Selection of Genetic Markers for Association Analyses, Using Linkage Disequilibrium and Haplotypes

Zhaoling Meng,[1,2] Dmitri V. Zaykin,[2] Chun-Fang Xu,[2] Michael Wagner,[2] and Margaret G. Ehm[2]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh; and [2]GlaxoSmithKline, Department of Population Genetics, Research Triangle Park, NC

The genotyping of closely spaced single-nucleotide polymorphism (SNP) markers frequently yields highly correlated data, owing to extensive linkage disequilibrium (LD) between markers. The extent of LD varies widely across the genome and drives the number of frequent haplotypes observed in small regions. Several studies have illustrated the possibility that LD or haplotype data could be used to select a subset of SNPs that optimize the information retained in a genomic region while reducing the genotyping effort and simplifying the analysis. We propose a method based on the spectral decomposition of the matrices of pairwise LD between markers, and we select markers on the basis of their contributions to the total genetic variation. We also modify Clayton's "haplotype tagging SNP" selection method, which utilizes haplotype information. For both methods, we propose sliding window–based algorithms that allow the methods to be applied to large chromosomal regions. Our procedures require genotype information about a small number of individuals for an initial set of SNPs and selection of an optimum subset of SNPs that could be efficiently genotyped on larger numbers of samples while retaining most of the genetic variation in samples. We identify suitable parameter combinations for the procedures, and we show that a sample size of 50–100 individuals achieves consistent results in studies of simulated data sets in linkage equilibrium and LD. When applied to experimental data sets, both procedures were similarly effective at reducing the genotyping requirement while maintaining the genetic information content throughout the regions. We also show that haplotype-association results that Hosking et al. obtained near *CYP2D6* were almost identical before and after marker selection.

Introduction

Efforts to positionally clone susceptibility genes for common, oligogenic diseases have led to the development of high-density maps of SNPs distributed across the human genome (Sachidanandam et al. 2001). Theoretical studies have suggested that association tests employing such high-density SNP maps, either as a primary approach or as a follow-up to family-based linkage studies, should be more powerful in the detection of disease-susceptibility genes than in traditional linkage approaches (Risch and Merikangas 1996). However, the precise numerical meaning of "high density" is a matter of debate and has significant implications for the cost and practicality of conducting SNP association studies. An optimum strategy would be to genotype enough SNPs to capture the large majority of information on genetic variation within a defined chromosomal region while avoiding the typing of SNPs that yield redundant information because of extensive linkage disequilibrium (LD) between nearby SNPs. Defining the optimum set of SNPs will require knowledge of the patterns of LD across the human genome.

Recently, a common pattern has emerged from several studies that investigated the empirical distribution of LD in a number of different human chromosomal regions. LD appears to be organized in blocklike structures, in which a contiguous group of SNPs that constitute a block show high levels of pairwise LD between SNPs and in which there is little LD between SNPs in different blocks. These blocklike LD structures show considerable spatial variation across different genomic regions, extending from a few kilobases to several hundred kilobases and exhibiting differing boundaries in samples from different ethnic groups (Daly et al. 2001; Johnson et al. 2001; Patil et al. 2001; Subrahmanyan et al. 2001; Dawson et al. 2002; Gabriel et al. 2002). Reduced measures of haplotype diversity within blocks (as compared to expectations under linkage equilibrium [LE] based on the number of SNPs involved) are observed not only for numerically inferred haplotypes derived from genotype data but also for experimentally determined haplotypes (Patil et al. 2001). The reduction of haplotype diversities suggests the possibility of identifying the minimum number of SNPs needed

Received February 14, 2003; accepted for publication April 24, 2003; electronically published June 5, 2003.

Address for correspondence and reprints: Dr. Zhaoling Meng, Bioinformatics Research Center, Campus Box 7566, North Carolina State University, Raleigh, NC 27695-7566. E-mail: zmeng@unity.ncsu.edu

0002-9297/2003/7301-0012$15.00

to define the common haplotypes, thereby reducing the number of markers needed to capture the majority of the genetic information about the region. A procedure that utilizes genotype information on a small number of samples to prioritize SNPs for typing on the basis of a large number of samples could be useful in increasing the experimental efficiency in any project involving a high-density map of SNPs; examples of such a procedure include testing multiple SNPs within a candidate gene for association, fine mapping a region identified through linkage analysis, and testing thousands of SNPs as part of a genomewide association study. Furthermore, identification of the most independent and informative SNPs could be helpful in interpreting analyses across a region where a large number of highly correlated SNPs have been typed, resulting in a large amount of redundant information; on the other hand, any marker-selection procedure based on common SNPs or haplotypes relies on the arguable assumption that common SNP variation can provide high predictive values for risks associated with complex diseases (Couzin 2002). Thus, these procedures are only valuable to the extent that the original set of SNPs is useful for association-mapping purposes. Nevertheless, marker selection can be viewed as a procedure for identifying the polymorphisms most characteristic of underlying populations.

Several algorithms have been proposed for the detection of haplotype blocks and/or the selection of markers. Patil et al. (2001) utilized a greedy algorithm to partition an entire chromosome into a set of contiguous haplotype blocks while minimizing the total number of representative SNPs that distinguish at least α percent of the unambiguous haplotypes in each block. Zhang et al. (2002) extended the greedy algorithm of Patil et al. to a dynamic programming algorithm, which can guarantee an optimal solution for haplotype partitioning. These algorithms rely on the definitions of block boundaries for the selection of markers and require data in which haplotype phase is known. In the data set analyzed by Patil et al. and Zhang et al., haplotypes were determined experimentally, but this type of information is usually not available. Other block-defining algorithms, such as those described by Daly et al. (2002) and Gabriel et al. (2002), do not require haplotype-phase–known data. Daly et al. used a combination of methods (including familial data and the EM algorithm) to estimate haplotype frequencies, identified lower-diversity haplotype within consecutive five-marker windows, and applied a hidden Markov model to formally define the blocks. Gabriel et al. used D', together with associated CIs, as a measure of the historical recombination and defined blocks. However, both of the latter methods appear to be specific to the particular data sets used, and their general applicability is not known. Johnson et al. (2001) proposed two methods to select markers within genes on the basis of haplotypes constructed using either family data or the expectation-maximization (EM) algorithm in unrelated individuals: one method orders the haplotypes by their similarities and requires that SNPs be selected by eye; the other suggests a haplotype tagging SNP (htSNP) diversity method, proposed by Clayton (2001) for the selection of htSNPs, to best extract the haplotype information in a gene. The first method is difficult to automate, and the second can become quite computationally intensive when a large number of markers are considered in a region (for detailed reasons, see the "Discussion" section). Both methods require the predetermination of haplotypes for the region considered, which is difficult when the region contains a large number of markers.

Notice that all the block-detecting methods mentioned may result in differing block boundaries. Given the diversity of methods used to define blocks and the conflicting assertions as to whether they exist at all (Couzin 2002), we choose to develop marker-selection procedures that do not rely on the definition of blocks. Instead, we select a set of SNPs that retain haplotype information similar to an original, presumably larger set of SNPs. The procedure is applicable to regions with a large number of SNPs and to data sets without haplotype-phase information or family information. We propose a method based on the spectral decomposition (spD) of the matrix of the pairwise LD coefficients of the markers and compare it with the htSNP diversity (div) method proposed by Clayton (2001). In addition, we propose a procedure summarizing the information obtained from a sliding window approach, to allow both methods to be applied in large chromosomal regions. Our procedures are local, in that they are applied to genetically proximal sets of markers by considering relatively short windows of markers covering distances that are generally <500 kb. None of the existing marker-selection methods have been evaluated using quantitative criteria describing the proportion of the information retained in the selected marker sets. We compare two local div measures—the haplotype heterozygosity and the number of frequent haplotypes before and after application of the procedures—to assess the information retained and to measure the success of these procedures. We summarize the results from two simulation studies to evaluate the performance of these procedures, and we apply them to two experimental data sets as examples, as well as to markers typed around *CYP2D6*, where an association has been identified, to show the impact that marker-selection procedures have on the results of these association studies (Hosking et al. 2002).

