Genetic Epidemiology 31: 553–564 (2007)

Sequential Haplotype Scan Methods for Association Analysis

Zhaoxia Yu and Daniel J. Schaid*

Division of Biostatistics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, Minnesota

Multi-locus association analyses, including haplotype-based analyses, can sometimes provide greater power than single-locus analyses for detecting disease susceptibility loci. This potential gain, however, can be compromised by the large number of degrees of freedom caused by irrelevant markers. Exhaustive search for the optimal set of markers might be possible for a small number of markers, yet it is computationally inefficient. In this paper, we present a sequential haplotype scan method to search for combinations of adjacent markers that are jointly associated with disease status. When evaluating each marker, we add markers close to it in a sequential manner: a marker is added if its contribution to the haplotype association with disease is warranted, conditional on current haplotypes. This conditional evaluation is based on the well-known Mantel-Haenszel statistic. We propose two permutation based methods to evaluate the growing haplotypes: a haplotype method for the combined markers, and a summary method that sums conditional statistics. We compared our proposed methods, the single-locus method, and a sliding window method using simulated data. We also applied our sequential haplotype scan algorithm to experimental data for CYP2D6. The results indicate that the sequential scan procedure can identify a set of adjacent markers whose haplotypes might have strong genetic effects or be in linkage disequilibrium with disease predisposing variants. As a result, our methods can achieve greater power than the single-locus method, yet is much more computationally efficient than sliding window methods. *Genet. Epidemiol.*31:553–564, 2007.

Key words: case-control; single nucleotide polymorphism (SNP); gene-gene interaction; single-locus analysis; Mantel-Haenszel statistic

Contract grant sponsor: US Public Health Service, National Institutes of Health; Contract grant number: GM065450.
*Correspondence to: Dr. Daniel J. Schaid, Harwick 775, Section of Biostatistics, Mayo Clinic, 200 First Street, SW, Rochester, MN 55905.
E-mail: schaid@mayo.edu
Received 3 November 2006; Revised 13 February 2007; Accepted 11 March 2007
Published online 8 May 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20228

INTRODUCTION

With the development of DNA microchip technology, we are now able to measure single nucleotide polymorphisms (SNPs) on a genome-wide scale. However, detecting disease predisposing variants is still challenging, especially for complex diseases. Haplotypes, which are combinations of alleles on the same chromosome, may provide more information for association analyses than individual markers. Therefore, haplotype-based analyses have received much attention. Clark [2004] stated three reasons why haplotype analyses might be helpful: (1) particular combinations of amino acids might be responsible for folding of polypeptide chains; (2) haplotypes are pieces of chromosomes that are passed down from ancestors; (3) haplotype analyses can have greater power than single-locus analyses because they naturally incorporate the linkage disequilibrium (LD) among loci and potentially reduce the degrees of freedom of test statistics. In addition, with the increase of marker density, both the number of markers and the dependency among them will make the single-locus analyses with the Bonferroni correction too conservative. As a result, there have been numerous approaches for haplotype association analyses based on different designs and assumptions.

Despite the obvious benefits of haplotype-based analyses, there are debates about the relative power of haplotype-based analyses compared to single-locus analyses or multi-locus analyses that ignore the phase of alleles on haplotypes. Long and Langley [1999] and Kaplan and Morris [2001]

found that single-locus analyses have at least the same power as haplotype analyses. Clayton et al. [2004] showed that a genotype-based multi-locus method that does not use phase information can sometime be more powerful than haplotype methods. Studies by others [Akey et al., 2001; Zaykin et al., 2002; Fallin et al., 2001; Morris and Kaplan, 2002; Zhang et al., 2002; Epstein and Satten, 2003] demonstrated that haplotype-based methods can have greater power than single-locus analyses. These contradicting conclusions arise from different assumptions, underlying LD structures, and ways to analyze markers simultaneously [Schaid, 2004].

To improve the power of haplotype-based analyses, two difficulties need to be addressed: how to handle rare haplotypes and which markers should be combined for haplotype analyses. Statistical techniques such as cladistic analyses [Durrant et al., 2004], clustering [Liu et al., 2001; Molitor et al., 2003; Yu et al., 2004; Tzeng, 2005; Morris, 2005; Waldron et al., 2006], coalescent methods [Zollner and Pritchard, 2005], pattern mining [Toivonen et al., 2000], and haplotype sharing [McPeek and Strahs, 1999; Beckman et al., 2005] have been proposed, many of which reduce the degrees of freedom, leading to an increase of power. Still, choosing the proper set of markers for haplotype analyses is essential to improve power, as it is impractical and often not necessary to analyze haplotypes constructed from all markers over a long genomic region because the core associated haplotype might be short, yet a longer haplotype region would have a large number of possible haplotypes, especially when markers are in weak LD. In practice, many investigators use a sliding window method with arbitrarily chosen window length. This strategy is time consuming and it increases the number of comparisons. Cheng et al. [2005] used an exhaustive search of all possible sets of adjacent markers that flank a marker. More recently, for each putative disease locus, Bahlo et al. [2006] calculated a statistic that sums a series of nested χ^2 statistics. Although these approaches avoid arbitrarily choosing window length, they are either computationally demanding or might only outperform the traditional single-locus method in some specific scenarios.

In this paper, we propose a sequential haplotype scan procedure, which improves power of single-locus analyses. In our methods, a SNP will not be combined with a chosen set of SNPs unless it contributes a significant amount of information to disease association. This assures that the power gained from haplotype analyses is not compromised by an increase in the degrees of freedom. Methods based on both the haplotype test for sequentially chosen SNPs and the sum of a series of conditional χ^2 statistics are proposed. Furthermore, the sequential procedure we developed is computationally efficient, compared with some sliding window methods. We applied our methods to data simulated under different genetic models and to experimental data for CYP2D6. With either minor loss or substantial gain in power for most situations, and its computational efficiency, we recommend use of our procedure for case-control analyses, in addition to single-locus analyses.

