DIVERSE SNP SELECTION FOR EPISTASIS TEST PRIORITIZATION

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
Gizem Çaylak
August 2019

Diverse SNP Selection for Epistasis Test Prioritization
By Gizem Çaylak
August 2019

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)

Can Alkan

R. Gökberk Cinbiş

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

DIVERSE SNP SELECTION FOR EPISTASIS TEST PRIORITIZATION

Gizem Çaylak
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
August 2019

Genome-wide association studies explain a fraction of the underlying heritability of genetic diseases. Epistatic interactions between two or more loci help closing the gap and identifying those complex interactions provides a promising road to a better understanding of complex traits. Unfortunately, sheer number of loci combinations to consider and hypotheses to test prohibit the process both computationally and statistically. This is true even if only pairs of loci are considered. Epistasis prioritization algorithms have proven useful for reducing the computational burden and limiting the number of tests to perform. While current methods aim at avoiding linkage disequilibrium and covering the case cohort, none aims at diversifying the topological layout of the selected SNPs which can detect complementary variants. In this thesis, a two stage pipeline to prioritize epistasis test is proposed. In the first step, a submodular set function is optimized to select a diverse set of SNPs that span the underlying genome to (i) avoid linkage disequilibrium and (ii) pair SNPs that relate to complementary function. In the second step, selected SNPs are used as seeds to a fast epistasis detection algorithm. The algorithm is compared with the state-of-the-art method LinDen on three datasets retrieved from Wellcome Trust Case Control Consortium: type two diabates, hypertension and bipolar disorder. The results show that the pipeline drastically reduces the number of tests to perform while the number of statistically significant epistatic pairs discovered increases.

Keywords: GWAS, Epistasis Test Prioritization, SNP Selection.

ÖZET

EPİSTATİK TEST ÖNCELİKLENDİRME İÇİN ÇEŞİTLİ SNP SEÇİLİMİ

Gizem Çaylak
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Ağustos 2019

Genom çapında ilişkilendirme çalışmaları (Genome-Wide Association Studies - GWAS) genetik hastalıkların temelini teşkil eden kalıtsallığın altında yatan sebeplerin sadece bir kısmını açıklayabilmektedir. İki ya da daha fazla lokusun arasındaki epistatik etkileşimler açıklama gücündeki boşluğu kapatmaya yardımcı olduğu gibi kompleks etkileşimleri de tespit ederek kompleks karakterlerin daha iyi çözümlenebilmesi için gelecek vaat etmektedir. Fakat değerlendirilmesi ve hipotez için test edilmesi gereken çok sayıdaki lokus kombinasyonları, hem algoritma karmaşıklığı hem de istatiksel olarak çalışmaları engellemektedir. Sadece ikili etkileşimler göz önüne alındığında dahi bu durum düzelmemektedir. Epistasis önceliklendirme algoritmalarının hem hesaplama yükünü hem de yapılması gereken test sayısını azalttığı kanıtlanmıştır. Güncel metotlar bağlantı dengesizliğinden kaçınmayı ve vaka kohortunu kapsamayı amaçlasa da, metotların hiçbiri seçilen lokusların topolojik düzenini çeşitlendirmeyi amaçlamamıştır. Bu tezde, epistatik testleri önceliklendirmek için iki aşamalı ardışık düzen algoritması önerilmiştir. İlk aşamada çeşitli lokusları seçmek için altmodüler bir fonksiyon optimize edilmiştir. Bu aşama (i) bağlantı dengesizliğinden kaçınmayı ve (ii) birbirini fonksiyonel olarak tamamlayan lokus ikilileri seçmeyi amaçlamaktadır. İkinci aşamada, seçilen lokuslar hızlı epistatik etkileşim tespit eden bir algoritmada girdi olarak kullanılmıştır. Deneylerimizde, metot modern yöntemlerden biri olan LINDEN ile Wellcome Trust Case Control Consortium'dan alınan tip 2 diyabet, hipertansiyon, bipolar bozukluk olmak üzere üç veriseti üzerinde karşılaştırılmıştır. Sonuçlar göstermektedir ki epistatik çiftleri bulmak için yapılan testlerin sayısında önemli bir düşüş gözlenirken aynı zamanda keşfedilen istatiksel olarak önemli epistatik çift sayısı da artmıştır.

Anahtar sözcükler: GWAS, Epistatik Test Önceliklendirme, SNP seçimi.

Acknowledgement

First and foremost, I would like to thank my advisor Asst. Prof. A. Ercüment Çiçek for his support, help and patience throughout my master's studies. His guidance on this work, not only made completing this thesis possible, but helped me to become a good researcher. I will be forever grateful for his extensive support.

I would also like to thank my jury members Asst. Prof. Can Alkan and Asst. Prof. R. Gökberk Cinbiş for reading my thesis and kindly accepting to be in my thesis committee. I thank TUBITAK for supporting this research via Career Grant 116E148 awarded to A. Ercument Cicek.

I would like to thank my friends and collegues in Bilkent. Firstly, I would like to thank Alper, Cihan and Onur for their valuable friendships and support. Then, I would like to thank Çağlar, Furkan, Miray, and Ömer for all the good memories we had, especially during the coffee breaks. Their constructive feedbacks, perspectives on academia, life and many other matters were nonesuch.

Additionally, I would like to express my gratitude to Esma, Musa, and Taha. Their support during my academic life was of great value. I consider myself a person of great luck to call these people as my friends.

Finally and uttermost, I would like to express my gratitude to my mother, sister and brother for their continuous moral support, unparalleled love and belief; I could not have succeed this without them. They have presented me opportunities for my education which I cannot repay all along.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genome-wide association studies (GWAS) have been leading the susceptibility gene discovery for numerous genetic diseases. Analyzing single loci associations has provided many valuable insights but they alone do not account for the full heritability [1]. Instead of analyzing single loci associations, investigating the interplay among multiple loci with respect to a phenotype has helped to bridge the missing heritability gap. Such interactions between two or more loci is called epistasis and it has been shown to contribute to complex genetic traits such as cancer [2]. Given a million variants in a genome, a trillion tests are required to process all single nucleotide polymorphism (SNP) pairs. This number exponentially increases as the order of the interactions increase. Thus, this procedure is not only computationally prohibitive, but also lacks statistical power due to multiple hypothesis testing issue. Also, due to associations between loci in a population, which is referred as Linkage Disequilibrium (LD), many of these statistical tests are redundant [3].

Many methods have been developed to circumvent these problems. TEAM and BOOST are exhaustive models which exploit data structures and efficient data representations to improve up on the brute force performance [4, 5]. However,

these methods still perform many tests and do not scale for large tasks. For instance, BOOST reports a runtime of 60h for 360k SNPs. Another approach is to reduce the search space. Several methods aim to filter pairs based on statistical thresholds: SNPHarvester [6], SNPRuler [7] and IBBFS [8]. However, as noted in Piriyapongsa et al. [9], these methods mostly do not follow a biological reasoning and tend to detect interactions that are in close proximity. Thus, even though detected pairs are statistically meaningful, they might not be so biologically, as they are in linkage disequilibrium. Incorporating biological background for pruning the search space by testing the SNP pairs that are functionally associated has also proven useful [10, 11, 12, 13, 14]. However, this approach requires most SNPs to be discarded as many are quite far away from any gene to be associated. Moreover, this introduces a literature bias in the selection of the algorithms. A popular approach is to prioritize the tests to be performed rather than discarding pairs from the search space and control for type I error. In this approach, the user can keep performing tests, in the order specified by the algorithm, until a desired number of significant pairs are found. While false negatives may arise, the idea is to provide the user a tractable number of true positives with minimum number of tests performed to ensure statistical power. The first algorithm in this kind is iLOCi [9], which ranks the SNP pairs by performing a dependence test. It avoids pairs that are unrelated to disease but might be related to linkage disequilibrium (LD). This work was followed by Ayati and Koyutürk [15] who proposed testing pairs of SNPs in population covering locus sets - PoCos. Algorithm first greedily selects multiple exclusive groups of SNPs that cover all affected individuals. Epistasis tests then are performed across PoCos with the hope that independent coverage of the cases will lead different PoCos to include complementary SNPs and thus, will lead to the test of epistatic pairs [15, 16]. Finally, Cowman and Koyutürk , introduced the state-of-the-art LINDEN algorithm [17]. The method first generates trees that represent genomic regions (LD forest). Then, by comparing the roots of these trees, it decides if this pair of genomic regions is a promising candidate for epistasis test. Trees are parsed in depth first manner and leaf pairs are tested only if the estimation at higher levels provides a value larger than a threshold. All three methods aim at avoiding testing pairs that are in LD.

The fundamental problem in all of the above-mentioned algorithms is the high number of false positives (FP) (i.e., LinDen's false discovery rate (FDR) is ∼0.99, 5 TP, ∼1800 FP for type 2 diabetes). FPs are SNPs that are being tested and not crossing the Bonferroni-corrected significance threshold. In an orthogonal study, Yilmaz et al. avoids LD in a different manner and for phenotype prediction problem [18]. They show that while looking for a small set of loci (i.e., 1000) that is the most predictive of a continuous phenotype in GWA study, selecting SNPs further away from each other results in better predictive power. This method (SPADIS) is designed for feature selection for multiple regression and as the SNP set it generates contains diverse and complementary SNPs, it results in better R^2 values.

In this study, we conjecture that we can minimize the number of FPs by guiding the prioritization algorithms using SPADIS. The hypothesis is that the set of SNPs selected by SPADIS are likely to be epistatic, since the algorithm is designed to diversify the set and select complementary SNPs. We created a pipeline that first uses SPADIS to generate its candidate set for epistasis test. Instead of using this set for all-pairs epistasis testing which would still return a large number of FPs, we use it to guide the state of the art example of these algorithms: LinDen. We let it only form LD trees over SPADIS-selected regions (selected SNPs and a small number of neighbors) to pick likely epistatic pairs from this set. Thus, LinDen does not have perform still a sheer number of tests that cover the genome but a pruned search space of likely epistatic pairs. We also hypothesize that integrating regulatory regions such as enhancers which can affect the gene expression level into the pipeline will improve the algorithm since disruptions on these regions have ties to genetic diseases and disorders.

We compare our algorithm with the state-of-the-art LinDen on Wellcome Trust Case Control Consortium datasets, type 2 diabetes (T2D), hypertension (HT) and bipolar disorder (BD) datasets, and show that we drastically reduce the number of tests to perform to discover epistatic pairs (i.e. down to 36% for T2D). Moreover, we improved the precision substantially, i.e. from 0.3% , 0.3% and 0.4% to 42%

, 34% and 29% for T2D, BD and HT datasets respectively. When we integrate regulatory regions to the pipeline precision values improved up to 48% , 82% and 32% for T2D, BD and HT datasets respectively. Moreover, the total runtime of the pipeline is only one forth of the LinDen-only run (15min vs 1+ hour).

The rest of the thesis is organized as follows. In Chapter 2, terminologies such as epistasis and linkage disequilibrium that are used throughout this study are introduced and illustrated. Chapter 3 defines the notation and describe the related methods from the literature in detail. In Chapter 4, the problem of epistasis test prioritization is described and the methods we proposed as the solution to the problem are described. In Chapter 5, we describe the datasets on which we used to evaluate our algorithm as well as interaction network we use. Then, we describe experimental framework that we used to evaluate our algorithm and present the results of the algorithm. Finally, in Chapter 6 we discuss the results of the algorithm and draw our conclusions.

Chapter 2

Background Information

2.1 Epistasis

The effect of a given genomic variant on a specific phenotype can be dependent on other genomic variants. This idea first presented by Bateson [19] as epistasis to explain the segregation distortions in Mendelian inheritance model when allele frequency of a locus deviates from the expected Mendelian segregation ratio. Statistical interpretation of this idea, a latter definition, came from Fisher [20], a statistician and geneticist, as statistical deviation from the additivity of two loci in terms of their synergistic effects on a given phenotype. In the later years, besides additive model two more models commonly used to define epistasis: multiplicative and heterogeneity model [21].

In this thesis, we will use the latter, statistical and additive, definition of the epistasis since in general GWA studies use statistical approaches to detect epistatic interactions. Statistical significance may not imply biologically interesting finding; however, it is expected that this statistical definition helps inferring biological pathways corroborating disease.

2.2 Epistasis Test

In the explanation of complex diseases, the role of epistatic interactions is major [22, 23]. Thus, several tools and algorithms have been developed to discover epistatic interactions ranging from exhaustive methods to Bayesian models to various machine learning techniques [5, 24, 25, 26, 27, 28, 29, 30]. However, discovering those epistatic interactions is challenging in terms of validating identified statistical interactions biologically, and finding the balance between model complexity and computational efficiency. Even to detect pairwise interactions between a million SNPs, in total $5x10^{11}$ statistical tests are needed to be performed. Commonly epistasis tests base on linear statistical models in which interactions are described between predictor variables and an outcome variable [21]. In linear models in which the outcome is quantitative, we can define a function y of a predictor variable x such that $y = \beta x + \beta_0$. Here, β_0 corresponds to the intercept term and β corresponds to the slope of the best fit line. We want to find β and β_0 such that our observations, as data points (x, y), fit the line $y = \beta x + \beta_0$ with least error, average distance of observed values from the line. To consider the interactions between variants, we could use a multiple regression model in which the formula is extended to $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_0$. Here, to assess the interaction between x_1 and x_2, we constrain $x_3 = x_1 x_2$ and use β_3 as the interaction coefficient term. In the case where outcome variable is binary, e.g. sample has the disease or not, we convert our model to a logistic regression model via log scaling so instead of a quantitative outcome y, $\log(\frac{p}{1-p})$, where p is the probability of having the phenotype of interest, is used to model the problem. For example, assume we have a logistic model such that

$$\log(\frac{p}{1-p}) = \beta_1 A + \beta_2 B + \beta_3 AB + \beta_0 \tag{2.1}$$

where A and B correspond to binary values representing existence of genetic variation at locus A and B respectively, β_1 and β_2 represents the main effects of variances at A and B, and the coefficient β_3 represents the interaction term [31]. In essence, epistasis test corresponds to checking whether the interaction term is zero or not in the equation 2.1. In the example, 1 degrees of freedom (df) test of $\beta_3 = 0$ would be equal to interaction test. If a model in which only the intercept

term β_0 is included is compared to a model in which the interactions, and effects of A and B are included, then this would correspond to a 3df test if allellic coding is used and 8df test in the case of a saturated model [5].

2.3 Linkage Disequilibrium

Linkage Disequilibrium (LD) arises from the deterministic association of the genotypes at different loci and occurs between topologically close loci as the DNA is inherited in chunks rather than nucleotide by nucleotide [32]. As depicted in Figure 2.1, if locus 1 has alleles C or T and locus 2 has alleles A or G, and if these two loci are in LD, then one is more likely to have allele T at locus 1 when locus 2 has allele A. There exists a statistical association between two loci which means we may infer the genotype at one locus by observing the genotype at another location.

Figure 2.1: This example demonstrates the linkage disequilibrium between two loci, locus 1 and locus 2, each with two alleles. If locus 1 can have alleles C or T and locus 2 can have alleles A or G and if these loci are in LD then a sample is more likely to have allele T at locus 1 when locus 2 has allele A. We may infer that there exists a statistical association between the two loci.

Mathematically, the degree of LD between the alleles A and B is formulated as:

$$D(AB) = P(AB) - P(A)P(B)$$

where $P(A)$ is the rate of allele A at the first locus, $P(B)$ is the rate of allele B at the second locus, and $P(AB)$ is the joint frequency of genotype AB. In the

case of absolute independence between the two alleles, D becomes zero. Then these two alleles are said to be in linkage equilibrium. In the cases where $D \neq 0$, two loci are in LD. In the main literature, there are two derived version of this LD measure D to make it more robust and convenient: normalizing D [33] and defining a correlation coefficient by scaling D [34].

These correlations both offer challenges and opportunities in the detection of epistasis. When loci are in LD their association may appear statistically significant due to their physical link on a chromosome [35]. But if these strong correlated loci are grouped together, it would be sufficient to only use one representative locus from each group for testing epistasis. Since in general extend of LD is limited with a few thousand base pair around common loci, LD groups can be created from the neighboring loci [32].

