E-mail: serhan.yilmaz@bilkent.edu.tr.*
- *O. Tastan is with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey. E-mail: otastan@sabanciuniv.edu.*
- *A.E. Cicek is with the Computer Engineering Department, Bilkent University, Ankara 06800, Turkey, and also with the Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA. E-mail: cicek@cs.bilkent.edu.tr.*

Manuscript received 17 Jan. 2019; revised 8 Aug. 2019; accepted 12 Aug. 2019. Date of publication 20 Aug. 2019; date of current version 3 June 2021. (Corresponding authors: Oznur Tastan and A. Ercument Cicek.) Recommended for acceptance by C. Wu. Digital Object Identifier no. 10.1109/TCBB.2019.2935437

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 leads to selecting functionally redundant SNPs and miss variants that cover different 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 [42]. While there are numerous examples of long-range (trans) genetic interactions for transcription control [43] and long-range epistasis is evident in complex genetic diseases such as type 2 diabetes [44], 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 [45].

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 and select SNPs by maximizing this set function. To maximize our proposed submodular set function, which encourages selecting a diverse SNP set with high predictive power, SPADIS uses a greedy algorithm [46]. This algorithm is guaranteed to return a solution that approximates the optimal solution up to a constant factor $(1 - 1/e)$ [46]. Submodular optimization for the task of searching for a good subset from a ground set has been used in other domains such as sensor placement [47], document summarization [48], and active learning [49]. However, studies which resort to submodular optimization for tackling biological problems are much more limited. Most notably, Libbrecht et al. [50] used this to create a non-redundant representative subset of protein sequences. Also, Wei et al. [51] choose a panel of genomic assays using submodular optimization. To the best of our knowledge, this is the first application in selecting a subset of SNPs.

We compare our algorithm to the state-of-the-art method SConES, on a GWAS of *Arabidopsis Thaliana (AT)* with 17 continuous phenotypes related to flowering time [52]. We show that SPADIS has better average regression performance in 15 out of 17 phenotypes with better runtime performance. Moreover, our method always identifies more candidate genes (up to 50 percent) and always hits more Gene Ontology (GO) terms (up to 20 percent) on average, indicating that selection of SPADIS is more diverse.

2 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) [53]. In the second step, our goal is to maximize the total score of SNP set while ensuring the selected set consists of SNPs that are remotely 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 **c** is the scoring vector such that $c_i \in \mathbb{R}_{\geq 0}$ indicates the level of the ith SNP's association with the phenotype. $D \in \mathbb{R}_{>0}$ 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 $d(i,j) \leq D$. Note that, $d(i,j) = \infty$ if j is not reachable from i, thus, does not affect the penalization due to the hard threshold of 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

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

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: $\quad S \leftarrow S \cup \arg\max_{x \in V \setminus S} F(S \cup x)$
4: **end while**

Subset selection problem with cardinality constraint is NP-hard. Thus, exhaustive search is infeasible when k or V is not small. We make use of the fact that the function defined in Equation (1) is submodular. Although submodular optimization itself is NP-hard as well [54], the greedy algorithm given in Algorithm 1, proposed by [46], 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 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.

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

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

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 $S = \{v_1, v_2, \ldots, v_n\}$ with cardinality n, let S^i denote the subset of S that contains elements up to the ith element, i.e., $S^i = \{v_1, v_2, \ldots, v_i\}$ and $S^i = \emptyset$ 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 $|S| \leq k$. □

3 Results

3.1 Dataset

We use *AT* genotype and phenotype data from [52]. 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 (PPI) 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 [55]. Candidate genes pertaining to each phenotype is retrieved from [56] and used for validating the models. Gene Ontology (GO) annotations are obtained from TAIR [57]. We obtain the Hi-C data for *AT* from [58] and process the intra-chromosomal contact matrices using the Fit-Hi-C method [59].

3.2 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 [35]: The *gene sequence (GS) network* links loci that are adjacent on the DNA sequence. The *gene membership (GM) 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

