Published in final edited form as:

Genet Epidemiol. 2010 April ; 34(3): 201–212. doi:10.1002/gepi.20448.

Comparisons of Multi-Marker Association Methods to Detect Association Between a Candidate Region and Disease

David H. Ballard^{1,*}, Judy Cho^{2,3}, and Hongyu Zhao^{3,4}

¹Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut

²IBD Center, Division of Gastroenterology, Department of Medicine, Yale University, New Haven, Connecticut

³Department of Genetics, Yale University, New Haven, Connecticut

⁴Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut

Abstract

The joint use of information from multiple markers may be more effective to reveal association between a genomic region and a trait than single marker analysis. In this article, we compare the performance of seven multi-marker methods. These methods include (1) single marker analysis (either the best-scoring single nucleotide polymorphism in a candidate region or a combined test based on Fisher's method); (2) fixed effects regression models where the predictors are either the observed genotypes in the region, principal components that explain a proportion of the genetic variation, or predictors based on Fourier transformation for the genotypes; and (3) variance components analysis. In our simulation studies, we consider genetic models where the association is due to one, two, or three markers, and the disease-causing markers have varying allele frequencies. We use information from either all the markers in a region or information only from tagging markers. Our simulation results suggest that when there is one disease-causing variant, the best-scoring marker method is preferred whereas the variance components method and the principal components method work well for more common disease-causing variants. When there is more than one disease-causing variant, the principal components method seems to perform well over all the scenarios studied. When these methods are applied to analyze associations between all the markers in or near a gene and disease status for an inflammatory bowel disease data set, the analysis based on the principal components method leads to biologically more consistent discoveries than other methods.

Keywords

multi-marker; association; power

INTRODUCTION

Genome-wide association studies (GWAS) have proven to be an effective approach for identifying the genetic components of complex diseases. According to the catalog of published GWAS [Hindorff et al., 2008], 1317 single nucleotide polymorphisms (SNPs) have been identified as being associated/causal of various diseases/phenotypes as of April

29, 2009. These studies represent GWAS using at least 100,000 SNPs genotyped on standard genotyping platforms analyzed predominantly through single SNPs. As technology continues to progress, these standard genotyping platforms are being improved in the amount of genomic coverage provided and the number of genetic markers interrogated. To address the need to analyze such rich information, many statistical methods have been proposed to more accurately identify associations between the genetic probes and disease state.

Despite these methodological developments, most published GWAS only employ single SNP analysis. In a typical GWAS data analysis pipeline, data are first filtered by removing markers with low call rates, low minor allele frequencies, or are not in Hardy-Weinberg equilibrium (HWE). Individuals not with a high proportion of missing genotypes are also removed. After such pre-processing, disease association analysis is performed for each SNP individually. If a researcher is fortunate, a SNP(s) located in a coding region of the genome is found to be significant after a multiple testing adjustment. However, this is generally not the case. Significant SNPs are often located in regions outside of annotated genes and/or no SNP will surpass the multiple testing threshold. The odds of no significant finding may increase as the coverage and number of SNPs increase on standard genotyping chips due to a higher penalty for multiple comparisons [Nannya et al., 2007]. In order to alleviate these issues, researchers have considered to use prior biological knowledge such as pathway and gene annotation in their analyses as a means to increase the statistical power for association studies. In order to appropriately incorporate this information, however, several issues must be addressed as detailed below.

Pathway-based analysis in gene expression research has been shown to provide great insight into the biological phenomena under investigation. Recently, this type of analysis has been adopted and applied to GWAS [e.g. Peng et al., 2008]. However, unlike gene expression studies where it is relatively easy to summarize gene expression levels from a single gene, GWAS requires calculating a gene score based on a few up to, perhaps, hundreds of SNPs. It is a non-trivial problem to develop a gene level score based on the available genotype information from these many SNPs. Wang and Elston [2007] were the first to publish a pathway-based analysis of a GWAS, and their gene level score was based on using the most significant SNP within the gene plus or minus 500 kb. Similar types of studies have followed and used a variety of gene level scores. For example, Peng et al. [2008] used several combination statistics (Fisher, Sidak, and Simes) where the overall evidence is summarized by combining information from all individual markers. The authors showed that despite the exclusion of linkage disequilibrium (LD) information among the markers, the gene level analysis had the ability to identify associations found using single SNP analysis and to identify new associations. Torkamani et al. [2008] described pathway analysis of seven common diseases studied by the Wellcome Trust Consortium. Their analysis used the same method of calculating gene significance, and identified biologically plausible pathogenic pathways for each disease. One must question how the results might change and possibly improve in these studies, if a more powerful gene level analysis is performed.

In order to create the best gene level score, a method should capitalize on all available information present. In 2005, the International Hapmap Project first described the extent of LD in the human genome in different populations [International HapMap Consortium, 2005]. The results suggested that SNPs were correlated with each other into block-like structures (haplotypes). These dependencies allow researchers to choose an informative subset of representative SNPs (also called tagging SNPs) for interrogation on genotyping platforms in order to provide the best coverage of the genome. In the context of genetic association studies, Akey et al. [2001] showed that exploiting LD information in testing for association could provide greater coverage of the genome and more power than single SNP

analysis to identify associations. The advantage of haplotype-based analysis is perhaps even more important when the disease is due to multiple loci within the gene or region [Morris and Kaplan, 2002], especially if those causal loci are uncorrelated. Applying LD information to gene level tests is an active area of research and many haplotype association methods have been proposed [Liu et al., 2008]. One limitation of haplotype analysis is that haplotypes are often unknown and have to be inferred for the analysis. Several studies [Chapman et al., 2003; Clayton et al., 2004] have shown that using multi-marker SNP genotypes directly in a regression setting without haplotype phasing may provide as much power to identify an association, as does haplotype analysis. Because haplotype inference is computationally expensive to do on the genome wide scale, and has relatively similar or poorer performance, we will not consider haplotype analysis in this manuscript.

