Published in final edited form as:

ACM BCB. 2012 October ; 2012: 313–320. doi:10.1145/2382936.2382976.

A Collective Ranking Method for Genome-wide Association Studies

Jie Liu,

Dept. of Computer Sciences, Univ. of Wisconsin-Madison

Humberto Vidaillet,

Marshfield Clinic Research Foundation

Elizabeth Burnside,

Department of Radiology, Univ. of Wisconsin-Madison

David Page

Biostat. & Med. Info. Dept. Univ. of Wisconsin-Madison

Abstract

Genome-wide association studies (GWAS) analyze genetic variation (SNPs) across the entire human genome, searching for SNPs that are associated with certain phenotypes, most often diseases, such as breast cancer. In GWAS, we seek a ranking of SNPs in terms of their relevance to the given phenotype. However, because certain SNPs are known to be highly correlated with one another across individuals, it can be beneficial to take into account these correlations when ranking. If a SNP appears associated with the phenotype, and we question whether this association is real, the extent to which its neighbors (correlated SNPs) also appear associated can be informative. Therefore, we propose CollectRank, a ranking approach which allows SNPs to reinforce one another via the correlation structure. CollectRank is loosely analogous to the well-known PageRank algorithm. We first evaluate CollectRank on synthetic data generated from a variety of genetic models under different settings. The numerical results suggest CollectRank can significantly outperform common GWAS methods at the cost of a small amount of extra computation. We further evaluate CollectRank on two real-world GWAS on breast cancer and atrial fibrillation/flutter, and CollectRank performs well in both studies. We finally provide a theoretical analysis that also suggests CollectRank's advantages.

Keywords

Genome-wide association studies, Ranking; Breast cancer; Atrial fibrillation/flutter

1. Introduction

Genome-wide association studies (GWAS) analyze genetic variation, typically single-nucleotide polymorphisms (SNPs), across the entire human genome, searching for variations

that are associated with certain diseases or other phenotypes. GWAS can be naturally cast as a ranking problem, in which we seek a ranking of the SNPs in terms of the relevance to the given disease or phenotype. However, ranking SNPs in GWAS presents three challenges. First, GWAS are characterized by *high-dimension*. State-of-the-art technology enables measurement of a million SNPs in an individual in one experiment. Nevertheless, typical GWAS have at least two orders of magnitude fewer samples—cases and controls—than SNPs. While a million SNPs is a large number, these account for only 10% of the estimated number of common SNPs across the population [10]. Although the full set of known SNPs cannot be measured in one experiment at present, SNPs that are close together on the genome are often highly correlated, namely in high linkage disequilibrium (LD). Therefore, we can pick a set of tagSNPs such that the ungenotyped SNPs are in LD with the tagSNPs, either covered by one single tagSNPs or multiple tagSNPs [4, 6, 28, 11]. Hence the omission of some SNPs is not as much of a problem as one might first think. Instead, we have the problem of *strong-correlation* among our SNPs: most SNPs are highly correlated with one or more nearby SNPs, with r^2 (squared correlation coefficient) values well above 0.8. A third problem that makes GWAS especially challenging is the *weak-association* of the relevant SNPs with the phenotype, which makes the relevance measure for each SNP unreliable. One reason for weak association is that most diseases have both a genetic and environmental component. For example, genetics has been estimated to account for approximately 27% of breast cancer risk [14], and therefore a large fraction of the response variable is essentially noise when using genetic data only. Another reason for weak association is that many diseases, such as breast cancer, are polygenic, so that the genetic component is spread over multiple genes. Based on these two observations, we should expect the contribution from any one SNP toward predicting disease to be quite small.

Many practical GWAS [12, 7] employ hypothesis testing, such as Cochran-Armitage's trend test [5, 2, 20, 9], to test the association of each SNP and then rank the SNPs by the P -values from the tests. However in such a procedure, the correlations between the SNPs are totally ignored in the ranking. Here is a motivating example of our CollectRank approach, which shows that taking into account the correlation structure can provide a more efficient ranking. Suppose that we have already selected a small number of SNPs (e.g. the top 5%) according to the P -values from Cochran-Armitage's trend test, and SNP₁ and SNP₂ are in this top segment of the list. Suppose further that SNP₁ is highly correlated with SNP₃, and SNP₂ is highly correlated with SNP₄. However, SNP₃ is also in the top 5%, but SNP₄ is not. In the case, we would like to guess SNP₁ is more likely to be associated with the target class, whereas SNP₂ is more likely to be false positive that just appears promising by chance. The reason is that if one SNP is truly associated and there are several other SNPs correlated with it, they should be expected to show association with the class variable to some level too. Therefore, we would like to enable these SNPs to reinforce one another via the linkage disequilibrium structure. This is loosely analogous to the well-known PageRank algorithm [3] which allows the web-pages to boost each other by the links among them. We design our CollectRank algorithm to accomplish this through a HITS type of computation from the information retrieval literature, where correlations between SNPs are handled in a manner analogous to links between the webpages.

