Algorithms for Detecting Complementary SNPs within a Region of Interest That Are Associated with Diseases

Sinan Erten*
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio
sinan.erten@case.edu

Marzieh Ayati
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio
mxa401@case.edu

Yu Liu
Center for Proteomics and Bioinformatics
Case Western Reserve University
Cleveland, Ohio
yu.liu3@case.edu

Mark R. Chance
Center for Proteomics and Bioinformatics
Case Western Reserve University
Cleveland, Ohio
mark.chance@case.edu

Mehmet Koyutürk
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio
mxk331@case.edu

ABSTRACT

Genome Wide Association Studies (GWAS) comprehensively compare common genetic variants in affected and control populations to identify variants that are potentially associated with diseases. In recent years, GWAS successfully identified susceptible genes for many diseases. However, limitations of GWAS in uncovering the cellular mechanisms of complex diseases have been increasingly pronounced. In particular, GWAS analyze disease associations at the single variant level (e.g., single nucleotide polymorphism – SNP), however the functional links between these variants and the disease manifest at the level of genes, their products, and interactions. Since many genes are associated with multiple SNPs (within their coding and regulatory regions, i.e., regions of interest), it is not straightforward to characterize the association of individual genes with diseases based on SNP-level genotype data. Many of the existing studies that study functional implications of GWAS assess disease-gene association by simply taking the most statistically significant SNP in the gene's region of interest. Recently, some alternate approaches have been proposed to integrate the genotypes of all SNPs within the region of interest. In this study, we take an algorithmic approach to the problem and identify the optimal subset of SNPs that provide the maximum disease association score within each region of interest. The proposed algorithms represent the "genotype" of a gene as a combination of a subset of SNPs within its region of interest and search for the subset that maximizes the test statistic comparing this representative genotype in case and control samples. In order to handle the multiple testing problem, we compute the statistical significance of these scores by using permutation tests and using a background population that takes into account the number of variants lying in the region of interest (gene). We apply the proposed algorithms on a GWAS dataset for Type 2 Diabetes (T2D). To assess the performance of different algorithms, we use a manually curated set of genes known to be associated with T2D and compare different algorithms using ROC curves. Our experimental results show that the proposed algorithms are able to identify disease genes missed by other methods, with better sensitivity against the false positive rate.

Categories and Subject Descriptors

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

General Terms

Algorithms, Design, Experimentation

Keywords

GWAS, case-control studies, summary statistics

1. INTRODUCTION

One of the most common type of genomic variations among humans are Single Nucleotide Polymorphisms (SNP). With the emerging sequencing technologies, it is possible to genotype up to a million SNPs in parallel [4, 19]. Studying SNP patterns among populations are of great interest to researchers as they reveal associations between genotype and phenotype effectively. Genome Wide Association Studies (GWAS) commonly use statistical tests on case and control populations to discover SNPs that are associated with

*Corresponding Author.

ACM-BCB '12, October 7-10, 2012, Orlando, FL, USA

diseases [16]. The main objective in many genome wide association studies is to compare the allelic frequency differences across affected and unaffected samples in similar populations [10].

Although recent studies revealed hundreds of robustly replicated significant disease loci associations [6], GWAS are recently criticised due to several limitations they possess. These limitations include the following:

- The number of SNPs monitored is usually in the order of millions and rare variants with small effect are unlikely to be detected [21]. Existence of huge number of SNPs also bring computational challenges to multiple hypothesis testing [15].
- Most of the SNPs identified by GWAS do not show significant association with diseases, thus do not have clear functional implications [6].
- Individual SNPs identified to be associated with diseases show very limited success when used in classification of samples [16, 24].
- GWAS analyze associations at the SNP level, however the underlying functional mechanisms are best understood by analysis at the level of genes (and their products) [11].

As a result of these limitations, approaches that characterize disease progression based solely on genomic level are quite limited. Consequently, integrating genomic data with other sources of biological data such as PPI networks is crucial in complementing the GWAS and effectively utilizing genomic data. In order to achieve this, a necessary step is to move from SNP to gene level, by effectively combining the association scores of SNPs lying on a gene. The most common approach for this task is to directly use the association level of the most significant SNP [23, 7, 5]. Other simple approaches include taking mean of the χ^2 statistic of all SNPs in the region of interest [14], or considering only the top quartile of the variants. A recent survey compares the performances of these three basic summary statistics [11] and proposes a method of computing empirical p-values for these approaches.

Since the effect of individual variants might be very low (especially if they have low minor allele frequency), and the standard statistical methods such as logistic regression and Cochrane-Armitage trend tests are not applicable for those rare variants [3], scientists are often interested in pooling the scores of multiple SNPs within a genomic region. Moreover, Bansal *et al.* previously discussed the reasons to believe that multiple less common variants (either both on the same or different genes) collectively contribute to disease susceptibility in humans [1]. *Collapsing* is one such strategy for combining the effect of variants, where the variants generally with a low minor allele frequency (MAF) are pooled [12, 20, 17, 2]. Collapsing all variants naturally causes variants that are not associated with the disease to be included in the computations, thus they might hide the effect of association of actual variants with the phenotype of interest. Consequently, more advanced collapsing algorithms are proposed. The Combined Multivariate and Collapsing (CMC) method groups SNPs based on some criteria (such as MAF) and applies a multivariate test (such as Hotelling's $\mathcal{T}^2$) on the resulting collapsed sets of SNPs [12]. Dai *et al.* propose a greedy approach that first chooses the most significant common SNP within the region and greedily adds relatively rare SNPs, by choosing SNPs that provide the highest improvement of the association score, until there is no more improvement with the remaining rare SNPs [2]. Association levels of the collapsed variants are then computed using a univariate test.