Methods

We describe two methods, a procedure extending them to a large chromosomal region, two simulation studies, and criteria by which we evaluate the procedure's performance when applied to experimental data sets. Our selection procedures study relatively polymorphic SNPs with minor-allele frequency (MAF) $\geqslant 0.05$.

spD

Population genetics theory predicts that the LD associated with alleles from three or more markers decays more rapidly than that from two markers (Bennett 1954). Moreover, the precision of estimates and the power to detect LD associated with alleles from three or more markers quickly diminishes with their order. Therefore, it is reasonable to describe dependencies between markers by pairwise correlations. In essence, spD represents an entire variance-covariance matrix (in the present article, the LD matrix) in terms of its eigenvalues and eigenvectors. The spD-based method that we propose takes into account all pairwise disequilibria for a set of markers. It assumes that the LD associated with alleles from three or more markers is negligible and that the practically available haplotype information can be recovered from pairwise LD and single-marker characteristics. The spD method is also the basis for the principal-component analysis (PCA). In PCA, the sample variation is represented by a few linear combinations (the eigenvectors) of all original variables (i.e., SNPs), taken with different weights (the eigenvalues) reflecting their importance. In contrast, we examine all eigenvectors (linear combinations of the marker contributions) and eigenvalues (the importance of the corresponding combinations) and retain only a subset of the original variables that contribute more to the more important weights. This procedure allows us to consider the pairwise LD coefficients of all markers at once, instead of considering the LD measure for only one pair of markers at a time. Let L be the number of markers evaluated. For a set of markers, $m_1,\ldots,m_L$, the LD matrix is $\mathbf{R}$ with the pairwise correlation r_{ij} as components, where Δ_{ij} is the composite LD (Weir 1996) between markers i and j (see appendix A);

$$r_{ij}=\frac{\hat{\Delta}_{ij}}{\sqrt{\mathrm{Var}\,(\hat{\Delta}_{ij})}}\ .$$

Applying the spD technique, $\mathbf{R}$ can be written as $\sum_{i=1}^{L}\lambda_i\mathbf{e}_i\mathbf{e}_i^T$, where $\mathbf{e}_i$ and λ_i are eigenvectors and eigenvalues of $\mathbf{R}$, $i=1,\ldots,L$, and $\lambda_1\geqslant\lambda_2\cdots\geqslant\lambda_L$. The variables (markers) that contribute more to the eigenvectors associated with the first several large eigenvalues are considered to be the more influential variables (markers) for that LD matrix, $\mathbf{R}$. Variables that contribute more to the eigenvectors associated with subsequent eigenvalues are considered to be less influential.

To determine if there are variables or markers that have little or no influence on the LD matrix, we calculate the L_r index as follows (for details, see appendix B):

$$L_r=L\frac{\sum\lambda_i^2}{(\sum\lambda_i)^2}-1\ .$$

$L_r=0$ indicates that all the markers in the set provide important information and the whole set should be kept. This measure is derived by examining the conditions of extreme disequilibrium and complete independence. We find it useful in determining when no SNPs should be eliminated from the set. If $L_r>0$, then the actual number of markers to be retained, x, is most precisely determined from the inequality

$$\frac{\sum_{i=1}^{x}\lambda_i}{\sum_{i=1}^{L}\lambda_i}\geqslant\alpha\ ,$$

where α is the proportion of information retained (proportion of variation explained). Therefore, we retain markers while the sum of the eigenvalues corresponding to the eigenvectors to which they contribute more is a high proportion of the sum of all eigenvalues. Appropriate levels for α will be investigated in the "Simulation Studies" section.

It is not always clear which marker contributes more to which eigenvalue or eigenvector. To sharpen marker loadings to particular eigenvectors, we apply the varimax-rotation procedure to the original set of eigenvectors, $\mathbf{E}=\{\mathbf{e}_1,\ldots,\mathbf{e}_L\}$. This procedure finds an orthogonal transformation $\mathbf{T}$, $\mathbf{E}^*=\mathbf{ET}$, that will confine the influence of each marker to a particular eigenvector. The varimax rotation is chosen according to the recommendation of Jackson (1991). For each marker, m_j, compute the following:

$$\Gamma_j=\frac{1}{x}\sum_{v=1}^{x}|\mathbf{e}_{jv}^*|$$

and

$$\gamma_j=\frac{1}{L-x}\sum_{v=x+1}^{L}|\mathbf{e}_{jv}^*|\ ,$$

where $\mathbf{e}_{jv}^*$ is the jth element of the vth column of E^*. A marker, m_j, is selected if $\Gamma_j>\gamma_j$—that is, if this marker contributes mostly to eigenvectors associated with the main part of the variation in the data.

div

Clayton (2001) proposed a method to select a subset of SNPs by using haplotype information. Let N be the total number of haplotypes in the sample, which is twice the number of individuals for a diploid population. For L diallelic-marker haplotypes, each haplotype can be written as a vector $z_i = \{z_{ij}, j = 1, \ldots, L, i = 1, \ldots, N\}$, where z_{ij} is either 0 or 1, representing one of the two alleles. The div measure can be defined as the total number of differences in all N^2 pairwise comparisons between a pair of haplotypes. If haplotypes i and k are the same at locus j, then $z_{ij} - z_{kj} = 0$; if they differ, then $z_{ij} - z_{kj} = \pm 1$. At locus j, div is calculated as

$$D_j = \sum_{i=1}^{N}\sum_{k=1}^{N}(z_{ij} - z_{kj})^2 = 2\left[N\sum_{i=1}^{N}z_{ij}^2 - (\sum_{i=1}^{N}z_{ij})^2\right] .$$

Clayton proposed the calculation of the total div as the summation over all loci, which is analogous to the total sum of squares in an analysis-of-variance setting:

$$D = \sum_{i=1}^{N}\sum_{k=1}^{N}(\mathbf{Z}_{\mathrm{i}} - \mathbf{Z}_{\mathrm{k}})^T(\mathbf{Z}_{\mathrm{i}} - \mathbf{Z}_{\mathrm{k}}) = \sum_{j=1}^{L}D_j , \qquad (1)$$

where L is the number of loci.

htSNPs are a set of SNPs that retain most of the information available in the full haplotype. After the selection of a set of htSNPs, N haplotypes are collapsed into groups according to allele combinations for htSNPs. If H of L SNPs under study are selected as candidate htSNPs, then any haplotypes will belong to the same group as long as they have the same alleles at these H loci. Then, the N full haplotypes are divided into $G = 2^H$ (at most) groups. Within each group, a similar diversity measure to that above (eq. [1]) is computed. Within-group div is then summed over all groups, which is analogous to the residual sums of squares:

$$R = \sum_{g=1}^{G}\left[\sum_{i\in G_g}\sum_{k\in G_g}(\mathbf{Z}_i - \mathbf{Z}_k)^T(\mathbf{Z}_i - \mathbf{Z}_k)\right] .$$

Then, Clayton (2001) calculated the proportion of diversity explained by a set of htSNPs as $p = 1 - (R/D)$. The preferred value of R/D is as close to 0 as possible, indicating that there is little diversity left when the haplotype is represented by the subset of htSNPs. The optimal htSNP set is obtained by an exhaustive search from the possible $2^L - 1$ candidate sets. Since Clayton (2001) does not provide guidance on obtaining a good set of htSNPs, we propose selecting a set of htSNPs by first minimizing the number of SNPs selected when maintaining p greater than a desired value, say α, and then picking a marker set provided the maximum p among them.

We have simplified the expressions for both D and R in Clayton's (2001) formula:

$$D = \sum_{j=1}^{L}D_j = \sum_{j=1}^{L}2(Nn_{0j} - n_{1j}^2)$$
$$= \sum_{j=1}^{L}2n_{0j}n_{1j} = N^2\sum_{j=1}^{L}2p_{0j}p_{1j} ,$$

and

$$R = \sum_{j=1}^{L}\sum_{g=1}^{G}2n_{0jg}n_{1jg} = N^2\sum_{j=1}^{L}\sum_{g=1}^{G}2p_{0jg}p_{1jg} ,$$

where n_{0j} and n_{1j} are the number of "0's" and "1's" at locus j and p_{0j} and p_{1j} are the frequencies of "0" and "1" alleles at locus j. Here, $2p_{0j}p_{1j}$ is the expected heterozygosity measure for the jth locus. Correspondingly,

$$p = 1 - \frac{\sum_{j=1}^{L}\sum_{g=1}^{G}2p_{0jg}p_{1jg}}{\sum_{j=1}^{L}2p_{0j}p_{1j}} > \alpha . \qquad (2)$$

Therefore, htSNPs are selected by trying to minimize the within-group locus heterozygosity. After the simplification, the above measure (eq. [2]) can be extended to analyze multiallelic markers; p_0 and p_1 are extended to p_i, where $i = 0, \ldots, T$ and T is the total number of alleles at this marker. Clayton (2001) suggested a κ statistic, which corrects for the fact that selecting a set of ht SNPs will always reduce the residual diversity. When haplotypes are not known with certainty, the EM algorithm is used to infer haplotype frequencies.

