SPADIS: SELECTING PREDICTIVE AND DIVERSE SNPS IN GWAS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Serhan Yılmaz
May 2018

SPADIS: Selecting Predictive and Diverse SNPs in GWAS
By Serhan Yılmaz
May 2018

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercüment Çiçek(Advisor)

Öznur Taştan Okan(Co-Advisor)

Mehmet Koyutürk

Tolga Can

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

SPADIS: SELECTING PREDICTIVE AND DIVERSE SNPS IN GWAS

Serhan Yılmaz
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
May 2018

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 Single Nucleotide Polymorphisms (SNPs) on SNP-SNP networks has 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. To this end, 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 $(1 - 1/e)$ approximation to the optimal solution. We compare SPADIS to the state-of-the-art method SConES, on a dataset of *Arabidopsis Thaliana* 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. 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 improvements in regression performance across all methods.

Keywords: GWAS, SNP Selection, SNP-SNP Networks, Hi-C, Submodularity.

ÖZET

SPADIS: GWAS ÇALIŞMALARINDA AÇIKLAYICI VE ÇEŞİTLİ SNP SEÇİMİ

Serhan Yılmaz
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Mayıs 2018

Genom çapında ilişkilendirme çalışmalarında (Genome-Wide Association Studies - GWAS) saptanan genetik varyasyonlar nadiren tek başlarına karmaşık hastalıkların kalıtsal aktarımını açıklamakta başarılı olabilmektedirler. Şimdiye kadar, fenotiple ilişkili olan varyasyonların bir alt kümesini seçmek amacıyla çeşitli yöntemler geliştirilmiştir. Bu yöntemlerden bazılarında, tekil nükleotit polimorfizmlerini (Single Nucleotide Polymorphism - SNP) bir SNP-SNP ağında bağlı şekilde seçmeyi ödüllendiren bir yaklaşım izlenmiştir. Bu yaklaşımın fenotipi açıklayıcı ve biyolojik anlamda yorumlanabilir SNP'leri bulmakta başarılı sonuçlar elde ettiği de gösterilmiştir. Fakat, bizim hipotezimize göre, ağ üzerinde bağlılık kısıtlaması yapmak benzer biyolojik süreçleri etkileyen, ihtiyaç fazlası SNP'lerin seçimini destekler ve bu da fenotipi açıklama gücünde potensiyel bir kayba sebep olabilir. Bu doğrultudaki çalışmamızda, birbirini tamamlayıcı etkiye sahip olması adına, ağ üzerinde yakın SNP'leri seçmekten kaçınan SPADIS adında yeni bir yöntem sunulmaktadır. SPADIS bu işlevini, altmodüler bir fonksiyonun azami değerine yakınlığını bir sabit çarpan $(1 - 1/e)$ ile garanti edebilen açgözlü (greedy) bir algoritma ile yerine getirmektedir. SPADIS, deneylerimizde, modern yöntemlerden biri olan SConES ile *Arabidopsis Thaliana* verisinde karşılaştırılmıştır: Fenotip açıklayabilme ölçütünde ortalama olarak 17 fenotipin 15'inde daha iyi sonuçlar elde edilmekle birlikte, çeşitli ağ ve kurgular arasında istikrarlı gelişmeler de sağlanmıştır. Üstelik, SPADIS'in fenotip ile ilişkili daha fazla sayıda gen saptadığı ve çalışmasını daha kısa sürede tamamladığı gösterilmiştir. Ayrıca, deneylerimizde, Hi-C verisinin SNP seçimi problemi çerçevesinde SNP-SNP ağı oluşturmadaki kullanımı incelenmiş ve bunun test edilen tüm yöntemlerin fenotipi açıklamasına katkıda bulunduğu gözlemlenmiştir.

Anahtar sözcükler: GWAS, SNP seçimi, SNP-SNP ağları, Hi-C, Altmodülerlik.

Acknowledgement

First and foremost, I would like to thank my advisor A. Ercüment Çiçek for his guidance, patience and support in overcoming numerous obstacles that I have faced throughout this work. He provided me the freedom to shape my research while steering me in the right direction whenever I needed it. I am truly grateful for that.

Additionally, I would like thank Öznur Taştan Okan for her support and valuable contributions in this project. I am grateful to Mehmet Koyutürk and Tolga Can for reading my thesis and showing the kindness to be in my thesis committee. I thank TUBITAK for supporting this research via Career Grant #116E148 and I thank Chloé-Agathe Azencott and Dominik Grimm for their help on running SConES.

I would also like to thank my fellow graduate students for their feedback, cooperation and friendship. Especially, I am grateful to Utku Norman for his invaluable comments, productive criticism and overall support.

Finally, I would like to express my utmost gratitude to my parents and my sister for their unconditional love and continuous encouragement throughout my years of study. I would not be who I am and this thesis would not have been possible without them. Thank you.

Contents

List of Figures

List of Tables

Chapter 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 [8, 7, 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, 16] and greedy search algorithms have been limited to searching for small combinations of SNPs – mostly around 3 [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. Multivariate regression-based approaches have been used [33, 34, 35, 36, 37]. 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 [38].

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 [39, 40]. 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 [41] 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 [42, 43]. 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 [38, 44]. 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 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 [45]. While there are numerous examples of long-range (trans) genetic interactions for transcription control [46] and long-range epistasis is evident in complex genetic diseases such as type 2 diabetes [47], 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 [48].

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 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 method SConES, on a GWAS of *Arabidopsis Thaliana (AT)* with 17 continuous phenotypes related to flowering time [49]. 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%) and always hits more Gene Ontology (GO) terms (up to 20%) on average, indicating that selection of SPADIS is more diverse.

Finally, we 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 [50] and contacts in 3D have been shown to affect the phenotype [51, 52]. Hi-C technology can detect the 3D conformation genome-wide and yield contact maps which show loci that reside nearby in 3D [53]. 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.

Chapter 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) [54]. 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 2.1 to encode this intuition.

$$F(S) = \sum_{i \in S} \left(c_i + \beta \left(1 - \sum_{j \in S} \frac{K(i,j)}{2k} \right) \right) \tag{2.1}$$

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

Here $\mathbf{c}$ is the scoring vector such that $c_i \in \mathbb{R}_{\geq 0}$ indicates the level of the i-th 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 $i, j \in V$. Note that, $d(i, j) = \infty$ if j is not reachable from i. $K(i, j)$ 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:

$$S^* = \underset{S \subseteq V, |S| \leq k}{\operatorname{argmax}} F(S) \tag{2.2}$$

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 2.1 is submodular. Although submodular optimization itself is NP-hard as well [55], the greedy algorithm given in **Algorithm 1**, proposed by [56], 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.

Algorithm 1 Greedy Algorithm

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

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 2.1, we can rewrite G.

$$\begin{aligned} G(S,x) &= F(S \cup \{x\}) - F(S) \\ &= \sum_{i \in S \cup \{x\}} c_i + \beta \sum_{i \in S \cup \{x\}} \left(1 - \sum_{j \in S \cup \{x\}} \left(\frac{K(i,j)}{2k} \right) \right) \\ &\quad - \left(\sum_{i \in S} c_i + \beta \sum_{i \in S} \left(1 - \sum_{j \in S} \left(\frac{K(i,j)}{2k} \right) \right) \right) \\ &= c_x + \beta - \frac{\beta}{2k} \sum_{i \in S} (K(i,x) + K(x,i)) \end{aligned} \tag{2.3}$$

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 2.1 is submodular.*

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

$$G(A, x) - G(B, x) \geq 0 \tag{2.4}$$

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

$$\begin{aligned} H(A,B,x) &= G(A,x) - G(B,x) \\ &= \left(c_x + \beta - \frac{\beta}{2k}\left(\sum_{i \in A}(K(i,x) + K(x,i))\right)\right) \\ &\quad - \left(c_x + \beta - \frac{\beta}{2k}\left(\sum_{i \in B}(K(i,x) + K(x,i))\right)\right) \\ &= \frac{\beta}{2k}\left(\sum_{i \in B \setminus A}(K(i,x) + K(x,i))\right) \end{aligned} \tag{2.5}$$

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 2.1 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;

$$\begin{aligned} G(S,x) &\geq c_x + \beta - \frac{\beta}{2k}\sum_{i \in S}(1+1) \\ &\geq c_x + \beta - \frac{\beta}{2k}2|S| \\ &\geq c_x + \beta(1 - |S|/k) \\ &\geq (1 - |S|/k) \\ &\geq 0 \end{aligned} \tag{2.6}$$

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

Lemma 3. $F(S)$ given in Equation 2.1 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. $S^i = \{v_1, v_2, ..., v_i\}$ and $S^i = \emptyset$ for $i = 0$. $F(S)$ can be decomposed as the summation of marginal gain functions:

$$F(S) = F(\emptyset) + \sum_{i=1}^{n} G(S^{i-1}, v_i) \tag{2.7}$$

$F(\emptyset) = 0$ 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 $|S|\leq k$. □

Chapter 3

Results

3.1 Dataset

We use *AT* genotype and phenotype data from [49]. 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 [57]. Candidate genes pertaining to each phenotype is retrieved from [58] and used for validating the models. Gene Ontology (GO) annotations are obtained from TAIR [59]. We obtain the Hi-C data for *AT* from [60] and process the intra-chromosomal contact matrices using the Fit-Hi-C method [61].

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

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 [38]: The *GS (gene sequence) network* links loci that are adjacent on the DNA sequence. 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). 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 [38].
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) [62]. 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 [63] that minimizes the prediction error with the ℓ1-regularizer of the coefficient vectors. We use the SLEP implementation [64].

GraphLasso and GroupLasso: We also compare our method to GraphLasso and GroupLasso [42] through simulations, using the implementation in the SLEP package. Due to the prohibitive runtimes of these algorithms, they are excluded from the comparison on *AT* dataset (see *Time Performance* section). 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) 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 that do not participate in the gene 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.

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.

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 [38] using their setting. Then, we compare SPADIS with SConES and other methods using another evaluation scheme.

3.4.1 Parameter Selection

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 found using the consistency index as described in [65], and *regression performance*, denoted with (R), measured using Pearson's squared correlation coefficient. The details on parameter selection for each method are given as follows:

For stability based parameter selection (S), the common set of SNPs consistently selected across all training folds are chosen. In regression performance based parameter selection (R), SNP set is selected by a single run with the best parameter set on the training data even though the regression performance is still measured via 10-fold cross-validation. 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.

The parameter selection process differs for tight and maximum cardinality constraints. In both cases, measurements of regression performance in methods denoted with (R), are done by applying ridge regression using 10-fold cross-validation on the training data set. However, in maximum cardinality constraint for $k = 1733$, due to memory constraints, a parameter value that results in SNP sets with cardinality > 3466 (1733×2) in at least one training fold, is deemed invalid.

When tight cardinality constraint is applied with k target SNPs, parameter selection for each method are performed as follows:

SPADIS: The distance parameter (D) and the penalty parameter (β) are selected by applying two consecutive line searches. First, seven different values of D within $[D_{min}, D_{max}]$ varying in logarithmic scale are tested when $\beta = \infty$ and the D value that maximizes the training regression performance is selected. D_{min} is the minimum edge weight i.e. $D_{min} = \min_{e \in E} w(e)$. D_{max} is the maximum distance such that penalty can be avoided i.e. $K(i, j) = 0$, $\forall i, j \in S^*$ where S^* is the selected set when $\beta = \infty$ and $D = D_{max}$. We find D_{max} by binary search. Having set the D value, 16 β values within $[10^{-4}, 1]\beta_{max}$ in logarithmic scale are tested where $\beta_{max} = 2kD \max(\mathbf{c})$. The β value that maximizes the training regression performance is selected.

SConES(S): We perform a line search for the connectivity parameter (λ) within the range $\left[\frac{\min(\mathbf{c})}{\delta}, \frac{\max(\mathbf{c})}{\delta}\right]$, in logarithmic scale where δ is the average degree of the graph. For each λ value, the sparsity parameter (η) is set such that the number of selected SNPs is close to k. This is achieved by a binary search. Then, the most *stable* λ is selected using the stability criteria [65]. Note that this is the parameter selection method described in [38]. We call the SConES runs with this parameter selection technique SConES(S) to distinguish it from the following version.

SConES(R): Since, we select the parameters of SPADIS with respect to regression performance, for a fair comparison, we select SConES' parameters with respect to regression performance as well and call the runs of SConES with this technique SConES(R). As in SConES(S), we perform line search for λ and binary search for η parameters. The λ value that maximizes training regression performance is selected.

Lasso, GroupLasso and GraphLasso: The regularization parameter (λ) is determined by binary search such that the number of selected SNPs is close to k.

When maximum cardinality constraint of k SNPs is applied, parameter selection is performed as follows:
SPADIS: It is the same as when tight cardinality constraint is applied. The number of SNPs to be selected is set to k.

SConES(S): First, binary search targeting k SNPs is done on η to find a valid lower bound for η. Then, $7x7$ grid search experiments are conducted and the most *stable* parameter pair across the training folds is selected.

SConES(R): First, a valid lower bound (that yields k SNPs selected) for η is found using binary search. Then, $7x7$ grid search experiments are conducted and the parameter pair that performs best in regression is selected.

Lasso(S): First the lower bound of the regularization parameter (λ) that yields k SNPs is found using binary search. Then, seven values of λ in logarithmic scale are tested and the most *stable* one is selected.

Lasso(R): First the lower bound of the regularization parameter (λ) that yields k SNPs is found using binary search. Then, seven values of λ in logarithmic scale are tested and the one that maximizes the regression performance on training data is selected.

3.4.2 Replicating Results of SConES

Here, we use SConES' setting explained in [38]. 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 [38] 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 the 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 Figure 3.1 and Figure 3.2, respectively. In addition, we run SConES(R), SConES(S) and SPADIS for the tight cardinality constraint of $k = 500$ using this setting. The corresponding results suggest that SPADIS performs better in regression performance in this setting—see Figure 3.3.

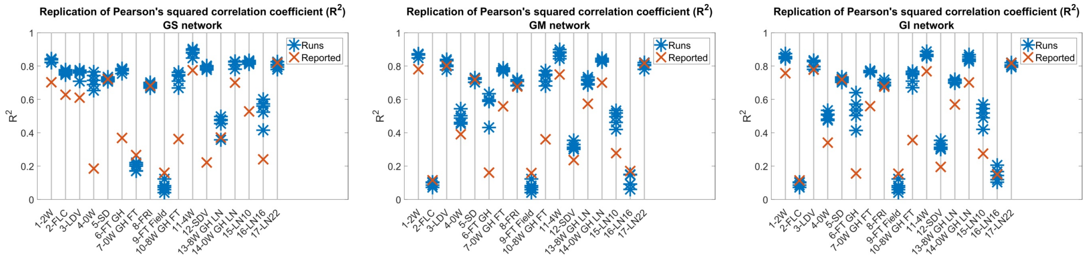


Figure 3.1: The replication of the Pearson's squared correlation coefficients (R^2) results of Azencott et. al, 2013 successfully. We do not have the actual folds used in the paper. Using binary search, we adjust the parameters so that SConES selects the same the number of SNPs used in the paper to produce the reported R^2 value. Each sub-figure represents the underlying network: GS (left), GM (center), and GI (right). x and y axis show phenotypes and R^2 values for each phenotype respectively. Cross signs indicate the value reported in Azencott et. al, 2013, for that phenotype. Blue stars show the results we obtained for 5 different 10-fold cross validation splits.

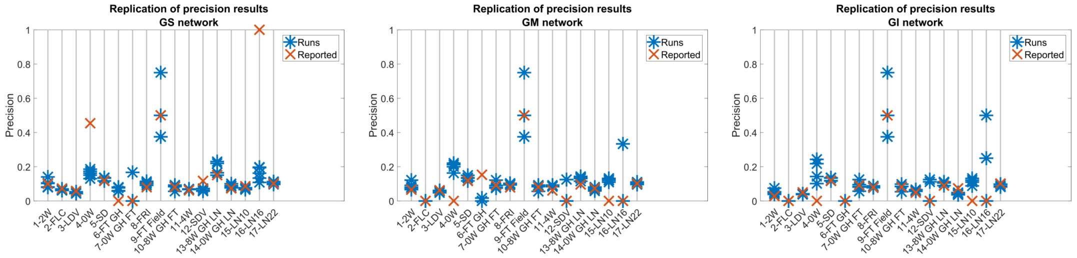
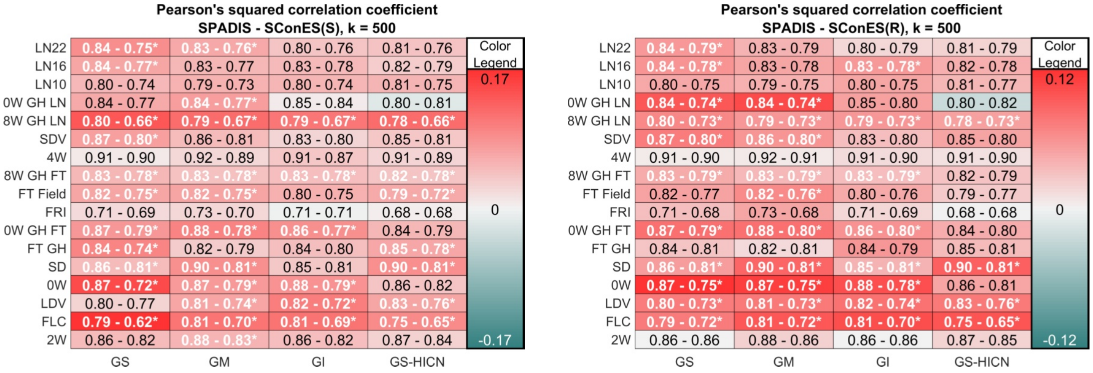


Figure 3.2: The replication of the precision results of Azencott et. al, 2013. Precision is defined as the ratio of the number of SNPs selected that are near candidate genes and total number of SNPs selected. We do not have the actual folds used in the paper. Using binary search, we adjust the parameters so that SConES selects the same the number of SNPs used in the paper to produce the reported precision value. Each sub-figure represents the underlying network: GS (left), GM (center), and GI (right). x and y axis show phenotypes and precision for each phenotype respectively. Cross signs indicate the reported precision value in Azencott et. al, 2013, for that phenotype. Blue stars show the results we obtained for 5 different 10-fold cross validation splits.