An open question remains. Given the haplotype structure, should all genotyped SNPs within the gene be used, or only those SNPs that provide non-redundant information (tagging SNPs). The tagging process is also computationally expensive albeit less so than computing haplotypes. Roeder et al. [2005] reported a simulation study showing that the difference between using all SNPs available and only tagging SNPs was low. Ideally, we would like to use an association method that is able to handle all SNPs available within the gene without much computational overhead. Our analysis will compare the use of tagging SNPs to using all available SNPs in the gene region.

In an effort to address these issues surrounding gene level scores for GWAS, we compare seven multi-marker methods proposed in the literature to identify an association in the case-control study setting. These seven methods are briefly described below. We conducted both simulation and real data analysis to compare these methods. In our simulations, we used empirical haplotype data to represent the LD patterns at the candidate gene and assume that the disease is caused by one, two, or three disease-causing SNPs. The causal SNPs can be either tagging SNPs or non-tagging ones, and we compared using only tagging SNPs to identify the association to using all available SNPs within the gene. Our real data came from a GWAS performed on inflammatory bowel disease.

METHODS

The multi-marker association methods we evaluated can be divided into roughly three general categories: single SNP-based tests, tests based on fixed effects models, and tests based on random effects models. Each of these categories differs on whether and how they handle the correlation structure of the SNPs located within and near a gene. For example, the fixed effects methods either directly regress disease status on the SNP genotypes, or reduce the amount of variation in the SNPs into a lower dimensional space and use the main information-containing axes of variation to predict disease status. Secondly, some of the methods are score tests having a single degree of freedom despite the number of SNPs present in the gene. This is in comparison to the regression procedures where an additional degree of freedom is required for each SNP in the model. In the following, we briefly summarize the seven methods considered in this manuscript.

SINGLE SNP-BASED TESTS

The simplest method to determine association between a candidate gene and disease status is to use the best-scoring SNP as a summary of the overall association evidence. However, there are some comparison issues if the raw *P*-value from this best-scoring SNP is used without taking into account different complexities among genes, e.g. some genes may have only a few SNPs whereas others may have hundreds of or more SNPs. In our analysis, we used permutations to make multiple comparison adjustments. This was done by permuting the disease status and recording the best-scoring SNP for each permuted sample. Then the

empirical P -value can be estimated by comparing the P -value from the observed best-scoring SNP from real data to the distribution of the best-scoring SNPs from the permuted samples. This permutation procedure can account for the LD structure of the gene and the number of SNPs per gene. MinSNP denotes this test in our following discussion.

An alternative, based on single SNP evidence, is to combine the individual single degree of freedom tests of association for each SNP within the gene. Fisher's combination test summarizes information across SNPs by $-2\sum \log(P\text{-value})$. This procedure was used by Peng et al. [2008]. Because of the LD structure among SNPs, standard distribution theory used to calculate the P -value for this statistic does not apply here. Instead, we used permutations to obtain an empirical P -value. This test is denoted by COMBO in our following discussion.

FIXED EFFECTS METHODS

The fixed effects methods include regression based on individual genotypes [Chapman et al., 2003], transformed predictors based on principal components [Gauderman et al., 2007; Wang and Abbott, 2008], and transformed predictors based on Fourier transformations [Wang and Elston, 2007].

The most straightforward approach to jointly consider multiple SNPs is to use individual SNP genotypes as predictors, and it was found that this procedure had similar performance to haplotype methods [Chapman et al., 2003]. For ease of computation we used ordinary least squares (OLS) regression, not logistic regression. To assess the adequacy of the OLS method in this context, we performed both regression methods on 10% of the files using both tagging SNPs and all SNPs to ensure choice was appropriate. We computed a P -value for logistic regression using a likelihood ratio test, and then computed that average difference in P -value between the methods. Using all SNPs within the gene the average difference was 0.006, for tagging SNPs the difference was 0.001. We preferred OLS regression due to its speed, though it may not perform well due to multicollinearity problems when the SNPs within the gene are highly correlated. We expect that this will be less of an issue when using only tagging SNPs, but comparison between the results using all vs. tagging SNPs is one of our research questions. This method is denoted by REG in our following discussion.

Gauderman et al. [2007] and Wang and Abbot [2008] proposed to first reduce the number of predictors through principal components analysis (PCA) and then regress disease status only on the first few principal components. In contrast to REG, this method will not suffer from multicollinearity because the resulting components are orthogonal to each other. To implement PCA regression, the genotypes of the subjects must be converted to genotype scores. The scoring method can be count of the minor allele, but other scoring methods can also be used. PCA regression uses the correlation matrix of these centered genotype scores to identify the directions in which most of the variability in the data occurs. New variables are created as linear combinations of the observed data such that the new variables are uncorrelated and projected onto new axes ordered by the amount of variation they explain. The maximum number of new variables equals the number of SNP included in the correlation matrix, but given that the new variables are ordered by the amount of variation they describe, few principal components are necessary. Gauderman et al. [2007] and Wang and Abbot [2008] showed that using enough principal components to explain 80–90% of the variation is sufficient. For our PCA regression procedure, we used the number of principal components that explained at least 85% of the observed variation. This method will be denoted by PCA in our following discussion. In addition, we also consider doing regression based solely on the first principal component and this method will be denoted by PCA1 in the following.