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

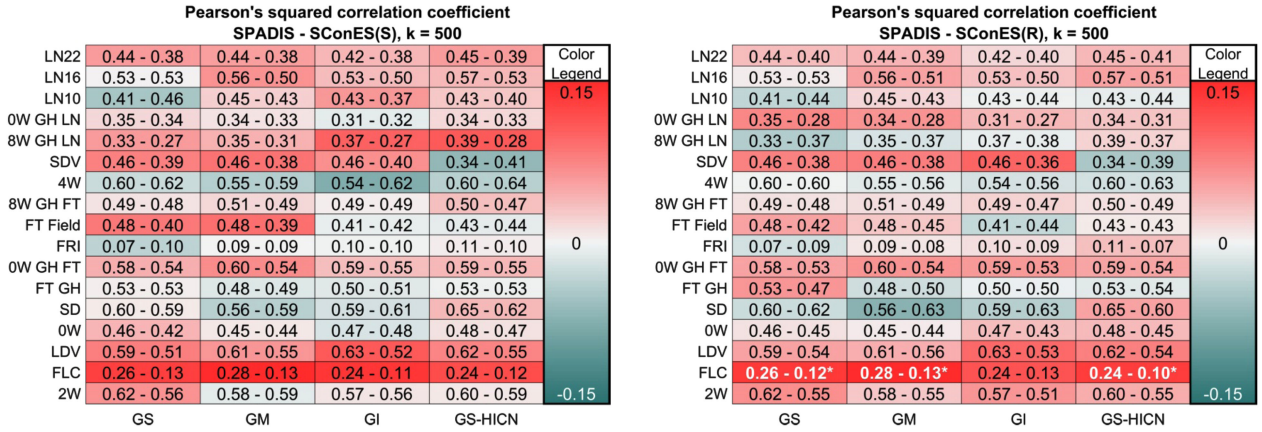

Fig. 1. The regression performance comparisons of SPADIS with SConES(S) and SConES(R) on AT data for tight cardinality constraint of $k = 500$. The rows denote phenotypes and the columns denote networks. The numbers in each cell show Pearson's squared correlation coefficients attained by SPADIS and SConES respectively. The background color encodes the difference in correlation coefficients. Red indicates SPADIS performs better than SConES while blue indicates otherwise. Differences that are found to be statistically significant are shown in bold, white font and marked with star (*). The phenotype descriptions are provided in Supplementary Table 1, available online.

20 000 bp. The *gene interaction (GI) network* also links any two loci if their nearby genes are interacting in the protein interaction network. Note that, $\mathrm{GS} \subset \mathrm{GM} \subset \mathrm{GI}$. 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). That is, 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). All networks contain 173 219 vertices. The number of (undirected) edges are as follows: GS: 173 214, GM: 11 661 166, GI: 18 134 516, GS-HICN: 2 919 607.

3.3 Compared Methods

We compare SPADIS with the following methods:

SConES. A network-constrained SNP selection method with a max-flow based solution [35].

*Univariate.*We run univariate linear regression and select SNPs that are found to be significantly associated with the phenotype (FDR-adjusted p-value ≤ 0.05) [60]. If the number of SNPs found to be associated is larger than a cardinality constraint of k (the maximum number of SNPs to be selected), only the most significant k SNPs are picked.

*Lasso.*The Lasso regression [61] that minimizes the prediction error with the $\ell 1$-regularizer of the coefficient vectors. We use the SLEP implementation [62].

3.4 Experimental Setup

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 such differences, we compare the methods using either of the following constraints: (1) Tight cardinality constraint where all methods select a fixed number of SNPs which is k, and (2) maximum cardinality constraint where the methods are allowed to select SNP sets of different sizes as long as the set sizes are smaller than an upper bound k. In both cases, SPADIS selects k SNPs.

Some of the methods that we compare SPADIS to, such as SConES and Lasso, do not operate with a cardinality constraint directly. In order to satisfy the tight cardinality constraint, during parameter selection of these methods, we apply binary search over a range of sparsity parameter values that yields numbers close to k. For the rest of the parameters or all parameters in the case of maximum cardinality constraint (including sparsity parameter), we select them using two metrics separately: *stability*, denoted with (S) and measured using the consistency index as described in [63], and *regression performance*, denoted with (R), measured using Pearson's squared correlation coefficient. The details on parameter selection for each method are provided in Supplementary Text 3.1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2019.2935437.

Since we compare SPADIS with SConES in various settings, as a first step, we verify that we make use of SConES properly by replicating the results reported in [35] using their setting. Then, we compare SPADIS with SConES and other methods using another evaluation scheme.

3.4.1 Replicating Results of SConES

Here, we use SConES ' setting explained in [35]. First, using 10-fold cross validation, the desired objective function (i.e. stability for SConES(S), regression performance for SConES (R)) are measured for all parameters tested. The parameter values that maximize the desired objective are selected, and the final SNP set is determined with these parameters. Then, for evaluation, a ridge regression is performed on the complete dataset in a 10-fold cross validated setting using this SNP set and Pearson's squared correlation coefficient is calculated for regression performance. Although this strategy is adopted by [35] due to the limited dataset size, it also implicates that the test data is used during the parameter selection step which might lead to memorization. In order to reproduce these results, we apply tight cardinality constraint during parameter selection, targeted at the number of SNPs that are reported in the paper. We show that our replicated results are on par with the reported R^2 and ratio of SNPs near candidate genes, respectively, indicating that we are able to replicate their results. These results are shown in Supplementary Figs. 1 and 2, respectively, available online. In addition, we run SConES(R), SConES(S) and SPADIS for the tight cardinality constraint of $k = 500$ using