Pearson's squared correlation coefficient
SPADIS - SConES(S), k = 500

	GS	GM	GI	GS-HICN
LN22	**0.84 - 0.75***	**0.83 - 0.76***	0.80 - 0.76	0.81 - 0.76
LN16	**0.84 - 0.77***	0.83 - 0.77	0.83 - 0.78	0.82 - 0.79
LN10	0.80 - 0.74	0.79 - 0.73	0.80 - 0.74	0.81 - 0.75
0W GH LN	0.84 - 0.77	**0.84 - 0.77***	0.85 - 0.84	0.80 - 0.81
8W GH LN	**0.80 - 0.66***	**0.79 - 0.67***	**0.79 - 0.67***	**0.78 - 0.66***
SDV	**0.87 - 0.80***	0.86 - 0.81	0.83 - 0.80	0.85 - 0.81
4W	0.91 - 0.90	0.92 - 0.89	0.91 - 0.87	0.91 - 0.89
8W GH FT	**0.83 - 0.78***	**0.83 - 0.78***	**0.83 - 0.78***	**0.82 - 0.78***
FT Field	**0.82 - 0.75***	**0.82 - 0.75***	0.80 - 0.75	**0.79 - 0.72***
FRI	0.71 - 0.69	0.73 - 0.70	0.71 - 0.71	0.68 - 0.68
0W GH FT	**0.87 - 0.79***	**0.88 - 0.78***	**0.86 - 0.77***	0.84 - 0.79
FT GH	**0.84 - 0.74***	0.82 - 0.79	0.84 - 0.80	**0.85 - 0.78***
SD	**0.86 - 0.81***	**0.90 - 0.81***	0.85 - 0.81	**0.90 - 0.81***
0W	**0.87 - 0.72***	**0.87 - 0.79***	**0.88 - 0.79***	0.86 - 0.82
LDV	0.80 - 0.77	**0.81 - 0.74***	**0.82 - 0.72***	**0.83 - 0.76***
FLC	**0.79 - 0.62***	**0.81 - 0.70***	**0.81 - 0.69***	**0.75 - 0.65***
2W	0.86 - 0.82	**0.88 - 0.83***	0.86 - 0.82	0.87 - 0.84

Pearson's squared correlation coefficient
SPADIS - SConES(R), k = 500

	GS	GM	GI	GS-HICN
LN22	**0.84 - 0.79***	0.83 - 0.79	0.80 - 0.79	0.81 - 0.79
LN16	**0.84 - 0.78***	0.83 - 0.78	**0.83 - 0.78***	0.82 - 0.78
LN10	0.80 - 0.75	0.79 - 0.75	0.80 - 0.75	0.81 - 0.77
0W GH LN	**0.84 - 0.74***	**0.84 - 0.74***	0.85 - 0.80	0.80 - 0.82
8W GH LN	**0.80 - 0.73***	**0.79 - 0.73***	**0.79 - 0.73***	**0.78 - 0.73***
SDV	**0.87 - 0.80***	**0.86 - 0.80***	0.83 - 0.80	0.85 - 0.80
4W	0.91 - 0.90	0.92 - 0.91	0.91 - 0.90	0.91 - 0.90
8W GH FT	**0.83 - 0.79***	**0.83 - 0.79***	**0.83 - 0.79***	0.82 - 0.79
FT Field	0.82 - 0.77	**0.82 - 0.76***	0.80 - 0.76	0.79 - 0.77
FRI	0.71 - 0.68	0.73 - 0.68	0.71 - 0.69	0.68 - 0.68
0W GH FT	**0.87 - 0.79***	**0.88 - 0.80***	**0.86 - 0.80***	0.84 - 0.80
FT GH	0.84 - 0.81	0.82 - 0.81	0.84 - 0.79	0.85 - 0.81
SD	**0.86 - 0.81***	**0.90 - 0.81***	**0.85 - 0.81***	**0.90 - 0.81***
0W	**0.87 - 0.75***	**0.87 - 0.75***	**0.88 - 0.78***	0.86 - 0.81
LDV	**0.80 - 0.73***	**0.81 - 0.73***	**0.82 - 0.74***	**0.83 - 0.76***
FLC	**0.79 - 0.72***	**0.81 - 0.72***	**0.81 - 0.70***	**0.75 - 0.65***
2W	0.86 - 0.86	0.88 - 0.86	0.86 - 0.86	0.87 - 0.85

Figure 3.3: The regression performances of SPADIS, SConES(R) and SConES(S) on AT data when tight cardinality constraint is applied for $k = 500$, using the procedure described in the *Replicating Results of SConES* section. The rows denote phenotypes and columns denote networks. The numbers in each cell show Pearson's squared correlation coefficients (R^2) attained by SPADIS and SConES, respectively. The background color reflects the difference in Pearson's squared 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 (*).

3.4.3 Evaluation of SPADIS and Compared Methods

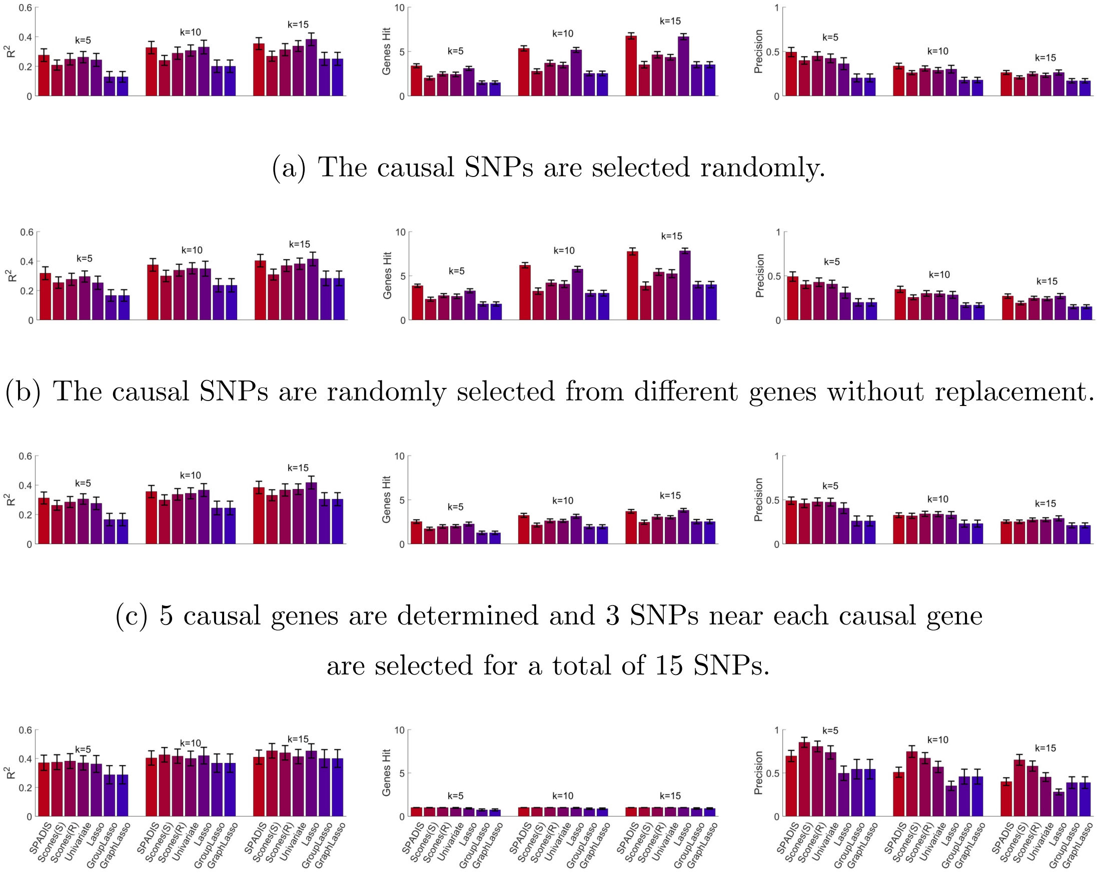
In this study, we use nested cross-validation for evaluation. The outer 10-fold cross-validation splits the data into training and test data, and the inner 10-fold cross-validation selects the parameters using the training data only. For each fold in the outer cross-validation, a separate SNP set is selected and the test data is not seen by the algorithms. Unless otherwise stated, we use this setting in our experiments.

3.5 Simulation Experiments

To assess the performance of the methods in a controlled manner, we conduct simulation experiments. We randomly choose 200 samples (out of 1307) in *AT* data. We select 500 random SNPs with MAF $> 10\%$ as follows: We first select 25 genes randomly. Then, we select 20 random SNPs near (< 20 kbp) each gene. In each experiment, we designate 15 SNPs to be causal and 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 the weight vector for each SNP, and ϵ is the error term. Both $\mathbf{w}$ and ϵ are normally distributed. We sample the weights of the causal SNPs from a standard normal distribution. We argue that in a real-life setting, there is no clear separation between causal and non-causal SNPs i.e. all SNPs play some part in explaining the phenotype at varying degrees. Hence, we sample the weights of the non-causal SNPs from a normal distribution with zero mean and 0.1 standard deviation instead of setting the standard deviations directly to zero. In our tests, we use the GS network as the SNP-SNP network.

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.

In simulation settings (a), (b) and (c), SPADIS outperforms other methods when k is less than the number of causal SNPs—see Figure 3.4. When k is equal to the number of causal SNPs, Lasso catches up to SPADIS and they outperform all other methods. In setting (d) where the assumptions of SPADIS are violated, SPADIS underperforms compared to others in terms of Precision. Regardless, its regression performance is on a par with other methods. Note that, this is the setting where methods with graph connectivity assumption should perform well (casual SNPs are close). However, this scenario is not realistic since all associated SNPs are rarely that close to each other for complex traits.



(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..

Figure 3.4: The simulation results of SPADIS, SConES(S), SConES(R), Univariate, Lasso, GroupLasso and GraphLasso for $k = 5$, $k = 10$ and $k = 15$. (Left) Pearson's squared correlation coefficient, (Middle) Number of causal genes hit, (Right) Precision calculated for causal SNPs hit. Black bars indicate the 95% confidence intervals.

Next, we check the number of causal genes hit (GenesHit) for all methods. In all simulation settings, we observe a correlation between GenesHit and R^2 (i.e. methods that perform well in GenesHit perform well in R^2 as well). We argue that high number of hit genes indicates high regression performance because when the selected SNPs fall into different genes, they are likely to contain complementary information and can explain the phenotype better. This constitutes the core idea of SPADIS.

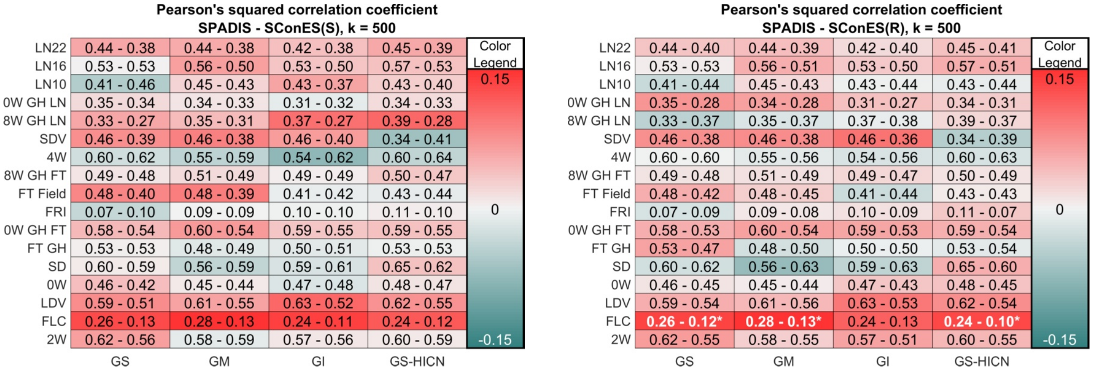
3.6 Phenotype Prediction Performance

3.6.1 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 investigate the results for $k = 500$ which we consider representative—see Figure 3.5. The results for $k = 100$, 250 and 1000 are provided in Figure B.1, Figure B.2 and Figure B.3 respectively.

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 the method described in [66]. The multiple hypothesis correction is conducted as in [62]. 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 Table C.1. 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 Figure 3.6 for each network and k value tested. Note that, the improvement 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.



Pearson's squared correlation coefficient
SPADIS - SConES(S), k = 500

	GS	GM	GI	GS-HICN
LN22	0.44 - 0.38	0.44 - 0.38	0.42 - 0.38	0.45 - 0.39
LN16	0.53 - 0.53	0.56 - 0.50	0.53 - 0.50	0.57 - 0.53
LN10	0.41 - 0.46	0.45 - 0.43	0.43 - 0.37	0.43 - 0.40
0W GH LN	0.35 - 0.34	0.34 - 0.33	0.31 - 0.32	0.34 - 0.33
8W GH LN	0.33 - 0.27	0.35 - 0.31	0.37 - 0.27	0.39 - 0.28
SDV	0.46 - 0.39	0.46 - 0.38	0.46 - 0.40	0.34 - 0.41
4W	0.60 - 0.62	0.55 - 0.59	0.54 - 0.62	0.60 - 0.64
8W GH FT	0.49 - 0.48	0.51 - 0.49	0.49 - 0.49	0.50 - 0.47
FT Field	0.48 - 0.40	0.48 - 0.39	0.41 - 0.42	0.43 - 0.44
FRI	0.07 - 0.10	0.09 - 0.09	0.10 - 0.10	0.11 - 0.10
0W GH FT	0.58 - 0.54	0.60 - 0.54	0.59 - 0.55	0.59 - 0.55
FT GH	0.53 - 0.53	0.48 - 0.49	0.50 - 0.51	0.53 - 0.53
SD	0.60 - 0.59	0.56 - 0.59	0.59 - 0.61	0.65 - 0.62
0W	0.46 - 0.42	0.45 - 0.44	0.47 - 0.48	0.48 - 0.47
LDV	0.59 - 0.51	0.61 - 0.55	0.63 - 0.52	0.62 - 0.55
FLC	0.26 - 0.13	0.28 - 0.13	0.24 - 0.11	0.24 - 0.12
2W	0.62 - 0.56	0.58 - 0.59	0.57 - 0.56	0.60 - 0.59

Pearson's squared correlation coefficient
SPADIS - SConES(R), k = 500

	GS	GM	GI	GS-HICN
LN22	0.44 - 0.40	0.44 - 0.39	0.42 - 0.40	0.45 - 0.41
LN16	0.53 - 0.53	0.56 - 0.51	0.53 - 0.50	0.57 - 0.51
LN10	0.41 - 0.44	0.45 - 0.43	0.43 - 0.44	0.43 - 0.44
0W GH LN	0.35 - 0.28	0.34 - 0.28	0.31 - 0.27	0.34 - 0.31
8W GH LN	0.33 - 0.37	0.35 - 0.37	0.37 - 0.38	0.39 - 0.37
SDV	0.46 - 0.38	0.46 - 0.38	0.46 - 0.36	0.34 - 0.39
4W	0.60 - 0.60	0.55 - 0.56	0.54 - 0.56	0.60 - 0.63
8W GH FT	0.49 - 0.48	0.51 - 0.49	0.49 - 0.47	0.50 - 0.49
FT Field	0.48 - 0.42	0.48 - 0.45	0.41 - 0.44	0.43 - 0.43
FRI	0.07 - 0.09	0.09 - 0.08	0.10 - 0.09	0.11 - 0.07
0W GH FT	0.58 - 0.53	0.60 - 0.54	0.59 - 0.53	0.59 - 0.54
FT GH	0.53 - 0.47	0.48 - 0.50	0.50 - 0.50	0.53 - 0.54
SD	0.60 - 0.62	0.56 - 0.63	0.59 - 0.63	0.65 - 0.60
0W	0.46 - 0.45	0.45 - 0.44	0.47 - 0.43	0.48 - 0.45
LDV	0.59 - 0.54	0.61 - 0.56	0.63 - 0.53	0.62 - 0.54
FLC	**0.26 - 0.12***	**0.28 - 0.13***	0.24 - 0.13	**0.24 - 0.10***
2W	0.62 - 0.55	0.58 - 0.55	0.57 - 0.51	0.60 - 0.55

Figure 3.5: 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 (*).

