Assessing The Collective Disease Association of Multiple Genomic Loci

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

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

ABSTRACT

Genome-wide association studies (GWAS) facilitate large-scale identification of genomic variants that are associated with complex traits. However, susceptibility loci identified by GWAS so far generally account for a limited fraction of the genotypic variation in patient populations. Predictive models based on identified loci also have modest success in risk assessment and therefore are of limited practical use. In this paper, we propose a new method to identify sets of loci that are collectively associated with a trait of interest. We call such sets of loci "population covering locus sets" (PoCos). The main contribution of the proposed approach is three-fold: 1) We consider all possible genotype models for each locus, thereby enabling identification of combinatorial relationships between multiple loci. 2) We use a network model to incorporate the functional relationships among genomic loci to drive the search for PoCos. 3) We develop a novel method to integrate the genotypes of multiple loci in a PoCo into a representative genotype to be used in risk assessment. We test the proposed framework in the context of risk assessment for two complex diseases, Psoriasis (PS) and Type 2Diabetes (T2D). Our results show that the proposed method significantly outperforms individual variant based risk assessment models.

Categories and Subject Descriptors

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

General Terms

Algorithms, Design, Experimentation

Keywords

Genome-wide association studies, risk assessment, protein protein interaction networks

1. INTRODUCTION

Genome-wide association studies (GWAS) had a transformative effect on the search for genetic variants that are associated with complex traits, since they enable screening of hundreds of thousands of genomic variants for their association with trait of interest [1]. Recently published GWAS led to the discovery of susceptibility loci for many complex diseases, including type 2 diabetes [2], psoriasis [3], multiple sclerosis [4], and prostate cancer [5], to name a few. For improved identification of risk variants, researchers have drawn information from clinical, microarray, copy number, and single nucleotide polymorphism (SNP) data to build disease risk models, which are then used to predict an individual's susceptibility to the disease of interest [6, 7]. Several companies, such as deCODE genetics (`http://www.decodeme.com`) and 23andme (`https://www.23andme.com`) have started using SNPs identified by GWAS, to provide personal genomic test services in the United States and health related genomic test services in Canada and the United Kingdom.

An important problem with GWAS is that the identified variants account for little heritability [8, 9]. Empirical evidence from model organisms [10] and human studies [11] suggests that the interplay among multiple genetic variants contribute to complex traits. Epistasis among pairs of loci, i.e., significantly improved association with the phenotype when two loci are considered together, is also shown to provide provide further insights into disease mechanisms [12, 13, 14]. Therefore, recent studies focus on identifying the interactions among pairs of genomic loci, as well as among multiple genomic loci [15, 16, 17]. These studies suggest that consideration of more than one locus together can better capture the relationship between genotype and phenotype. For this reason, genetic markers that involve the aggregation of multiple genomic loci can be used to design effective strategies for risk assessment and guiding treatment decisions [18]. The Polygenic score is a commonly used method that has been used to identify the joint association of a large mass of the loci to predict disease risk. The first application of polygenic score on GWAS data showed the genetic risk for schizophrenia is a predictor of bipolar disorder [19]. There are also several studies that demonstrate that the polygenic risk score is a powerful tool in risk prediction [19, 20, 21].

The importance of epistasis or higher order interactions among multiple loci is commonly recognized. However, detecting such interactions requires tremendous amount of computational resources. So detecting epistatic interactions on GWAS data containing hundreds of thousands of loci is not

BCB'15, September 9–12, 2015, Atlanta, GA, USA.

http://dx.doi.org/10.1145/2808719.2808758.

usually feasible. Identifying the optimal number of loci for high-order interactions also makes the problem even more computationally expensive. Furthermore, the issue of multiple hypothesis testing leads to challenging statistical problems. Therefore, several methods and software packages are developed to identify the statistical interactions between pairs of genomic loci [22, 23, 24]. It is often observed that the risk loci for many diseases are clustered in common pathways [25]. Based on this observation, many methods overcome the multiple testing issue by filtering SNP pairs based on prior biological knowledge [26, 27]. However, such filtering-based approaches are biased toward established functional relationships, therefore they can miss important causal variants. In contrast, some approaches prioritize pairs of loci based on simpler statistics on the relationship between the genotypes of loci [28, 29].

Since detection of epistasis and higher order interactions is costly, many methods first assess the disease association of individual loci and then use functional knowledge to integrate these associations [25, 30]. The key idea behind these methods is that functionally related variants, e.g., those that induce dense subnetworks in protein-protein interaction (PPI) networks, can provide stronger statistical signals when they are considered together [31]. Based on similar insights, some researchers integrate GWAS with pathway information to identify the statistically significant pathways that are associated with the disease [32, 33]. Azencott *et al* propose a method to discover sets of genomic loci that are associated with a phenotype while being connected in an underlying biological network which is defined between SNPs [34]. They use an additive model to integrate the genotypes of loci and they build a network by integrating the proximity of genetic loci as well as interacting genes. Then they use graph structured features to select a set of disease associated SNPs.

In this paper, we propose a novel criterion to assess the collective disease association of multiple genomic loci. The proposed method builds on the concept of "Population Covering Locus Sets" (PoCos), which is introduced in our previous work and used to prioritize pairs of genomic loci for testing epistasis [29]. Namely, a PoCo is a set of loci that harbor at least one susceptibility allele in samples with the phenotype of interest. Here, we extend the notion of PoCos to enable the adaptive identification of "susceptibility genotype" (as opposed to susceptibility allele) for each locus. We also develop a method for aggregating the genotypes of multiple loci in a PoCo to compute representative genotypes for use in risk assessment. Finally, in order to capture the functional relationship between genomic loci, we incorporate the human protein-protein interaction (PPI) network in the identification procedure of PoCos.