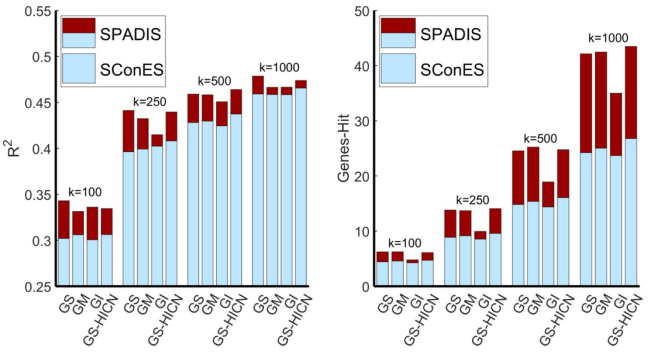

Fig. 2. The improvement of SPADIS over SConES in terms of (left) Pearson's squared correlation coefficient and (right) number of distinct candidate genes-hit for different tight cardinality constaints k. All values shown are averages over 17 phenotypes. Blue bar indicates the maximum of SConES(S) and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

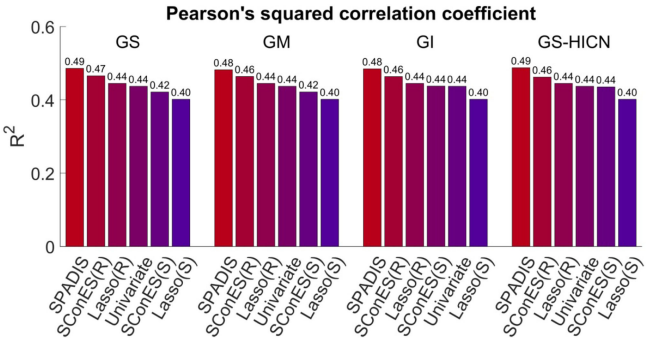

Fig. 3. Regression performances of SPADIS, SConES(S), SConES(R), Univariate, Lasso(S) and Lasso(R) averaged over 17 AT phenotypes for maximum cardinality constraint of 1,733. X-axis shows the compared methods and Y-axis shows the Pearson's squared correlation coefficient (R^2). For each network, methods are ordered in descending order of R^2.

this setting. The corresponding results suggest that SPADIS performs better in regression performance in this setting —see Supplementary Fig. 3, available online.

3.4.2 Evaluation of SPADIS and Compared Methods

We use a the nested cross-validation procedure: An outer cross-validation for evaluation, and a nested inner cross-validation for tuning the parameters. In the outer loop, the data is divided into 10 cross-validation folds. For each of the ith run in this loop corresponding to ith evaluation fold, the test data Te_i consists of the samples in the ith fold and the training data Tr_i consists of the samples in the remaining 9 folds. For tuning the parameters, the training data (Tr_i) is again split again into 10 cross-validation folds: The methods are run for each parameter on 9/10th portion of Tr_i and the performance of each parameter is measured on the held-out validation fold (remaining $1/10$th portion of Tr_i). After repeating this process 10 times for all validation folds, the parameters with the best average performance on validation folds are determined. Once the best performing parameters are set, the methods are run on the complete training data Tr_i to acquire selected SNP set S_i that is used for evaluating on fold i. Then, the Pearson's squared correlation coefficient (R^2) on the ith evaluation fold is measured after performing ridge regression using the samples in Te_i and the selected SNPs in S_i. This process is repeated for all i from 1 to 10, and average R^2 performance for all Te_i is reported.

In order to assess the performance of the methods in a controlled manner, we first conduct simulation experiments where the phenotypes are randomly generated using the real genotypes of AT. We show that the average regression performance of SPADIS is better or on a par with other methods in all simulation settings tested. See Supplementary Text 3.2, available online, for the details on simulation experiments and Supplementary Fig. 4, available online, for the corresponding results.

3.5 Phenotype Prediction Performance

3.5.1 Experiments with Tight Cardinality Constraint

First, we compare the regression performances of SConES(S), SConES(R) and SPADIS in AT data using the Pearson's squared correlation coefficient (R^2) by constraining them to select close to k SNPs (tight cardinality constraint). Here, we report results for $k = 500$ which we consider representative —see Fig. 1. The results for k = 100, 250, and 1,000 are provided in Supplementary Figs. 5, 6 and 7, respectively, available online.

Out of 68 tests that is performed for $k = 500$ over 17 phenotypes using 4 different networks separately 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 averaged over all networks. Next, we test whether the differences in R^2 are statistically significant (FDR adjusted p-value ≤ 0.05) using Fisher-Pitman permutation test with 1,000 Monte-Carlo simulations [64]. The multiple hypothesis correction is performed using BenjaminiYekutieli procedure [65]. 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.

When averaged over all k values tested and all networks, SPADIS performs better than SConES in terms of Pearson's squared correlation coefficient in 15 out of 17 phenotypes —see Supplementary Table 2, available online. Moreover, SPADIS provides a consistent improvement in regression performance over SConES when averaged over all phenotypes. This improvement of SPADIS over SConES is summarized in Fig. 2 for each network and k value tested. We also test the collective difference (averaged over all phenotypes and networks) between SPADIS and SConES (maximum of SConES (S) and SConES(R)) using two-tailed t-test and show that SPADIS is collectively better than both SConES(S) and SConES(R) for all k values tested (100, 250, 500 and 1,000) in a statistically significant manner at 0.05 level (see Supplementary Tables 3-4, available online). Note that, the improvement of SPADIS is particularly prevalent when k is smaller. On the other hand, we observe that average 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. That is why we perform the experiments with tight cardinality constraints.

