Published in final edited form as:
J Comput Biol. 2008 April ; 15(3): 241–257. doi:10.1089/cmb.2007.0090.

Prioritize and Select SNPs for Association Studies with Multi-Stage Designs

JING LI
Electrical Engineering and Computer Science Department, Case Western Reserve University, Cleveland, Ohio.

Abstract

Large-scale whole genome association studies are increasingly common, due in large part to recent advances in genotyping technology. With this change in paradigm for genetic studies of complex diseases, it is vital to develop valid, powerful, and efficient statistical tools and approaches to evaluate such data. Despite a dramatic drop in genotyping costs, it is still expensive to genotype thousands of individuals for hundreds of thousands single nucleotide polymorphisms (SNPs) for large-scale whole genome association studies. A multi-stage (or two-stage) design has been a promising alternative: in the first stage, only a fraction of samples are genotyped and tested using a dense set of SNPs, and only a small subset of markers that show moderate associations with the disease will be genotyped in later stages. Multi-stage designs have also been used in candidate gene association studies, usually in regions that have shown strong signals by linkage studies. To decide which set of SNPs to be genotyped in the next stage, a common practice is to utilize a simple test (such as a χ^2 test for case-control data) and a liberal significance level without corrections for multiple testing, to ensure that no true signals will be filtered out. In this paper, I have developed a novel SNP selection procedure within the framework of multi-stage designs. Based on data from stage 1, the method explicitly explores correlations (linkage disequilibrium) among SNPs and their possible interactions in determining the disease phenotype. Comparing with a regular multi-stage design, the approach can select a much reduced set of SNPs with high discriminative power for later stages. Therefore, not only does it reduce the genotyping cost in later stages, it also increases the statistical power by reducing the number of tests. Combined analysis is proposed to further improve power, and the theoretical significance level of the combined statistic is derived. Extensive simulations have been performed, and results have shown that the procedure can reduce the number of SNPs required in later stages, with improved power to detect associations. The procedure has also been applied to a real data set from a genome-wide association study of the sporadic amyotrophic lateral sclerosis (ALS) disease, and an interesting set of candidate SNPs has been identified.

Keywords

genome wide association studies; SNP selection; two-stage design

Address reprint requests to: Dr. Jing Li EECS Department Case Western Reserve University 10900 Euclide Ave. Cleveland, OH 44106 jingli@case.edu.

DISCLOSURE STATEMENT: No competing financial interests exist.

1. INTRODUCTION

Two grand challenges in the post-genomic era are to develop a detailed understanding of heritable variation in the human genome and to develop robust strategies for identifying genetic contributions to diseases and drug responses (Collins et al., 2003). Central to the understanding of such complex systems and the role of genes underlying diseases are effective mathematical models and software tools that can facilitate the characterization of genetic variation in human populations based on genomic data, mostly single nucleotide polymorphisms (SNPs) data, which has been available mainly due to the international efforts driven by the HapMap project (International HapMap Consortium, 2005). Genome-wide association studies are now feasible, and some are underway. On the other hand, studies have shown that a two-stage or multi-stage design might be able to achieve a similar power but with much reduced genotyping costs, comparing with a single-stage design (Hirschhorn and Daly, 2005; Satagopan and Elston, 2003; Skol et al., 2006; Thomas et al., 2004; Wang et al., 2006). An optimal two-stage/multi-stage design to achieve a minimum cost with a similar overall significance level and statistical power depends on many factors (such as disease allele frequencies, disease effects, the fraction of samples genotyped in stage 1, the fraction of markers genotyped in stage 2, and the genotyping cost ratio in stages 1 and 2). Several groups have investigated this issue using different statistical tests under different assumptions (Satagopan and Elston, 2003; Skol et al., 2006; Thomas et al., 2004; Wang et al., 2006).

The purpose of the first stage in a multi-stage design is to select a much smaller but promising subset of SNPs from all available SNPs for further tests. A common practice is to utilize a simple statistic to test the association of each SNP with the disease, and adopt a liberal significance level (per test) to ensure that no true signals will be filtered out. Because in principle, methods used in the first stage for SNP selection in a two-stage design can be applied to any later stages in a multi-stage design, I will focus my discussion on a two-stage design for simplicity. After the set of SNPs has been determined based on stage 1, there are usually three test strategies that can be adopted in stage 2, namely, replication-based analysis, joint analysis assuming homogeneity between stages, and joint analysis that allows heterogeneity between stages (Satagopan and Elston, 2003; Skol et al., 2006; Thomas et al., 2004; Wang et al., 2006). In a replication-based analysis, data in stage 2 are considered alone, and a positive association is reported if a statistical score reaches its significance level. In a joint analysis, subjects in stage 1 as well as those in stage 2 will be considered together at the end, while raw data from two stages are combined first to obtain an overall statistical score if assuming homogeneity, and the two statistics from two stages are combined if assuming heterogeneity (Skol et al., 2006). From now on, I use the term “joint analysis” for the former and “combined analysis” for the latter. For both types of analyses, the significance level of the new test statistic of each SNP has to be calculated conditional on the fact that the SNP has passed the screen in stage 1 (Skol et al., 2006). In addition, to obtain an experiment-wise significance level, Bonferroni corrections for multiple testing have been commonly used for all three methods (replicate-based, joint, and combined).

However, it is well-known that the method using Bonferroni corrections is highly conservative because, essentially, it assumes all SNPs are independent and in linkage equilibrium. The assumption does not hold any more when using SNP arrays with hundreds of thousands markers. It has been observed that many nearby SNPs are indeed in high linkage disequilibrium (Cancer Genetic Markers of Susceptibility project [CGEMS] at http://cgems.cancer.gov). Furthermore, when a SNP shows moderate association with the disease and is selected for stage 2, it is very likely that other SNPs in high correlations with this SNP will also be selected for stage 2. In this paper, I propose a novel procedure within the framework of a two-stage design to select a subset of highly discriminative SNPs based

on data in stage 1. The procedure consists of three phases. In the first phase, all the SNPs are ranked based on their associations with the disease as usual. Correlations among SNPs will then be explored using a clustering algorithm in phase two. In the third phase, potential haplotype effects and/or gene-gene interactions will be considered using an information theory–based approach. A SNP will be genotyped and tested in stage 2 only if it passes all three phases.

The procedure is flexible because it can incorporate different measures/approaches in each phase. More importantly, it is efficient for genome-wide association studies. It is different from tag SNP selection methods because it is used in a two-stage design, and phenotype information can be incorporated into the SNP selection problem. The efficiency and power thus can be improved for each specific study, while a regular two-stage analysis only consists of phase 1 and is a special case of the proposed procedure.

Extensive experimental studies using simulated and a real data set have been performed to systematically evaluate the proposed algorithm. In the simulation of a candidate gene study, I investigate the power, the number of SNPs prompted to stage 2, and the prediction errors of the proposed algorithm, and I compare its performance with the single-stage analysis and the regular two-stage analysis. The framework has also been applied to a genome-wide association study of the sporadic amyotrophic lateral sclerosis (ALS) disease (Schymick et al., 2007). The dataset is one of the first sets of publicly available genome wide SNP data. The genotypes of 555k SNPs from 276 cases and 271 controls, as well as phenotypes, are downloaded from Coriell Institute for Medical Research (http://ccr.coriell.org). Because of the small sample size of the dataset, a special treatment using the re-sampling technique has been adopted. An interesting set of candidate SNPs has been identified. The rest of the paper is organized as follows. Some notations and definitions are introduced in Section 2. The details of the algorithm are given in Section 3. Experimental results are presented in Section 4. I conclude the paper with some remarks about possible future work in Section 5.