We use the PoCos identified by the proposed framework to develop models for risk assessment. For this purpose, we perform nested cross-validation and use a feature selection algorithms to select PoCos used in the model. We evaluate the performance of PoCos in risk assessment on GWAS datasets for Type 2 Diabetes (T2D) and Psoriasis (PS), using Area Under ROC Curve (AUC) in a cross-validation setting. We compare the risk assessment performance of models built using PoCos to that of models built using individual loci. Our experimental results show that PoCos significantly outperform individual loci in risk assessment. Furthermore, we investigate the effect of adaptively selecting "susceptibility genotypes" on improving risk assessment. Finally, we compare the risk assessment performance of PoCos identified using the network with PoCos that are identified independent of network. In order to show the significant improvement of risk assessment using the proposed framework, we also compare the performance of our method with polygenic risk score.

In the next section, we describe the proposed procedure for modeling the genotypes and identifying PoCos. Then we describe how we use PoCos to develop a model for risk assessment. In section 3, we present comprehensive experimental results on two GWAS data sets for T2D and PS. We conclude with a discussion of our results and future research in section 4.

2. METHODS

In this section, we first present the set-up for genome-wide association analysis. We then define "Population Covering Locus Sets" (PoCos) and describe the algorithm we use to identify PoCos. Finally, we describe our framework for risk assessment using PoCos and the feature selection method we use to identify an optimal set of PoCos to be used for risk assessment.

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 the 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 ($f(s) = 0$). We denote the set of case samples with S_1 and the set of control samples with S_0, where $S_1 \cup S_0 = S$. 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).

2.2 Identifying Genotypes of Interest

The minor allele for a locus is usually defined as the allele that is less frequent in the population. While it is common to focus on the minor allele to assess the effect of the SNPs on the phenotype, specific genotypes can also be associated with a phenotype [35, 36, 37]. Here, we argue that considering the effect of all possible genotype combinations can provide more information in distinguishing case samples from control samples. This notion is particularly useful when the genotypes of multiple loci are being integrated. For example, heterozygosity on one locus can be associated with increased susceptibility to a disease, while homozygous minor allele on another locus may be protective at the presence of heterozygosity in the former locus [38]. In this case, the interaction between the two loci can be detected by considering the association of all possible genotype combinations with the phenotype.

In this paper, we adaptively binarize the genotypes of each locus by considering all possible allele combinations. Given

the genotype of a locus, we consider five different binary genotype models $m^{(i)}, i \in \{1, \ldots 5\}$. Based on each model, we generate a binary genotype profile for the locus. We then separately assess the association of the resulting five genotype profiles with the phenotype of interest. Subsequently, we choose the model that leads to greatest discrimination between cases and controls, and use the respective binary genotype profile as the representative genotype of that locus. This process is illustrated in Figure 1. Namely, we consider the following genotype models:

1. Homozygous Minor Allele: This corresponds to the case when the possible effect of the minor allele is "recessive", i.e., the locus is considered to harbor a genotype of interest if both copies contain the minor allele.

$$m^{(1)}(c,s) = \begin{cases} 1 & \text{if } g(c,s) \in \{aa\} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2. Heterozygous Genotype: The locus is considered to harbor a genotype of interest if the two copies contain different alleles.

$$m^{(2)}(c,s) = \begin{cases} 1 & \text{if } g(c,s) \in \{Aa\} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

3. Homozygous Major Allele: The locus is considered to harbor a genotype of interest if both copies contain the major allele.

$$m^{(3)}(c,s) = \begin{cases} 1 & \text{if } g(c,s) \in \{AA\} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

4. Presence of Minor Allele: This corresponds to the case when the possible effect of the minor allele is "dominant", i.e., the locus is considered to harbor a genotype of interest if at least one copy contains the minor allele. This is the complement of $m^{(3)}$.

$$m^{(4)}(c,s) = \begin{cases} 1 & \text{if } g(c,s) \in \{Aa, aa\} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

5. Presence of Major Allele: The locus is considered to harbor a genotype of interest if at least one copy contains the major allele. This is the complement of $m^{(1)}$.

$$m^{(5)}(c,s) = \begin{cases} 1 & \text{if } g(c,s) \in \{Aa, AA\} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

These five models represent all possible allele combinations for a single locus. Note that, although models m_4 and m_5 are complements of other models, we consider them separately. This is because, as we discuss in the next section, the 1s and 0s in the binary genotype profiles are considered asymmetrically while integrating the genotypes of multiple loci.

Given the five $|S|$-dimensional binary genotype profiles $m^{(i)}(c), i \in \{1, \ldots 5\}$, we compute the difference in the fraction of case and control samples that harbor the genotype of interest as follows:

$$D^{(i)}(c) = \frac{\left\langle f, m^{(i)}(c) \right\rangle}{|S_1|} - \frac{\left\langle \mathbf{1} - f, m^{(i)}(c) \right\rangle}{|S_0|}. \quad (6)$$

where $\mathbf{1}$ denotes a vector of all 1's and $< . >$ denotes the inner product of two vectors. We then determine the binary genotype model for each locus as the model that maximizes the difference of relative coverage between case samples and control samples, i.e.:

$$k(c) = \text{argmax}_{i \in \{1 \ldots 5\}} \{D^{(i)}(c)\}. \quad (7)$$

Based on the selected model for each locus, we compute the binary genotype profile for each locus accordingly:

$$M(c,s) = m^{(k(c))}(c,s). \quad (8)$$

2.3 Population Covering Locus Sets (PoCos)

Once we compute the binary genotype profiles for all loci, we identify Population Covering Locus Sets (PoCos). In previous work, we define and use PoCos in the context of prioritizing locus pairs for testing epistasis [29]. In this earlier definition, the genotypes of interest are limited to the presence of the minor or major allele; i.e., only the last two models described in the previous section are used to determine the binary genotype profile of each locus. Here, we generalize the concept of PoCo to utilize five different models for determining the genotypes of interest, as described in section 2.2.

