Nonparametric Disequilibrium Mapping of Functional Sites Using Haplotypes of Multiple Tightly Linked Single-Nucleotide Polymorphism Markers

Rong Cheng,*1 Jennie Z. Ma,*1† Fred A. Wright,1,2 Shili Lin,§ Xin Gao,1 Daolong Wang,1,2 Robert C. Elston** and Ming D. Li*3

*Department of Psychiatry and 1Center for Biostatistics and Epidemiology, The University of Texas Health Science Center, San Antonio, Texas 78229, 1Division of Human Cancer Genetics and 3Department of Statistics, Ohio State University, Columbus, Ohio 43210 and **Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44109

ABSTRACT

As the speed and efficiency of genotyping single-nucleotide polymorphisms (SNPs) increase, using the SNP map, it becomes possible to evaluate the extent to which a common haplotype contributes to the risk of disease. In this study we propose a new procedure for mapping functional sites or regions of a candidate gene of interest using multiple linked SNPs. Based on a case-parent trio family design, we use expectation-maximization (EM) algorithm-derived haplotype frequency estimates of multiple tightly linked SNPs from both unambiguous and ambiguous families to construct a contingency statistic for linkage disequilibrium (LD) analysis. In the procedure, a moving-window scan for functional SNP sites or regions can cover an unlimited number of loci except for the limitation of computer storage. Within a window, all possible widths of haplotypes are utilized to find the maximum statistic for each site (or locus). Furthermore, this method can be applied to regional or genome-wide scanning for determining linkage disequilibrium using SNPs. The sensitivity of the proposed procedure was examined on the simulated data set from the Genetic Analysis Workshop (GAW) 12. Compared with the conventional and generalized TDT methods, our procedure is more flexible and powerful.

Most human disorders of interest likely result from the cumulative effect of alleles at multiple susceptibility loci, none of which on its own is either necessary or sufficient to cause the disease. Because of this, the classical strategies of analyzing monogenic disorders have been unsuccessful. Alternative approaches such as genome-wide linkage and association analysis have been proposed and utilized in many studies (e.g., see Risch 2000). Large-scale association studies would employ a dense single-nucleotide polymorphism (SNP) map to detect association between markers and disease. In the causal hypothesis, most genetic determinants of a disease are SNPs that are likely to be selected as markers. In the proximity hypothesis, most disease determinants will not be included among markers but may be detected through linkage disequilibrium with other SNPs. Simulation studies based on monotonic population expansion suggest that useful association does not usually extend beyond 3 kb along the genome (Collins et al. 1999; Kruglyak 1999). Recent advances in the identification of SNP haplotype blocks in the human genome (Patil et al. 2001; Gabriel et al. 2002; Zhang et al. 2002a) showed that using selected representative SNPs by haplotype block partition algorithms can dramatically reduce the time and effort for genotyping without losing much haplotype information. Therefore, association studies using SNP markers represent an important tool in identifying susceptibility loci for human disorders because of their ability to identify unique chromosomal segments likely to harbor disease-predisposing genes.

Many haplotype analysis methods in the literature require phase information inferred from genotype data. However, as the number of loci increases, the information loss due to haplotype ambiguity would increase rapidly (Hodge et al. 1999). Several strategies that make use of the expectation-maximization (EM) algorithm (Dempster et al. 1977) have been proposed to overcome the problem of missing phase information for estimating haplotype frequencies from unphased diploid genotype data (Excoffier and Slatkin 1995, 1998; Hawley and Kidd 1995; Long et al. 1995; Slatkin and Excoffier 1996; Chiano and Clayton 1998; Clayton 1999). Recently, several research groups have discussed and compared the accuracy of haplotype frequency estimation using the EM algorithm (Fallin and Schork 2000; Tishkoff et al. 2000; Stephens et al. 2001; Niu et al. 2002; Qin et al. 2002). It was suggested that even in the worst case, individual haplotype frequency estimates obtained by the EM algorithm from a sample size of 100

1Present address: Columbia Genome Center, Columbia University, 1150 S. Nicholas Ave., New York, NY 10032.
2Present address: Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599.
3Corresponding author: Department of Psychiatry, Mail Code 7792, 7703 Floyd Curl Dr., San Antonio, TX 78229.
E-mail: lim2@uthscsa.edu
or more would not deviate >5% from their true values among the sampled individuals (Fallin and Schork 2000). Furthermore, Fallin and Schork (2000) suggested that one should be more concerned about the quality of sampling than the possibility of estimation errors when assessing haplotypes among unphased individuals. Therefore, EM estimation of haplotype frequencies for multiple diallelic genotypes may represent a viable alternative to the recruitment of additional family members or intensive laboratory haplotyping for haplotype-based genetic studies.