Weighted Sum Statistic (WSS) proposed by Madsen and Browning computes an association score for each individual by assigning weights to each variant and aggregating them [15]. Scores for all case and control samples are then sorted and WSS is computed as the sum of the rank of affected samples. P-values of these scores are then estimated by permuting the disease status among samples 1000 times and computing standardized score using this background distribution. Similarly, Zawistowski *et al.* propose a pooling strategy called cumulative minor-allele test (CMAT), which utilizes aggregate allele counts of SNPs rather than collapsing them [25]. In this approach, rare allele counts are assigned weights and the weighted sum is computed for case and control samples separately. The weighting factor enables filtering out some variants or emphasizing some variants that are known to be functionally related to the trait, by assigning higher weights. Association level is later computed by a test analogous to Pearson's χ^2 statistic.

In this study, we propose two novel algorithms for effectively choosing a set of SNPs to be collapsed, to achieve a combined association score of the gene with the disease of interest. Our algorithms are based on the idea that, different variants might be responsible for the phenotype for different samples, *i.e.* they might complement each other in explaining genotypic variability in the affected population. First, we propose an *adaptive collapsing* algorithm, in which a set of SNPs are greedily chosen and added to the collapsed set, based on their improvement of the test statistic with respect to the phenotype. This algorithm is similar to the one proposed by Dai *et al.* [2], but there is no such restriction as collapsing rare SNPs on top of a common SNP. All SNPs, possibly after applying a MAF threshold filter, are treated equally during the search. We further improve upon this approach by developing a *set–cover based* algorithm that aims to explicitly optimize the complementarity of chosen SNPs in explaining the genotypic variability in case samples. Namely, the *set–cover based* algorithm chooses the representative set of SNPs based on how they present the relatively rare genotype in the case and control samples. This way, information about SNPs that are not strongly associated with the disease individually is also incorporated to the computations. Our algorithms do not necessarily focus on rare variants. After possibly filtering out very common variants (with a MAF above a user-defined threshold), we work on a set of relatively less frequent SNPs.

We start our discussion in the next section by introducing the genetic model used in this study. Subsequently, the formal definitions of the proposed algorithms are presented as well as the proposed approach for assessing the statistical significance of the scores calculated with these algorithms. We then present experimental results obtained by applying these methods on a GWAS dataset for Type 2 Diabetes (T2D) and a manually curated set of genes associated with T2D, to compare different algorithms using ROC curves. We discuss the results achieved and conclude the work in the Conclusion section.

2. METHODS

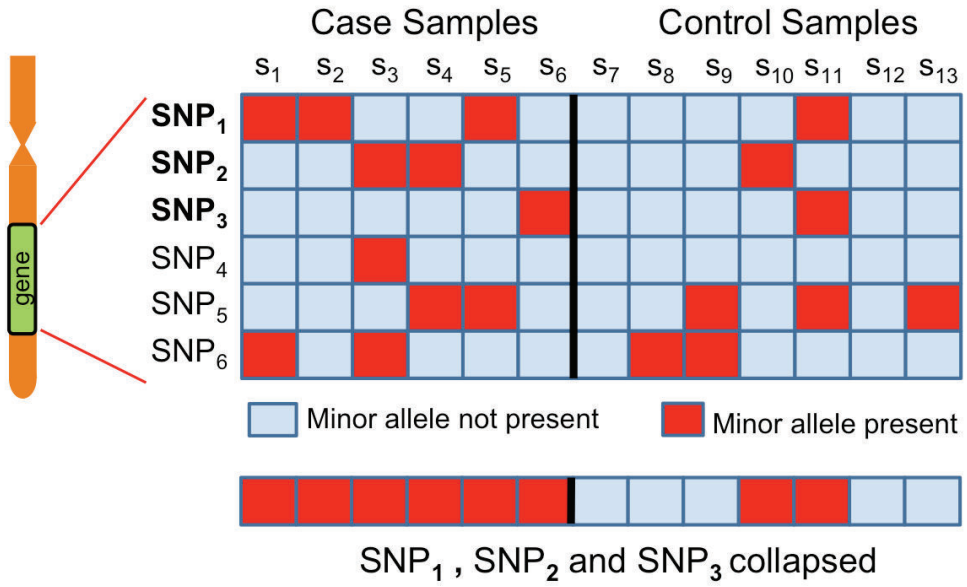


Figure 1: Illustration of the concept of covering SNPs. In this figure, each row represents the coverage of samples with those 6 SNPs across 13 samples. Red color indicates the existence of the minor allele, which means that sample is covered by the corresponding SNP. SNPs 1,2 and 3 collectively cover all case samples. In order to identify these covering variants, *set–cover based* algorithm would first add the SNP with highest coverage score($\mathcal{D}$) to the covering set and search greedily for other SNPs that maximally increase $\mathcal{D}$.