The next section makes this informal definition precise and gives a detailed description of the CollectRank approach. In Section 3, we evaluate CollectRank with synthetic data generated from a variety of genetic models under different settings, and show that the ranking provided by CollectRank is more efficient than common approaches. In Section 4, we further evaluate CollectRank on two real-world GWAS on breast cancer and atrial fibrillation/flutter respectively, and CollectRank shows outstanding performance in both studies. Section 5 provides a theoretical analysis in context of hypothesis testing. We finally conclude in Section 6.

2. Algorithm

When CollectRank (see Algorithm 1) evaluates the relevance of a SNP with the phenotype, it does not merely consider its individual relevance, but also the relevance of other SNPs with which it strongly correlates. We term these SNPs *neighbors*. CollectRank first evaluates the relevance of every single SNP with the phenotype, constructing an individual relevance vector just as common GWAS approaches do. Then it constructs a correlation matrix among SNPs from their pairwise correlations; these pairwise correlations may be taken from an external source, such as the HapMap database, or computed from the data itself. Based on the correlation matrix, CollectRank computes the collaborative relevance of the SNPs (collaborative relevance vector) from their neighbors, and eventually ranks SNPs according to the collaborative relevance vector.

Algorithm 1 CollectRank

Input: Labeled data with p SNPs, natural number k
Output: Top k SNPs sorted by importance
Procedure:
 for $i = 1$ to p do
 $IRV_i = \text{importance}(\text{SNP}_i)$
 end for
 Rank SNPs using IRV and keep top k SNPs
 for $i = 1$ to k do
 for $j = 1$ to k do
 $CM_{(i),(j)} = \text{correlation}(\text{SNP}_{(i)}, \text{SNP}_{(j)})$
 end for
 end for
 $CRV = CM \cdot IRV$
 Rank SNPs by CRV

The individual relevance vector (IRV) of p SNPs is the vector of length p whose i -th entry is SNP_i 's relevance with the class variable (phenotype) according to some measure of relevance. The most common measure in GWAS is the Cochran-Armitage's trend test (CATT) which is equivalent to logistic regression followed by a score test. Accordingly, we can use the absolute value of the test statistic from CATT to construct the IRV. Another commonly used measure is the logistic regression followed by a likelihood ratio test which is equivalent to the mutual information between the SNP and phenotype.

The correlation matrix (CM) of p SNPs is the $p \times p$ matrix whose (i, j) entry measures the correlation between SNP_i and SNP_j . Ideally, we would like to use the covariance matrix of IRV_i and IRV_j . However, it is usually not easy to derive it. In this case, we can approximate it with other correlation measures of SNP_i and SNP_j such as the absolute Pearson's correlation coefficient. When the sample size is small, the estimation of the Pearson's correlation coefficient is usually unreliable. The Pearson's correlation coefficient between the SNPs can be alternatively downloaded from HapMap [22].

The collaborative relevance vector (CRV) depicts our assessment of SNP relevance after we consider the correlations among them. The CRV of p SNPs is the vector of length p whose i -th entry is the weighted sum of the relevance of feature i and of all its neighbors. The weights are the entries in the corresponding row vector in the correlation matrix CM . Thus, the collaborative relevance vector can be calculated as:

$$CRV = CM \cdot IRV \quad (1)$$

Finally we use CRV to rank the SNPs.