Applying spD or div to a Large Chromosomal Region

In the selection of markers that maintain haplotype information, it is important to consider the haplotype information in the context of nearby markers, rather than that of any marker regardless of its position. That is, we decide not to include markers not only if they provide similar information but also if they are fairly close to each other. Therefore, we propose the following procedure to apply either spD or div to a chromosomal region with a large number of SNPs. First, we assume that the markers are arranged according to the map order. Next, a sliding window with a relatively small window size is moved along the map. Either spD or div is used to select informative SNPs in each window. The event of selecting or failing to select a SNP is recorded in a vector $\mathbf{Wi} = \{w_{ij}, j = 1, \ldots, L\}$, where L is the number of SNPs in a window (or the window size); $w_{ij} = 1$ indicates that the jth SNP is not selected in the ith window, and $w_{ij} = 0$ otherwise. Most SNPs appear in multiple windows. Each marker's relative redundancy is computed by averaging its corresponding

w_{ij} over all the windows in which it appears, and it is recorded in another vector $\mathbf{RR} = \{rr_m, m = 1, \ldots, M\}$, where M is the total number of markers in the region and rr_m is the relative redundancy of the mth SNP. A SNP is dropped from the final list when its relative redundancy is above a predetermined threshold. Note that the window size (L) and the relative-redundancy threshold are adjustable parameters.

Ideally, the sliding window size should be changed to reflect differing amounts of LD in the data. More SNPs should be included in a window and examined together when they are in regions of extensive LD, and fewer SNPs should be examined together in regions of more limited LD. Practically, it is difficult to identify regions of high and low LD and choose the window sizes accordingly. Therefore, we propose applying the procedure multiple times with a fixed window size, using the selected set of SNPs resulting from each run as the input set for the subsequent run. When a contiguous group of SNPs in LD with each other exceeds the window size, some of those SNPs will likely be dropped in the first run, bringing more of the group's SNPs within a window's length of each other in each subsequent run. We will refer to this setup as "repeated runs"; additional runs can be repeated until the procedure converges, and convergence is achieved when the difference between the number of markers before and after selection represents ⩽5% of the markers before selection. We will refer to the procedures based on spD and div in uppercase, as "SPD" and "DIV," respectively.

Simulation Studies

Two simulation studies are designed to study the performance of SPD and DIV. The first study investigates how the procedures behave when applied to SNPs in LE; the second study investigates what sample sizes provide consistent selection results. If we put these procedures into a hypothesis-testing framework, then the first study is similar to controlling the false-positive rate under the null hypothesis, "How often do we drop important markers that should be included in our set?" Also note that we are interested in the true-positive rate; or, "How often do we drop markers and maintain the desired information when there are redundancies among them?" This will be addressed by using the experimental data, rather than by using a simulation approach.

Simulation Study I: Will the SNP-Selection Procedure Drop "Important" SNPs?

When the SNPs are in LE, all SNPs should be selected if both SPD and DIV drop only SNPs that are redundant. The performance of both SPD and DIV will be affected by the set of parameters used, such as the sliding window size, the percentage of the variation explained, and the relative-redundancy threshold for each marker. We identify suitable parameter combinations that allow us to limit the drop rate of SNPs in LE (i.e., the false-positive rate) to <5%. First, genotype data for 50 SNPs are simulated. Each SNP's population allele frequency is randomly and independently drawn from the uniform distribution in (0,1) and is truncated to be between 5% and 95%. Then, after obtaining a random sample of subject genotypes on the basis of the population allele frequencies, we apply either the SPD method or the DIV method, record the percentage of markers (of 50) dropped, and average this percentage over 100 simulation runs. Note that the repeated-runs setup is not used here, since the purpose is to find parameter combinations that ensure that the proportion of nonredundant SNPs dropped at each single run is <5%. Note also that this threshold is the convergence criterion for repeated runs. If the drop percentage in any one run is <5% (i.e., below the false-positive rate), then we stop the procedure and declare convergence, to prevent the dropping of informative SNPs. For SPD, we investigate the parameter combinations: sliding window sizes of 2, 5, 10, 15, and 30 and percentages of the variation explained of 85% and 90%. For DIV, we investigate sliding window sizes of 2, 3, 5, and 7 and percentages of the variation explained of 92% and 96%. Note that the values for the percentages of the variation explained are calculated using different methods for SPD and DIV, have different interpretations, and cannot be directly related to each other. For both SPD and DIV, we investigate relative-redundancy thresholds of 50%, 70%, and 90% and sample sizes of 10, 50, 100, and 200 individuals. In addition, we look at the effect of availability of haplotype-phase information by providing the same data in both haplotype-phase–known and haplotype-phase–unknown forms.

The percentage of variation explained, as a parameter, significantly determines the proportion of the LD information retained. We need to set this parameter high enough to conserve the required amount of LD to map susceptibility genes successfully but low enough to achieve a useful reduction in SNP numbers. However, the optimal amount of information is affected by many factors, including the effect size and phenotype model of the susceptibility gene being mapped, the local LD pattern, marker-allele frequencies, and distribution of markers. Therefore, an optimal value for the percentage of variation explained for all cases will not exist. To provide some guidance as to acceptable values for this parameter, we examined the effects that a wider range of values for the percentage of variation explained have on SNPs in LE while keeping window size, relative-redundancy threshold, and sample size constant. As mentioned above, the variation-explained percentages for SPD and DIV cannot be compared to each other directly. However, we can find values for the parameter for each method that result in roughly similar behavior in the simulated data set. This allows us to calibrate the behavior of the methods under the null hypothesis of high importance of all markers and is analogous to setting a common rejection region for power comparisons of statistical tests. The SPD procedure, as described in the "Methods" section, calculates an L_r measure that essentially prevents dropping SNPs under LE, thus making it impossible to find the corresponding DIV variation-explained percentages. We simulated data as described above, applied SPD (without L_r index) or DIV with a range of the variation-explained percentages, and computed the average percentage of SNPs dropped. The sample sizes considered were 50 and 100. The rest of the parameters were chosen on the basis of the results from the first part of simulation study I.

Simulation Study II: What Sample Size Is Required in Order to Ensure Consistent Results across a Sample?

Ideally, we would like to obtain information about the most informative markers representing the population of interest with as few samples as possible, to keep genotyping costs low. Therefore, the consistency of our procedures is investigated as a function of the sample size required for marker selection. First, we simulate a large diploid data set, later referred to as "the population," containing a chromosomal region with 50 SNPs and 20,000 individuals by using a forward simulation model that assumes constant population size, nonoverlapping generations, random mating, and no other disturbing forces except recombination. Initial LD in the data is created by mixing two populations with discrepant allele frequencies and no control of the frequency range. The number of generations, the recombination rate, and the initial LD determine the degree of LD in the final generation. To ensure that our simulated data is realistic, we study the LD patterns in several regions for which we have experimental data, select high- and low-LD regions (with criteria defined later, in the "Results" section), and adjust the parameters in our simulation program to mimic these patterns in our simulated data sets. Then, we sample a certain number of individuals without replacement from the population, and we apply either SPD or DIV to each sample. For both methods, we fix the relative-redundancy threshold at 75% and the sliding window size at 5, and we test sample sizes of 10, 50, 100, and 200 individuals by using the variation-explained percentages 85% and 90% for SPD and 92% and 96% for DIV. We also apply repeated runs until the procedure meets the convergence criterion. For each parameter combination, we record, in a vector $v_1, \ldots, v_{50}$, the percentage of times that each SNP, 1,2,…,50, is dropped in 100 nonoverlapping samples from the population. Note that v_i values close to 0 or 1 are preferred, since, respectively, they indicate that the SNP gets dropped or kept each time in the simulations. The consistency for each marker is evaluated using the mean square error (MSE) of its dropping:

$$\mathrm{MSE}_i = \frac{1}{100}\sum_{k=1}^{100}(y_{ik} - v_i)^2 = v_i(1 - v_i) \ ,$$

where y_{ik} is an indicator variable, indicating whether the ith marker gets dropped in the kth sample, and v_i is the drop percentage for the ith marker over 100 simulations. Then, the overall consistency, the average MSE of the drop percentage for all markers, is calculated as

$$\mathrm{MSE} = \frac{1}{50}\sum_{i=1}^{50} v_i(1 - v_i) \ .$$

We also provide the same data in both haplotype-phase–known and haplotype-phase–unknown formats, to study the effect that haplotype information has on our results.