Haplotype frequency estimates from tightly linked multilocus genotyping data have been used for linkage disequilibrium (LD) analysis (Clayton 1999; Toivonen et al. 2000; Zhao et al. 2000; Fallin et al. 2001; Schaids et al. 2002). For example, Zhao et al. (2000) proposed a new statistical method for the transmission/disequilibrium test (TDT) using EM-derived haplotype frequency estimates of multiple tightly linked markers from both unambiguous and ambiguous families. Their results suggest that the power of the method was higher than that of the conventional TDT method (Spielman et al. 1993) or other TDT methods, such as using estimated haplotype frequencies of multiple markers from only unambiguous families (Wilson 1997; Clayton and Jones 1999). Clayton (1999) also proposed a new TDT statistic for the situation in which some haplotype phases are unknown or some parental genotypes are missing. Dudbridge et al. (2000) showed that, in some families for which haplotypes are known, a potentially serious bias is introduced into the TDT if the loss of information from families with ambiguous haplotypes is not taken into consideration. The methods proposed by Clayton (1999) and Zhao et al. (2000) can avoid loss of information from families with ambiguous haplotypes, but both require one to preassign a window width prior to analyzing multisite parental transmission data under this given fixed width.

Sachidanandam et al. (2001) pointed out that in practice, when a gene has been implicated in causing the disease of interest as a result of linkage analysis, known biological function(s), or an expression pattern, it is desirable to survey allelic variation within it exhaustively for any potential association to the disease. Using SNP markers, it becomes possible to evaluate the extent to which common haplotypes contribute to disease risk. In this study, we apply EM-derived haplotype frequency estimates of multiple tightly linked SNP markers from both unambiguous and ambiguous families to construct a contingency table statistic, namely $S(a, b)$, for LD analysis. Our procedure assumes no or very rare recombination among multiple tightly linked markers. In our procedure, a moving-window scan for functional sites or regions can cover any number of loci without limitation except for that imposed by computer storage. Within a given maximum window width, all possible widths of haplotypes are utilized to find the maximum statistic $S^*$ for each site (or locus). This represents a significant difference between the procedure reported here and the published methods in the literature (Clayton 1999; Zhao et al. 2000). We apply the proposed method to scan the simulated sequence data from the Genetic Analysis Workshop (GAW) 12 (Almasy et al. 2001) to identify the functional sites of candidate genes with tightly linked SNP markers. Comparing the results from our procedure with those from the conventional and generalized TDT methods, it appears that our method is more flexible and powerful.

**METHODS**

We describe a nonparametric method of detecting functional SNP site(s) of haplotypes associated with the disease of interest in a LD study. The method naturally incorporates tightly linked multilocus genotype data, in the sense that multiallelic loci are utilized and the existing disequilibrium among markers (in both haplotypes that are transmitted and those that are not transmitted to affected offspring) is built into the test. The approach searches for evidence of ancestral haplotypes that are shared more often among chromosomes transmitted than among chromosomes not transmitted to affected offspring, followed by comparison of the observed data to the distribution of data expected under the null hypothesis.

**EM algorithm estimation:** When a haplotype segment with certain marker(s) occurs in a chromosome transmitted to affected offspring with a frequency higher than that in the chromosome not transmitted, there exists an association of the haplotype with the disease. This finding would be considered striking and consistent with the hypothesis that the marker positions are near a disease gene or located at a disease locus (e.g., a functional site lies within the sequence of a disease gene), explaining the haplotype as being identical by descent (IBD) with that from a common founding ancestor. We consider $N$ randomly sampled trios, each containing an affected offspring along with two parents, with each member genotyped for multiple tightly linked markers. In some trios, it may be impossible to unambiguously reconstruct haplotypes from the given genotypes. In such a design with data from an affected singleton offspring and both parents, if the haplotypes can be reconstructed from the known genotypes unambiguously, we let $H$ be the total number of possible haplotypes and $n_j$ and $n'_j$ be numbers of a particular haplotype $h_j (j = 1, 2, \ldots, H)$ transmitted and not transmitted to the affected offspring, respectively (see Table 1). If the haplotypes cannot be reconstructed unambiguously, then, to fully use the information of the multiple tightly linked markers, the EM algorithm is employed to estimate the transmitted and not transmitted haplotype frequencies $f_j$ and $f'_j (j = 1, 2, \ldots, H)$. In Table 1, we use $2N \times f_j$ and $2N \times f'_j$ to represent directly counted numbers $n_j$ and $n'_j$, respectively, even when some of the actual numbers are ambiguous.
Suppose \( G_i (= g_o, g_m, g_f) \) is the trio for the \( i \)th family, where \( g_o, g_m, \) and \( g_f \) represent the genotypes of the child, the mother, and the father, respectively. Let \( h_o \) be paternal haplotype transmitted to the affected offspring \( h_m \) be paternal haplotype not transmitted to the affected offspring \( h_f \) be maternal haplotype transmitted to the affected offspring \( h_m \) be maternal haplotype not transmitted to the affected offspring.