3. Simulations

3.1 Simulation Settings

We simulate the linkage disequilibrium structure of a segment on chromosome 22 of humans, and compare the ranking yielded by our CollectRank approach and the ranking yielded by the common GWAS approaches such as the Armitage's trend test. We follow the simulation settings in the work of Wu et al. [25]. We use HAPGEN2 [21] and the CEU sample of HapMap [22] (Release 22) to generate SNP genotype data at each of the 2,420 loci between bp 14431347 and bp 17999745 on Chromosome 22. A total of 685 out of 2,420 SNPs can be genotyped with the Affymetrix 6.0 array. These are the typed SNPs that we use for our simulations. Within the overall 2,420 SNPs, we randomly select 10 SNPs to be the *causal* SNPs for our phenotype. All the SNPs on the Affymetrix 6.0 array whose r^2 values with any of the causal SNPs are above t are set to be the *associated* SNPs. In the simulations, we report results for three different t values, namely 0.8, 0.5 and 0.25. We also simulate three different models (additive model, dominant model, and recessive model) with different levels of homozygous relative risk (1.2, 1.3, and 1.4). In total, we simulate 250 cases and 250 controls. The experiment is replicated 100 times and the average result is provided. With the simulated data, we apply Cochran-Armitage's trend test and use the absolute value of the test statistic to construct the IRV. Since we can handle 685 SNPs in the calculation, we do not remove any SNP in the first step, namely we keep $k = 685$. Linkage disequilibrium information from HapMap contains the Pearson's correlation coefficients among the SNPs, and we use them to construct the correlation matrix in CollectRank algorithm. The CollectRank algorithm then computes the CRV. Finally, we compare the ranking yielded by IRV and the ranking yielded by CRV. We regard a correctly recovered associated SNP as a true positive and evaluate the rankings with ROC curves and PR curves. Since the experiments have been repeated 100 times, we vertically average the ROC curves and we vertically average the PR curves from the replicates.

3.2 Simulations Results

The ROC curves and PR curves from the rankings yielded by IRV and CRV in additive models with three levels of relative risk are shown in Figure 1. The first observation is that when the relative risk is low (such as 1.2), the recovery of associated SNPs is very difficult using CATT, whose ROC curves are like random guessing. However after a simple computation of CollectRank, CRV provides a more efficient ranking, namely a higher true positive rate while controlling false positive rate on ROC curves, or a higher precision while

maintaining recall on PR curves. For all three levels of relative risk and all three values of t , the ROC and PR curves from CRV (our approach) completely dominate the curves from IRV. We further calculate the areas under the ROC curves and PR curves under different levels of relative risk. The results are summarized in Table 1. We conclude that the rankings from CRV are more efficient than the rankings from IRV in additive models.

For the dominant models, the ROC curves and PR curves from IRV and CRV are shown in Figure 2. The ROC curves and PR curves from IRV and CRV in recessive models are shown in Figure 3. From the ROC curves, we again observe that the ranking yielded by IRV is almost as same as random guessing, but by a simple computation of CollectRank, CRV provides a more efficient ranking. The areas under the ROC and PR curves in dominant models and recessive models are summarized in Table 2. The performances of IRV and CRV in dominant models and recessive models are comparable to those in additive models. The only difference is that when the $t = 0.8$ and $t = 0.5$, the PR curves of IRV dominate the PR curves of CRV in the low recall region.

3.3 Replications under Different Settings

We further replicate the simulations in two different ways to check whether the performance of CollectRank is stable. First, we replicate all the experiments under the exact same settings (all the three genetic models, different levels of relative risk, and different t values) except we use another testbed (a different segment of the human genome). The different segment has 3,025 loci between bp 62765 and bp 2499283 on Chromosome 10. A total of 837 out of 3,025 SNPs can be genotyped with the Affymetrix 6.0 array. These are the typed SNPs that we use for our simulations. Within the overall 3,025 SNPs, we randomly select 10 SNPs to be the causal SNPs. All the SNPs on the Affymetrix 6.0 array whose r^2 values with any of the causal SNPs are above t are set to be the associated SNPs. The results from the new testbed are almost the same as the results in Figure 1, Figure 2, Figure 3, Table 1 and Table 2. Second, we replicate the experiments under the exact same settings except we replace CATT with the two-proportion z -test in the construction of the IRV. The two-proportion z -test is another commonly used test in GWAS and is especially useful for theoretical analysis such as power analysis. More details about the two-proportion z -test are provided in Section 5.

The rest of the computation of CollectRank stays the same, and we compare the ranking from IRV with the ranking from CRV. Again the results remain comparable with our first results.

4. Real-World Applications

4.1 GWAS on Breast Cancer

The GWAS dataset on breast cancer for our experiment comes from the NCI's Cancer Genetic Markers of Susceptibility website (<http://cgems.cancer.gov/data/>). It includes 528, 173 SNPs for 1, 145 patients and 1,142 controls. A detailed description about this dataset can be found in the original paper [12]. The GWAS on breast cancer exhibits weak-association. The relative risk of the several identified SNPs are between 1.07 and 1.26 [17].