Evaluation Criteria

We propose to evaluate the information retained about a region by using two metrics that summarize haplotype information: the number of frequent haplotypes and their haplotype frequencies. Unfortunately, haplotype phase is unobservable in most cases. We use the EM algorithm with a sliding window to infer the haplotype frequencies. Our procedure is as follows: apply a sliding window with window size equal to 5, and estimate the haplotypes by using the EM algorithm in each window. We chose a window size of 5 because the calculation of EM frequencies for five SNPs is computationally feasible and sample sizes of 50 and 100 individuals provide enough genotype information to get reasonable estimates. Some unique situations may require other window sizes. Compute two measures to evaluate the information: count the number of the frequent haplotypes, defined as the haplotypes with frequencies >5%, and calculate the heterozygosity as $1 - \sum p_i^2$ for these haplotypes. Then, either SPD or DIV is applied to select "informative" SNPs. If the selected SNPs can represent most of the information in each window, then we expect to observe nearly the same number of frequent haplotypes and the same frequencies when we only use the selected SNPs to infer these. Therefore, we use only the selected SNPs to estimate the haplotype frequencies within the previously defined windows, we compute the two above measures again, and we compare the measures before and after selection. We define an acceptable difference for haplotype heterozygosity as >90% of the windows having a heterozygosity difference $\leqslant 0.1$. When repeated runs are used, we evaluate the final selected marker set against the initial full data set (for details, see the "Discussion" section). We judge the procedure performance by the

Table 1

Percentage of SNPs Dropped in LE—Using SPD, with Variation Explained 85% and Haplotype Phase Unknown

Sliding Window Size	SNPs Dropped When (%)					
	Sample Size Is 10 and			Sample Size Is 50 and		
	RR = 50%	RR = 70%	RR = 90%	RR = 50%	RR = 70%	RR = 90%
2	0	0	0	0	0	0
5	7	2	1	0	0	0
10	45	21	6	0	0	0
15	70	41	12	0	0	0
30	91	74	28	36	15	5

NOTE.—Results averaged over 100 simulations. RR = relative redundancy.

differences along the chromosomal region, as well as by their overall distributions.

Data Sets

Using linkage analysis, we identified a 12-cM region on chromosome 12 centered at D12S853 as likely to contain a susceptibility gene for type 2 diabetes (Ehm et al. 2000). Six hundred forty-nine SNPs distributed across this region were genotyped in 138 unrelated white individuals. The SNPs have been placed on a 12-Mb composite map by using a combination of STS content mapping and sequence analysis. Of these 649 SNPs, 604 have MAF >5%.

To illustrate the impact that marker selection has on an association study, we used data from the study by Hosking et al. (2002), in which 32 markers surrounding the *CYP2D6* gene on chromosome 22 were typed in 1,018 white individuals. Twenty-seven SNPs had MAF >5%. All SNPs were mapped to an 879-kb contig flanking the *CYP2D6* locus. Hosking et al. reported significant associations between SNPs and the poor-drug-metabolizing phenotype. We first reproduced the results from the study by Hosking et al. by using the same association tests; we then applied SPD and DIV to select markers by using 100 randomly selected controls. Using only the markers selected by SPD and all 1,018 individuals, we conducted Fisher's exact test (used by Hosking et al.) for single-marker genotypic tests, used only selected SNPs to estimate haplotype frequencies within the previously defined windows, and applied the regression-based haplotype tests (used by Hosking et al.) with the estimated haplotype frequencies. We plotted the test *P* values versus the marker positions of the full data set and the selected set, and we compare their patterns.

Results

Simulation Study I

Table 1 shows the percentage of SNPs dropped when markers are in LE as determined by using SPD, with the percentage of variation explained set to 85% when the haplotype phase is unknown, for sample sizes of 10 and 50. No markers were dropped for sample sizes of 100 and 200, and, therefore, these percentages are not shown. The pattern of markers dropped was similar for a variation-explained parameter of 90%, except that fewer markers are dropped (data not shown). The results show that SPD will not drop SNPs in LE under most parameter combinations, unless the sample size is small relative to the window size, when some independencies of SNPs will not be represented in the sample. The small percentages of SNPs dropped are probably the result of using the L_r measure to determine if there is redundant information in the sample and dropping SNPs only if redundancy appears to exist. In conclusion, for sample sizes of 50 individuals or more, SPD retains important SNPs for window sizes ≤15.

Table 2 shows results for DIV when the percentage of variation explained is 92%. The results are similar when the percentage of variation explained is 96%, except that fewer markers are dropped. Compared with SPD, DIV drops more SNPs in LE. A similar pattern is observed, in which more SNPs are dropped when window size is large and sample size is relatively small. The percentage of SNPs dropped is relatively stable when the sample size is >50 and the relative-redundancy threshold is >70%. The results guide us in choosing a sliding window size: it should be large enough to include a substantial amount of variation but small enough that large sample sizes are not required in order to capture the important variation. When the window size is 5 and the relative-redundancy threshold is >70%, the percentage of SNP dropped is close to

Table 2

Percentage of SNPs Dropped in LE—Using DIV, with Variation Explained 92% and Haplotype Phase Unknown

Sliding Window Size	SNPs Dropped When (%)											
	Sample Size Is 10 and			Sample Size Is 50 and			Sample Size Is 100 and			Sample Size Is 200 and		
	RR = 50%	RR = 70%	RR = 90%	RR = 50%	RR = 70%	RR = 90%	RR = 50%	RR = 70%	RR = 90%	RR = 50%	RR = 70%	RR = 90%
2	3	<1	<1	<1	0	0	0	0	0	0	0	0
3	6	2	2	<1	0	0	<1	0	0	0	0	0
5	24	14	9	10	6	4	7	5	3	5	4	2
7	39	28	15	21	16	8	15	11	6	13	10	6

NOTE.—Results averaged over 100 simulations. RR = relative redundancy.

5%, even with a small sample size (e.g., 50). On the basis of these simulations, we selected a sliding window size of 5 and a relative-redundancy threshold of 75% for many of our further analyses. To make SPD and DIV easier to compare, we used the same values for these two parameters for SPD as well.

We also explored the behavior of the two methods when haplotype phase is known. The percentages of SNPs dropped are smaller when the haplotype phase is known, because inferring this information by use of a mathematical algorithm such as EM results in an information loss (data not shown). Nevertheless, the percentages show a similar pattern, in which, to control the percentages of SNPs dropped, relatively small window sizes are needed when the sample size is small, although it may be possible to apply the algorithms by using a smaller sample size when the phase is known.

Table 3 shows the percentages of SNPs dropped for SNPs in LE, for a range of variation-explained percentages, as determined by using either SPD or DIV. The variation-explained percentages should be ⩾94% for DIV and 75% for SPD when the sample size is 50 and 92% for DIV and 70% for SPD when sample size is 100, if the false-positive rate for the markers in LE is controlled to be <5%. We propose these values as the "safe" starting point, to avoid dropping important markers when the degree of LD in the data is either hard to measure or complex. The variation-explained percentage could be determined by testing a large range of values for the experimental data and a value chosen when the haplotype information before and after selection is similar. However, the results may become too complicated to interpret when the real data consist of regions with differing degrees of LD and the repeated-runs setup is applied. Therefore, using a variation-explained parameter that we know is simpler, because it does not result in dropping too many SNPs under LE.

Table 3

Percentage of SNPs Dropped in LE—Using Different Variation-Explained Values

	SNPs Dropped When (%)			
	Sample Size Is 50, for		Sample Size Is 100, for	
Variation Explained (%)	DIV	SPD	DIV	SPD
65	ND	12.6	ND	11.8
70	ND	8.9	ND	4.2
75	ND	3.9	ND	3.4
80	19.7	3.4	17.0	3.4
85	14.7	3.2	13.3	1.7
90	10.0	.1	7.9	0
92	6.5	ND	4.8	ND
94	3.4	ND	1.7	ND
95	ND	0	ND	0
96	1.0	ND	.3	ND
98	0	ND	0	ND

Note.— Results averaged over 100 simulations. ND = not done.

Simulation Study II

To study the sample sizes needed to obtain consistent selection results, we simulated data sets containing differing degrees of LD with patterns similar to our experimental data. We divided our data on chromosome 12 into six regions (each containing ~110 markers) and treated chromosome 22 as a single region, to study the LD patterns of these seven regions. D' was calculated for each marker pair within each region and was averaged according to the number of intervening markers. The averaged D' was plotted versus the number of intervening markers. As expected, for all regions, the average D' decreases as the number of intervening markers increases (graphs not shown). Two regions, one with the fastest and one with the slowest decrease in LD, with increasing distance between markers were selected as high- and low-LD regions, respectively. The average D' drops under 0.5 when the markers considered are separated by 30 intervening markers in the high-LD region; for the low-LD region, the average D' drops under 0.5 after a separation of five markers. Two data sets, each with 50 SNPs, were generated to produce similar LD patterns by adjusting the number of generations evolved and the recombination rate used in the simulation.