METHODS

ALGORITHM

One general strategy to evaluate the association of haplotypes with disease is to use a forward selection procedure. Beginning with a single locus, we test the association of the alleles at this locus with disease status, using the usual Pearson χ^2 test. We then test the association of its nearest locus with disease, conditional on the first locus. This sequential process ''grows'' haplotypes while testing the contribution of an additional allele to a set of haplotypes, conditional on the current set of haplotypes. To achieve these conditional tests, we use the Mantel-Haenszel (MH) test [Mantel and Haenszel, 1959].

The MH procedure can be used to test if two binary variables are conditionally independent given a third variable. To simplify the presentation of the sequential haplotype scan algorithm, we first assume haplotypes are observed directly, although we drop this requirement later. Therefore, for a sample with N subjects, there are $2N$ haplotypes. Let Y denote the ''disease status'' of the $2N$ haplotypes, with values of 0 or 1 according to whether a haplotype came from a control or case, respectively. Let X denote the allele of a new SNP we wish to add to the current set of SNPs, with X having values of 0 or 1 for two different alleles. Let H be the number of distinguishable current haplotypes. Thus, the H unique haplotypes divide the $2N$ haplotypes into H strata. For the hth stratum, let n_{ijh} denote the number of haplotypes with $X = \mathrm{i}$, $Y = j$, as displayed in Table I. It is well known that, conditional on fixed row and column margins, the entry n_{11h}

TABLE I. Observed contingency table for stratum *h*

	Control (0)	Case (1)	
Allele			
0	n_{00h}	n_{01h}	n_{0+h}
1	n_{10h}	n_{11h}	n_{1+h}
	n_{+0h}	n_{+1h}	n_{++h}

Controls are labeled by "0" and cases are labeled by "1". The entries are numbers of haplotypes stratified by disease status and the alleles of a new SNP. More specifically, n_{i0h} and n_{i1h} are the numbers of haplotypes that have allele "i" at the new SNP in stratum h, among the controls and the cases, respectively. Entries in the fourth row and column are the marginal counts.

has a hypergeometric distribution with mean and variance

$$\mu_{11h} = E(n_{11h}) = \frac{n_{1+h}n_{+1h}}{n_{++h}},$$
$$\mathrm{Var}(n_{11h}) = \frac{n_{1+h}n_{0+h}n_{+1h}n_{+0h}}{n_{++h}^2(n_{++h}-1)}.$$

Testing whether a new SNP(X) contributes to disease status (Y) conditional on current haplotypes is equivalent to testing the null hypothesis that X and Y are not associated in any stratum. Under the null hypothesis, the Mantel-Haenszel statistic,

$$\mathrm{MH} = \frac{\left[\sum_h (n_{11h} - \mu_{11h})\right]^2}{\sum_h Var(n_{11h})},$$

has an asymptotic χ^2 distribution with one degree of freedom. By adopting the MH approach sequentially, we decide which markers should be added to a variable length haplotype.

When scanning SNP X_0, we examine SNPs close to it on both sides, to determine if they can help to detect a local haplotype frequency difference between the case and the control groups. The two alleles 0 and 1 of SNP X_0 separate the sample of alleles into two strata: those with allele 1 and those with allele 0. We first examine if at least one of the nearest SNPs on each side (left and right) of X_0 provides information for association, conditional on X_0. If not, we do not combine any SNP with X_0. Otherwise, we combine X_0 with the SNP(s) into a multi-locus haplotype variable and test if the second nearest marker on each side of X_0 should be added. The process is continued until no SNP should be combined. Let X_{L_i} and X_{R_i} denote the ith nearest SNP on the left and right sides of X_0, respectively. Let $\chi^2_{b|a}$ denote the MH statistic that tests whether X_b provides additional information for association, conditional on X_a. The algorithm is summarized below:

1) Start with SNP X_0 at position x. Let $\chi^2_0(x)$ be the traditional single-locus χ^2 statistic. Set $\mathrm{Sum}(x) = \chi^2_0(x)$. Assume X_{L_1} is closer to X_0 than X_{R_1}.

a. If $\chi^2_{L_1|0}(x) \le \lambda$, calculate $\chi^2_{R_1|0}(x)$. If $\chi^2_{R_1|0}(x) > \lambda$, update $\mathrm{Sum}(x)$ by $\mathrm{Sum}(x) = \mathrm{Sum}(x) + \chi^2_{R_1|0}(x)$, combine X_0 and X_{R_1} to form a haplotype variable X_{01} and go to step 2; else stop here.
b. If $\chi^2_{L_1|0}(x) > \lambda$, update $\mathrm{Sum}(x)$ by $\mathrm{Sum}(x) = \mathrm{Sum}(x) + \chi^2_{L_1|0}(x)$, combine X_0 and X_{L_1} into a two-locus haplotype variable $X_{(0,L_1)}$, and calculate $\chi^2_{R_1|(0,L_1)}(x)$. If $\chi^2_{R_1|(0,L_1)}(x) > \lambda$, update $\mathrm{Sum}(x)$ by $\mathrm{Sum}(x) = \mathrm{Sum}(x) + \chi^2_{R_1|(0,L_1)}(x)$ and denote the haplotype variable formed by X_0, X_{L_1} and X_{R_1} as X_{01}; else denote the haplotype variable formed by X_0 and X_{L_1} as X_{01}. Go to step 2.

If X_{R_1} is closer than X_{L_1} to X_0, we first evaluate X_{R_1}.

2) Stratify the $2N$ haplotypes according to the haplotype variable X_{01} constructed in step 1. The second nearest SNPs on both sides of X_0, namely X_{L_2} and X_{R_2}, are then examined for additional information for association, conditional on X_{01}, in the same way that is described in step 1.

