Prioritization of Genomic Locus Pairs for Testing Epistasis

Marzieh Ayati
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, OH
marzieh.ayati@case.edu

Mehmet Koyutürk
(1) Department of Electrical Engineering and Computer Science
(2) Center for Proteomics and Bioinformatics
Case Western Reserve University
Cleveland, OH
mehmet.koyuturk@case.edu

ABSTRACT

In recent years, genome-wide association studies (GWAS) have successfully identified loci that harbor genetic variants associated with complex diseases. However, susceptibility loci identified by GWAS so far generally account for a limited fraction of heritability in patient populations. More recently, there has been considerable attention on identifying epistatic interactions. However, the large number of pairs to be tested for epistasis poses significant challenges, in terms of both computational (run-time) and statistical (multiple hypothesis testing) considerations.

In this paper, we propose a new method to reduce the number of tests required to identify epistatic pairs of genomic loci. The key idea of the proposed algorithm is to reduce the data by identifying sets of loci that may be complementary in their association with the disease. Namely, we identify population covering locus sets (PoCos), i.e., sets of loci that harbor at least one susceptibility allele in samples with the phenotype of interest. Then we compute representative genotypes for PoCos, and assess the significance of the interactions between pairs of PoCos. We use the results of this assessment to prioritize pairs of loci to be tested for epistasis. We test the proposed method on two independent GWAS data sets of Type 2 Diabetes (T2D). Our experimental results show that the proposed method reduces the number of hypotheses to be tested drastically, enabling efficient identification of more epistatic loci that are statistically significant. Moreover, some of the identified epistatic pairs of loci are reproducible between the two datasets. We also show that the proposed method outperforms an existing method for prioritization of locus pairs.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Algorithms, Design, Experimentation

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.

http://dx.doi.org/10.1145/2649387.2649449.

Keywords

Genome-wide association studies, epistasis, statistical significance

1. INTRODUCTION AND BACKGROUND

In recent years, there have been rapid developments in our understanding of the genetic etiology of various complex diseases [17]. Further developments in biotechnology and genomic sciences, including the explosion of genome-wide association studies (GWAS), are improving the identification of susceptibility loci and the underlying genetic mechanisms of complex diseases. Recent genome-wide association studies lead to the discovery of susceptibility loci for many complex diseases, including type 2 diabetes [25], psoriasis [15], multiple sclerosis [2], and prostate cancer [8].

Earlier GWAS focus on identifying individual loci associated with diseases using standard statistical tests comparing the distribution of genotypes or minor alleles in case and control populations. However, increasing empirical evidence from model organisms [20] and human studies [26] suggests that complex interactions among two or more loci contribute broadly to complex traits. Indeed, the individual locus associations identified in GWAS are often not reproducible. These observations lead to a growing interest in identifying genetic interactions among two or more genomic loci.

Most of the studies that aim to identify genetic interactions focus on pairwise epistasis, which is usually defined as a masking effect between two genomic loci; i.e., a variant at one of the loci masks the phenotypic effect of a variant at the other locus [4, 13]. An increasing number of studies report the presence of statistically significant epistatic interactions in complex diseases [19], while some of the observed epistatic interactions remain dominated by individual variants with strong disease association [15].

Nevertheless, identifying epistatic interactions remains a computationally and statistically challenging problem since one needs to test the interaction between all pairs of loci, which amounts to $O(10^{12})$ tests for today's genome-wide screening platforms (assuming screening of 1M loci in a single assay). With the advent of whole-genome association studies through massively parallel sequencing, these numbers are bound to grow. Motivated by these considerations, many researchers claim that exhaustive methods find epistatic loci pairs are infeasible from a computational perspective [19].

The statistical challenges involve the multiple hypothesis testing problem which greatly degrades statistical power. To alleviate this problem, the number of hypotheses being

tested are usually reduced by focusing on pairs of loci that are functionally associated through regulatory elements, pathways, protein interactions, and other functional annotations [11, 18]. Some algorithms also prune out certain pairs of loci based on their allelic distributions in the case and control populations, but without explicitly testing them for epistasis (e.g., TEAM [27], QMDR [9], SNPHarvester [24]). These methods perform reasonably well for certain models of epistasis; however, they usually test a very large number of hypotheses.

Some methods, instead of controlling type 1 error, try to prioritize the SNP pairs [16]. The benefit to prioritization is that it may enable identification of interactions with modest, yet potentially relevant statistical significance that becomes insignificant after correction for multiple testing. Reducing the search space helps control for false positive errors while rendering the computation feasible. However, this may also lead to false negatives since many locus pairs are not tested at all.

It is also important to note that, although genomic loci that are in linkage disequilibrium (LD) tend to have a strong interaction, making LD an important confounding factor. However, some of the existing algorithms do not handle LD properly and report epistatic pairs of loci which are in LD [7].

In this paper, we propose a novel method for prioritizing the pairs of loci to be tested for epistasis. The proposed method prioritizes the locus pairs using Population Covering Locus Sets (PoCos). PoCos are sets of loci that harbor at least one susceptibility allele in samples with the phenotype of interest. The main idea behind the proposed approach is to group genomic loci that complement each other in describing the relationship between genotype and phenotype in the affected population. This notion is not to be confused with "genetic complementation", which refers to the case where two recessive mutations produce the wild type phenotype when combined. Here, we use the term "complementarity" in a more general and abstract sense, referring to the statistical observation that at least one of the genomic loci harbors a susceptibility allele in all affected individuals. Since such complementary genomic loci may be functionally linked to the phenotype in ways similar to each other, multiple pairs of loci in different groups may exhibit similar patterns of interaction. Indeed, our previous studies show that accounting for complementarity of genomic loci enables effective integration of the disease-association of loci that are related to a single gene, thereby improving the identification of disease-associated genes [5].

The proposed approach can be thought of as "coarsening" the set of variables (genomic loci) by grouping those that are potentially "similar" in terms of their effect on phenotype. We assess the statistical interactions among these "coarsened" variables (PoCos), which requires testing orders of magnitude fewer models than testing the interactions between the individual loci themselves. We use the outcome of this assessment to prioritize the pairs of loci that are contained in interacting pairs of PoCos.