2. PRELIMINARIES

2.1. Notations

I first introduce some notations and definitions that will be used later. Only diallelic SNPs will be considered. For a locus A, let A_1 and A_0 denote the two alleles, where $A_1, A_0 \in \{0, 1\}$. A genotype g at a locus is represented by 2, 1, or 0, where 2 means heterozygous genotypes, and 1 (or 0) means homozygous genotypes with the same allele 1 (or 0). A haplotype h over m loci is just a binary string with the length m. The allele at the kth position is denoted as $h(k)$. The frequencies of alleles at a locus A are denoted as p_{A_1} and $p_{A_0} = 1 - p_{A_1}$. The frequency of a two-loci haplotype $h = 11$ is denoted as p_{11}. Frequencies of other haplotypes can be defined similarly.

2.2. Haplotype estimation

Allele frequencies can be easily estimated using the maximum likelihood method based on genotype counts (Weir, 1996). The estimation of haplotype frequencies is a computationally challenging problem and has been intensively investigated recently. Despite the fact that some progress has been made, the calculation of haplotype frequencies for large scale genome-wide association studies requires tremendous computational resources and is not feasible in many cases. The proposed method in this paper will use the estimated haplotypes whenever available. But one of the advantages of the algorithm is that, if haplotype information is not available, it only uses haplotype frequencies over two or three loci that can be estimated easily (Weir, 1996).

2.3. Linkage disequilibrium measures

The linkage disequilibrium (LD) coefficient for two alleles A_1 and B_1 at two loci A and B is defined as the difference of the joint haplotype frequency and the product of the two allele frequencies: $D = p_{11} - p_{A1}p_{B1} = p_{11}p_{00} - p_{10}p_{01}$. Two normalized LD coefficients (D' and r^2) are in common use in practice, where

$$D' = \begin{cases} D/\min\left(p_{A_1}p_{B_0}, p_{A_0}p_{B_1}\right) & \text{if } D>0 \\ D/\min\left(p_{A_1}p_{B_1}, p_{A_0}p_{B_0}\right) & \text{if } D<0 \end{cases} \quad \text{and} \quad r^2 = D^2/\left(p_{A_1}p_{B_1}p_{A_0}p_{B_0}\right)^2.$$

Both measures can be used in the proposed algorithm. A detailed description of these two and some other measures can be found in Devlin and Risch (1995).

2.4. A close look at two-stage designs

When performing genome-wide association studies, not all existing SNPs (e.g., from dbSNP database) will be used. As a matter of fact, a set of pre-fixed array of SNPs, denoted as tag SNPs, across the genome is available to researchers from commercial companies (e.g., Affymetrix and Illumina), and it will be used for any disease association studies in any population. Figure 1 (top) illustrates all the steps in a two-stage design. The number of tag SNPs on those SNP chips is currently at the level of 300–500k, which is much smaller than the total number of discovered SNPs in the dbSNP database (about 11 million). The set of tag SNPs is selected primarily based on knowledge about human variation across the genome from HapMap data. The rationale of using a set of tag SNPs to represent all variation is based on the fact that the distributions of alleles at nearby loci are highly correlated and linkage disequilibrium across the genome observes a block-like structure with low haplotype diversity within each block. So a small subset of tag SNPs can approximate most common variation and can distinguish most common haplotypes within each block. A variety of algorithms have been proposed to select an optimal set of tag SNPs, and most of them (including the commercially available high-density SNP chips) are solely based on available HapMap data. On the other hand, although the total number of tag SNPs is much smaller than the total number of known SNPs, these tag SNPs are still not independent. First of all, tag SNPs are primarily selected based on linkage disequilibrium and a high threshold (e.g., $r^2 = 0.85$) is commonly used. In addition, tag SNPs are selected based on HapMap data, which consist of a relatively small number of samples from four populations. The problem of “over-fitting” might occur (i.e., tag SNPs selected from one population may not be able to represent or predict non-tag SNPs in another population, because of the differences of allele frequencies in different populations). The predictability of tag SNPs might be even worse in affected individuals for a particular disease. It has been reported elsewhere (Ahmadi et al., 2005; de Bakker et al., 2005) that loss of power might occur when using a prefixed set of tag SNPs for association studies in other populations. To overcome drawbacks of decreasing power using tag SNPs in other populations, one has to increase the density of tag SNPs. Therefore, many tag SNPs for genome-wide associations are themselves in high LD. And this has been observed in various studies such as the CGEMS (Cancer Genetic Markers of Susceptibility) project. For a candidate gene study, researchers can choose their tag SNPs based on available SNPs in dbSNP database and/or HapMap datasets, functional annotations in the candidate regions, and sometimes by resequencing the candidate regions. To ensure that statistical power will not be deteriorated because of the selection of tag SNPs, researchers might choose all available SNPs or a set of nonredundant SNPs for their studies. Therefore, high LD will also be observed in candidate gene studies.

3. METHODS

3.1. The framework

The proposed algorithm for SNP selection and prioritization is based on a two-stage design that aims to reduce genotyping costs and to possibly increase power of an association study by reducing the number of tests. Both genome-wide and candidate gene studies will be discussed. For a genome-wide association study, the algorithm mainly focuses on data in stage 1, which are sampled from the population under study with phenotype information for each individual. This is because the set of m_2 tag SNPs (Fig. 1) has been fixed beforehand, and researchers usually cannot choose tag SNPs by themselves. In a candidate gene study, researchers can select their own tag SNPs, and there are many existing tag selection algorithms. In this study, I take a simple approach and select all nonredundent SNPs as tag SNPs for candidate gene studies (Section 3.2). Then a fraction of samples (say π%, including both cases and controls) will be genotyped for all the m_2 tag SNPs in stage 1. A much smaller subset of m_3 SNPs is then selected from the m_2 SNPs based on correlations of SNPs and the phenotype/disease, as well as correlations among SNP themselves (Section 3.3). Only a subset of SNPs with low redundancy and high discriminative power will be kept. The remaining $(100 - \pi)$% samples will then be genotyped, but only for the m_3 SNPs. The power analysis will then be performed by combining the samples in the two stages (Section 3.4).

3.2. Tag SNP selection

Genome-wide tag SNP selection has been intensively studied recently, and the set of tag SNPs has been fixed in commercial SNP chips. Therefore, tag SNP selection is not an issue/option for genome-wide association studies. For a candidate gene study, researchers can choose their tag SNPs based on all available SNPs using any selection algorithm. For example, one can use the HapMap data as the reference panel, and a nonredundant set of m_2 SNPs can be selected from all m_1 SNPs in the regions. The most conservative approach is taken here to ensure that the statistical power will not be deteriorated because of the selection of tag SNPs. The algorithm for this step is simple. When haplotype information is available, one can directly check whether two SNPs are the same or different across all haplotypes. Only one SNP is kept from each set of such "equivalent" SNPs. When only genotypes are available, one can use the pairwise LD measure r^2 to remove all redundant SNPs. Because r^2 is directly related to the power of Pearson's χ^2 test, no information will be lost when removing a SNP if the value of r^2 between the SNP and a tag SNP is 1. A similar approach has been used in de Bakker et al. (2005). The pairwise haplotype frequencies, which are required in calculating r^2, can be efficiently estimated using the maximum likelihood approach discussed in Subsection 2.2.