3) Keep adding markers until neither X_{L_K} nor X_{R_K} should be combined based on the MH test. Two statistics are then calculated:

a. A $-\log_{10}$(P-value) for the haplotype-based χ^2 test for the contingency table of disease status and the haplotypes constructed from all makers in the sequentially chosen SNPs. Denote this statistic by $\chi^2_{\mathrm{H}}(x)$.
b. A $-\log_{10}$(P-value) for the sum of conditional χ^2 statistics, $\mathrm{Sum}(x)$. Because conditioning creates independence among the χ^2 statistics, the sum has an asymptotic χ^2 distribution with the degrees of freedom equal to the number of variables combined. Denote this statistic by $\chi^2_{\mathrm{S}}(x)$.

Thus, for the marker at physical position x, in addition to the traditional single-locus χ^2 statistic $\chi^2_0(x)$ that uses the starting marker only, we have two other statistics: (1) the $-\log_{10}$(P-value) based on haplotype test, $\chi^2_{\mathrm{H}}(x)$, and (2) the $-\log_{10}$ (P-value) based on the summary statistic, $\chi^2_{\mathrm{S}}(x)$. Note that the number of distinguishable haplotypes H varies based on different sets of markers.

Therefore, instead of using the χ^2 statistic as the test statistic, we used the $-\log_{10}(P\text{-value})$ to test the association for a $2 \times H$ table, constructed by disease status (for haplotypes) and H distinguishable haplotypes. The degrees of freedom of the summary statistic depend on both the number of SNPs combined and the LD level among them. When there is no LD among the SNPs combined, the degrees of freedom equal the number of SNPs combined. If some SNPs are in complete LD, the degrees of freedom will be smaller than the number of SNPs combined. For moderate LD, our simulation results indicate that the degrees of freedom can still be approximated by the number of SNPs combined, with a group of SNPs in complete LD being treated as a single SNP. Therefore, we first calculated the P-value for the summary statistic using the number of SNPs combined as the degrees of freedom and then calculated $-\log_{10}(P\text{-value})$. We also considered a sliding window method. In this method, we calculated all the asymptotic P-values for haplotypes constructed from 2 to 5 adjacent SNPs using sliding window methods. The largest $-\log_{10}(P\text{-value})$ among them was used as the test statistic. Denote the corresponding statistic by $\chi^2_{\text{Slide}}(x)$.

The threshold λ is used to decide when SNPs should be combined for haplotype analyses. A small value of λ causes SNPs to be combined even if they do not contribute a substantial amount of new information to disease association. On the other hand, a large value of λ prohibits SNPs from being combined into haplotypes even though they might contribute substantial information. Here, we reported results using both the 99th and 95th percentiles of the χ^2 distribution with one degree of freedom, which correspond to λ with values 6.63 and 3.84, respectively. Using these values, when examining whether a marker should be added to the current list, there are approximately 1% and 5% chances that a new marker will be added to the current list under the null hypothesis that the new marker does not contribute additional information to disease.

Because the sequential scan procedure uses the MH test multiple times, the multiple testing can inflate the Type I error rate. Another source of multiple testing occurs when a region under consideration contains many markers. To correct for multiple comparisons, we use permutation tests to evaluate both pointwise and regional P-values. We randomly permute disease status a large number of times to compute P-values for a test statistic at position x. The corresponding pointwise P-value is simply the fraction of times that the statistic for the permuted data is larger than that for the observed data. The regional P-values are the chances of observing the permutation statistics, maximized over a region, that are greater than those for the observed data. Notice that the sliding window method does not ''scan'' marker by marker, therefore, there is no definition of pointwise P-values.

To compare the relative power of different methods, we considered six methods: the single-locus method (χ^2_0), the sequential haplotype method with λ equal to the 99th and 95th percentiles of χ^2 distribution with one degree of freedom (χ^2_{H99} and χ^2_{H95}), the sequential summary method with λ equal to the 99th and 95th percentiles of χ^2 distribution with one degree of freedom (χ^2_{S99} and χ^2_{S95}), and the sliding window method (χ^2_{Slide}) and present results for each of them. When haplotypes are not observed directly, we first estimate the posterior probabilities of all possible pairs of haplotypes for each subject by applying the expectation maximization (EM) algorithm to the pool of all subjects. The estimated posteriors are then used in a weighted fashion. In this situation, n_{ijh} is an expected count based on the sum of posterior probabilities for all estimated haplotypes with $X = i$, $Y = j$, in the hth stratum.

SIMULATION METHODS

Simulations were used to examine both Type I error rates and power of the three statistics. A total of 2,000 haplotypes across a 200 kb region were simulated using the coalescent-based program ms [Hudson, 2002] with scaled mutation rate $4N_e\rho/bp = 6 \times 10^{-5}$ and recombination rate $4N_e r/bp = 4 \times 10^{-4}$. We kept 36 SNPs with minor allele frequencies no less than 0.05. The 2,000 haplotypes based on the 36 SNPs were treated as the genetic pool. When creating case-control data, pairs of haplotypes were drawn from the genetic pool with replacement until 1,000 cases and 1,000 controls were sampled. Let g be the coding of the genetic covariant and π be the probability for a subject with covariate g to be affected. We consider the logistic model, $\text{logit}(\pi/(1-\pi)) = \beta_0 + \beta_1 g$. In one-locus models and two-locus haplotype models, additive effects were assumed, i.e., g is the number of copies of a risk allele or haplotype. We also used the three models discussed by Marchini et al. [2005]. On the logit scale,

our additive model is equivalent to their multiplicative model on the penetrance scale. Therefore, we call the three models additive, interaction, and threshold models. The values of g based on phase-unknown two-locus genotypes are shown in Table II (a). Using methods in Schaid [2006], we chose parameters β_0 and β_1 such that the population prevalence would be approximately 0.1 and the statistical power for two-locus haplotype analyses would be around 80%. The corresponding odds ratios, $\exp(\beta_1)$, were less than 1.4 for two-locus additive and interaction models; between 1.3 and 2.0 for two-locus threshold and haplotype models. The odds ratios for all models are listed in Table II (b, c).