Fast fourier transform (FFT) decomposes a function representing the minor allele frequencies of multiple SNPs into a set of more simple functions. These simple functions can be used to reconstruct the original function; an approximation of the original function can be made if only a few of the more informative simpler functions are used. Wang and Elston [2007] proposed a weighted score test using the real parts of the FFT components where weights are determined by the minor allele frequency (MAF). In order to maximize the weight for SNPs in high LD, it is necessary to ensure that the correlation matrix is as positive as possible. With this in mind, we implemented an algorithm to make the correlation matrix as positive as possible before calculating the test statistic. The algorithm involved computing the correlation matrix, tallying the total number of negative elements, then identifying the column with the most negative elements and reversing its genotype score as described in the original paper. We repeated this procedure until reversing the genotype of the column containing the most negative elements no longer reduced the total number of negative elements in the matrix. Chapman and Whittaker [2008] pointed out that the weighting scheme described by Wang and Elston [2007] may not be optimal, and that using only the first FFT component may be optimal. In this manuscript, we used the original implementation of the procedure, the test was performed using the R code provided at the original authors' website (<http://darwin.cwru.edu/~twang/wst>). This method will be denoted by FFT in the following discussion.

RANDOM EFFECTS METHODS

Tzeng and Zhang [2007] described a score test using a variance components method for detecting association between haplotypes and disease status. Their method uses a generalized linear mixed model (GLMM) that incorporates the correlation structure among observed haplotypes to test whether a haplotype effect exists. The GLMM allows for the inclusion of fixed effects (environmental covariates) and random effects (the haplotype effect), and allows for more accurate modeling of the biological phenomena that produce the disease-causing mutations. Their model is based on the assumption that haplotypes carrying disease variants are more likely to be similar to each other, and this can be captured in the covariance matrix for the haplotype random effects. In keeping with other methods, the authors suggest a correlation structure based on a haplotype similarity measure; however others are possible. The haplotype effect is assumed to be multivariate normal with mean equal to zero and variance equal to a constant multiplied by the haplotype correlation structure. Similar to FFT, a high degree of correlation among SNPs and with disease status is key to the method's ability to identify an association. Here we use the haplotype similarity (count of alleles shared among haplotypes) as the correlation structure and test whether the constant is equal to zero. The choice of this similarity matrix allows for the use of unphased SNPs. We used the code provided at the author's web site

(<http://www4.stat.ncsu.edu/~tzeng/Softwares/Hap-VC/R>). This method will be denoted by VC in the following discussion. We note that the method described in Goeman et al. [2006] and implemented by Chapman and Whittaker [2008] gives the same results as VC.

SIMULATION SETUP

We used the haplotype information for genes CHI3L2 and NAT2 studied by Kwee et al. [2008] as the basis of our simulations. These authors computed the haplotype frequencies by applying PHASE [Stephens et al., 2001] to genotype data obtained from Centre d'Etude du Polymorphisme Humain (CEU) genotypes of the International HapMap Project. Before applying PHASE, the authors identified tagging SNPs for each gene using the Tagger program [De Bakker et al., 2005] using an $R^2 > 0.8$. PHASE inferred a total of 45 haplotypes for NAT2. These resulting haplotypes contained 20 SNPs, 7 of which were considered tagging SNPs. The non-tagging SNPs had minor allele frequencies ranging from 29 to 44%, while the tagging SNPs' frequencies ranged from 1 to 41%. CHI3L2 had 66

distinct haplotypes containing 37 SNPs, 10 of which were considered tagging SNPs. The individual SNP frequencies for the non-tagging SNPs ranged from 1 to 56%, and the tagging SNPs ranged from 1 to 45%. CHI3L2 is nearly twice the size of NAT2, 15.8 kb compared to 9.9 kb, and has a more varied LD structure than NAT2. CHI3L2 is the same gene used in simulations by authors of other multi-marker methods [e.g. Wang and Elston, 2007].

For each gene we assumed HWE in the general population and simulated two sets of case-control samples, one assuming the causal variant is a tagging SNP; the other assuming it is a SNP not in the tagging SNP set. Different case-control samples were simulated assuming one, two, or three disease-causing SNPs. When one SNP was assumed to be disease causing, its MAF is equal to the sum of haplotype frequencies containing that SNP allele. We chose SNPs with MAF of at least 0.01 to be the causal SNP. For each selected diseases causing SNP, tagging or non-tagging, we simulated 100 samples each with 600 cases and 600 controls using three disease prevalence $k = (0.05, 0.1, 0.2)$ and 14 different effect sizes $\beta = (1.0, 1.05, 1.1, 1.15, 1.2, 1.25, 1.3, 1.35, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9)$. There were a total 84 sets of 100 simulated case-control samples. The power of each method is the proportion of the 100 case-control samples where the association was detected at a statistical significance level of 0.05.

By setting the disease prevalence k , the effect size β , and assuming a multiplicative genetic model, the baseline disease risk (β_0), i.e. those individuals with two normal alleles, can be calculated as $k/[\beta^2P^2 + 2\beta P(1-P) + (1-P)^2]$ under HWE for the disease-causing SNP with a risk allele frequency of P in the general population. For an individual with two haplotypes i and j , the haplotype effect (h_{ij}) for disease risk is β_0 , $\beta_0*\beta$, or $\beta_0*\beta^2$, respectively, depending on the number of disease-causing alleles present at the causal locus for the two haplotypes carried by this person. We can calculate the probability of each haplotype pair given disease status by standardizing each haplotype pair by the population prevalence: $P(h_ih_j|\text{Disease}) = h_{ij}*P(h_{ij})/k$ and $P(h_ih_j|\text{Disease}) = (1-h_{ij})*P(h_{ij})/(1-k)$, where $P(h_{ij})$ is the probability that a person carries haplotypes i and j . These probabilities were used for multinomial sampling of haplotypes of the 600 cases and 600 controls. For each simulated individual, we recorded the disease status and the genotypes at all the SNPs. Only SNP genotypes, not haplotypes, were used in association analysis.