Figure 1 shows each SNP's average drop percentage across 100 simulations when SPD is applied with variation-explained percentage 90% and window size 5, for the high-LD data. Figure 2 shows the results when DIV is used with variation-explained percentage 92% and window size 5, for the high-LD data. Note that SPD and DIV may select different markers across different samples from the population, because these markers are highly correlated and provide somewhat equivalent information, and which marker(s) provides more information in each sample is affected by the statistical-sampling variation. Therefore, for the two methods, we expect either patterns that are relatively stable but not identical or patterns with 0% or 100% drop for each marker. Both figures 1 and 2 suggest that sample sizes ⩾50 are needed to achieve consistent results. When the sample size is increased from 50 to 100 or from 100 to 200, the consistency improves a little. With higher variation-explained values or a low degree of LD in the data, a similar pattern is obtained with even more consistent results (data not shown). Knowledge of haplotype-phase information does improve the consistency, but its effect is small and can be compensated for with a slight increase of the variation-explained percentage or the sam-

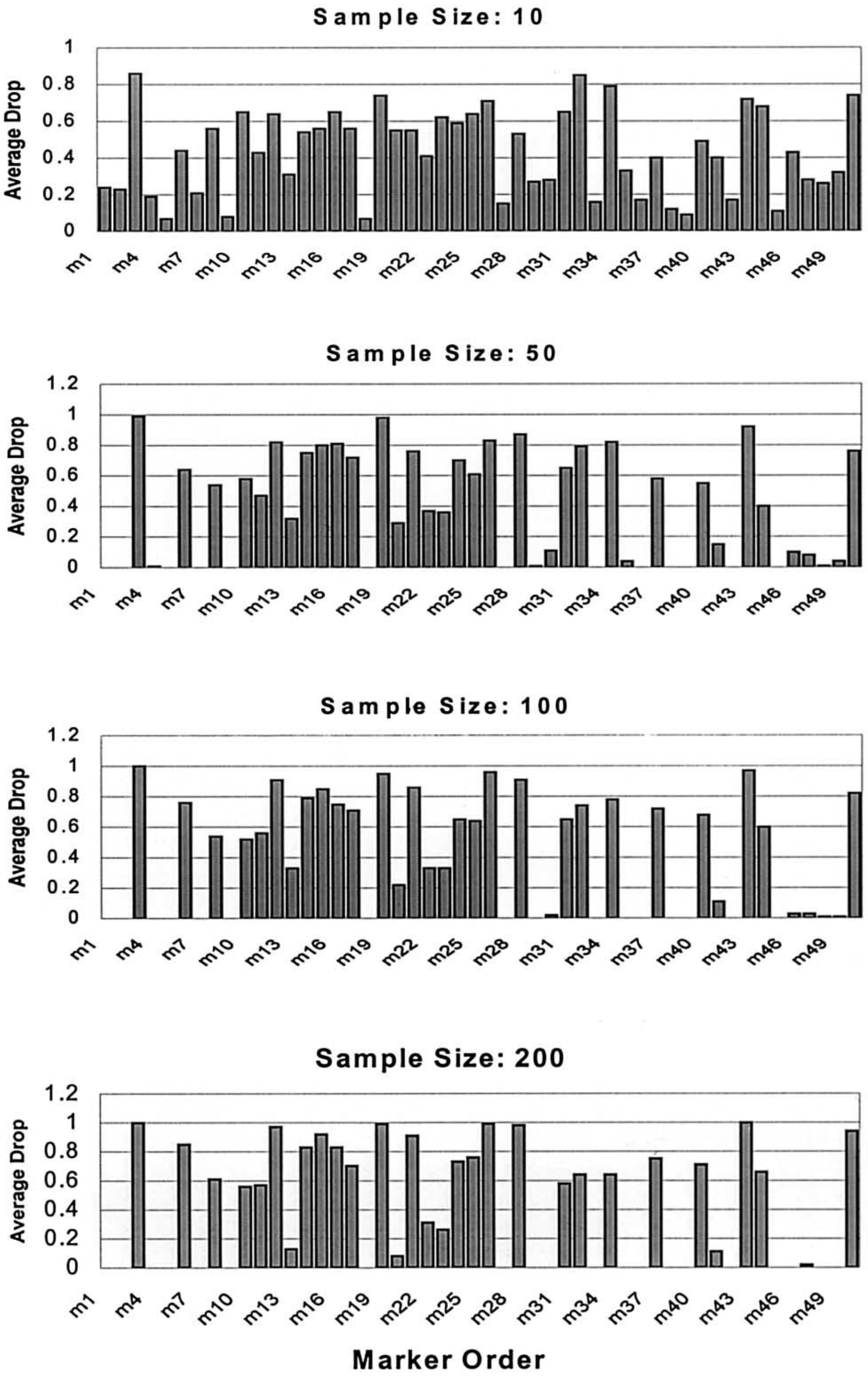

Figure 1 Average drop percentage of 50 SNPs across 100 simulations on the high-LD data when haplotype phase is unknown. SPD is used with variation explained 90% and window size 5. From top to bottom, the graphs are for sample sizes equal to 10, 50, 100, and 200; for each respective graph, the average number of SNPs dropped is 25.5, 19.7, 20.1, and 20.2, and the average MSE of dropping is 0.17, 0.11, 0.09, and 0.08.

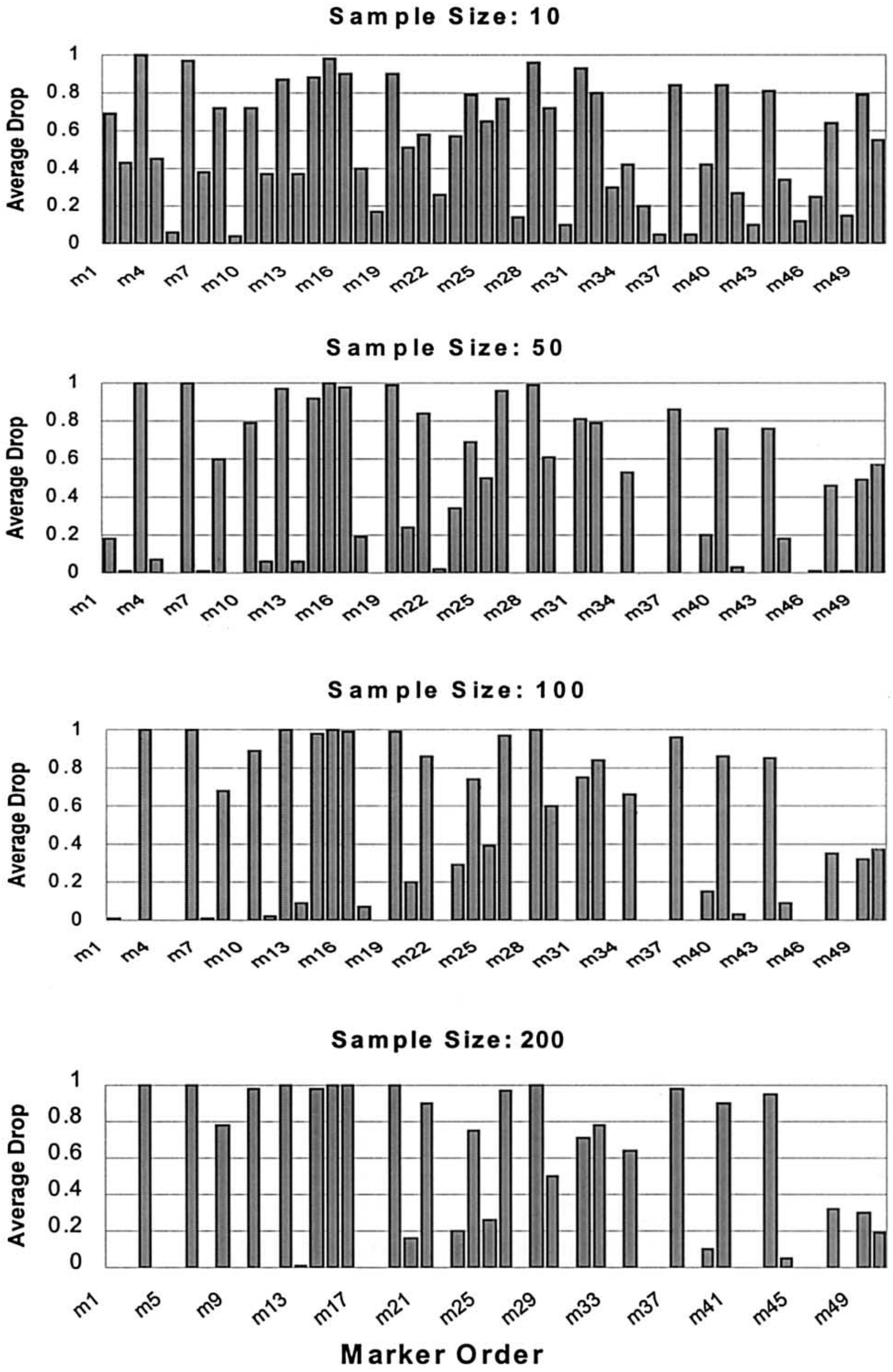

Figure 2 Average drop percentage of 50 SNPs across 100 simulations on the high-LD data when haplotype phase is unknown. DIV is used with variation explained 92% and window size 5. From top to bottom, the graphs are for sample sizes equal to 10, 50, 100, and 200; for each respective graph, the average number of SNPs dropped is 26.2, 20.4, 20.0, and 19.4, and the average MSE of dropping is 0.16, 0.09, 0.07, and 0.06.

ple size. For both SPD and DIV, almost the same number of markers and the same markers are dropped when the same parameters are used and the sample size is large enough (⩾50), regardless of whether the haplotype information is observable. On the basis of the results of simulation study II, we recommend a sample size of 50–100 individuals, depending on the expected amount of missing data.