The reasons for weak association are that (i) it is estimated genetics only accounts for about 27% of breast cancer risk and the rest is caused by environment [14] and (ii) breast cancer and many other diseases are polygenic, namely the genetic component is spread over multiple genes. Therefore, given equal numbers of breast cancer patients and controls without breast cancer, the highest predictive accuracy we can reasonably expect from genetic features alone is about 63.5%, obtainable by correctly predicting the controls and correctly recognizing 27% of cancer cases based on genetic predisposition. For purposes of comparison, if we select SNPs based on published literature alone—SNPs that are already identified to be associated with breast cancer by other studies (the closest feature set we have the ground truth for this task)—we get a set of 19 SNPs. Using these 19 SNPs as input to leading classification algorithms, such as support vector machines, results in at most a 55% predictive accuracy.

We convert SNP value GG into 1, gG into 0, and gg into 1 where G stands for the common allele at this locus and g stands for the rare allele. This encoding method has been used in many similar studies such as [23]. We use symmetrical uncertainty [18] to calculate IRV so as to be consistent with the Fast Correlation-Based Filter (FCBF) algorithm [27]. We use absolute Pearson's correlation coefficient from HapMap to construct the CM. The correlation matrix is blocked into 23 sub-matrices with each one corresponding to one human chromosome.

Since we do not have complete ground truth of which SNPs are associated with the phenotype, we are unable to carry out the ROC/PR evaluation we used in our simulation experiments of Section 3. Instead, we treat CollectRank as a feature selection algorithm by selecting the top f SNPs from the ranking given by CollectRank. We compare the performance of supervised learning when coupled with CollectRank against performance of the same supervised learning approach coupled instead with other feature selection algorithms, including FCBF, mutual information, Relief [13], symmetrical uncertainty and Lasso penalized logistic regression [26]. Because linear SVMs have been shown to perform particularly well on high-dimensional data such as genetic data [23, 24], we employ it as our machine learning algorithm to test the performance of feature selection methods. All the experiments are run in a stratified 10-fold cross-validation fashion, using the same folds for each approach, and each feature selection method is paired with the linear SVM. For each fold, the entire training process (feature selection and supervised learning) is repeated using only the training data in that fold before predictions are made on the test set of that fold, to ensure a fair evaluation.

For all feature selection approaches, we tune the parameters in a nested cross-validation fashion [1]. In each training-testing experiment of the 10-fold cross-validation, we have 9 folds for training and 1 fold for testing. On the 9 folds training data, we carry out a 9-fold cross-validation (8 folds for training and 1 fold for tuning) to select the best parameters. All of FCBF, mutual information, Relief, symmetrical uncertainty and Lasso penalized logistic regression have one implicit parameter f , the number of features used for supervised learning. Similarly, CollectRank also has the f parameter in addition to the k parameter ($f < k$). We use grid search to tune k and f in CollectRank. Since domain experts estimate relevant features to spread roughly among the top 20% of SNPs, for the k parameter we try

5,000 (~1%), 10,000 (~2%), 15,000 (~3%), ..., 100,000 (~20%), in CollectRank. For the f parameter, we try 50, 100, ..., and 1000. Since we have almost equal numbers of cases and controls, and for computational efficiency, we use accuracy to measure the classification performance for both inner and outer cross-validation.

As listed in Table 3, linear SVM's average accuracy is 54.04% when CollectRank is used. When FCBF, mutual information, Relief, symmetrical uncertainty and Lasso penalized logistic regression are used, linear SVM's average accuracies are 51.68%, 50.64%, 50.90%, 50.63% and 48.75% respectively. We perform significance testing on the 10 accuracies from cross-validation, and CollectRank significantly outperforms all the other feature selection filters significantly at 0.05 level.