In this section, we first introduce the genetic model used in this study. Next, we present the two algorithms proposed, namely, *adaptive collapsing* and the *set–cover based* algorithms for combining the genotypes of SNPs that are within the region of interest of a gene. As mentioned previously, the proposed algorithms depend on the idea that SNPs might be complementary, in that different SNPs within the same genomic region might be associated with the phenotype of different subsets of samples. We use these algorithms to compute a representative genotype for each gene in the genome and compute test statistics for each gene using this representative genotype. Finally, we discuss how we compute the empirical p-values, from the gene-level test statistics computed using the proposed approaches.

2.1 Genetic Model

Consider a GWAS in which m SNPs are genotyped across n_a case and n_u control samples with $n_a + n_u = n$. In the simplest case, the genotype of the ith locus on the jth sample can take three values, AA, Aa and aa, where A and a denote alleles harbored on that locus in the population. Let $X_i(j) \in \{0, 1\}$ indicate the existence of the minor allele on ith locus of jth sample, *i.e.* $X_i(j) = 1$ if the minor allele is present and $X_i(j) = 0$ otherwise. Also let $P(j)$ denote the phenotype of the jth sample, such that $P(j) = 1$ if the sample is affected and $P(j) = 0$ if this sample belongs to the control population (assuming a dichotomous trait for the sake of simplicity). The association of a single SNP with the phenotype of interest is commonly calculated by comparing the frequency of the minor allele across the case and control populations, using Pearson's χ^2 statistics.

2.1.1 Collapsing Variants

As mentioned previously, effects of individual variants lying within a genomic region are often combined using various techniques. One such pooling strategy is *collapsing*. Let $\mathcal{L}$ denote a set of genomic variants, *e.g.* a subset of the SNPs within a region of interest. Also let $\mathcal{C}_\mathcal{L}$ denote the indicator vector of the collapsed variants $\mathcal{L}$, lying within the genomic region. Formally for sample j,

$$\mathcal{C}_\mathcal{L}(j) = \begin{cases} 1 & \text{if } \exists r_i \in \mathcal{L} : X_i(j) = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

This formula simply indicates whether a minor allele is present in one of the collapsed variants or not for the sample of interest. Pearson's χ^2 statistic (1df) can then be applied on the collapsed vector.

2.2 Adaptive Collapsing Method

Collapsing all variants in a genomic region might result in inclusion of many false variants, thus hiding the effect of SNPs that are actually associated with the phenotype. Consequently, choosing a representative subset of SNPs that capture the effect of the phenotype among the population might provide a better representation of the variants in that genomic region. Since enumeration of all possible combinations of variants assigned to a gene is intractable, we here propose a greedy algorithm for finding such set of variants. Let $\mathcal{K}$ denote the set of SNPs lying within the gene of interest. Since we are often interested in identifying combinations of SNPs with a low MAF instead of the very common ones, the first step is to filter the variants with a user defined MAF threshold(δ). Complete list of the steps of the algorithm is presented below:

1. Initialize the candidate set $\mathcal{Y}$ of SNPs to be collapsed: $\mathcal{Y} \leftarrow \{r_i \in \mathcal{K} : MAF(r_i) < \delta\}$, where δ is a threshold on MAF, for filtering common variants.
2. Initialize the set of SNPs to be collapsed: $\mathcal{L} \leftarrow \emptyset$.
3. For all SNPs r_i in $\mathcal{Y}$, compute the χ^2 statistic of $\mathcal{C}_{\mathcal{L} \cup \{r_i\}}$.

4. Add SNP r_i that provides highest improvement of χ^2 in previous step, to the collapsed set: $\mathcal{L} \leftarrow \mathcal{L} \cup \{r_i\}$.

5. Update the candidate set of SNPs: $\mathcal{Y} \leftarrow \mathcal{Y} \setminus \{r_i\}$.

6. If $\mathcal{Y} = \emptyset$ or none of the SNPs in $\mathcal{Y}$ provide an increase in χ^2 of the collapsed vector $\mathcal{C}$, return $\mathcal{C}$; otherwise, go to step 3.

This algorithm terminates when there are no more SNPs exist that improve the score of the collapsed vector.

2.3 Set-Cover Based Method

Here, instead of optimizing the statistical test score (χ^2) at each step, we propose the *set–cover based* algorithm that maximizes the difference between the frequency of case and control samples that harbor the minor allele. For this purpose, we first identify the set of samples that are "covered" by a SNP as follows. A SNP $r_i \in \mathcal{K}$ is said to cover a sample s_j if $X_i(j) = 1$. In other words, a SNP covers all samples in which the minor allele is present. We define $\mathcal{A}_i$ and $\mathcal{U}_i$ as the set of samples covered in case and control populations respectively, with r_i. We can generalize this definition directly to a set of SNPs: $\mathcal{V} \subset \mathcal{K}$ is said to cover a sample s_j if $\exists r_i \in \mathcal{V} : X_i(j) = 1$, *i.e.* the set of samples covered by a set of SNPs is equal to the union of the samples covered by each SNP. Observe that the cover of a given set of SNPs can be computed by collapsing them. The idea of SNP cover is illustrated in Figure 1.