2.4 Hardy-Weinberg Equilibrium

Hardy-Weinberg equilibrium (HWE) or principle defines a mathematical model between allele frequencies and genotypes. This model predicts the change in gene and allele frequencies from one generation to the next under a set of assumptions such as no natural selection, no mutation, no immigration [36]. A population which satisfies these assumptions is considered in HWE. The deviation from HWE is measured using a statistical test, either goodness-of-fit or χ^2 test [37]. For example, assume a locus with two possible alleles A and T with given frequencies $f_A = 0.8$, $f_T = 0.2$ respectively. Then, there are three possible genotypes: AA, TT and AT. Consider a large gene pool in which the eggs and sperm enter randomly. We calculate the expected frequency of having an individual with an AA genotype as the probability of a sperm with allele A fertilizes an egg with allele A, which is $0.8 \times 0.8 = 0.64$. The expected frequency of having an individual with a TT genotype is calculated in a similar way ($0.2 \times 0.2 = 0.04$). To calculate the frequency of having an individual with a genotype AT we consider two cases: a sperm with allele A fertilizes an egg with allele T ($0.8 \times 0.2 = 0.16$) and a sperm

with allele T fertilizes an egg with allele A ($0.2 \times 0.8 = 0.16$). Hence the expected frequency is $0.2 \times 0.8 \times 2 = 0.32$. Then, we can predict the genotypes of the next generation, i.e a population with 100 individuals, as given in Table 2.1.

Table 2.1: Calculation of expected genotype frequencies for the next generation on a locus with two possible alleles A and T with given frequencies $f_A = 0.8$, $f_T = 0.2$ respectively. # Alleles indicates the allele counts for the next generation, a population with 100 individuals.

Allele	**Frequency**	**Genotype**	**Expected Frequency**	**# Alleles (the next generation)**
A	$f_A = 0.8$	AA	0.64	64
		AT	0.32	32
T	$f_T = 0.2$			
		TT	0.04	4

In association studies, deviations from HWE are assumed to be an indicator of genotyping errors. Thus, deviations from HWE should be considered as a part of quality control on genotype data to filter out of SNPs. In this thesis, we perform a χ^2 HWE test on SNPs and exclude SNPs that are above a HWE p-value threshold 1e$-$6 using PLINK tool [25].

2.5 Regulatory and Coding Regions

Coding regions of genes are first transcribed (gene expression) and then translated into proteins. Thus, they are directly related to the function (phenotype). Regulatory regions are non-coding sequences of DNA that enable regulation of gene expression when bound by regulatory proteins called transcription factors [38]. Promoter regions are an example of these regions which are just upstream of transcription start sites. Disruptions on these regions have ties to genetic diseases and disorders. Enhancers are regulatory DNA sequences that are distant-acting (far from the gene on the genome). When they bound by transcription factors, enhance the transcription of its target gene [39]. In complex diseases such as

invasive breast cancer, disruptions in enhancer regions have pivotal roles in terms of explaining the beginning and progress of the disease [40]. In particular distant-acting transcriptional enhancers are believed to be involved in the progression of complex diseases [41].

In this thesis, we promote selection of SNPs from the following these regions to prioritize epistasis tests with the conjecture that they might be better candidates for epistatis testing given their functional effects: (i) distant-acting transcriptional enhancers; (ii) promoter regions (1kb downstream and upstream of the transcription start sites - TSS); and (iii) coding regions.

Chapter 3

Related Work

In this chapter, we define the notation and describe related methods from the literature.

3.1 Notation

A GWAS dataset consists of genotypes of a set of samples S who are associated with a binary phenotype, e.g., a sample either has the disease or not. In the dataset, V denotes the SNP set. Function h which is denoted as the genotype of sample $s \in S$ at locus $v \in V$ is encoded as:

$$h(s,v) = \begin{cases} 0, & \text{if genotype of sample } s \text{ at } v \text{ is Homozygous major} \\ 1, & \text{if genotype of sample } s \text{ at } v \text{ is Hetorozygous} \\ 2, & \text{if genotype of sample } s \text{ at } v \text{ is Homozygous minor} \end{cases}$$

The genotype vector of locus v is denoted by h_v and consists of genotypes of sample set S at locus v.

Let $f(s)$ be a function that corresponds to phenotype of a sample $s \in S$, i.e.:

$$f(s) = \begin{cases} 1, & \text{if sample } s \text{ is case} \\ 0, & \text{if sample } s \text{ is control} \end{cases}$$

3.2 iLOCi

iLOCi is a locus interaction prioritization algorithm proposed by Piriyapongsa et al [9], which ranks the locus pairs by performing a *dependence test*. The method consists of two-stage: calculation of dependencies on locus pairs for case and control samples separately, and prioritization of locus pairs based on the *dependence test*s.

3.2.1 Calculation of dependencies

In total, there are $\frac{n(n-1)}{2}$ possible pairwise locus combinations, where n is the number of loci. For one thousand SNPs, $\sim$500000 interactions are needed to be considered. Thus, it is crucial to identify pairs to be prioritized first. In this step which is called as *dependence test*, iLOCi calculates two separate scores, ρ_{case} and $\rho_{control}$, for the case and control samples respectively which capture the correlation between locus pairs. However, among this many number of interactions, it is important to identify which ones are unrelated to the disease such as SNP pairs in linkage disequilibrium (LD). However, the methods to calculate LD such as using Hardy-Weinberg Equilibrium model is computationally very expensive especially for large datasets. iLOCi proves that there exists a relationship between ρ values calculated and LD obtained from allelic information [9]. Then, it detects LD using the dependence test based on the LD contrast method which can identify the disease signal above the background noise of dependent variants that are unrelated to the disease [42]. This method is computationally much more efficient compared with the HWE model.

To calculate ρ values, each locus is treated as a discrete random variable and genotype of each locus is encoded as 1, 0 and -1 which correspond to homozygous variant (v), heterozygous (h), and homozygous wild (w) respectively. Then, joint probabilities are calculated for each genotype combination which results in nine possible probabilities in total. Let i and j correspond to the two different loci,

discrete random variables. Then, genotype probability mass function $P_{(i,j)}$ for these loci is calculated as:

$$P_{(i,j)} = \begin{bmatrix} P_{ww} & P_{wh} & P_{wv} \\ P_{hw} & P_{hh} & P_{hv} \\ P_{vw} & P_{vh} & P_{vv} \end{bmatrix}$$

For the case samples, each of these probabilities in the matrix is calculated as:

$$P^{case}_{xy} = P^{case}_{(i=x,j=y)} = \frac{N^{case}_{(i=x,j=y)}}{N_{case}}$$

where $x, y \in \{w, h, v\}$ and N_{case} is the total number of case samples. For the control groups, the formula is the same except instead of case samples control samples are used. Based on these genotype probabilities, iLOCi calculates ρ_{case} and $\rho_{control}$ scores for each locus pair.

3.2.2 Prioritization of locus pairs

This step of the algorithm, calculating the difference between ρ scores, is named as *difference test*. After calculating ρ scores, iLOCi calculates the absolute difference between ρ_{case} and $\rho_{control}$, denoted by ρ_{diff}. Then, all locus pairs are sorted based on their ρ_{diff} values. iLOCi prefers to sort SNP pairs rather than a p-value cut-off to avoid FP pairs. Even though with this prioritization strategy, iLOCi improves epistatis detection computationally while considering SNPs with modest effect, it takes 19 hours to process a ∼500k SNPs dataset completely even with parallel computing.

3.3 PoCos

PoCo is an algorithm proposed by Ayati and Koyutürk [15] to test pairs of SNPs in population covering locus sets, PoCos, which may complement each other in terms of their association with the phenotype of interest. Algorithm first greedily selects multiple exclusive groups of SNPs that cover all affected

individuals as much as possible. Epistasis tests then are performed across PoCos with the hope that independent coverage of the cases will lead different PoCos to include complementary SNPs and thus, will lead to the test of epistatic pairs [15, 16].

In general, the allele that is less frequent in a population is referred as the minor allele and in most interaction analysis minor allele frequency (MAF) is used to determine association significance. But in PoCos, *allele of interest* is used as a term that is useful in distinguishing between case and control samples to consider the combinatory effects of different loci.

Given the allele of interest of each locus, function m is defined as follows given the genotype of sample $s \in S$ at locus $v \in V$:

$$m(s,v) = \begin{cases} 0, & \text{if genotype of sample } s \text{ at } v \text{ is Homozygous of allele of interest} \\ 1, & \text{if genotype of sample } s \text{ at } v \text{ is Hetorozygous} \\ 2, & \text{otherwise} \end{cases}$$

3.3.1 Population Covering Locus Set (PoCo)

Population Covering Locus Set (PoCo) is defined as locus set such that (i) at least one case sample contains an allele of interest at a locus within this set, and (ii) the number of control samples that contain an allele of interest at a loci within this set is minimized. Based on these two constraints, formally PoCo is formulated as:

Definition 3.3.1. PoCos Let define $\alpha(v)$ as the allele of interest for a locus $v \in V$ and $E(v) \subset S$ and $T(v) \subset S$ as the subset of case and control samples that contain the allele of interest in locus v respectively. Mathematically, these sets are formulated as:

$$E(v) = \{s \in S : f(s) = 1 \text{ and } h(v,s) = a(v)\}$$

$$T(v) = \{s \in S : f(s) = 0 \text{ and } h(v, s) = a(v)\}$$

Based on these definitions, the purpose is to find a PoCo, a set $P \subset V$, that solves the following problem:

$$\begin{aligned} \underset{P}{\text{minimize}} \quad & |\bigcup_{v \in P} T(v)| \\ \text{subject to} \quad & \bigcup_{v \in P} E(v) = \{s \in S : f(s) = 1\} \end{aligned} \tag{3.1}$$

Even though the problem is defined as a constrained optimization problem, the aim is to find POCOs that satisfy the optimization problem 3.1 locally rather than finding a single optimal POCO for the problem. Thus, they use the equation 3.2 to discover PoCos.

$$\delta(P) = \frac{|\bigcup_{v \in P} E(v)|}{\{s \in S : f(s) = 1\}} - \frac{|\bigcup_{v \in P} T(v)|}{\{s \in S : f(s) = 0\}} \tag{3.2}$$

With a greedy approach, in each iteration they select the locus that maximizes $\delta(P)$, the difference of the case and control ratios covered by PoCo P, and add the locus to the set P until no locus can enrich the set P more in terms of maximizing $\delta(P)$ or all cases are covered.

3.3.2 Epistatic Pair Priorization

Construction of representative genotypes

The idea behind this method is to reduce the number of epistasis tests performed by prioritizing locus pairs based on the epistatic interactions between the corresponding PoCos in which they belong. Since these PoCos do not correspond to exact genotype loci but rather are representatives of correlated locus sets, representative genotypes should be calculated for each PoCo. Representative genotype for each PoCo $p \in P$ is calculated as

$$H(P, s) = \sum_{c \in P} h(c, s) \quad \forall s \in S$$

Epistasis Test Model

A logistic regression model is used to test pairwise interactions between PoCo pairs, i.e:

$$f = \beta_0 + \beta_i H(P_i) + \beta_j H(P_j) + B_{ij} H(P_i) H(P_j) \quad \forall P_i, P_j \in P$$

Using this model, they assign statistical significance value of the interaction term B_{ij} to all locus pairs within these PoCos and loci that are not in any PoCo is labeled as *unscored*. Scored values are sorted in descending order and tested based on that order.

Chapter 4

Methods

In this section we formulate the epistasis test prioritization problem and describe our method in context with the literature.

4.1 Problem Description

In this study, the goal is to perform feature selection to guide epistasis test prioritization to minimize false positive findings. We propose the following pipeline. Given a GWAS dataset and a corresponding SNP set V, the first step is to select a diverse SNP subset M such that $|M| = k$, in which cardinality $k \ll |V|$ on a SNP-SNP interaction network G. Second step is to find statistically significant epistatic pairs within the selected SNP subset M. Next, we describe the details of the proposed approach and the methods we rely on.

4.2 LinDen

Cowman and Koyutürk propose a fast epistasis detection algorithm, LinDen, that exploits LD structure to decrease the number of tests in epistasis detection [17]. In the proposed formulation, they declare the problem as discovering the most statistically significant epistatic partner for each locus. In consideration of the problem statement, LinDen proposes following definitions to limit the search space for epistatic interactions.

Definition 4.2.1. *Most significant epistatic partner for a locus* For each locus $v_i \in V$, let $v_j \in V \setminus v_i$ be the most significant epistatic partner for v_i. Then, v_i and v_j satisfy the condition that $\chi^2(v_i, v_j) > \chi^2(v_i, v_k) \quad \forall v_k \in V \setminus \{v_i, v_j\}$ in which χ^2 test of the $v_i, v_j \in V$ is denoted by $\chi^2(v_i, v_j)$.

Definition 4.2.2. *Reciprocally significant epistatic pairs* If two different loci $v_i, v_j \in V$ are most significant epistatic partners for each other, then they are reciprocally significant epistatic pairs.

Definition 4.2 guarantees that only one epistatic partner is outputted for each locus. The idea behind this approach is to reduce noise due to large marginal effects. Based on those definitions LinDen proposes a framework which is described in figure 4.1.

4.2.1 Construction of Linkage Disequilibrium Trees

LinDen uses a tree-representation to group loci that are in linkage disequilibrium (LD) to reduce the number of tests performed.

Definition 4.2.3. LD-Tree LinDen defines LD-Tree as a full binary tree where each node t corresponds to a set $L(t) \subset V$ of genomic region and is represented with a genotype vector, R_t. There exist one-to-one relationship between genomic loci and leaf nodes, i.e. $R_t = h_c$ for a leaf node t where $L(t) = \{c\}$. For an

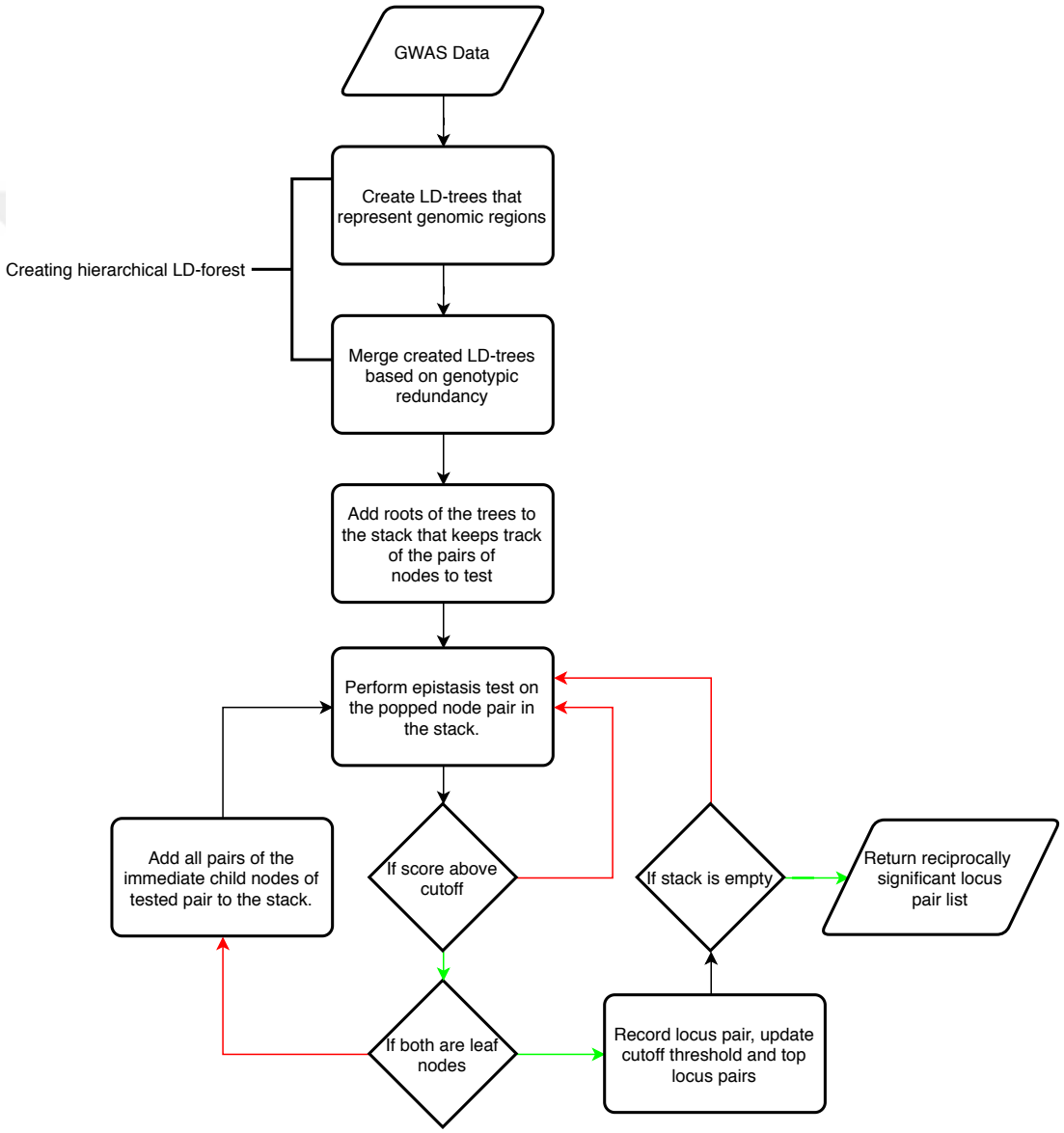

Figure 4.1: LINDEN work-flow for fast epistasis detection. Decision nodes are represented by rhombus, and green and red arrows represent yes and no respectively.

internal node t, in which left and right children are represented with t_l and t_r respectively, the representative genotype vector R_t is defined as:

$$R_t(s) = \begin{cases} R_{t_l}(s), & \text{if } R_{t_l}(s) = R_{t_r}(s) \\ NIL, & \text{otherwise} \end{cases} \quad \forall s \in S$$

With this formulation, LinDen aims to group loci that are in LD by controlling ambiguity in genotype vector. In this context, ambiguity refers to number of NIL values in the genotype vector.