We test the proposed method on two independent GWAS data sets of Type 2 Diabetes (T2D). We investigate the performance of prioritization from the perspective of optimizing the number of tests to be performed to capture the largest number of significant interactions by avoiding too many tests that will degrade statistical power. Our experimental results show that the proposed method ranks potentially epistatic pairs of loci higher, and drastically reduces the number of tests to be performed to capture most of the interactions. This also translates into earnings in runtime. Furthermore, we compare the patterns identified on the two different datasets and observe that the loci recruited by PoCos as well as the epistatic loci identified by the proposed prioritization algorithm are reproducible across datasets. We also compare the proposed method with an existing prioritization method called iLOCi.

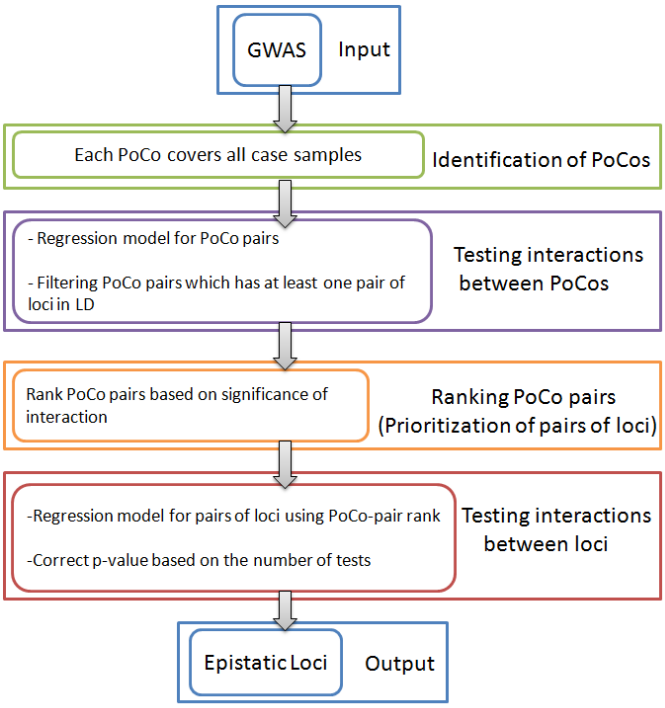


Figure 1: **The workflow of the proposed method for the prioritization of locus pairs for testing epistasis.**

In the next section, we describe the proposed procedure for identifying PoCos and using them to prioritize the locus pairs. Subsequently, in Section 3, we present comprehensive experimental results on two independent datasets for T2D. We conclude with a discussion of the implications of our results, limitations of this study, and avenues for future research in Section 4.

2. METHODS

In this section, we first present the general setup for genome-wide association analysis and motivate the problem of prioritizing locus pairs for test of epistasis. Subsequently, we introduce the notion of "Population Covering Locus Sets" (PoCos) and describe the algorithm we use to identify PoCos. Finally, we describe how we use the PoCos to prioritize pairs of loci to be tested for epistasis. The workflow of the proposed method is presented in Figure 1.

2.1 Problem Formulation

The input to the problem is a genome-wide association (GWA) dataset $D = (C, S, g, f)$, where C denotes the set of genomic loci that harbor certain genetic variants (e.g., single nucleotide polymorphisms or copy number variants) that are

assayed, S denotes the set of samples, $g(c, s)$ denotes the genotype of locus $c \in C$ in sample $s \in S$, and $f(s)$ denotes the phenotype of sample $s \in S$. Here, we assume that the phenotype variable is dichotomous, i.e., $f(s)$ can take only two values: if sample s is associated with the phenotype of interest (e.g. was diagnosed with the disease, responds to a certain drug etc.), s is called a "case" sample ($f(s) = 1$), otherwise (e.g., was not diagnosed with the disease, does not respond to a certain drug etc.), s is called a "control" sample ($f(s) = 0$). While we focus on qualitative traits here for brevity, the proposed methodology can also be extended to quantitative traits (i.e., when $f(s)$ is a continuous phenotype variable).

Association analysis. The main objective in genome-wide association studies (GWAS) is to find genomic variants whose genotypes significantly correlate with phenotype. Usually, standard statistical tests, such as χ^2 test [14] or Fisher's exact test [6] are applied to identify individual variants that are significantly associated with the phenotype.

Allele of interest. Association analysis focuses either on the genotypes (i.e., a specific combination of alleles) or the presence/frequency of the "minor allele" . The minor allele for a locus is usually defined as the allele that is less frequent in the population. In this paper, we use the more general notion of "allele of interest", where the allele of interest is not necessarily the less frequent but is useful in distinguishing case samples from control samples. This notion is particularly useful when the genotypes of multiple loci are being integrated, since alleles on different loci can act together to have a particular phenotypic effect but their effects may be in opposite directions. For example, the minor allele on one locus can be associated with increased susceptibility to the phenotype of interest while the major allele on another locus may be protective, and these two loci can be related in their association with the phenotype. In such cases, it may be more informative to consider them together.

Genotype coding. Given the allele of interest for each locus, the genotype can be coded as a $|C| \times |S|$ matrix m such that $m(c, s)$ denotes the number of copies of the allele of interest in locus c, i.e.:

$$m(c,s) = \begin{cases} 2 & \text{if } g(c,s) \text{is Homozygous of allele of interest} \\ 1 & \text{if } g(c,s) \text{is Heterozygous} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Epistasis. Since individual variants can explain a limited fraction of heritability for complex phenotypes, significant amount of research focuses on identifying epistatic pairs of genomic loci. Epistasis is defined in general as a statistical interaction between two genomic loci in the context of association with a phenotype interest. Two commonly used models for epistasis are multiplicative models and genotype-based models. In the multiplicative model, marginal effects of the two individual loci and the multiplication of their genotypes (coded as described above) are entered into a linear model. The significance of the interaction between the two loci is then assessed in terms of the significance of the multiplicative term in the model. Genotype-based models also use a linear model, but they represent each possible genotype combination with a dichotomous variable, thereby enabling identification of interactions where a specific combination of genotypes in two loci have a significant effect on phenotype. These two models are based on different assumptions on the relationship between the two loci, but an interaction may be significant according to both models as well. In this paper, we focus on the multiplicative model of epistasis to develop a prioritization algorithm based on complementarity of different loci. The proposed method can also be extended to genotype-based models, but the extension is not straightforward.