We define the cover-based score $\mathcal{D}_\mathcal{V}$ of a set of variants as the difference of the fraction of covered case and control samples, *i.e.* $\mathcal{D}_\mathcal{V} = |\mathcal{A}_\mathcal{V}|/n_a - |\mathcal{U}_\mathcal{V}|/n_u$. The *set–cover based* algorithm iteratively adds the SNP that provides the maximal increase to the cover-based score to the covering set of SNPs.

1. Initialize the candidate set $\mathcal{Y}$ of SNPs to be collapsed: $\mathcal{Y} \leftarrow \{r_i \in \mathcal{K} : MAF(r_i) < \delta\}$.

2. Initialize the covering set of SNPs: $\mathcal{V} \leftarrow \emptyset$.

3. For all SNPs r_i in $\mathcal{Y}$, compute the cover-based score of the covering set of SNPs after adding r_i to $\mathcal{V}$, *i.e.* let $\mathcal{V}\prime = \mathcal{V} \cup \{r_i\}$, then $\mathcal{D}_{\mathcal{V}\prime} = |\mathcal{A}_{\mathcal{V}\prime}|/n_a - |\mathcal{U}_{\mathcal{V}\prime}|/n_u$.

4. Add SNP r_k that provides the highest increase to the cover-based score $\mathcal{D}$ calculated in the previous step to the covering set of SNPs: $\mathcal{V} \leftarrow \mathcal{V} \cup \{r_k\}$.

5. Update the candidate set of SNPs: $\mathcal{Y} \leftarrow \mathcal{Y} \setminus \{r_k\}$.

6. If $\mathcal{Y} = \emptyset$ or none of the SNPs in $\mathcal{Y}$ provide an increase to the cover-based score $\mathcal{D}$ in step 3, return $\mathcal{V}$; otherwise, go to step 3. The score of the final covering set is computed by using χ^2 statistic of the collapsed vector as in the *adaptive collapsing* algorithm.

This algorithm is based on the hypothesis that the cover-based score is an indicator of the separation between case and control samples. In order to validate this hypothesis, we first identify the covering set of SNPs for all genes using the *set–cover based* algorithm (see Section 3.1.1 for detailed information about the gene and GWAS datasets used). We then plot the cover-based score versus the association score (test statistic) of the covering set of SNPs for all genes in Figure 2 (association score is calculated on the collapsed set of covering SNPs). The relationship between the association and the cover-based scores supports the hypothesis that the cover-based score identifies the separation between case and control samples effectively. However, the relationship between the cover-based score and association score is not linear; therefore we expect that greedily maximizing the cover-based score (as opposed to maximizing the association score) may provide global awareness for the algorithm.

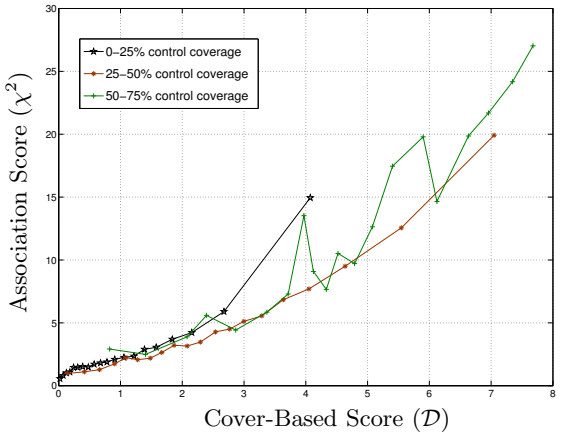


Figure 2: In this figure, each curve represents a different range of the fraction of covered control samples. Cover-based score of each gene is sorted and split into bins with equal number of genes. x and y-axes respectively represent the average cover-based score and the association score (test statistic) of all genes fall that into the corresponding bins.

2.4 Deriving Empirical P-values

In order to assess the statistical significance of the scores computed using any of the proposed algorithms, we apply permutation tests. For this purpose, we generate a large number of random datasets by permuting the disease status of the samples. Next, we run the proposed algorithms on each gene in the same way with the original dataset to get a background distribution of achievable disease association scores. We then compute the statistical significance of the original association scores by comparing them to this background distribution, as described below.

Due to the nature of the proposed algorithms, computed association scores depend on the number of SNPs that are within the region of interest for the respective gene. In other words, if we randomly assign SNPs to genes, we would expect genes with higher number of SNPs to have higher scores for both algorithms. This is the case for most of the existing algorithms as well [11]. We show the effect of the number of SNPs to the gene scores calculated using the *set–cover based* algorithm in Figure 3. A similar observation is made on the scores computed using *adaptive collapsing* (data not shown). Note that in this figure, we only consider those SNPs that have a MAF$< \delta$ when computing the number of variants within the genomic region (for our experiments, we use $\delta = 0.1$ as explained in the Results section.