When multiple causal SNPs are assumed to be associated with disease, each SNP is assumed to be in HWE and these SNPs jointly have multiplicative effects. For each model assumption, we can derive the conditional distributions of all the haplotype pairs given the disease status following a similar procedure shown above for the single disease-causing SNP case. SNPs were divided into one of three categories: high, medium, low based on the disease-causing allele frequency. A SNP was considered to have a high allele frequency (H) if its disease-causing allele was greater than 29%, medium (M) if greater than 10% and less than 29%, and low (L) if its allele frequency was less than 10%. We used a prevalence of 0.1 for all simulations with two or three causal SNPs. When two causal SNPs were simulated, we set one SNP with a constant effect size of 1.2 while three different effect sizes (1.1, 1.2, 1.3) were used for the second SNP. The two SNPs were then swapped and the simulations were performed again. This resulted in a set of six joint effects considered. This process was repeated for non-tagging causal SNPs using SNPs with comparable allele frequencies. The non-tagging causal SNP results were comparable to the tagging causal SNPs and are not reported here. The power results reported are for a single scenario. The simulations using three tagging causal SNPs are shown using a single effect size (1.2) for each of the SNPs involved. In order to compare the impact of adding more SNPs, we concentrate on reporting the results from SNPs that were used in all three categories.

After a data set was simulated, we applied the seven methods discussed above to test for association. In order to ensure adequate comparison of the P -values associated with each method, we report P -values based on permutation, not asymptotic distributions. For permutation analysis, we created 500 permuted data sets by permuting the disease status among the sampled 1,200 individuals. Each method used the same 500 permutations to determine its empirical P -value.

REAL DATA ANALYSIS

Evaluating how these methods work in a simulated environment allows us to identify the factors that affect the power of identification. Often, however, their power on real data is not directly analogous. Hence, we applied each of the methods described to a GWAS on inflammatory bowel disease [Duerr et al., 2006] to see which method was able to identify the most disease-related genes. This study includes two cohorts (Jewish, Non-Jewish) each of which we analyzed separately. SNPs with a call rate greater than 0.9, MAF > 0.01 and HWE P -value greater than 0.001 were kept in the analysis. Subjects with a call rate less than 0.95 were removed from the analysis. SNPs were considered mapped to a gene if their physical location was ± 10 kb from the start/end point of the gene as given by Refseq annotation at the NCBI website. The SNPs for each gene and disease status were then used for each of the methods described above. Missing values were removed for every method. A P -value was obtained via permutation for MinSNP and VC, otherwise the asymptotic P -value was used. In order to compare the methods we identified a set of 34 known Crohn's related genes from the literature [Duerr et al., 2006; Barrett et al., 2008] and counted the number of known genes that had a P -value less than 0.002 (value chosen due to the number of permutations performed). A statistical significance of overlap between the known disease genes and the genes identified for each method, for each cohort, was calculated using the hypergeometric distribution.

RESULTS

SINGLE CAUSAL SNP

The power of each method to identify an association was highly dependent on the allele frequency of the causal locus. Figure 1–Figure 3 show the typical results observed. To focus our comparison of the findings among the methods we will report results for CHI3L2, as the same trends were observed for NAT2. Each of the figures is comprised of two plots: the top plot shows the power of using all SNPs within the gene; the bottom shows the power using only the tagging SNPs. The figures are for populations where the prevalence was equal to 0.1. The power to identify an association increased with the prevalence. We do not plot the results when the prevalence was 0.05 and 0.2 because different methods had similar general patterns of relative performance, both for tagging SNPs and all SNPs.

In Figure 1, we show the power to identify a tagging causal SNP with a disease-causing allele frequency of 44%. FFT has the greatest power when all SNPs are used regardless of the effect size; however, it is closely followed by VC and PCA using only the first principal component (PCA1). The combination test (COMBO) is in the middle of all of the methods, and is followed by PCA, MinSNP, and REG. When only the tagging SNPs are used, there is a greater separation between the methods, and most methods see a slight improvement in power irrespective of the effective size. The order of the methods according to their ability to identify the association, however, does remain the same when compared to using all SNPs. At prevalence 0.05, using tagging SNPs has more power than using all SNPs except for a large difference for PCA when the effect size equals 1.25. As the prevalence increases, the power difference for PCA decreases.

Figure 2 demonstrates the power when a tagging SNP with a disease-causing allele frequency of 30% is the causal variant. This drop in allele frequency has the largest impact on FFT. When using all SNPs, FFT retains a high degree of power, but is outperformed by PCA1 and VC across all prevalences. The remaining methods perform worse, with REG having the least power. When tagging SNPs are used, FFT has a greater reduction in power. Now it is among the worst performing methods, while VC and PCA1 still retain their power. FFT, COMBO, and PCA1 have much greater power when using all SNPs within the gene. COMBO performs similarly to FFT in this situation: a moderate amount of power using all SNPs, but a large drop when tagging SNPs are used. REG is the only method to show an increase in power when using tagging SNPs across the three prevalences.

The plots in Figure 3 are based on a tagging causal SNP with disease-causing allele frequency of 2%. The power of all methods is greatly reduced in this scenario. When all SNPs are used, MinSNP performs best followed closely by REG; all others have minimal to no power to identify the association. When tagging SNPs are used, MinSNP and REG retain their power and PCA and COMBO become more able to identify the associations. The power of each of these methods does depend on the effect size. A causal allele with an effect size less than 1.5 is unable to be identified by any method.