4.2 GWAS on Atrial Fibrillation/Flutter

Our GWAS dataset on atrial fibrillation/flutter (AFF) comes from Marshfield Clinic. The Personalized Medicine Research Project [15], sponsored by Marshfield Clinic, was used as the sampling frame to identify 314 atrial fibrillation/flutter cases and 314 controls. The project was reviewed and approved by the Marshfield Clinic IRB. Subjects were selected using clinical data from the Marshfield Clinic Cancer Registry and Data Warehouse. Cases were defined as patients having a specialist diagnosis of AFF and an electrocardiographic diagnosis of either atrial fibrillation and/or atrial flutter. Controls were defined to be patients who had had at least one electrocardiogram and no evidence of atrial fibrillation or atrial flutter. The subjects also had DNA samples that were genotyped using the Illumina HumanHap660 array, as part of the eMERGE (electronic MEDical Records and Genomics) network by [16]. In total 551,776 SNPs have been genotyped after the quality assurance step.

We directly apply CollectRank in the AFF study. We use mutual information to construct IRV, which is equivalent to using logistic regression followed by a likelihood ratio test. We use the linkage disequilibrium data (Pearson's correlation coefficients) from HapMap to construct the correlation matrix. We set parameter $k = 50,000$ and calculate CRV. Table 4 summarizes the details about the top 25 SNPs from CollectRank, including the chromosomes (Chr), the physical positions (PhyPos), and the per-allele odds-ratios calculated from the population.

Eleven of the SNPs in Table 4, including the top 5, all are in the gene SLC44A5 (also known as CTL5). CTL5 is a choline transporter. Choline is important for the synthesis of acetylcholine, and acetylcholine is already known to either induce or terminate atrial fibrillation under different conditions (known from work on animal models of AFF going back as far as 1954). Cholinergic causes have been identified as one of six major causes of AFF and are the only type of cause without an identified AFF-associated SNP as yet. Our result is the first known evidence for association between CTL5 and AFF and therefore deserves confirmation in another cohort and/or further study in the laboratory for possible confirmation.

The remaining fourteen top SNPs are all in the 9p21.3 locus. This locus already has been associated with two other conditions with some relationship to AFF—myocardial infarction

(MI, essentially heart attack) and diabetes—although our fourteen SNPs are roughly a megabase away from these previously-identified MI and diabetes SNPs. This specific region of this locus therefore also deserves further investigation.

We also evaluate CollectRank by coupling it with supervised learning. Using the same SVM approach as in the last section and the same cross-validation methodology, but keeping only the top 25 SNPs per fold, we achieve 55% accuracy in predicting AFF, obviously based on genetic predisposition alone.

5. Theoretical Analysis

In this section, we show theoretical aspects of CollectRank. We explain how the collaborative relevance vector in the calculation of CollectRank is an approximation of the relevance of the hidden SNPs. We also show that the power of the hypothesis test can increase after the calculation of CollectRank. We carry out the analysis in the context of the two-proportion z -test which is commonly used in theoretical analysis in GWAS [8]. The derivation of the two-proportion z -test requires the assumption that the two copies of a chromosome are independent, which allows us to use Bernoulli features (we can regard a minor allele at a locus as a head in coin flipping).

Suppose we have two Bernoulli features F_A and F_B (locus A and locus B on the chromosome) in a binary classification problem. Further suppose we have n i.i.d. positive examples (copies of chromosome in cases) and n i.i.d. negative examples (copies of chromosome in controls). F_A^+ denotes the random variable of F_A in the positive samples. F_A^- denotes the random variable of F_A in negative samples.

$$F_A^+ \sim \text{Bernoulli}(p_A^+), F_A^- \sim \text{Bernoulli}(p_A^-) \quad (2)$$

p_A^+ and p_A^- are the population probability that F_A is 1 in the positive and negative population, respectively. p_A is the population probability that F_A is 1 in the whole population. Accordingly, $\hat{p}_A^+$, $\hat{p}_A^-$ and $\hat{p}_A$ are sample-based versions of p_A^+ , p_A^- and p_A . Analogously we define F_B^+ , F_B^- , p_B^+ , p_B^- , p_B , $\hat{p}_B^+$, $\hat{p}_B^-$ and $\hat{p}_B$ in the obvious manner. In addition, the correlation coefficient of F_A and F_B is r and we assume that it is the same for the positive and negative population.