A PoCo is a set of genomic loci that collectively "cover" a larger fraction of case samples while minimally covering control samples. Namely for a given set $P \subseteq C$ of loci, we define the set of case and control samples covered by P respectively as

$$E(P) = \cup_{c \in P} \{s \in S_1 : M(c,s) = 1\} \quad (9)$$

and

$$T(P) = \cup_{c \in P} \{s \in S_0 : M(c,s) = 1\}. \quad (10)$$

Given a parameter α that defines the population coverage threshold, we define a PoCo as a set P of loci that satisfies $|E(P)| \geq \alpha |S_1|$ while minimizing $|T(P)|$. Note that, since we are interested in finding all sets of loci with potential relationship in their association with phenotype, we do not define an optimization problem that aims to find a single PoCo with minimum $|T(P)|$. We rather develop an algorithm to heuristically identify all non-overlapping PoCos with minimal $|T(P)|$.

2.4 Identification of PoCos

To identify all non-overlapping PoCos, we use a greedy algorithm that progressively grows a set of loci to maximize the difference of the fraction of case and control samples covered by the loci that are recruited in a PoCo. In another words, we initialize P to $\emptyset$ and at each step, add to P the locus that maximize

$$\delta(c) = \frac{|E(\{c\}) \cap S'|}{|S_1|} - \frac{|T(\{c\}) \cap S'|}{|S_0|} \quad (11)$$

where $S' = S \setminus (E(P) \cup T(P))$. The algorithm stops when a sufficient number of case samples are covered, i.e., when $|E(P)| \geq \alpha |S|$. We then record P, remove the loci in P from the dataset, and identify another PoCo. This process continues until it is not possible to find a set of loci that covers a sufficient fraction of case samples.

Since we are trying to find the sets of variants that are related to each other in their association with the disease, utilizing interaction data can be a powerful tool to provide a functional context for PoCos. Motivated by this observation, we use two different methods to guide the search for

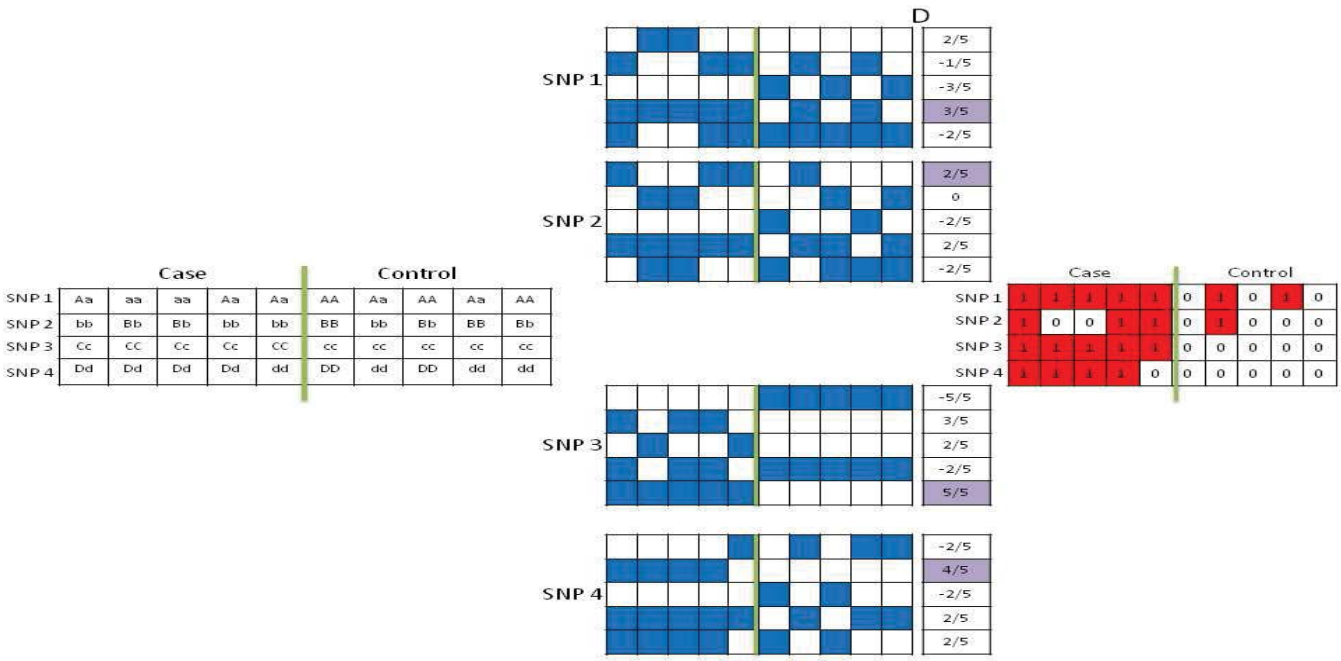


Figure 1: **Model selection and computation of binary genotype profiles for each genomic locus.** The genotypes of four loci on a hypothetical case-control dataset are shown on the left. The five possible binary genotype profiles for each SNP are computed, as shown in the middle. Blue squares indicate the presence of the genotype of interest in the respective sample for each model (respectively, homozygous minor allele, heterozygous, homozygous major allele, presence of minor allele, presence of major allele). The resulting binary genotype profiles for each locus are shown on the right. Red squares indicate the existence of genotype of interest according the selected model. In this example, models $m^{(4)}$, $m^{(1)}$, $m^{(5)}$, and $m^{(2)}$ are respectively selected for the four loci. .

PoCos: (i) Network-free PoCos and (ii) Network-guided PoCos (NetPocos).

2.4.1 *Network-Free* PoCos

For network-free PoCos, the search space for the problem contains all the loci that are genotyped and no restriction is applied on the search space. We use $\delta(.)$ to guide the search for PoCos, and require the search to proceed until $\alpha|S_1|$ case samples are covered.

2.4.2 NetPocos

We also identify PoCos by restricting the search space to the human protein-protein interaction network (PPI). The basic idea is inspired by the NetCover algorithm that is used to identify dysregulated subnetworks [39]. The inputs to this problem are GWAS data and also a network $G = (V \cup U, E \cup F)$, which represents the functional relationships among genomic loci through a backbone that is derived from the PPI network. The two types of nodes in the network represent proteins and genomic loci. Namely, V denotes the set of proteins and U denotes the set of loci that are genotyped in the GWAS. The interactions and associations between these two different types of nodes are also represented by two different sets of edges. The set of pairwise interactions between proteins is represented by E. The association between genomic loci and proteins are represented by edge set F. Namely, there is an edge between a locus and a protein if the locus is in the region of interest (RoI) for the coding gene (in our experiments, RoI is defined to be within 20Kb of the coding region). Therefore, in this network, two loci that are in the RoI of two functionally associated genes are 3 hops apart from each other. The constructed subnetwork is more sparse compared to the networks between loci that is used by Azencott *et al* [34].