The prioritization problem for testing epistasis. With a brute-force approach, the number of the pairs of loci to be tested for epistasis is $\binom{|C|}{2}$. This is problematic from computational, as well as statistical, perspectives. Computationally, running such a large number of tests is not practically feasible. Statistically, performing a very large number of tests reduces statistical power drastically. For these reasons, it may be quite useful to prioritize the tests to be performed by ranking pairs of loci based on their promise in revealing significant interactions. In the next two sections, we propose a solution to the prioritization problem. The workflow of the proposed method is shown in Figure 1.

The proposed method is based on identifying Population Covering Locus Sets, i.e., sets of loci that complement each other in distinguishing case and control. We represent such sets of complementary loci as composite variables representing all loci in the set, thereby providing a coarser set of variables for which interactions can be tested. We then test the interactions between these composite variables to assess the likelihood of interactions between the loci in these sets.

2.2 Population Covering Locus Sets (PoCos)

We define a Population Covering Locus Set (PoCo) as a subset of individual loci such that (i) all case samples harbor an allele of interest in at least one of these loci, and (ii) the number of control samples that harbor an allele of interest in at least one of these loci is minimized. Note that the allele of interest for each locus is not specified *a priori*. As described below, we rather define allele of interest as part of the solution to the problem of identifying PoCos. Although the allele of interest is often one that is more frequent in cases than in controls, this may not be always true since we allow inclusion of alleles that are more frequent in the controls to satisfy (i).

Formal definition of PoCos. Let $a(c)$ denote a possible designation of the allele of interest for a locus $c \in C$. We define $E(c) \subseteq S$ and $T(c) \subseteq S$ as respectively the set of case and control samples that harbor the allele of interest in c, i.e.:

$$\begin{aligned} E(c) &= \{s \in S : f(s) = 1 \text{ and } g(c,s) = a(c)\} \\ T(c) &= \{s \in S : f(s) = 0 \text{ and } g(c,s) = a(c)\} \end{aligned} \tag{2}$$

A PoCo is then defined as a set $P \subseteq C$ of genomic loci accompanied by a designation of alleles of interest for all loci in P such that

$$t(P) = |\bigcup_{c \in P} T(c)| \tag{3}$$

is minimized, under the constraint

$$\bigcup_{c \in P} E(c) = \{s \in S : f(s) = 1\}. \tag{4}$$

From a biological perspective, a PoCo can be considered as a set of loci that complement each other in their association with the phenotype of interest. This is because, the case samples contain an allele of interest (which becomes the

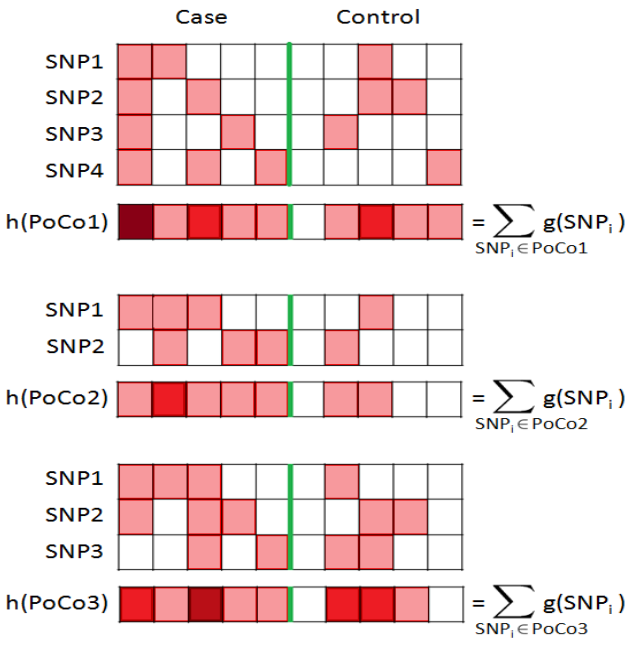


Figure 2: **Illustration of the concept of Population Covering Locus Sets (PoCos).** Three PoCos on a hypothetical case-control genotype dataset are shown on this figure. Pink squares indicate the existence of the allele of interest (major or minor, chosen adaptively) in a sample. The bottom row for each PoCo shows the "genotype" of the respective PoCo, computed as the summation of the rows for all loci in the PoCo. Each PoCo has at least one allele of interest in all the case samples, while minimizing the number of such control samples.

susceptibility allele for that locus) in at least one of these loci (the PoCo "covers" the entire case population), whereas this is true for only a minimal number of control samples. Therefore, the existence of a susceptibility allele in any of these loci may have a similar effect on phenotype, driving a possible association and functional link between these set of loci and the phenotype of interest.

Identification of PoCos. Although we define the problem of identifying PoCos as an optimization problem, we are not interested in identifying a single (and possibly the optimal) PoCo. We are rather interested in identifying multiple PoCos, since our objective is to use the interactions between PoCos to prioritize the interactions between individual loci. For this reason, we use a greedy algorithm that identifies all non-overlapping sets satisfying the constraint in (4) while providing a local minimum of the objective function in (3). For this purpose, for a given set $P \subseteq C$ of loci and a designation of alleles of interest for the loci in P, we define

$$\delta(P) = \frac{|\bigcup_{c \in P} E(c)|}{\{s \in S : f(s) = 1\}} - \frac{|\bigcup_{c \in P} T(c)|}{\{s \in S : f(s) = 0\}} \quad (5)$$

as the difference of the fraction of case and control samples "covered" by P.