3.5.2 Experiments with Maximum Cardinality Constraint

A more natural setting for SConES and other compared methods is to let them decide the number of SNPs based on

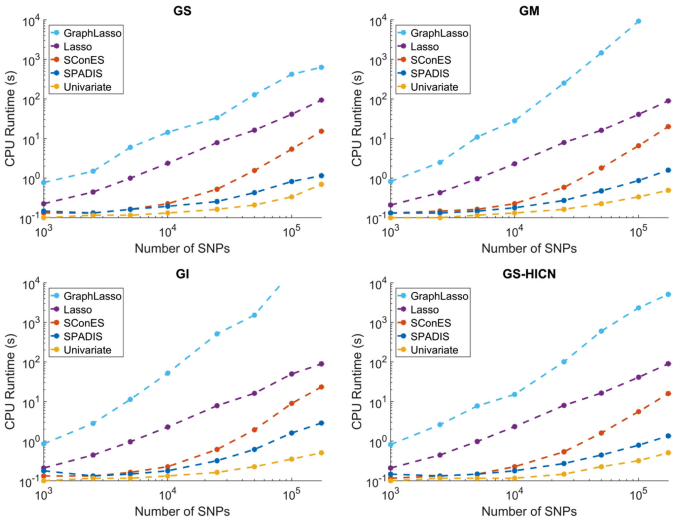

Fig. 4. CPU time measurements of SPADIS, SConES, Univariate, Lasso, and GraphLasso from 1.000 to 173.219 SNPs on four networks: (Top left) GS, (Top right) GM, (Bottom left) GI, and (Bottom right) GS-HICN.

their parameter search procedure. Hence, we perform a second set of experiments in which we allow methods to pick the SNP set size as long as the set sizes are bounded from above by 1,733 i.e., 1 percent of the number of all SNPs as done in [35]. Here, we compare SPADIS with SConES(S), SConES(R), Univariate, Lasso(S) and Lasso(R) on all phenotypes and all networks.

SPADIS is the best performing method in 8 out of 17 phenotypes on GS and GI networks and the best in 9 phenotypes on GM and GS-HICN networks (see Supplementary Figs. 8-11, available online). When regression performances (R^2) are averaged over all phenotypes for each method, SPADIS outperforms all other methods on every network (see Fig. 3). The next two best performing methods are SConES(R) and Lasso (R) respectively. Unsurprisingly, the methods that directly optimize or are tuned based on R^2 are better in regression than their stability optimizing versions on average.

Next, we check whether the differences between SPADIS and other methods are statistically significant. Out of 68 experiments of SPADIS (17 phenotypes $\times$ 4 networks) SPADIS is found to be significantly better than (i) SConES(R) in 2 experiments, (ii) Lasso(R) in 6 experiments, (iii) SConES (S) in 14 experiments, (iv) Univariate in 17 experiments, and finally, (v) Lasso(S) in 28 experiments. In none of the experiments, SPADIS is found to be significantly worse than its counterparts. See Supplementary Figs. 12-14, available online, for the corresponding results.

3.5.3 Significance of the Improvement over SConES

When we consider the R^2 distributions of all 68 tests (17 phenotypes x over 4 networks), SPADIS' improvement over SConES is statistically significant at 0.05 level with respect to two-tailed Students t-test for tight constraints when k = 100, 250, 500 and 1,000 (Supplementary Table 3, available online) as well as for the maximum constraint when k = 1733 (Supplementary Table 4, available online). Typically, there is around 25 to 40 percent overlap of the selected SNP sets between and SPADIS and SConES (Supplementary Table 5, available online).

TABLE 1
Table Shows Statistics about the Genes and Biological Processes Hit by the Selected SNPs Sets by SPADIS, SConES(S), SConES(R), Univariate, and Lasso

Metric	k	SPADIS	SConES(S)	SConES(R)	Univariate	Lasso
Genes-Hit	100	**5.9**	4.4	4.5	3.8	5.5
	250	**12.9**	8.7	9.0	7.6	10.9
	500	**23.4**	14.3	15.0	13.8	18.3
	1000	**40.8**	24.7	23.6	24.2	27.9
GO-Hit	100	**151**	114	117	137	144
	250	**306**	230	236	266	280
	500	**491**	373	382	424	441
	1000	**747**	597	581	659	636
Within-Genes-Hit	100	7.0%	**11.0%**	10.9%	8.6%	7.3%
	250	6.3%	**9.4%**	**9.4%**	7.4%	6.1%
	500	6.2%	8.3%	**8.5%**	6.9%	5.9%
	1000	6.3%	7.5%	**7.6%**	6.7%	5.8%

Tight cardinality constraint is applied for the following k values: $k = 100$, 250, 500, and 1000. The reported results are averages over all 17 phenotypes and four networks. The best result for each k is marked as bold.