Conditional on \( G_i \) the probability or weight to a transmitted haplotype \( h_i \) from the \( i \)th family is

\[
\begin{align*}
  w_i^{(0)} &= \frac{\sum_{(h_o, h_m, h_f) \in G_i} \epsilon_{\tilde{h}_i}(f, f', f''_n)}{\sum_{(h_o, h_m, h_f) \in G_i} f, f', f''_n},
\end{align*}
\]

where \((h_o, h_m, h_f) \in G_i\) denotes the haplotype group \((h_o, h_m, h_f)\) that is compatible with the genotype group of the \( i \)th family and the factor \( \epsilon_{\tilde{h}_i} (=0, 1, \text{or } 2) \) depends on the counts of haplotype frequency \( f_i \) occurring in the pair of haplotype frequencies \((f, f')\). The summation in the denominator is over all haplotype groups that are compatible with \( G_i \). The \( w_i^{(0)} \) are defined as the estimated weight of the \( i \)th family when the parent(s) has haplotype \((h_o, h_m, h_f)\) and/or \((h_o, h_m, h_f)\), and \( h_i \) is transmitted. Similarly,

\[
\begin{align*}
  w_i^{(0)} &= \frac{\sum_{(h_o, h_m, h_f) \in G_i} \epsilon_{\tilde{h}_i}(f, f', f')}{\sum_{(h_o, h_m, h_f) \in G_i} f, f', f'},
\end{align*}
\]

is the weight for a not-transmitted haplotype \( h_i \).

Given the genotypes, the likelihood function for the data is

\[
L(\text{data}|G_1, G_2, \ldots, G_N) \propto \prod_{i=1}^{N} P_i(g_o, g_m, g_f),
\]

where

\[
 P_i(g_o, g_m, g_f) = \sum_{(h_o, h_m, h_f) \in G_i} \epsilon_{\tilde{h}_i}(f, f', f')
\]

So the log-likelihood is

\[
\log L(\text{data}|G_1, G_2, \ldots, G_N) = \sum_{i=1}^{N} \log |P_i(g_o, g_m, g_f)| + \text{constant},
\]

Next we consider estimation of the haplotype frequencies \( f_i \) and \( f'_i \) \((j = 1, 2, \ldots, H)\). The EM algorithm (see, e.g., Excoffier and Slatkin 1995, 1998; Hawley and Kidd 1995; Long et al. 1995; Slatkin and Excoffier 1996; Chiano and Clayton 1998; Clayton 1999) is a way of attempting to find the \( f_i \) and \( f'_i \) \((j = 1, 2, \ldots, H)\) that maximize the log-likelihood \( \log L(\text{data}|G_1, G_2, \ldots, G_N) \) in a single pass through the \( H \) possible haplotypes as indicated below:

\[
\begin{align*}
  w_i^{(1)} &= \sum_{(h_o, h_m, h_f) \in G_i} \epsilon_{\tilde{h}_i}(f, f', f''_n),
  w_i^{(2)} &= \sum_{(h_o, h_m, h_f) \in G_i} f, f', f''_n,
  w_i^{(3)} &= \sum_{(h_o, h_m, h_f) \in G_i} f, f', f''_n,
\end{align*}
\]