First, it constructs a tree for each locus, so initially there is a collection of $|V|$ trees. Then, iteratively, it scans all trees and perform pairwise tree merge such that the number of NILs in the genotype vector of the root of the new merged tree is at most $d*$. Here, d is a parameter to control the number of ambiguous samples and $d* \leq d$ is a dynamic threshold which increases in each iteration. In each iteration, LinDen compares each tree to its nearest b-neighboring trees, in which parameter b controls the range of LD in terms of topological closeness. In the first iteration, $d* = 0$ to merge identical genotypes. In following iterations, $d*$ is incremented by 1% and the merging process continues until $d* = d$.

4.2.2 Discovery of Epistatic Interactions

There are two types of tests LinDen performs to detect significant epistatic interactions: (i) Tests containing internal nodes, and (ii) Tests between leaf node pairs. Since each leaf node corresponds to a locus, a standard χ^2 test on 9×2 contingency table of all genotype combinations between cases and controls is performed between the leaf nodes. Each genotype class is treated as a fixed effect and 8 df test is conducted. Only the significant pairs detected in result of this test is used as output since only leaf nodes correspond to SNPs. The tests between internal nodes aim to find subtrees that may involve significant leaf pairs. Thus, an estimation function is used to determine interaction significance between internal nodes, considering that representative genotype vectors are not

ideal representative of their children nodes. For this purpose, they offer to use an estimation function that provides a heuristic bound on χ^2 statistics. To determine if an internal test should pass or not, they set a *cut-off* threshold χ^* which is a dynamic significance threshold. Then, in a top-down manner, they perform tests in accordance with the type of the node (leaf or internal) and finally, they output a list of reciprocally significant pairs that contains discovered significant leaf nodes.

4.3 SPADIS

Yilmaz et al. propose an algorithm, SPADIS, to select predictive and diverse genetic variants [18]. It favors selection of distant variants, which are associated with the phenotype, in a given biological network $G(V, E)$ to avoid selecting redundant variants that have similar functionalities. In directed/undirected graph G, SNPs are represented by vertices V and edges E indicates the relationship, functional or topological proximity, between SNPs. SPADIS follows a two-step approach to select a SNP set M. Initially, it assigns a score for each SNP based on its association with the given phenotype via the Sequence Kernel Association Test (SKAT) by regressing phenotype on the covariates using a flexible semiparametric linear model [43]. Instead of directly associating genotypes of the variants with the phenotype, SKAT uses a nonparametric function of the genotypes that is possibly contained in a vector space generated by a positive semi-definite kernel function. Using the kernel functions provides flexibility and increased model complexity. The score assigned to i-th SNP is indicated with $c_i \in R_{\geq 0}$. Then it maximizes a submodular set function F with a greedy approach to select diverse SNPs while maximizing the total score of SNP set M. The set function F is defined as:

$$F(M) = \sum_{i \in M} (c_i + \beta(1 - \sum_{j \in M} (\frac{K(i,j)}{2k})) \tag{4.1}$$

$$K(i,j) = \begin{cases} 1 - d(i,j)/D, & \text{if } d(i,j) \leq D, i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

In this formulation, $d(i, j)$ represents the shortest distance between i-th and j-th SNPs. $D \in \mathbb{R}_{>0}$ is a maximum distance parameter. Here, $K(i, j) \in [0, 1], \forall i, j \in V$ is a function which is used to penalize proximity between SNPs and parameter $\beta \in \mathbb{R}_{\geq 0}$ determines the degree of this penalization. To solve this NP-hard subset selection problem, a heuristic approach is proposed by Nemhauser et al [44] is used. This algorithm, given in **Algorithm 1**, ensures to return an approximation close to the optimal solution within a constant bound $1-(1/e)$ for monotonically non-decreasing and nonnegative submodular functions.

Algorithm 1 Greedy algorithm

Input: Ground set V, submodular set function F and cardinality constraint $k \leq |V|$
Output:

1: $M \leftarrow \emptyset$
2: **for** $|M| < k$ **do**
3: $\quad M \leftarrow M \cup \text{argmax}_{x \in V \setminus M} F(M \cup x)$
4: **end**

4.4 Proposed Algorithms

4.4.1 Guiding LinDen with SPADIS

SPADIS selects diverse and explanatory SNP set without introducing a literature bias. Our hypothesis is that those selected SNPs are likely to be epistatic since SPADIS is designed to diversify the set and select complementary SNPs. Thus, it provides a pruned search space of likely epistatic pairs for LinDen. Based on our hypothesis, we propose a pipeline that first uses SPADIS to generate its candidate set for epistasis test. Then, LinDen form LD trees only over SPADIS-selected regions (selected SNPs and a small number of neighbors) to pick likely epistatic pairs from this set. The work-flow of the proposed algorithm is described in Figure 4.2.

Initially, datasets are preprocessed which is explained in Chapter 5 in detail. Using these datasets, SPADIS selects k SNPs further away from each other and associated with the phenotype. By default, SPADIS uses a continuous SKAT scoring. However, since our phenotype of interest is dichotomous, SKAT scoring which measures the association between phenotype and variant should change accordingly. To construct a relationship between the binary phenotype and the covariates, we replace semiparametric linear regression model with the semiparametric logistic regression model which also utilizes kernel functions. We add n, in our experiments $n = 9$, topologically nearest neighbors of each selected SNP to the set. We modified the initial merging step of LinDen such that in the first and second iteration of merging, each SPADIS-selected region can be merged only within themselves to form a new tree as illustrated in Figure 4.3. Then, in each successive iteration, LinDen continues to compare each tree to the b, in our experiments $b = 10$, closest trees, as it does originally.

4.4.2 Integrating Regulatory and Coding Regions

Binding of proteins to regulatory regions affects the expression level of a gene as well as mutations on the coding regions themselves. A mutation on such regions may have a significant effect on the protein production and might be related to disease. Given that there are millions of possible combinations for locus-locus interactions and it is computationally infeasible to test all such pairs, prioritization process might benefit from using mutations falling into these regions. Thus, we conjecture that we can find more statistically significant and biologically meaningful SNP pairs via promoting mutations in regulatory and coding regions. In this study, we propose to use three region types which are (i) distant-acting transcriptional enhancer regions, (ii) promoter regions (1kb downstream and upstream of the transcription start sites - TSS) and (iii) coding regions into the two different stages of the algorithm separately. We integrate them to the algorithms and report the performance in three scenarios: SPADIS-only, to LinDen-only, and both to SPADIS and LinDen.

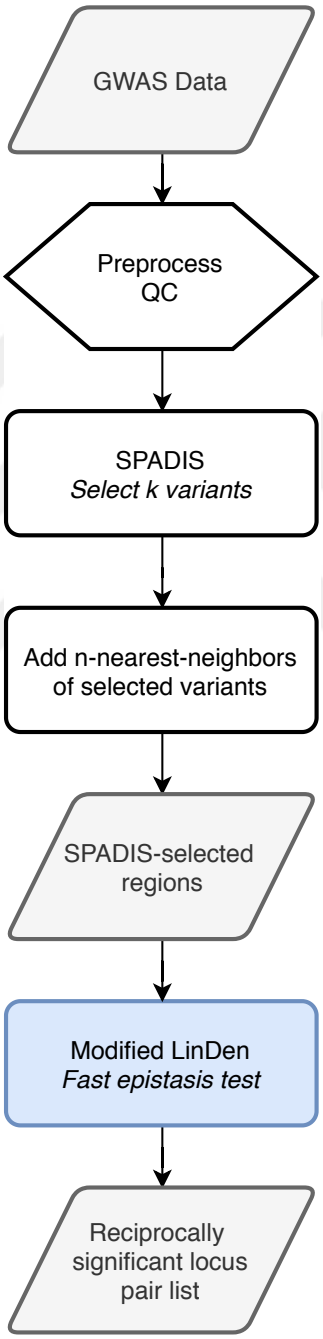

Figure 4.2: SPADIS+LinDen work-flow. LinDen is modified to initially form LD-trees on each SPADIS-selected region by merging each selected SNP with its n-neighbors. The GWAS data may be filtered based on subject-based quality measures and variant-based quality measures.

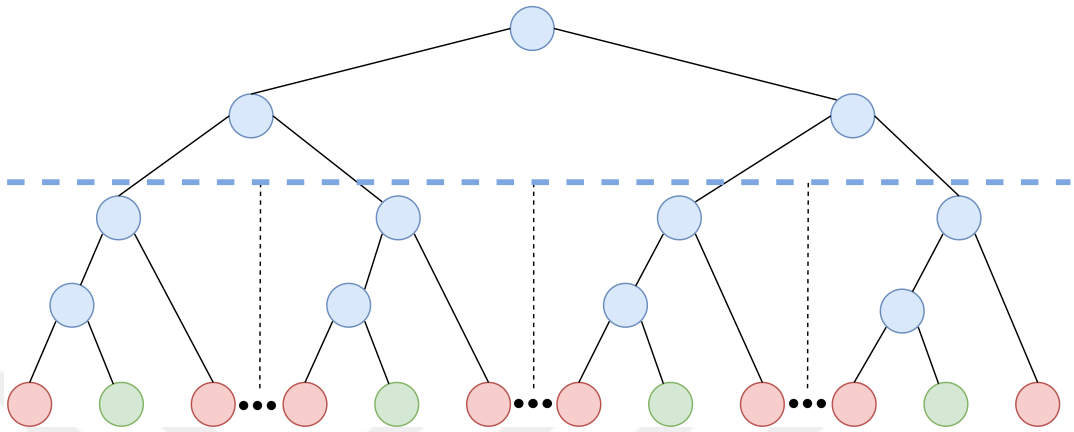

Figure 4.3: An example of modified merging procedure. Green, red and blue nodes denote the SPADIS-selected SNPs, neighbors of selected SNPs and merged nodes, respectively. Vertical dotted-lines separate each SPADIS-selected region. Above blue dotted-line merging procedure continues as in the original algorithm.

Integrating SPADIS with regulatory and coding regions

SPADIS favors the selection of distant and explanatory SNPs [18]. It does that by rewarding maximizing a submodular function F as stated in equation 4.1. By preserving the submodularity of the function F, we can integrate the idea of favoring selection of SNPs that are in regulatory regions by manually increasing the scores of those SNPs. Let $w_i \in \{0, 1\}$ be a binary indicator for i-th SNP such that:

$$w_i = \begin{cases} 1, & \text{if SNP } i \text{ is in at least one of the regulatory/coding regions} \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

Then, we can modify set function F such that:

$$F(M) = \sum_{i \in M}((1 + w_i)c_i + \beta(1 - \sum_{j \in M}(\frac{K(i,j)}{2k})) \tag{4.4}$$

Lemma 4.4.1. If the functions f, g are submodular, $\lambda, \beta \in \mathbb{R}$ and M is a set, then $h(M) \coloneqq \lambda f(M) + \beta g(M)$ is submodular.

Lemma 4.4.2. The function $F(M)$ which is given in equation 4.4 is submodular.

Proof. F is a submodular function iff

$$(F(A \cup \{x\}) - F(A)) - (F(B \cup \{x\}) - F(B)) \geq 0$$
$$\forall A, B \ni A \subset B \subset V \text{ and } x \in V \setminus B.$$

We can rewrite the function $F(M)$ in the form of two sums, i.e.:

$$F(M) = \sum_{i \in M} (c_i + \beta(1 - \sum_{j \in M} (\frac{K(i,j)}{2k})) + \sum_{i \in M} w_i c_i$$

The submodularity of the first sum is proven by Yilmaz et al [18]. By Lemma 4.4.1, we only need to show that second sum is submodular to prove that F is submodular. Let denote the second sum as $H(M) \coloneqq \sum_{i \in M} w_i c_i$ and define a function T by

$$T(A, B, x) \coloneqq H(A \cup \{x\}) - H(A) - H(B \cup \{x\}) + H(B)$$

To prove H is submodular, we need to show that $T(A, B, x) \geq 0 \ \forall A, B \ni A \subset B \subset V$ and $x \in V \setminus B$. Then,

$$T(A, B, x) = \sum_{i \in A} w_i c_i + w_x c_x - \sum_{i \in A} w_i c_i$$
$$- \sum_{i \in B} w_i c_i - w_x c_x + \sum_{i \in B} w_i c_i = 0$$

Since $T(A, B, x) = 0$, H is submodular. By using lemma 4.4.1, we can conclude that F, the sum of two submodular functions, is submodular. □

F without adding the new term w_i is non-negative and monotonically non-decreasing as proven by Yilmaz et al [18]. In our formulation, we only positively weight the score term with a nonnegative term, w_i, thus we preserve these two properties of the function. Thus, we can use the greedy algorithm, given in **Algorithm 1**, which ensures to return an approximation close to the optimal solution within a constant bound $1 - (1/e)$ for non-negative and monotonically non-decreasing submodular functions.

Integrating LinDen with regulatory and coding regions

To guide LinDen further, we propose to integrate regulatory and coding regions as weights to the LD-trees. As in the SPADIS integration above, for each locus, we check if the regulatory/coding regions contain that locus. Initially, for each locus i at the leaf nodes, we assign a weight $w_i \in [0, 1]$. Initial weights are assigned based on the definition given in the equation 4.3. Then, at each merging iteration, the weight of the new root is calculated as the average weight of its children nodes, i.e: $w_r = \frac{w_{lc}+w_{rc}}{2}$ where w_r, w_{lc}, w_{rc} are weights of the root, left child, and right child, respectively. Then, while performing an internal test for nodes i and j with weights w_i, w_j, the dynamic significance threshold χ^* is decreased based on the average of the weights, i.e.: $\chi^*_{new} = \chi^*(1 - (w_1 + w_2)/4)$. Then, any node pairs which have χ^2 statistic less than χ^*_{new} are eliminated during internal tests and any children nodes of the eliminated nodes are not tested. Via this method, we aim to minimize the number of false positives with the hypothesis that the trees containing SNPs in regulatory/coding regions are more likely to contain epistatic pairs.

By combining these two upgraded version of SPADIS and LinDen as in the Figure 4.2, we create a new pipeline which enables to favor SPADIS-selected SNPs that are in regulatory/coding regions as candidate epistatic pairs.

Chapter 5

Results

5.1 Datasets

We use three GWAS datasets obtained from WTCCC: Type 2 diabetes (T2D), Hypertension (HT), Bipolar disorder (BD) cohorts [45]. As exercised in numerous articles [46, 47], quality control is done on the datasets using PLINK tool [25] following a number of steps:

- ***Gender assignment check:*** The subjects, in which the chromosome X data conflicts with the gender reported, are removed from the datasets.
- ***Removing individuals with high missing genotype data:*** We removed the individuals with more than 10 percent missing genotype rate from the datasets.
- ***Removing rare SNPs:*** Once we excluded the individuals with poor quality, we removed SNPs with less than 5 percent minor allele frequency.
- ***Removing SNPs with high missing genotype rate:*** We removed SNPs with more than 10 percent missing genotype rate from the datasets.
- ***Removing SNPs that fail Hardy-Weinberg Equilibrium (HWE)***

test: We removed SNPs that fail to pass HWE with a nominal p-value threshold 1e−6.