It is straightforward to see that any $P \subseteq C$ that satisfies the constraint in (4) and maximizes $\delta(P)$ also minimizes $t(P)$. Motivated by this observation, we use $\delta(.)$ to guide the search for PoCos, while requiring the search to proceed until all case samples are covered. To be more precise, our algorithm seeds the search for a PoCo by selecting the locus c that maximizes $\delta(\{c\})$ and setting $P = \{c\}$. Subsequently, it considers adding each remaining locus to P and selects the locus whose addition to P improves $\delta(P)$ best. This process continues until after all case samples are covered by P or no locus can improve $\delta(P)$. In the former case, P is stored as a PoCo, all loci in P are removed from the dataset and the algorithm restarts to discover another PoCo. In the latter case, P is dismissed and the algorithm stops searching for PoCos. Using $\delta(.)$ instead of $t(.)$ during the search helps make more desirable local decisions, since the algorithm effectively tries to maximize the set of covered case samples while selecting loci to add to the growing set of loci,

When the algorithm terminates, it returns the set P of all discovered PoCos along with the designation of alleles of interest for all loci in these PoCos. As we discuss in Section 3, each identified PoCo in practice contains multiple loci and most of the loci in the dataset are not assigned to any of the PoCos. For this reason, we usually have $|P| << |C|$, i.e., the number of PoCos is orders of magnitude smaller than the number of loci.

2.3 Prioritization of Pairs of Loci as Candidates for Epistasis

Computing representative genotypes for PoCos. As stated previously, the key idea of the proposed method is to prioritize pairs of loci for testing epistasis based on the interactions between the PoCos that contain them. In order to test the interactions between pairs of PoCos, we need to compute a "genotype" for each PoCo that is representative of the genotypes of the loci in the PoCo. Since we focus on the multiplicative model of epistasis, we use an additive function to integrate the genotypes of the loci within a PoCo. This enables using a multiplicative model to test the interactions between the PoCos as well.

To be more precise, for each PoCo $P \in P$, we compute the genotype of P as

$$h(P, s) = \sum_{c \in P} g(c, s) \quad (6)$$

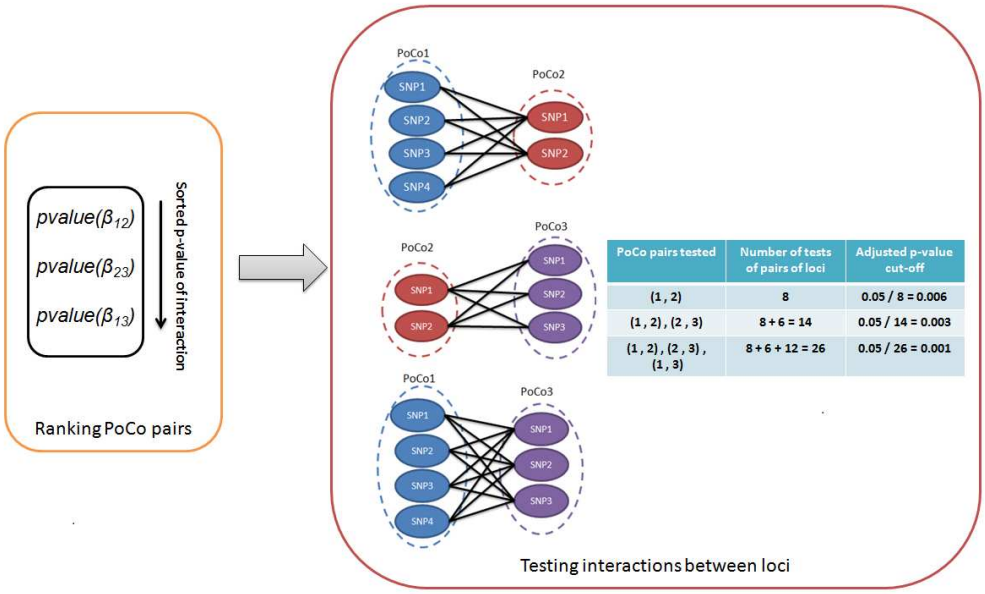
for all samples $s \in S$. This is illustrated in Figure 2. For notational convenience, we denote the genotypes of PoCo P across all samples with $h(P)$.

Testing interactions between pairs of PoCos. Once the genotypes of all PoCosare computed, we assess the interaction between any pair $P_i, P_j \in P$ of PoCos using a logistic regression model, i.e:

$$f = \beta_0 + \beta_i h(P_i) + \beta_j h(P_j) + \beta_{ij} h(P_i) h(P_j). \quad (7)$$

Since $|P| << |C|$ in practice, testing these models for all pairs of PoCos is much faster than testing epistasis for all pairs of loci.

Prioritization of pairs of loci. After the model in (7) is computed, we assign the p-value of the term β_{ij} as the score of all pairs of loci in these PoCos, i.e., we assign score $\pi(c_k, c_\ell) = pvalue(\beta_{ij})$ to all locus pairs $c_k, c_\ell \in C$ such that $c_k \in P_i$ and $c_\ell \in P_j$. This is illustrated in Figure 3. The loci that are not assigned to any PoCo remain as "unscored". Subsequently, we sort all "scored" pairs of loci in ascending



PoCo pairs tested	Number of tests of pairs of loci	Adjusted p-value cut-off
(1 , 2)	8	0.05 / 8 = 0.006
(1 , 2) , (2 , 3)	8 + 6 = 14	0.05 / 14 = 0.003
(1 , 2) , (2 , 3) , (1 , 3)	8 + 6 + 12 = 26	0.05 / 26 = 0.001

Figure 3: **Using interactions between pairs of PoCos to prioritize pairs of loci.** The prioritization of the pairs of loci in Figure 2 is shown. The interactions among the three pairs of PoCos are tested. Subsequently, these PoCo pairs are ranked based on the significance of their interaction. This ranking is directly translated into the the prioritization of pairs of loci contained within the respective PoCos, as shown by bipartite graphs on the right panel. The table on the right panel shows the number of hypothesis and the effective Bonferroni-corrected p-value threshold at each step.

order of their scores ($\pi(.)$), and test them for epistasis in the resulting order.

Reducing the number of hypotheses. An important benefit of prioritizing pairs of loci for test of epistasis is the potential reduction in the number of hypotheses that are tested. For this purpose, it is desirable to select a threshold π^* on the score $\pi(c_k, c_\ell)$ of a pair to be tested by epistasis and test only the pairs that have $\pi(c_k, c_\ell) \leq \pi^*$. Here, rather than applying a threshold, we test the performance of the prioritization by considering all possible values of π^*. Namely, after the pairs of loci are prioritized, we consider each p-value of the interaction testes for p in increasing order and use that score as π^*. We perform a test of epistasis for all pairs of loci with score less than π^* and adjust the p-values of these tests using Bonferroni correction, where the number of hypotheses tested is equal to the number of locus pairs with scores less than π^*. We then assess the number of locus pairs that are significant at a reasonable cut-off (we use 0.05 in our experiments) for these adjusted p-values. This process is illustrated in Figure 3.