Experimental Data Results

In applying the procedure to the experimental data, we also used a sliding window size of 5 and retained a SNP when its relative redundancy was <75%. We used repeated runs and utilized the convergence criterion mentioned above (see the "Simulation Studies" section). On the basis of the results in table 3, we chose starting variation-explained values of 70% for SPD and 92% for DIV, respectively, since we used sample sizes close to 100 individuals for all experimental data. To evaluate the outcome of marker selection, we used a sliding window size of 5 when calculating the number of frequent haplotypes and the haplotype heterozygosity. We included all the available SNPs even when their MAFs were <5%, to achieve a relatively comprehensive picture of LD before selecting markers, regardless of the SNP allele frequencies. We adjusted the variation explained for each method until >90% of the windows had haplotype heterozygosity differences <0.1. We chose the haplotype heterozygosity because we found the number of frequent haplotypes to be unreliable in certain situations; for example, the number of frequent haplotypes can change owing to a small change of one haplotype frequency (e.g., from 4% to 5.5%). These haplotype-frequency changes contribute little to the difference in heterozygosity. For all experimental data sets, we presented results as long as one procedure (SPD or DIV) achieved this goal; for the other procedure, we adjusted the variation-explained percentage so that it retained a similar number of markers, to make our comparisons fair.

For the chromosome 12 data, the above procedure resulted in variation-explained values of 90% for SPD and 96% for DIV. Using these settings, from a total of 649 markers, we selected 415 (63.9%) by SPD and 412 (63.5%) by DIV. Histograms of the differences in the haplotype heterozygosity and the number of frequent haplotypes for each window are shown in figure 3. The proportions of windows with differences <0.1 in the heterozygosity are 85.2% for the SPD method and 92.2% for the DIV method. The proportions of windows with differences ⩽1 in the number of frequent haplotypes are 93.4% for SPD and 95.3% for DIV. For the chromosome 22 region, we have a much larger sample size than that required for marker selection. We randomly selected 100 controls and applied our procedure as if this were the sample size collected for marker-selection purposes. Variation-explained percentages of 75% and 92% were used for SPD and DIV, respectively, and, of the 32 markers reported by Hosking et al. (2002), 20 (62.5%) were selected in both instances. The proportions of windows with differences <0.1 in heterozygosity are 88.9% and 92.9% for SPD and DIV, respectively. We used relatively low variation-explained percentages for each method, compared with those for the chromosome 12 data, and we achieved smaller differences before and after selection. One explanation for this finding may be the homogeneous nature of the LD pattern in this data set. The procedures were applied to several other data sets (data not shown), in different samples with different LD patterns. The results in all the data sets tested showed that the majority of haplotype information could be maintained while achieving substantial reduction in the number of SNPs that need to be analyzed.

We analyzed the markers selected by SPD from the chromosome 22 data for association with the *CYP2D6* poor-metabolizer phenotype. Note that, because almost one-half of the markers show strong association with the phenotype, it does not make sense to evaluate the procedure on the basis of whether the selected marker set includes the significant signals. Therefore, we chose to compare the test *P* values' patterns by using the full data and the selected data. Note that 20 of 27 markers were selected and that the markers with the most significant genotypic tests were among those selected (data not shown). Figure 4 contains the haplotype test results for the full marker set, with a sliding window size of 5, and for the selected marker set, within the context of the windows defined by the full data; the two curves almost overlap, which is not surprising, since the selected markers preserved the information content of the full data well. Therefore, marker selection had a negligible impact on the results of this association study. Similar results were duplicated in other association studies (data not shown).

Discussion

There are several fundamental differences between the two methods, spD and div. The spD method relies on two-locus LD (i.e., pairwise correlation) and single-marker characteristics, whereas the div method relies on haplotype frequencies, which involve not only two-locus LD coefficients and allele frequencies but also LD coefficients pertaining to alleles at three or more markers (Bennett 1954). Haplotypes will provide more information than do pairwise LD measures if the LD measures involving three or more markers make a significant contribution; otherwise, haplotype frequencies are just linear combinations or summaries of pairwise LD and allele frequencies. M. G. Ehm, D. M. Nielsen,

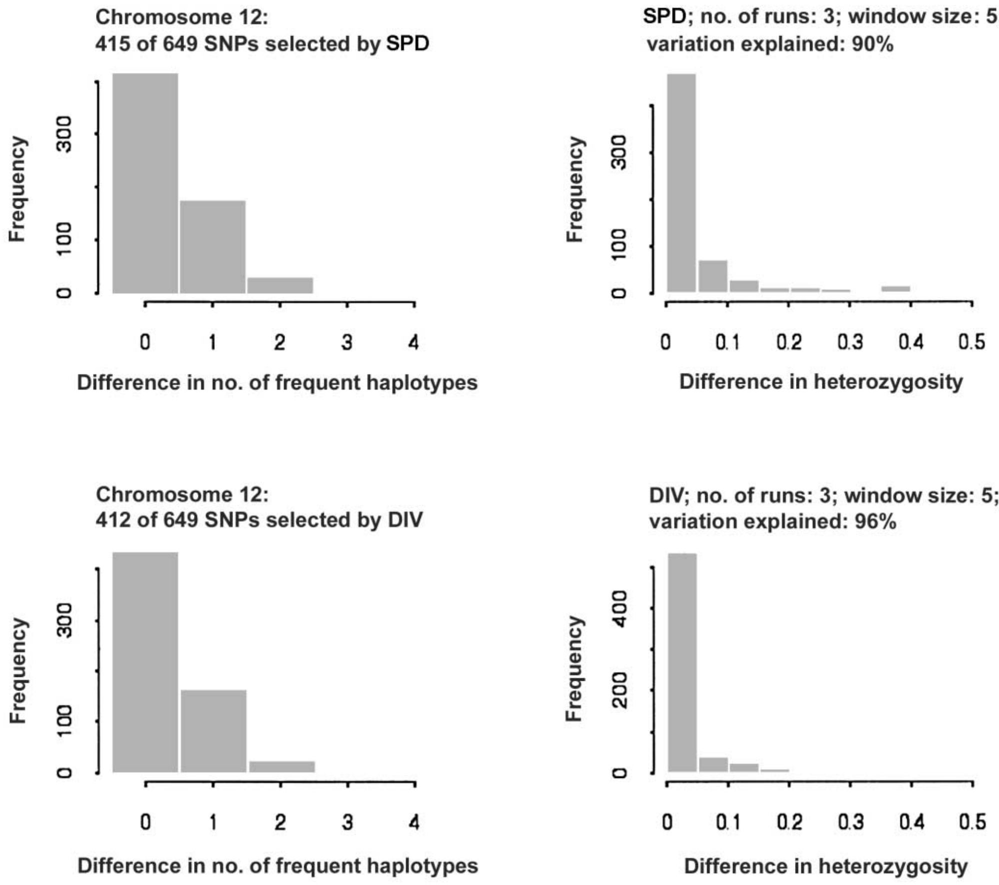

Figure 3 Overall evaluation for the procedure—with SPD using variation explained 90% (*top*) and DIV using variation explained 96% (*bottom*), on the chromosome 12 region with 649 SNPs. The histograms show the differences in the number of frequent haplotypes (*left*) and the differences in heterozygosity (*right*), before and after marker selection. For both methods, sliding window size 5 was used, and the procedure converged on the third run.