Motivated by these observations, we assess the statistical significance of the association scores by also taking into consideration the number of variants that are mapped to the gene of interest. More precisely, the empirical p-value of an association score for a gene is calculated as the fraction of higher scores for the genes with similar number of SNPs in the background population. In our experiments, we permute

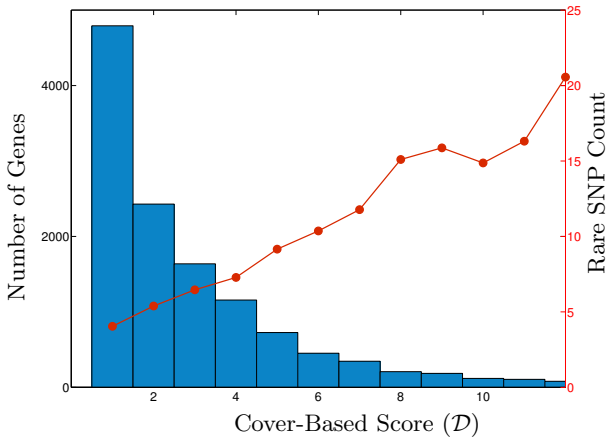


Figure 3: Effect of SNP count on the score computed using *set–cover based* algorithm. The bars show the histogram representing the distribution of the cover-based scores The red curve shows the average number of SNPs (with a MAF lower than a specific threshold) for the genes that are within the corresponding score range. The relationship between the association score and average number of SNPs mapped to the gene can be easily observed.

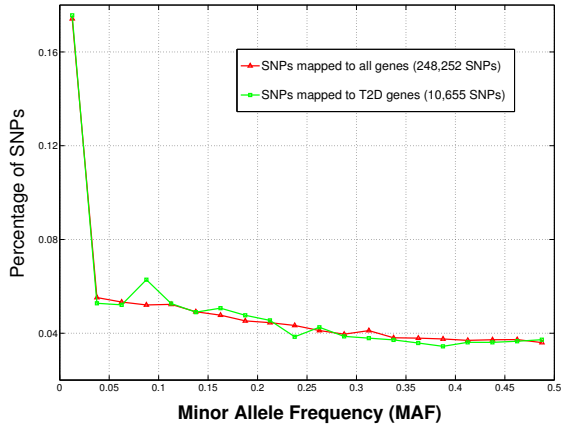


Figure 4: Distribution of the minor allele frequency values for SNPs mapped to all genes and SNPs mapped to 286 T2D related genes are shown.

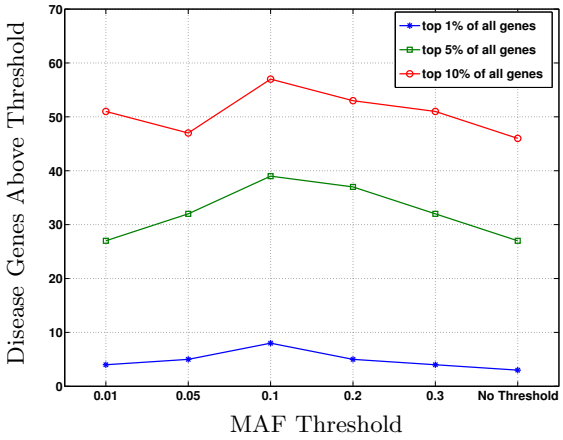


Figure 5: Number of correctly identified T2D related genes using the *set–cover based* algorithm for different values of the minor allele frequency threshold (δ) is shown. Different curves refer to different values of rank threshold used for a gene to be predicted as disease associated.

the disease status for 10^3 times and compute the association scores for all ~17000 genes (see the next section for details of the gene set used) for each of the permuted dataset. This provides ~10^7 association scores in the background, for a range of SNP counts. The empirical p-value for a gene is then computed by comparing it to the 10^4 background scores achieved with similar number of SNPs. Please note that, if the algorithm utilizes variants with a low MAF (if there is a filtering step of SNPs based on MAF threshold), we use the number of SNPs that pass the filtering stage, instead of all SNPs mapped to that gene, during the computation of the significance scores.

3. RESULTS

In this section, we start by describing the dataset and the experimental setup used for assessing the performance of alternate approaches to the problem. We then present and discuss in detail the results achieved with the proposed algorithms, as well as existing approaches.

3.1 Experimental Setup

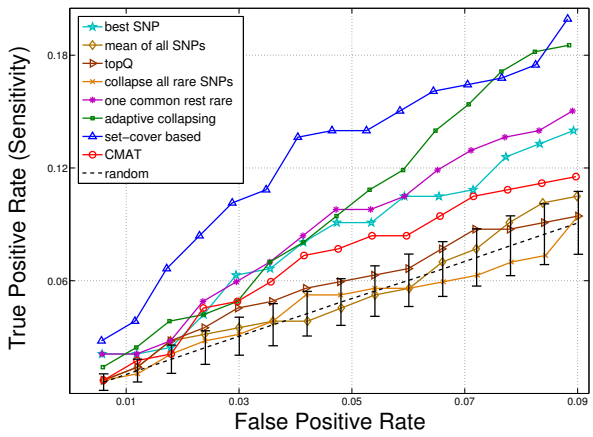
3.1.1 Datasets Used

We focus on analyzing the Type 2 Diabetes (T2D) samples (1999 samples), using the project samples from the 1958 British Birth Cohort (1504 samples) as control population. This dataset is provided by Wellcome Trust Case-Control Consortium (WTCCC) [22] and it contains ~500000 SNPs among the case and control populations. We define the region of interest for each gene as the region that extends 20kb upstream or downstream of the coding region for a gene. Thus, some SNPs might be mapped to multiple genes.