In the two-proportion z -test, the null hypothesis $\mathcal{H}_0$ is that the feature is not associated. The alternative hypothesis $\mathcal{H}_1$ is that the feature is associated. The test statistic for F_A is

$$S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/n\sqrt{\hat{p}_A(1 - \hat{p}_A)}}}. \quad (3)$$

S_A is approximately normally distributed with variance 1 and mean $\lambda_A\sqrt{n}$ where

$$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2\sqrt{p_A(1-p_A)}}} \quad (4)$$

$\lambda_A\sqrt{n}$ is termed the *non-centrality parameter*. Under $\mathcal{H}_0$, S_A is approximately standard normally distributed. Under $\mathcal{H}_1$, S_A is approximately normally distributed with variance 1 and mean $\lambda_A\sqrt{n}$. The power of the test, the probability of identifying associated feature F_A , at some significance level α is,

$$P(\alpha, \lambda_A\sqrt{n}) = 1 - \frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{n}}^{\Phi^{-1}(1-\alpha/2) + \lambda_A\sqrt{n}} e^{-\frac{x^2}{2}} dx \quad (5)$$

where Φ^{-1} is quantile function of the standard normal distribution [8]. For any given significance level α , the power of the test is entirely determined by the non-centrality parameter. For a given sample set, the larger λ_A we have, the larger the power of the test.

Suppose F_A is strongly correlated with another binary feature F_H , and F_A is independent of the class variable given F_H . In other words, F_A works as a proxy for F_H . One result is that if we only observe F_A , and F_H is hidden, we will have the relation

$$\lambda_A = \lambda_H \sqrt{r^2} \quad (6)$$

where r is the correlation coefficient of F_A and F_H [19, 28]. By formula (5), if we only observe F_A we need more samples so as to achieve the same power.

Now suppose both F_A and F_B are proxies for F_H , and $(S_A; S_B)'$ is approximately distributed as a bivariate normal $N(\mu, \Sigma)$. Here $\mu = \sqrt{n} \begin{pmatrix} \lambda_A \\ \lambda_B \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where ρ is correlation

coefficient of S_A and S_B . Under our assumption, $\rho = r$. Suppose we want to build the test statistic S_H with the linear combination of S_A and S_B , S_H is also approximately normally distributed and it is on the direction where the variance is maximized. It is the largest principal component of $(S_A; S_B)'$, and the linear transformation of $(S_A; S_B)'$ is the leading eigenvector of the covariance matrix Σ . This result matches formula (1) if we set the individual relevance vector to be $(S_A; S_B)'$ and the correlation matrix to be the covariance matrix of the statistic. Therefore, the collaborative relevance of feature j in CollectRank is an approximation of the relevance of the hidden variable which is responsible for feature j and its neighbors.

It is easy to verify that the test statistic for F_H is

$$S_H = \frac{S_A + \text{sign}(\rho)S_B}{\sqrt{2(1+|\rho|)}} \quad (7)$$

and the non-centrality parameter for the hidden variable F_H is as formula (8).

$$\lambda_H \sqrt{n} = \frac{\lambda_A + \text{sign}(\rho) \lambda_B}{\sqrt{2(1+|\rho|)}} \sqrt{n} \quad (8)$$

λ_H is no less than the average of λ_A and λ_B . As long as F_A and F_B are not perfectly dependent on each other, we can have a larger non-centrality parameter and a larger power for any given significance level.

6. Conclusion

The key insight of CollectRank is to use the strong correlation among SNPs to magnify the weak association of individual SNPs. The extra computation of CollectRank (compared with common GWAS procedures such as CATT) is to use the correlation matrix to compute the CRV; this step has a computation time of $O(k^2)$. We have shown that this modest increase in computation enables CRV to provide a significantly more efficient ranking than IRV. In the simulations in Section 3 when relative risk is 1.2 or 1.3, IRV is no better than random guessing whereas CRV provides a ranking that captures a substantial signal. The results from the two real-world GWAS in Section 4 further confirm the performance of CollectRank. Section 5 explains the empirical improvement that CollectRank provides; this explanation comes within the framework of hypothesis testing, showing that CollectRank provides an increase in power for any given amount of data and specified significance level.

Acknowledgments

The authors gratefully acknowledge the support of the grant 1UL1RR025011 from the Clinical and Translational Science Award (CTSA) program of the National Center for Research Resources, National Institutes of Health, and the support from the State of Wisconsin through the Wisconsin Genomics Initiative. The authors also acknowledge NCI grant R01CA127379-01 and its ARRA supplement 3R01CA127379-03S1, NIGMS grant R01GM097618-01, NLM grant R01LM011028-01 and the UW Carbone Cancer Center.