Observe that, with this method, the number of significant pairs discovered may goes down as more pairs are tested, since the number of hypotheses grows with increasing π^*. Consequently, besides providing a useful methodology for assessing the ability of the prioritization algorithm in extracting significantly epistatic pairs, this analysis leads to useful insights on how π^* should be selected.

Filtering pairs of loci that are in disequilibrium. A confounding factor in the identification of epistatic loci is Linkage Disequilibrium (LD). Since SNPs that are in Linkage Disequilibrium (LD) may exhibit significant interaction effects, they may be considered to be in epistasis. In another words, pairs of loci that are not the etiological variants but are both in LD may exhibit significant interactions. If proper care is not given, this type of interactions may predominates identification of epistatic interactions [4]. In order to avoid identifying loci that are in LD, we do not test PoCo pairs that share at least one pair of loci that are in LD. This ensures that none of the pairs of loci that are in the prioritized list are in LD. In the experiments reported in the next section, we use r^2 to assess the LD between pairs of loci and consider a pair of loci in LD if they have $r^2 > 0.05$.

3. RESULTS AND DISCUSSION

In order to assess the ability of our algorithm in identifying epistatic SNP pairs, we use two independent GWAS datasets for Type 2 Diabetes (T2D). We first evaluate the performance of the proposed algorithm on the two datasets separately and then investigate the reproducibility of identified epistatic interactions. We also compare the proposed algorithm against an existing algorithm for fast discovery of epistatic pairs, iLOCi.

3.1 Datasets

We use two GWAS datasest for T2D. The first dataset is obtained from the Wellcome Trust Case-Control Consortium (WTCCC) [3]. The second dataset is the eMERGE dataset obtained from the database of Genotypes and Phenotypes (dbGaP) [12]. We filter out the SNPs with MAF greater than 5%. Moreover, in order to avoid marginal effect of SNPs, we filter those SNPs with nominal p-value of individual association less than 10^{-6}. Since we are interested in assessing the reproducibility of identified epistatic pairs, we work on the genomic loci for which genotype information is available in both data sets. In order to extend the set of common SNPs between the data sets, we also map

Table 1: Descriptive statistics of the PoCos identified on two different GWAS datasets for T2D.

	GWAS dataset	
	WTCCC	eMERGE
Number of PoCos	1258	2431
Average size (SNPs) of PoCos	4.46	4.14
Common loci between datasets	497	
Significance of overlap	6.28E-12	

SNPs that are in strong LD to each other across datasets ($r^2 > 0.9$). We use the genotype calls for 258553 loci provided by WTCCC on 1999 case and 1504 control samples. The eMERGE dataset contains genotype calls for 152831 loci on 1007 case and 983 control samples.

3.2 Identification of PoCos

We identify all the PoCos in the two datasets using the method described in Section 2.2. Descriptive statistics of the PoCos identified on each datasets are shown in Table 1. We identify 1258 PoCos containing 5618 loci on the WTCCC data set and 2431 PoCos containing 10084 loci in the eMERGE data set. The number of loci that are included in a PoCo in both datasets is 497. Using the hypergeometric model, the p-value of this overlap is found to be $6.28E-12$ which is highly significant.

3.3 Prioritization Performance

After identifying the PoCos, we rank locus pairs based on the significance of the interaction between the PoCos that contain them. Then we test the epistasis between locus pairs going through this ranking. This approach is based on the hypothesis that there are more epistatic locus pairs between PoCo pairs that are more significant in their interaction. In order to assess the validity of this hypothesis, we investigate the relationship between the significance of the interaction between a PoCo pair and the number of epistatic locus pairs in these PoCos. For this analysis, we consider two loci epistatic if their nominal p-value is less than 0.05. Figure 4 shows the results of this analysis. As can be seen in the figure, as the rank of PoCo pairs goes down in terms of the significance of their interaction, the number of epistatic locus pairs in these PoCos also decreases. In another words, the PoCo pairs that are ranked higher according to the significance of their interaction contain more epistatic locus pairs. Indeed, the Pearson correlation between the rank of PoCo pairs and the number of epistatic locus pairs contained within is respectively -0.27 and -0.24 in the WTCCC and eMERGE datasets. This result demonstrates that the significantly interacting pairs of PoCos are indeed more likely to contain epistatic pairs of genomic loci.

Recall that, given the prioritization of locus pairs, we use a moving threshold to select the pairs to be tested for epistasis. As we consider locus pairs that are ranked lower, the Bonferroni-adjusted p-value cut-off becomes stricter since the number of hypotheses goes up. The blue curves in Figure 5 show how the number of significant interactions ($p < 0.05$ after Bonferroni correction) changes as we test more locus pairs by relaxing the threshold. Since the adjusted p-value cut-off is less stringent at the beginning and high ranked PoCo pairs have more significant interactions, the number of significantly epistatic loci grows quickly at the beginning. However, as more pairs (hypotheses) are tested, the p-value threshold becomes stringent enough that the number of significant pairs starts declining.

In this analysis, the location of the peak of the curve is an important indicator of the performance of a prioritization algorithm in extracting epistatic loci (higher on the y-axis) while minimizing the number of tests performed (to the left on the x-axis). Since the blue curve grows steadily until it reaches a peak at 1K (out of potentially $\approx 10^{11}$) and 100K (out of potentially $\approx 10^{11}$) tests on respectively the WTCCC and eMERGE datasets, we can conclude that the prioritization provided by the proposed algorithm can indeed be useful in practice. It is likely that the discrepancy in the performance of the algorithm between the two datasets is because of the number of samples that are available. Since there are nearly twice the samples available in WTCCC than in eMERGE, the PoCos discovered on WTCCC are statistically more powerful.