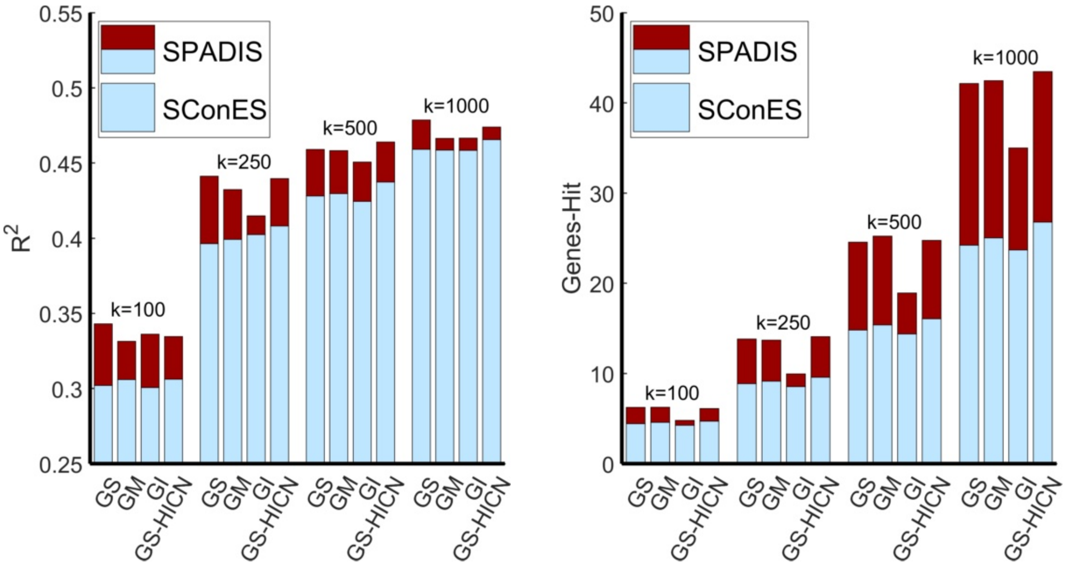


Figure 3.6: The improvement of SPADIS over SConES in terms of (left) Pearson's squared correlation coefficient and (right) number of distinct candidate genes-hit for different tight cardinality constaints k. All values shown are averages over 17 phenotypes. Blue bar indicates the maximum of SConES(S) and SConES(R) for the corresponding network and k value. The red bar indicates the amount of improvement of SPADIS over SConES.

3.6.2 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 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 1733 i.e. 1% of the number of all SNPs as done in [38]. 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 Figure B.4 and Figure B.5. When regression performances (R^2) are averaged over all phenotypes for each method, SPADIS outperforms all other methods on every network—see Figure 3.7. 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 × 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 Figure B.6, Figure B.7 and Figure B.8 for the corresponding results.

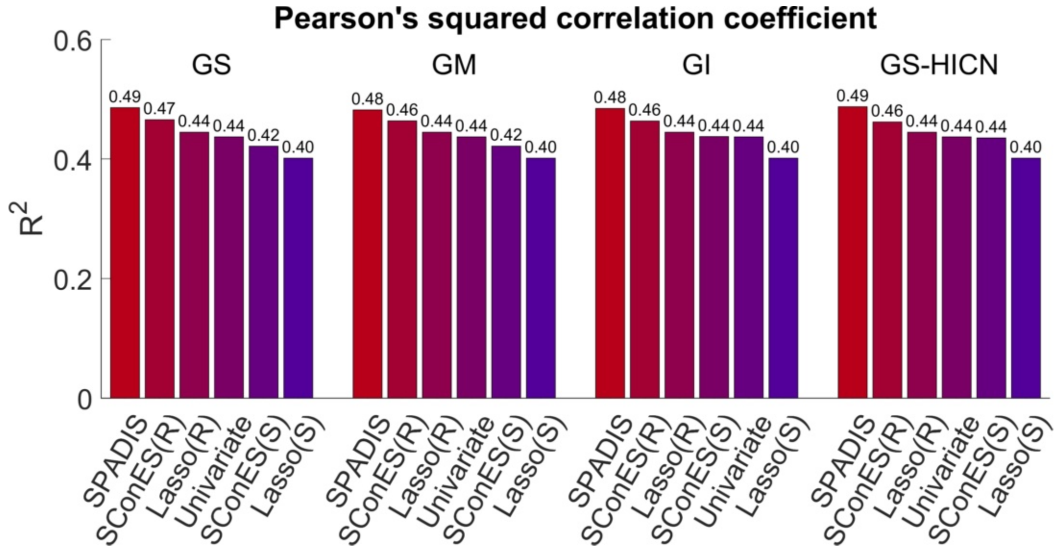


Figure 3.7: Regression performances of SPADIS, SConES(S), SConES(R), Univariate, Lasso(S) and Lasso(R) averaged over 17 AT phenotypes for maximum cardinality constraint of 1733. X-axis shows the compared methods and Y-axis shows the Pearson's squared correlation coefficient (R^2). For each network, methods are ordered in descending order of R^2.

3.7 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) Precision, 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 1000.

3.7.1 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% more distinct candidate genes compared to the next best performing method on average, over different cardinality constraints—see Table 3.1. This is an indication that SPADIS realizes one of its goals which is to spatially *cover* the network and genome.

3.7.2 Evaluation with Precision metric

We compare SPADIS and other methods with respect to the ratio of the number of selected SNPs that are near a candidate gene to the total number of selected SNPs, as done in [38]. This metric measures the *precision* of the selected SNPs, hence we denote it as such. As shown in Table 3.1, 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. Consider the following extreme case: a method that selects solely a set of SNPs near a single candidate gene can achieve a precision of 1. Hence, precision indirectly rewards the selection of SNPs that fall into a smaller number of genes. On the other hand, the diversification of 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 and the number of biological processes hit.

Table 3.1: Statistics about the genes and biological processes hit by the selected SNPs sets by SPADIS, SConES(S), SConES(R), Univariate and Lasso. 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 4 networks. The best result for each k is marked as bold.

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
Precision	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%

3.7.3 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 3.1, 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% 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.

Table 3.2: The average Pearson's squared correlation coefficient obtained for all networks and experiments that are tested. The results are averaged over all 17 phenotypes and all methods (SPADIS, SConES(S) and SConES(R)). The best result for each experiment is marked as bold.

Experiment		Network			
Constraint	k	GS	GM	GI	GS-HICN
Tight	100	0.310	0.311	0.309	**0.314**
Tight	250	0.403	0.406	0.398	**0.415**
Tight	500	0.438	0.438	0.432	**0.445**
Tight	1000	0.461	0.461	0.459	**0.467**
Maximum	1733	0.457	0.456	**0.462**	0.461
Average		0.414	0.414	0.412	**0.420**

3.8 Contribution of the Hi-C Data

We evaluate the information leveraged by using the Hi-C data via comparing the regression performances obtained when using GS-HICN compared to using other networks (GS, GM, GI). Tests are performed for all 17 phenotypes with SPADIS, SConES(S) and SConES(R). We compared the methods over five experiments: four experiments with tight cardinality constraint applied for $k = 100, 250, 500$ and 1000, and one experiment with maximum cardinality constraint applied for $k = 1733$. As shown in Table 3.2, Hi-C data provides improvements in regression performance on average: 1.4% higher than GS and GM and 1.9% higher than GI. Moreover, vement can be considered consistent since GS-HICN performs better than other networks on average in 4 out of 5 experiments. Moreover, GS-HICN hits 3.0% to 6.6% more genes and 2.7% to 21.9% more biological processes compared to other networks, on average— see Table C.2, Table C.3, Table C.4 and Table C.5 for the corresponding results. For comparisons of GS-HICN with other networks per individual phenotype in terms of regression performance—see Figure B.9 and Figure B.10.

3.9 Time Performance

We report the CPU runtime of all methods, across a range of number of SNPs (from 1000 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 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 (baseline) method—see Figure 3.8. GroupLasso and GraphLasso do not scale to SNP selection problem in GWAS. For this reason, they are not included in the experiments performed on *AT* data.

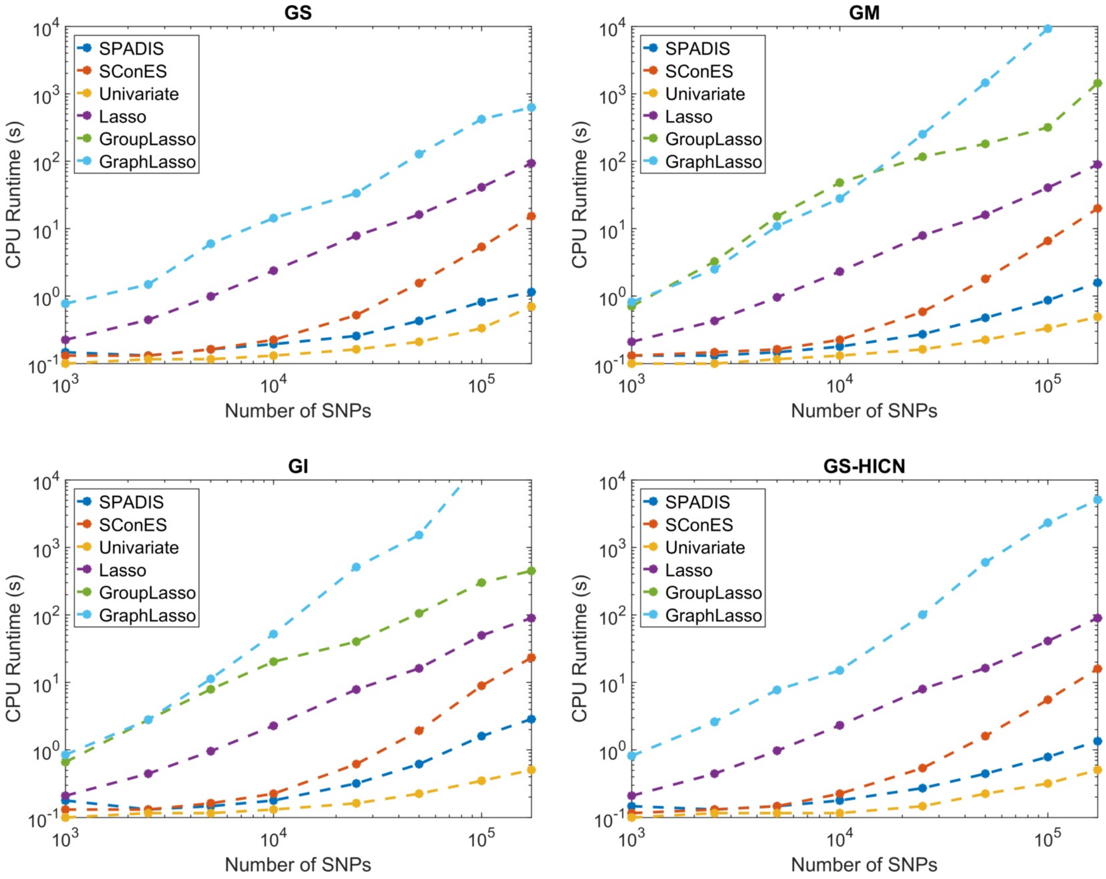


Figure 3.8: CPU time measurements of SPADIS, SConES, Univariate, Lasso, GroupLasso and GraphLasso from 1.000 to 173.219 SNPs on four networks: (Top left) GS, (Top right) GM, (Bottom left) GI and (Bottom right) GS-HICN. Note that, runtimes of GroupLasso and GraphLasso are the same for GS and GS-HICN networks by construction.

Chapter 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 thesis 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 SNP's 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 [67], 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 the first time, we investigate the utility of Hi-C data for selecting a SNP set. Our results show that Hi-C data consistently provides 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.

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 [68, 69]. However, close to 90% of the point mutations fall outside of the coding regions [41] 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 [70], we plan to apply SPADIS on autism, which sould help explain the heterogeneity in wide spectrum of phenotypes.

Bibliography

[1] P. M. Visscher *et al.*, “10 years of gwas discovery: biology, function, and translation,” *The American Journal of 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 Journal of Medicine*, vol. 360, no. 17, p. 1696, 2009.

[4] P. Kraft and D. J. Hunter, “Genetic risk prediction, are we there yet?,” *New England Journal of Medicine*, 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 Journal of Medicine*, vol. 356, no. 11, pp. 1094–1097, 2007. PMID: 17360987.

[6] J. H. Moore *et al.*, “Bioinformatics challenges for genome-wide association studies,” *Bioinformatics*, vol. 26, no. 4, pp. 445–455, 2010.

[7] H. J. Cordell, “Detecting gene–gene interactions that underlie human diseases,” *Nature Reviews 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 Reviews 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 Reviews 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 research*, vol. 11, no. 3, pp. 458–470, 2001.

[12] M. D. Ritchie *et al.*, “Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer,” *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[13] X.-Y. Lou *et al.*, “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*, vol. 80, no. 6, pp. 1125–1137, 2007.

[14] J. Lehár *et al.*, “High-order combination effects and biological robustness,” *Molecular Systems Biology*, vol. 4, no. 1, p. 215, 2008.

[15] X. Hua *et al.*, “Testing multiple gene interactions by the ordered combinatorial partitioning method in case–control studies,” *Bioinformatics*, vol. 26, no. 15, pp. 1871–1878, 2010.

[16] 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, p. e33531, 2012.

[17] J. D. Storey *et al.*, “Multiple locus linkage analysis of genomewide expression in yeast,” *PLoS biology*, vol. 3, no. 8, p. e267, 2005.

[18] D. M. Evans *et al.*, “Two-stage two-locus models in genome-wide association,” *PLoS Genetics*, vol. 2, no. 9, p. e157, 2006.

[19] N. Yosef *et al.*, “A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data,” *Bioinformatics*, vol. 23, no. 2, pp. e91–e98, 2007.

[20] V. Varadan and D. Anastassiou, “Inference of disease-related molecular logic from systems-based microarray analysis,” *PLoS computational biology*, vol. 2, no. 6, p. e68, 2006.

[21] V. Varadan *et al.*, “Computational inference of the molecular logic for synaptic connectivity in c. elegans,” *Bioinformatics*, vol. 22, no. 14, pp. e497–e506, 2006.

[22] 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.

[23] C. Herold *et al.*, “Intersnp: genome-wide interaction analysis guided by a priori information,” *Bioinformatics*, vol. 25, no. 24, pp. 3275–3281, 2009.

[24] 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, p. e1000464, 2009.

[25] R. Jiang *et al.*, “A random forest approach to the detection of epistatic interactions in case-control studies,” *BMC bioinformatics*, vol. 10, no. 1, p. S65, 2009.

[26] W. Zhang *et al.*, “A bayesian partition method for detecting pleiotropic and epistatic eqtl modules,” *PLoS computational biology*, vol. 6, no. 1, p. e1000642, 2010.

[27] Z. Wang *et al.*, “A general model for multilocus epistatic interactions in case-control studies,” *PLoS One*, vol. 5, no. 8, p. e11384, 2010.

[28] X. Wan *et al.*, “Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies,” *The American Journal of Human Genetics*, vol. 87, no. 3, pp. 325–340, 2010.

[29] X. Guo *et al.*, “Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering,” *BMC bioinformatics*, vol. 15, no. 1, p. 102, 2014.

[30] X. Ding *et al.*, "Searching high-order snp combinations for complex diseases based on energy distribution difference," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 3, pp. 695–704, 2015.

[31] M. Ayati and M. Koyutürk, "Pocos: Population covering locus sets for risk assessment in complex diseases," *PLoS computational biology*, vol. 12, no. 11, p. e1005195, 2016.

[32] S. Tuo *et al.*, "Niche harmony search algorithm for detecting complex disease associated high-order snp combinations," *Scientific Reports*, vol. 7, no. 1, p. 11529, 2017.

[33] W. Shi *et al.*, "Lasso-patternsearch algorithm with application to ophthalmology and genomic data," *Statistics and its Interface*, vol. 1, no. 1, p. 137, 2008.

[34] T. T. Wu *et al.*, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[35] S. Cho *et al.*, "Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis," *Annals of human genetics*, vol. 74, no. 5, pp. 416–428, 2010.

[36] D. Wang *et al.*, "Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso," *Journal of agricultural, biological, and environmental statistics*, vol. 16, no. 2, pp. 170–184, 2011.

[37] B. Rakitsch *et al.*, "A lasso multi-marker mixed model for association mapping with population structure correction," *Bioinformatics*, vol. 29, no. 2, pp. 206–214, 2012.

[38] C.-A. Azencott *et al.*, "Efficient network-guided multi-locus association mapping with graph cuts," *Bioinformatics*, vol. 29, no. 13, pp. i171–i179, 2013.

[39] L. Wang *et al.*, "An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies," *Bioinformatics*, vol. 27, no. 5, pp. 686–692, 2011.

[40] C. A. de Leeuw *et al.*, "Magma: generalized gene-set analysis of gwas data," *PLoS computational biology*, vol. 11, no. 4, p. e1004219, 2015.