3.3. Prioritize and select SNPs in stage 1

The goal of SNP selection in stage 1 is to select a minimum set of SNPs with highest discriminative power. It can be viewed as the classical problem of variable selection. Therefore, the principle of many existing variable selection methods can be applied here. However, because the number of variables (i.e., the number of SNPs) considered here is extremely large (about half million), most traditional variable selection approaches are not computationally feasible. This motivates us to develop efficient but effective heuristics for this problem.

For simplicity, assume that there are n individuals to be sampled with equal number of cases and controls. Among them, $n\pi$% individuals with equal number of cases and controls will be genotyped for all the m_2 SNPs (the extension to unequal numbers of cases and controls is straightforward). The selection procedure based on data in stage 1 consists of three steps

with different emphases. In the first step, SNPs are ranked based on their associations with the disease. A regular two-stage design selects SNPs solely based on this ranking. But such a strategy is not optimal because of sampling errors and potential strong associations of selected SNPs among themselves. In the second step, nearby SNPs will form clusters based on their pairwise correlations, and only one representative from each cluster will be considered for subsequent analysis. At last, in order to take into considerations joint effects from multiple SNPs and their potential interactions, the proposed algorithm employs an entropy-based subset selection technique in the third step. Not only does the algorithm select a much reduced set of SNPs for stage 2, it also provides an order of SNPs according to their joint discriminative power with respect to the disease. It is very flexible and can incorporate different algorithms in each step. But to ensure efficiency for genome-wide studies, some efficient and easy-to-implement algorithms are chosen for each step in this study and will be discussed.

3.3.1. Single SNP ranking—One simple method to assess the association of a single SNP and the disease under study is Pearson's χ^2 test based on a 2×2 contingency table generated by the counts of two alleles that occur in cases and controls. The goal is not to perform statistical tests here, but to provide a quantitative measure of association to rank SNPs. It is expected that the values might have large variances due to a small sample size in stage 1. More complicated ranking mechanisms are also possible. For example, if samples are haplotyped, haplotype information can be taken into consideration, using feature selection techniques. Extensions along this direction are currently under investigations.

3.3.2. SNP clustering—It is not wise to take the SNPs with high ranks from the previous step as the set of SNPs for stage 2. First, highly ranked SNPs themselves might be highly correlated. Therefore, they are redundant for mapping disease susceptibility loci. Second, SNPs that are close to disease genes might not have the highest ranks due to a variety of reasons (e.g., sampling errors). To address these two questions, I propose a simple clustering algorithm that explicitly explores correlations among SNPs. More specifically, starting from the SNP with the highest rank, all the SNPs that are highly correlated with it (e.g., if the pairwise LD measure D' is greater than a predefined threshold) will be grouped into a cluster, conditional on that they are within a certain physical distance. The cluster will be represented by the SNP with the highest rank. The process will continue in the decreasing order of SNP ranks until all the SNPs have been included. At the end, the algorithm returns a set of clusters, each represented by a SNP with the highest rank within its cluster. The clusters are different from haplotype blocks because it does not require all SNPs in a cluster being consecutive. This flexibility is necessary given the small sample size in stage 1 and some inconsistency in haplotype block structures. Further variation can be added to this basic algorithm. For example, when adding a SNP to a cluster, one may also require that the SNP must be in high or moderate correlations with all the SNPs that have been selected in the cluster, instead of only using the correlation with the representative SNP.

3.3.3. Subset selection—The previous two steps mainly focus on the correlation of two SNPs, or the correlation of one SNP and the disease. It works best if the disease is caused by a single mutation. But it is well known that, for most complex diseases, multiple DS genes with low individual effects might be involved, and haplotype effects or gene-gene interactions might play a key role in the development of a disease. Explicit modeling of gene-gene interactions in genome-wide association studies is in general not feasible, because it requires an extremely large sample size to obtain some statistical significant results. On the other hand, it is unwise not to consider the issue when designing association studies. We explicitly investigate joint contributions to the disease from a subset of representative SNPs obtained in the previous step using an entropy-based approach. Entropy is a measure of

uncertainty of a random variable. The concept originates in information theory and has been widely used in many applications. Hampe et al. (2003) have proposed an entropy-based SNP selection algorithm. In their paper, the usefulness of a SNP is defined with respect to a disease locus. Because both the location and the allele status of the disease locus are unknown, the authors defined a "mapping utility" function as an approximation. In this paper, the usefulness of a SNP is defined directly based on its relationship with the disease status. Formally, for a locus **A**, its entropy $H(A)$ is defined as:

$$H(A) = -p_{A_0} \log\left(p_{A_0}\right) - p_{A_1} \log\left(p_{A_1}\right).$$

Let **Y** denote the disease status, the joint entropy $H(Y, A)$ of **Y** and **A** is defined analogously based on their joint distribution. The conditional entropy $H(Y|A)$ of **Y** given **A** is defined as:

$$H(Y|A) = H(Y, A) - H(A).$$

It represents the uncertainty of **Y** given the fact that **A** is known. So one can measure how much information **A** contributing to **Y** using the difference of **Y**'s entropy and the conditional entropy of **Y** given **A**:

$$I(Y|A) = H(Y) - H(Y|A).$$

In general, suppose a set of markers $\mathbf{A}_1$, $\mathbf{A}_2$, …, $\mathbf{A}_i$ has been selected, the next marker **B** to be included should be the one that maximizes the information gain about **Y**, i.e., the one that maximizes

$$I(Y|B, A_1, A_2, \ldots, A_i) - I(Y|A_1, A_2, \ldots, Ai).$$

The principle used here is similar to the one used in Hampe et al. (2003). But with the disease status instead of the disease locus, the computation of the quantities is now feasible when haplotype frequencies are known. When haplotype frequencies across a larger number of markers are not available, we propose two strategies that use pairwise haplotypes or haplotypes of three loci instead. The first strategy uses pairwise haplotypes of **B** and an $\mathbf{A}_j$. The information gain about **Y** by further including a marker **B** critically depends on the association of **B** with other markers that have already been included (**A** markers). If **B** is highly associated with a marker $\mathbf{A}_j$, it will contribute small amount information to **Y** since $\mathbf{A}_j$ has already been included. Formally, one can choose a marker **B** that maximizes the minimum information gain by utilizing all pairwise haplotypes (**B** and $\mathbf{A}_j$):

$$\max_B \left(\min_j \left(I\left(Y|B, A_j\right) - I\left(Y|A_j\right) \right) \right). \quad (1)$$

The second strategy is also to consider the association of a marker **B** with those markers that have already been included. But this time, instead of examining all pairwise relationships, we propose to simultaneously examine the relationship of **B** with the two markers $\mathbf{A}_l$ and $\mathbf{A}_r$ that are adjacent to **B**. In general, these two flanking markers are likely to contain more information about **B**. If **B** is not represented by the two markers, including **B** may provide much more information on the disease. So one should choose a marker **B** that maximizes the information gain by utilizing haplotypes of three loci:

$$\max_{B} \left(I\left(Y|A_l, B, A_r\right) - I\left(Y|A_l, A_r\right) \right).$$

For both strategies, a global threshold σ can be specified. Only markers with scores larger than σ will be included. In some cases, it is hard to define a threshold beforehand. As an alternative, an ad hoc stop criterion is adopted. Let σ_i denote the quantity calculated using above formula when there are i markers being selected. We suggest that the procedure should stop when σ_i is smaller than a fraction of σ_1. Such a fraction can be chosen based on the distribution of σ_is. Furthermore, this stop criterion can be adjusted by the total number of markers already being selected. For example, the selection should stop at the first i such that

$$\sigma_i / \log(i) < \sigma_1 / c, \tag{2}$$

where $c \geq 1$ is a constant. Essentially, one wants to achieve a balance between the model complexity (the number of variables/SNPs) and the prediction accuracy on training data (data in stage 1), so that the selected variables/SNPs will have a great generalization power on test data (data in stage 2). For both strategies, the selection of markers is in the decreasing order of their informativeness with respect to the disease. Therefore, this approach provides a way of prioritizing SNPs to be typed in stage 2. The above discussion assumes the SNPs are linked so their haplotypes are used in the calculation. It is designed so that if there are haplotype effects on the disease, it will be captured by considering the joint information from different SNPs. For SNPs/genes on different chromosomes, one can use genotype frequencies instead of haplotype frequencies in the above calculation, so that gene-gene interactions might be captured.