In principle, the maximum-likelihood (ML) estimates of the haplotype frequencies could be found analytically by solving a set of equations with the Lagrange multipliers \( \lambda_i \) and \( \lambda_j \):

\[
\frac{\partial}{\partial \log (f, f')_j} \left( \sum_{i=1}^{N} \sum_{(h_o, h_m, h_f) \in G_i} f, f', f''_n \right)
+ \lambda_j (1 - \sum_{j=1}^{H} f_j) + \lambda_2 (1 - \sum_{j=1}^{H} f'_j) = 0,
\]

\[
  j = 1, 2, \ldots, H,
\]

and the additional partial derivatives with respect to \( \lambda_i \).
and λk. We obtain the ML estimators of \( f_i \) and \( f^*_j \) (\( j = 1, 2, \ldots, H \)) in the \((p + 1)\)st EM iteration,

\[
\hat{f}^{(p+1)}_j = \frac{1}{2N} \sum_{i=1}^{N} w^*_i^{(p+1)} \quad \text{and} \quad \hat{f}^{(p+1)}_j = \frac{1}{2N} \sum_{i=1}^{N} w^*_i^{(p+1)}.
\]

Initially, all haplotypes are set equally frequent, so that all possible complementary haplotype pairs are equally likely. In the \((p + 1)\)st expectation step (E-step), the weights of the transmitted and not-transmitted haplotypes, \( w^*(j) \) and \( w^*_i^{(p+1)} \), can be obtained from the given genotype information and current estimates of the haplotype frequencies, \( \hat{f}^{(p)}_j \) and \( \hat{f}^{(p)}_j \). The \((p + 1)\)st maximization step (M-step) gives the maximum-likelihood estimates of \( \hat{f}^{(p+1)}_j \) and \( \hat{f}^{(p+1)}_j \).

When the difference between the values of the haplotype estimates in the previous M-step and current one becomes less than a predetermined quantity (e.g., \( 10^{-7} \)), the iteration is stopped and the final estimates are obtained. The estimated numbers \( \hat{n}_i = \sum_{j=1}^{H} \hat{f}^{(p+1)}_j i \) and \( \hat{n}_i = \sum_{j=1}^{H} \hat{f}^{(p+1)}_j i \) of transmitted and not-transmitted haplotypes are then used for further computing the statistic \( S \) and estimating the empirical 100(1 − \( \alpha \)) percentiles and \( P \) value by simulation permutation (Churchill and Doerge 1994).

**Statistic S**: The key feature of this method is encompassed by defining a statistic \( S(a, b) \) (Gao and Wright 1999) that is computed on the set of haplotypes beginning at marker position \( a \) and ending at position \( b \). The association evidence at position \( x \) is

\[
S^*(x) = \max_{a | b, a \leq x} S(a, b).
\]

The statistic \( S \) is chosen to reflect a striking association of all haplotypes with the disease. In other words, for an arbitrary statistic \( S \) and at position \( x \), we search over all haplotype widths (containing \( x \)) to find the haplotype in most striking association with the disease. Under the null hypothesis, this haplotype has no association with the disease. Under the assumption that haplotypes are sampled independently, we compute a \( P \) value for the \( \chi^2 \) contingency table test of haplotype vs. disease status and refer to the statistic \( S(a, b) = -\log_{10} \) (\( P \) value) as \( \hat{S} \). The \( P \) value corresponds to the \( k \times 2 \) standard \( \chi^2 \) contingency table test of \( k (1 < k \leq H) \) unique haplotypes (with \( k - 1 \) d.f.) that begin at position \( a \) and end at \( b \) vs. disease status, under the assumption of independence. The number \( k \) of unique haplotypes will depend on the choice of \( a \) and \( b \).

**Permutation for empirical \( P \) value**: To evaluate type I error accurately, an empirical \( P \) value that appropriately corrects for the testing of multiple marker locations is obtained from the EM estimates of haplotype numbers and frequencies (\( \hat{n}_i, \hat{n}^*_i, \hat{f}_j, \) and \( \hat{f}^*_j; \) \( j = 1, 2, \ldots, H \)), by permuting the transmitted status of individual haplotypes, and computing \( S^*(x) \) for each permutation (Zhao et al. 2000). Briefly, we first generate haplotypes conditional on each family’s genotypes. If multiple possible compatible haplotype groups are conditional on a given family’s genotypes, first a compatible haplotype group is randomly selected. For each family, haplotypes are then randomly permuted within each transmitted/nontransmitted haplotype pair for each parent to form a new set of family genotypes. The transmitted and nontransmitted haplotype frequencies are then reestimated using the EM algorithm for each random selection and permutation of the data. By using this procedure, we obtain the null distribution of the maximum of \( S^* \) from 10,000 randomized permutations. Then it is easy to obtain the 100(1 − \( \alpha \)) percentile and appropriate \( P \) value for the maximum of \( S^* \) over the region with multiple tightly linked markers. A computer program for the proposed method is available upon request.