[41] L. A. Hindorff *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[42] L. Jacob *et al.*, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th annual international conference on machine learning*, pp. 433–440, ACM, 2009.

[43] J. Huang *et al.*, "Learning with structured sparsity," *Journal of Machine Learning Research*, vol. 12, no. Nov, pp. 3371–3412, 2011.

[44] M. Sugiyama *et al.*, "Multi-task feature selection on multiple networks via maximum flows," in *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 199–207, SIAM, 2014.

[45] J. R. S. Fincham, "Genetic complementation," *Science Progress (1933-)*, pp. 165–177, 1968.

[46] A. Miele and J. Dekker, "Long-range chromosomal interactions and gene regulation," *Molecular biosystems*, vol. 4, no. 11, pp. 1046–1057, 2008.

[47] S. Wiltshire *et al.*, "Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21-25 and 10q23-26 in northern europeans," *Annals of human genetics*, vol. 70, no. 6, pp. 726–737, 2006.

[48] D. H. Geschwind, "Autism: many genes, common pathways?," *Cell*, vol. 135, no. 3, pp. 391–395, 2008.

[49] 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.

[50] W. A. Bickmore, "The spatial organization of the human genome," *Annual review of genomics and human genetics*, vol. 14, pp. 67–84, 2013.

[51] P. Martin *et al.*, "Capture hi-c reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci," *Nature communications*, vol. 6, p. 10069, 2015.

[52] R. Jäger *et al.*, "Capture hi-c identifies the chromatin interactome of colorectal cancer risk loci," *Nature communications*, vol. 6, 2015.

[53] N. L. van Berkum *et al.*, "Hi-c: a method to study the three-dimensional architecture of genomes.," *J Vis Exp*, 2010 2010.

[54] M. C. Wu *et al.*, "Rare-variant association testing for sequencing data with the sequence kernel association test," *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.

[55] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, (Arlington, Virginia, United States), pp. 324–331, AUAI Press, 2005.

[56] G. L. Nemhauser *et al.*, "An analysis of approximations for maximizing submodular set functions-i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[57] 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.

[58] 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.

[59] T. Z. Berardini *et al.*, "Functional annotation of the arabidopsis genome using controlled vocabularies," *Plant Physiology*, vol. 135, no. 2, pp. 745–755, 2004.

[60] C. Wang *et al.*, "Genome-wide analysis of local chromatin packing in arabidopsis thaliana," *Genome research*, vol. 25, no. 2, pp. 246–256, 2015.

[61] F. Ay *et al.*, "Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts," *Genome research*, vol. 24, no. 6, pp. 999–1011, 2014.

[62] D. Yekutieli and Y. Benjamini, "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics," *Journal of Statistical Planning and Inference*, vol. 82, no. 1, pp. 171–196, 1999.

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

[64] J. Liu *et al.*, "Slep: Sparse learning with efficient projections," *Arizona State University*, vol. 6, no. 491, p. 7, 2009.

[65] L. I. Kuncheva, "A stability index for feature selection.," in *Artificial intelligence and applications*, pp. 421–427, 2007.

[66] J. B. Hittner, K. May, and N. C. Silver, "A monte carlo evaluation of tests for comparing dependent correlations," *The Journal of general psychology*, vol. 130, no. 2, pp. 149–168, 2003.

[67] D. N. Reshef *et al.*, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[68] S. De Rubeis *et al.*, "Synaptic, transcriptional and chromatin genes disrupted in autism," *Nature*, vol. 515, no. 7526, pp. 209–215, 2014.

[69] I. Iossifov *et al.*, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[70] R. K. Yuen *et al.*, "Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder," *Nature Neuroscience*, vol. 20, no. 4, pp. 602–611, 2017.

Appendix A

Definitions

A.1 Single Nucleotide Polymorphism (SNP)

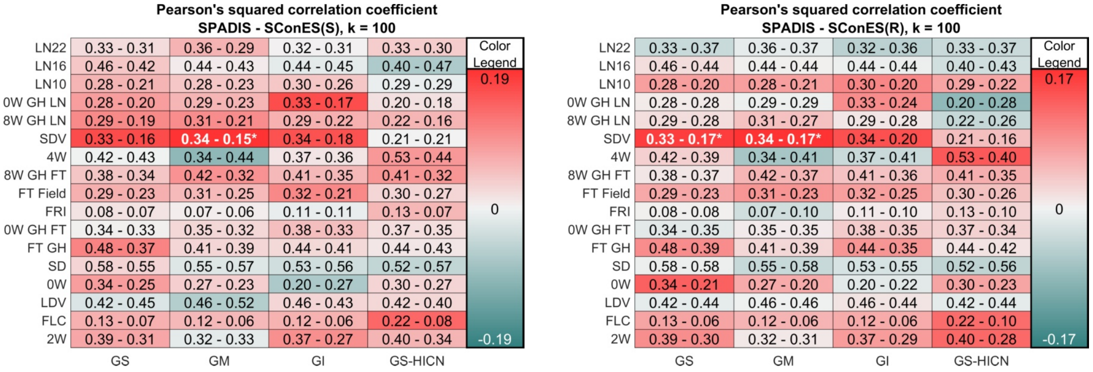
A single nucleotide polymorphism (SNP) is a DNA sequence variation that occurs when a single nucleotide (adenine, thymine, cytosine or guanine) differs among individuals. Such a variation can only be classified as a SNP if it is observed in more than 1% of a population. For example, at a specific position in the DNA sequence, cytosine (C) might be observed in most members of a population, yet in some portion of the population ($> 1\%$), the position may be occupied by thymine (T). In this case, there is a SNP at this location and the two nucleotide variations (C and T) are called alleles for this locus.

A.2 Minor Allele Frequency (MAF)

For a given population, typically two different alleles appear at a single locus and such loci are said to be biallelic. For a biallelic SNP position, the minor allele frequency (MAF) is equal to the frequency of the less frequent allele. If more than two alleles are observed at a specific locus, that position is called multiallelic. The MAF of a multiallelic position is equal to the frequency of the second most common allele observed. For example, suppose that at a specific position, there are three different alleles which are adenine, thymine and cytosine, and they appear with frequencies of 70%, 25% and 5% respectively. Then, the MAF of this position is equal to the frequency of thymine which is 25%.

Appendix B

Supplementary Figures



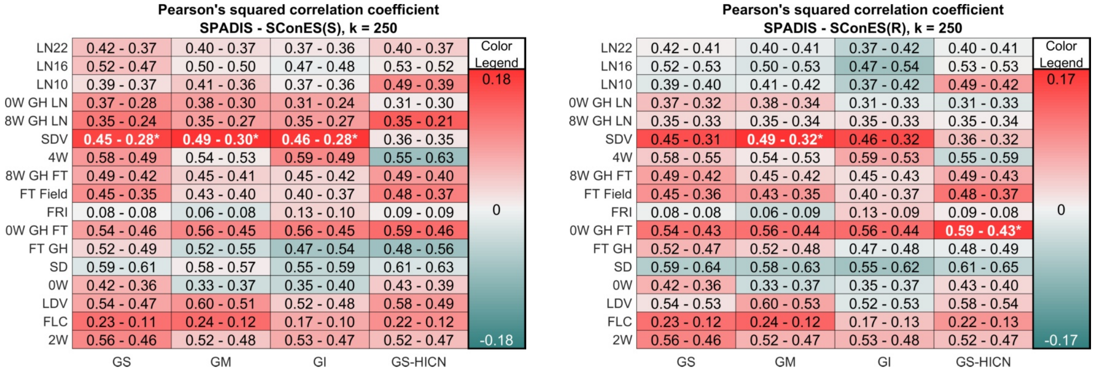
Pearson's squared correlation coefficient
SPADIS - SConES(S), k = 100

	GS	GM	GI	GS-HICN
LN22	0.33 - 0.31	0.36 - 0.29	0.32 - 0.31	0.33 - 0.30
LN16	0.46 - 0.42	0.44 - 0.43	0.44 - 0.45	0.40 - 0.47
LN10	0.28 - 0.21	0.28 - 0.23	0.30 - 0.26	0.29 - 0.29
0W GH LN	0.28 - 0.20	0.29 - 0.23	0.33 - 0.17	0.20 - 0.18
8W GH LN	0.29 - 0.19	0.31 - 0.21	0.29 - 0.22	0.22 - 0.16
SDV	0.33 - 0.16	**0.34 - 0.15***	0.34 - 0.18	0.21 - 0.21
4W	0.42 - 0.43	0.34 - 0.44	0.37 - 0.36	0.53 - 0.44
8W GH FT	0.38 - 0.34	0.42 - 0.32	0.41 - 0.35	0.41 - 0.32
FT Field	0.29 - 0.23	0.31 - 0.25	0.32 - 0.21	0.30 - 0.27
FRI	0.08 - 0.07	0.07 - 0.06	0.11 - 0.11	0.13 - 0.07
0W GH FT	0.34 - 0.33	0.35 - 0.32	0.38 - 0.33	0.37 - 0.35
FT GH	0.48 - 0.37	0.41 - 0.39	0.44 - 0.41	0.44 - 0.43
SD	0.58 - 0.55	0.55 - 0.57	0.53 - 0.56	0.52 - 0.57
0W	0.34 - 0.25	0.27 - 0.23	0.20 - 0.27	0.30 - 0.27
LDV	0.42 - 0.45	0.46 - 0.52	0.46 - 0.43	0.42 - 0.40
FLC	0.13 - 0.07	0.12 - 0.06	0.12 - 0.06	0.22 - 0.08
2W	0.39 - 0.31	0.32 - 0.33	0.37 - 0.27	0.40 - 0.34

Pearson's squared correlation coefficient
SPADIS - SConES(R), k = 100

	GS	GM	GI	GS-HICN
LN22	0.33 - 0.37	0.36 - 0.37	0.32 - 0.36	0.33 - 0.37
LN16	0.46 - 0.44	0.44 - 0.44	0.44 - 0.44	0.40 - 0.43
LN10	0.28 - 0.20	0.28 - 0.21	0.30 - 0.20	0.29 - 0.22
0W GH LN	0.28 - 0.28	0.29 - 0.29	0.33 - 0.24	0.20 - 0.28
8W GH LN	0.29 - 0.28	0.31 - 0.27	0.29 - 0.28	0.22 - 0.26
SDV	**0.33 - 0.17***	**0.34 - 0.17***	0.34 - 0.20	0.21 - 0.16
4W	0.42 - 0.39	0.34 - 0.41	0.37 - 0.41	0.53 - 0.40
8W GH FT	0.38 - 0.37	0.42 - 0.37	0.41 - 0.36	0.41 - 0.35
FT Field	0.29 - 0.23	0.31 - 0.23	0.32 - 0.25	0.30 - 0.26
FRI	0.08 - 0.08	0.07 - 0.10	0.11 - 0.10	0.13 - 0.10
0W GH FT	0.34 - 0.35	0.35 - 0.35	0.38 - 0.35	0.37 - 0.34
FT GH	0.48 - 0.39	0.41 - 0.39	0.44 - 0.35	0.44 - 0.42
SD	0.58 - 0.58	0.55 - 0.58	0.53 - 0.55	0.52 - 0.56
0W	0.34 - 0.21	0.27 - 0.20	0.20 - 0.22	0.30 - 0.23
LDV	0.42 - 0.44	0.46 - 0.46	0.46 - 0.44	0.42 - 0.44
FLC	0.13 - 0.06	0.12 - 0.06	0.12 - 0.06	0.22 - 0.10
2W	0.39 - 0.30	0.32 - 0.31	0.37 - 0.29	0.40 - 0.28

Figure B.1: The regression performance comparisons of SPADIS with SConES(S) and SConES(R) on AT data for tight cardinality constraint of $k = 100$. 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 (*).



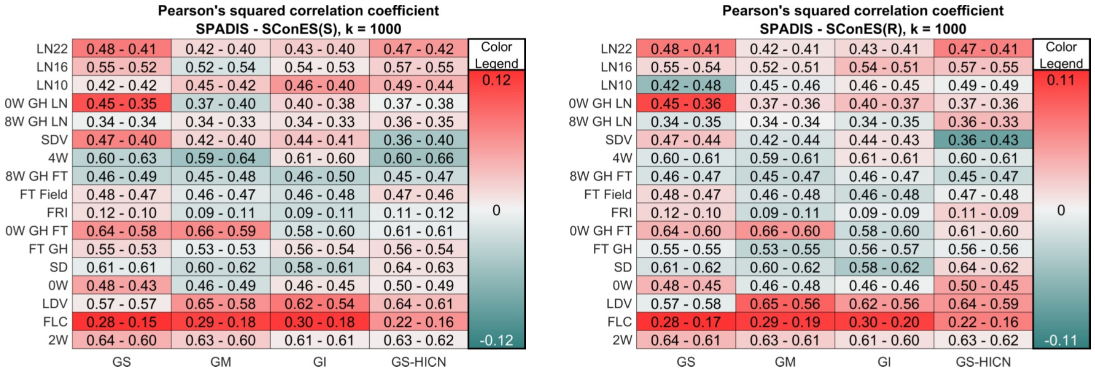
Pearson's squared correlation coefficient
SPADIS - SConES(S), k = 250

	GS	GM	GI	GS-HICN
LN22	0.42 - 0.37	0.40 - 0.37	0.37 - 0.36	0.40 - 0.37
LN16	0.52 - 0.47	0.50 - 0.50	0.47 - 0.48	0.53 - 0.52
LN10	0.39 - 0.37	0.41 - 0.36	0.37 - 0.36	0.49 - 0.39
0W GH LN	0.37 - 0.28	0.38 - 0.30	0.31 - 0.24	0.31 - 0.30
8W GH LN	0.35 - 0.24	0.35 - 0.27	0.35 - 0.27	0.35 - 0.21
SDV	**0.45 - 0.28***	**0.49 - 0.30***	**0.46 - 0.28***	0.36 - 0.35
4W	0.58 - 0.49	0.54 - 0.53	0.59 - 0.49	0.55 - 0.63
8W GH FT	0.49 - 0.42	0.45 - 0.41	0.45 - 0.42	0.49 - 0.40
FT Field	0.45 - 0.35	0.43 - 0.40	0.40 - 0.37	0.48 - 0.37
FRI	0.08 - 0.08	0.06 - 0.08	0.13 - 0.10	0.09 - 0.09
0W GH FT	0.54 - 0.46	0.56 - 0.45	0.56 - 0.45	0.59 - 0.46
FT GH	0.52 - 0.49	0.52 - 0.55	0.47 - 0.54	0.48 - 0.56
SD	0.59 - 0.61	0.58 - 0.57	0.55 - 0.59	0.61 - 0.63
0W	0.42 - 0.36	0.33 - 0.37	0.35 - 0.40	0.43 - 0.39
LDV	0.54 - 0.47	0.60 - 0.51	0.52 - 0.48	0.58 - 0.49
FLC	0.23 - 0.11	0.24 - 0.12	0.17 - 0.10	0.22 - 0.12
2W	0.56 - 0.46	0.52 - 0.48	0.53 - 0.47	0.52 - 0.47

Pearson's squared correlation coefficient
SPADIS - SConES(R), k = 250

	GS	GM	GI	GS-HICN
LN22	0.42 - 0.41	0.40 - 0.41	0.37 - 0.42	0.40 - 0.41
LN16	0.52 - 0.53	0.50 - 0.53	0.47 - 0.54	0.53 - 0.53
LN10	0.39 - 0.40	0.41 - 0.42	0.37 - 0.42	0.49 - 0.42
0W GH LN	0.37 - 0.32	0.38 - 0.34	0.31 - 0.33	0.31 - 0.33
8W GH LN	0.35 - 0.33	0.35 - 0.34	0.35 - 0.33	0.35 - 0.34
SDV	0.45 - 0.31	**0.49 - 0.32***	0.46 - 0.32	0.36 - 0.32
4W	0.58 - 0.55	0.54 - 0.53	0.59 - 0.53	0.55 - 0.59
8W GH FT	0.49 - 0.42	0.45 - 0.42	0.45 - 0.43	0.49 - 0.43
FT Field	0.45 - 0.36	0.43 - 0.35	0.40 - 0.37	0.48 - 0.37
FRI	0.08 - 0.08	0.06 - 0.09	0.13 - 0.09	0.09 - 0.08
0W GH FT	0.54 - 0.43	0.56 - 0.44	0.56 - 0.44	**0.59 - 0.43***
FT GH	0.52 - 0.47	0.52 - 0.48	0.47 - 0.48	0.48 - 0.49
SD	0.59 - 0.64	0.58 - 0.63	0.55 - 0.62	0.61 - 0.65
0W	0.42 - 0.36	0.33 - 0.37	0.35 - 0.37	0.43 - 0.40
LDV	0.54 - 0.53	0.60 - 0.53	0.52 - 0.53	0.58 - 0.54
FLC	0.23 - 0.12	0.24 - 0.12	0.17 - 0.13	0.22 - 0.13
2W	0.56 - 0.46	0.52 - 0.47	0.53 - 0.48	0.52 - 0.47

Figure B.2: The regression performance comparisons of SPADIS with SConES(S) and SConES(R) on AT data for tight cardinality constraint of $k = 250$. 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 (*).