The algorithm for identifying NetPocos is illustrated in Figure 2. This algorithm proceeds similarly to the original algorithm, but the set of loci that can be recruited is restricted by the network. Namely, at any step of the algorithm, only loci that are at most three hops away from at least one locus in P are considered as candidates for addition into P. This ensures that the loci in a NetPoco are functionally associated with each other; i.e., each locus in a NetPoco is associated with a protein that interacts with a protein that is associated with another locus in the NetPoco.

When the algorithm terminates, it returns the set Π of all discovered 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 $|\Pi| << |C|$.

2.5 Model Development for Risk Assessment

One potential utility of the PoCos is risk assessment. Since each PoCo provides a means to aggregate the effects of multiple loci in their association with the disease, these PoCos may provide more robust and reproducible features to be used in predictive models, as compared to individual variants. To investigate the utility of these multi-locus fea-

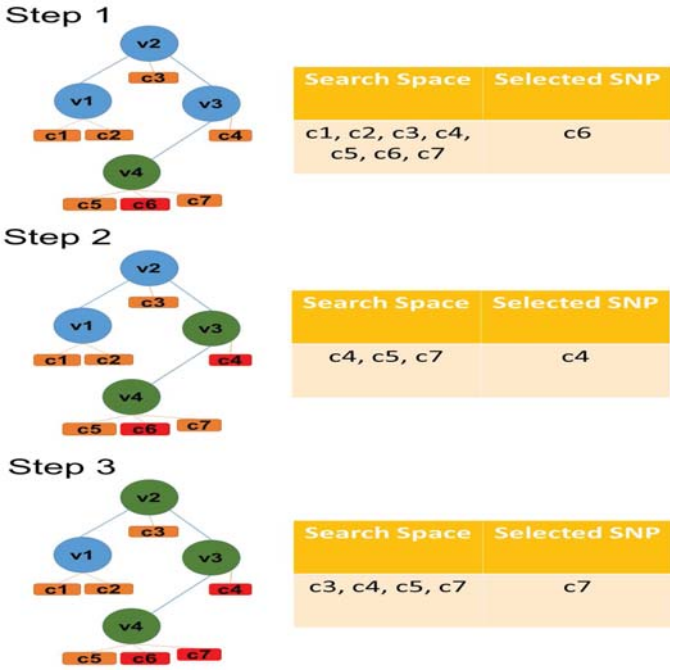


Figure 2: **Identification of NetPocos.** Each v_i represents a protein (V) and each c_j represents a genomic locus (U). Blue edges represent the interactions between proteins (E) and orange edges indicate that the respective locus is in the RoI of the coding gene for the respective protein (F). Initially, P is empty and all loci are considered and the locus (c_6) that maximizes $\delta(.)$ is added to P. After this point, the search space is restricted to loci that are three hops away from c_6. We continue this procedure until the set of selected loci cover a sufficient fraction of the case samples. Red nodes and green nodes show the selected loci and proteins respectively.

tures in risk assessment, we build a model for risk assessment using a Naïve Bayes classifier.

2.5.1 *Representative Genotypes of* PoCos

To facilitate the use the PoCos for risk assessment, we compute a representative genotype for each PoCo. For this purpose, we use the fraction of the loci in the PoCo that harbor a genotype of interest in the respective sample. To be more precise, for each PoCo $P \in \Pi$, we compute the profile of P as

$$h(P,s) = \frac{\sum_{c \in P} M(c,s)}{|P|} \tag{12}$$

The set of features utilized by the classifier is comprised of $h(P,s)$ for all $P \in \Pi$. Next, we perform feature selection to identify a parsimonious set of PoCos to be used in risk assessment.

2.5.2 *Feature Selection*

In order to find the optimal set of PoCos to be used for risk assessment, we use a forward selection based wrapper method. To avoid overfitting, we use nested 5-fold cross-validation. Namely, we divide the set of samples into 5 groups $\{T_1, \ldots, T_5\}$ while keeping the proportion of case and control samples fixed across groups. We use T_1 as an independent test group and the rest of population for training the model. We divide the training group further into 5 groups and use this partitioning to perform feature selection. The objective function that we try to maximize within the inner fold is the area under the ROC curve (AUC), described in the next subsection.

We start with an empty model and select the PoCo that provides the best AUC score in cross-validation to be added to the model. We then add the next PoCo that provides the best improvement in the AUC. We repeat these steps until adding a new PoCo does not improve the AUC. We use the final set of selected PoCos in the final model to be tested on an independent part T_1. We then repeat the same procedure by using T_2, T_3, T_4, T_5 as the test group.

2.5.3 *Performance Evaluation for Risk Asssessment*

Risk assessment models produce quantitative predictions of susceptibility to the disease of interest. To evaluate the predictive ability of these risk assessment models, we apply different thresholds on the predicted risk to obtain a binary prediction for each test sample. Using these binary predictions, we obtain the counts of true positives (predicted to be in risk, has the disease), false positives (predicted to be in risk, does not have the disease), and false negatives (predicted not to be in risk, has the disease), and compute the precision (fraction of true positives among all predicted to have risk) and recall (fraction of true positives among all who have the disease) figures based on these counts. We assess the performance of each risk assessment model based on the area under the ROC curve (AUC), which characterizes the ability of the model in trading off precision and recall for varying thresholds on the quantitative prediction.

3. RESULTS AND DISCUSSION

To assess the ability of PoCos in producing informative multi-locus features, we evaluate their utility in the context of risk assessment. For this purpose, we use GWA data for two different complex diseases: Type 2 Diabetes (T2D) and Psoriasis (PS). On each dataset, we first identify PoCos, select features to build a model for risk assessment, and then evaluate the performance of the resulting model. To avoid overfitting and to ensure that the performance figures are not biased, we use nested cross validation.