After preprocessing the datasets, remaining number of SNPs, as well as number of cases and number of controls used are given in Table 5.1.

Table 5.1: Information of the T2D, BD and HT datasets which are used in our experiments.

Dataset	**# SNPs**	**# Cases**	**# Controls**
T2D	378016	1973	1498
HT	377547	1996	1504
BD	378095	1993	1504

We collected distant-acting transcriptional enhancer dataset from VISTA Enhancer Browser [48], and transcription start sites (TSSs) and coding regions from UCSC Genome Browser [49]. VISTA enhancer dataset contains 1912 human non-coding fragments with gene enhancer activity. To obtain gene locations, we used UCSC Genome Browser. From this system, we chose Ensembl Genes as gene annotation track [50]. We posed the following query to UCSC Table Browser to obtain the genes locations:

Clade: Mammal
Organism: Human
Assembly: Mar. 2006 (NCBI36/hg18)
Group: Genes and Gene Predictions
Track: Ensembl Genes
Table: ensGene
Region: genome

The number and types of genes obtained from the Ensembl dataset are given in Table 5.2. We defined one-kb downstream and upstream of each TSS as the regulatory region. The coding regions are defined as the start of the first exon till the end of the last exon.

Table 5.2: The number of genes in the Ensembl dataset and gene predictions dataset obtained from UCSC Genome Browser.

	Gene Counts
Known protein-coding genes:	21370
Novel protein-coding genes:	46
Pseudogenes:	9899
RNA genes:	5732
Genscan gene predictions:	49796

5.2 Networks

There are three different networks constructed in SPADIS to determine locus-locus interactions: Gene Interaction (GI), Gene Sequence (GS), Gene Membership (GM) networks [18]. In the GS network, loci that are next to each other in genomic sequence are connected. In the GM network which is a superset of GS network, loci that are on the same gene or both close to the same gene within 20k base pair distance, are connected. In the GI Network which is a superset of the GM network, two loci are connected if they interact in a protein-protein interaction network. All of these networks consider spatial proximity in 1-dimensional DNA sequence. To exploit the information in the 3D organization of the DNA, Yilmaz et al introduce a new network: GS-HICN. This network considers interactions between genomic loci that are close to each other in 3D space as well as being next to each other in the 1D sequence (GS network). In this study, we only utilize GS network since in SPADIS it yields better or comparable results among other networks [18].

5.3 Experimental Setup

We run the proposed pipeline on the three aforementioned WTCCC dataset after doing the quality control on these datasets. In order to evaluate performance of the pipeline, we construct four different experimental setups.

Setup 1: LinDen. The experimental setting for LinDen is described by Cowman et al [17]. We set the parameter d, which determines the fraction of ambiguous samples, as 0.45 and parameter b, which determines the extent of LD, as 10 based on the observations in [17]. We run LinDen by setting the maximum number of threads to 20 in parallel setting.

Setup 2: SPADIS. The experimental settings for SPADIS is explained by Yilmaz et al [18]. In our experiments, we also use 10-fold cross validation. However, since phenotypes which we consider are binary, while evaluating the performance of SPADIS and selecting the best parameter set, we perform ridge penalized logistic regression. Also, in our experiments SKAT scores are calculated considering dichotomous phenotypes instead of continuous traits which is the case in the default SPADIS formulation. We run SPADIS for five different k values: 500, 750, 1000, 1500, 2000. Based on our observations, these values provide best accuracies as the result of ridge penalized logistic regression. When k is 5000 or 10000, the accuracy decreases dramatically up to $\sim$22%, compared to selected k values.

Setup 3: SPADIS + LinDen. For the pipeline, we keep the parameters same as the Setup 1 and 2 which are explained above. Only, we limit the extend of LD in the first two iterations of the merging procedure of LinDen as explained in Section 4.2.1 and set the parameter n to 9.

Setup 4: SPADIS + LinDen with integrated regulatory/coding regions. For the pipeline, we keep the parameters same as the Setup 3 which are explained above. Only, we integrate regulatory/coding regions into different stages of the pipeline as explained in Section 4.2.2 and present the results for each stage separately.

To quantify the performance of the proposed algorithms, we used precision ($TP/(TP + FP)$) as the evaluation metric in which true positives (TP) refer to the reciprocally significant epistatic pairs that pass the Bonferroni adjusted threshold and false positives (FP) refer to the reciprocally significant epistatic pairs that are below the Bonferroni adjusted threshold. We set the significance level as 10% throughout experiments and adjust the significance level using the Bonferroni correction based on the number of test performed by each approach.

5.4 Precision Improvement Guiding LinDen with SPADIS

We measure the improvement in precision using the pipeline approach on the WTCCC datasets. First, we run LinDen on these datasets and, it returns 1786, 906 and 1135 reciprocally significant epistatic pairs for T2D, BD, and HT datasets, respectively. Only 5, 30, and 5 are statistically significant at the 0.1 level after Bonferroni adjustment, respectively. These correspond to precision values of 0.003, 0.033, and 0.004, respectively. This sets our baseline. Then, we first input the same datasets to SPADIS and let it select top k SNPs, where $k = 500, 750, 1000, 1500$ and 2000. The parameters are optimized to maximize classification accuracy and individual SNP scoring method is configured to work with discrete labels. Next, for each selected SNP, we also gather nearest 9 upstream and 9 downstream SNPs on the genome. Finally, we input these SNPs into LinDen and let it find epistatic pairs with default parameters. In this version, we prohibited LinDen to merge far away trees with respect to the distance on genome. Complete results are shown in Tables 5.3, 5.4, 5.5 for T2D, BD and HT datasets, respectively. The guidance of SPADIS improves the precision substantially, from 0.003 up to 0.421. This is achived by drastically reducing the number of false positives, while maintaining of increasing the number of true positives. Our pipeline outperforms LinDen for all k values on all datasets, but we observe that the ideal k values are 500, 750 and 1000. For k values 1500 and 2000, the precision values decrease drastically compared to results for lower k

values. This is inline with the diminishing returns property of the optimization function of SPADIS that as the set is enlarged the gain decreases. So, as k is increased beyond 1000, the gain in diversification is marginal and the guidance of SPADIS no longer helps to guide LinDen.

Table 5.3: Results for T2D dataset. Number of pairs reported is the total number of reciprocally significant pairs returned by LinDen with and without the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. Table shows that the guidance of SPADIS increases the precision substantially as compared to LinDen only.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
LinDen with all T2D		378016	1786 (5)	0.003
SPADIS + LinDen	*k = 500*	9250	19 (8)	0.421
	k = 750	14146	43 (13)	0.302
	k = 1000	18943	55 (21)	0.382
	k = 1500	28014	79 (18)	0.228
	k = 2000	37927	95 (11)	0.116

Table 5.4: Results for BD dataset. Number of pairs reported is the total number of reciprocally significant pairs returned by LinDen with and without the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. Table shows that the guidance of SPADIS increases the precision substantially as compared to LinDen only.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
LinDen with all BD		378095	906 (30)	0.033
SPADIS + LinDen	*k = 500*	9387	35 (3)	0.085
	k = 750	14098	21 (5)	0.238
	k = 1000	17048	23 (8)	0.348
	k = 1500	28161	52 (15)	0.289
	k = 2000	36323	69 (18)	0.261

Table 5.5: Results for HT dataset. Number of pairs reported is the total number of reciprocally significant pairs returned by LinDen with and without the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. Table shows that the guidance of SPADIS increases the precision substantially as compared to LinDen only.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
LinDen with all HT		377547	1135 (5)	0.004
SPADIS + LinDen	*k = 500*	8744	15 (4)	0.267
	k = 750	11886	27 (8)	0.296
	k = 1000	18953	34 (8)	0.235
	k = 1500	27275	50 (6)	0.120
	k = 2000	37979	132 (6)	0.045

We also compare SPADIS+LinDen with LinDen in Figures 5.1, 5.2, 5.3 for the T2D, BD and HT datasets, respectively. The green lines denote the significance level (0.1) to be passed for each approach ($k = 1000$) after Bonferroni correction. The dots below and above the threshold represent false positives and true positives, respectively. It is clear that the pipeline drastically reduces the number of false positives while increasing the number true positives. Also, we can observe the importance of the number of tests performed by looking at the difference between Bonferroni thresholds. Since SPADIS provides a pruned search space for LinDen by eliminating SNPs that are most likely irrelevant to the disease, it also reduces number of tests that will be performed during epistasis test. Indirectly it eliminates the negative effect of multiple hypothesis testing which reduces the statistical power of the tests performed, thus making Bonferroni correction less conservative. As seen in the figures, due to low Bonferroni threshold the pipeline is able to discover more true positives compared to LinDen. We also show that our pipeline not only minimizes the number of false positives but is also able to maintain the significance level of the returned pairs. Figures show that SPADIS+LinDen's top picked pairs have similar p-values in T2D dataset. Despite lower significance levels in BD and HT datasets, both LinDen and our pipeline is able to return a single pair which stands out in terms of its p-value.

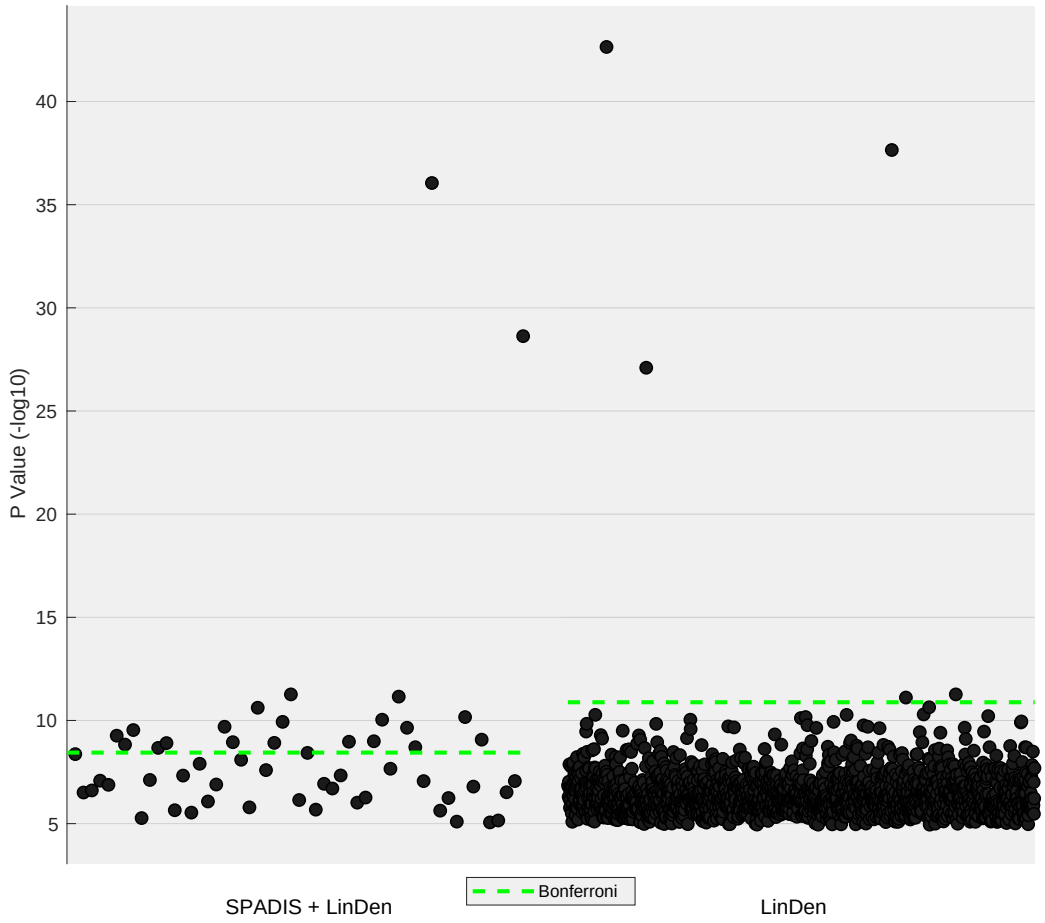

Figure 5.1: On the T2D dataset, for each approach, SPADIS+LinDen, LinDen only, we show the significance levels (y-axis) of each reported pair (dots) given the Bonferroni corrected significance threshold (0.1, green line). X – axis is just randomly assigned values to pairs for visualization for $k = 1000$. SPADIS clearly minimizes FPs by guiding LinDen.

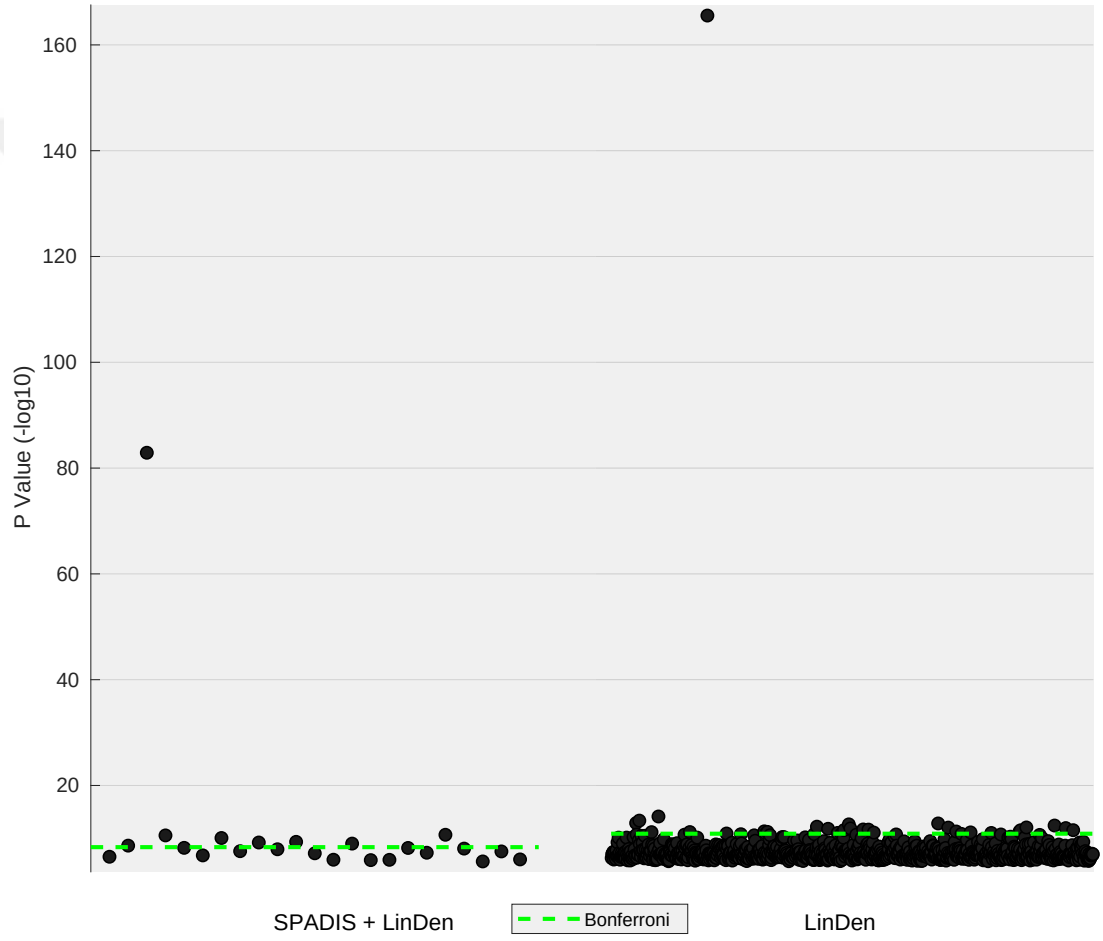

Figure 5.2: On the BD dataset, for each approach, SPADIS+LinDen, LinDen only, we show the significance levels (y-axis) of each reported pair (dots) given the Bonferroni corrected significance threshold (0.1, green line). X – axis is just randomly assigned values to pairs for visualization for $k = 1000$. SPADIS clearly minimizes FPs by guiding LinDen.