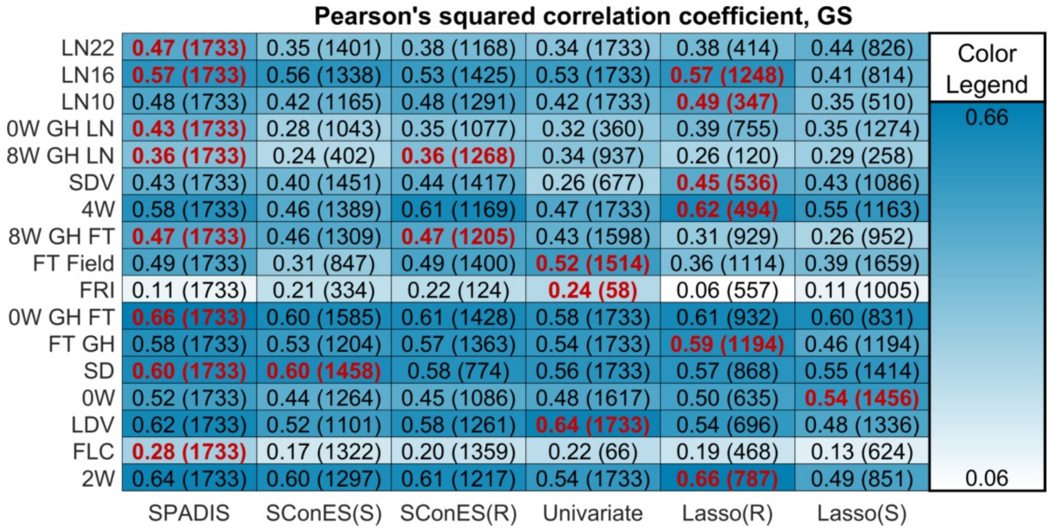
Pearson's squared correlation coefficient
SPADIS - SConES(S), k = 1000

	GS	GM	GI	GS-HICN
LN22	0.48 - 0.41	0.42 - 0.40	0.43 - 0.40	0.47 - 0.42
LN16	0.55 - 0.52	0.52 - 0.54	0.54 - 0.53	0.57 - 0.55
LN10	0.42 - 0.42	0.45 - 0.42	0.46 - 0.40	0.49 - 0.44
0W GH LN	0.45 - 0.35	0.37 - 0.40	0.40 - 0.38	0.37 - 0.38
8W GH LN	0.34 - 0.34	0.34 - 0.33	0.34 - 0.33	0.36 - 0.35
SDV	0.47 - 0.40	0.42 - 0.40	0.44 - 0.41	0.36 - 0.40
4W	0.60 - 0.63	0.59 - 0.64	0.61 - 0.60	0.60 - 0.66
8W GH FT	0.46 - 0.49	0.45 - 0.48	0.46 - 0.50	0.45 - 0.47
FT Field	0.48 - 0.47	0.46 - 0.47	0.46 - 0.48	0.47 - 0.46
FRI	0.12 - 0.10	0.09 - 0.11	0.09 - 0.11	0.11 - 0.12
0W GH FT	0.64 - 0.58	0.66 - 0.59	0.58 - 0.60	0.61 - 0.61
FT GH	0.55 - 0.53	0.53 - 0.53	0.56 - 0.54	0.56 - 0.54
SD	0.61 - 0.61	0.60 - 0.62	0.58 - 0.61	0.64 - 0.63
0W	0.48 - 0.43	0.46 - 0.49	0.46 - 0.45	0.50 - 0.49
LDV	0.57 - 0.57	0.65 - 0.58	0.62 - 0.54	0.64 - 0.61
FLC	0.28 - 0.15	0.29 - 0.18	0.30 - 0.18	0.22 - 0.16
2W	0.64 - 0.60	0.63 - 0.60	0.61 - 0.61	0.63 - 0.62

Pearson's squared correlation coefficient
SPADIS - SConES(R), k = 1000

	GS	GM	GI	GS-HICN
LN22	0.48 - 0.41	0.42 - 0.41	0.43 - 0.41	0.47 - 0.41
LN16	0.55 - 0.54	0.52 - 0.51	0.54 - 0.51	0.57 - 0.55
LN10	0.42 - 0.48	0.45 - 0.46	0.46 - 0.45	0.49 - 0.49
0W GH LN	0.45 - 0.36	0.37 - 0.36	0.40 - 0.37	0.37 - 0.36
8W GH LN	0.34 - 0.35	0.34 - 0.34	0.34 - 0.35	0.36 - 0.33
SDV	0.47 - 0.44	0.42 - 0.44	0.44 - 0.43	0.36 - 0.43
4W	0.60 - 0.61	0.59 - 0.61	0.61 - 0.61	0.60 - 0.61
8W GH FT	0.46 - 0.47	0.45 - 0.47	0.46 - 0.47	0.45 - 0.47
FT Field	0.48 - 0.47	0.46 - 0.48	0.46 - 0.48	0.47 - 0.48
FRI	0.12 - 0.10	0.09 - 0.11	0.09 - 0.09	0.11 - 0.09
0W GH FT	0.64 - 0.60	0.66 - 0.60	0.58 - 0.60	0.61 - 0.60
FT GH	0.55 - 0.55	0.53 - 0.55	0.56 - 0.57	0.56 - 0.56
SD	0.61 - 0.62	0.60 - 0.62	0.58 - 0.62	0.64 - 0.62
0W	0.48 - 0.45	0.46 - 0.48	0.46 - 0.46	0.50 - 0.45
LDV	0.57 - 0.58	0.65 - 0.56	0.62 - 0.56	0.64 - 0.59
FLC	0.28 - 0.17	0.29 - 0.19	0.30 - 0.20	0.22 - 0.16
2W	0.64 - 0.61	0.63 - 0.61	0.61 - 0.60	0.63 - 0.62

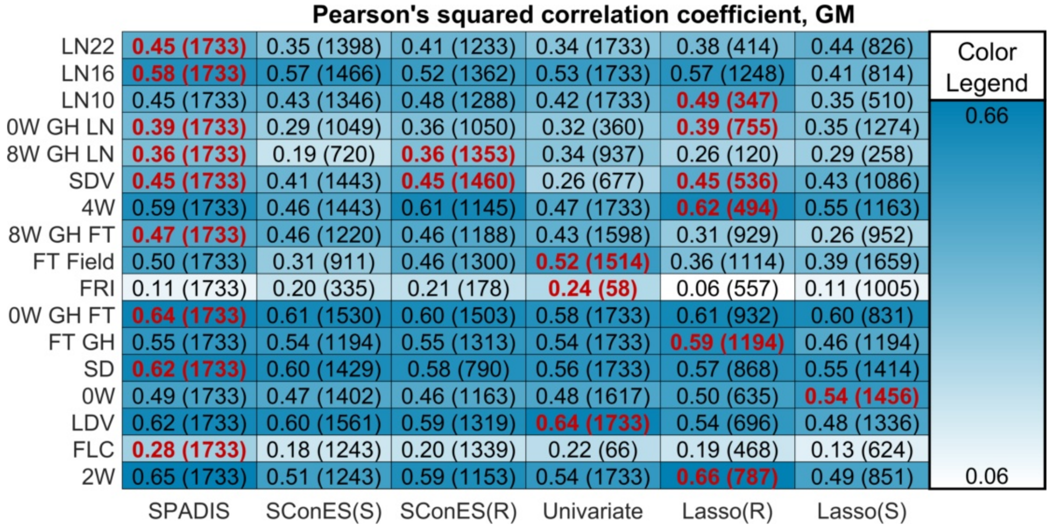
Figure B.3: The regression performance comparisons of SPADIS with SConES(S) and SConES(R) on AT data for tight cardinality constraint of $k = 1000$. The rows denote phenotypes and the columns denote networks. The numbers in each cell show Pearson's squared correlation coefficients attained by SPADIS 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 (*).



Pearson's squared correlation coefficient, GS

	SPADIS	SConES(S)	SConES(R)	Univariate	Lasso(R)	Lasso(S)
LN22	**0.47 (1733)**	0.35 (1401)	0.38 (1168)	0.34 (1733)	0.38 (414)	0.44 (826)
LN16	**0.57 (1733)**	0.56 (1338)	0.53 (1425)	0.53 (1733)	**0.57 (1248)**	0.41 (814)
LN10	0.48 (1733)	0.42 (1165)	0.48 (1291)	0.42 (1733)	**0.49 (347)**	0.35 (510)
0W GH LN	**0.43 (1733)**	0.28 (1043)	0.35 (1077)	0.32 (360)	0.39 (755)	0.35 (1274)
8W GH LN	**0.36 (1733)**	0.24 (402)	**0.36 (1268)**	0.34 (937)	0.26 (120)	0.29 (258)
SDV	0.43 (1733)	0.40 (1451)	0.44 (1417)	0.26 (677)	**0.45 (536)**	0.43 (1086)
4W	0.58 (1733)	0.46 (1389)	0.61 (1169)	0.47 (1733)	**0.62 (494)**	0.55 (1163)
8W GH FT	**0.47 (1733)**	0.46 (1309)	**0.47 (1205)**	0.43 (1598)	0.31 (929)	0.26 (952)
FT Field	0.49 (1733)	0.31 (847)	0.49 (1400)	**0.52 (1514)**	0.36 (1114)	0.39 (1659)
FRI	0.11 (1733)	0.21 (334)	0.22 (124)	**0.24 (58)**	0.06 (557)	0.11 (1005)
0W GH FT	**0.66 (1733)**	0.60 (1585)	0.61 (1428)	0.58 (1733)	0.61 (932)	0.60 (831)
FT GH	0.58 (1733)	0.53 (1204)	0.57 (1363)	0.54 (1733)	**0.59 (1194)**	0.46 (1194)
SD	**0.60 (1733)**	**0.60 (1458)**	0.58 (774)	0.56 (1733)	0.57 (868)	0.55 (1414)
0W	0.52 (1733)	0.44 (1264)	0.45 (1086)	0.48 (1617)	0.50 (635)	**0.54 (1456)**
LDV	0.62 (1733)	0.52 (1101)	0.58 (1261)	**0.64 (1733)**	0.54 (696)	0.48 (1336)
FLC	**0.28 (1733)**	0.17 (1322)	0.20 (1359)	0.22 (66)	0.19 (468)	0.13 (624)
2W	0.64 (1733)	0.60 (1297)	0.61 (1217)	0.54 (1733)	**0.66 (787)**	0.49 (851)

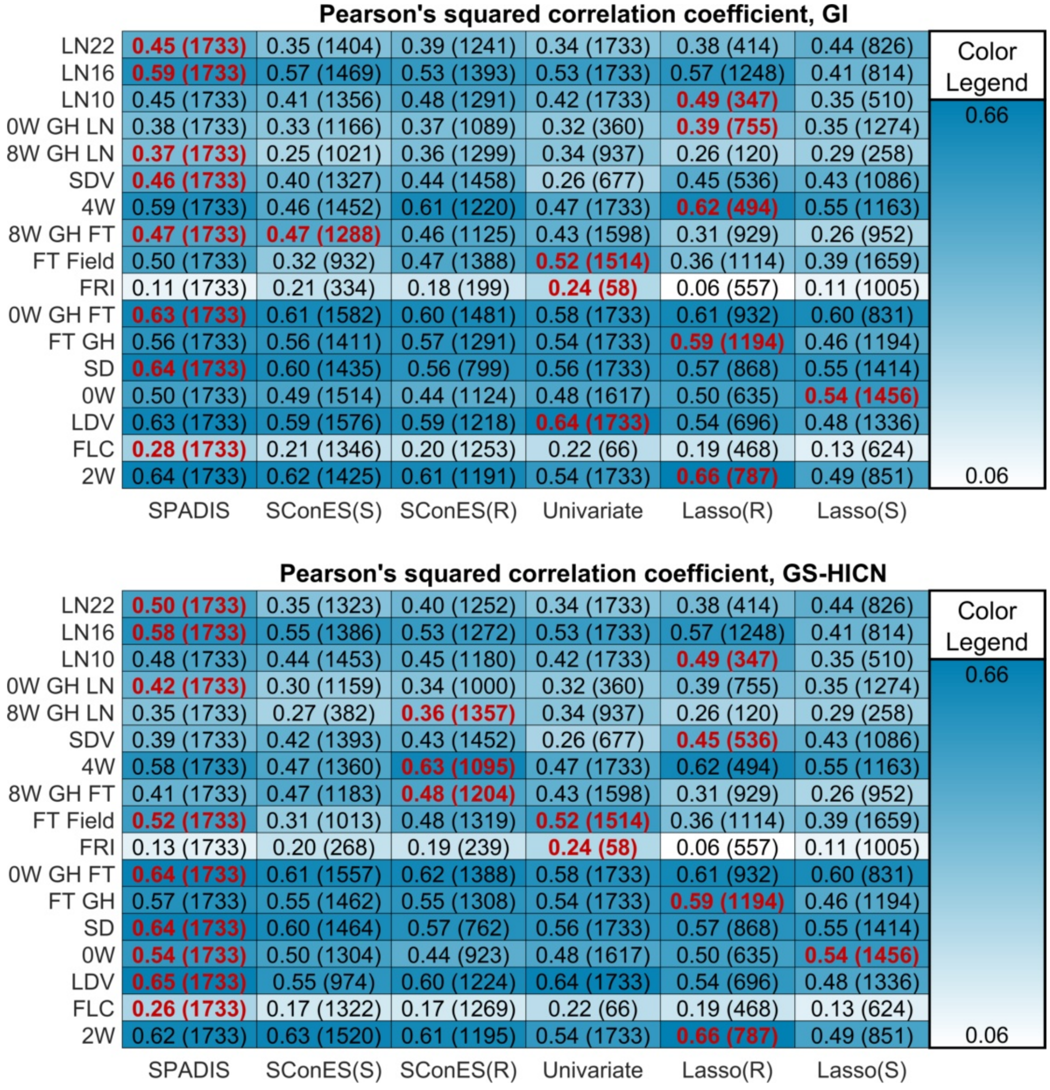
Color Legend
0.66
0.06



Pearson's squared correlation coefficient, GM

	SPADIS	SConES(S)	SConES(R)	Univariate	Lasso(R)	Lasso(S)
LN22	**0.45 (1733)**	0.35 (1398)	0.41 (1233)	0.34 (1733)	0.38 (414)	0.44 (826)
LN16	**0.58 (1733)**	0.57 (1466)	0.52 (1362)	0.53 (1733)	0.57 (1248)	0.41 (814)
LN10	0.45 (1733)	0.43 (1346)	0.48 (1288)	0.42 (1733)	**0.49 (347)**	0.35 (510)
0W GH LN	**0.39 (1733)**	0.29 (1049)	0.36 (1050)	0.32 (360)	**0.39 (755)**	0.35 (1274)
8W GH LN	**0.36 (1733)**	0.19 (720)	**0.36 (1353)**	0.34 (937)	0.26 (120)	0.29 (258)
SDV	**0.45 (1733)**	0.41 (1443)	**0.45 (1460)**	0.26 (677)	**0.45 (536)**	0.43 (1086)
4W	0.59 (1733)	0.46 (1443)	0.61 (1145)	0.47 (1733)	**0.62 (494)**	0.55 (1163)
8W GH FT	**0.47 (1733)**	0.46 (1220)	0.46 (1188)	0.43 (1598)	0.31 (929)	0.26 (952)
FT Field	0.50 (1733)	0.31 (911)	0.46 (1300)	**0.52 (1514)**	0.36 (1114)	0.39 (1659)
FRI	0.11 (1733)	0.20 (335)	0.21 (178)	**0.24 (58)**	0.06 (557)	0.11 (1005)
0W GH FT	**0.64 (1733)**	0.61 (1530)	0.60 (1503)	0.58 (1733)	0.61 (932)	0.60 (831)
FT GH	0.55 (1733)	0.54 (1194)	0.55 (1313)	0.54 (1733)	**0.59 (1194)**	0.46 (1194)
SD	**0.62 (1733)**	0.60 (1429)	0.58 (790)	0.56 (1733)	0.57 (868)	0.55 (1414)
0W	0.49 (1733)	0.47 (1402)	0.46 (1163)	0.48 (1617)	0.50 (635)	**0.54 (1456)**
LDV	0.62 (1733)	0.60 (1561)	0.59 (1319)	**0.64 (1733)**	0.54 (696)	0.48 (1336)
FLC	**0.28 (1733)**	0.18 (1243)	0.20 (1339)	0.22 (66)	0.19 (468)	0.13 (624)
2W	0.65 (1733)	0.51 (1243)	0.59 (1153)	0.54 (1733)	**0.66 (787)**	0.49 (851)

Figure B.4: The regression performances of SPADIS, SConES(S), SConES(R), Univariate, Lasso(R) and Lasso(S) when maximum cardinality constraint of 1733 is applied, for (Top) GS network and (Bottom) GM network. Rows are phenotypes and the columns are compared methods. The numbers in each cell show Pearson's squared correlation coefficients for the corresponding phenotype and method. The average cardinality of the selected SNP sets (over 10 evaluation folds) are given in parentheses. For each phenotype, the best performing method(s) are shown with bold and red font. While determining the best performing method(s), differences smaller than two significant digits are disregarded.