We first compare the risk assessment performance of the multi-locus features against the standard approach of using individually significant loci. Subsequently, to gain insights into the effects of genotype models and network information we also compare the performance of NetPocos *vs.* network-free PoCos, and minor-allele based PoCos *vs.* multiple genotype model based PoCos. Moreover, we compare the performance of NetPocos *vs.* polygenic score which is a commonly used method for risk assessment. Finally, we assess the potential biological relevance of the performance improvement provided by PoCos by repeating our experiments on datasets with permuted phenotypes and permuted genotypes. To facilitate fair comparisons, we use the classification and feature selection methods described in Section 2.5 identically for all types of multi-locus and individual-locus based features.

3.1 Experimental Setup

Datasets: We use GWA data for type II diabetes (T2D) and Psiorasis (PS), both obtained from the Wellcome Trust Case-Control Consortium (WTCCC) [40]. We filter out the loci with minor allele frequency (MAF) $\geq 5\%$. While identifying the PoCos, in order to avoid marginal effect of individual loci, we filter the loci with nominal p-value of individual association less than $\leq 10^{-7}$ (this corresponds to a corrected p-value threshold of 0.05). Since we utilize the PPI network to identify NetPoCos, we focus the SNPs that are in $20kb$ upstream and downstream of the gene intervals. After all filters are applied, the remaining T2D dataset contains genotype calls for 86181 loci on 1999 case and 1504 control samples. The final PS dataset contains genotype calls for 52346 loci on 2178 case and 5175 control samples. We use a human PPI network downloaded from the HIPPIE (Human Integrated Protein-Protein Interaction rEference) database, which integrates multiple experimental PPI datasets [41]. The HIPPIE PPI network contains 160561 interactions among 14611 proteins.

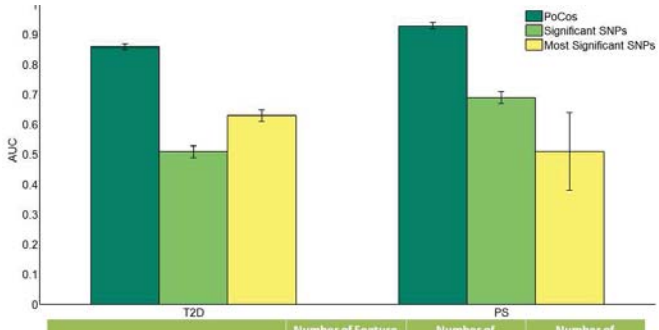
SNP-gene mapping: We do not use gene information to identify network-free PoCos. To facilitate the identification of NetPoCos, we map SNPs to genes by defining the region of interest (RoI) for a gene as the genomic region that extends from 20kb upstream to 20kb downstream of the coding region for that gene.

Association analysis for individual loci: In order to compare the multi-locus features with single-locus features, we need to identify the individually significant loci. For this purpose, we use PLINK [42], a well-established toolkit for GWA analysis. We assess the disease-association of all loci in each dataset based on minor allele frequency, obtaining a p-value for the association of each locus with the disease. We adjust the p-values for multiple hypothesis testing using Bonferroni correction and set a threshold of 0.05 to identify the individually significant loci.

3.2 Performance of PoCos in Risk Assessment

We identify PoCos in the two datasets using the method described in Section 2.4.2. After identifying the NetPoCos across all samples, we divide the population to 5 groups while preserving the proportion of case and control samples in each group. Then we reserve one group for testing, and train the model on the remaining four groups for feature selection as described in Section 2.5. Then we test the performance on the group reserved for testing. Note that the NetPoCos are identified using all samples. However, when we compare the performance of different multi-locus based features and individual locus based features, all of these features are identified using all samples as well. Here, we compare the risk prediction performance of different multi-locus based features and individual locus based features based on this nested cross-validation framework.

PoCos *vs. Individual Loci:* To investigate the benefit of using multiple-locus features (PoCos) in risk assessment, we compare the performance of PoCo-based risk assessment models against that of individual locus based models. As described in Section 2.5, we select individual locus based features by identifying loci with statistically significant association with the disease ($p < 0.05$ after correction for multiple hypothesis testing). Since the number of statistically significant individual loci is smaller than number of multiple-locus features, we also run the feature selection algorithm on the a set of individual loci with the same size as the set of multiple-locus features. For this purpose, we sort the loci corresponding to their p-value and pick the top k loci such that k is the number of multiple-locus features. We then perform feature selection for individual loci using the forward selection algorithm as for multiple-locus features. We then test the final model using cross-validation. The results of this analysis are shown in Figure 3. Note that, the PoCos utilized in this analysis are NetPoCos that are identified using multiple genotype models. We compare these PoCos against individual locus based features, since as we discuss in the rest of this section, this combination outperforms any other combination on both datasets.

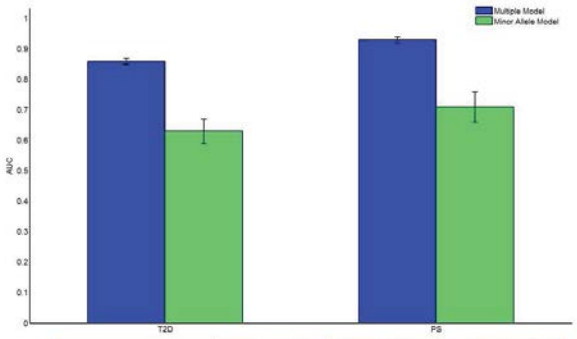


		Number of Feature	Number of Selected Features	Number of Selected SNPs
T2D	PoCos	7999	225 ± 17.75	1781 ± 135.74
	Significant SNPs	242	2	2
	Most Significant SNPs	7999	152 ± 9.2	152 ± 9.2
PS	PoCos	4163	286.8 ± 14.02	2206 ± 129.96
	Significant SNPs	216	27.8 ± 16.65	27.8 ± 16.65
	Most Significant SNPs	4163	33.5 ± 31.13	33.5 ± 31.13

Figure 3: **Comparison of the risk assessment performance of PoCos and individual locus based features on T2D and PS datasets.** The colored bars show the average AUC score and error bars show the standard deviation of AUC score across 10 different runs. The table shows descriptive statistics of the features in the full and final models. Significant SNPs are the SNPs that are significant after Bonferroni correction (p-value < 0.05). Most significant SNPs is a set of significant SNPs (before correction) that has the same size with number of PoCos.