**RESULTS**

We applied the proposed procedure to the simulated sequence data set of GAW 12. This data set contains computer-simulated sequence data with multiple SNP markers for seven candidate genes in 23 extended pedigrees with a total of 1497 individuals for two populations: a general and an isolated population. There were 50 replicates for each population model. We randomly sampled 10 replicates from the 50 replicates of the general population and separated them into two groups each with five replicates. Group 1 contained replicates 1, 10, 18, 42, and 48, while group 2 contained replicates 21, 23, 33, 34, and 38. For each extended pedigree in a replicate, we randomly sampled a trio with one affected offspring and two parents (affected or unaffected), for whom genotyping information was available. In total, 95 and 97 trios were in groups 1 and 2, respectively. Last, we pooled the two samples together to form sample 3, with a total of 192 trios. All the SNP variants present in a sample were counted, and only those with a frequency >1% were chosen as possible SNP sites (Broder and Venter 2000).

Theoretically, the EM algorithm can be applied to an unlimited number of loci with any number of alleles. However, in practice, implementation of this algorithm is limited by the need to store the estimated haplotype frequencies for every possible haplotype contained in the sample. These storage requirements increase exponentially with the number of loci under investigation. For example, the numbers of SNPs for candidate gene 1 are 155, 158, and 152 in samples 1, 2, and 3, respectively. If any individual is heterozygous at 150 loci, then the number of possible haplotypes in that sample is \( 2^{150} \). Also, as the number of markers increases, there will be an increased variance in the estimates. In this study, we first selected SNP markers at \( \sim 1000 \)-bp intervals and set the window width at 5 (i.e., \( m = 5 \); five tightly linked markers) to scan every selected SNP marker for all candidate gene sequences. The procedure of window scanning implemented in the program can be described as
TABLE 2

Chromosomal locations and sequence lengths of seven candidate genes in the simulation model

<table>
<thead>
<tr>
<th>Candidate gene</th>
<th>Chromosome</th>
<th>Location (cM)</th>
<th>Length (bp)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (MG6)*</td>
<td>6</td>
<td>30.50</td>
<td>20,000</td>
<td>150*</td>
</tr>
<tr>
<td>2 (MG5)</td>
<td>1</td>
<td>137.10</td>
<td>13,000</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>111.00</td>
<td>16,000</td>
<td>113</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>138.30</td>
<td>20,000</td>
<td>104</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>9.80</td>
<td>17,000</td>
<td>35</td>
</tr>
<tr>
<td>6 (MG1)</td>
<td>19</td>
<td>42.10</td>
<td>17,000</td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>47.90</td>
<td>20,000</td>
<td>185</td>
</tr>
</tbody>
</table>

*a Chromosome number.
*b The location in centimorgans from p-ter of the chromosome.
*c The length of the sequence for a candidate gene in base pairs.
*d MG indicates major gene in the original simulation model.
*e Number of SNPs for each gene (1% less than one allele frequency at a locus <99%).

Linkage disequilibrium analysis was performed for all selected SNPs of these seven candidate genes. The distributions of the statistic $S^*$ for these candidate genes are shown in Figure 1. The highest peak locations of $S^*$ and their corresponding $P$ values for all candidate genes are summarized in Table 3. As described in methods, for an arbitrary statistic $S$ and at the position $x$, we searched over all haplotype widths. If we set $m = 5$ and $x = 5$, the varied pairs of $(a, b)$ containing the $x$ will be $(1, 5)$, $(2, 6)$, $(3, 7)$, $(4, 8)$, $(5, 9)$, $(2, 5)$, $(3, 6)$, $(4, 7)$, $(5, 8)$, $(3, 5)$, $(4, 6)$, $(5, 7)$, $(4, 5)$, $(5, 6)$, and $(5, 5)$, and the statistic $S$ calculated from the last pair of $(5, 5)$ is equivalent to a conventional single-marker test. Some of the peak locations of $S^*$ in Table 3 contained two linked markers with an equal maximum $S^*$, while others might contain three, four, or five linked markers. One functional SNP was identified for candidate genes 1 and 6. No functional SNP site was detected at the 0.01 significance level for the other five candidate genes (Table 3). In fact, for candidate gene 2 multiple simulated functional sites directly affected a quantitative trait, but no single functional site directly affected disease status. For the other four candidate genes, no functional sites were simulated in the original model (Almasy et al. 2001). For candidate gene 1, the functional site at the nucleotide 5782, was detected in both samples 1 and 3. For candidate gene 1 in sample 1 and candidate gene 6 in sample 2, the statistic $S^*$ peaks were detected at the same locations as in the other samples, except that the corresponding $P$ values did not reach the 0.01 level. It should be noted that, after pooling samples 1 and 2 to form sample 3, the statistic $S^*$ values at the peak locations increased as the sample size became larger for both candidate genes 1 and 6, but did not increase, in fact even decreased, for the other five candidate genes at some peak locations (Figure 1).