Similar to the use of causal tagging SNP results, VC, FFT, and PCA1 have the highest power to identify an association across all effect sizes and prevalences when a non-tagging SNP is used as the disease-causing SNP. COMBO's performance is in the middle of the pack, and MinSNP, PCA, and REG have the least power. Using only tagging SNPs does generally provide greater power for the methods, with the exception of PCA when the penetrance is 0.1 and the causal allele has an effect size of 1.2. Lastly, for a non-tagging SNP with a disease-causing allele frequency of 17%, we see the broadest range of power to identify an association. PCA and MinSNP have the highest power, whereas FFT and PCA1 now have the lowest power. REG, VC, and COMBO fall in the middle. Using only the tagging SNPs, the ranks of the methods remain constant, but the difference between the methods is much smaller. Using tagging SNPs is more powerful for every method across the three prevalences. The largest benefit of tagging SNPs is for FFT and PCA1.

TWO CAUSAL SNPs

Table IA and B show the power of each method using all SNPs and only tagging SNPs to identify an association that is due to two causal tagging SNPs (SNP1 and SNP2). The results are broken down into 7 (A–F) combinations of high (H), medium (M), and low (L) disease-causing allele frequency SNPs: HH (44%, 30%), 2 HM (44%, 12%) and (44%, 17%), HL (44%, 2%), MM (12%, 17%), ML (17%, 2%), and LL (2%, 2%). When using all SNPs within the gene to identify the association, the power to identify the association for HH, which has a correlation of 0.73, ranges from 57–92%, 85–100%, and 98–100% when SNP1 has a set effect size of 0.1 and SNP2's effect size is equal to 1.1, 1.2, and 1.3, respectively. Similarly, the power of tagging SNPs ranges from 63–92%, 86–100%, and 99–100%. FFT, VC, and PCA1 have the highest power, whereas REG has the lowest power to identify the association in this scenario. There are two HM scenarios: one has a positive correlation of 0.42 (Table IA and B, section B), and the other a negative correlation (−0.41, Table IA and B, section G). The positively correlated scenario has more power at each effect size than the negatively correlated scenario. Using all SNPs, PCA is generally the most powerful for most effect sizes; however, using only tagging SNPs, VC and FFT are generally the most powerful. The power is almost twice as high for each method when the correlation is positive. In the negative correlation scenario, the power decreases as the effect size of the oppositely coded SNP increases. When an L SNP and an H SNP are the causal SNPs, the power is less than that of the HM positively correlated scenario, but greater than when HM is negatively correlated. The power among the methods is highly variable. For example, FFT

has a power of 71/77% (all/tagging SNPs) at an effect size of 1.2 for both SNPs, while REG has a power of 34/38%. The best methods in HL are FFT, VC, and PCA1, while REG, PCA, and MinSNP are the poorest performers. MM SNPs have a negative correlation of -0.19 and the most powerful methods are different than when there is a single causal SNP with a high frequency disease-causing allele. In this scenario, PCA is the most powerful followed by REG and Min SNP. The difference between using all SNPs and tagging SNPs is minimal for most methods, except for VC. For ML SNPs and a correlation of -0.08 , the power of the worst performing methods is slightly better than random. The most powerful methods in this scenario are PCA, MinSNP, and REG. Also, there is not a large difference between using all SNPs vs. tagging SNPs. Several methods have no power to identify an association when two causal SNPs had low allele frequencies. FFT, VC, PCA1, and COMBO have power less than or equal to 6% for each effect size. REG, PCA, and MinSNP perform better, with MinSNP and REG having the least difference between using all SNPs and tagging SNPs.

While FFT was not the best performing method in the two causal SNP case, we were curious to understand the practical impact of the algorithm used to increase the number of positive correlations in the correlation matrix. With that in mind, we also ran the FFT test on the positive and negatively correlated SNP discussed above (Table IA and B, sections B and G) without the algorithm to increase the number of positive correlations. We found that when the correlation was negative, the algorithm had no appreciable effect on identifying the association when only tagging SNPs were used. The average decrease in P -value with the algorithm was 0.002. If all SNPs were used the algorithm improved the identification of the association, an average decrease in P -value of 0.112 was observed. If the correlation between the two causal SNPs was positive, the algorithm greatly improved the identification of the association. For tagging SNPs the average decrease in P -value was 0.39, when all SNPs were used an improvement of 0.21 was observed.

THREE CAUSAL SNPs

For each of the three causal SNP scenarios, the effect size of every SNP was set to 1.2 and the prevalence was 0.1. The SNPs used varied in their disease-causing allele frequencies similar to the two causal SNP scenarios. Here we consider five different cases: LLL, LLM, LMM, MMM, and HMM. The difference in power between using all SNPs and tagging SNPs is minimal for all cases except for HMM. Here, the methods vary by a difference of 35% favoring tagging SNPs for PCA1 and by a 29% difference in favor of tagging SNPs for VC. In comparison, PCA has a 6% difference in favor of using all SNPs. There is no overall best performing method when three causal SNPs are used, and we will concentrate our discussion on using all SNPs. The combination of low frequency SNPs (LLL) results in low power for all methods, with REG and PCA performing best. LLM and MMM have similar levels of power where VC narrowly outperforms PCA, VC, and PCA1. Interestingly, LMM has over 20% greater power across the methods than MMM. Additionally, FFT performs exceptionally well with 62% power, the next best has power in the mid 50s. MinSNP performs the worst with 28% power to identify the association. In the HMM scenario, PCA has 70% power to identify the association, strongly outperforming the other methods. PCA1 has limited power (16%), while the other methods have mediocre power in the 40% range.