As seen in Figure 3, models that utilize multi-locus features significantly outperform individual locus based features in risk assessment for both T2D ($p < 1.24E-8$) and PS ($p < 2.3E-6$). Here, the significance of the performance gap between two different methods is assessed using standard t-test.

Note that, particularly for T2D, non-genetic risk factors including age, sex, and body-mass index (BMI) play an important role in risk. These factors can be also combined with genetic factors to obtain better performance in risk assessment [43]. Janipalli et al. [44] combine 32 genomic loci with other conventional risk factors to obtain an AUC of 0.63 in an Indian population. Therefore the performance improvement provided by the multi-locus features as compared to the individual locus based features in a genetic factor only setting suggests that combination of multi-locus genomic features with other factors may lead to an even greater predictive performance in risk assessment.



		Number of PoCos	Size of PoCos	Number of Selected Features
T2D	Multiple Model	7999	7.35 ± 2.01	225 ± 17.75
	MA Model	5808	8.57 ± 2.43	145 ± 25.64
PS	Multiple Model	4163	7.38 ± 1.82	286.8 ± 14.02
	MA Model	3324	8.44 ± 2.07	137.8 ± 15.65

Figure 4: **Risk assessment performance of** PoCos **identified using multiple genotype models vs. minor allele model based** PoCos **on T2D and PS.** The colored bars show the average AUC score and the error bars shows the standard deviation of AUC score across 10 different runs. The table shows descriptive statistics of the features in the full and final models.

Multiple Genotype Models vs. Minor Allele Based Model. One of the important contributions of the proposed framework is the ability to adaptively choose a genotype model among all possible models while integrating the effects of multiple loci. To understand whether this added flexibility has an effect on the performance of risk assessment, we compare the performance of multiple-genotype based PoCos to PoCos identified by considering only the presence of the minor allele. The results of this comparison are shown in Figure 4. As seen in the figure, multiple allele based PoCos significantly outperform minor allele based PoCos for both T2D ($p < 1.9E-8$) and PS ($p < 5E-5$). Therefore, for the rest of the paper, we use PoCos identified using multiple models.

Risk Assessment on Randomized Datasets: Since the proposed framework provides the flexibility to choose the genotype model that provides the best disease association for each locus, it may be prone to overfitting. Also, the performance gain provided by the multi-locus based features can be purely methodological. More specifically, while our goal here is to identify loci that coordinately describe the genotypic variability in the patient population, the use of multiple loci in constructing a feature can be the main factor that results in more robust risk assessment. To investigate whether the performance improvement provided by PoCos is biologically relevant, we compare the risk assessment performance of PoCos on the original datasets to that on randomized datasets. For this purpose, we generate two sets of randomized datasets:

- Permuted phenotype: We generate 10 randomized datasets by randomly permuting the phenotype labels of the samples. Using this permutation, we assess the statistical significance of the association of PoCos with the disease.

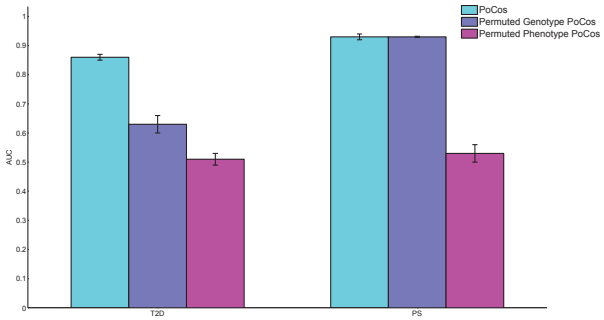


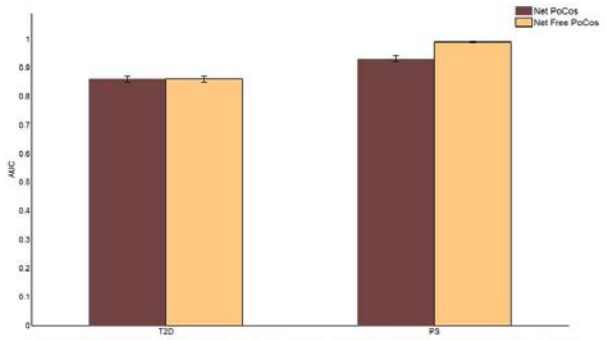
Figure 5: **Comparison of the risk assessment performance of** PoCos **on the original dataset and randomized datasets with permuted phenotypes and permuted genotypes for T2D and PS.** The colored bars show the average AUC score and the error bars shows the standard deviation of AUC score across 10 different runs. The table shows descriptive statistics of the features in the full and final models.

- Permuted genotype: We generate 10 randomized datasets by randomly permuting the genotype of loci while preserving the phenotype labels. This permutation preserves the individual disease association of each locus, but randomizes the relationship between the genotypes of different loci. We use this permutation to assess the significance of the performance improvement provided by PoCos.

As on the original dataset, we identify PoCos in each randomized dataset and use the resulting PoCos to build models for risk assessment. The results of this analysis are shown in Figure 5. As seen in the figure, for both diseases, the PoCos do not provide predictions better than random guess on datasets with permuted phenotypes. This observation confirms that the use of multiple genotype models does not lead to overfitting.