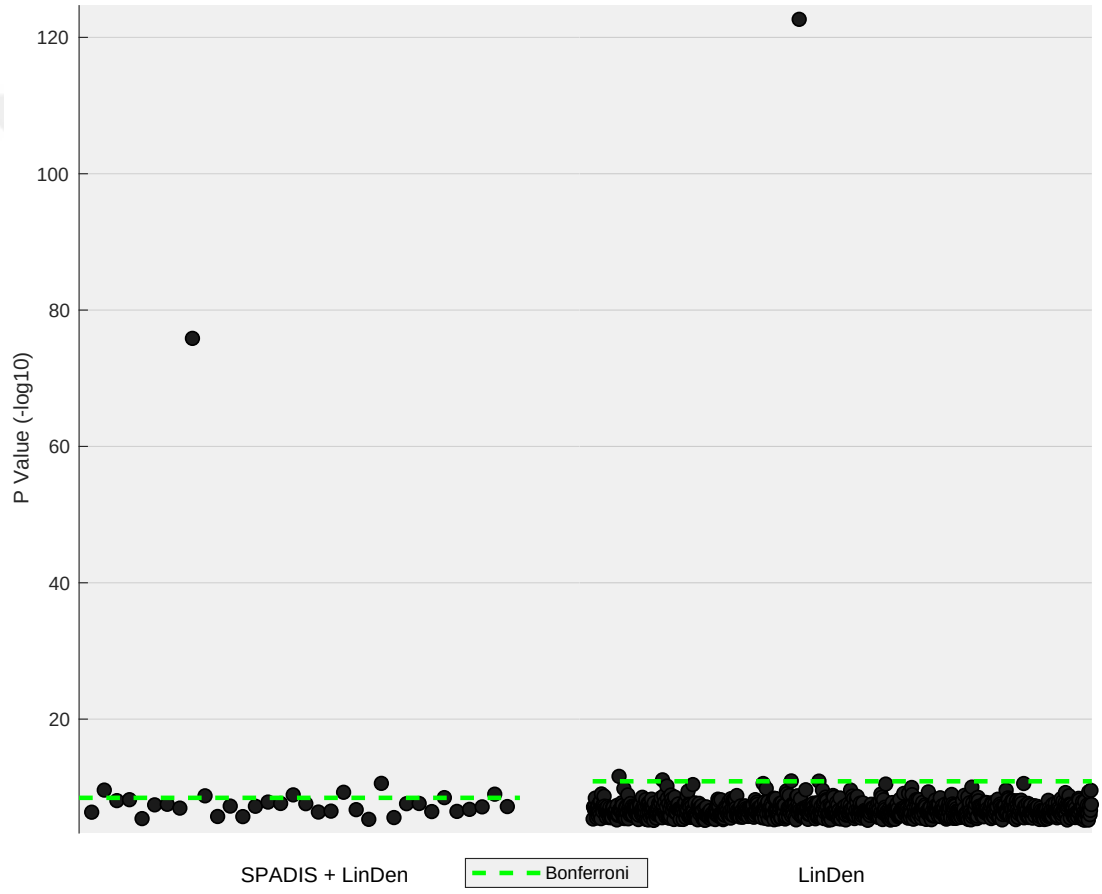

Figure 5.3: On the HT dataset, for each approach, SPADIS+LinDen, LinDen only, we show the significance levels (y-axis) of each reported pair (dots) given the Bonferroni corrected significance threshold (0.1, green line). X – axis is just randomly assigned values to pairs for visualization for $k = 1000$. SPADIS clearly minimizes FPs by guiding LinDen.

5.5 Integrating Regulatory and Coding Regions

In this section, we present the results of integrating regulatory and coding regions to LinDen and to the overall SPADIS + LinDen pipeline and check whether these regions contain epistatic pair candidates that are associated with the phenotype of interest. We also demonstrate that weighting only SPADIS with regulatory/coding elements improves the accuracy in terms of regression performance.

5.5.1 Integrating SPADIS with Regulatory and Coding Regions

In this section, we check if integrating regulatory/coding regions into the SPADIS algorithm improves the phenotype prediction performance. We find that when we reward regulatory/coding regions in SPADIS SNP selection process and then apply ridge regularized logistic regression via using those selected SNPs, the accuracy of SPADIS increases. The Tables 5.6, 5.7 and 5.8 compares original SPADIS to the SPADIS which regulatory/coding elements are rewarded for T2D, BD and HT datasets, respectively.

Table 5.6: Accuracy comparison for SPADIS vs SPADIS with integrated regulatory/coding regions on T2D dataset for various k values.

	Accuracy	
# SNPs	**SPADIS**	**Regulatory/Coding SPADIS**
k = 500	0.985	0.988
k = 750	0.978	0.981
k = 1000	0.967	0.972
k = 1500	0.956	0.955
k = 2000	0.933	0.934

Table 5.7: Accuracy comparison for SPADIS vs SPADIS with integrated regulatory/coding regions on BD dataset for various k values.

	Accuracy	
# SNPs	**SPADIS**	**Regulatory/Coding SPADIS**
k = 500	0.998	0.999
k = 750	0.998	0.999
k = 1000	0.997	0.998
k = 1500	0.991	0.992
k = 2000	0.978	0.978

Table 5.8: Accuracy comparison for SPADIS vs SPADIS with integrated regulatory/coding regions on HT dataset for various k values.

	Accuracy	
# SNPs	**SPADIS**	**Regulatory/Coding SPADIS**
k = 500	0.995	0.995
k = 750	0.990	0.991
k = 1000	0.983	0.985
k = 1500	0.971	0.973
k = 2000	0.955	0.959

5.5.2 Integrating LinDen with Regulatory and Coding Regions

We analyze the effect of integrating regulatory/coding regions into only LinDen stage of the pipeline by weighting LD-trees that are in those regions. We compare this approach with the original SPADIS+LinDen pipeline. The results of this analysis are shown in Tables 5.9, 5.10 and 5.11 for T2D, BD and HT datasets, respectively. Major improvements over the original SPADIS + LinDen are observed in the precision values of BD and HT datasets up to 47% ($k = 500$) and 42%($k = 2000$) respectively. Also, for T2D dataset we may observe light improvements up to 28% ($k = 2000$). Green cells indicate improvements compared with Tables 5.3, 5.4, and 5.5, respectively.

Table 5.9: Results for T2D dataset. Number of pairs reported is the total number of reciprocally significant pairs returned by weighted LinDen with the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to the LinDen and SNPs in regulatory/coding regions are weighted. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. The green cells show that the guidance of regulatory/coding regions increases the precision as compared to the original pipeline.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
SPADIS + LinDen	*k = 500*	9250	19 (6)	0.316
	k = 750	14146	**32 (10)**	**0.313**
	k = 1000	18943	**47 (19)**	**0.404**
	k = 1500	28014	**67 (16)**	**0.239**
	k = 2000	37927	**94 (14)**	**0.149**

Table 5.10: Results for BD dataset. Number of pairs reported is the total number of reciprocally significant pairs returned by weighted LinDen with the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen and SNPs in regulatory/coding regions are weighted. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. The green cells show that the guidance of regulatory/coding regions increases the precision as compared to the original pipeline.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
SPADIS + LinDen	*k = 500*	9387	**32 (4)**	**0.125**
	k = 750	14098	**19 (5)**	**0.263**
	k = 1000	17048	**17 (7)**	**0.412**
	k = 1500	28161	**51 (19)**	**0.373**
	k = 2000	36323	**65 (20)**	**0.308**

Table 5.11: Results for HT dataset. Number of pairs reported is the total number of reciprocally significant pairs returned by weighted LinDen with the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen and SNPs in regulatory/coding regions are weighted. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. The green cells show that the guidance of regulatory/coding regions increases the precision as compared to the original pipeline.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
SPADIS + LinDen	*k = 500*	8744	**15 (4)**	**0.267**
	k = 750	11886	**27 (11)**	**0.407**
	k = 1000	18953	**29 (7)**	**0.241**
	k = 1500	27275	**49 (6)**	**0.122**
	k = 2000	37979	**110 (7)**	**0.064**

We also illustrate the power of integrating regulatory/coding regions into LinDen in terms of discovering more significant epistatic pairs by comparing it with the default pipeline in Figures 5.4, 5.5, 5.6 for the T2D, BD and HT datasets, respectively (for $k = 1000$). To highlight the false positives in the figures, the p-values of reported pairs are sorted in ascending order. Since Bonferroni corrected thresholds of each approach are too tight that cannot be distinguished in the figure, we draw the Bonferroni threshold as the average of both thresholds.

The green line denotes the Bonferroni corrected significance level to be passed as the average of the actual thresholds of each approach. The actual Bonferroni corrected thresholds are indicated in the respective figure caption on each dataset.

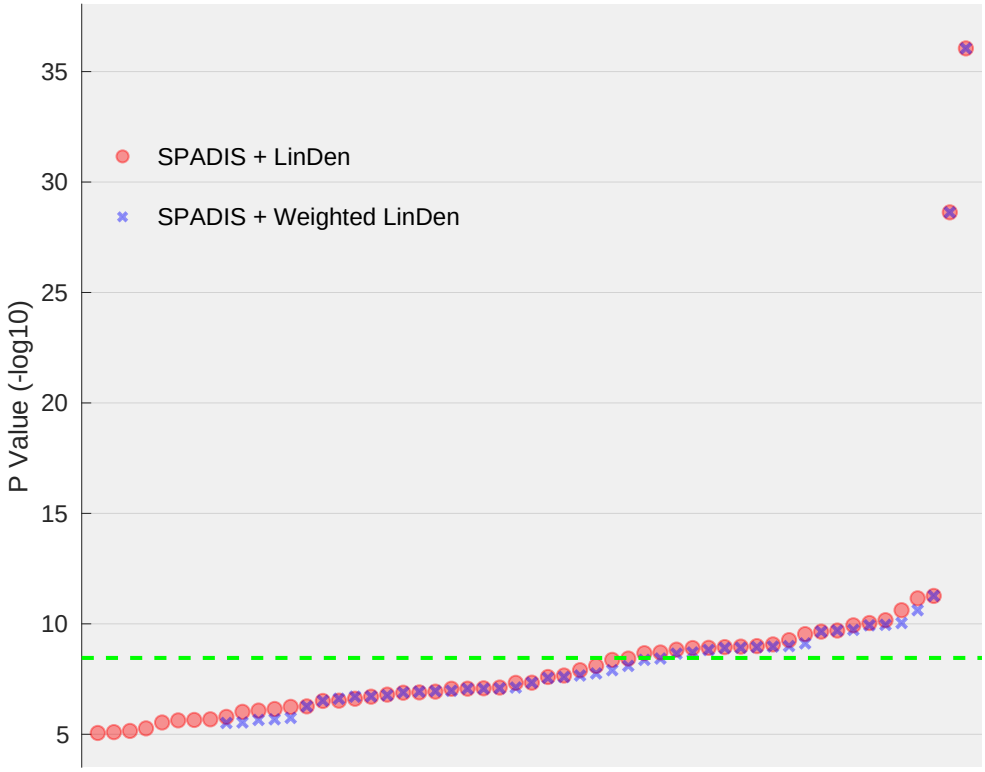

Figure 5.4: On the T2D dataset, for each approach, SPADIS+LinDen, SPADIS+Weighted LinDen, we show the significance levels (y-axis) of each reported pair (dots). Since the difference of the number of tests conducted for each approach is too small we draw average of the Bonferroni corrected significance thresholds (0.1, green line). Bonferroni corrected threshold for SPADIS+LinDen is 8.443 and for SPADIS+Weighted LinDen is 8.470. X – axis is just randomly assigned values to pairs for visualization for $k = 1000$. Integrating regulatory/coding regions to LinDen clearly minimizes FPs.

Figures 5.4, 5.5 and 5.6, demonstrate the decrease in the number of FPs by further guiding of LinDen with regulatory/coding regions. When the SPADIS + weighted LinDen is used, the number of TPs decreases slightly (i.e. a decrease from 21 to 19 for T2D) compared to the original pipeline. However, those TPs that make the difference are accumulated around Bonferroni threshold and we observe a slight difference between the p-values of SPADIS + LinDen dots and

SPADIS + weighted LinDen dots around Bonferroni threshold. On the other side, with this weighted version, we eliminate FPs that are farther away from Bonferroni threshold compared to those TPs in terms of their p-values. Thus, our improvement is not only better in terms of precision but also in terms of gain in the strength of p-values. While the figures clearly illustrate the decrease in FPs (for $k = 1000$), we can observe a trend to decrease in FPs for other k values as well over all datasets in Tables 5.9, 5.10 and 5.11.

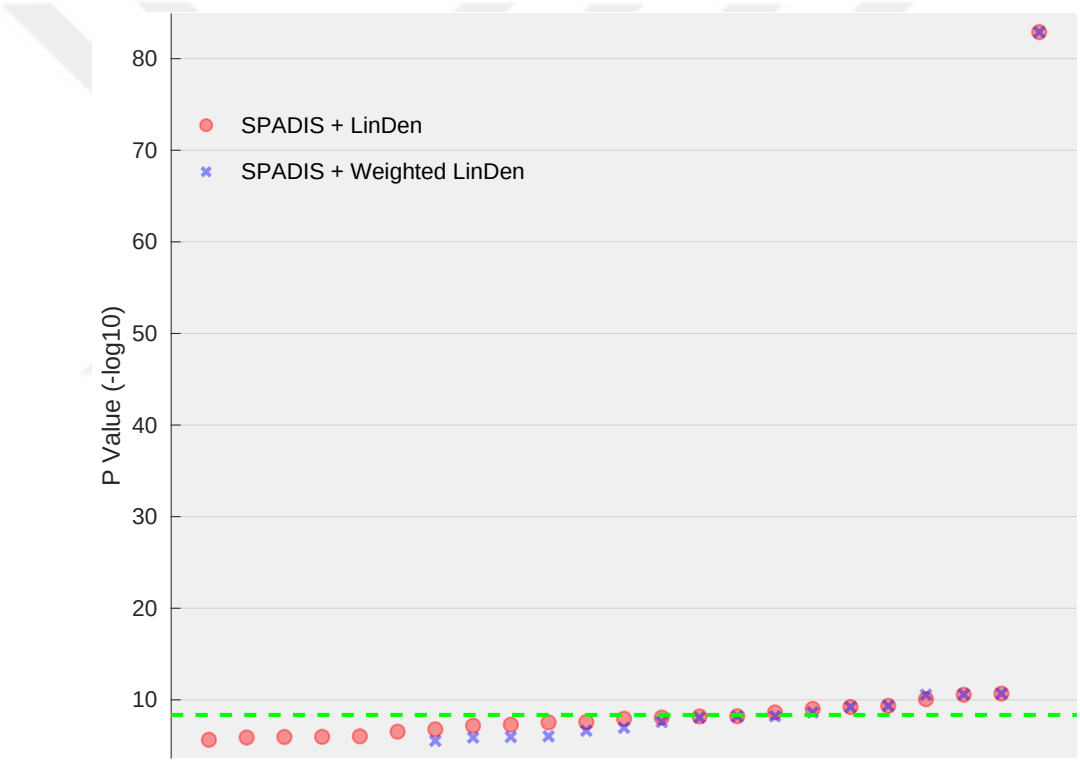

Figure 5.5: On the BD dataset, for each approach, SPADIS+LinDen, SPADIS+Weighted LinDen, we show the significance levels (y-axis) of each reported pair (dots). Since the difference of the number of tests conducted for each approach is too small we draw average of the Bonferroni corrected significance thresholds (0.1, green line). Bonferroni corrected threshold for SPADIS+LinDen is 8.342 and for SPADIS+Weighted LinDen is 8.369. X – axis is just randomly assigned values to pairs for visualization for $k = 1000$. Integrating regulatory/coding regions to LinDen clearly minimizes FPs.