Next, we compared our procedure to the generalized TDT (Clayton 1999) and the conventional TDT (Spielman et al. 1993) methods for candidate genes 1 and 6 in samples 1–3. For the conventional TDT method, each SNP site was studied separately, leading to a $P$ value for each SNP site. For the generalized TDT method implemented in the TRANSMIT program (Clayton 1999), we fixed the window widths at two, three, four, or five linked SNPs followed by choosing the smallest $P$ value for each SNP site from the multiple-locus and single-locus analysis results under different windows. As shown in Figure 2, for candidate gene 1, the highest peaks of the statistic $S^*$ from our procedure are greater than (or at least equal to) those obtained from the generalized or conventional TDT methods for all samples. For candidate gene 6, the largest statistics obtained from our procedure are all greater than those from the TRANSMIT or TDT procedures. The highest peaks of statistic $S^*$ from the TRANSMIT program are greater than or equal to those from the conventional TDT method.

Figure 3 compares the results obtained from our procedure with the generalized and conventional TDT methods for candidate gene 1 in sample 3 ($N = 192$ trios) under different SNP densities. A total of 41, 23, 8, and 5 SNPs were selected from the original SNPs for candidate gene 1, which gave an average SNP interval of 500, 1000, 3000, and 5000 bp, respectively. The
simulated functional site (i.e., 557 bp) was included for only the first two high densities. As shown in Figure 3, the highest peaks of statistic $S^*$ detected by these methods are all at or around the 557-bp functional site. Compared to what is found with high densities, the highest peaks of statistic $S^*$’s detected at low densities are smaller. Similarly, we compared the results from these three methods for candidate gene 6 under three different SNP densities. As shown in Figure 4, the functional site is detected by all three methods for candidate gene 6 under the SNP density of 1000- and 2000-bp intervals, with the highest peak at functional site 5782 bp for
TABLE 3

The peak locations of the statistic $S^*$ and $P$ values (each obtained empirically from 10,000 permutations) for seven candidate genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>SNP location (bp)</th>
<th>Statistic $S^*$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>557</td>
<td>2.608</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>557</td>
<td>5.495</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>557</td>
<td>7.209</td>
<td>0.0004</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5961</td>
<td>0.943</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8657, 9034, 10106, 11030, 12311</td>
<td>1.531</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11030</td>
<td>1.065</td>
<td>0.340</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10595</td>
<td>2.796</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3072</td>
<td>1.086</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3072</td>
<td>1.482</td>
<td>0.169</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>12721, 13666, 14790</td>
<td>1.761</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11910</td>
<td>0.640</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>72, 1192</td>
<td>1.042</td>
<td>0.185</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>12</td>
<td>0.726</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>834, 1987</td>
<td>3.075</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>834</td>
<td>2.385</td>
<td>0.014</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5782, 7332</td>
<td>4.099</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5782, 7332, 8226, 9954</td>
<td>1.918</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5782, 7332</td>
<td>5.142</td>
<td>0.0001</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9682, 10863, 12132</td>
<td>1.593</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12, 987, 2364, 3541, 4508</td>
<td>1.618</td>
<td>0.115</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5508, 6472</td>
<td>1.398</td>
<td>0.260</td>
</tr>
</tbody>
</table>

1000-bp intervals and at SNP site 7332 bp (which is 1550 bp away from the functional 5782-bp site) for 2000-bp intervals. Maximum statistics $S^*$ obtained from our procedure are always greater than those obtained from the generalized or conventional TDT (see Figure 4, A and B). When 5 SNPs at ~4000-bp intervals were used, no linkage disequilibrium was detected by any method, but our procedure and the generalized TDT method still produced smaller $P$ values than those from the conventional TDT (Figure 4C).