For the randomized genotypes, it is expected for the risk assessment models to perform better than random guess since the relationship between the genotypes of each locus and the phenotype is preserved. However, the predictive performance of PoCos on the original T2D dataset is significantly ($p < 5.8E-7$) better than that on the T2D dataset with randomized genotypes. This result suggests that the proposed framework captures relationships among multiple loci that is relevant in the context of susceptibility to T2D. In contrast, the model based on PoCos performs slightly better on the original PS dataset than on the randomized PS dataset, and the performance gap is not significant. The reasonably good performance of PoCos identified from permuted genotype data can be attributed to the strong association between psoriasis and individual loci in the genomic region surrounding HLA [45].

NetPocos *vs. network-free* PoCos*:* Many computational methods have been developed to integrate the GWAS data with other biological datasets that provide information on the functional relationships between individual biological entities (here, genomic loci). Here, we utilize the human PPI network in the identification of NetPocos. Since the identified PoCos are guided by the PPI network, we expect that



		Number of PoCos	Size of PoCos	Number of Selected Features
T2D	Net PoCos	7999	7.35 ± 2.01	225 ± 17.75
	Net Free PoCos	14054	5.93 ± 1.80	60 ± 28
PS	Net PoCos	4163	7.38 ± 1.82	286.8 ± 14.02
	Net Free PoCos	8779	5.96 ± 1.89	80.6 ± 34.43

Figure 6: **Comparison of the risk assessment performance of NetPoCos and network-free PoCos on T2D and PS.** The colored bars show the average AUC score and the error bars shows the standard deviation of AUC score across 10 different runs.

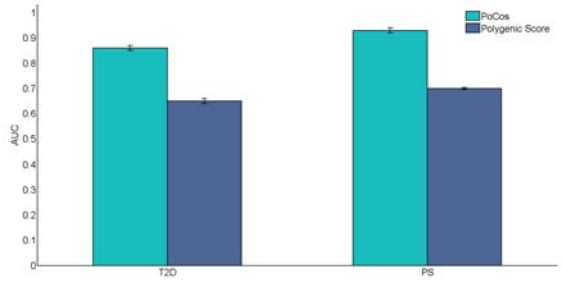


Figure 7: **Comparison of the risk assessment performance of NetPoCos and polygenic score on T2D and PS.** The colored bars show the average AUC score and the error bars shows the standard deviation of AUC score across 10 different runs.

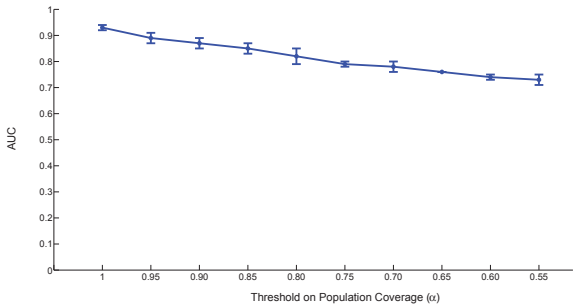


Figure 8: **Effect of threshold on population coverage on the risk assessment performance of PoCos on PS.** The x-axis shows α (the threshold on the coverage of case samples) and the y-axis shows the Area Under ROC Curve (AUC) provided by the PoCos identified using the respective threshold.

NetPoCos would be more informative and robust, since they are composed of functionally related loci. To investigate whether this hypothesis is supported by empirical results, we compare the performance of NetPoCos in risk assessment with that of network-free PoCos. The results of this analysis are shown in Figure 6. As seen in the figure, constraining the search space by functional interactions based on PPIs results in reduced predictive power of PoCos, and network-free PoCos provide more parsimonious final models. This result suggests that interactions among proteins may be limited in capturing the functional relationship between genomic loci. For more effective utilization of functional information, it may be more useful to incorporate regulatory interactions (e.g., ENCODE [46]) as well.

NetPoCos *vs. polygenic score:* A polygenic risk score is a sum of associated loci, weighted by effect sizes which are estimated using the training set. The features are selected using the p-value threshold in training samples and they are used to score the individuals in test samples. In order to find the best performance of the polygenic score, we test different p-value thresholds for a set of associated loci and pick the threshold with the best performance in the risk prediction. The results of the performance of risk assessment are shown in Figure 7. As seen in the figure, the model that uses PoCos significantly outperforms polygenic risk score performance in risk assessment for both T2D ($p < 1.4E-6$) and PS ($p < 5.57E-7$).

Effect of Threshold on Population Coverage: As discussed in the previous section, we define PoCos as a set of loci that cover at least a sufficient number of case samples while minimizing the number of covered control samples. We also investigate the effect of the threshold (α) we use to decide on what level of coverage is considered sufficient. For each $\alpha \in \{0.55, \ldots, 0.95, 1\}$, we generate PoCos with limited coverage of case samples and assess their performance in prediction of risk. The results of this analysis for the PS dataset are shown in Figure 8. As seen in the figure, reducing the coverage of case samples results in reduced performance PoCos in risk assessment. As we reduce the coverage of case samples, the number of loci in PoCos also goes down. Effectively, at a coverage threshold of around 55%, most PoCos contain a single locus. Therefore, the risk assessment performance of PoCos identified with a coverage threshold of 55% is similar to that of individual locus based models.

4. CONCLUSION

In this paper, we propose a novel criterion to assess the collective disease-association of multiple genomic loci (PoCos) and investigate the utility of these multiple-loci features in risk assessment. We also perform extensive experiments to evaluate the effect of using multiple genotype models, using network information to drive the search for multi-locus features, and the coverage of control samples on risk assessment. Moreover, we compare the proposed method with the polygenic score which has been shown to be successful in different studies. The result shows that our method is significantly more powerful in risk assessment.

Our results show that multi-locus features provide improved prediction performance as compared to individual locus based features. Interestingly, however, we observe

that integrating functional information provided by protein protein interaction data does not provide significant performance improvement. This is most likely to the limitation of PPI data to physical interactions among proteins; whereas many genomic variants may have regulatory effects on the function of proteins, as well as other molecules such as miRNA and lncRNA. Since PPI network-based search limits the focus on genomic loci that are in close proximity of genomic regions, relevant information may be lost by using only PPIs to incorporate functional information. An important benefit of using PPI networks, however, is that it reduces the search space to make the problem computationally feasible.