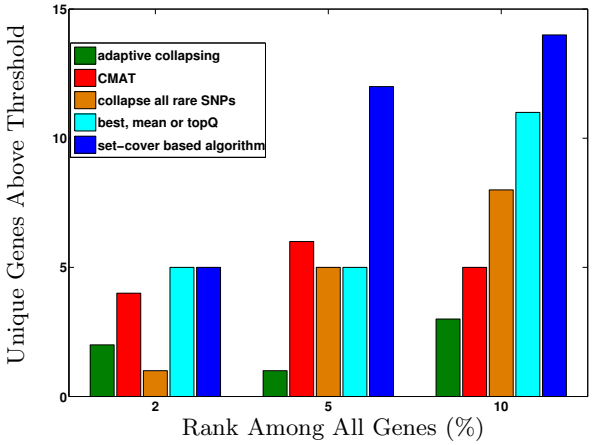
We use a manually curated database [13] to obtain a set of genes known to be associated with T2D. After removing the genes with only negative associations from this dataset, we have 286 genes with at least 1 SNP mapped from the GWAS data used. We use other 17119 human genes and that are not known to be associated with T2D. This gene list and the chromosomal location of each gene (used for mapping SNPs) are from Human Genome Assembly [9] provided by UCSC Genome Browser website (last access date: 02/02/2012) [8]. A total of 248252 SNPs are mapped to all genes tested, where 10655 of them are mapped to those 286 T2D related genes. Distribution of the MAF values of the SNPs used in our experiments is shown in Figure 4.

3.1.2 Performance Evaluation

In order to assess the performance, we compare the partial ROC curves of different algorithms. More specifically, we calculate the score for each gene using different algorithms, and plot the true positive rate (sensitivity) against false positive rate (1−specificity) with a varying rank threshold among the set of all genes scored (rank threshold is varied between 1%-10% of all genes). Sensitivity is defined as the proportion of true disease genes that are ranked above the particular threshold, whereas specificity is defined as the percentage of genes not known to be associated with the dis-



(a) Total number of disease genes identified



(b) Number of uniquely identified disease genes

Figure 6: In (a), fraction of the true disease genes identified with varying threshold, for different algorithms is shown. Genes are ranked based on the raw scores calculated using each method, thus not corrected with the number of variants in the genomic region of interest. In order to monitor the amount of overlap between identified disease genes using different algorithms, in (b), we show the number disease genes uniquely identified with different approaches.

ease, that are ranked below the threshold. We also present the number of disease related genes uniquely identified by each method.

3.2 Effect of the Minor Allele Frequency Threshold

Most of the existing algorithms for combining SNPs in a region of interest, focus on variants with a low MAF (usually less than 5%). Thus, in this section, we investigate the effect of the MAF threshold on the performance of the *set–cover based* algorithm. This threshold (δ) is used for filtering those SNPs with a high MAF in the first step of both of the proposed algorithms, as explained in Methods section. The number of identified true positives for different values of δ using the *set–cover based* algorithm is shown in Figure 5. In this figure, different curves refer to different values of the rank threshold used for a gene to be predicted as being disease associated. For all three rank thresholds (1%, 5%, 10%), $\delta = 0.1$ provides the highest number of correctly identified T2D related genes. Consequently, we use this threshold value for the rest of the experiments presented. Please note that the performance of *adaptive collapsing* algorithm yields a similar trend with varying δ (data not shown).

This result can be interpreted as follows: Using a very low threshold for δ causes many of the variants to be filtered out, thus resulting with too few information to be used. Allowing most of the SNPs on the other hand (using a high threshold, or no threshold) causes inclusion of false variants, adding bias to the scoring. Consequently, a threshold around 0.1 provides optimal results for the data in hand.

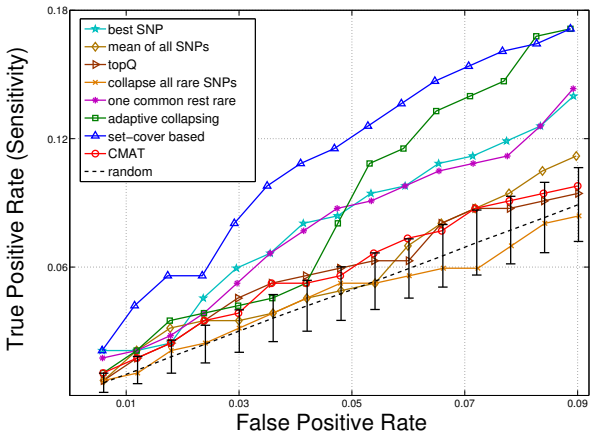
3.3 Performance Comparison of Different Algorithms

We compare the proposed algorithms and our implementation of existing methods, by plotting the fraction of correctly identified "true disease genes" (sensitivity) against the false positive rate (1−specificity) for each method tested. In other words, after computing the association scores for each gene and the statistical significance of these association scores, we rank all genes according to these scores and calculate the fraction of "true disease genes" that are ranked above a certain threshold among all genes *vs.* the false positive rate. We report the fraction of known T2D genes that are ranked in the top 1% to 10% by each algorithm.