DISCUSSION

To date, ~2.1 million SNPs have been identified in the human genome (Venter et al. 2001). This rich resource should allow the initiation of genome-wide linkage disequilibrium mapping of disease genes in the human population (Lander et al. 2001). In the past, many LD analysis methods have been proposed (e.g., Falk and Rubinstein 1987; Ott 1989; Thomson et al. 1989; Terwilliger and Ott 1992; Spielman et al. 1993; Sham and Curtiss 1995; Thomson 1995; Schaid 1996; Spielman and Ewens 1996, 1998; Martin et al. 1997, 2000; Sham 1997; Xiong and Guo 1997; Lazzeroni and Lange 1998; Cordell and Elston 1999; Sun et al. 1999; Teng and Risch 1999; Zheng and Elston 1999; Zhu and Elston 2000, 2001) and proved to be powerful approaches to identify susceptibility genes for the disease of interest. However, to apply most, if not all, of these methods, one needs to either discard families with ambiguous haplotypes or analyze the markers separately, resulting in potential loss of power.

In this study, we propose a new procedure for nonparametric disequilibrium mapping at a functional site (or region) within a candidate gene by using multiple tightly linked SNPs. For an arbitrary statistic $S$ at position $x$, we search over all haplotype widths (containing $x$) and find some of the peak locations of the maximum statistic $S^*$ contain two linked markers while others might contain three, four, or five linked markers (see Table 3). With our procedure, researchers do not need to know prior to analysis what window width of haplotype transmission data under this fixed width. Therefore, our procedure is more flexible, and it can find the maximum statistic $S^*$ with an appropriate window width of haplotype transmission data within a preset range ($m$) spanning a sin-
Figure 2—Comparison of linkage disequilibrium analyses among our procedure (nonparametric disequilibrium, ND), the generalized TDT (using the program TRANSMIT; CLAYTON 1999), and the conventional TDT method, for candidate genes 1 and 6 on samples 1, 2, and 3. The lines with open circles, triangles, and squares indicate the results from ND, TRANSMIT, and TDT, respectively. The average SNP interval for all candidate genes is ~1000 bp.

gle, two, or more multiple linked markers (≤m). Although it is generally thought that a multiple-locus approach may be more powerful than single-locus analysis, this does not mean that this is true in all cases. For example, for candidate gene 1 in the simulated GAW 12 data set, it seems that the single-marker approach is better than that of multiple markers. In contrast, for candidate gene 6, we found that using the haplotype transmission data of two linked markers appears to be better than using those of a single marker, or three, four, or five linked markers.

The simulation model implemented in the GAW 12 data set was rather complex (ALMASY et al. 2001). Seven major genes (MG) were assumed to influence one or more of five quantitative traits (Q1–Q5), disease liability, and age at onset. The disease liability was constructed as a function of Q1–Q5, MG6, and a mitochondrial component. MG6 provides a baseline liability value, which is adjusted for Q1–Q5. MG6 and the mitochondrial component have direct effects on liability and account for 22 and 10% of its variance, respectively, conditional on other covariates like gender, age, and environmental factors. MG1–MG5 affect liability only through the quantitative traits. MG1 accounts for 4.6%
of the variance in liability, MG2 for 4.4%, MG3 for 8%, MG4 for 0.3%, and MG5 for 9%. Our procedure could detect the functional site for candidate gene 1 that directly affects liability and the functional site for candidate gene 6, where only one functional site directly contributes to Q1 and Q2 and indirectly affects the affection status. But we could not detect the multiple functional sites for candidate gene 2, a quantitative trait gene with multiple functional sites. Therefore, for quantitative traits controlled by multiple genes with relatively small effects or multiple functional sites each with a small genetic effect, such as candidate gene 2, other methods of using haplotype information should be considered. Several methods focusing on this type of data have been proposed (Clayton 1999; Li et al. 2001; Schaid et al. 2002).

Using the procedure reported in this study, linkage disequilibrium due to the two simulated functional sites of candidate genes 1 and 6 (defined as MG6 and MG1, respectively, in the original model) were detected. Multiple alleles were in candidate gene 2; all changes in regulatory elements or in the first or second base pair of a codon leading to amino acid substitutions were functional. However, this was not the case for candidate gene 6. Furthermore, candidate gene 2 directly contributes to Q5 from multiple functional sites and then indirectly affects the affection status.