Based on the success of multi-locus genomic features in risk assessment, we conclude that combining these features with non-genetic risk factors and other biological data may lead to further improvements in risk assessment.

Acknowledgments

We would like to thank Matthew Ruffalo and Dan Savel for useful discussions. This work was supported in part by US National Institutes of Health (NIH) award R01-LM011247. 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.

5. REFERENCES

[1] Visscher PM, Brown MA, McCarthy MI, and Yang J. Five years of gwas discovery. *Am J Hum Genet.*, 2012.

[2] E. Zeggini, LJ. 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.

[3] 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.

[4] Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat Genet*, 41, 2009.

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

[6] Conde L., Bracci P.M., Richardson R., Montgomery S.B., and Skibola C.F. Integrating gwas and expression data for functional characterization of disease-associated snps: an application to follicular lymphoma. *Am. J. Hum. Genet.*, 2013.

[7] Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, and et al. Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nature Gentic*, 2008.

[8] Manolio TA, Collins FS, and etc. Finding the missing heritability of complex diseases. *Nature*, 2009.

[9] D.B. Goldstein. Common genetic variation and human traits. *N. Engl. J. Med*, 2009.

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

[11] K.E. Zerba, R.E. 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.

[12] M.M. Carrasquillo, A.S. McCallion, E.G. Puffenberger, C.S. Kashuk, N. Nouri, and A. Chakravarti. Genome-wide association study and mouse model identify interaction between ret and ednrb pathways in hirschsprung disease. *Nature Gentic*, 2002.

[13] Vawter MP1, Mamdani F, and Macciardi F. An integrative functional genomics approach for discovering biomarkers in schizophr. *Brief Funct Genomics*, 2011.

[14] Wan X and et al. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, 2007.

[15] J. Gui, JH 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.

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

[17] Gang Fang, Majda Haznadar, Wen Wang, Haoyu Yu, Michael Steinbach, Timothy R Church, William S Oetting, Brian Van Ness, and Vipin Kumar. High-order snp combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. *PloS one*, 7(4):e33531, 2012.

[18] Pierce BL and Ahsan H. Case-only genome-wide interaction study of disease risk. prognosis and treatment. *Genet Epidemiol.*, 2010.

[19] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, Douglas M Ruderfer, Andrew McQuillin, Derek W Morris, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

[20] International Multiple Sclerosis Genetics Consortium et al. Evidence for polygenic susceptibility to multiple sclerosisâĂŤthe shape of things to come. *The American Journal of Human Genetics*, 86(4):621–625, 2010.

[21] Matthew A Simonson, Amanda G Wills, Matthew C Keller, and Matthew B McQueen. Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC medical genetics*, 12(1):146, 2011.

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

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

[24] Zhang Y. and Liu J.S. Bayesian inference of epistatic interactions in caseâĂŞcontrol studies. *Nature Genetic*, 39, 2007.

[25] S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, and et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, 18:2078–2090, 2009.

[26] Emily M., Mailund T., Hein J., Schauser L., and M. H. Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17, 2009.

[27] Yu Liu, Sean Maxwell, Tao Feng, Xiaofeng Zhu, Robert C Elston, Mehmet Koyutürk, and Mark R Chance. Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from gwas data. *BMC systems biology*, 6(Suppl 3):S15, 2012.

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

[29] Marzieh Ayati and Mehmet Koyutürk. Prioritization of genomic locus pairs for testing epistasis. *Proceedings of ACM-BCB*, 2014.

[30] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao. dmGWAS: dense module searching for genome-wide

association studies in protein-protein interaction networks. *Bioinformatics*, 27:95–102, 2011.

[31] Marzieh Ayati, Sinan Erten, and Mehmet Koyutürk. What do we learn from network-based analysis of genome-wide association data? *Proceedings of Applications of Evolutionary Computation*, 2014.

[32] Holmans P, Green EK, Pahwa JS, Ferreira MA, and et al. Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet.*, 2009.

[33] Lingjie Weng, Fabio Macciardi, Aravind Subramanian, Guia Guffanti, Steven G Potkin, Zhaoxia Yu, and Xiaohui Xie. Snp-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 2011.

[34] Chloé-Agathe Azencott, Dominik Grimm, Mahito Sugiyama, Yoshinobu Kawahara, and Karsten M Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–i179, 2013.

[35] W. Li, B. Hu, G.L. Li, X.Q. Zhao, B.Z. Xin, and et al. Heterozygote genotypes at rs2222823 and rs2811712 snp loci are associated with cerebral small vessel disease in han chinese population. *CNS Neurosci. Ther.*, 2012.

[36] Zhang K, Wang YY, Liu QJ, Wang H, Liu FF, Ma ZY, Gong YQ, and Li L. Two single nucleotide polymorphisms in ALOX15 are associated with risk of coronary artery disease in a chinese han population. *Heart Vessels*, 2010.

[37] Huang R, Huang J, Cathcart H, Smith S, and Poduslo SE. Genetic variants in brain-derived neurotrophic factor associated with alzheimer's disease. *J Med Genet*, 2007.

[38] Can Yang, Xiang Wan, Qiang Yang, Hong Xue, and Weichuan Yu. Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. *BMC Bioinformatics*, 11, 2010.

[39] Salim A Chowdhury and Mehmet Koyutürk. Identification of coordinately dysregulated subnetworks in complex phenotypes. In *Pacific Symposium on Biocomputing*, volume 15, pages 133–144. World Scientific, 2010.

[40] W. T. C. C. Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 2007.

[41] Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, and et al. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 2012.

[42] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, and et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81, 2007.

[43] H. Lango, C. N.A Palmer, and et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Nature genetics*, 57, 2008.

[44] C. S. Janipallian, M. V. Kumar, and et al. Analysis of 32 common susceptibility genetic variants and their combined effect in predicting risk of type 2 diabetes and related traits in indians. *Diabetic Medicine*, 29(1), 2011.

[45] T.J. Russell, L.M. Schultes, and et al. Histocompatibility (HLA) antigens associated with psoriasis. . *N. Engl. J. Med.*, 287, 1972.

[46] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 2004.