3.6 Diverse Selection of SNPs

The goal of SPADIS is to select a diverse set of SNPs over the SNP-SNP network. We hypothesize that SNPs selected with SPADIS overlap with more diverse biological processes and that the prediction performance is reinforced by this effect. Here, we investigate whether this hypothesis is supported by empirical values on the 17 flowering time phenotypes of *AT*. To this end, we utilize three metrics: (1) Genes-Hit, (2) GO-Hit, and (3) Within-Genes-Hit, which are explained in the following subsections. Since the performance with respect to these metrics typically depends on the number of SNPs selected, we apply tight cardinality constraint and report the results for k = 100, 250, 500 and 1,000.

3.6.1 Evaluation with Genes-Hit Metric

First, we compare the average number of candidate genes hit by each method (out of 165 candidate genes related with flowering time). A gene is considered *hit* if the method selects a SNP *near* the gene ($\leq$20 kbp). SPADIS hits 7-46 percent more distinct candidate genes compared to the next best performing method on average, over different cardinality constraints —see Table 1. This is an indication that SPADIS realizes one of its goals which is to spatially *cover* the network and genome.

3.6.2 Evaluation with GO-Hit Metric

Here, we check how many distinct GO biological processes are hit by the SNPs selected by each method. A process is considered hit if the method chooses a SNP near a gene which is annotated with that biological term.

As shown in Table 1, SNPs discovered by SPADIS covers 151, 306, 491 and 747 GO-terms on average for $k = 100, 250, 500$ and 1000 respectively. This is an increase of 5 to 17 percent compared to the next best performing method, over different cardinality constraints. It supports our intuition that SPADIS discovers SNPs that are related to diverse processes.

3.6.3 Evaluation with Within-Genes-Hit Metric

Finally, for the sake of completeness, we compare SPADIS and other methods with respect to the *within-genes-hit* metric: the ratio of the number of selected SNPs nearby a candidate gene (SNP-hits) to the total number of selected SNPs; as done in [35]. As shown in Table 1, SPADIS consistently underperforms with this metric. Nevertheless, we believe that this particular metric is not well suited for the intended purpose. Consider the following extreme case: a method that exclusively selects a set of SNPs near a single candidate gene could trivially achieve a within-genes-hit of 1. On the other hand, the diversification of SNPs in terms of genes and biological processes can help explain the phenotype better. Furthermore, this metric ignores the fact that informative SNPs can be in the noncoding regions. Indeed, the majority of the SNPs that are near a candidate gene (dubbed *positive* SNPs) have a minor or near-zero association with the phenotype (Supplementary Fig. 17, available online). Moreover, as we compare the individual R^2 values, there is little differences between positive SNPs and *negative* SNPs (SNPs that are not near a candidate gene) —see Supplementary Fig. 17, available online.

3.7 Comparison of Networks

Here, we evaluate the performance of SPADIS and SConES when different networks are used. Fig. 2 shows that SPADIS mostly achieves the best regression performance when the GS network is used. In other words, when the genomic locations of the selected SNPs are diversified, the phenotype is predicted better. Adding functional information such as gene (GM), gene interaction (GI) and Hi-C information act as noise and slightly decreases the regression performance. This is mostly observed in the GI network. The most possible explanation is the small world property of the PPI network. That is, once this information is added SNPs all become tightly interconnected which leads SPADIS to miss some informative SNPs. Fig. 2 and Supplementary Tables 6-9, available online, also confirm that once GI is used, the number of genes-hit also goes down along with the regression performance. However, this issue with GI does not counteract against the motivation of the study as we show that the regression performance goes up once the diverse selection of SNPs is achieved (better Genes-Hit performance is an indication) and SPADIS is able to attain higher performance with the GS network. On the other hand, for SConES, functional information improves the prediction performance and it achieves the best results when the GS-HICN network is used. This result indicates that Hi-C data are informative and can help gather related SNPs together. This way it becomes possible to leverage the performance of SConES now that it can can reach out to longer distances while obtaining a connected set of SNPs. We also observe that Hi-C data leads to hitting more genes and covering more GO terms for SConES. This increase is subtle for SPADIS. A detailed account is provided in Supplementary Tables 6-15, available online, for each network and for each k value.

3.8 Time Performance

We report the CPU runtime of all methods, across a range of number of SNPs (from 1,000 to 173 219) and all 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 x10 evaluation folds x10 training folds x at least 7 parameters). Results show that SPADIS is more efficient than all other methods except the Univariate (baseline) method —see Fig. 4. GraphLasso described in [39] do not scale to SNP selection problem in GWAS. For this reason, it is not included in the experiments performed on AT data.