POWER AND TYPE I ERROR RATE FOR SIMULATED DATA

We compared power of the six methods (the single-locus method, two sequential haplotype methods, two sequential summary methods, and the sliding window method) using data simulated under different genetic models. For each genetic model, 1,000 simulations were used to evaluate the power. In each simulation, both pointwise and regional P-values were calculated by permuting the case/control status 1,000 times. The power for each statistic was defined as the frequency of observing a regional P-value less than 0.05 among the 1,000 simulations. To mimic what we expect in practice, when analyzing data simulated by one-locus models, we assumed that the SNP carrying the risk allele was not observed. When examining Type I error rates, we generated data under the null hypothesis, i.e., the haplotype pairs were drawn randomly from the haplotype pool, followed by 1,000 permutations. To study possible bias by estimating haplotypes using the EM algorithm when phase information is not available, we also estimated the Type I error rate assuming that haplotypes were not observed directly. Computation time for the six methods was also measured and compared.

TABLE II. Genetic models used to simulate data

(a) The coding of the covariate g for two-locus models

	Additive			Interaction			Threshold		
	00	01	11	00	01	11	00	01	11
00	0	1	2	0	0	0	0	0	0
01	1	2	3	0	1	2	0	1	1
11	2	3	4	0	2	4	0	1	1

(b) Odds ratios used in one-locus models

SNP ID	9	16	18	23	27	31	33	35
$e^{\beta 1}$	2.23	1.72	1.49	2.01	1.65	2.27	2.46	2.27

(c) Odds ratios used in two-locus models

SNP IDs	Additive	Interaction	Threshold	Haplotype
3,4	1.26	1.25	1.73	1.51
4,5	1.25	1.38	1.68	1.95
12,13	1.14	1.18	1.45	1.30
28,29	1.14	1.16	1.49	1.32

The first column and the second row of (a) show the alleles for the first and second markers, respectively.

APPLICATION TO CYP2D6 DATA

It has been reported that the enzyme CYP2D6 metabolizes about 20% of market drugs. The poor drug metabolizer (PM) phenotype exists at a frequency between 5% and 10% in Caucasians and is attributed to CYP2D6 polymorphisms. We applied our sequential scan methods to the data used by Hosking et al. [2002]. In a sample of 1,018 unrelated Caucasians, 41 individuals were labeled as PM cases based on typed functional polymorphisms of the CYP2D6 gene. After removing SNPs that were rare (minor allele frequency < 0.05) or with poor quality (deviation from Hardy-Weinberg equilibrium), we used 27 SNPs in our analysis. Because our sequential scan algorithm requires haplotype data, the phase information was estimated by the EM algorithm in the haplo.em package implemented in Splus [Schaid et al., 2002]. Hosking et al. [2002] reported that the SNP with the strongest association with the PM phenotype had a Fisher's exact P-value less than 10^{-34}, which is difficult to evaluate by permutations. Therefore, we did not perform the permutation test. We reported the P-values based on asymptotic χ^2 distributions. Both the 99th and 95th percentiles of the χ^2 distribution with one degree of freedom were used as thresholds in the sequential procedure.

RESULTS

TYPE I ERROR

Using data simulated under the null hypothesis, the Type I error rates using the single-marker (χ^2_0), sequential haplotype (χ^2_{H99} and χ^2_{H95}), sequential summary (χ^2_{S99} and χ^2_{S95}), and sliding window (χ^2_{Slide}) methods were: 0.053, 0.047, 0.052, 0.054,

0.048, and 0.051 based on 1,000 permutations, indicating that all tests have proper Type I error rates. When the phase information was unknown and the estimated posterior probabilities were used to compute expected counts, the Type I error rates of the six methods gave similar results (0.046, 0.050, 0.048, 0.052, 0.049, and 0.053). This suggests that using the estimated posterior probabilities does not bias the results.

COMPUTATION TIME

We evaluated the CPU time used for a data set generated by assuming SNP 16 carried the disease causal variant. Again, SNP 16 was eliminated in the analysis, which resulted in 35 flanking markers. The data contains 1,000 cases and 1,000 controls. We assumed that the haplotype phase was known. All methods were tested on a Sun workstation (Sun 420r) with four 450 MHz CPUs and 4 GB of memory. For 1,000 permutations, the number of minutes to complete χ^2_0, χ^2_{H99}, χ^2_{H95}, χ^2_{S99}, χ^2_{S95}, and χ^2_{Slide} were about 2, 4, 4, 4, 4, and 13 min, respectively. We also found that the computation time required for them is linear in both number of permutations and number of SNPs. Therefore, the computational time for 500 k SNPs for the six methods will be approximately 20, 40, 40, 40, 40, and 130 days. Because our sequential methods only combine SNPs together for association analysis when necessary, it is much more computationally efficient than the sliding window method.

ONE-LOCUS MODELS

Because disease causal variants were assumed unobserved when analyzing data generated by one-locus models, one SNP was considered to carry the causal variant and the remaining 35 SNPs were used to test statistical associations. To cover different scenarios, e.g., different LD structure surrounding the disease causal variant, each SNP with minor allele frequency between 0.05 and 0.1 was considered as causal variant. The minor allele frequency of each hypothesized disease causal SNP and its maximum r^2 (the square of the Pearson's correlation) with all the other SNPs are described in Table III.