To better understand the effect of using multiple SNPs, we can pay closer attention to how the power of these methods works when the same SNP is used in all three simulations scenarios. For example, rs2255089 is a SNP with high MAF (44%). The power to detect rs2255089 in the single causal SNP scenario with 10% prevalence and effect size 1.2 using all SNPs within the gene has the following power: FFT 56%, MinSNP 37%, VC 53%, PCA 41%, PCA1 50%, REG 31%, and COMBO 46% (Fig. 1). When a positively correlated SNP (rs12070867) of medium MAF and effect size 1.2 is used as the second causal SNP (Table IA and B, section B), the power of all methods increases: FFT 80%, MinSNP 67%, VC

75%, PCA 81%, REG 68%, COMBO 69%, except for PCA1 50%. When a medium MAF, negatively correlated SNP (rs6685226) of effect size 1.2 is the second causal SNP, the power of all methods drops. FFT 41%, MinSNP 23%, VC 38%, PCA 32%, PCA1 32%, REG 24%, and COMBO 30% (Table IA and B, section G), although the most powerful methods remain the same. Lastly, we use each of the three above as causal SNPs. The HMM case combines the high MAF rs2255089 with the positively correlated rs6668814 and the negatively correlated rs6685226. Here the power of the methods varies widely: FFT 41%, MinSNP 49%, VC 44%, PCA 70%, PCA1 16%, REG 54%, and COMBO 40%. The power of FFT, VC, PCA1 has dropped dramatically, while the power of PCA is greater than the previous scenarios. The power of REG and COMBO remains mediocre.

Comparably, using rs1077059, a low MAF SNP, as the causal SNP, the methods have no power to identify the association when the effect size is 1.2 and penetrances is 0.1: FFT 6%, MinSNP 5%, VC 2%, PCA 3%, PCA1 3%, REG 4%, and COMBO 4% (Fig. 3). When a mildly negatively correlated (−0.08) SNP (rs961364) is used in combination to create an association, the power of most methods increases: FFT 7%, MinSNP 23%, VC 13%, PCA 26%, PCA1 6%, REG 19%, and COMBO 11% (Table 1A/B section E). PCA and MinSNP are the best at capitalizing on the new information. When another medium MAF SNP (rs961364), that is well correlated with rs6685226, is added to create a new association, we can see that FFT and VC are now the most powerful: FFT 62%, MinSNP 28%, VC 53%, PCA 48%, PCA1 54%, REG 36%, and COMBO 44% (Table II, column 3).

In order to get a view as to which method is best, we calculated the total number of associations identified by each method. This will neither give a definite answer as to which method is best considering the scenarios simulated cover only a very small fraction of the model space, nor does it contain an equal number of each type of combination, but it will give a feeling of which methods retain their power when multiple SNPs are the cause of the disease association. With that in mind, the best methods for the single causal SNP case in order of best to worst are: MinSNP, VC, PCA, COMBO, REG, FFT, and PCA1. When two causal SNPs are used the order from best to worst performing is: PCA, VC, MinSNP, FFT, REG, COMBO, and PCA1. Lastly, using three causal SNPs the ranking is: PCA, VC, FFT, REG, COMBO, PCA1, and MinSNP.

REAL DATA

The seven methods were applied to the Crohn's data sets, and the results are summarized in Table III. The power of all methods to identify the known genes was far better than random, based on the *P*-values calculated from the hypergeometric distribution. The rank of the methods from best to worse was close to the patterns emerging from our simulations. These ranks were also similar across the two cohorts studied. In the Jewish cohort, PCA identified 7 known genes in the 109 genes below the threshold of 0.002. This resulted in a *P*-value of 1.96×10^{-10} , the best result among all the methods. Interestingly, REG was the second best method, followed by PCA1, FFT, and MinSNP, these three had roughly the same performance. The worst performing method was VC. For the Non-Jewish cohort, PCA performed the best; REG, PCA1, MinSNP, and VC and comparable performance; and FFT had the worst performance.

DISCUSSION

In this paper, we have examined the power of seven methods proposed in the literature to identify an association between SNPs within a genomic region of interest (e.g. gene) with a binary phenotype. Our simulations considered associations due to one, two, or three SNPs under a variety of MAF and LD scenarios. When the association is due to a single causal SNP, the power of the methods was highly dependent upon the MAF of the causal SNP.

High allele frequency SNPs were best identified by FFT, VC, and PCA1; however, as the effect size grew above 1.3 and the prevalence was greater than 0.1, all methods performed well. While using tagging SNPs generally had more power than using all SNPs within a gene, the difference was generally acceptable compared to the added computation of determining the tagging SNPs on a genome-wide scale. When two SNPs were used to create the association, like the single SNP case, all methods performed well when allele frequencies were high. When the disease-causing alleles at the two SNPs were positively correlated, VC and FFT performed the best. It is here, though, that we can also investigate the impact of correlation among the SNPs on the ability to identify an association. A negative correlation between the SNPs reduced the power of some methods to identify associations. Primarily, the methods that had the highest power in the high MAF case had the greatest loss of power with a negatively correlated SNP (FFT, VC). Several methods had an increase in power (PCA, REG), while others retained comparable power as the effect size increased (MinSNP, COMBO). In the hardest situation, low MAF SNPs and negative correlation, MinSNP, REG, and PCA were the only methods that had power greater than random. Lastly, in situations where the MAF was at the medium level, PCA performed best and seemed less dependent on the sign of correlation among the SNPs. The interaction of MAF and correlation was also seen when an association was generated due to three SNPs. When all three SNPs had low MAF, only REG and PCA were able to identify an association, the others have no power. As the MAF increased, power increased. The interplay of correlation among the SNPs was an equally strong factor to determine the association. In general, if high MAF SNPs shared the majority of the correlation, it was not difficult to identify an association. Across all of the simulation scenarios we looked at, no method identified the most associations in every scenario, but summing across the scenarios, PCA performed best. Table IV summarizes our results.