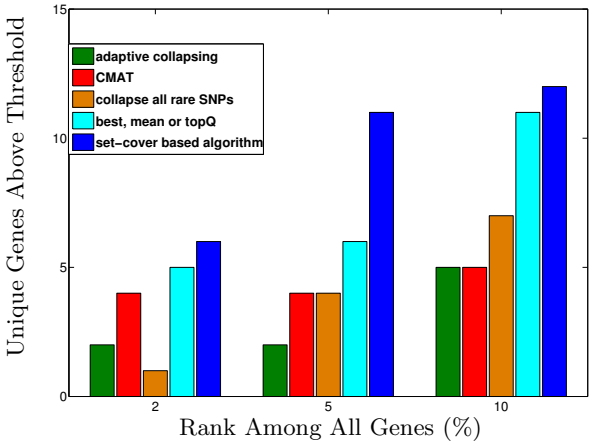
First, we briefly explain our implementation of existing algorithms. In the definitions below, SNPs with MAF $< \delta$ and MAF $\geq \delta$ are referred as *rare* and *common* SNPs respectively.

- *best SNP*: Use directly the most significant SNP as the association score of the gene.
- *mean of all SNPs*: Use the mean of the χ^2 statistic of all SNPs as the score of the gene.
- *topQ*: Sort all SNPs with respect to individual χ^2 statistic, then use the mean of the test statistic of SNPs in the top quartile.
- *collapse all rare SNPs*: Collapse all rare SNPs and apply a univariate statistical test on the collapsed vector.
- *one common rest rare SNPs*: Choose the most significant common SNP and collapse greedily other rare SNPs until the statistic of the collapsed vector does not improve.
- *CMAT*: Calculate the weighted minor allele count for case and control separately. Here, we follow the original study [25] and use weight as a inclusion parameter for rare SNPs and filter out common variants. We then use the formula analogous to χ^2 as introduced in [25] to assess the association scores with disease.

Please note that, the original versions of these algorithms might have some differences to our implementations. However, our aim in this study is to compare the core ideas of algorithms combining SNPs. Thus, we ignore the minor details and compare different ideas in a fair framework by using the same dataset. MAF threshold parameter (δ) and the method to calculate empirical p-values are similar for all algorithms.



(a) Total number of disease genes identified



(b) Number of uniquely identified disease genes

Figure 7: Empirical p-values used in this ranking are corrected with respect to the number of variants in the genomic region. This correction is done either using the number of all variants or only the rare ones, depending on the algorithm. In (a), fraction of the true disease genes identified with varying threshold, for different algorithms is shown. Again, in (b) we show the number disease genes uniquely identified with different approaches.

We present the results for both the raw and corrected scores with the number of SNPs assigned to the gene of interest. In Figure 6(a), we compare different algorithms by plotting the partial ROC curve with a varying rank threshold. In order to check the overlap between the identified genes and investigate the success of these algorithms in terms of identifying genes missed by other approaches, we also show the number of unique disease associated genes correctly identified by each algorithm in Figure 6(b). In this experiment, we use the raw association scores calculated by the corresponding algorithms, without correcting with respect to the background distribution of scores. Next, we rank the genes using the statistical significance scores corrected with respect to the number of variants (see Methods section for details). Results achieved using the corrected scores are shown in Figures 7(a) and 7(b). Observe that the performance of most of the different algorithms degrade with the correction. This is because the T2D associated genes have a higher number of assigned SNPs in average, thus assigned higher raw scores. This bias is partially removed with the correction with respect to the number of variants assigned to the genes. Both corrected and uncorrected results show that, the *set–cover based* algorithm is able to identify highest fraction of disease related genes, compared to other algorithms. It also performs better in terms of identifying (unique) disease genes missed by other approaches.

3.4 Linkage Disequilibrium of Covering SNPs

It was previously argued that the Linkage Disequilibrium (LD) between two variants with a low MAF is usually very low (often negligible) [12]. Moreover, collapsing algorithms are shown to be more robust in terms of power to the inclusion of variants in LD, compared to single-gene marker methods [12]. In this section, we investigate the existence of LD between the combined SNPs identified by the *set–cover based* algorithm, although it wouldn't affect the validity of the method. In Figure 8, we present the histogram of LD scores (correlation scores not squared) for two SNP pair set: (i) all SNP pairs lying on the same gene, and (ii) the SNP pairs combined by the *set–cover based* algorithm. LD scores for SNP pairs are calculated using PLINK [18]. It can clearly be observed that there are very few negative LD scores for the set (ii). This is because the LD score of two variants with a low MAF is very unlikely to be negative as LD here is calculated using correlations. Variants combined with the proposed algorithm tend to have a lower fraction of pairs with a high LD. We believe there are two reasons behind this observation. First, combined pairs consist of SNPs with a low MAF (less than 0.1 in our experiments) which in general, causes a lower LD between variants. Second, the *set–cover based* algorithm specifically chooses variants covering different groups of samples in case population, which naturally results in lower LD.