7. References

- [1]. Ambroise C and McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6562–6566, 2002. [PubMed: 11983868]
- [2]. Armitage P. Tests for linear trends in proportions and frequencies. *BIOMETRICS*, 11:375–386, 1955.
- [3]. Brin S and Page L. The anatomy of a large-scale hypertextual web search engine In *Proceedings of the Seventh International Conference on World Wide Web 7, WWW7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998 Elsevier Science Publishers B. V.
- [4]. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, and Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74(1):106–120, 1 2004. [PubMed: 14681826]
- [5]. Cochran WG. Some methods for strengthening the common chi-square tests. *BIOMETRICS*, 10:417–451, 1954.
- [6]. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, and Altshuler D. Efficiency and power in genetic association studies. *Nature Genetics*, 37(11):1217–1223, 11 2005. [PubMed: 16244653]
- [7]. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struwing JP, ..., Mannermaa A, Kosma V-M, Kataja V, Hartikainen J, Day NE, Cox DR, and Ponder BAJ.

Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1093, 5 2007. [PubMed: 17529967]

- [8]. Eskin E. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Research*, 18(4):653–660, 2008. [PubMed: 18353808]
- [9]. Freidlin B, Zheng G, Li Z, and Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *HUM HERED*, 53(3):146–152, 2002. [PubMed: 12145550]
- [10]. Halperin E and Stephan DA. Maximizing power in association studies. *Nature Biotechnology*, 27(3):255–256, 3 2009.
- [11]. Han B, Kang HM, Seo MS, Zaitlen N, and Eskin E. Efficient association study design via power-optimized tag SNP selection. *Annals of Human Genetics*, 72(6):834–847, 11 2008. [PubMed: 18702637]
- [12]. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF, Hoover RN, Thomas G, and Chanock SJ. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–874, 5 2007. [PubMed: 17529973]
- [13]. Kira K and Rendell LA. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, pages 249–256, 1992.
- [14]. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, and Hemminki K. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*, 343:78–85, 2000. [PubMed: 10891514]
- [15]. McCarty C, Wilke R, Giampietro P, Wesbrook S, and Caldwell M. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *PERS MED*, 2:49–79, 2005.
- [16]. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struwing JP, Wolf WA, and eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC MED GENET*, 4(1):13, 2011.
- [17]. Pharoah PDP, Antoniou AC, Easton DF, and Ponder BAJ. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med*, 358(26):2796–2803, 6 2008. [PubMed: 18579814]
- [18]. Press W, Flannery B, Teukolsky S, and Vetterling W. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.
- [19]. Pritchard JK and Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 2001. [PubMed: 11410837]
- [20]. Slager SL and Schaid DJ. Case-control studies of genetic markers: power and sample size approximations for Armitage’s test for trend. *HUM HERED*, 52(3):149–153, 2001. [PubMed: 11588398]
- [21]. Su Z, Marchini J, and Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *BIOINFORMATICS*, 2011.
- [22]. The International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, 2003. [PubMed: 14685227]
- [23]. Waddell M, Page D, Zhan F, Barlogie B, and Shaughnessy J Jr. Predicting cancer susceptibility from single-nucleotide polymorphism data: A case study in multiple myeloma. In *Proceedings of BIOKDD ‘05*, Chicago, Illinois, August 2005, 8 2005.
- [24]. Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, and Hakonarson H. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, 5:e1000678, 2009. [PubMed: 19816555]

- [25]. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, and Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*, 86(6):929–942, 6 2010. [PubMed: 20560208]
- [26]. Wu TT, Chen YF, Hastie T, Sobel EM, and Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009. [PubMed: 19176549]
- [27]. Yu L and Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.
- [28]. Zaitlen N, Kang HM, Eskin E, and Halperin E. Leveraging the hapmap correlation structure in association studies. *Am J Hum Genet*, 80(4):683–691, 4 2007. [PubMed: 17357074]

Figure 1:

Performance of CRV (our approach, red solid line) and IRV (green dashed line) under additive models with different levels of homozygous relative risk and under different t values. (1a)-(1f) are for models with relative risk 1.2. (2a)-(2f) are for models with relative risk 1.3. (3a)-(3f) are for models with relative risk 1.4. (a)'s, (c)'s, and (e)'s are ROC curves with $t = 0.8$, $t = 0.5$ and $t = 0.25$ respectively. (b)'s, (d)'s and (f)'s are PR curves with $t = 0.8$, $t = 0.5$ and $t = 0.25$ respectively.