Z. Meng, M. Karnoub, C.-F. Xu, D. Zaykin, E. H. Lai, M. J. Wagner, D. K. Burns, and B. S. Weir (unpublished data) summarized the decay and extent of two- and three-locus LD in several genomic regions, including the chromosome 12 and chromosome 22 regions described here. They found that LD based on alleles at three loci decays more quickly than two-locus LD. The extent of three-locus LD is relatively small across the chromosome 12 locus, except for the central region, near 5,000 or 6,000 kb, where there is a large amount of three-locus LD. Therefore, it is reasonable to investigate marker-selection procedures based on two-locus LD and single-marker characteristics. Our observation that there is no major difference in the overall performance of the procedures based on spD and div supports the above hypothesis. A minor observation is that spD tends to drop markers with closer or higher MAFs, because it relies on the correlation *r,* which achieves higher values when marker MAFs are close or relatively high. In contrast, div tends to drop markers with disparate allele frequencies, because it is based on the heterozygosity and markers with less-frequent alleles contribute less to this measure. Although we did not see significant differences in the performance of the methods when the results were evaluated using our evaluation criteria, this difference may explain why the methods do select different markers. With both methods, we preselect markers according to their allele frequencies and retain SNPs only when their MAFs are >5%. This may make the two procedures more comparable, since SNPs with very disparate allele frequencies will not be retained. Furthermore, the typing of markers with frequencies <5% may be less efficient in the association-study design. Finally, spD is less computationally intensive and can be applied to analyze a larger number of SNPs (e.g., candidate genes typed for several dozen mark-

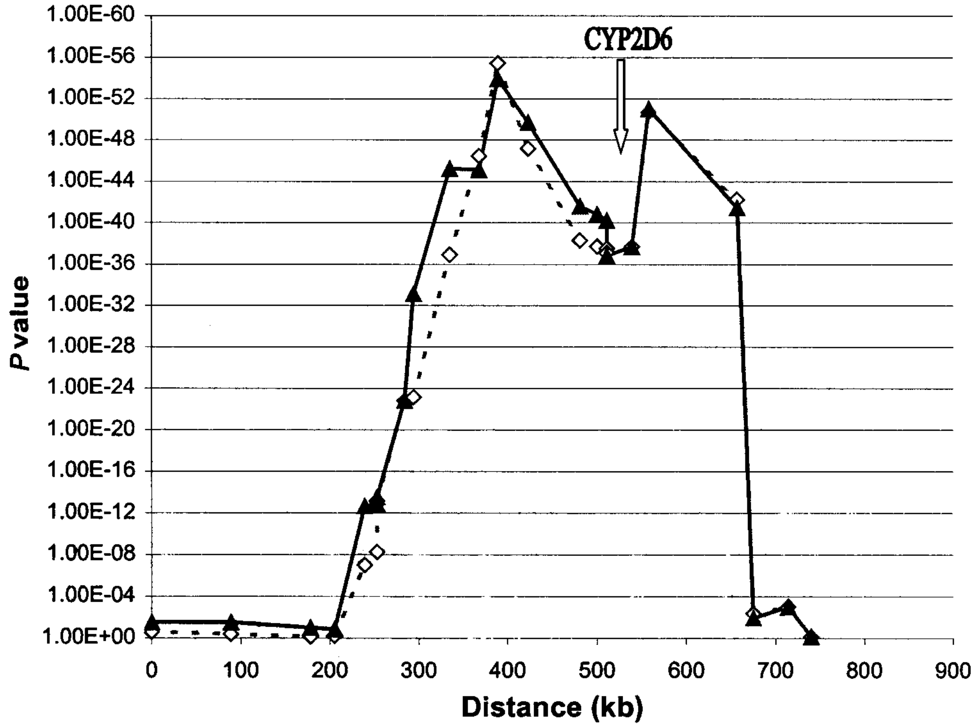

Figure 4 Association between *CYP2D6* poor-metabolizer phenotype and haplotypes. Haplotypes were derived by the EM algorithm, from windows of consecutive SNPs. Symbols indicate the first markers in a window: unblackened diamonds (◇) represent haplotype test results of 27 markers with window size 5; blackened triangles (▲) represent haplotype test results of 20 selected markers with markers placed in previously defined windows.

ers) without using sliding windows, since it is based on a summary of pairwise LD and does not require haplotype-phase information. The div approach is constrained by its computational limitations, because it relies on the haplotype information that has to be estimated if unknown. The required computational time to estimate haplotype frequencies in unrelated individuals by using a numerical algorithm such as EM increases dramatically as the number of markers increases. Furthermore, as described in the "Methods" section, the optimal htSNP set is found by an exhaustive search of all the possibilities. Therefore, div, by itself, is quite time-consuming when more than seven SNPs are considered. It may be possible to reduce the computational burden of div by substituting its exhaustive search step with a preapplication of spD. In brief, the order of markers' contributions to the total variation can be obtained by applying the spD method. SNPs in the ordered marker set can be included in the htSNP set of the div method, gradually, until a criterion is met (details still under study).

The LD matrix that spD is based on is the correlation matrix of marker genotypes, which is semipositive when there is complete data. However, when there is missing genotype data, the correlation (or covariance) matrix may occasionally be negative and may present a problem in the spD analysis. This problem is more likely to occur as the proportion of missing data or the number of markers for a given sample size increases. Although the problem did not occur in the present study, caution is needed. The normalized measure of LD (D') is not recommended for use with the spD method, since it may exaggerate the problem. Moreover, the resulting matrix is neither the correlation nor the covariance; therefore, the eigenvalue-eigenvector pairs lack clear statistical interpretation.

Zhang et al. (2002) proposed a dynamic programming method for the performance of haplotype-block partitioning, to minimize the SNPs that are needed to represent common haplotypes. This method can utilize different measures of the block quality, including the ratio of the number of SNPs in the block to the minimum number of SNPs required in order to define haplotypes or the proportion of div explained by a subset of the SNPs in a block (i.e., the htSNP method [Clayton 2001]). Similarly, another possible measure of the block quality could be the number of SNPs that are needed to achieve a previously specified measure of the variation explained using the spD method. Whereas Zhang et al. defined blocks with a minimum number of SNPs by using any of a number of measures of block quality, our goal is to select a subset of SNPs, preserving local haplotype frequencies within a relatively short genetic distance and thus maintaining similar measures of haplotype information before and after marker selection. This leads us to the sliding window approach, which does not rely on the block definitions. Furthermore, our procedure is designed to be applicable to data when haplotype-phase information is not available.

There are several approaches that could be considered in defining the convergence for repeated runs. One choice is to run the procedure until no more markers are dropped. We found that this worked well for a homogeneous LD region, but, when a region exhibits a mixed amount of LD, too many repeated runs can lead to the dropping of informative markers and information loss. One possible reason is that markers in LE get dropped if the procedure is applied to data with less and less correlation. Therefore, for a complex chromosomal region, we suggested the 5% cutoff, as described in the "Methods" section. One might consider whether it makes sense to use repeated runs at all. We found that using a higher variation-explained percentage in combination with repeated runs was preferable (i.e., had less information loss) to one with a lower variation-explained value with no repeated runs.

In evaluating our procedures, we used the heterozygosity and the number of the frequent haplotypes to measure the haplotype information content. We felt that these measures captured aspects of haplotypes important in association studies. Any other suitable measures, such as matching haplotype frequencies before and after selection, could be used. Furthermore, the evaluation procedure can vary according to the different requirements of the studies. We used a fixed variation-explained percentage, ran the procedures repeatedly until they converged, and evaluated the information content of the selected marker set against that of its initial full data within the initially defined sliding windows. One advantage of this approach is that it provides an overall evaluation for all repeated runs. However, it may be too conservative, since markers are selected on the basis of their LD structure in the sliding windows and some of the windows may lose their initial meanings after the repeated marker reductions. This can result in some large differences in the haplotype information measures as a result of the selection process, but these differences do not necessarily indicate serious information loss. An alternative approach would be to evaluate the information content of the selected marker set against that of its immediate input data for each repeated run. There are two advantages for the latter approach: First, all selected markers are evaluated in the windows they are selected. Second, the variation-explained value could be adjusted to control the information loss for each repeated run. However, it is not clear how we summarize the overall performance of such a procedure.

In the present article, the selection procedures are applied to the markers discovered and typed in population controls—samples chosen regardless of phenotype. We have found it useful to ensure that SNPs are polymorphic in the ethnic group of interest before typing them on expensive disease samples. Currently, we are typing a large number of SNPs in a panel of 100 population samples and using these data in the marker selection. Since we are studying several diseases, using a population sample allows us to use the marker selected in further studies, regardless of the disease. However, this approach does assume the common-disease, common-allele hypothesis: If a rare allele at a marker is responsible for the disease, then it is unlikely to be selected in such an approach. Another choice would be to apply marker selection to markers discovered in case-control samples and typed in cases only. This way, the disease alleles, although rare in the population, will have increased frequency in the sample, and the selection favoring polymorphic markers would have good justification. When case-control samples are available, procedures similar to the ones described here can be useful in selecting the most discriminative subsets of markers among those that show frequency differences between cases and controls. Analogously, the marker-selection procedure might need to be conducted separately, using different samples from different populations, if we wish to study different populations that may have somewhat different haplotype structures.