4 Discussion

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 other network based methods for tackling the same problem, 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 selecting connected SNPs together might not provide additional predictive power as they can be in haplotype blocks and bring redundant information. Moreover, a method that highlights different parts of the network could be useful because it can potentially recover different biological processes: SNPs affecting diverse biological processes would be complementary and explain the phenotype better. To address these issues, 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 process 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 locations lead the diverse selection.

In our experiments, 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 literature. There are other alternatives such as Pearson's correlation coefficient, or maximal information coefficient [66], which can easily be used with SPADIS as long as the computed scores are non-negative or are transformed to a non-negative range.

For a SNP selection method like SPADIS, we believe the parameters determining the size of the final SNP set (e.g.,

cardinality constraint k in SPADIS or η in SConES) are critical parameters, ultimately determining the tradeoff between computational resources and performance. Thus, we believe those parameters are best determined by the users who can make a judgement about the available computational resources and time at their disposal. Therefore, in our framework, we assume the users can decide at most how many SNPs they would like to acquire, i.e., the maximum cardinality constraint. A similar constraint is also employed by other SNP selection methods like SConES. For example, Azencott et al. (2013) allowed at most 1,733 SNPs (corresponding to 1 percent of all SNPs) to be selected in their experiments for the evaluation of SConES.

We suggest that the users of SPADIS select the highest maximum cardinality constraint they can afford, given their time constraints considering the costs and limitations of follow-up experiments. Having decided the maximum cardinality constraint, users can also optimize k using cross-validation as we have shown in the experiment with maximum cardinality constraint (MCC) of 1,500 (see Supplementary Figures 15 and 16). However, we do not suggest this. It is a time-consuming process and we think that well-established regression methods with shrinkage parameters, such as ridge regression, can implicitly acquire a reduced SNP set to perform prediction sufficiently well. Moreover, as it can be seen from our experiment with MCC of 1,500, optimization of k parameter in SPADIS tends to favor larger k values and we do not estimate a significant improvement of optimizing k with cross-validation over selecting the maximum allowed k. Therefore, we suggest that the users select k equal to the maximum cardinality constraint they determine according to their time constraints.

In this study, we also investigate the utility of Hi-C data for selecting a SNP set for the first time as far as we are aware. Our results show that Hi-C data helps linking related mutations and especially helps SConES to reach out to informative SNPs. 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 versus 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.

We benchmark the performance of SPADIS on flowering time phenotypes of *AT*. Alternatively, SPADIS can be used for discovering associated SNP sets for complex genetic disorders as well. For instance in autism, research efforts have mostly focused on identifying risk genes through whole exome sequencing studies [67], [68]. However, close to 90 percent of the point mutations fall outside of the coding regions [38] and discovering a set of non-coding risk mutations would certainly help to uncover the genetic architecture. In future work, using the GWAS data of autism families that are reported in [69], we plan to apply SPADIS on autism, which should help explain the heterogeneity in wide spectrum of phenotypes.

Acknowledgments

The authors thank Chloé-Agathe Azencott and Dominik Grimm for their help on running SConES, Mehmet Koyuturk and Utku Norman for their feedback on SPADIS. They also thank TUBITAK for supporting this research via Career Grant #116E148 to AEC.