As shown in Table III, when using SNP 9, 16, 27, 31, 33, or 35 as the disease causal variant, the power provided by our sequential methods and the sliding window method are much greater than that of the single-locus method. When a rare variant is unobserved but can be predicted by multiple measured markers close to it, it is usually beneficial to use a multi-locus approach. For example, when SNP 16 was selected as disease causal variant, the mutant allele 1 is frequently associated with the haplotype 011 of SNPs 13–15. Thus, even if the causal variant is not observed, it can be accurately predicted by haplotypes based on SNPs nearby. The frequencies of haplotypes based on SNPs 13–16 in both cases and controls

TABLE III. Simulated power for one-locus models

SNP ID	Allele freq.	Max r^2	χ^2_0	χ^2_{H99}	χ^2_{H95}	χ^2_{S99}	χ^2_{S95}	χ^2_{Slide}
9	0.056	0.17	0.66	0.66	0.80	0.69	0.71	0.92
16	0.062	0.43	0.62	0.73	0.76	0.68	0.66	0.79
18	0.082	0.61	0.65	0.66	0.65	0.63	0.61	0.64
23	0.057	0.20	0.57	0.57	0.57	0.59	0.60	0.60
27	0.085	0.35	0.61	0.83	0.85	0.82	0.84	0.87
31	0.069	0.09	0.63	0.73	0.84	0.68	0.74	0.93
33	0.064	0.09	0.64	0.90	0.95	0.90	0.92	0.97
35	0.081	0.09	0.67	0.81	0.90	0.77	0.91	0.96

The first column indicates which SNP carries the risk allele, the second column shows the corresponding frequency of the risk allele, and the third column gives the maximum value of the squared Pearson's correlation of the SNP in column 1 and all the other SNPs. Entries in the fourth to the ninth columns are the power based on the single-locus method (χ^2_0), the sequential haplotype methods with the 99th and 95th percentiles (χ^2_{H99} and χ^2_{H95}), the sequential summary methods with the 99th and 95th percentiles (χ^2_{S99} and χ^2_{S95}), and the sliding window method (χ^2_{Slide}).

TABLE IV. Frequencies of haplotypes

Haplotype	Frequency in cases	Frequency in controls
(a) Frequencies of haplotypes constructed from SNPs 13–16		
0000	0.584	0.609
0110	0.043	0.045
0111	0.096	0.057
1110	0.277	0.289
(b) Frequencies of haplotypes constructed from SNPs 17–19		
000	0.815	0.851
001	0.020	0.021
100	0.026	0.027
101	0.023	0.024
111	0.116	0.077
(c) Frequencies of haplotypes constructed from SNPs 6–9		
0000	0.350	0.371
0001	0.105	0.050
0010	0.094	0.100
0110	0.128	0.136
1000	0.323	0.343

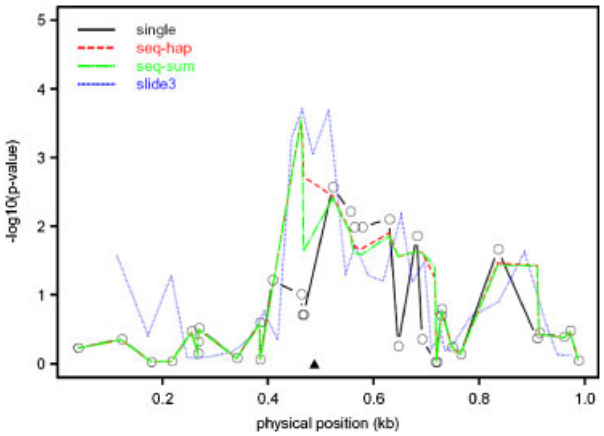

Fig. 1. Results for one simulation of the one-locus disease model. SNP16, whose position is indicated by the filled triangle, carries the risk allele. The pointwise *P*-values on the $-\log_{10}$ scale for χ^2_0 (solid line), χ^2_{H99} (dashed line), χ^2_{S99} (dot-dashed line) and χ^2_{Slide3} (dotted line). When carrying out the sliding window methods we used window lengths 2–5 and treated the most extreme statistic as our test statistic; however, it is difficult to define and represent the pointwise *P*-value of this sliding window method, so we presented the pointwise *P*-values when using window length 3.

are listed in Table IV (a). We randomly selected one simulation from the 1,000 simulations and plotted the pointwise *P*-values in Figure 1. Because our algorithm can automatically combine markers that jointly have a stronger signal than individual markers, both the sequential haplotype and summary methods can provide more significant pointwise and regional evidence for association than the single-locus method. However, when using SNP 18 or 23 as the disease predisposing variant, neither of our proposed methods nor the sliding window method increased power, compared with the single-locus method. To understand why our methods did not improve power in some situations, we examined the LD structure of our simulated genetic pool. For SNPs 9, 16, 27, 31, 33, and 35, their risk alleles were weakly associated with individual SNPs but strongly associated with several haplotypes of SNPs nearby. In contrast, for SNPs 18, and 23, their risk alleles were moderately or strongly associated with some individual SNPs but weakly associated with sets of SNPs around them. For example, as shown in Table IV (b), because both SNPs 17 and 19 are highly correlated with SNP 18, therefore, either of them is sufficient to predict the unobserved causal variant. When a risk allele can be well predicted by a single SNP, single-locus analyses can provide enough power and multi-locus analyses are not necessary.

For the four methods we proposed, for a chosen value of the threshold λ, the sequential haplotype method and the sequential summary method had similar power. Table III also shows that the power of using $\lambda = 3.84$, usually resulted in greater power than that of $\lambda = 6.63$. This is because that even if haplotype tests have stronger evidence for association, the MH statistic may not reach the 99th percentile of the χ^2 distribution with one degree of freedom. These four methods also provide comparable power with the sliding window method in most situations. However, if two non-adjacent SNPs jointly have a strong genetic effect, our proposed methods may fail to detect the combination. This is true when using SNP 9 or 31 as the disease causal variant. For example, Table IV (c) shows that neither SNPs 6–7 nor SNPs 7–8 can predict SNP 9 accurately. However, the haplotype 00 of SNPs 6 and 8 is frequently associated with allele 1 of SNP 9. In this situation, the sliding window method can reach greater power than our sequential methods.