3.4. Power analysis in stage 2

To compare the power of the proposed method, the method using the regular two-stage design, and the method using one-stage design, many statistical tests can be utilized. For simplicity, the Pearson's χ^2 test is used to compare the difference of allele frequencies among cases and controls for each SNP. Although it is not ideal for the proposed method to use a single locus-based method, it is one of the few feasible for genome-wide association studies. As mentioned earlier, three strategies (i.e., replication-based, joint, and combined) for two-stage analysis have been proposed in the literature, and any of them can be used here. It has been shown that both joint analysis and combined analysis have higher power than replication-based analysis. Combined analysis has been chosen in this study because it is more general than the joint analysis by allowing heterogeneity among stages. However, unlike the regular two-stage design, the power of the proposed method cannot be obtained directly by analytic methods because the selection method proposed here is not a pure test-based approach. Instead, we perform power comparisons using simulation studies. Four different approaches have been considered, namely, single stage analysis (SS), regular two-stage analysis (TS), two-stage with clustering (TS-C), and two-stage with clustering and subset selection (TS-CS). For the three approaches that consist of two stages, the combined analysis adds the two statistics from the two stages together. Because only SNPs with large statistics in stage 1 are evaluated using the combined statistics, the combined statistics are biased even under the null assumption of no association. One needs to derive significance levels for the new statistic. Suppose a liberal significance level α_1 with the critical value c_1 is used in stage 1. Let X_1 denote the χ^2 test statistic based on samples in stage 1. Only markers with $X_1 > c_1$ will be further considered for all three methods (TS, TS-C, and TS-CS). TS-C and TS-CS will further limit the number of SNPs to be genotyped by considering their LD and their discriminative power. For the set of markers to be genotyped in stage 2, let X_2 denote the test statistic using samples from stage 2. Under the null hypothesis of no

association, X_1 and X_2 are independent and follow χ^2 distribution with 1 degree of freedom. For the combined analysis, the statistic X is equal to the summation of X_1 and X_2. Notice that X and X_1 are not independent even under the null hypothesis. Let $f(x)$ and $F(x)$ denote the probability density function and the cumulative distribution function of χ^2 distribution with 1 degree of freedom, the significance level of X with a value c should be defined as the conditional probability of $X > c$ given the fact of $X_1 > c_1$. By some simple algebraic manipulations, it can be shown that the significance level can be calculated based on the following formula through numerical methods (Press et al., 1992):

$$Prob\,(X>c|X_1>c_1) = \left\{\int_{c_1}^{c} (1 - F\,(c - x_1))\, f\,(x_1)\, dx_1 + (1 - F\,(c))\right\} / \alpha_1.$$

For a fixed overall significance level α, if the above value is smaller than α/n_i, where n_i (i = 1, 2, 3) is the number of SNPs selected by TS, TS-C, and TS-CS respectively, a positive finding is declared. Basically, Bonferroni correction is applied here to obtain an overall p-value. Bonferroni correction is less problematic for TS-C and TS-CS because SNPs with high LD have been grouped into clusters. Another approach to obtain an empirical p-value is via Monte Carlo simulation. For case-control studies, a permutation test can be easily performed by shuffling the phenotypes among all the individuals to obtain an empirical p-value. By randomly shuffling the phenotype values (disease status), it is expected that associations between the disease locus and the trait will be broken. The association mapping analysis is performed on each shuffled data set and the values of the resulting statistics are recorded. The process is repeated for a sufficiently large number of times to mimic the distribution of the original data. The proportion of the data sets whose statistic values are equal to or more extreme than the statistic produced by the original data set is regarded as the empirical p-value. But this usually can only be done for candidate gene studies, but not genome-wide studies because it is time consuming.

4. RESULTS

The above algorithm has been implemented using C++ and has been tested under two different scenarios: candidate gene studies and genome-wide association studies. First, a simulation study is performed in the context of candidate gene studies. The algorithm is then applied to a publicly available genome wide association data set for sporadic ALS (Schymick et al., 2007). For a simulation study, one can either generate genotypes based on population genetics models, or by perturbing real data with some noise. Due to simplified assumptions, simulations based on population genetics models may not be able to capture the true property of LD in human populations. Therefore, the second approach is taken in this study, and large numbers of case and control samples are generated based on data from the HapMap ENCODE project using a tool developed recently by our group (Li and Chen, 2008).

4.1. Generating empirical data sets based on HapMap data

The ENCODE project has selected ten 500-kb regions to discover all SNPs within those regions by resequencing. SNP genotypes were obtained for all the 269 HapMap samples from four different populations. We randomly take one region (ENr112 on chromosome 2p16.3) from one population (30 trios with northern and western European ancestry) to generate our empirical data. The number of SNPs in region ENr112 is 1157. We start with the phased genotypes (120 haplotypes of parents) from HapMap website. We first partition all the SNPs into sets of equivalent classes, where a SNP in a class can be totally determined/predicted by any other SNPs in the class. In other words, we take the data from HapMap project as our reference panel, and the set of nonredundant SNPs is kept and typed

in stage 1. Four hundred forty-one SNPs are retained after this step. To generate case-control samples, we consider four different disease models (dominant, recessive, multiplicative, and additive) with a small genotype relative risk γ (1.2–1.5) and a single high-risk variant at the disease locus. The population prevalence θ is fixed to be 0.1 for a common disease. The disease allele frequency (DAF) f varies from 0.1 to 0.5. The penetrance (the probability of an individual being affected given its genotype) can be calculated based on disease models, prevalence, genotype relative risk and allele frequencies under the assumption of Hardy-Weinberger equilibrium. For example, for the dominant model, assuming the penetrance for homozygous wild genotype is ρ, we have:

$$\theta=\rho(1-f)^2+2\rho\lambda(1-f)f+\rho\lambda f^2.$$

For each disease allele frequency, we select one locus with the minor allele frequency approximately equal to the DAF. We extend the approach used by Durrant et al. (2004) in generating the genotypes at the disease locus and other loci in the region. To generate a case, the distribution of all three possible genotypes at the disease locus t can be obtained based on genotype frequencies and penetrances using Bayes theorem:

$$P(g_i|\text{case}) = (P(g_i)\,P(\text{case}|g_i)) \,/\, \left(\sum_{g_j}\left(P\left(g_j\right)P\left(\text{case}|g_j\right)\right)\right).$$