Color Legend
0.66
0.06

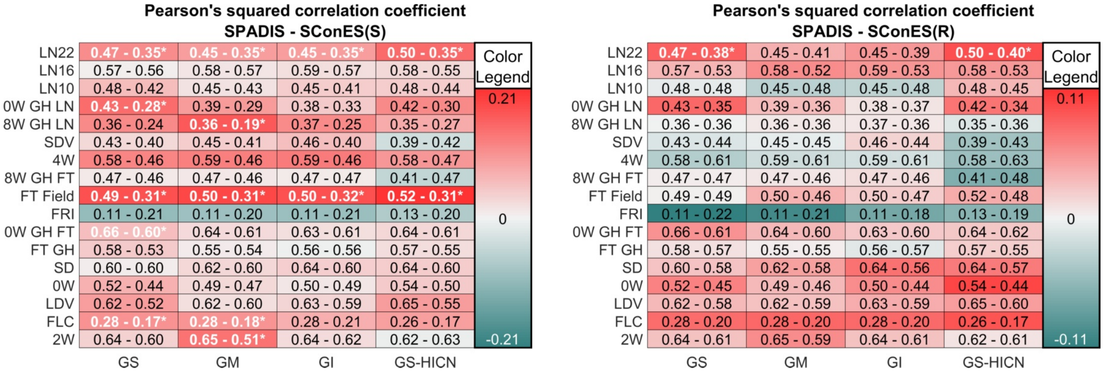
Pearson's squared correlation coefficient, GI

	SPADIS	SConES(S)	SConES(R)	Univariate	Lasso(R)	Lasso(S)
LN22	**0.45 (1733)**	0.35 (1404)	0.39 (1241)	0.34 (1733)	0.38 (414)	0.44 (826)
LN16	**0.59 (1733)**	0.57 (1469)	0.53 (1393)	0.53 (1733)	0.57 (1248)	0.41 (814)
LN10	0.45 (1733)	0.41 (1356)	0.48 (1291)	0.42 (1733)	**0.49 (347)**	0.35 (510)
0W GH LN	0.38 (1733)	0.33 (1166)	0.37 (1089)	0.32 (360)	**0.39 (755)**	0.35 (1274)
8W GH LN	**0.37 (1733)**	0.25 (1021)	0.36 (1299)	0.34 (937)	0.26 (120)	0.29 (258)
SDV	**0.46 (1733)**	0.40 (1327)	0.44 (1458)	0.26 (677)	0.45 (536)	0.43 (1086)
4W	0.59 (1733)	0.46 (1452)	0.61 (1220)	0.47 (1733)	**0.62 (494)**	0.55 (1163)
8W GH FT	**0.47 (1733)**	**0.47 (1288)**	0.46 (1125)	0.43 (1598)	0.31 (929)	0.26 (952)
FT Field	0.50 (1733)	0.32 (932)	0.47 (1388)	**0.52 (1514)**	0.36 (1114)	0.39 (1659)
FRI	0.11 (1733)	0.21 (334)	0.18 (199)	**0.24 (58)**	0.06 (557)	0.11 (1005)
0W GH FT	**0.63 (1733)**	0.61 (1582)	0.60 (1481)	0.58 (1733)	0.61 (932)	0.60 (831)
FT GH	0.56 (1733)	0.56 (1411)	0.57 (1291)	0.54 (1733)	**0.59 (1194)**	0.46 (1194)
SD	**0.64 (1733)**	0.60 (1435)	0.56 (799)	0.56 (1733)	0.57 (868)	0.55 (1414)
0W	0.50 (1733)	0.49 (1514)	0.44 (1124)	0.48 (1617)	0.50 (635)	**0.54 (1456)**
LDV	0.63 (1733)	0.59 (1576)	0.59 (1218)	**0.64 (1733)**	0.54 (696)	0.48 (1336)
FLC	**0.28 (1733)**	0.21 (1346)	0.20 (1253)	0.22 (66)	0.19 (468)	0.13 (624)
2W	0.64 (1733)	0.62 (1425)	0.61 (1191)	0.54 (1733)	**0.66 (787)**	0.49 (851)

Pearson's squared correlation coefficient, GS-HICN

	SPADIS	SConES(S)	SConES(R)	Univariate	Lasso(R)	Lasso(S)
LN22	**0.50 (1733)**	0.35 (1323)	0.40 (1252)	0.34 (1733)	0.38 (414)	0.44 (826)
LN16	**0.58 (1733)**	0.55 (1386)	0.53 (1272)	0.53 (1733)	0.57 (1248)	0.41 (814)
LN10	0.48 (1733)	0.44 (1453)	0.45 (1180)	0.42 (1733)	**0.49 (347)**	0.35 (510)
0W GH LN	**0.42 (1733)**	0.30 (1159)	0.34 (1000)	0.32 (360)	0.39 (755)	0.35 (1274)
8W GH LN	0.35 (1733)	0.27 (382)	**0.36 (1357)**	0.34 (937)	0.26 (120)	0.29 (258)
SDV	0.39 (1733)	0.42 (1393)	0.43 (1452)	0.26 (677)	**0.45 (536)**	0.43 (1086)
4W	0.58 (1733)	0.47 (1360)	**0.63 (1095)**	0.47 (1733)	0.62 (494)	0.55 (1163)
8W GH FT	0.41 (1733)	0.47 (1183)	**0.48 (1204)**	0.43 (1598)	0.31 (929)	0.26 (952)
FT Field	**0.52 (1733)**	0.31 (1013)	0.48 (1319)	**0.52 (1514)**	0.36 (1114)	0.39 (1659)
FRI	0.13 (1733)	0.20 (268)	0.19 (239)	**0.24 (58)**	0.06 (557)	0.11 (1005)
0W GH FT	**0.64 (1733)**	0.61 (1557)	0.62 (1388)	0.58 (1733)	0.61 (932)	0.60 (831)
FT GH	0.57 (1733)	0.55 (1462)	0.55 (1308)	0.54 (1733)	**0.59 (1194)**	0.46 (1194)
SD	**0.64 (1733)**	0.60 (1464)	0.57 (762)	0.56 (1733)	0.57 (868)	0.55 (1414)
0W	**0.54 (1733)**	0.50 (1304)	0.44 (923)	0.48 (1617)	0.50 (635)	**0.54 (1456)**
LDV	**0.65 (1733)**	0.55 (974)	0.60 (1224)	0.64 (1733)	0.54 (696)	0.48 (1336)
FLC	**0.26 (1733)**	0.17 (1322)	0.17 (1269)	0.22 (66)	0.19 (468)	0.13 (624)
2W	0.62 (1733)	0.63 (1520)	0.61 (1195)	0.54 (1733)	**0.66 (787)**	0.49 (851)

Figure B.5: The regression performances of SPADIS, SConES(S), SConES(R), Univariate, Lasso(R) and Lasso(S) when maximum cardinality constraint of 1733 is applied, for (Top) GI network and (Bottom) GS-HICN network. Rows are phenotypes and the columns are compared methods. The numbers in each cell show Pearson's squared correlation coefficients for the corresponding phenotype and method. The average cardinality of the selected SNP sets (over 10 evaluation folds) are given in parentheses. For each phenotype, the best performing method(s) are shown with bold and red font. While determining the best performing method(s), differences smaller than two significant digits are disregarded.



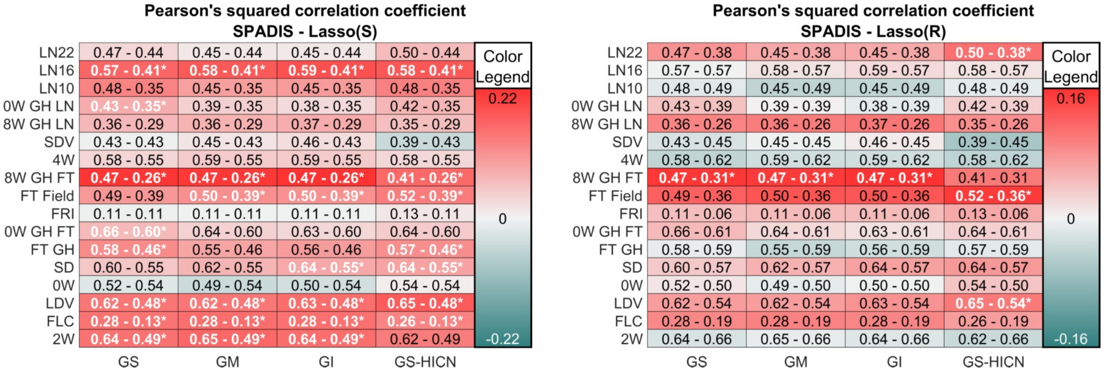
Pearson's squared correlation coefficient
SPADIS - SConES(S)

	GS	GM	GI	GS-HICN
LN22	**0.47 - 0.35***	**0.45 - 0.35***	**0.45 - 0.35***	**0.50 - 0.35***
LN16	0.57 - 0.56	0.58 - 0.57	0.59 - 0.57	0.58 - 0.55
LN10	0.48 - 0.42	0.45 - 0.43	0.45 - 0.41	0.48 - 0.44
0W GH LN	**0.43 - 0.28***	0.39 - 0.29	0.38 - 0.33	0.42 - 0.30
8W GH LN	0.36 - 0.24	**0.36 - 0.19***	0.37 - 0.25	0.35 - 0.27
SDV	0.43 - 0.40	0.45 - 0.41	0.46 - 0.40	0.39 - 0.42
4W	0.58 - 0.46	0.59 - 0.46	0.59 - 0.46	0.58 - 0.47
8W GH FT	0.47 - 0.46	0.47 - 0.46	0.47 - 0.47	0.41 - 0.47
FT Field	**0.49 - 0.31***	**0.50 - 0.31***	**0.50 - 0.32***	**0.52 - 0.31***
FRI	0.11 - 0.21	0.11 - 0.20	0.11 - 0.21	0.13 - 0.20
0W GH FT	**0.66 - 0.60***	0.64 - 0.61	0.63 - 0.61	0.64 - 0.61
FT GH	0.58 - 0.53	0.55 - 0.54	0.56 - 0.56	0.57 - 0.55
SD	0.60 - 0.60	0.62 - 0.60	0.64 - 0.60	0.64 - 0.60
0W	0.52 - 0.44	0.49 - 0.47	0.50 - 0.49	0.54 - 0.50
LDV	0.62 - 0.52	0.62 - 0.60	0.63 - 0.59	0.65 - 0.55
FLC	**0.28 - 0.17***	**0.28 - 0.18***	0.28 - 0.21	0.26 - 0.17
2W	0.64 - 0.60	**0.65 - 0.51***	0.64 - 0.62	0.62 - 0.63

Pearson's squared correlation coefficient
SPADIS - SConES(R)

	GS	GM	GI	GS-HICN
LN22	**0.47 - 0.38***	0.45 - 0.41	0.45 - 0.39	**0.50 - 0.40***
LN16	0.57 - 0.53	0.58 - 0.52	0.59 - 0.53	0.58 - 0.53
LN10	0.48 - 0.48	0.45 - 0.48	0.45 - 0.48	0.48 - 0.45
0W GH LN	0.43 - 0.35	0.39 - 0.36	0.38 - 0.37	0.42 - 0.34
8W GH LN	0.36 - 0.36	0.36 - 0.36	0.37 - 0.36	0.35 - 0.36
SDV	0.43 - 0.44	0.45 - 0.45	0.46 - 0.44	0.39 - 0.43
4W	0.58 - 0.61	0.59 - 0.61	0.59 - 0.61	0.58 - 0.63
8W GH FT	0.47 - 0.47	0.47 - 0.46	0.47 - 0.46	0.41 - 0.48
FT Field	0.49 - 0.49	0.50 - 0.46	0.50 - 0.47	0.52 - 0.48
FRI	0.11 - 0.22	0.11 - 0.21	0.11 - 0.18	0.13 - 0.19
0W GH FT	0.66 - 0.61	0.64 - 0.60	0.63 - 0.60	0.64 - 0.62
FT GH	0.58 - 0.57	0.55 - 0.55	0.56 - 0.57	0.57 - 0.55
SD	0.60 - 0.58	0.62 - 0.58	0.64 - 0.56	0.64 - 0.57
0W	0.52 - 0.45	0.49 - 0.46	0.50 - 0.44	0.54 - 0.44
LDV	0.62 - 0.58	0.62 - 0.59	0.63 - 0.59	0.65 - 0.60
FLC	0.28 - 0.20	0.28 - 0.20	0.28 - 0.20	0.26 - 0.17
2W	0.64 - 0.61	0.65 - 0.59	0.64 - 0.61	0.62 - 0.61

Figure B.6: The regression performances of SPADIS, SConES(R) and SConES(S) on AT data when maximum cardinality constraint of 1733 is applied. Rows represent phenotypes and columns represent networks. The numbers in each cell show Pearson's squared correlation coefficients achieved by SPADIS and SConES, respectively (separated with a dash). The tone of background color reflects the difference 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 (*).



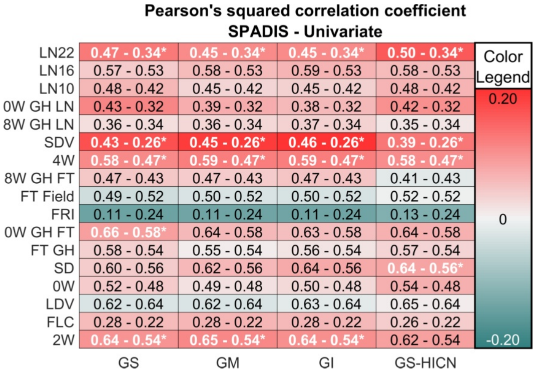
Pearson's squared correlation coefficient
SPADIS - Lasso(S)

	GS	GM	GI	GS-HICN
LN22	0.47 - 0.44	0.45 - 0.44	0.45 - 0.44	0.50 - 0.44
LN16	**0.57 - 0.41***	**0.58 - 0.41***	**0.59 - 0.41***	**0.58 - 0.41***
LN10	0.48 - 0.35	0.45 - 0.35	0.45 - 0.35	0.48 - 0.35
0W GH LN	**0.43 - 0.35***	0.39 - 0.35	0.38 - 0.35	0.42 - 0.35
8W GH LN	0.36 - 0.29	0.36 - 0.29	0.37 - 0.29	0.35 - 0.29
SDV	0.43 - 0.43	0.45 - 0.43	0.46 - 0.43	0.39 - 0.43
4W	0.58 - 0.55	0.59 - 0.55	0.59 - 0.55	0.58 - 0.55
8W GH FT	**0.47 - 0.26***	**0.47 - 0.26***	**0.47 - 0.26***	**0.41 - 0.26***
FT Field	0.49 - 0.39	**0.50 - 0.39***	**0.50 - 0.39***	**0.52 - 0.39***
FRI	0.11 - 0.11	0.11 - 0.11	0.11 - 0.11	0.13 - 0.11
0W GH FT	**0.66 - 0.60***	0.64 - 0.60	0.63 - 0.60	0.64 - 0.60
FT GH	**0.58 - 0.46***	0.55 - 0.46	0.56 - 0.46	**0.57 - 0.46***
SD	0.60 - 0.55	0.62 - 0.55	**0.64 - 0.55***	**0.64 - 0.55***
0W	0.52 - 0.54	0.49 - 0.54	0.50 - 0.54	0.54 - 0.54
LDV	**0.62 - 0.48***	**0.62 - 0.48***	**0.63 - 0.48***	**0.65 - 0.48***
FLC	**0.28 - 0.13***	**0.28 - 0.13***	**0.28 - 0.13***	**0.26 - 0.13***
2W	**0.64 - 0.49***	**0.65 - 0.49***	**0.64 - 0.49***	0.62 - 0.49

Pearson's squared correlation coefficient
SPADIS - Lasso(R)