TWO-LOCUS MODELS

We selected four adjacent SNP pairs as disease causal variants to study the power of the six methods in settings more complicated than a single causal variant. Because of the moderate recombination rate and the relatively high mutation rate used in our simulations, most of the pairs were in LD with various levels. The results are illustrated in Table V. When the correlation between the pair of causal SNPs in a two-locus model was not high, e.g., with statistical correlations (r^2) equal to 0.18 for pair 3 and 4, and 0.10 for pair 4 and 5, the four methods we proposed and the sliding window method had increased power. Sometimes the increase can be close to twofold. In contrast, when SNPs were moderately or highly correlated, e.g., with statistical correlations equal to 0.85 for pair 12 and 13, and 0.41 for pair 28 and 29, our methods still reached power similar to that of the single-locus method with negligible power loss; however, the sliding window method usually had decreased power by 0.05 and greater. When a pair of SNPs was highly correlated, the joint effect was not that striking compared to the main effects. Consequently, it is not beneficial to use multi-locus analyses. Table V also indicates that the four methods we proposed (χ^2_{H99}, χ^2_{H95}, χ^2_{S99}, and χ^2_{S95}) gave similar power in most situations.

Genet. Epidemiol. DOI 10.1002/gepi

TABLE V. Simulated power for two-locus models

			Additive						Interaction					
SNP IDs	Allele freq.	r^2	χ^2_0	χ^2_{H99}	χ^2_{H95}	χ^2_{S99}	χ^2_{S95}	χ^2_{Slide}	χ^2_0	χ^2_{H99}	χ^2_{H95}	χ^2_{S99}	χ^2_{S95}	χ^2_{Slide}
3	P1 = 0.4805													
4	P2 = 0.4835	0.18	0.52	0.76	0.76	0.81	0.81	0.77	0.42	0.71	0.72	0.68	0.69	0.73
4	P1 = 0.4835													
5	P2 = 0.234	0.10	0.63	0.75	0.74	0.72	0.74	0.74	0.54	0.68	0.69	0.64	0.68	0.70
12	P1 = 0.257													
13	P2 = 0.289	0.85	0.83	0.81	0.81	0.81	0.79	0.76	0.82	0.81	0.79	0.80	0.77	0.73
28	P1 = 0.3745													
29	P2 = 0.3665	0.41	0.79	0.77	0.77	0.78	0.77	0.71	0.81	0.78	0.77	0.81	0.80	0.76
			Threshold						Haplotype					
SNP IDs	Allele freq.	r^2	χ^2_0	χ^2_{H99}	χ^2_{H95}	χ^2_{S99}	χ^2_{S95}	χ^2_{Slide}	χ^2_0	χ^2_{H99}	χ^2_{H95}	χ^2_{S99}	χ^2_{S95}	χ^2_{Slide}
3	P1 = 0.4805													
4	P2 = 0.4835	0.18	0.43	0.67	0.67	0.73	0.75	0.68	0.47	0.78	0.84	0.66	0.66	0.84
4	P1 = 0.4835													
5	P2 = 0.234	0.10	0.64	0.72	0.72	0.70	0.72	0.70	0.53	0.81	0.87	0.74	0.78	0.84
12	P1 = 0.257													
13	P2 = 0.289	0.85	0.82	0.81	0.79	0.80	0.77	0.74	0.82	0.80	0.80	0.79	0.77	0.74
28	P1 = 0.3745													
29	P2 = 0.3665	0.41	0.80	0.78	0.77	0.81	0.80	0.75	0.83	0.83	0.80	0.83	0.81	0.79

The first three columns indicate the pairs of SNPs carrying the risk variants, the frequencies of risk alleles, and the squared Pearson's correlation between each pair. Entries in the other columns are power of association analyses using the single-locus method (χ^2_0), the sequential haplotype methods with the 99th and 95th percentiles (χ^2_{H99} and χ^2_{H95}), the sequential summary methods with the 99th and 95th percentiles (χ^2_{S99} and χ^2_{S95}), and the sliding window method (χ^2_{Slide}).

CYP2D6 DATA

We found that the results using the two thresholds ($\lambda = 6.63$ and 3.84) were very similar. Therefore, we only reported the results when using $\lambda = 6.63$. The sequential haplotype and sequential summary methods also gave similar results. To compare our results with those reported by Hosking et al. [2002], we reported the results using the sequential haplotype method. As displayed in Figure 2 and Table VI, the smallest asymptotic *P*-value (the largest $-\log_{10}$(*P*-value)) was reached by the combination of SNPs 14 and 15, with an asymptotic *P*-value of 9.3 E-49. As shown in Table VI and Figure 2, each of the SNPs 14 and 15 only provided relatively moderate evidence for association; however, the combination jointly provided much stronger evidence. It is interesting to see from Figure 2 that SNPs 14 and 15 are closer to the physical position of the gene that encodes the drug metabolizing enzyme (indicated by a filled triangle in Figure 2) than the most significant single locus (SNP 20). Besides the combination including SNPs 14–15, other combinations that provided very strong evidence for association can be found in Table VI.