These results agree with previously reported studies. Specifically, Chapman and Whittaker [2008] showed how the power of many of the methods used here plus Goeman's score test was similarly dependent on MAF. Using CTLA4 (average MAF = 0.27), IL21R (average MAF = 0.24), and CHI3L2 (SNPs with MAF greater than 20%, average MAF 0.34) as the basis of simulation, they showed that no matter which method was used, the power to identify association was highest in CHI3L2, followed by CTLA4, then IL21R. They also demonstrated a small difference in power between using all SNPs and only tagging SNPs. Lastly, based on their simulations, which used a single causal SNP, these authors concluded that Goeman's score test or the MinSNP performs the best in general and that MinSNP is a good default choice. Because the Goeman's method and the VC yielded essentially the same results, we made the same conclusion for the MinSNP and VC procedures based on our simulations. However, PCA, one of the best overall methods found in our study, was not included in their comparisons. In addition, neither these authors nor the authors of the original studies evaluated their procedures when the association was due to multiple SNPs. In haplotype association analysis context, Lin and Schaid [2008] compared a number of methods and concluded that distance-based methods are more robust, especially when there are many markers. This is consistent with our observation of the relatively good performance of the VC method. Additionally, Bacanu et al. [2008] found that PCA combined with their comprehensive single locus test, which tests three genetic models and adjusts the *P*-value by the number tests performed, had better performance than the single locus test alone. Their simulations also compared the use of several principal component tests independently; the best performing was the use of the first two principal components. They did not state the percent of variation these components explained. Their study also showed the poor performance of T^2 , which is the same as our REG method.

Of course, the biggest question remains—which multi-marker method is best? Given the simulations performed here, PCA performs the best when looking across all three scenarios.

The interpretation may, however, depend on the belief on how prevalent multiple causal mutations are expected to exist within a gene and the distribution of their MAF and the correlation among them. To get a better grasp of the MAF distribution, we examined the allele frequency data from a catalog of the successful GWAS studies to date maintained at <http://www.genome.gov/gwastudies>. The studies listed provide the allele frequencies of the SNPs found to be in association with one or more diseases. Figure 4 shows that most causal associations identified so far have a MAF greater than 0.05. Given the large MAF and the fact that this is the situation where MinSNP did not perform as well as other methods in our single SNP simulations, MinSNP may not in fact be the best method to use for identifying associations in the published studies. If we remove our 2% MAF simulation from the calculation for best method, the method ranks change to: VC, PCA, MinSNP, COMBO, FFT, PCA1, REG. MinSNP drops to third and VC and PCA's ranks improve. Hence with ranks of 1, 1, and 3 (or 2), PCA appears to be the most powerful method. However, the interpretation of Figure 4 must be taken with caution, as most published studies do not have adequate power to detect genes with small MAFs. Therefore, the lack of SNPs below 10% may simply reflect the lack of power in the published studies.

The availability of reliable gene level tests of association will allow researchers to aggregate the genetic information and provide biologically relevant interpretations to their findings. Our simulation and real data results demonstrate that choosing a more powerful multi-locus test may identify gene associations missed by single marker tests and may lead to more robust tests when incorporated into pathway or gene set enrichment. Our real data findings showed that PCA was the most powerful method, even though the set of known genes consists of genes identified by single marker tests of association. We believe that as multi-marker tests become better understood, the genes they identify will be worthy of biological followups.

Acknowledgments

We appreciate the insightful comments received by the reviewers that helped improve our manuscript. We thank Lydia Kwee and Michael Epstein for use of their haplotype data, and we also thank the Yale University Biomedical High Performance Computing Center and NIH grant: RR19895, which funded the instrumentation. This research was supported in part by NIH grants T15 LMG7055 from the National Library of Medicine, GM 59507, U01 DK062422, 1R01DK072373, and UL1 RR024139.

REFERENCES

- Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet.* 2001; 9:291–300. [PubMed: 11313774]
- Bacanu SA, Nelson MR, Ehm HG. Comparison of association methods for dense marker data. *Genet Epidemiol.* 2008; 32:791–799. [PubMed: 18551558]
- Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AH, Kistner EO, Murtha MT, Regueiro MD, Rotter JJ, Schumm LP, Steinhardt AH, Targan SR, Xavier RJ, NIDDK IBD Genetics Consortium; Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet.* 2008; 40:955–962. [PubMed: 18587394]
- Chapman J, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. *Genet Epidemiol.* 2008; 32:560–566. [PubMed: 18428428]

- Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* 2003; 56:18–31. [PubMed: 14614235]
- Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol.* 2004; 24:415–428. [PubMed: 15481099]
- De Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet.* 2005; 37:1217–1223. [PubMed: 16244653]
- Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science.* 2006; 314:1461–1463. [PubMed: 17068223]
- Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol.* 2007; 31:383–395. [PubMed: 17410554]
- Goeman JJ, van de Geer SA, van Houwelingen HC. Testing against a high dimensional alternative. *J R Stat Soc B.* 2006; 68:477–493.
- Hindorf LA, Junkins HA, Manolio TA. A Catalog of Published Genome-Wide Association Studies. 2008 Available from: <http://www.genome.gov/gwastudies/>.
- International HapMap Consortium. A haplotype map of the human genome. *Nature.* 2005; 437:1299–1320. [PubMed: 16255080]
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet.* 2008; 82:386–397. [PubMed: 18252219]
- Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet Epidemiol.* 2008; 33:183–197. [PubMed: 18814307]
- Liu N, Zhang K, Zhao H. Haplotype association analysis. *Advances in Genetics.* 2008; 60:335–405. [PubMed: 18358327]
- Morris RW, Kaplan N. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol.* 2002; 23:221–233. [PubMed: 12384975]
- Nannya Y, Taura K, Kurokawa M, Chiba S, Ogawa S. Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project. *Hum Mol Genet.* 2007; 16:2494–2505. [PubMed: 17666406]
- Peng, G.; Luo, L.; Siu, H.; Zhu, Y.; Hu, P.; Hong, S.; Zhao, J.; Zhou, X.; Reveille, JD.; Fin, L.; Amos, CI.; Xiong, M. Gene and Pathway-based analysis second wave of genome-wide association studies; *Nature Precedings*; 2008. Available from <http://hdl.handle.net/10101/npre.2008.2068.1>
- Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol.* 2005; 28:207–219. [PubMed: 15637715]
- Stevens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001; 68:978–989. [PubMed: 11254454]
- Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics.* 2008; 92:265–272. [PubMed: 18722519]
- Tzeng JY, Zhang D. Haplotype-based association analysis via variance-components score test. *Am J Hum Genet.* 2007; 81:927–938. [PubMed: 17924336]
- Wang K, Abbott D. A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol.* 2008; 32:108–118. [PubMed: 17849491]
- Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium Mapping. *Am J Hum Genet.* 2007; 80:360–355.