To generate an empirical haplotype pair (h_1, h_2) across the region, we first randomly select two haplotypes (h_3, h_4) from the 120 haplotypes with the required genotype at the disease locus t—that is, ($h_1(t)$, $h_2(t)$) = ($h_3(t)$, $h_4(t)$). In their original paper (Durrant et al., 2004), the haplotype h_1 is given the same alleles as h_3 from locus $t - l$ to $t + l$, where l is a parameter that can be specified by users. To extend h_1 to the right for one more locus, it randomly selects another haplotype h_5 that has the same alleles as h_1 from locus $t - l + 1$ to locus $t + l$, and let $h_1(t + l + 1) = h_5(t + l + 1)$. By iterating the above process, one can extend h_1 to the right and to the left. We found that LD patterns from samples generated this way greatly depend on the parameter l (data not shown). But it is difficult for users to select a proper l. We have extended the above method by introducing two parameters, $l_{\min}$ and $l_{\max}$. The overlapped length for both the initial assignment and the extension of h_1 will be stochastically determined by two values l_l and l_r ($l_{\min} <= l_l, l_r <= l_{\max}$), one for each direction. The values of l_l and l_r depend upon the strength of local LD. By using two parameters, our method takes both long-range LD and short-range LD into considerations. A detailed description of the procedure can be found on our website at www.eecs.case.edu/~jxl175/gs.html. Alleles of h_2 can also be constructed similarly. The required number of cases can be generated by iterating the above process. Controls can also be generated similarly. The disease locus will be removed before further analysis. Figure 1 (bottom) illustrates the procedures in the experiment. By taking such an approach, we can generate a large number of samples based on the 120 haplotypes. These samples are not exact copies of human data, so large scale simulation studies can be performed. Also, the samples have similar local LD patterns as those in real human data (data not shown). The results obtained from these synthetic data are very likely to hold in real data.

4.2. Selection results

We compared the power, the number of SNPs selected for stage 2, and the distance from predicted positions to the true disease SNP position for the four strategies (SS, TS, TS-C, and TS-CS) over a wide range of parameters, including four different disease models (dominant, recessive, multiplicative, and additive), three different levels of genotype relative

risks (γ = 1.2, 1.35, 1.5), and three different disease allele frequencies (f = 0.1, 0.3, 0.5). With the correction of bias introduced from selection in stage 1 (Subsection 3.4), all three two-stage methods have correct type one errors (data not shown). We have tried different sample sizes (total number of individuals n = 500, 1000, 2000, with equal numbers of cases and controls), and selected a sample size of 1000 to avoid the situation that all methods have very low or very high power. We have also tried different split fractions (30%, 40%, 50%) of samples genotyped in stage 1, and no major differences have been observed. Therefore, we fixed the split fraction as 50%. The LD threshold for the clustering step is fixed as D' = 0.8. Formula (1) is used in the subset selection step, and the procedure stops when the inequity (2) (with c = 1) holds for the first time. For each parameter combination, 100 replicates were generated. Results of the four methods were recorded. As shown in Figure 2, the power of each method greatly depends on the parameters such as disease models, genotype relative risks and allele frequencies. For each fixed set of parameter, the original two-stage design (TS) always has the lowest power. The two-stage design with clustering and subset selection (TS-CS) always has the highest power. The two-stage design with clustering only (TS-C) and the single stage design (SS) have similar performance. This means that with further processing by clustering and/or subset selection, the proposed approach achieves higher power comparing with the original two-stage design (TS). In addition, a noticeable result from the simulation is that TS-CS, with much reduced genotyping costs, always performs better than SS. The increase in power is due to smaller number of SNPs examined by TS-C and TS-CS in their second stage. In terms of the total number of SNPs genotyped, SS needs to consider all the 440 SNPs for all the 1000 individuals. The other three methods need to genotype the 440 SNPs for only 500 individuals in their first stage. In stage 2, TS needs to genotype about 50 SNPs on average (with a liberal significance level of 0.1 in stage 1). By clustering nearby SNPs that are in high LD and subset selection, one can further reduce genotyping costs in stage 2. For example, TS-C needs to genotype about 15 SNPs, and TS-CS only needs about seven SNPs across the whole region of 500 kb (Fig. 3, up row). Even with a much reduced number of SNPs, the prediction errors (the distance from the most significant SNP to the true disease SNP) are almost the same for all four methods (Fig. 3, bottom row), which demonstrates that the proposed approach can effectively remove irrelevant SNPs.

4.3. Genome wide association of amyotrophic lateral sclerosis (ALS)

ALS is a fatal neurological disease with unknown causes. Recently, a genome-wide association study has been carried out with the aim to identify genetic variants that are associated with an increased or decreased risk for developing sporadic ALS (Schymick et al., 2007). As the first stage, the study genotyped 555k SNPs using the Illumina Infinium II SNP chip, but only for a very small sample size (276 patients and 271 controls). The study identified a set of candidate SNPs that are associated with the disease, but, not surprisingly, none of them are significant after Bonferroni correction. This is not an ideal test data set for the proposed algorithm because of its small sample size. Nevertheless, the data represent one of the first sets of publicly available genome wide association data, and I decide to test the new proposed framework using it.

To determine proper parameter values for clustering (LD threshold) and subset selection (σ) for this genome-wide association study, preliminary tests were first performed on one chromosome. Chromosome 10 was chosen because it contains the strongest signal based on allelic association analysis in the original paper by Schymick et al. (2007). The total length of chromosome 10 is 135 Mbps and the total number of SNPs is 28,818. The average interval distance between adjacent markers is about 4.7 kbps. Because of the small sample size of the dataset, multiple (100) random runs were performed to selection a subset of samples in stage 1. Results averaging across all random runs were used in determining

proper parameter values. For each run of the two-stage analysis, an individual from all 276 patients and 271 controls is selected as a sample in stage 1 with a probability of 0.5. The average sizes of clusters were obtained using both D' and r^2 ranging from 0.9 to 0.1. Results showed that the average number of SNPs within each cluster is always smaller than 2 (Fig. 4A), comparing with the average size of 3 in the candidate gene study with $D' = 0.8$. This difference is mainly due to the difference of SNP densities of the two studies. Because the average size of clusters does not change much with different threshold of D' and r^2, and some clusters actually only consist of a single SNP, the same threshold ($D' = 0.8$) was chosen for the genome-wide study. To determine a proper threshold value for σ, I examined the distribution of the maximum of minimum information gains (σ_i, calculated based on Equation (1)) and compared their values to σ_1. Figure 4B provides a typical case in illustrating how the values ($\sigma_i - \sigma_1$, $i \geq 2$) change when including more markers. A clear exponential decay was observed and the trend held consistently for different runs examined (data not shown). Based on this observation, the global threshold σ was set to be σ_1 in this study.