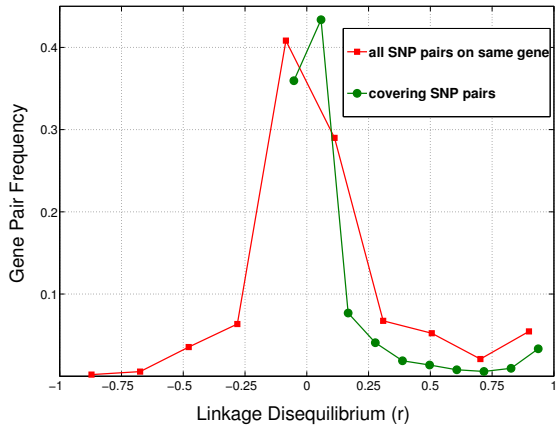


Figure 8: In this figure, histograms of LD scores for all SNP pairs lying on the same gene, as well as the SNP pairs combined by *set–cover based* algorithm are shown.

4. CONCLUSION

In this study, we proposed two novel algorithms for combining variants within a genomic region (usually a single

gene). We first presented a greedy algorithm that iteratively collapses SNPs based on the improvement they provide, in terms of the test statistic with respect to the phenotype. Second, we presented a set-cover based approach, aiming to choose those SNPs that cover all samples in case population, with as little control sample coverage as possible. For assessing statistical significance, we corrected the combined scores with the number of variants lying on the gene. As an application, we used the GWAS data provided by WTCCC for T2D, and applied proposed algorithms to identify genes associated with T2D. We showed that the proposed algorithms provide better performance in terms of identifying known T2D related genes, compared to existing algorithms. *set–cover based* also outperforms all other approaches in terms of identifying T2D related genes, missed by all other approaches.

5. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation (NSF) grant CCF-0953195, National Institutes of Health (NIH) grant R01LM011247 from the National Libraries of Medicine (NLM) and the Choose Ohio First Scholarship. 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.

6. REFERENCES

[1] V. Bansal, O. Libiger, A. Torkamani, and N. J. Schork. Statistical analysis strategies for association studies involving rare variants. *Nature reviews. Genetics*, 11(11):773–785, Nov. 2010.

[2] Y. Dai, L. Guo, J. Dong, and R. Jiang. Improved power by collapsing rare and common variants based on a data-adaptive forward selection strategy. *BMC Proc*, 5 Suppl 9, 2011.

[3] C. Dering, C. Hemmelmann, E. Pugh, and A. Ziegler. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genetic Epidemiology*, 35(S1):S12–S17, 2011.

[4] H. J. Edenberg and Y. Liu. Laboratory methods for high-throughput genotyping. *Cold Spring Harbor Protocols*, 2009(11):pdb.top62, 2009.

[5] C. C. Elbers, ..., and N. C. Onland-Moret. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genetic Epidemiology*, 33(5):419–431, 2009.

[6] L. A. Hindorff, ..., and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[7] M. Holden, S. Deng, L. Wojnowski, and B. Kulle. Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–2785, 2008.

[8] W. J. Kent, ..., Haussler, and D. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.

[9] E. S. Lander and et. al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001.

[10] C. C. Laurie, ..., B. S. Weir, and GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic epidemiology*, 34(6):591–602, Sept. 2010.

[11] B. Lehne, C. M. Lewis, and T. Schlitt. From SNPs to Genes: Disease Association at the Gene Level. *PLoS ONE*, 6(6):e20133+, June 2011.

[12] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *AJHG*, 83(3):311 – 321, 2008.

[13] 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*, 10(76), 2010.

[14] J. Z. Liu, ..., and S. Macgregor. A versatile gene-based test for genome-wide association studies. *AJHG*, 87(1):139–145, 07 2010.

[15] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 02 2009.

[16] T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

[17] S. Morgenthaler and W. G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615:28 – 56, 2007.

[18] S. Purcell, ..., and P. C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *AJHG*, 81(3):559–575, Sept. 2007.

[19] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5):e1000477, 05 2009.

[20] Y. Sung, T. Rice, and D. Rao. Application of collapsing methods for continuous traits to the genetic analysis workshop 17 exome sequence data. *BMC Proc*, 5 Suppl 9, 2011.

[21] E.-K. Tan. Genome-wide association studies: Promises and pitfalls. *Annals Academy of Medicine Singapore*, 39(2):77–78, 2010.

[22] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.

[23] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *AJHG*, 81(6):1278–1283, 2007.

[24] N. R. Wray, M. E. Goddard, and P. M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17(10):1520–1528, 2007.

[25] M. Zawistowski, ..., and S. ZÃűllner. Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. *AJHG*, (5):604–607.