References

[1] P. M. Visscher, et al., "10 years of GWAS discovery: Biology, function, and translation," *Amer. J. Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.
[2] T. A. Manolio, et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.
[3] D. B. Goldstein, et al., "Common genetic variation and human traits," *New England J. Med.*, vol. 360, no. 17, 2009, Art. no. 1696.
[4] P. Kraft and D. J. Hunter, "Genetic risk predictionare we there yet?" *New England J. Med.*, vol. 360, no. 17, pp. 1701–1703, 2009.
[5] K. Christensen and J. C. Murray, "What Genome-wide association studies can do for medicine," *New England J. Med.*, vol. 356, no. 11, pp. 1094–1097, 2007, pMID: 17360987. [Online]. Available: http://dx.doi.org/10.1056/NEJMp068126
[6] J. H. Moore, et al., "Bioinformatics challenges for genome-wide association studies," *Bioinf.*, vol. 26, no. 4, pp. 445–455, 2010.
[7] H. J. Cordell, "Detecting gene–gene interactions that underlie human diseases," *Nature Rev. Genetics*, vol. 10, no. 6, pp. 392–404, 2009.
[8] P. C. Phillips, "Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems," *Nature Rev. Genetics*, vol. 9, no. 11, pp. 855–867, 2008.
[9] X. Wang, et al., "The meaning of interaction," *Human Heredity*, vol. 70, no. 4, pp. 269–277, 2010.
[10] W.-H. Wei, et al., "Detecting epistasis in human complex traits," *Nature Rev. Genetics*, vol. 15, no. 11, pp. 722–733, 2014.
[11] M. Nelson, et al., "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation," *Genome Res.*, vol. 11, no. 3, pp. 458–470, 2001.
[12] X.-Y. Lou, et al., "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence," *Amer. J. Human Genetics*, vol. 80, no. 6, pp. 1125–1137, 2007.
[13] J. Lehár, et al., "High-order combination effects and biological robustness," *Mol. Syst. Biol.*, vol. 4, no. 1, 2008, Art. no. 215.
[14] X. Hua, et al., "Testing multiple gene interactions by the ordered combinatorial partitioning method in case–control studies," *Bioinf.*, vol. 26, no. 15, pp. 1871–1878, 2010.
[15] G. Fang, et al., "High-order SNP combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions," *PloS One*, vol. 7, no. 4, 2012, Art. no. e33531.
[16] J. D. Storey, et al., "Multiple locus linkage analysis of genomewide expression in yeast," *PLoS Biol.*, vol. 3, no. 8, 2005, Art. no. e267.
[17] D. M. Evans, et al., "Two-stage two-locus models in genome-wide association," *PLoS Genetics*, vol. 2, no. 9, 2006, Art. no. e157.
[18] V. Varadan and D. Anastassiou, "Inference of disease-related molecular logic from systems-based microarray analysis," *PLoS Comput. Biol.*, vol. 2, no. 6, 2006, Art. no. e68.
[19] V. Varadan, et al., "Computational inference of the molecular logic for synaptic connectivity in C. Elegans," *Bioinf.*, vol. 22, no. 14, pp. e497–e506, 2006.
[20] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, 2007.
[21] C. Herold, et al., "INTERSNP: Genome-wide interaction analysis guided by a priori information," *Bioinf.*, vol. 25, no. 24, pp. 3275–3281, 2009.
[22] W. Tang, et al., "Epistatic module detection for case-control studies: A Bayesian model with a Gibbs sampling strategy," *PLoS Genetics*, vol. 5, no. 5, 2009, Art. no. e1000464.
[23] R. Jiang, et al., "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. S65.
[24] Z. Wang, et al., "A general model for multilocus epistatic interactions in case-control studies," *PLoS One*, vol. 5, no. 8, 2010, Art. no. e11384.
[25] X. Wan, et al., "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *Amer. J. Human Genetics*, vol. 87, no. 3, pp. 325–340, 2010.
[26] X. Guo, et al., "Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering," *BMC Bioinf.*, vol. 15, no. 1, 2014, Art. no. 102.