Because of the small sample size of the current data set, it is unlikely to obtain globally significant SNPs for any method as illustrated by the original paper. Furthermore, the ranks of SNPs based on one random run are expected to have large variances. Therefore, I took a different approach in evaluating the importance of each SNP using this dataset. The approach resembles the principle of boosting methods in determining the importance of each variable in a classification analysis. More specifically, for each run, the set of SNPs selected in stage 1, as well as their ranks, are recorded. Such experiments will be performed for a large number of times (here $N = 1000$), and the number of occurrences that a SNP being prompted to stage 2, along with its average rank will be used to evaluate the significance of the SNP. As usual, such an analysis was first performed on chromosome 10. To assess the soundness of the approach, I compared the results obtained using this approach and those obtained by other two approaches. First, a permutation test with 1000 runs was performed to examine the distribution of the number of a SNP being prompted to stage 2 totally by random. The results from the permutation and the results from the real data are shown in Figure 4C. None of the SNPs in the permutation test have occurrences greater than 100 times. There are 93 SNPs from the real dataset with occurrences great than 100. There are probably many false positives among the set of SNPs, but that can be tolerated because they will be further assessed in stage 2. Table 1 summarizes all the SNPs with the number of occurrences greater than 300 (out of 1000 runs). It is very unlikely that all these top candidates returned by the method are totally due to random. I also compared the results with results using the single SNP-based allelic association test (Fig. 4D) obtained directly from Schymick et al. (2007). The two methods returned the same SNP (rs4363506) as their top candidate. This further shows that the number of occurrences of a SNP from a large number of runs of the algorithm is indeed a good indicator of its significance. The allelic association test reported only two additional SNPs (rs10765118 and rs10830099) on chromosome 10 with a p-value smaller than 10^{-4}. All three SNPs are in the same genomic region within a 10-kb interval on 10q26.13 (Table 1 in Schymick et al., 2007). The other two SNPs were not in the top candidates returned by our approach (Table 1). Further examinations reveal that they are not in the cluster represented by SNP rs4363506 neither, under the current parameter setting of $D' = 0.8$. It seems that they are not significant any more after rs4363506 has been included. The second SNP (rs7902011) on the list of Table 1 is within an intron region of the gene APBB1IP. According to its Gene Ontology annotation, this gene involves in the neuropeptide signalling pathway. The minor allele frequencies at this SNP are 0.291 and 0.389 in cases and controls, respectively. But this SNP has a large missing rate (81 normal individuals and 49 controls have missing genotypes). Additional investigation is needed for this and other candidate SNPs.

A similar analysis was carried out for each of the 22 autosomes individually. Sex chromosomes were not included because of a confusing format in the original data.1 The autosomes were analyzed individually because a joint analysis including all chromosomes is not feasible due to high computational costs. Figure 5 illustrates the combined results from all autosomes. All SNPs with the number of occurrences greater than one were included and were ranked according to their occurrences. Those ordered SNPs were then aligned in the *x*-axis with their occurrences on the *y*-axis in Figure 5A. A SNP on chromosome 3 (rs11915402) has the maximum number of occurrences (993 out of 1000) with the average rank of 2.57 within SNPs on chromosome 3. It is within an intron region of a predicted gene with no functional annotations. The number of occurrences decays rather rapidly. There are only about 500 SNPs with occurrences greater than 200. Figure 5B shows the histogram of SNPs according to the number of their occurrences, and it follows an exponential decay in general. However, the number of SNPs with extreme values (the number of occurrences greater than 800) seems enriched. This can be clearly seen in Figure 5C, in which a linear regression has been used in fitting the number of SNPs (after a logarithmic transformation) with occurrences smaller than 800. This was done by first separating SNPs into 10 equally sized bins according to the numbers of their occurrences (possibly from 1 to 1000). The R^2 of the linear regression is 0.9206. The SNPs with occurrences greater than 800 clearly do not follow the trend. There are total 11 such SNPs, and their functional annotations are given in Table 2. Incidently, in the final list, there are three genes (GTSE1, CDC25C, CDC42) involving cell cycles and two genes (SLC39A11, ZNHIT2) involving ion transporting/ binding. It is well known that ALS was caused by the gradual degeneration and death of motor neurons. It has also been shown that familial ALS is linked to SOD1 gene, which encodes proteins that bind copper and zinc ions. There are two predicted genes (FLJ42117 and C7orf42) with no functional annotations, and the remaining SNPs are located in intergenic regions. Complete results across the genome are available from our website (www.eecs.case.edu/~jxl175/PS_SNP.html). Although definite links between these SNPs and ALS cannot be established here because of the small sample size, the set of top ranked SNPs, together with the set of candidate SNPs identified by single-SNP based approaches provide a promising set of SNPs for further association studies or functional analysis of ALS.

5. DISCUSSION

For genome-wide association studies or candidate gene studies with dense SNP arrays, SNPs within a short distance are not independent from each other. In this paper, I have presented a novel framework for prioritizing and selecting SNPs for association analysis with two-stage designs. The framework is unique in the sense that it takes into consideration correlations among SNPs, as well as correlations between SNPs and the disease. Simulation results from a candidate gene study have shown that by selecting a compact and discriminative subset of SNPs for stage 2, the framework indeed achieves higher power than a regular two-stage design. It also performs better than the single-stage design, with much reduced genotyping costs. I have also applied the method using a boosting strategy to a real data set generated by a genome-wide association study of ALS and identified an interesting candidate set of SNPs. Like any results from genome-wide association studies, this set of SNPs also needs to be validated by further replication studies.

Two-stage designs have been mostly studied in the context of optimizing genotyping costs with desired power, given a set of parameters such as allele frequencies and disease effects (Satagopan and Elston, 2003; Thomas et al., 2004; Wang et al., 2006). However, prior

[1]A query has been sent to the data provider, but no explanation was given.

information about those parameters are usually not available for many complex diseases. Actually, not requiring prior information about diseases has been viewed as one of the advantages of genome-wide association studies. In practice, when a two-stage or multi-stage strategy is being adopted in genome-wide association studies, the ability that a researcher can make choices on options (such as sample split fractions among stages or number of markers to be prompted to later stages) is usually limited by many other factors such as budgets and the availability of cases and/or controls. Therefore, in many cases, it is unrealistic for researchers to have an optimal design beforehand, and this is not the focus of the current study. Instead, our goal is to compare the performance of the new strategy with the performance of the original two-stage analysis. The results indeed show that the new strategy consistently performs much better than the original two-stage analysis, regardless of disease allele frequencies and gene effects, for instance.

The model in generating the simulated data in this study is under overly simplified assumptions. This problem was compensated by a study of real data set. However, because of the small sample size of the ALS dataset, it is hard to draw a conclusion based on the analysis. This problem will be alleviated by emerging genome-wide association data sets concerning many other complex diseases (e.g., diabetes, breast cancer, prostate cancer), as well as data for ALS from additional studies. Some of the projects such as CGEMS have promised to provide raw data, including SNP genotypes and clinical phenotypes, to the community. Such datasets will be invaluable resources for the whole community, not only for the study of the diseases themselves, but also for the advance of methodology developments.

The proposed framework is flexible and can incorporate different algorithms or measures in each step. An important consideration in implementation is efficiency for genome-wide association studies. I have implemented some simple algorithms/measures in each step in this study so that the final boosting and permutation analysis can be performed. Still, the algorithm has to run on each chromosome individually. The experiments were performed using a cluster with 42 nodes provided by the high-performance computing service at Case Western Reserve University. Each node either has a dual 3.2-GHz Pentium-4 Xeon processors with 4 gigabytes of memory, or has two Intel Pentium-4 Xeon EM64T processors running at 3.8 GHz, 4 Gb of main memory. It took 2-3 days to analyze one chromosome on one node (the actual time depends on the number of markers on each chromosome).

Although the problem has been formulated to identify a minimum subset of SNPs with low redundancy and maximum discriminative power from data in stage 1, redundancy may not be useless in real data analysis. For example, the (most) significant SNPs are not necessarily causal SNPs themselves, but may be in LD with them. Including more relevant SNPs might be helpful in identifying functional elements that are involved in the development of diseases. The framework can be easily modified to include all SNPs in clusters that have been prompted to stage 2. Additional SNPs can be selected based on HapMap information when needed.