In order to investigate whether the complementary nature of the loci in the PoCos has an effect on prioritization performance, we also compare the performance of prioritization provided by PoCos to the performance of prioritization provided by random sets of loci. For this purpose, we repeat the proposed procedure, described in Section 2.3, 100 times by using random sets of loci instead of PoCos. We select these random sets to mirror the number and size distribution of identified PoCos and use the same prioritization method, but by replacing the PoCos with the random sets of loci. In Figure 5, the red curves show the performance of random sets in prioritizating epistatic locus pairs.

As seen in Figure 5, the number of significant loci identified using the random sets is much lower than the number of significant loci identified using PoCos. This observation suggests that a locus that is recruited into a PoCo is more likely than a random locus to be involved in an epistatic interaction. This result clearly demonstrates that PoCos may indeed be biologically relevant in terms of how they capture the loci that complement each other in their association with the phenotype.

3.4 Biological Relevance and Reproducibility

To investigate whether the identified epistatic pairs are reproducible, we cross-check the significance of the identified pairs on the two datasets. Our results show that among 22 epistatic pairs of loci identified on the WTCCC dataset, 6 of them are also significant at $p < 0.05$ in eMERGE dataset. Furthermore, 3 out of 63 epistatic pairs discovered in the eMERGE dataset are also significant at $p < 0.05$ on the WTCCC dataset. All of these 9 reproducible pairs and their p-values are presented in Table 2. Note that, due to the population difference between two dataset, the PoCos in WTCCC and eMERGE datasets overlap only to a certain extent, some of the loci in one dataset are not even tested in another dataset. This result is particularly encouraging since significance of individual association and epistasis is often not reproducible across GWAS datasets.

We also estimate the power of interaction between identified pairs of loci using odds ratio test. The odds ratio test is based on the analysis of interaction in contingency tables. The odds ratio of the most significant pairs of loci identified on each dataset are shown in Figure 6.

In order to assess the functional relevance and biological validity of identified epistatic pairs, we map identified pairs

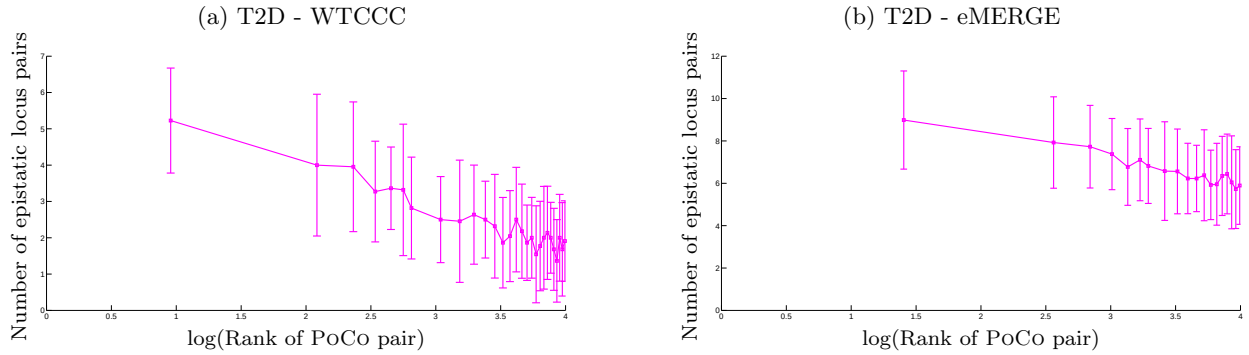


Figure 4: **The relationship between the significance of interactions among PoCo pairs and the number of epistatic locus pairs contained within.** The PoCos are binned into 30 groups to visualize the correlation between the two variables. The x-axis shows the log-scale of the rank of the PoCo pairs in terms of the significance of their interaction, the y-axis shows the 95% confidence interval for the rank of the number of epistatic pairs ($p < 0.05$) within each PoCo pair in the corresponding bin.

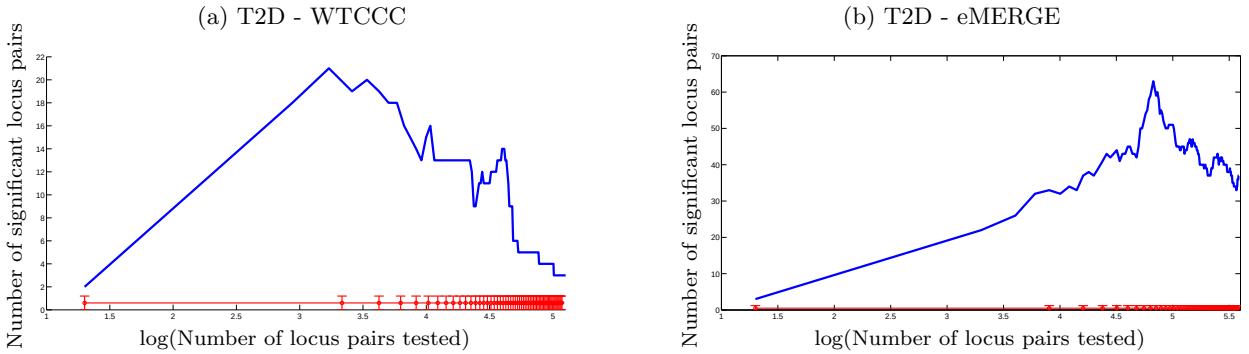


Figure 5: **The performance of prioritization in handling multiple hypothesis testing.** The x-axis shows the number of locus pairs tested for epistasis in the order provided by the prioritization, the y-axis shows the number of locus pairs that are significant at $p < 0.05$ after Bonferroni correction for multiple hypothesis testing. The blue curve shows the number of epistatic loci identified by the proposed algorithm, the red curve shows the 95% confidence interval for the number of epistatic loci identified by 100 runs of the same prioritization algorithm that works with random sets of loci instead of PoCos.

of loci to genes. For this purpose, we define the region of interest for a gene as the genomic region that extends from 20kb upstream to 20kb downstream of the coding region for that gene. Using this definition of region of interest, we are able to map 23 loci involved in epistatic interactions identified on the WTCCC dataset to the region of interest of 14 genes. 6 loci out of these 14 loci are mapped to genes that are well known to be associated with T2D, including *BAZ2B*, *TCF7L2*, *CDKN2B* and *CNTN1* [10]. Similary, on the eMERGE data set, 67 of epistatic loci are mapped to 49 genes and 16 of these genes are previously reported to be associated with T2D [22]. These results suggest that novel genes found among the prioritized pairs of loci can also be potential genes involved in T2D and gene-gene interactions associated with T2D. Furthermore, for 5 of the reproducible epistatic pairs presented in table 2, both loci in the pair map to the region of interest of a gene. Pathway analysis shows that two of these five pairs are involved in G protein-coupled receptor signalling pathway, which has a role in type 2 diabetes [1].