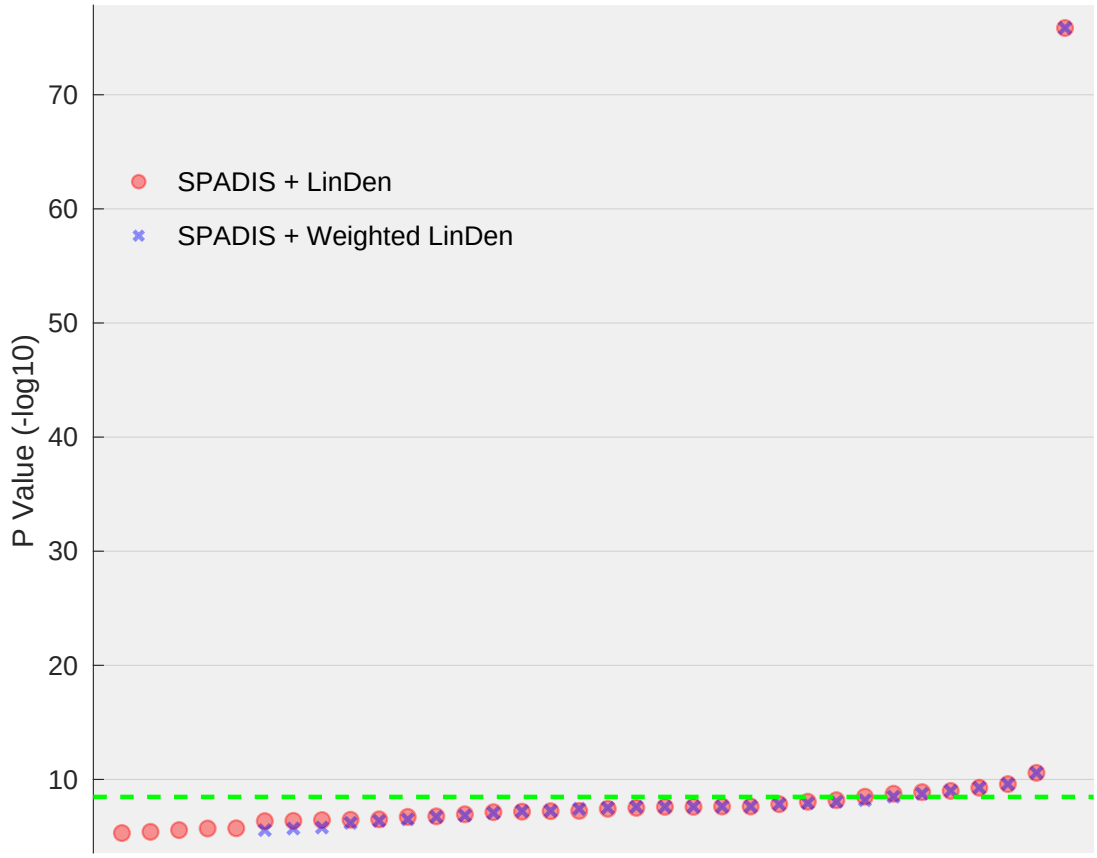

Figure 5.6: On the HT dataset, for each approach, SPADIS+LinDen, SPADIS+Weighted LinDen, we show the significance levels (y-axis) of each reported pair (dots). Since the difference of the number of tests conducted for each approach is too small we draw average of the Bonferroni corrected significance thresholds (0.1, green line). Bonferroni corrected threshold for SPADIS+LinDen is 8.454 and for SPADIS+Weighted LinDen is 8.485. X – axis is just randomly assigned values to pairs for visualization for $k = 1000$. Integrating regulatory/coding regions toLinDenclearly minimizes FPs.

5.5.3 Integrating SPADIS+LinDen with Regulatory and Coding Regions

We analyze the effect of regulatory/coding elements when those regions are integrated into the both stages of the pipeline: SPADIS and LinDen. In the previous two subsections, the effect of these integrations to the pipeline separately are observed. Here, using the pipeline, we pass the SNPs selected by the version of SPADIS which rewards regulatory/coding regions to the weighted LinDen which also highlights those regulatory/coding regions more in terms of marking them as potentially epistatic. We compare this approach with best performance of the original SPADIS + LinDen and SPADIS + weighted LinDen for each k value. The results for this approach are demonstrated in Tables 5.12, 5.13 and 5.14 for T2D, BD and HT datasets respectively.

The results demonstrate that integrating regulatory/coding regions improves the precision for T2D, BD and HT datasets up to 48% ($k = 750$), 82% ($k = 750$) and 32% ($k = 500$) respectively. However, for some k values especially on T2D and HT datasets, the results get worse such that the decrease in precision is up to 54% ($k = 1000$) and 60% ($k = 1500$) respectively. When the regions are integrated into both algorithms SPADIS and LinDen, the pipeline is over-constrained such that it is forced to select SNPs largely from those regions. Thus, integrating only LinDen with regulatory/coding regions is more advantageous compared with this approach considering the number of improved cases and the decrease percentage in precision values.

Table 5.12: Results for T2D dataset for SPADIS + LinDen with both integrated with regulatory/coding regions. Number of pairs reported is the total number of reciprocally significant pairs returned by LinDen with the guidance of SPADIS with rewarded regulatory/coding regions for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen and SNPs in regulatory/coding regions are weighted. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. The green cells show that the k values of regulatory/coding regions increases the precision as compared with the previous pipeline versions.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
SPADIS + LinDen	*k = 500*	9250	15 (6)	0.400
	k = 750	14146	26 (12)	0.462
	k = 1000	18943	48 (9)	0.188
	k = 1500	28014	86 (15)	0.174
	k = 2000	37927	103 (16)	0.155

Table 5.13: Results for BD dataset for SPADIS + LinDen with both integrated with regulatory/coding regions. Number of pairs reported is the total number of reciprocally significant pairs returned by Linden with the guidance of SPADIS with rewarded regulatory/coding regions for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to Linden and SNPs in regulatory/coding regions are weighted. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. The green cells show the k values that the guidance of regulatory/coding regions increases the precision as compared with the previous pipeline versions.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
SPADIS + LinDen	*k = 500*	9387	21 (4)	0.191
	k = 750	14098	23 (11)	0.478
	k = 1000	17048	32 (15)	0.469
	k = 1500	28161	59 (19)	0.322
	k = 2000	36323	49 (14)	0.286

Table 5.14: Results for HT dataset for SPADIS + LinDen with both integrated with regulatory/coding regions. Number of pairs reported is the total number of reciprocally significant pairs returned by LinDen with the guidance of SPADIS with rewarded regulatory/coding regions for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to LinDen and SNPs in regulatory/coding regions are weighted. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by each method. The green cells show the k values that the guidance of regulatory/coding regions increases the precision as compared with the previous pipeline versions.

Method		**# Tested Loci**	**# Pairs Reported**	**Precision**
SPADIS + LinDen	*k = 500*	8744	17 (6)	0.353
	k = 750	11886	23 (7)	0.304
	k = 1000	18953	26 (7)	0.269
	k = 1500	27275	102 (5)	0.049
	k = 2000	37979	125 (9)	0.072

We also illustrate the power of integrating regulatory/coding regions into both SPADIS and LinDen in terms of discovering more significant epistatic pairs by comparing it with the default pipeline in Figures 5.7, 5.8, 5.9 for the T2D, BD and HT datasets, respectively (for $k = 500$). To highlight the false positives in the figures, the p-values of reported pairs are sorted in ascending order. Since Bonferroni corrected thresholds of each approach are too tight that cannot be distinguished in the figure, we draw the Bonferroni threshold as the average of both thresholds. The green line denotes the Bonferroni corrected significance level to be passed as the average of the actual thresholds of each approach. The actual Bonferroni corrected thresholds are indicated in the respective figure caption on each dataset.

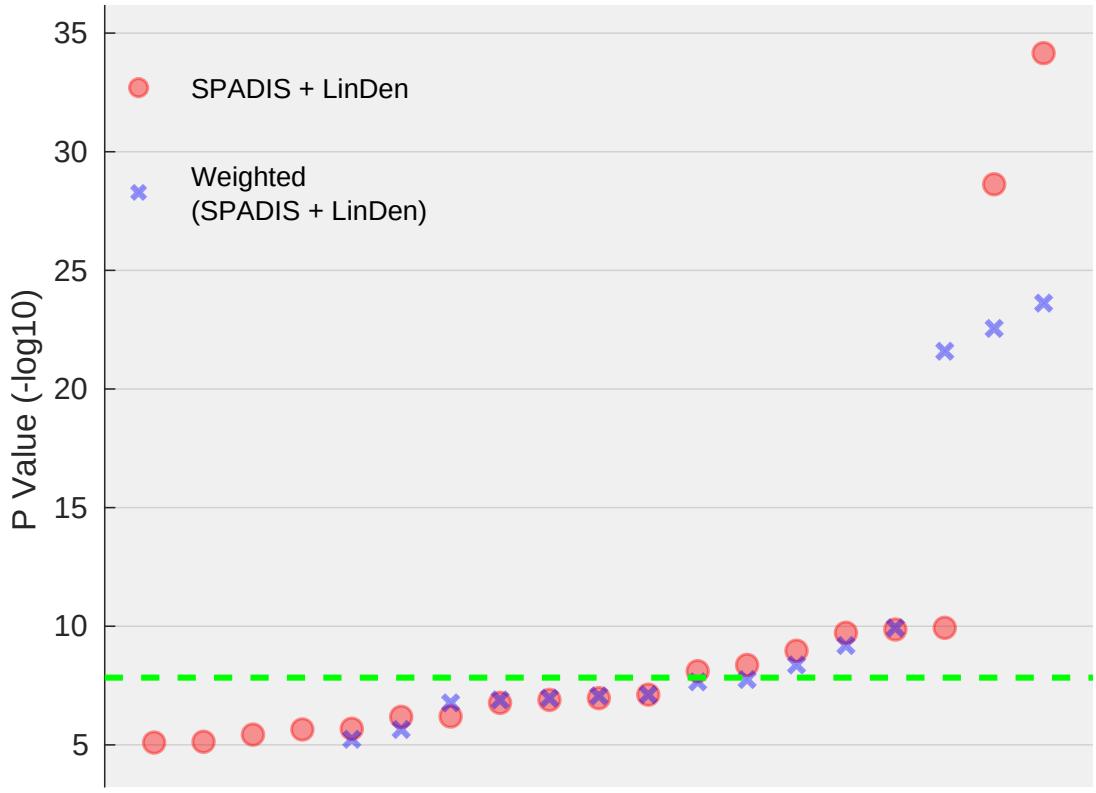

Figure 5.7: On the BD dataset, for each approach, SPADIS+LinDen, SPADIS+Weighted LinDen, we show the significance levels (y-axis) of each reported pair (dots). Since the difference of the number of tests conducted for each approach is too small we draw average of the Bonferroni corrected significance thresholds (0.1, green line). Bonferroni corrected threshold for SPADIS+LinDen is 7.804 and for Weighted (SPADIS + LinDen) is 7.866. X – axis is just randomly assigned values to pairs for visualization for $k = 500$. Integrating regulatory/coding regions to LinDen clearly minimizes FPs.

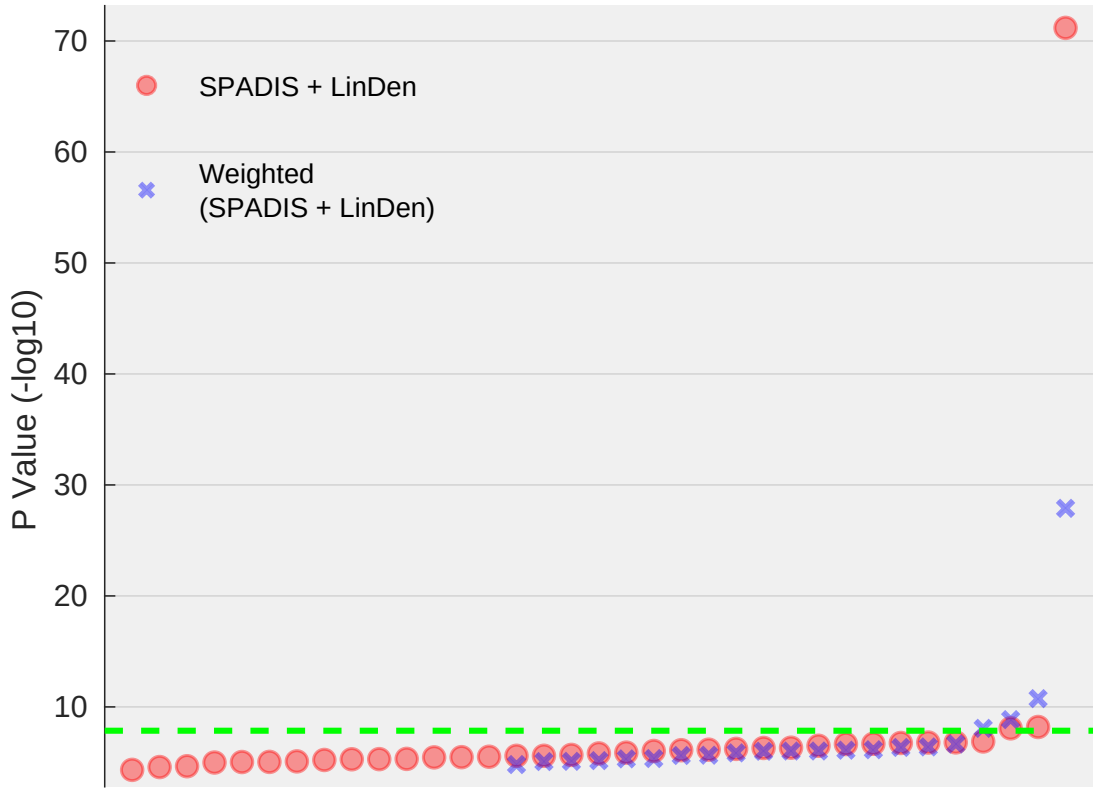

Figure 5.8: On the BD dataset, for each approach, SPADIS+LinDen, SPADIS+Weighted LinDen, we show the significance levels (y-axis) of each reported pair (dots). Since the difference of the number of tests conducted for each approach is too small we draw average of the Bonferroni corrected significance thresholds (0.1, green line). Bonferroni corrected threshold for SPADIS+LinDen is 7.821 and for Weighted (SPADIS + LinDen) is 7.891. X – axis is just randomly assigned values to pairs for visualization for $k = 500$. Integrating regulatory/coding regions to LinDen clearly minimizes FPs.

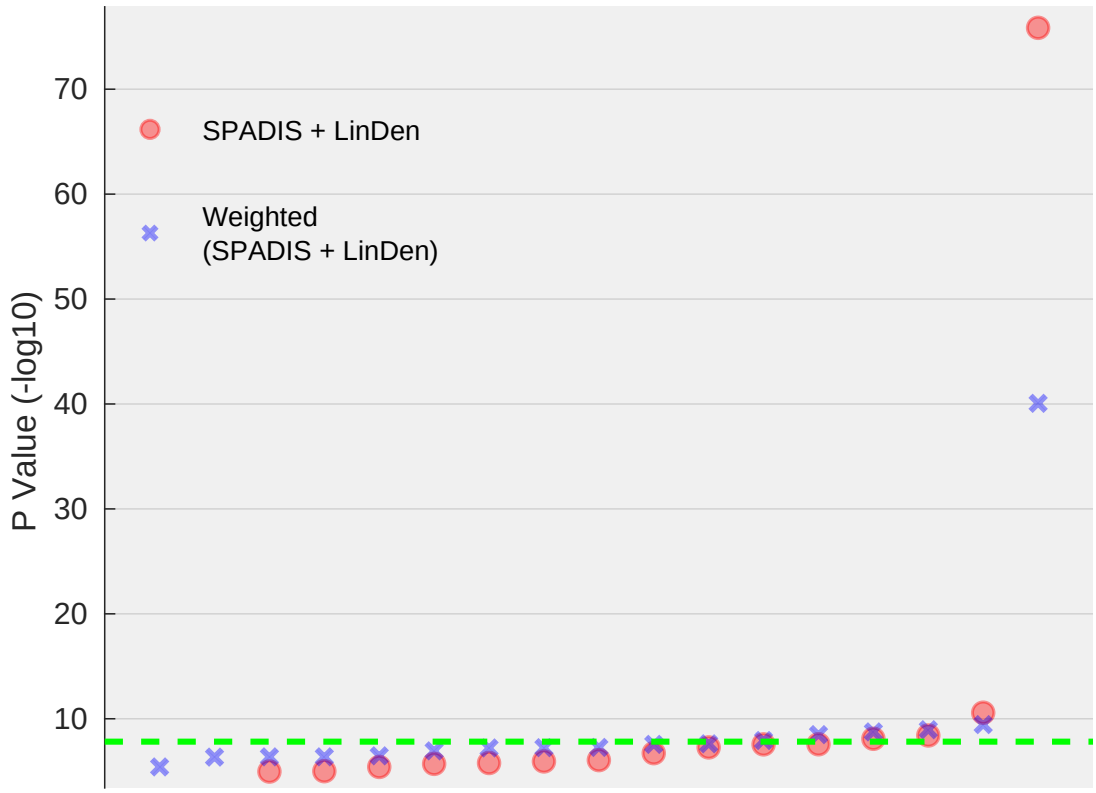

Figure 5.9: On the HT dataset, for each approach, SPADIS+LinDen, Weighted SPADIS+ WeightedLinDen, we show the significance levels (y-axis) of each reported pair (dots). Since the difference of the number of tests conducted for each approach is too small we draw average of the Bonferroni corrected significance thresholds (0.1, green line). Bonferroni corrected threshold for SPADIS+LinDen is 7.557 and for Weighted (SPADIS + LinDen) is 7.887. X – axis is just randomly assigned values to pairs for visualization for $k = 500$.