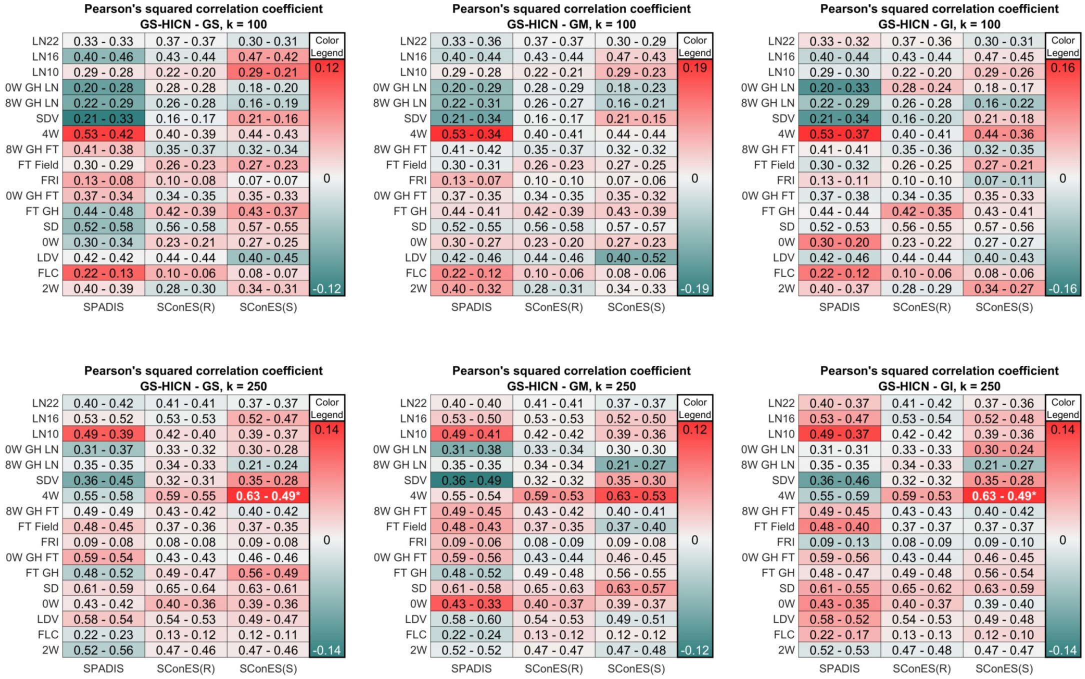
	GS	GM	GI	GS-HICN
LN22	0.47 - 0.38	0.45 - 0.38	0.45 - 0.38	**0.50 - 0.38***
LN16	0.57 - 0.57	0.58 - 0.57	0.59 - 0.57	0.58 - 0.57
LN10	0.48 - 0.49	0.45 - 0.49	0.45 - 0.49	0.48 - 0.49
0W GH LN	0.43 - 0.39	0.39 - 0.39	0.38 - 0.39	0.42 - 0.39
8W GH LN	0.36 - 0.26	0.36 - 0.26	0.37 - 0.26	0.35 - 0.26
SDV	0.43 - 0.45	0.45 - 0.45	0.46 - 0.45	0.39 - 0.45
4W	0.58 - 0.62	0.59 - 0.62	0.59 - 0.62	0.58 - 0.62
8W GH FT	**0.47 - 0.31***	**0.47 - 0.31***	**0.47 - 0.31***	0.41 - 0.31
FT Field	0.49 - 0.36	0.50 - 0.36	0.50 - 0.36	**0.52 - 0.36***
FRI	0.11 - 0.06	0.11 - 0.06	0.11 - 0.06	0.13 - 0.06
0W GH FT	0.66 - 0.61	0.64 - 0.61	0.63 - 0.61	0.64 - 0.61
FT GH	0.58 - 0.59	0.55 - 0.59	0.56 - 0.59	0.57 - 0.59
SD	0.60 - 0.57	0.62 - 0.57	0.64 - 0.57	0.64 - 0.57
0W	0.52 - 0.50	0.49 - 0.50	0.50 - 0.50	0.54 - 0.50
LDV	0.62 - 0.54	0.62 - 0.54	0.63 - 0.54	**0.65 - 0.54***
FLC	0.28 - 0.19	0.28 - 0.19	0.28 - 0.19	0.26 - 0.19
2W	0.64 - 0.66	0.65 - 0.66	0.64 - 0.66	0.62 - 0.66

Figure B.7: The regression performances of SPADIS, Lasso(R) and Lasso(S) on AT data when maximum cardinality constraint of 1733 is applied. Rows represent phenotypes and columns represent networks. The numbers in each cell show Pearson's squared correlation coefficients achieved by SPADIS and Lasso, respectively (separated with a dash). The tone of background color reflects the difference between SPADIS and Lasso. Red indicates SPADIS performs better than Lasso while blue indicates otherwise. Differences that are found to be statistically significant are shown in bold and white and marked with star (*).



	GS	GM	GI	GS-HICN
LN22	**0.47 - 0.34***	**0.45 - 0.34***	**0.45 - 0.34***	**0.50 - 0.34***
LN16	0.57 - 0.53	0.58 - 0.53	0.59 - 0.53	0.58 - 0.53
LN10	0.48 - 0.42	0.45 - 0.42	0.45 - 0.42	0.48 - 0.42
0W GH LN	0.43 - 0.32	0.39 - 0.32	0.38 - 0.32	0.42 - 0.32
8W GH LN	0.36 - 0.34	0.36 - 0.34	0.37 - 0.34	0.35 - 0.34
SDV	**0.43 - 0.26***	**0.45 - 0.26***	**0.46 - 0.26***	**0.39 - 0.26***
4W	**0.58 - 0.47***	**0.59 - 0.47***	**0.59 - 0.47***	**0.58 - 0.47***
8W GH FT	0.47 - 0.43	0.47 - 0.43	0.47 - 0.43	0.41 - 0.43
FT Field	0.49 - 0.52	0.50 - 0.52	0.50 - 0.52	0.52 - 0.52
FRI	0.11 - 0.24	0.11 - 0.24	0.11 - 0.24	0.13 - 0.24
0W GH FT	**0.66 - 0.58***	0.64 - 0.58	0.63 - 0.58	0.64 - 0.58
FT GH	0.58 - 0.54	0.55 - 0.54	0.56 - 0.54	0.57 - 0.54
SD	0.60 - 0.56	0.62 - 0.56	0.64 - 0.56	**0.64 - 0.56***
0W	0.52 - 0.48	0.49 - 0.48	0.50 - 0.48	0.54 - 0.48
LDV	0.62 - 0.64	0.62 - 0.64	0.63 - 0.64	0.65 - 0.64
FLC	0.28 - 0.22	0.28 - 0.22	0.28 - 0.22	0.26 - 0.22
2W	**0.64 - 0.54***	**0.65 - 0.54***	**0.64 - 0.54***	0.62 - 0.54

Figure B.8: The regression performances of SPADIS and Univariate on AT data when maximum cardinality constraint of 1733 is applied. Rows represent phenotypes and columns represent networks. The numbers in each cell show Pearson's squared correlation coefficients achieved by SPADIS and Univariate, respectively (separated with a dash). The tone of background color reflects the difference between SPADIS and Univariate. Red indicates SPADIS performs better than Univariate while blue indicates otherwise. Differences that are found to be statistically significant are shown in bold and white and marked with star (*).



Pearson's squared correlation coefficient
GS-HICN - GS, k = 100

	SPADIS	SConES(R)	SConES(S)
LN22	0.33 - 0.33	0.37 - 0.37	0.30 - 0.31
LN16	0.40 - 0.46	0.43 - 0.44	0.47 - 0.42
LN10	0.29 - 0.28	0.22 - 0.20	0.29 - 0.21
0W GH LN	0.20 - 0.28	0.28 - 0.28	0.18 - 0.20
8W GH LN	0.22 - 0.29	0.26 - 0.28	0.16 - 0.19
SDV	0.21 - 0.33	0.16 - 0.17	0.21 - 0.16
4W	0.53 - 0.42	0.40 - 0.39	0.44 - 0.43
8W GH FT	0.41 - 0.38	0.35 - 0.37	0.32 - 0.34
FT Field	0.30 - 0.29	0.26 - 0.23	0.27 - 0.23
FRI	0.13 - 0.08	0.10 - 0.08	0.07 - 0.07
0W GH FT	0.37 - 0.34	0.34 - 0.35	0.35 - 0.33
FT GH	0.44 - 0.48	0.42 - 0.39	0.43 - 0.37
SD	0.52 - 0.58	0.56 - 0.58	0.57 - 0.55
0W	0.30 - 0.34	0.23 - 0.21	0.27 - 0.25
LDV	0.42 - 0.42	0.44 - 0.44	0.40 - 0.45
FLC	0.22 - 0.13	0.10 - 0.06	0.08 - 0.07
2W	0.40 - 0.39	0.28 - 0.30	0.34 - 0.31

Pearson's squared correlation coefficient
GS-HICN - GM, k = 100

	SPADIS	SConES(R)	SConES(S)
LN22	0.33 - 0.36	0.37 - 0.37	0.30 - 0.29
LN16	0.40 - 0.44	0.43 - 0.44	0.47 - 0.43
LN10	0.29 - 0.28	0.22 - 0.21	0.29 - 0.23
0W GH LN	0.20 - 0.29	0.28 - 0.29	0.18 - 0.23
8W GH LN	0.22 - 0.31	0.26 - 0.27	0.16 - 0.21
SDV	0.21 - 0.34	0.16 - 0.17	0.21 - 0.15
4W	0.53 - 0.34	0.40 - 0.41	0.44 - 0.44
8W GH FT	0.41 - 0.42	0.35 - 0.37	0.32 - 0.32
FT Field	0.30 - 0.31	0.26 - 0.23	0.27 - 0.25
FRI	0.13 - 0.07	0.10 - 0.10	0.07 - 0.06
0W GH FT	0.37 - 0.35	0.34 - 0.35	0.35 - 0.32
FT GH	0.44 - 0.41	0.42 - 0.39	0.43 - 0.39
SD	0.52 - 0.55	0.56 - 0.58	0.57 - 0.57
0W	0.30 - 0.27	0.23 - 0.20	0.27 - 0.23
LDV	0.42 - 0.46	0.44 - 0.46	0.40 - 0.52
FLC	0.22 - 0.12	0.10 - 0.06	0.08 - 0.06
2W	0.40 - 0.32	0.28 - 0.31	0.34 - 0.33

Pearson's squared correlation coefficient
GS-HICN - GI, k = 100

	SPADIS	SConES(R)	SConES(S)
LN22	0.33 - 0.32	0.37 - 0.36	0.30 - 0.31
LN16	0.40 - 0.44	0.43 - 0.44	0.47 - 0.45
LN10	0.29 - 0.30	0.22 - 0.20	0.29 - 0.26
0W GH LN	0.20 - 0.33	0.28 - 0.24	0.18 - 0.17
8W GH LN	0.22 - 0.29	0.26 - 0.28	0.16 - 0.22
SDV	0.21 - 0.34	0.16 - 0.20	0.21 - 0.18
4W	0.53 - 0.37	0.40 - 0.41	0.44 - 0.36
8W GH FT	0.41 - 0.41	0.35 - 0.36	0.32 - 0.35
FT Field	0.30 - 0.32	0.26 - 0.25	0.27 - 0.21
FRI	0.13 - 0.11	0.10 - 0.10	0.07 - 0.11
0W GH FT	0.37 - 0.38	0.34 - 0.35	0.35 - 0.33
FT GH	0.44 - 0.44	0.42 - 0.35	0.43 - 0.41
SD	0.52 - 0.53	0.56 - 0.55	0.57 - 0.56
0W	0.30 - 0.20	0.23 - 0.22	0.27 - 0.27
LDV	0.42 - 0.46	0.44 - 0.44	0.40 - 0.43
FLC	0.22 - 0.12	0.10 - 0.06	0.08 - 0.06
2W	0.40 - 0.37	0.28 - 0.29	0.34 - 0.27

Pearson's squared correlation coefficient
GS-HICN - GS, k = 250

	SPADIS	SConES(R)	SConES(S)
LN22	0.40 - 0.42	0.41 - 0.41	0.37 - 0.37
LN16	0.53 - 0.52	0.53 - 0.53	0.52 - 0.47
LN10	0.49 - 0.39	0.42 - 0.40	0.39 - 0.37
0W GH LN	0.31 - 0.37	0.33 - 0.32	0.30 - 0.28
8W GH LN	0.35 - 0.35	0.34 - 0.33	0.21 - 0.24
SDV	0.36 - 0.45	0.32 - 0.31	0.35 - 0.28
4W	0.55 - 0.58	0.59 - 0.55	**0.63 - 0.49***
8W GH FT	0.49 - 0.49	0.43 - 0.42	0.40 - 0.42
FT Field	0.48 - 0.45	0.37 - 0.36	0.37 - 0.35
FRI	0.09 - 0.08	0.08 - 0.08	0.09 - 0.08
0W GH FT	0.59 - 0.54	0.43 - 0.43	0.46 - 0.46
FT GH	0.48 - 0.52	0.49 - 0.47	0.56 - 0.49
SD	0.61 - 0.59	0.65 - 0.64	0.63 - 0.61
0W	0.43 - 0.42	0.40 - 0.36	0.39 - 0.36
LDV	0.58 - 0.54	0.54 - 0.53	0.49 - 0.47
FLC	0.22 - 0.23	0.13 - 0.12	0.12 - 0.11
2W	0.52 - 0.56	0.47 - 0.46	0.47 - 0.46

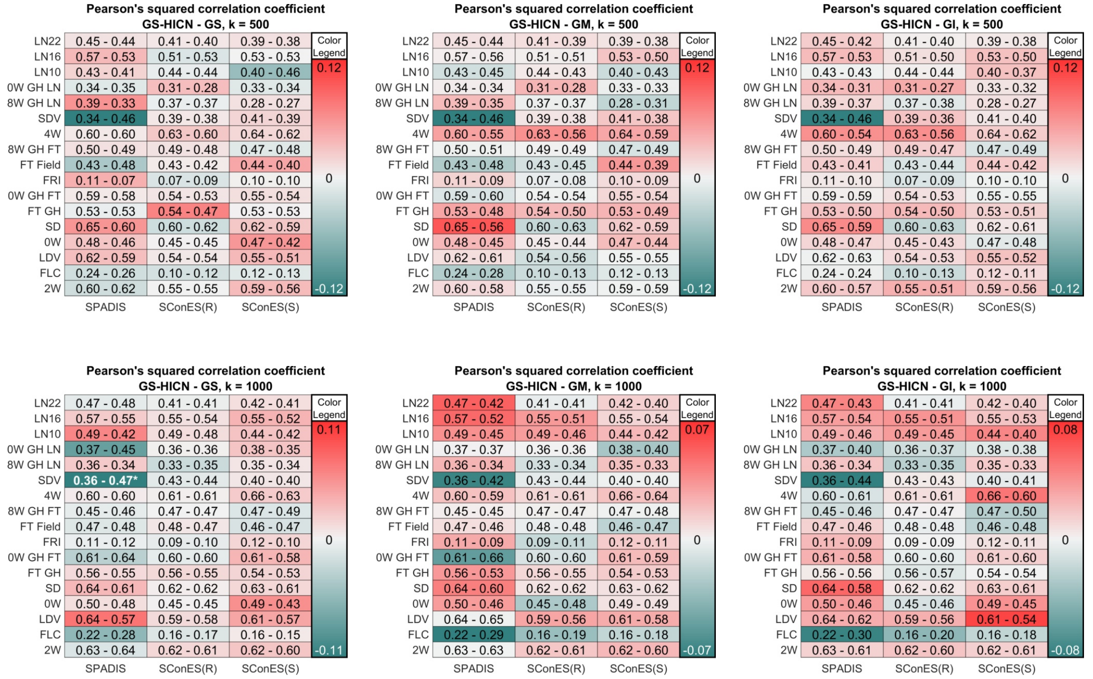
Pearson's squared correlation coefficient
GS-HICN - GM, k = 250

	SPADIS	SConES(R)	SConES(S)
LN22	0.40 - 0.40	0.41 - 0.41	0.37 - 0.37
LN16	0.53 - 0.50	0.53 - 0.53	0.52 - 0.50
LN10	0.49 - 0.41	0.42 - 0.42	0.39 - 0.36
0W GH LN	0.31 - 0.38	0.33 - 0.34	0.30 - 0.30
8W GH LN	0.35 - 0.35	0.34 - 0.34	0.21 - 0.27
SDV	0.36 - 0.49	0.32 - 0.32	0.35 - 0.30
4W	0.55 - 0.54	0.59 - 0.53	0.63 - 0.53
8W GH FT	0.49 - 0.45	0.43 - 0.42	0.40 - 0.41
FT Field	0.48 - 0.43	0.37 - 0.35	0.37 - 0.40
FRI	0.09 - 0.06	0.08 - 0.09	0.09 - 0.08
0W GH FT	0.59 - 0.56	0.43 - 0.44	0.46 - 0.45
FT GH	0.48 - 0.52	0.49 - 0.48	0.56 - 0.55
SD	0.61 - 0.58	0.65 - 0.63	0.63 - 0.57
0W	0.43 - 0.33	0.40 - 0.37	0.39 - 0.37
LDV	0.58 - 0.60	0.54 - 0.53	0.49 - 0.51
FLC	0.22 - 0.24	0.13 - 0.12	0.12 - 0.12
2W	0.52 - 0.52	0.47 - 0.47	0.47 - 0.48

Pearson's squared correlation coefficient
GS-HICN - GI, k = 250

	SPADIS	SConES(R)	SConES(S)
LN22	0.40 - 0.37	0.41 - 0.42	0.37 - 0.36
LN16	0.53 - 0.47	0.53 - 0.54	0.52 - 0.48
LN10	0.49 - 0.37	0.42 - 0.42	0.39 - 0.36
0W GH LN	0.31 - 0.31	0.33 - 0.33	0.30 - 0.24
8W GH LN	0.35 - 0.35	0.34 - 0.33	0.21 - 0.27
SDV	0.36 - 0.46	0.32 - 0.32	0.35 - 0.28
4W	0.55 - 0.59	0.59 - 0.53	**0.63 - 0.49***
8W GH FT	0.49 - 0.45	0.43 - 0.43	0.40 - 0.42
FT Field	0.48 - 0.40	0.37 - 0.37	0.37 - 0.37
FRI	0.09 - 0.13	0.08 - 0.09	0.09 - 0.10
0W GH FT	0.59 - 0.56	0.43 - 0.44	0.46 - 0.45
FT GH	0.48 - 0.47	0.49 - 0.48	0.56 - 0.54
SD	0.61 - 0.55	0.65 - 0.62	0.63 - 0.59
0W	0.43 - 0.35	0.40 - 0.37	0.39 - 0.40
LDV	0.58 - 0.52	0.54 - 0.53	0.49 - 0.48
FLC	0.22 - 0.17	0.13 - 0.13	0.12 - 0.10
2W	0.52 - 0.53	0.47 - 0.48	0.47 - 0.47

Figure B.9: The effect of using Hi-C data on the regression performances of SPADIS, SConES(R) and SConES(S) as compared to other networks, on AT data when tight cardinality constraint is applied for (Top) $k = 100$ and (Bottom) $k = 250$. Rows are the phenotypes and columns are the methods. The Pearson's squared correlation coefficient differences between GS-HICN and other networks are shown: (Left) GS-HICN vs GS, (Center) GS-HICN vs GM, and (Right) GS-HICN vs GI. The tone of background color reflects the difference in obtained Pearson's correlation coefficients between using GS-HICN or other networks. Red indicates GS-HICN performs better while blue indicates otherwise. Differences that are found to be statistically significant are shown in bold and white and marked with star (*).