We do realize that procedures such as these have consequences. As with any statistical procedure, marker selection is always a gamble, since markers are selected mainly on the basis of LD structure, regardless of any phenotypes. Therefore, the impact that marker selection has on the results of an association study, in general, is not known. Although we have showed one example in which marker selection had a negligible effect on the association results, the impact can vary greatly from case to case. For instance, the required percentage of information retained in the association study might depend on effects of disease-susceptibility genes, which are hard to assess before actually conducting the association test. Furthermore, it is possible that a causal marker is not selected because it is highly correlated with nearby markers. Fortunately, it appears that analyzing the data by using haplotype analysis would reduce the impact of such a selection. In summary, marker selection should be viewed as providing a way to prioritize markers for a first genotyping screen, and more markers can always be typed in the regions of interest later.

Acknowledgments

We thank Bruce Weir, Mike Boehnke, and two reviewers, for their helpful comments on the manuscripts; Eric Lai, Clive Bowman, Mike Mosteller, Brian Browning, and Georgiy Bobashev, for their valuable advice; and Achamma Philip, for her technical support.

Appendix A

Obtaining the Matrix of Pairwise LD for Diallelic Markers

Calculation of matrices of pairwise LD is most straightforward for markers with two alleles and can be handled with a simple command by using standard statistical software—for example, by invoking the cor() function in R/Splus. The following method requires that marker genotypes are recoded as follows:

$$\text{New value} = \begin{cases} -1 & \text{if genotype is 11} \\ 0 & \text{if genotype is 12} \\ 1 & \text{if genotype is 22} \end{cases} .$$

A pair of SNPs will be represented by two vectors x and y with entries as just indicated. It can be easily shown that the usual sample covariance

$$C_{\mathrm{AB}}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum x_i y_i - \frac{1}{n^2}\sum x_i \sum y_i$$

is twice the composite LD, Δ_{AB}, reported by Weir (1996). To see this, the terms of the covariance can be written in terms of two-locus counts, as

$$\sum x_i y_i = n_{\mathrm{AABB}} - n_{\mathrm{AAbb}} - n_{\mathrm{aaBB}} + n_{\mathrm{aabb}}$$

and

$$\sum x_i \sum y_i = (n_{\mathrm{aa}} - n_{\mathrm{AA}})(n_{\mathrm{bb}} - n_{\mathrm{BB}}) .$$

Then, $C_{\mathrm{AB}}(\mathbf{x},\mathbf{y}) = 2\Delta_{\mathrm{AB}}$ follows from the relation

$$\begin{aligned} \Delta_{\mathrm{AB}} &= \frac{1}{4}(\Delta_{\mathrm{AB}} + \Delta_{\mathrm{ab}} - \Delta_{\mathrm{Ab}} - \Delta_{\mathrm{aB}}) \\ &= \frac{1}{2n}(n_{\mathrm{AABB}} - n_{\mathrm{AAbb}} - n_{\mathrm{aaBB}} + n_{\mathrm{aabb}}) \\ &\quad - \frac{1}{2n^2}(n_{\mathrm{aa}} - n_{\mathrm{AA}})(n_{\mathrm{bb}} - n_{\mathrm{BB}}) . \end{aligned}$$

These composite coefficients are unbiased estimates of the population LD under Hardy-Weinberg equilibrium (HWE). When HWE does not hold, they include an additional component that measures covariance between alleles between different haplotypes in an individual. The diagonal elements $C_{\mathrm{A}}(\mathbf{x},\mathbf{x})$ of the variance-covariance matrix $\mathbf{C}(\mathbf{x},\mathbf{x})$, are $4n$ times variances of allele frequencies, $\mathrm{Var}\,(p_{\mathrm{A}}) = [p_{\mathrm{A}}(1 - p_{\mathrm{A}}) + D_{\mathrm{A}}]/2n$, where p_{A} is the allele frequency of allele A, $D_{\mathrm{A}} = p_{\mathrm{AA}} - p_{\mathrm{A}}$ is the deviation from HWE, and p_{AA} is the frequency of genotype AA. In terms of recoded values, the allele frequencies are

$$\tilde{p}_{\mathrm{A}} = \frac{1}{2} - \frac{\sum x_i}{2n}$$

and

$$\tilde{p}_{\mathrm{B}} = \frac{1}{2} - \frac{\sum y_i}{2n} .$$

Finally, the correlation of Weir (1996), defined as

$$r_{\mathrm{AB}} = \frac{\hat{\Delta}_{\mathrm{AB}}}{\sqrt{[\tilde{p}_{\mathrm{A}}(1 - \tilde{p}_{\mathrm{A}}) + \hat{D}_{\mathrm{A}}][\tilde{p}_{\mathrm{B}}(1 - \tilde{p}_{\mathrm{B}}) + \hat{D}_{\mathrm{B}}]}} ,$$

can be computed from recoded values, as

$$r_{\mathrm{AB}} = \frac{C_{\mathrm{AB}}(\mathbf{x},\mathbf{y})}{\sqrt{C_{\mathrm{A}}(\mathbf{x},\mathbf{x})C_{\mathrm{B}}(\mathbf{y},\mathbf{y})}} ,$$

or by an R/Splus function call, cor().

Appendix B

Determining Effective/Redundant Numbers of Markers, L_e and L_r

Let L be the actual number of markers and $\{\lambda_i\}$ be a set of eigenvalues associated with the matrix of pairwise LD coefficients. From Cauchy-Schwarz inequality and noting that $\{\lambda_i\}$ are nonnegative, we have $\sum \lambda_i^2 \geqslant (\sum \lambda_i)^2/L$. This bound corresponds to the no-LD situation, for which $\lambda_i = \lambda_j, \forall i,j;$ in this case,

$$\sum \lambda_i^2 = \sum \lambda_i \frac{\sum \lambda_i}{L} = \frac{(\sum \lambda_i)^2}{L} \ .$$

By expanding $(\sum \lambda_i)^2$, we also see that $\sum \lambda_i^2 \leqslant (\sum \lambda_i)^2$. This bound corresponds to the maximum possible LD with a single nonzero eigenvalue; in this case, $\sum \lambda_i^2 = (\sum \lambda_i)^2$. Put together,

$$\frac{(\sum \lambda_i)^2}{L} \leqslant \sum \lambda_i^2 \leqslant \left(\sum \lambda_i\right)^2 \ .$$

Then, we have

$$0 \leqslant L \frac{\sum \lambda_I^2}{(\sum \lambda_i)^2} - 1 \leqslant L - 1 \ ,$$

so that the number of redundant markers can be defined as

$$L_{\mathrm{r}} = L \frac{\sum \lambda_i^2}{(\sum \lambda_i)^2} - 1$$

and the effective number of markers, $1 \leqslant L_{\mathrm{e}} \leqslant L$, reduced owing to LD, is

$$L_{\mathrm{e}} = 1 + L\left[1 - \frac{\sum \lambda_i^2}{(\sum \lambda_i)^2}\right] \ .$$

References

Bennett J (1954) On the theory of random mating. Ann Eugen 18:311–317

Clayton D (2001) Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf (accessed May 16, 2003)

Couzin J (2002) New mapping project splits the community. Science 296:1391–1393

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, et al (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544–548

Ehm MG, Karnoub MC, Sakul H, Gottschalk K, Holt DC, Weber JL, Vaske D, Briley D, Briley L, Kopf J, McMillen P, Nguyen Q, Reisman M, Lai EH, Joslyn G, Shepherd NS, Bell C, Wagner MJ, Burns DK (2000) Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet 66:1871–1881

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Hosking LK, Boyd PR, Xu CF, Nissum M, Cantone K, Purvis IJ, Khakhar R, Barnes MR, Liberwirth U, Hagen-Mann K, Ehm MG, Riley JH (2002) Linkage disequilibrium mapping identifies a 390 kb region flanking CYP2D6 associated with CYP2D6 poor drug metabolising activity. Pharmacogenomics J 2:165–175

Jackson JE (1991) A user's guide to principal components. Wiley & Sons, New York

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor (TCRB) locus. Am J Hum Genet 69:381–395

Weir BS (1996) Genetic Data Analysis II. Sinauer Associates, Sunderland, MA

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339