As an unbiased scan, genome-wide association studies have the potential to identify genetic risks with moderate effects for complex diseases. A multi-stage design has been shown effective and has been commonly used in practice. Our experiments (data not shown) indicate that combined analysis using statistics from both stages is more powerful than replication-based analysis, which is consistent with a previous report using a different statistic (Skol et al., 2006). Therefore, such a strategy should be adopted whenever possible. Results from genome-wide association studies have to be interpreted with caution. Independent studies usually are needed for validation. Even after SNP associations have been confirmed, it is still not easy to fully understand the functionality of those SNPs if the

mutations do not directly change their protein sequences. A comprehensive study that incorporates data from all different sources (such as gene/protein expressions, interactions, regulatory information) should be undertaken to provide new insights about functional roles of those genomic elements. Experiments then need to be performed to test biological hypotheses before diagnoses or treatments can be possible.

Acknowledgments

This research was supported by the NIH/NLM (grant LM008991), and in part by the NIH/NCRR (grant RR03655). The ALS data was obtained from Coriell Institute for Medical Research (http://ccr.coriell.org).

REFERENCES

Ahmadi KR, Weale ME, Xue ZY, et al. A single-nucleotide polymorphism tagging set for human drug metabolism and transport. Nat. Genet. 2005; 37:84–89. [PubMed: 15608640]

Collins FS, Green ED, Guttmacher AE, et al. A vision for the future of genomics research. Nature. 2003; 422:835–847. [PubMed: 12695777]

de Bakker PI, Yelensky R, Pe'er I, et al. Efficiency and power in genetic association studies. Nat. Genet. 2005; 37:1217–1223. [PubMed: 16244653]

Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics. 1995; 29:311–322. [PubMed: 8666377]

Durrant C, Zondervan KT, Cardon LR, et al. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. 2004; 75:35–43. [PubMed: 15148658]

Hampe J, Schreiber S, Krawczak M. Entropy-based SNP selection for genetic association studies. Hum. Genet. 2003; 114:36–43. [PubMed: 14505034]

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 2005; 6:95–108. [PubMed: 15716906]

International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]

Li J, Chen C. Generating samples for association studies based on HapMap data. BMC Bioinformatics. 2008; 9:44. [PubMed: 18218094]

Press, WH.; Flannery, BP.; Teukolsky, SA., et al. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press; Cambridge, UK: 1992.

Satagopan JM, Elston RC. Optimal two-stage genotyping in population-based association studies. Genet. Epidemiol. 2003; 25:149–157. [PubMed: 12916023]

Skol AD, Scott LJ, Abecasis GR, et al. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat. Genet. 2006; 38:209–213. [PubMed: 16415888]

Thomas D, Xie R, Gebregziabher M. Two-stage sampling designs for gene association studies. Genet. Epidemiol. 2004; 27:401–414. [PubMed: 15543639]

Wang H, Thomas DC, Pe'er I, et al. Optimal two-stage genotyping designs for genome-wide association scans. Genet. Epidemiol. 2006; 30:356–368. [PubMed: 16607626]

Weir, BS. Genetic Data Analysis II. Sinauer Associates; Sunderland, MA: 1996.

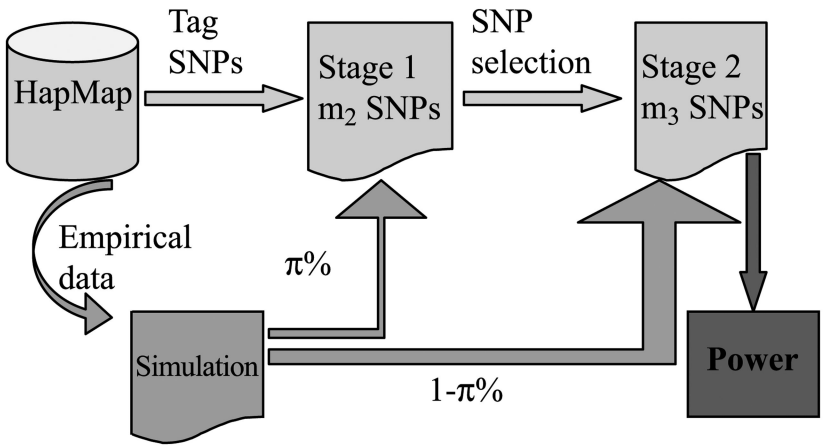

FIG. 1.
An illustration of a two-stage design for association studies (**top**) and the simulation strategy detailed in this paper (**bottom**).

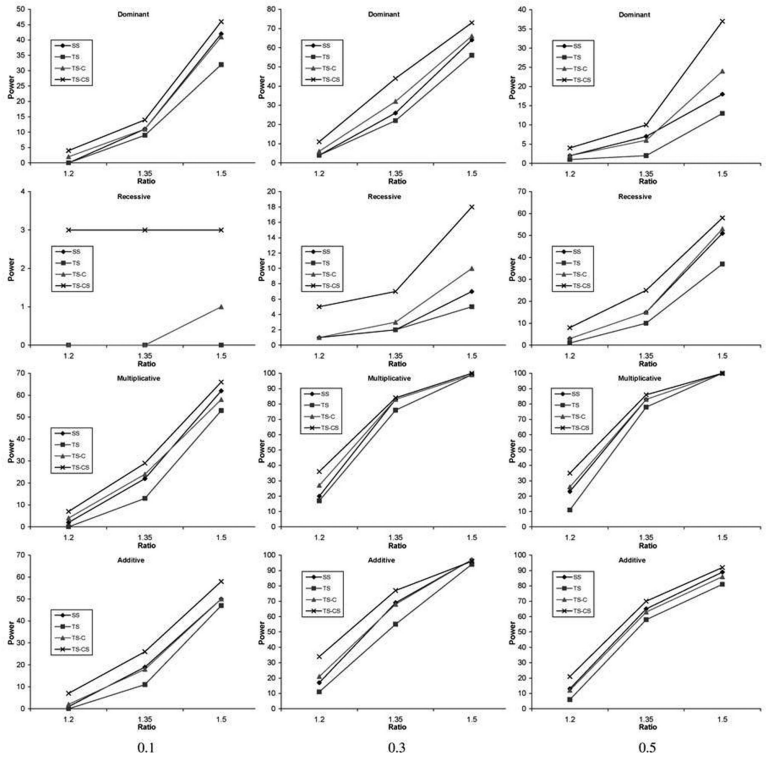

FIG. 2.

Power analysis of the four approaches (SS, single stage analysis; TS, regular two-stage analysis; TS-C, two-stage analysis with clustering; TS-CS, two-stage analysis with clustering and subset selection) across four different disease models (**rows**) with different disease allele frequencies (**columns**). The sample size is 1000 with equal number of cases and controls. Assume half of them are genotyped in stage 1 for the three 2-stage strategies.