5.6 Runtime Improvement

Next, we compare the runtime of each approach on the three datasets both as CPU time and clock-time. To assess the time performance, Intel(R) Xeon(R) CPU E5-2650 v3 with a 2×2.30GHz processors with 251 GB RAM is used in parallel setting. We run each approach on these processors and present the results in Tables 5.15, 5.16, 5.17 for T2D, BD and HT datasets, respectively. Results indicate that in all the datasets, the pipeline outperforms LinDen in terms of runtime. We report an almost 4-fold decrease in runtime. Integrating regulatory and coding regions to the pipeline does not change the time complexity of both SPADIS and LinDen, thus the pipeline.

Table 5.15: Runtime comparison for T2D dataset

		T2D	
Method		**CPU Time (log10)**	**Run Time**
LinDen		4.54	01:04:53
SPADIS + LinDen	*k = 500*	3.73	00:17:39
	k = 750	3.67	00:24:40
	k = 1000	3.79	00:16:50
	k = 1500	3.85	00:28:18
	k = 2000	3.88	00:30:20

Table 5.16: Runtime comparison for BD dataset

		BD	
Method		**CPU Time (log10)**	**Run Time**
LinDen		4,67	01:12:23
SPADIS + LinDen	*k = 500*	3.78	00:14:41
	k = 750	3.66	00:26:25
	k = 1000	3.85	00:16:28
	k = 1500	3.76	00:27:34
	k = 2000	3.82	00:31:41

Table 5.17: Runtime comparison for HT dataset

		HT	
Method		**CPU Time (log10)**	**Run Time**
LinDen		4,67	01:12:21
SPADIS + LinDen	*k = 500*	3.79	00:14:43
	k = 750	3.66	00:26:14
	k = 1000	3.86	00:17:26
	k = 1500	3.77	00:29:24
	k = 2000	3.83	00:32:11

5.7 Sanity Checks

5.7.1 Guiding LinDen with random SNPs

We tested if SPADIS really helps selecting complementary SNPs. Instead of inputting SPADIS-selected-SNP-set to LinDen, we input a randomly selected set of SNPs with their nearest 9 upstream and 9 downstream SNPs on the genome (matching number of SNPs, 10 run per each k value). Note that random SNPs are also expected to be away from each other on average. We compare the randomization with each of the proposed pipeline versions. The distribution of precision values per these runs are given in Figures 5.10, 5.11, 5.12 for T2D, BD and HT datasets respectively. These plots show SPADIS' guidance results in better precision values with just a few FPs compared to randomization.

The pipeline performs much better than randomization in 13 comparisons out of 15, but we observe one comparable ($k = 750$) and one poor ($k = 500$) result in BD dataset. When we integrate regulatory/coding regions into the pipeline, we observe that the precision values increase substantially especially for BD dataset from 0.238 to 0.478 (82%). Also, as can be observed from Figure 5.8, weighted (SPADIS+LinDen) is able to find more significant pairs while the number of FPs decrease substantially. This may indicate that the mutations on regulatory/coding regions contribute to bipolar disorder more compared with T2D and

HT datasets and integration of those region to SPADIS increases the explanatory power of the algorithm in terms of relating the variants to the bipolar disorder.

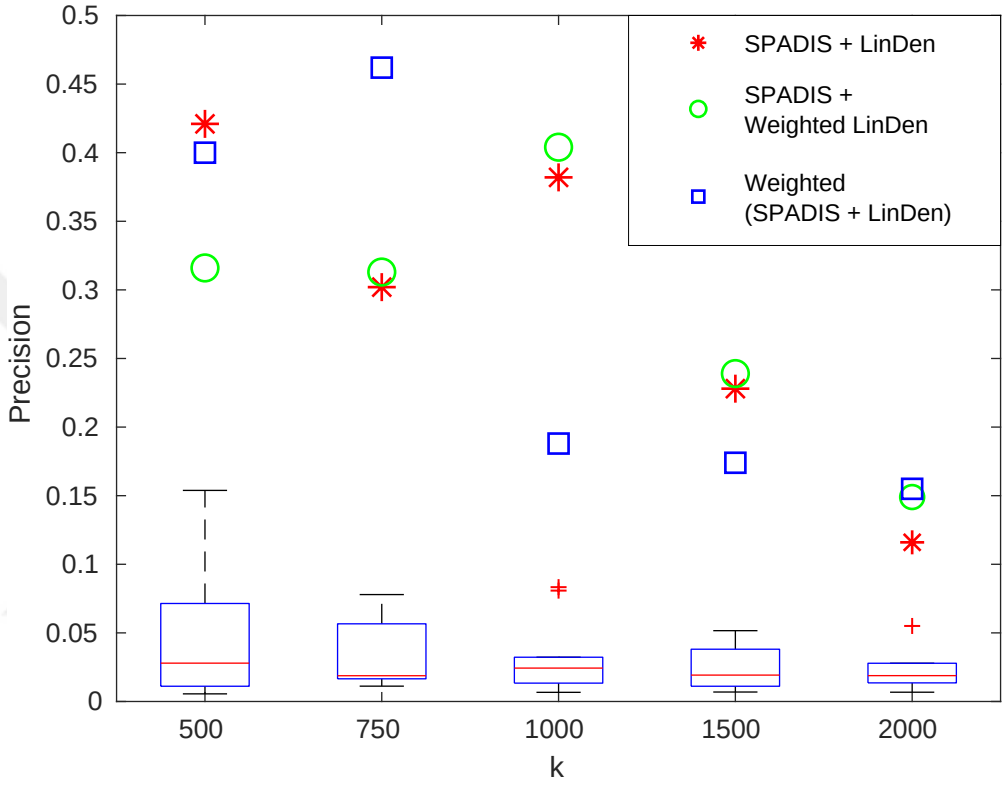

Figure 5.10: Box plot of T2D dataset shows the distribution of the precision values attained by LinDen when input by k random SNPs (10 runs). Star, circle and square indicate the value attained by the original pipeline, the pipeline with only LinDen integrated with regulatory/coding regions and the pipeline with both SPADIS and LinDen integrated with regulatory/coding regions respectively given k.

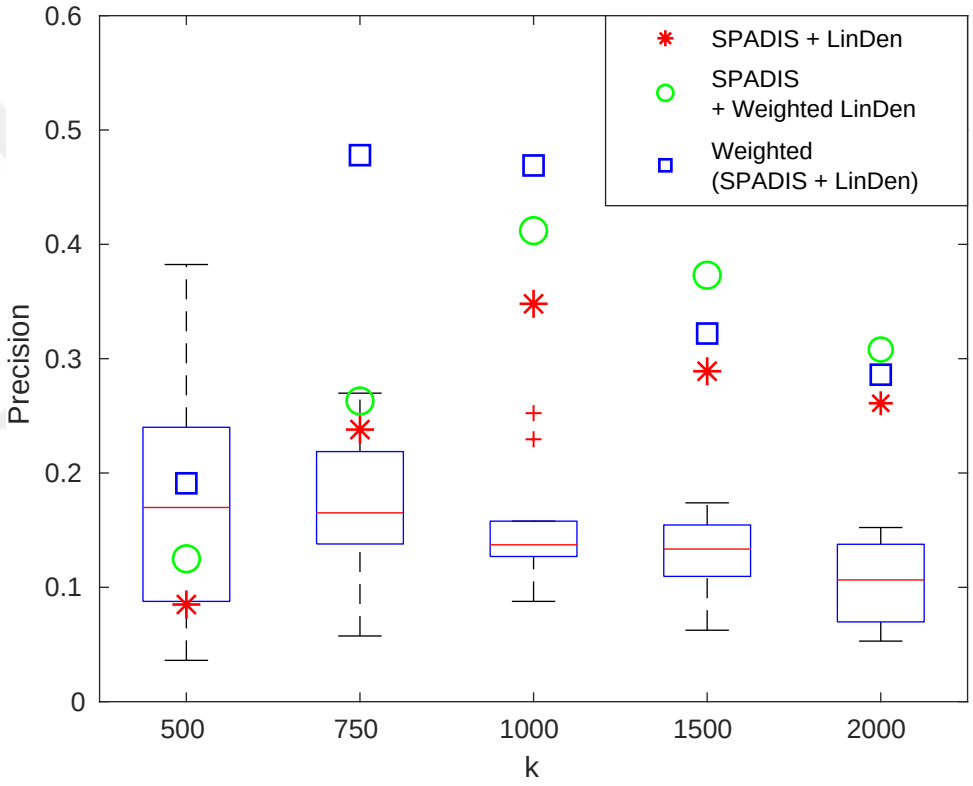

Figure 5.11: Box plot of BD dataset shows the distribution of the precision values attained by LinDen when input by k random SNPs (10 runs). Star, circle and square indicate the value attained by the original pipeline, the pipeline with only LinDen integrated with regulatory/coding regions and the pipeline with both SPADIS and LinDen integrated with regulatory/coding regions respectively given k.

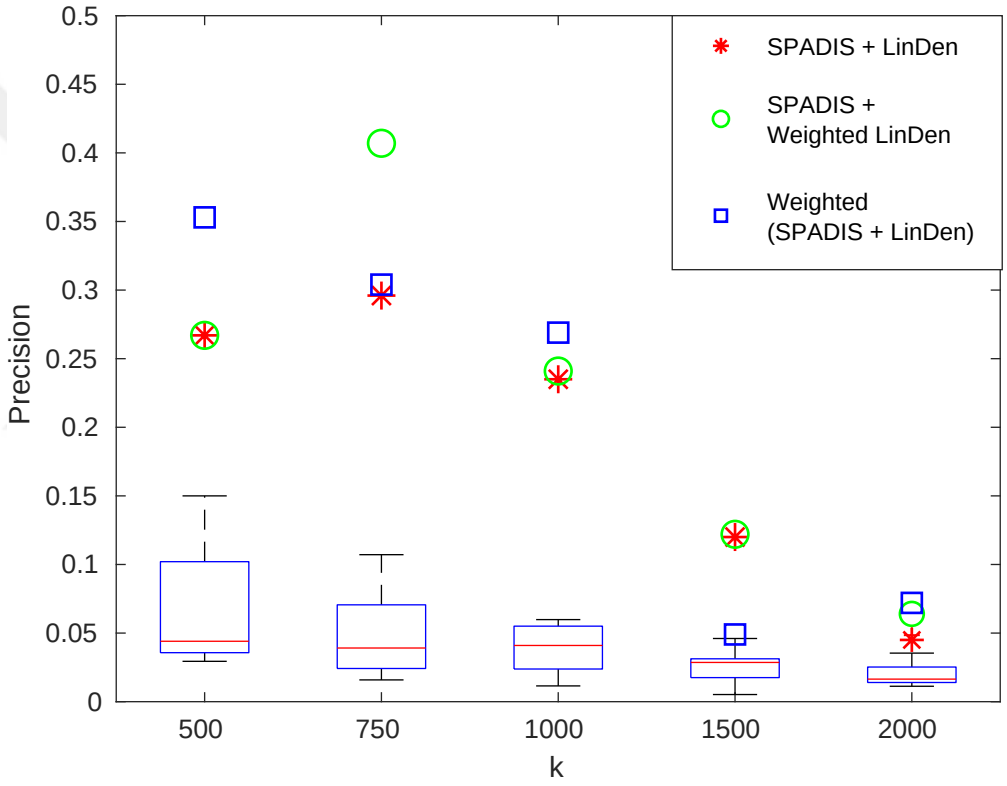

Figure 5.12: Box plot of HT dataset shows the distribution of the precision values attained by LinDen when input by k random SNPs (10 runs). Star, circle and square indicate the value attained by the original pipeline, the pipeline with only LinDen integrated with regulatory/coding regions and the pipeline with both SPADIS and LinDen integrated with regulatory/coding regions respectively given k.

5.7.2 Using SPADIS as an epistasis detection tool

We checked whether we would get better results if we directly test every pair SPADIS returned. For each dataset and each $k \in \{500, 750, 1000, 1500, 2000\}$ value, we performed standard pairwise exhaustive epistasis tests on all SPADIS-selected SNPs (i.e., LINDEN with $d = 0$). In each dataset, LINDEN only returned 1 reciprocally significant pair. Even though that pair passes the Bonferroni corrected significance threshold (0.1), directly using SPADIS remains very weak method in terms of number of true positives. Thus, we conclude that SPADIS' guidance helps LINDEN in finding significant pairs of epistatic pairs while minimizing false positives, but using it as an epistasis proioritization tool is not feasible. We also need LINDEN for further pruning.

5.7.3 Adjusting Parameter of LINDEN to Limit False Positives

We also checked if guidance of SPADIS is better than adjusting the parameter d of LINDEN to make it more conservative (i.e., smaller number of tests). While increasing d to 0.9 (instead of default 0.45) results in a similar number of reciprocally significant epistatic pairs as in the LINDEN with $d = 0.45$, but fewer number of pairs pass the threshold and LINDEN still needs too many tests (in the order of hundreds) compared to our pipeline. The performance results for three different d values are presented in Table 5.18. These results demonstrate that using SPADIS helps LINDEN not only in terms of decreasing the number of tests but also pinpointing significant pairs.

Table 5.18: LinDen results for T2D, BD and HT datasets. Number of pairs reported is the total number of reciprocally significant pairs returned by LinDen. The number in parentheses denotes the significant pairs passing significance threshold (0.1) after Bonferroni correction based on the number of tests performed by LinDen for each dataset. Table shows that conservatization of LinDen does not improve the precision as SPADIS does.

	d					
	0.45		**0.9**		**0.99**	
Dataset	**# Pairs Reported**	**Precision**	**# Pairs Reported**	**Precision**	**# Pairs Reported**	**Precision**
T2D	1786 (5)	0.0028	1717 (3)	0.0017	768 (5)	0.0065
BD	906 (30)	0.0331	669 (8)	0.0120	323 (9)	0.0279
HT	1135 (5)	0.0044	683 (2)	0.0029	328 (4)	0.0122

Chapter 6

Conclusion and Discussion

In this study, we develop a pipeline to prioritize epistasis test while minimizing the number of false positive tests. In our hypothesis, we claim that selecting diverse and explanatory SNPs would create a SNP set which contains pairs that are likely to be epistatic and yield a pruned search space for epistasis detection algorithms. We use LinDen as fast epistasis tool to perform epistasis test on this reduced search space. Via feature selection mechanism of SPADIS which eliminates redundant SNP pairs, i.e. SNPs in linkage disequilibrium, we aim to reduce the number of false discoveries returned by LinDen.

Our pipeline avoids testing SNP pair in LD at both stages of the algorithm. In the SPADIS stage, we avoid selecting SNPs that are close to each other as SPADIS diversifies genomic locations of the SNPs while awarding their individual associations with the phenotype. In the LinDen stage, even though we include neighbors of those SPADIS-selected SNPs, we avoid testing nearby SNPs by utilizing the LD-tree structure of LinDen. Thus, the pipeline tests more-likely-epistatic SNP pairs and avoids false positives.

The size of the search space is an another major challenge. With an exhaustive approach, half million tests are required to process thousand SNPs. This sheer

number of required tests causes two major problems: (i) Decrease in the statistical power of those tests, and (ii) Computational inefficiency. Correcting for multiple hypothesis testing results in highly conservative thresholds that lead to missing true positives. Our pipeline provides a solution for both problems by pruning the search space which leads to more relaxed significance thresholds that also comes with much shorter processing times. Thus, our algorithm is suitable for large datasets which may contain millions of interactions to be considered.