Pearson's squared correlation coefficient
GS-HICN - GS, k = 500

	SPADIS	SConES(R)	SConES(S)
LN22	0.45 - 0.44	0.41 - 0.40	0.39 - 0.38
LN16	0.57 - 0.53	0.51 - 0.53	0.53 - 0.53
LN10	0.43 - 0.41	0.44 - 0.44	0.40 - 0.46
0W GH LN	0.34 - 0.35	0.31 - 0.28	0.33 - 0.34
8W GH LN	0.39 - 0.33	0.37 - 0.37	0.28 - 0.27
SDV	0.34 - 0.46	0.39 - 0.38	0.41 - 0.39
4W	0.60 - 0.60	0.63 - 0.60	0.64 - 0.62
8W GH FT	0.50 - 0.49	0.49 - 0.48	0.47 - 0.48
FT Field	0.43 - 0.48	0.43 - 0.42	0.44 - 0.40
FRI	0.11 - 0.07	0.07 - 0.09	0.10 - 0.10
0W GH FT	0.59 - 0.58	0.54 - 0.53	0.55 - 0.54
FT GH	0.53 - 0.53	0.54 - 0.47	0.53 - 0.53
SD	0.65 - 0.60	0.60 - 0.62	0.62 - 0.59
0W	0.48 - 0.46	0.45 - 0.45	0.47 - 0.42
LDV	0.62 - 0.59	0.54 - 0.54	0.55 - 0.51
FLC	0.24 - 0.26	0.10 - 0.12	0.12 - 0.13
2W	0.60 - 0.62	0.55 - 0.55	0.59 - 0.56

Pearson's squared correlation coefficient
GS-HICN - GM, k = 500

	SPADIS	SConES(R)	SConES(S)
LN22	0.45 - 0.44	0.41 - 0.39	0.39 - 0.38
LN16	0.57 - 0.56	0.51 - 0.51	0.53 - 0.50
LN10	0.43 - 0.45	0.44 - 0.43	0.40 - 0.43
0W GH LN	0.34 - 0.34	0.31 - 0.28	0.33 - 0.33
8W GH LN	0.39 - 0.35	0.37 - 0.37	0.28 - 0.31
SDV	0.34 - 0.46	0.39 - 0.38	0.41 - 0.38
4W	0.60 - 0.55	0.63 - 0.56	0.64 - 0.59
8W GH FT	0.50 - 0.51	0.49 - 0.49	0.47 - 0.49
FT Field	0.43 - 0.48	0.43 - 0.45	0.44 - 0.39
FRI	0.11 - 0.09	0.07 - 0.08	0.10 - 0.09
0W GH FT	0.59 - 0.60	0.54 - 0.54	0.55 - 0.54
FT GH	0.53 - 0.48	0.54 - 0.50	0.53 - 0.49
SD	0.65 - 0.56	0.60 - 0.63	0.62 - 0.59
0W	0.48 - 0.45	0.45 - 0.44	0.47 - 0.44
LDV	0.62 - 0.61	0.54 - 0.56	0.55 - 0.55
FLC	0.24 - 0.28	0.10 - 0.13	0.12 - 0.13
2W	0.60 - 0.58	0.55 - 0.55	0.59 - 0.59

Pearson's squared correlation coefficient
GS-HICN - GI, k = 500

	SPADIS	SConES(R)	SConES(S)
LN22	0.45 - 0.42	0.41 - 0.40	0.39 - 0.38
LN16	0.57 - 0.53	0.51 - 0.50	0.53 - 0.50
LN10	0.43 - 0.43	0.44 - 0.44	0.40 - 0.37
0W GH LN	0.34 - 0.31	0.31 - 0.27	0.33 - 0.32
8W GH LN	0.39 - 0.37	0.37 - 0.38	0.28 - 0.27
SDV	0.34 - 0.46	0.39 - 0.36	0.41 - 0.40
4W	0.60 - 0.54	0.63 - 0.56	0.64 - 0.62
8W GH FT	0.50 - 0.49	0.49 - 0.47	0.47 - 0.49
FT Field	0.43 - 0.41	0.43 - 0.44	0.44 - 0.42
FRI	0.11 - 0.10	0.07 - 0.09	0.10 - 0.10
0W GH FT	0.59 - 0.59	0.54 - 0.53	0.55 - 0.55
FT GH	0.53 - 0.50	0.54 - 0.50	0.53 - 0.51
SD	0.65 - 0.59	0.60 - 0.63	0.62 - 0.61
0W	0.48 - 0.47	0.45 - 0.43	0.47 - 0.48
LDV	0.62 - 0.63	0.54 - 0.53	0.55 - 0.52
FLC	0.24 - 0.24	0.10 - 0.13	0.12 - 0.11
2W	0.60 - 0.57	0.55 - 0.51	0.59 - 0.56

Pearson's squared correlation coefficient
GS-HICN - GS, k = 1000

	SPADIS	SConES(R)	SConES(S)
LN22	0.47 - 0.48	0.41 - 0.41	0.42 - 0.41
LN16	0.57 - 0.55	0.55 - 0.54	0.55 - 0.52
LN10	0.49 - 0.42	0.49 - 0.48	0.44 - 0.42
0W GH LN	0.37 - 0.45	0.36 - 0.36	0.38 - 0.35
8W GH LN	0.36 - 0.34	0.33 - 0.35	0.35 - 0.34
SDV	**0.36 - 0.47***	0.43 - 0.44	0.40 - 0.40
4W	0.60 - 0.60	0.61 - 0.61	0.66 - 0.63
8W GH FT	0.45 - 0.46	0.47 - 0.47	0.47 - 0.49
FT Field	0.47 - 0.48	0.48 - 0.47	0.46 - 0.47
FRI	0.11 - 0.12	0.09 - 0.10	0.12 - 0.10
0W GH FT	0.61 - 0.64	0.60 - 0.60	0.61 - 0.58
FT GH	0.56 - 0.55	0.56 - 0.55	0.54 - 0.53
SD	0.64 - 0.61	0.62 - 0.62	0.63 - 0.61
0W	0.50 - 0.48	0.45 - 0.45	0.49 - 0.43
LDV	0.64 - 0.57	0.59 - 0.58	0.61 - 0.57
FLC	0.22 - 0.28	0.16 - 0.17	0.16 - 0.15
2W	0.63 - 0.64	0.62 - 0.61	0.62 - 0.60

Pearson's squared correlation coefficient
GS-HICN - GM, k = 1000

	SPADIS	SConES(R)	SConES(S)
LN22	0.47 - 0.42	0.41 - 0.41	0.42 - 0.40
LN16	0.57 - 0.52	0.55 - 0.51	0.55 - 0.54
LN10	0.49 - 0.45	0.49 - 0.46	0.44 - 0.42
0W GH LN	0.37 - 0.37	0.36 - 0.36	0.38 - 0.40
8W GH LN	0.36 - 0.34	0.33 - 0.34	0.35 - 0.33
SDV	0.36 - 0.42	0.43 - 0.44	0.40 - 0.40
4W	0.60 - 0.59	0.61 - 0.61	0.66 - 0.64
8W GH FT	0.45 - 0.45	0.47 - 0.47	0.47 - 0.48
FT Field	0.47 - 0.46	0.48 - 0.48	0.46 - 0.47
FRI	0.11 - 0.09	0.09 - 0.11	0.12 - 0.11
0W GH FT	0.61 - 0.66	0.60 - 0.60	0.61 - 0.59
FT GH	0.56 - 0.53	0.56 - 0.55	0.54 - 0.53
SD	0.64 - 0.60	0.62 - 0.62	0.63 - 0.62
0W	0.50 - 0.46	0.45 - 0.48	0.49 - 0.49
LDV	0.64 - 0.65	0.59 - 0.56	0.61 - 0.58
FLC	0.22 - 0.29	0.16 - 0.19	0.16 - 0.18
2W	0.63 - 0.63	0.62 - 0.61	0.62 - 0.60

Pearson's squared correlation coefficient
GS-HICN - GI, k = 1000

	SPADIS	SConES(R)	SConES(S)
LN22	0.47 - 0.43	0.41 - 0.41	0.42 - 0.40
LN16	0.57 - 0.54	0.55 - 0.51	0.55 - 0.53
LN10	0.49 - 0.46	0.49 - 0.45	0.44 - 0.40
0W GH LN	0.37 - 0.40	0.36 - 0.37	0.38 - 0.38
8W GH LN	0.36 - 0.34	0.33 - 0.35	0.35 - 0.33
SDV	0.36 - 0.44	0.43 - 0.43	0.40 - 0.41
4W	0.60 - 0.61	0.61 - 0.61	0.66 - 0.60
8W GH FT	0.45 - 0.46	0.47 - 0.47	0.47 - 0.50
FT Field	0.47 - 0.46	0.48 - 0.48	0.46 - 0.48
FRI	0.11 - 0.09	0.09 - 0.09	0.12 - 0.11
0W GH FT	0.61 - 0.58	0.60 - 0.60	0.61 - 0.60
FT GH	0.56 - 0.56	0.56 - 0.57	0.54 - 0.54
SD	0.64 - 0.58	0.62 - 0.62	0.63 - 0.61
0W	0.50 - 0.46	0.45 - 0.46	0.49 - 0.45
LDV	0.64 - 0.62	0.59 - 0.56	0.61 - 0.54
FLC	0.22 - 0.30	0.16 - 0.20	0.16 - 0.18
2W	0.63 - 0.61	0.62 - 0.60	0.62 - 0.61

Figure B.10: The effect of using Hi-C data on the regression performances of SPADIS, SConES(R) and SConES(S) as compared to other networks, on AT data when tight cardinality constraint is applied for (Top) $k = 500$ and (Bottom) $k = 1000$. Rows are the phenotypes and columns are the methods. The Pearson's squared correlation coefficient differences between GS-HICN and other networks are shown: (Left) GS-HICN vs GS, (Center) GS-HICN vs GM, and (Right) GS-HICN vs GI. The tone of background color reflects the difference in obtained Pearson's correlation coefficients between using GS-HICN or other networks. Red indicates GS-HICN performs better while blue indicates otherwise. Differences that are found to be statistically significant are shown in bold and white and marked with star (*).

Appendix C

Supplementary Tables

Table C.1: Pearson's squared correlation coefficient averaged over all networks (GS, GM, GI, GS-HICN) and all tight cardinality constraints applied (k = 100, 250, 500, 1000), for SPADIS, SConES(R), SConES(S) and all 17 phenotypes. For each phenotype, the best performing method is shown with bold text.

Phenotype	SPADIS	SConES(R)	SConES(S)
2W	**0.530**	0.479	0.491
FLC	**0.223**	0.124	0.118
LDV	**0.558**	0.524	0.511
0W	**0.400**	0.373	0.388
SD	0.582	**0.612**	0.596
FT GT	**0.499**	0.483	0.497
0W GH FT	**0.534**	0.480	0.483
FRI	**0.095**	0.091	0.094
FT Field	**0.416**	0.380	0.374
8W GH FT	**0.456**	0.436	0.428
4W	0.540	0.537	**0.552**
SDV	**0.400**	0.326	0.318
8W GH LN	**0.334**	0.331	0.266
0W GH LN	**0.340**	0.313	0.295
LN10	**0.396**	0.382	0.365
LN16	**0.509**	0.502	0.496
LN17	**0.405**	0.398	0.365

Table C.2: Statistics about the genes and functionalities that are hit by the selected SNP sets of all methods that utilize a SNP-SNP network i.e. SPADIS, SConES(S) and SConES(R), on AT data when tight cardinality constraint is applied for $k = 100$. All statistics given are averaged over all 17 phenotypes. GenesHit is the number of distinct candidate flowering time genes identified. GO-Hit is the number of distinct biological processes hit by hitting an associated gene with that GO term. Precision is the ratio of number of selected SNPs near candidate genes and 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	6.2	156	7.5%	0.343
	SConES(S)	4.2	107	11.4%	0.287
	SConES(R)	4.4	114	11.2%	0.302
GM	SPADIS	6.3	153	7.3%	0.331
	SConES(S)	4.5	116	11.2%	0.297
	SConES(R)	4.6	118	11.0%	0.306
GI	SPADIS	4.8	139	6.2%	0.336
	SConES(S)	4.0	114	10.4%	0.291
	SConES(R)	4.3	116	10.4%	0.301
GS-HICN	SPADIS	6.1	157	7.2%	0.335
	SConES(S)	4.7	118	11.1%	0.302
	SConES(R)	4.7	119	10.9%	0.306

Table C.3: Statistics about the genes and functionalities that are hit by the selected SNP sets of all methods that utilize a SNP-SNP network i.e. SPADIS, SConES(S) and SConES(R), on AT data when tight cardinality constraint is applied for $k = 250$. All statistics given are averaged over all 17 phenotypes. GenesHit is the number of distinct candidate flowering time genes identified. GO-Hit is the number of distinct biological processes hit by hitting an associated gene with that GO term. Precision is the ratio of number of selected SNPs near candidate genes and 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	13.8	312	6.8%	0.441
	SConES(S)	8.3	213	9.9%	0.371
	SConES(R)	8.9	232	9.6%	0.396
GM	SPADIS	13.7	308	6.5%	0.432
	SConES(S)	9.1	237	9.6%	0.388
	SConES(R)	9.1	237	9.4%	0.399
GI	SPADIS	10.0	289	5.1%	0.415
	SConES(S)	7.8	228	8.4%	0.376
	SConES(R)	8.5	235	8.9%	0.402
GS-HICN	SPADIS	14.1	317	6.8%	0.440
	SConES(S)	9.6	240	9.8%	0.397
	SConES(R)	9.3	240	9.5%	0.408

Table C.4: Statistics about the genes and functionalities that are hit by the selected SNP sets of all methods that utilize a SNP-SNP network i.e. SPADIS, SConES(S) and SConES(R), on AT data when tight cardinality constraint is applied for $k = 500$. GenesHit is the number of distinct candidate flowering time genes identified. GO-Hit is the number of distinct biological processes hit by hitting an associated gene with that GO term. Precision is the ratio of number of selected SNPs near candidate genes and 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
	SConES(S)	13.3	347	9.0%	0.426
	SConES(R)	14.8	375	8.7%	0.428
GM	SPADIS	25.2	489	6.4%	0.458
	SConES(S)	15.3	383	8.6%	0.425
	SConES(R)	15.4	385	8.6%	0.430
GI	SPADIS	18.9	462	5.5%	0.451
	SConES(S)	12.7	373	7.0%	0.424
	SConES(R)	14.4	381	8.0%	0.422
GS-HICN	SPADIS	24.8	511	6.6%	0.464
	SConES(S)	16.1	391	8.7%	0.437
	SConES(R)	15.6	387	8.6%	0.434

Table C.5: Statistics about the genes and functionalities that are hit by the selected SNP sets by each method, on AT data when tight cardinality constraint is applied for $k = 1000$. GenesHit is the number of distinct candidate flowering time genes identified. GO-Hit is the number of distinct biological processes hit by hitting an associated gene with that GO term. Precision is the ratio of number of selected SNPs near candidate genes and 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	42.1	771	6.1%	0.479
	SConES(S)	21.8	540	8.0%	0.447
	SConES(R)	24.2	584	7.8%	0.459
GM	SPADIS	42.5	733	6.4%	0.466
	SConES(S)	25.0	596	7.9%	0.459
	SConES(R)	25.0	601	7.7%	0.459
GI	SPADIS	35.0	697	6.3%	0.467
	SConES(S)	20.8	580	6.1%	0.452
	SConES(R)	23.7	596	7.1%	0.458
GS-HICN	SPADIS	43.5	789	6.5%	0.474
	SConES(S)	26.8	609	7.9%	0.466
	SConES(R)	25.9	607	7.8%	0.460