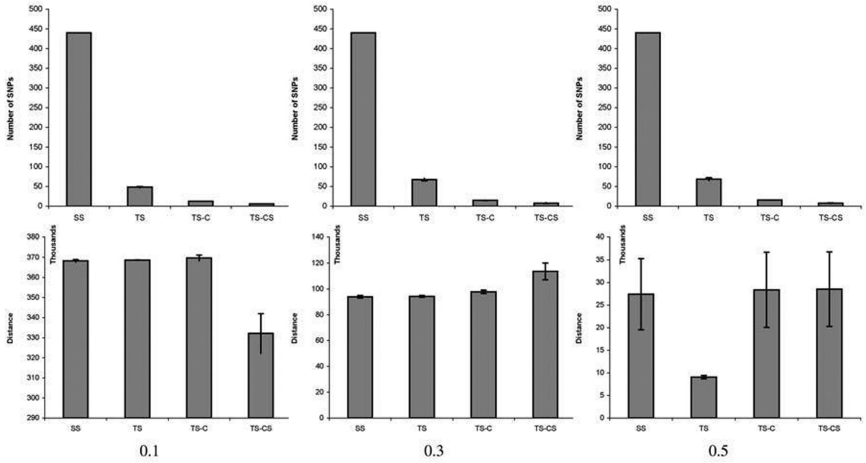

FIG. 3.

The average number of single nucleotide polymorphisms (SNPs) examined in stage 2 comparing with the total number of SNPs (**top**), and the average distance from predicted locus to the disease locus (**bottom**) for different disease allele frequencies.

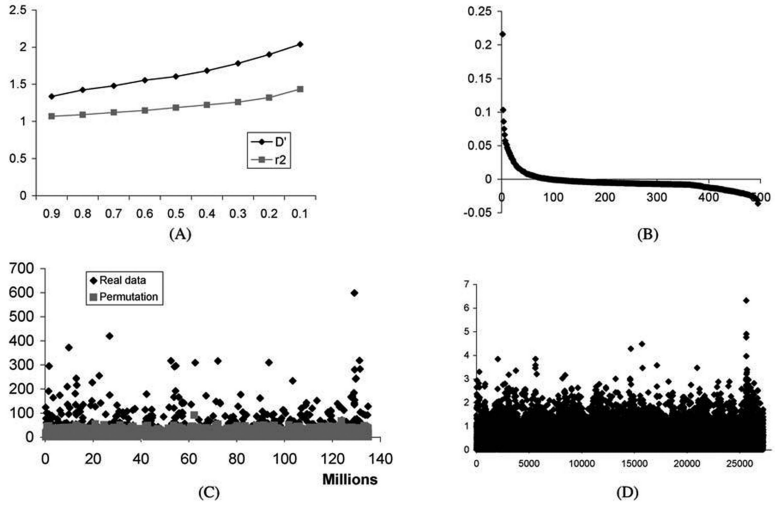

FIG. 4.

Experimental results on chromosome 10. (**A**) The average size of clusters for different levels of linkage disequilibrium (LD) threshold using D' and r^2. (**B**) The difference of the maximum of minimum information gains by including additional single nucleotide polymorphisms (SNPs) and the entropy of the first SNP ($\sigma_i - \sigma_1$) from one random run. (**C**) The number of times of a SNP being prompted to stage 2 out of 1000 runs from the original data (diamond) and permuted data (square). The *x*-axis represents the physical distance. (**D**) The $-\log(p)$ of the allelic association test obtained from the original paper. The *x*-axis represents the marker index.

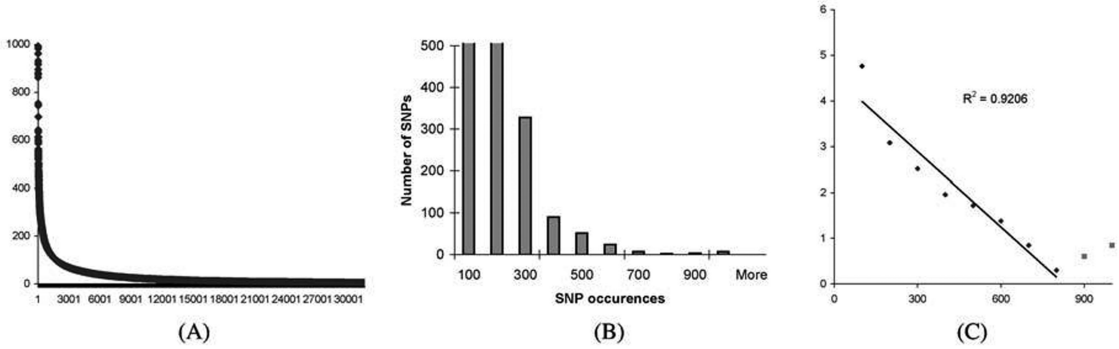

FIG. 5.

Distribution of occurrences across all autosomes. (**A**) Single nucleotide polymorphisms (SNPs) are ordered according to their occurrences. (**B**) SNP histogram based on the number of occurrences. The first two bins are truncated. (**C**) Linear regression analysis of the number of SNPs (*y*-axis, in a logarithmic scale) and the number of occurrences (*x*-axis). SNPs are grouped into 10 equally sized bins according to their occurrences. Regression is performed based on the first eight bins (with occurrences from 1 to 800). The number of SNPs with occurrences greater than 800 clearly shows a reversing trend.

Table 1

Top Ranked SNPs from Chromosome 10 and Their Functional Annotations

Chromosome	SNP ID	Location	No. of occurrences	Average rank	Gene	Function
10	rs4363506	129164492	598	20.6	Intergenic	—
10	rs7902011	26868894	420	1.9381	APBB1IP	Amyloid beta (A4) precursor protein-binding, family B, member 1 interacting protein
10	rs7907643	9869811	372	14.4651	Intergenic	—
10	rs567700	131313115	318	24.0849	MGMT	O-6-methylguanine-DNA methyltransferase
10	rs10508935	52536508	317	5.97476	PRKG1	Protein kinase, cGMP-dependent, type I
10	rs7092269	72152824	316	45.4652	ADAMTS14	ADAM metallopeptidase with thrombospondin type 1 motif, 14
10	rs1857060	93436051	310	21.2581	Intergenic	—
10	rs1906470	62675876	309	30.89	Intergenic	—

SNPs, single nucleotide polymorphisms.

Table 2

Top Ranked SNPs Across the Genome and Their Functional Annotations

Chromosome	SNP ID	Location	No. of occurrences	Average rank	Gene	Function
3	rs11915402†	58957114	993	2.56898	FLJ42117	Hypothetical protein
18	rs2941394	64428505	992	1.08972	Intergenic	—
22	rs140054‡	45101063	983	2.66328	GTSE1	G-2 and S-phase expressed 1; cell cycle; controls DNA damage-induced apoptosis by affecting p53 function
17	rs2567494†	68308488	962	1.41476	SLC39A11	Solute carrier family 39 (metal ion transporter), member 11
5	rs6596428†	137669928	931	1.2782	CDC25C	Cell division cycle 25 homolog C (S. pombe), regulation of cell division; suppress p53-induced growth arrest
7	rs6957895	117715999	920	6.09457	Intergenic	—
15	rs4924608	40120075	917	4.45474	Intergenic	—
6	rs9359255	77976705	895	4.51397	Intergenic	—
7	rs3800818¶	66058577	880	1.26477	C7orf42	Hypothetical protein
11	rs10501396§	64642020	874	4.49428	ZNHIT2	Zinc finger, HIT type 2, metal/zinc ion binding
1	rs2473323†	22261913	862	10.2749	CDC42	Cell division cycle 42 (GTP binding protein, 25 kDa)

The functional role of each SNP is also labeled

†(for intron,

‡for nonsynonymous SNPs,

§for 5′ UTR,

¶for 3′ UTR).

SNPs, single nucleotide polymorphisms; UTR, untranslated region.