Statistical significance may not indicate biological significance. For this reason incorporation of biological knowledge to the algorithm improves the power of epistasis detection tool in terms of interpreting the results. Promoting selection of regulatory regions could enable biological interpretation and coding regions can help interpretation of the affected functionality by the reported epistatic pairs in addition to promoting the selection of more statistically significant pairs.

We have shown that our proposed two-stage pipeline is able to detect more significant epistatic pairs while minimizing the false positive findings. The approach is computationally efficient and suitable for large datasets.

Bibliography

[1] T. Manolio, F. S Collins, N. J Cox, D. Goldstein, L. A Hindorff, D. J Hunter, M. I McCarthy, E. M Ramos, L. Cardon, A. Chakravarti, J. H Cho, A. E Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. Rotimi, M. Slatkin, D. Valle, A. S Whittemore, and P. M Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–53, 10 2009.

[2] X. Wang, A. Q. Fu, M. E. McNerney, and K. P. White, "Widespread genetic epistasis among cancer genes.," *Nature communications*, vol. 5, p. 4828, 2014.

[3] M. Slatkin, "Linkage disequilibrium - understanding the evolutionary past and mapping the medical future," *Nature reviews. Genetics*, vol. 9, pp. 477–85, 07 2008.

[4] X. Zhang, S. Huang, F. Zou, and W. Wang, "TEAM: efficient two-locus epistasis tests in human genome-wide association study," *Bioinformatics*, vol. 26, pp. i217–i227, 06 2010.

[5] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. Tang, and W. Yu, "Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *American journal of human genetics*, vol. 87, pp. 325–40, 09 2010.

[6] C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu, "Snpharvester: A filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics (Oxford, England)*, vol. 25, pp. 504–11, 02 2009.

[7] X. Wan, C. Yang, Q. Yang, H. Xue, N. Tang, and W. Yu, "Predictive rule inference for epistatic interaction detection in genome-wide association studies," *Bioinformatics (Oxford, England)*, vol. 26, pp. 30–7, 10 2009.

[8] L.-Y. Chuang, H.-W. Chang, M.-C. Lin, and C.-H. Yang, "Improved branch and bound algorithm for detecting snp-snp interactions in breast cancer," *Journal of Clinical Bioinformatics*, vol. 3, p. 4, Feb 2013.

[9] J. Piriyapongsa, C. Ngamphiw, A. Intarapanich, S. Kulawonganunchai, A. Assawamakin, C. Bootchai, P. J. Shaw, and S. Tongsima, "iloci: a snp interaction prioritization technique for detecting epistasis in genome-wide association studies," *BMC Genomics*, vol. 13, p. S2, Dec 2012.

[10] B. A. McKinney and N. M. Pajewski, "Six degrees of epistasis: Statistical network models for gwas," in *Front. Gene.*, 2011.

[11] P. Holmans, E. K. Green, J. S. Pahwa, M. A. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan, and N. Craddock, "Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder," *The American Journal of Human Genetics*, vol. 85, no. 1, pp. 13 – 24, 2009.

[12] L. Weng, F. Macciardi, A. Subramanian, G. Guffanti, S. G. Potkin, Z. Yu, and X. Xie, "Snp-based pathway enrichment analysis for genome-wide association studies," *BMC Bioinformatics*, vol. 12, p. 99, Apr 2011.

[13] Y. Liu, S. Maxwell, T. Feng, X. Zhu, R. C. Elston, M. Koyutürk, and M. R. Chance, "Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from gwas data," *BMC Systems Biology*, vol. 6, p. S15, Dec 2012.

[14] S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. Uitdehaag, L. Kappos, G. Consortium, C. H. Polman, P. M. Matthews, S. L. Hauser, R. A. Gibson, J. R. Oksenberg, and M. R. Barnes, "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis," *Human Molecular Genetics*, vol. 18, pp. 2078–2090, 03 2009.

[15] M. Ayati and M. Koyutürk, "Prioritization of genomic locus pairs for testing epistasis," in *BCB*, 2014.

[16] M. Ayati and M. Koyutürk, "Pocos: Population covering locus sets for risk assessment in complex diseases," in *PLoS Computational Biology*, 2016.

[17] T. Cowman and M. Koyutürk, "Prioritizing tests of epistasis through hierarchical representation of genomic redundancies," *Nucleic Acids Research*, vol. 45, pp. e131–e131, 06 2017.

[18] S. Yilmaz, O. Tastan, and A. E. Cicek, "Spadis: An algorithm for selecting predictive and diverse snps in gwas," *bioRxiv*, 2018.

[19] W. Bateson and G. Mendel, *Mendel's Principles of Heredity: A Defence, with a Translation of Mendel's Original Papers on Hybridisation*. Cambridge Library Collection - Darwin, Evolution and Genetics, Cambridge University Press, 2009.

[20] R. A. Fisher, "Xv.—the correlation between relatives on the supposition of mendelian inheritance.," *Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, p. 399–433, 1919.

[21] H. J. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans," *Human Molecular Genetics*, vol. 11, pp. 2463–2468, 10 2002.

[22] H. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nature reviews. Genetics*, vol. 10, pp. 392–404, 06 2009.

[23] P. Phillips, "Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems," *Nature reviews. Genetics*, vol. 9, pp. 855–67, 11 2008.

[24] Y. Chung, S. Y. Lee, R. C. Elston, and T. Park, "Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions," *Bioinformatics*, vol. 23 1, pp. 71–6, 2007.

[25] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I W de Bakker, M. Daly, and P. C Sham, "Plink: A tool set for whole-genome association and population-based linkage analyses," *American journal of human genetics*, vol. 81, pp. 559–75, 10 2007.

[26] J. H. Moore, "Computational analysis of gene-gene interactions using multifactor dimensionality reduction," *Expert Review of Molecular Diagnostics*, vol. 4, no. 6, pp. 795–803, 2004.

[27] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions," *Bioinformatics*, vol. 19, pp. 376–382, 02 2003.

[28] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.," *American journal of human genetics*, vol. 69 1, pp. 138–47, 2001.

[29] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, pp. 1167–1173, 2007.

[30] J. Gayán, A. González-Pérez, F. Bermudo, M. E. Sáez, J. L. Royo, A. Quintas, J. J. Galan, F. J. Morón, R. Ramirez-Lorca, L. M. Real, and A. Ruiz, "A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis," *BMC Genomics*, vol. 9, p. 360, Jul 2008.

[31] P. Kraft, Y.-C. Yen, D. O. Stram, J. T. Morrison, and W. J. Gauderman, "Exploiting gene-environment interaction to detect genetic associations.," *Human heredity*, vol. 63 2, pp. 111–9, 2007.

[32] D. E Reich, M. Cargill, S. Bolk, J. Ireland, P. Sabeti, D. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S Lander, "Linkage disequilibrium in the human genome," *Nature*, vol. 411, 06 2001.

[33] R. C. Lewontin, "The interaction of selection and linkage. i. general considerations; heterotic models," *Genetics*, vol. 49, no. 1, pp. 49–67, 1964.

[34] W. G. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *Theoretical and Applied Genetics*, vol. 38, pp. 226–231, Jun 1968.

[35] H. J. Cordell and D. G. Clayton, "Genetic association studies," *The Lancet*, vol. 366, no. 9491, pp. 1121 – 1131, 2005.

[36] D. P. Doolittle, *Conditions for Hardy-Weinberg Equilibrium*, pp. 8–11. Berlin, Heidelberg: Springer Berlin Heidelberg, 1987.

[37] J. Alghamdi and S. Padmanabhan, "Chapter 12 - fundamentals of complex trait genetics and association studies," in *Handbook of Pharmacogenomics and Stratified Medicine* (S. Padmanabhan, ed.), pp. 235 – 257, San Diego: Academic Press, 2014.

[38] D. S. Latchman, "Transcription factors: An overview," *The International Journal of Biochemistry Cell Biology*, vol. 29, no. 12, pp. 1305 – 1312, 1997.

[39] R. Grosschedl, "Enhancers," in *Encyclopedia of Genetics* (S. Brenner and J. H. Miller, eds.), pp. 624 – 625, New York: Academic Press, 2001.

[40] M. Kumar, R. DeVaux, and J. Herschkowitz, "Chapter thirteen - molecular and cellular changes in breast cancer and new roles of lncrnas in breast cancer initiation and progression," in *Molecular and Cellular Changes in the Cancer Cell* (K. Pruitt, ed.), vol. 144 of *Progress in Molecular Biology and Translational Science*, pp. 563 – 586, Academic Press, 2016.

[41] A. Visel, E. M. Rubin, and L. A. Pennacchio, "Genomic views of distant-acting enhancers," in *Nature*, 2009.

[42] T. Wang, X. Zhu, and R. C. Elston, "Improving power in contrasting linkage-disequilibrium patterns between cases and controls," *The American Journal of Human Genetics*, vol. 80, no. 5, pp. 911 – 920, 2007.

[43] M. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "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.

[44] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical Programming*, vol. 14, pp. 265–294, Dec 1978.

[45] N. J. Craddock, M. E. Hurles, N. J. Cardin, R. D. Pearson, V. Plagnol, S. C. Robson, D. V. Vukcevic, C. Barnes, D. F. Conrad, E. Giannoulatou, C. R. Holmes, J. Marchini, K. Stirrups, M. D. Tobin, L. V. Wain, C. Yau, J. Aerts, T. Ahmad, T. D. Andrews, H. Arbury, A. P. Attwood, A. Auton, S. Ball, A. J. Balmforth, J. C. Barrett, I. Barroso, A. Barton, A. J. Bennett, S. Bhaskar, K. Blaszczyk, J. Bowes, O. J. Brand, P. S. Braund, F. Bredin, G. Breen, M. Brown, I. N. Bruce, J. Bull, O. S. Burren, J. Burton, J. K. Byrnes, S. A. Caesar, C. Clee, A. J. Coffey, J. M. C. Connell, J. P. D. Cooper, A. F. Dominiczak, K. Downes, H. E. Drummond, D. Dudakia, A. Dunham, B. Ebbs, D. R. E. Eccles, S. Edkins, C. Edwards, A. Elliot, A. Doria, D. M. Evans, G. L. Evans, S. de Eyre, A. Farmer, I. N. Ferrier, L. Feuk, T. W. Fitzgerald, E. W. Flynn, A. H. Forbes, L. Forty, J. A. Franklyn, R. M. Freathy, P. Gibbs, P. B. Gilbert, O. Gokumen, K. Gordon-Smith, E. S. Gray, E. K. Green, C. J. Groves, D. Grozeva, R. Gwilliam, A. S. Hall, N. Hammond, M. Hardy, P. Harrison, N. Hassanali, H. Hebaishi, S. A. Hines, A. Hinks, G. A. Hitman, L. J. Hocking, E. K. Howard, P. Howard, J. M. M. Howson, D. Hughes, S. B. Hunt, J. D. Isaacs, M. Jain, D. P. Jewell, T. Johnson, J. Jolley, I. Jones, L. A. Jones, G. Kirov, C. Langford, H. Lango-Allen, G. M. Lathrop, J. Lee, K. L. Lee, C. W. Lees, K. Lewis, C. M. Lindgren, M. Maisuria-Armer, J. B. Maller, J. C. Mansfield, P. Martin, D. C. O. Massey, W. L. Mcardle, P. McGuffin, K. E. McLay, A. J. Mentzer, M. L. Mimmack, A. Morgan, A. P. Morris, C. Mowat, S. R. Myers, W. G. Newman, E. R. Nimmo, M. C. O'Donovan, A. K. Onipinla, I. Onyiah, N. R. Ovington, M. J. Owen, K. Palin, K. S. Parnell, D. Pernet, J. Perry, A. D. Phillips, D. Pinto, N. J. Prescott, I. Prokopenko, M. A. Quail, S. E. Rafelt, N. W. Rayner, R. Redon, D. M. Reid, A. Renwick, S. M. Ring, N. P. Robertson, E. Russell, D. S. Clair, J. G. Sambrook, J. A. Sanderson, H. Schuilenburg, C. E. Scott, R. J. Scott, S. Seal, S. Shaw-Hawkins, B. M. Shields, M. J. Simmonds, D. J. Smyth, E. Somaskantharajah, K. Spanova, S. Steer, J. Stephens, H. Stevens, M. Stone, Z. Su, D. Symmons, J. R. Thompson,

W. Thomson, M. E. Travers, C. Turnbull, A. Valsesia, M. C. Walker, N. M. Walker, C. Wallace, M. Warren-Perry, N. A. Watkins, J. A. Webster, M. N. Weedon, A. G. Wilson, M. Woodburn, B. P. Wordsworth, A. Young, E. Zeggini, N. P. Carter, T. M. Frayling, C. H. Lee, G. McVean, P. B. Munroe, A. Palotie, S. J. Sawcer, S. W. Scherer, D. P. Strachan, C. Tyler-Smith, M. A. Brown, P. R. Burton, M. J. Caulfield, A. D. Compston, M. Farrall, S. C. L. Gough, A. S. Hall, A. T. Hattersley, A. V. Hill, C. G. Mathew, M. Pembrey, J. Satsangi, M. R. Stratton, J. Worthington, P. Deloukas, A. Duncanson, D. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, M. Parkes, N. Rahman, J. A. Todd, N. J. Samani, and P. Donnelly, "Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls," 2010.

[46] C. Laurie, K. Doheny, D. B Mirel, E. Pugh, L. J Bierut, T. Bhangale, F. Boehm, N. Caporaso, M. C Cornelis, H. Edenberg, S. B Gabriel, E. L Harris, F. B Hu, K. Jacobs, P. Kraft, M. Teresa Landi, T. Lumley, T. Manolio, C. McHugh, and B. Weir, "Quality control and quality assurance in genotypic data for genome-wide association studies," *Genetic epidemiology*, vol. 34, pp. 591–602, 09 2010.

[47] P. R Burton, D. G Clayton, L. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P Kwiatkowski, M. I McCarthy, W. Ouwehand, N. J Samani, J. A Todd, P. Donnelly, J. C Barrett, D. Davison, D. Easton, D. Evans, H. T Leung, J. L Marchini, A. P Morris, and C. Mathew, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, 01 2007.

[48] A. Visel, S. Minovitsky, I. Dubchak, and L. A Pennacchio, "Vista enhancer browser—a database of tissue-specific human enhancers. nucleic acids res 35:d88-92," *Nucleic acids research*, vol. 35, pp. D88–92, 02 2007.

[49] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L. D. Ward, C. B. Lowe,

A. K. Holloway, M. Clamp, S. Gnerre, J. Alföldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M. J. Hubisz, D. B. Jaffe, I. Jungreis, W. J. Kent, D. Kostka, M. Lara, A. L. Martins, T. Massingham, I. Moltke, B. J. Raney, M. D. Rasmussen, J. Robinson, A. Stark, A. J. Vilella, J. Wen, X. Xie, M. C. Zody, B. I. S. P. Team, W. G. Assembly, J. Baldwin, T. Bloom, C. Whye Chin, D. Heiman, R. Nicol, C. Nusbaum, S. Young, J. Wilkinson, K. C. Worley, C. L. Kovar, D. M. Muzny, R. A. Gibbs, B. C. o. M. H. G. S. C. S. Team, A. Cree, H. H. Dihn, G. Fowler, S. Jhangiani, V. Joshi, S. Lee, L. R. Lewis, L. V. Nazareth, G. Okwuonu, J. Santibanez, W. C. Warren, E. R. Mardis, G. M. Weinstock, R. K. Wilson, G. I. a. W. University, K. Delehaunty, D. Dooling, C. Fronik, L. Fulton, B. Fulton, T. Graves, P. Minx, E. Sodergren, E. Birney, E. H. Margulies, J. Herrero, E. D. Green, D. Haussler, A. Siepel, N. Goldman, K. S. Pollard, J. S. Pedersen, E. S. Lander, and M. Kellis, "A high-resolution map of human evolutionary constraint using 29 mammals," *Nature*, vol. 478, pp. 476 EP –, Oct 2011. Article.

[50] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek, "Ensembl 2018," *Nucleic Acids Research*, vol. 46, pp. D754–D761, 11 2017.