Hosking et al. [2002] reported that, among all sliding windows with length five, SNPs 12–16 achieved highest significance. Based on data in their Table II, if we ignore rare haplotypes, it can be calculated that haplotypes from SNPs 12–16 provided a three-degree-of-freedom χ^2 statistic of value 221.1 (asymptotic *P*-value 1.1E-47). In contrast, our sequential haplotype scan approach rejected SNP 16 after combining SNP 15 with SNP 14, because SNP 16 did not provide sufficient additional information for the association. This resulted in a two-degree-of-freedom statistic with the same value (asymptotic *P*-value 9.3E-49). Therefore, the combination we identified, SNPs 14 and 15, provided more extreme statistical significance than SNPs 12–16. Another five-SNP set of SNPs reported in the Table II of Hosking et al. [2002] was SNPs 19–23. The asymptotic *P*-value was 6.7E-35. In contrast, our algorithm found several sets that had stronger evidence for association. These results show that by sequentially combing markers that jointly provide more information than individual SNPs about disease association, tests based on our sequential haplotype method is likely to achieve greater power than single-locus analyses.

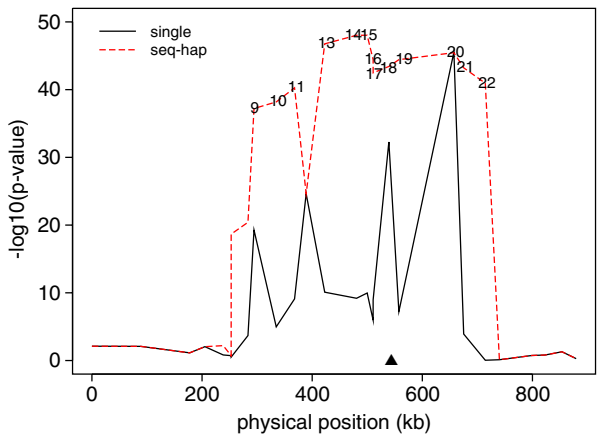

Fig. 2. The *P*-values on the $-\log_{10}$ scale of the two methods: single-locus (solid line) and sequential haplotype (dashed line) using CYP2D6 data. The filled triangle indicates the location of the *CYP2D6* gene, which is not mapped to any of the 27 SNPs.

TABLE VI. Results for the CYP2D6 data

	Single-locus method	Sequential haplotype method		
SNP ID	$-\log_{10}$ (*P*-value)	Start SNP	End SNP	$-\log_{10}$ (*P*-value)
1	2.11	1	1	2.11
2	2.09	2	2	2.09
3	1.13	3	3	1.13
4	2.05	4	4	2.05
5	0.82	4	5	2.20
6	0.72	6	6	0.72
7	0.51	7	9	18.66
8	3.70	8	9	20.41
9	19.27	8	12	36.59
10	4.96	9	13	38.19
11	9.10	10	13	40.27
12	24.64	12	12	24.64
13[a]	10.09	12	15	46.76
14	9.19	14	15	48.03
15	9.96	14	15	48.03
16	5.96	14	18	44.42
17[b]	8.96	14	19	42.78
18[c]	32.28	15	20	43.42
19	7.23	18	20	44.42
20	45.46	20	20	45.46
21	4.00	20	21	43.40
22	0.04	20	22	40.96
23	0.12	23	23	0.12
24	0.75	24	24	0.75
25	0.81	25	25	0.81
26	1.29	26	26	1.29
27	0.28	27	27	0.28

[a]SNP 14 was not selected when scanning SNP 13.
[b]SNPs 15 and 16 were not selected when scanning SNP 17.
[c]SNPs 16 and 17 were not selected when scanning SNP 18.

DISCUSSION

We provide a sequential haplotype scan procedure to combine nearby makers that can potentially increase the power to identify statistical association. This sequential haplotype scan procedure uses the MH statistic to examine if additional adjacent markers are informative for association, conditional on a current set of haplotypes constructed from sequentially chosen markers. The MH statistic has a large-sample χ^2 distribution with one degree of freedom and has been proven to be equivalent to the efficient score statistic of logistic regression [Day and Byar, 1979]. The advantage of using the MH approach is that we can compute the test statistic rapidly without iteratively estimating regression parameters. The MH test is also robust for tables with a large number of small strata and zero cell entries [Breslow and Day, 1980] and it is the uniformly most powerful unbiased test [Birch, 1964]. By adding markers sequentially based on the MH test, and scanning all makers, our sequential haplotype scan method can improve power under a variety of genetic models, even when disease predisposing loci are not genotyped. Its performance was also demonstrated by application to experimental data CYP2D6. Our simulations suggest that setting the threshold λ for the MH test to 3.84, which is approximately the 95th percentile of the χ^2 distribution with one degree of freedom achieves greater power than setting it to the 99th percentile.

We used the $-\log_{10}(P\text{-values})$ as test statistics in both the sequential haplotype and summary methods. For the sequential haplotype method, the degrees of freedom equals the number of different haplotypes minus one. For the sequential summary method, we used the number of SNPs combined as the degrees of freedom. When LD is not complete among SNPs, the summary statistic has an asymptotic χ^2 distribution with the degrees of freedom equal to the number of SNPs. Our simulations (results not shown) indicate the degrees of freedom can still be well approximated by the number of SNPs, even if two loci are in moderate LD, e.g., with r^2 ranges from 0 to 0.6. Based on our simulation results, this approximation still achieves proper Type I error rates and similar power with that of the sequential haplotype method.

Two neighboring markers that are highly correlated provide similar information. In this situation, the MH statistic conditional on the first marker will not be large, so the sequential scan procedure will not add the second marker given the first marker. But it is possible for the next nearest marker to provide critical information for disease association. In this situation, an informative marker can be blocked by the nearest marker that is in high LD with the marker being scanned. To avoid this, we combine markers that are highly correlated. If two markers have correlation r^2 greater than some large value, say, 0.95, we would force the sequential scan procedure to continue. A more preferable alternative to deal with highly correlated data is to pre-select ''tagging'' markers. Here, our purpose is not to reduce the number of markers, but rather to protect our sequential scan procedure from ceasing too early. Thus, we can ignore one of the two adjacent markers if their statistical correlation is very high.