Figure 2:

Performance of CRV (red solid line) and IRV (green dashed line) under dominant models with different levels of homozygous relative risk and under different t values. (1a)-(1f) are for models with relative risk 1.2. (2a)-(2f) are for models with relative risk 1.3. (a)'s, (c)'s, and (e)'s are ROC curves with $t = 0.8$, $t = 0.5$ and $t = 0.25$ respectively. (b)'s, (d)'s and (f)'s are PR curves with $t = 0.8$, $t = 0.5$ and $t = 0.25$ respectively.

Figure 3:

Performance of CRV (red solid line) and IRV (green dashed line) under recessive models with different levels of homozygous relative risk and under different t values. (1a)-(1f) are for models with relative risk 1.2. (2a)-(2f) are for models with relative risk 1.3. (a)'s, (c)'s, and (e)'s are ROC curves with $t = 0.8$, $t = 0.5$ and $t = 0.25$ respectively. (b)'s, (d)'s and (f)'s are PR curves with $t = 0.8$, $t = 0.5$ and $t = 0.25$ respectively.

Table 1:

Area under ROC curves (AUC-ROC) and area under PR curves (AUC-PR) of IRV and CRV (our approach) for additive models with different levels of homozygous relative risk and different t values

	AUC-ROC			AUC-PR		
	$t=0.8$	$t=0.5$	$t=0.25$	$t=0.8$	$t=0.5$	$t=0.25$
rr:1.2						
IRV	0.524	0.516	0.514	0.044	0.066	0.113
CRV	0.649	0.665	0.674	0.081	0.130	0.200
rr:1.3						
IRV	0.523	0.520	0.519	0.045	0.071	0.117
CRV	0.644	0.666	0.678	0.092	0.148	0.221
rr:1.4						
IRV	0.572	0.557	0.537	0.058	0.080	0.125
CRV	0.671	0.689	0.691	0.101	0.158	0.227

Table 2:

Area under ROC curves (AUC-ROC) and area under PR curves (AUC-PR) of IRV and CRV in dominant models and recessive models with different levels of homozygous relative risk

Dominant Models						Recessive Models					
AUC-ROC			AUC-PR			AUC-ROC			AUC-PR		
$r=0.8$	$r=0.5$	$r=0.25$	$r=0.8$	$r=0.5$	$r=0.25$	$r=0.8$	$r=0.5$	$r=0.25$	$r=0.8$	$r=0.5$	$r=0.25$
rr:1.2											
IRV	0.515	0.524	0.520	0.049	0.075	0.115	0.506	0.505	0.511	0.042	0.106
CRV	0.642	0.669	0.677	0.082	0.140	0.215	0.648	0.668	0.677	0.080	0.201
rr:1.3											
IRV	0.520	0.520	0.524	0.054	0.075	0.116	0.516	0.520	0.515	0.045	0.115
CRV	0.650	0.673	0.687	0.097	0.143	0.218	0.649	0.674	0.680	0.081	0.210

Table 3:

The classification accuracy (%) of linear SVM coupled with different feature selection methods, CollectRank, FCBF, mutual information (Info), Relief, symmetrical uncertainty (SU) and Lasso penalized logistic regression (LPLR) in the breast cancer study

	CollectRank	FCBF	Info	Relief	SU	LPLR
%	54.04	51.68	50.64	50.90	50.63	48.75
P	–	0.050	0.022	0.023	0.005	0.003

Table 4:

Details about the top 25 SNPs identified by CollectRank in the AFF study

dbSNP ID	Chr	PhyPos	Odds-ratio
rs1249657	1	75562929	1.358
rs1505258	1	75575376	1.358
rs1249659	1	75573516	1.358
rs11163185	1	75582352	1.358
rs1249845	1	75586659	1.367
rs1330314	9	21306649	1.637
rs10117202	9	21313089	1.665
rs7031048	9	21296318	1.595
rs1249655	1	75565057	1.231
rs6656537	1	75558471	1.239
rs10811527	9	21283829	1.616
rs10122881	9	21279869	1.601
rs1330312	9	21285828	1.601
rs2657206	1	75567126	1.214
rs11163154	1	75573366	1.231
rs7548993	1	75546049	1.204
rs12030510	1	75560001	1.247
rs614541	9	21342862	1.719
rs10124609	9	21186907	1.703
rs7849623	9	21181463	1.687
rs4977682	9	21184379	1.680
rs12348971	9	21188384	1.601
rs1360286	9	21283631	1.545
rs7032044	9	21266891	1.465
rs4978092	9	21279923	1.450