Fig. 1. Power comparisons of seven multi-locus association methods when the disease causing allele has a frequency of 44% and the disease prevalence is 10%.

Fig. 2. Power comparisons of seven multi-locus association methods when the disease causing allele has a frequency of 30% and the disease prevalence is 10%.

Fig. 3. Power comparisons of seven multi-locus association methods when the disease causing allele has a frequency of 2% and the disease prevalence is 10%.

Fig. 4.
The histogram of the minor allele frequency distribution of the SNPs that have been associated with one or more diseases through genome-wide association studies.

TABLE I

Power of seven multi-marker association methods to identify an association generated from two SNPs using only (A) all SNPs within the gene and (B) tagging SNPs

(A)		SNP2 (MAF 30%)			SNP2 (MAF 12%)			SNP2 (MAF 2%)		
A		1.1	1.2	1.3	B			C		
SNP 1 MAF 44% Effect 1.2		92	100	100	SNP 1 MAF 44% Effect 1.2			SNP 1 MAF 44% Effect 1.2		
FFT										
MinSNP		74	95	99						
VC		92	100	100						
PCA		70	95	99						
PCA1		92	100	100						
REG		57	85	98						
COMBO		81	98	100						
		SNP2 (MAF 17%)			SNP2 (MAF 17%)			SNP2 (MAF 2%)		
D		1.1	1.2	1.3	E			F		
SNP 1 MAF 12% Effect 1.2		4	4	15	SNP 1 MAF 2% Effect 1.2			SNP 1 MAF 2% Effect 1.2		
FFT										
MinSNP		10	16	42						
VC		6	14	39						
PCA		21	37	63						
PCA1		5	8	18						
REG		20	27	41						
COMBO		10	15	36						
		SNP2 (MAF 17%)			SNP2 (MAF 17%)			SNP2 (MAF 17%)		
G		1.1	1.2	1.3						
SNP 1 MAF 44% Effect 1.2		44	41	24						
FFT										
MinSNP		29	23	31						
VC		37	38	33						
PCA		30	32	44						
PCA1		37	32	42						
REG		19	24	35						

(A)	SNP2 (MAF 30%)				SNP2 (MAF 12%)				SNP2 (MAF 2%)																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
A					B					C																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
	1.1	1.2	1.3			1.1	1.2	1.3			1.1	1.2	1.3																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
(B)	COMBO	27	30	30	B																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																					

(A)	SNP2 (MAF 30%)			SNP2 (MAF 12%)			SNP2 (MAF 2%)			
	1.1	1.2	1.3	1.1	1.2	1.3	1.1	1.2	1.3	
A				B				C		
	COMBO			28	30	36				

TABLE II
Power of seven multi-marker association methods when the disease association is due to three SNPs

	LLL	LLM	LMM	MMM	HMM
FFT	2	27	62	37	41
FFT tag	3	25	54	31	60
MinSNP	7	28	28	22	49
MinSNP tag	7	27	29	20	56
VC	2	34	53	40	44
VC tag	3	33	63	41	73
PCA	10	32	48	35	70
PCA tag	13	25	46	34	64
PCAI	2	26	54	37	16
PCAI tag	2	26	61	38	51
REG	17	24	36	29	54
REG tag	15	28	38	25	58
COMBO	2	25	44	37	40
COMBO tag	12	27	46	34	55

The overlap between the validated genes associated with Crohn’s disease and those inferred from seven multi-marker association methods

TABLE III

Method	Jewish			Non-Jewish		
	# known genes	Total # genes	P-value	# known genes	Total # genes	P-value
REG	6	99	6.18×10^{-9}	3	98	3.94×10^{-4}
PCA	7	109	1.96×10^{-10}	4	104	1.57×10^{-5}
PCAI	5	90	1.96×10^{-7}	3	91	3.17×10^{-4}
FFT	5	88	1.75×10^{-7}	2	86	7.01×10^{-3}
MinSNP	6	173	1.74×10^{-7}	3	74	1.72×10^{-4}
VC	4	81	5.82×10^{-6}	3	83	2.42×10^{-4}

The Jewish and Non-Jewish data sets are analyzed separately. The *P*-value indicates the statistical significance of the overlap between the known gene set and the genes identified from the seven association analysis methods.

Summary of multi-marker association methods to identify associations on simulated and real data

TABLE IV

	Relies on high allele freq. of causal SNP	Good for low allele freq. causal SNP	Rank based on single causal SNP	Rank based on two causal SNPs	Rank based on three causal SNPs	Overall rank	Performance on real dataset Jewish	Performance on real dataset Non-Jewish
FFT	X		6	4	3	4	4	7
MinSNP		X	1	3	7	3	5	4
VC	X		2	2	2	2	7	5
PCA			3	1	1	1	1	1
PCAI			7	7	6	7	3	3
REG		X	5	5	4	5	2	2
COMBO			4	6	5	6	6	6