3.5 Comparison With Other Methods

While there are many algorithms developed to enable fast testing of epistasis or filtering of locus pairs to be tested, most algorithms do not follow the "prioritization" approach used here. iLOCi [16] is an algorithm that can be considered a prioritization algorithm. iLOCi accounts for marker dependencies separately in case and control groups. Phenotype-associated interactions are then prioritized according to a ranking score calculated from the difference in marker dependencies for every possible pair between case and control groups. It has been shown [16] that iLOCi algorithm outperforms FastEpistasis [21] in filtering pairs of loci for testing epistasis. Indeed, FastEpistasis, which is exhaustive search algorithm, takes about 89 days to run on WTCCC-sized data [7]. iLOCi is implemented to run in parallel and also uses GPU to quickly prioritize the pairs of loci.

We apply the analysis we use to assess the performance of our algorithm in the same way to assess the performance of iLOCi on the two T2D datasets. The results of this analysis are shown in Figure 7. As seen in the figure, the prioritization of pairs of loci by iLOCi leads to the identification of 3 significant loci pairs on WTCCC at its peak, which is much lower than the 22 pairs identified by the proposed algorithm. On the other hand, iLOCi is able to identify 575 epistatic pairs on the eMERGE dataset at its peak. However, 205 of these significant pairs of loci are in LD.

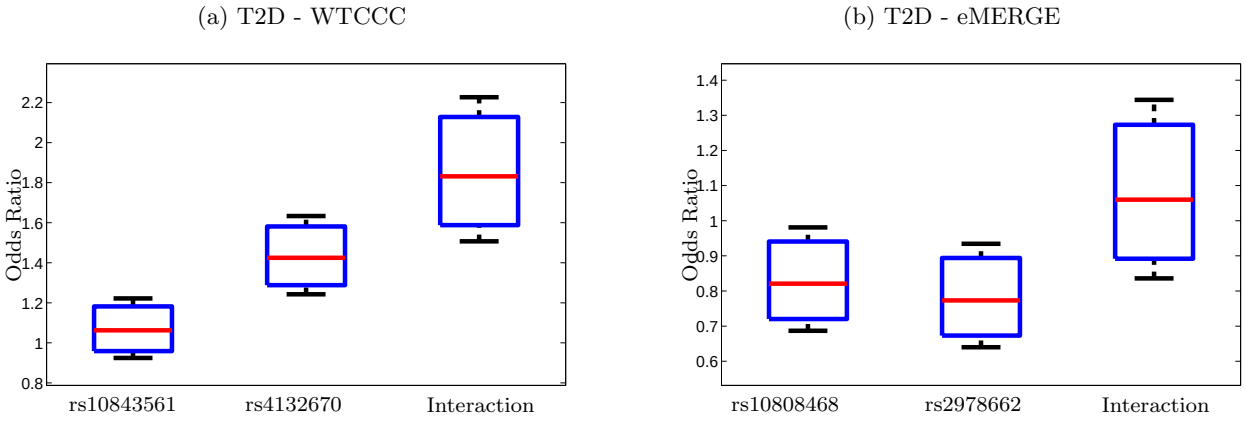


Figure 6: **Odds ratio of the most significant epistatic pairs of loci identified on each dataset.** Bars indicate 95% confidence intervals for the odds ratio of the marginal effects of each locus and the interaction effect.

Table 2: **Reproducibility of identified epistatic loci on two T2D datasets.** The first 6 rows show the pairs of loci that are discovered on the WTCCC dataset and are reproducible on the eMERGE dataset. The last three rows show the pairs of loci hat are discovered in eMERGE dataset and are reproducible in the WTCCC dataset.

Locus pair	p-value of interaction in WTCCC	eMERGE
(rs7852843 , rs3821136)	6.87E-6	0.02
(rs10965212 , rs6810719)	2.63E-5	0.02
(rs7852843 , rs3770699)	4.51E-6	0.04
(rs523096 , rs6810719)	8.38E-6	0.008
(rs564398 , rs6810719)	6.35E-6	0.009
(rs11070218 , rs9966798)	1.14E-5	9.53E-4
(rs7563869 , rs2426053)	0.022	3.99E-7
(rs584383 , rs12321799)	0.011	1.55E-7
(rs12199778 , rs368832)	0.018	1.87E-7

As we describe in 2, the proposed method particularly targets multiplicative models of epistasis. Therefore, comparison between this method and methods that target genotype-based epistasis, including SNPHarvester [24], BOOST [23] and GWIS [7], would not be informative, since these two models potentially capture different biological mechanisms of epistasis.

3.6 Runtime

iLOCi is implemented to run in parallel as well as using GPU to make the procedure faster. We use a server with a 2.2 GHz quad-core processor with 50 GB RAM. Running iLOCi takes about 11 hours to prioritize locus pairs on eMERGE and 20 hours on WTCCC. Our method, on the other hand, takes 6 hours to discover PoCos on WTCCC and 4 hours on eMERGE. Since we do not allow PoCos to overlap, this process cannot be straightforwardly parallelized. Computing the PoCo pair interactions using regression model takes about 4 hours for WTCCC and 6 hours for the eMERGE dataset. This stage of computation is parallelized using 12 workers in MATLAB, since each PoCo pair can be tested independently. In total, prioritization of pairs of loci takes 10 hours on WTCCC dataset and 10 hours on eMERGE dataset which is faster than iLOCi on both datasets.