Because our sequential haplotype scan method requires haplotype phase information, when haplotypes are not observed directly, we first estimated the posterior probabilities of all possible pairs of haplotypes that are consistent with the observed genotype data using the haplo.em package implemented in Splus [Schaid et al., 2002]. To make full use of the genotype information, similar to Schaid et al. [2002], Zaykin et al. [2002], and Stram et al. [2003], instead of using the most probable pair of haplotypes for each subject, we treat each estimated pair of haplotypes as a ''subject'' with fractional weight. In terms of efficiency of using data, this approach should be preferred over using only the most probable pair of haplotypes. Our proposed sequential procedure provides proper regional Type I error rates whether or not haplotypes are directly observed. When a region under study contains a larger number of markers, because it is impractical to infer haplotype phase information for all the markers using the traditional EM algorithm, one can divide a whole region into intervals with moderate number of markers, for example, 20 markers. Alternatively, one can also use the fastPHASE software [Scheet and Stephens, 2006] to infer haplotype phase.

A limitation of our sequential scan procedure is that when examining whether a marker should be added or not, the MH procedure is most powerful when the deviation from the null hypothesis is in the same direction across all strata. This ignores heterogeneous effects over strata. There are two reasons why we prefer not to test for homogeneity. First, tests for homogeneity usually have low power [Breslow and Day, 1980]. Second, this kind

of situation might be rare and difficult to interpret. Suppose that alleles of the first SNP divide the haplotypes into two strata: those with normal allele 0 and those with risk allele 1. When adding alleles of the second SNP, if allele 1 in stratum 0 has a negative effect while it has a positive effect in stratum 1, then both haplotypes 00 and haplotype 11 can be risk haplotypes. This kind of ''yin-yang'' effect [Zaykin et al., 2006; Culverhouse, 2004] might be rare in reality. Nonetheless, this type of epistasis can be evaluated by testing interaction terms in a loglinear model.

Another limitation of our procedure is that we can only detect association of haplotypes constructed from contiguous markers. This is why our methods have lower power than the sliding window method when using SNP 9 or 31 as the disease causal variant. It is possible that markers that are far apart can jointly have large effects. However, our procedure will fail in this situation. Indeed, detecting gene-gene interactions is a thorny issue because of the notorious computational burden and serious multiple-testing problems caused by the large number of tests. Our results show that the Mental-Haenszel test is computationally feasible. Further investigations are needed to consider gene–gene interactions among distant SNPs.

Finally, we emphasize that the sequential methods usually have greater power than single-locus method. Even compared with the computationally intensive sliding window method, our methods still achieve comparable power in most situations. However, our methods are much less computationally expensive and have more feasibility for genome-wide analysis than the sliding window method.

ACKNOWLEDGMENTS

This work was supported by the U.S. Public Health Service, National Institutes of Health, contract grant number GM065450. We are grateful to Louise K. Hosking, who provided the CYP2D6 data set.

REFERENCES

Akey J, Jin L, Xiong M. 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9: 291–300.

Bahlo M, Stankovich J, Speed TP, Rubio JP, Burfoot RK, Foote SJ. 2006. Detecting genome wide haplotype sharing using SNP or microsatellite haplotype data. Hum Genet 119:38–50.

Beckmann L, Thomas DC, Fischer C, Chang-Claude J. 2005. Haplotype sharing analysis using Mantel statistics. Hum Hered 59:67–78.

Birch MM. 1964. The detection of partial association. I: the 2-by-2 case. J R Stat Soc Ser B 27:111–124.

Breslow NE, Day NE. 1980. Statistical methods in cancer research. Vol. I: the analysis of case-control studies. International Agency for Research on Cancer. Lyon.

Cheng R, Ma JZ, Elston RC, Li MD. 2005. Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs. Ann Hum Genet 69: 102–112.

Clark AG. 2004. The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333.

Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415–428.

Culverhouse R, Klein T, Shannon W. 2004. Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27:141–152.

Day NE, Byar DP. 1979. Testing hypotheses in case-control studies—equivalence of Mantel-Haenszel statistics and Logit score testes. Biometrics 35:623–630.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151.

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 associated with CYP2D6 poor drug metabolising activity. Pharmacogenomics 2:165–175.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337–338.

Kaplan N, Morris R. 2001. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet Epidemiol 20:432–457.

Liu JS, Sabatti C, Teng J, Keats BJB, Risch N. 2001. Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res 11:1716–1724.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci influencing complex diseases. Nat Genet 37:413–417.

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J Nat Cancer Inst 22:719–748.

McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875.

Molitor J, Marjoram P, Thomas D. 2003. Fine-scale mapping of disease genes with multiple markers via spatial clustering techniques. Am J Hum Genet 73:1368–1384.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233.

Morris AP. 2005. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modeling of haplotypes. Genet Epidemiol 29:91–107.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425–434.

Schaid DJ. 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364.

Schaid DJ. 2006. Power and sample size for testing associations of haplotypes with complex traits. Ann Hum Genet 70:116–130.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190.

Toivonen HT, Onkamo P, Vasko K, Ollikainen V, Sevon P, Mannila H, Herr M, Kere J. 2000. Data mining applied to linkage disequilibrium mapping. Am J Hum Genet 67:133–145.

Tzeng Y-J. 2005. Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 28:220–231.

Waldron ERB, Whittaker JC, Balding DJ. 2006. Fine mapping of disease genes via haplotype clustering. Genet Epidemiol 30:170–179.

Yu K, Martin RB, Whittemore AS. 2004. Classifying disease chromosomes arising from multiple founders, with application to fine-scale haplotype mapping. Genet Epidemiol 27:173–181.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91.

Zaykin DV, Meng Z, Ehm MG. 2006. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 78:737–746.

Zhang K, Calabrese P, Nordborg M, Sun FZ. 2002. Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71:1386–1394.

Zollner S, Pritchard JK. 2005. Coalescent-based association mapping of complex trait loci. Genetics 169:1071–1092.