[27] X. Ding, et al., "Searching high-order SNP combinations for complex diseases based on energy distribution difference," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 3, pp. 695–704, May/Jun. 2015.
[28] M. Ayati and M. Koyutürk, "PoCos: Population covering locus sets for risk assessment in complex diseases," *PLoS Comput. Biol.*, vol. 12, no. 11, 2016, Art. no. e1005195.
[29] S. Tuo, et al., "Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 11529.
[30] W. Shi, et al., "LASSO-Patternsearch algorithm with application to ophthalmology and genomic data," *Statist. Interface*, vol. 1, no. 1, 2008, Art. no. 137.
[31] T. T. Wu, et al., "Genome-wide association analysis by lasso penalized logistic regression," *Bioinf.*, vol. 25, no. 6, pp. 714–721, 2009.
[32] S. Cho, et al., "Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis," *Ann. Human Genetics*, vol. 74, no. 5, pp. 416–428, 2010.
[33] D. Wang, et al., "Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO," *J. Agricultural Biol. Environ. Statist.*, vol. 16, no. 2, pp. 170–184, 2011.
[34] B. Rakitsch, et al., "A lasso multi-marker mixed model for association mapping with population structure correction," *Bioinf.*, vol. 29, no. 2, pp. 206–214, 2012.
[35] C.-A. Azencott, et al., "Efficient network-guided multi-locus association mapping with graph cuts," *Bioinf.*, vol. 29, no. 13, pp. i171–i179, 2013.
[36] L. Wang, et al., "An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies," *Bioinf.*, vol. 27, no. 5, pp. 686–692, 2011.
[37] C. A. de Leeuw, et al., "MAGMA: Generalized gene-set analysis of GWAS data," *PLoS Comput. Biol.*, vol. 11, no. 4, 2015, Art. no. e1004219.
[38] L. A. Hindorff, et al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proc. Nat. Academy Sci. United States America*, vol. 106, no. 23, pp. 9362–9367, 2009.
[39] L. Jacob, et al., "Group lasso with overlap and graph lasso," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 433–440.
[40] J. Huang, et al., "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, no. Nov., pp. 3371–3412, 2011.
[41] M. Sugiyama, et al., "Multi-task feature selection on multiple networks via maximum flows," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 199–207.
[42] J. R. S. Fincham, "Genetic complementation," *Sci. Progress*, vol. 56, pp. 165–177, 1968.
[43] A. Miele and J. Dekker, "Long-range chromosomal interactions and gene regulation," *Mol. Biosyst.*, vol. 4, no. 11, pp. 1046–1057, 2008.
[44] S. Wiltshire, et al., "Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21–25 and 10q23-26 in northern europeans," *Ann. Human Genetics*, vol. 70, no. 6, pp. 726–737, 2006.
[45] D. H. Geschwind, "Autism: Many genes, common pathways?" *Cell*, vol. 135, no. 3, pp. 391–395, 2008.
[46] G. L. Nemhauser, et al., "An analysis of approximations for maximizing submodular set functionsi," *Math. Program.*, vol. 14, no. 1, pp. 265–294, 1978.
[47] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, no. Feb., pp. 235–284, 2008.
[48] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Human Language Technol.-Volume 1*, 2011, pp. 510–520.
[49] C. Orhan and O. Tastan, "Active learning methods based on statistical leverage scores," *arXiv:1812.02497*, 2018.
[50] M. W. Libbrecht, J. A. Bilmes, and W. S. Noble, "Choosing non-redundant representative subsets of protein sequence data sets using submodular optimization," *Proteins: Structure Function Bioinf.*, vol. 86, no. 4, pp. 454–466, 2018.
[51] K. Wei, M. W. Libbrecht, J. A. Bilmes, and W. S. Noble, "Choosing panels of Genomics assays using submodular optimization," *Genome Biol.*, vol. 17, no. 1, 2016, Art. no. 229.
[52] S. Atwell, et al., "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines," *Nature*, vol. 465, no. 7298, pp. 627–631, 2010.
[53] M. C. Wu, et al., "Rare-variant association testing for sequencing data with the sequence kernel association test," *Amer. J. Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.
[54] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Proc. 21st Conf. Uncertainty Artif. Intell.*, 2005, pp. 324–331. [Online]. Available: http://dl.acm.org/citation.cfm?id=3020336.3020377
[55] A. L. Price, et al., "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.
[56] V. Segura, et al., "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations," *Nature Genetics*, vol. 44, no. 7, pp. 825–830, 2012.
[57] T. Z. Berardini, et al., "Functional annotation of the arabidopsis genome using controlled vocabularies," *Plant Physiology*, vol. 135, no. 2, pp. 745–755, 2004. [Online]. Available: http://www.plantphysiol.org/content/135/2/745
[58] C. Wang, et al., "Genome-wide analysis of local chromatin packing in arabidopsis thaliana," *Genome Res.*, vol. 25, no. 2, pp. 246–256, 2015.
[59] F. Ay, et al., "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts," *Genome Res.*, vol. 24, no. 6, pp. 999–1011, 2014.
[60] D. Yekutieli and Y. Benjamini, "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics," *J. Statistical Planning Inference*, vol. 82, no. 1, pp. 171–196, 1999.
[61] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.
[62] J. Liu, et al., "SLEP: Sparse learning with efficient projections," *Arizona State Univ.*, vol. 6, no. 491, 2009, Art. no. 7.
[63] L. I. Kuncheva, "A stability index for feature selection," in *Proc. IASTED Int. Multi-Conf. Artif. Intell. Appl.*, 2007, pp. 421–427.
[64] J. B. Hittner, K. May, and N. C. Silver, "A Monte Carlo evaluation of tests for comparing dependent correlations," *J. Gen. Psychology*, vol. 130, no. 2, pp. 149–168, 2003.
[65] Y. Benjamini, D. Yekutieli, et al., "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.
[66] D. N. Reshef, et al., "Detecting novel associations in large data sets," *Sci.*, vol. 334, no. 6062, pp. 1518–1524, 2011. [Online]. Available: http://science.sciencemag.org/content/334/6062/1518
[67] S. De Rubeis, et al., "Synaptic, transcriptional and chromatin genes disrupted in autism," *Nature*, vol. 515, no. 7526, pp. 209–215, 2014.
[68] I. Iossifov, et al., "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.
[69] R. K. Yuen, et al., "Whole Genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder," *Nature Neurosci.*, vol. 20, no. 4, pp. 602–611, 2017.

Serhan Yilmaz received the BS degree in electrical engineering from Middle East Technical University, in 2016, and the MS degree in computer engineering from Bilkent University, in 2018. Currently, he is working toward the PhD degree in the Electrical Engineering and Computer Science Department, Case Western Reserve University.

Oznur Tastan received the BS degree in biological sciences and bioengineering from Sabanci University, in 2004, and the MS and PhD degrees in computer science from Carnegie Mellon University, in 2007 and 2011, respectively. She worked as a postdoctoral researcher with Microsoft Research New England. She was an assistant professor with the Computer Engineering Department, Bilkent University, from 2012 to 2017. Since then, she is an assistant professor with the Faculty of Natural Sciences, Sabanci University, affiliated with the Computer Science and Engineering and Molecular Biology Genetics and Bioengineering programs.

A. Ercument Cicek received the BS and MS degrees in computer science and engineering from Sabanci University, in 2007 and 2009, respectively, and the PhD degree in computer science from Case Western Reserve University, in 2013. Then, he worked as a Lane fellow in computational biology with Carnegie Mellon University till 2015. Since then, he has been an assistant professor with the Computer Engineering Department, Bilkent University, and is an adjunct faculty member with the Computational Biology Department, Carnegie Mellon University.