4. CONCLUSION

Statistical tests always involve in a trade-off between false positives and false negatives. In testing epistasis, since the number of tests to be performed is quadratic in the number of genomic loci, multiple hypothesis testing poses significant challenges. In this paper, we propose a novel method to reduce the number of tests to identify epistatic pairs of genomic loci. Reducing the search space makes the problem computationally feasible and enables less conservative assessment of significance. However, this may also lead to false negatives (true interactions not chosen to be tested by the prioritization method) and false positives (spurious interactions with moderate nominal p-values that are deemed significant due to the lower number of tests performed).

Our comprehensive experiments on two T2D datasets suggest that the method identifies a considerable number of epistatic pairs that are biologically relevant and reproducible between the two datasets. These results are encouraging in terms of reducing the number false negatives (larger number of significant interactions identified as compared to other methods) and false positives (biological evidence indicating that interactions that are potentially relevant). However, a carefully designed simulation study can better characterize the method's performance in balancing the trade-off between false positives and false negatives.

Since our results suggests that PoCos may be biologically relevant, further investigation of the functional relationships among loci in the same PoCos may also reveal further insights into the mechanisms of the complementary relationships between genomic loci.

Acknowledgments

We would like to thank Thomas LaFramboise, Yu Liu, Pamela Clark, Mark Chance, Matthew Ruffalo, Dan Savel and Stevephen Hung for useful discussions. We would also like to thank anonymous reviewers whose comments and queries have helped significantly improve this paper. This work was supported in part by US National Science Foundation (NSF) award CCF-0953195 and US National Institutes of Health (NIH) award R01-LM011247.

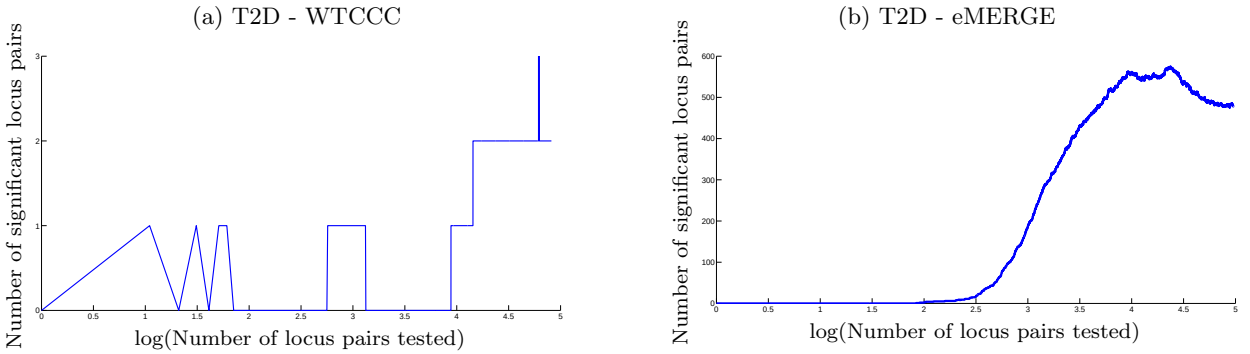


Figure 7: **iLOCi, Number of significant significant pairs of loci vs number of tests.** In order to handle multiple hypothesis testing, the significance level of interaction is corrected using Bonferroni method. The x-axis shows the log scale of number of tests, the y-axis shows number of significant epistasis according to Bonferroni corrected pvalue. The prioritization of locus pairs is the output of iLOCi method.

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from `www.wtccc.org.uk`. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. This study also uses samples and data provided by the NUgene Project (www.nugene.org).

5. REFERENCES

[1] B. Ahren. Islet G protein-coupled receptors as potential targets for treatment of type 2 diabetes. *Nat Rev Drug Discov*, 2009.

[2] Australia and N. Z. M. S. G. C. (ANZgene). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat Genet*, 41, 2009.

[3] G. Bader and C. Hogue. W. T. C. C. consortium. genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 2010.

[4] H. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, 11(20), 2002.

[5] S. Erten, M. Ayati, Y. Liu, M. R. Chance, and M. Koyutürk. Algorithms for detecting complementary snps within a region of interest that are associated with diseases. pages 194–201, 2012.

[6] R. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85, 1922.

[7] B. Goudey and et al. GWIS - model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomic*, 14(3), 2013.

[8] J. Gudmundsson, P. Sulem, and et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics*, 39, 2007.

[9] J. Gui, J. Moore, and et al. A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS One*, 8, 2013.

[10] J. Lim, K. Hong, H. Jin, Y. Kim, H. Park, and B. Oh. Type 2 diabetes genetic association database manually curated for the study design and odds ratio. *BMC Medical Informatics and Decision Making*, 2010.

[11] Y. Liu, S. Maxwell, and et al. Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from gwas data. *BMC Syst Biol*, 3, 2012.

[12] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, and et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*, 39, 2007.

[13] J. Marchini, P. Donnelly, and et al. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.*, 37, 2005.

[14] N. MS. Chi-square test for normality. *Proceedings of International Vilnius Conference on Probability Thepry and Mathematical. Statistics*, 2, 1973.

[15] R. P. Nair, K. C. Duffin, and et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kB pathways. *Nature genetics*, 2009.

[16] J. Piriyapongsa and et al. iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomic*, 13(7), 2012.

[17] N. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405, 2000.

[18] M. Ritchie. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Ann Hum Genet*, 75, 2011.

[19] M. ritchie and et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Hum. Genet.*, 69, 2001.

[20] D. Segre, A. Deluna, and et al. Modular epistasis in yeast metabolism. *Nature genetics*, 37, 2005.

[21] S. T and et al. FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26, 2010.

[22] N. Tiffin, E. Adie, F. Turner, and et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.*, 2006.

[23] X. Wan and et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-constrol studies. *Am J Hum Genet*, 87(3), 2010.

[24] C. Yang, Z. He, and et al. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, 25, 2009.

[25] E. Zeggini, L. Scott, and et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40, 2008.

[26] K. Zerba, R. Ferrell, and et al. Complex adaptive systems and human health: the influence of common genotypes of the apolipoprotein E (ApoE) gene polymorphism and age on the relational order within a field of lipid metabolism traits. *Hum. genetics*, 107, 2000.

[27] X. Zhang, S. Huang, and et al. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26, 2010.