On the basis of the mapping results from our method and the generalized and conventional TDT methods on the GAW 12 simulated data set, it appears that our procedure is as sensitive and powerful as the TDT methods for candidate gene 1, but is more powerful than the TDT methods for candidate gene 6. Additionally, we investigated how SNP density may affect the performance of our procedure. A functional SNP site for candidate gene 1 was detected by all three methods, regardless of SNP density and whether the functional SNP site was included or not. In contrast, a functional SNP site for candidate gene 6 was detected only at high SNP densities, not at low density. To explain this difference, we examined selected SNPs used under each density for both genes and found that a SNP at the 189-bp position of gene 1, which is only 368 bp away from the functional SNP site, was used for all analyses. However, for candidate gene 6 at the 4000-bp intervals, the closest SNP sites used in the analysis were at the 4848- and 9952-bp positions, which are either loosely linked or too far away from the functional site (5782 bp). Regardless of which SNP density was used, overall, we found that our procedure and the generalized TDT method are less dependent on the SNP density than the conven-

Figure 3.—Comparison of linkage disequilibrium analyses of our procedure with the generalized TDT (TRANSMIT; Clayton 1999) and the conventional TDT for candidate gene 1 under SNP densities of ~500-, 1000-, 3000-, and 5000-bp intervals. The lines with open circles, triangles, and squares represent the results from the ND procedure, the generalized TDT, and the conventional TDT method, respectively.
tional TDT method is. Moreover, our program is more user friendly and less time consuming than the TRANSMIT program with respect to file preparation before and after mapping analysis. For example, for candidate gene 1 under a density of 500-bp intervals, only 1 file was needed for our procedure, while for TRANSMIT we had to prepare 40, 39, 38, and 37 files for window widths of two, three, four, and five linked SNPs, respectively. Although these files can be prepared using a shell script for each selected window width, we still have to summarize a significant number of output files for comparing the results.

Theoretically, our search procedure can cover as many loci as one wishes without limitation, except for the computer storage requirement. As needed, our method can be used for regional or genome-wide scanning to study LD using SNP markers. This represents another advantage of our program compared to other currently available procedures. It should be pointed out that if genome-wide high-density SNP data were available, then the size of the data file would become significantly large. The newly proposed haplotype block partition algorithms (Patil et al. 2001; Zhang et al. 2002a) can be employed to find appropriate blocks and select representative SNPs. To overcome the difficulty of extreme time cost and the limitation of computer storage in estimating large haplotypes of SNPs, Niu et al. (2002) and Qin et al. (2002) proposed a partition-ligation (PL) strategy together with Gibbs sampling and an EM algorithm. Their PL-EM algorithm (Qin et al. 2002) can be applied to haplotype estimation within a window for a large number of SNPs (e.g., 20). These multimarker approaches can effectively extract information from a case-parent trio as well as case-control data and increase the efficiency and detection power in regional or genome-wide LD analysis.

An alternative approach to EM in inferring haplotypes has been suggested by Clark (1990) and used in several recent studies (Reich et al. 2001; Stephens et al. 2001). However, as discussed by Clark (1990), several problems may arise with his procedure, e.g., the possibility of never being able to start the iterative algorithm because of the absence of any unambiguous individuals in the sample. Instead of finding only a list of possible haplotypes by Clark’s algorithm, the EM approach has the advantage of estimating haplotype frequencies, which in turn can be used directly for further \( \chi^2 \) testing in LD analysis. Stephens et al. (2001) proposed a new statistical method, applicable for genotype data at linked loci, for haplotype reconstruction from population data, based on an iterative sampling algorithm using a pseudo-Gibbs sampler. Their simulated results showed that error rates were reduced by >50% relative to its nearest competitor (EM-derived haplotypes). However, Zhang et al. (2001) compared the two methods and pointed out that one limitation of Stephens et al.’s study is the fact that their simulations were based on coalescent models, which may not be good approximations of human population evolutionary histories. Furthermore, their method cannot be applied to estimate the transmission haplotype frequencies. The PL-Gibbs sampler strategy proposed by Niu et al. (2002) not only can infer the haplotype phases with a large number of SNPs, but is also supposed to run faster than the method proposed by Stephens et al. (2001). It is possible to extend their method to estimate the affected (case) and unaffected (control) haplotype frequencies from population data.

![Figure 4](image-url)

**Figure 4.**—Comparison of linkage disequilibrium analyses of the ND procedure with the generalized TDT and conventional TDT methods for candidate gene 6 under SNP densities of (A) 1000-, (B) 2000-, and (C) 4000-bp intervals, respectively. The lines with open circles, triangles, and squares represent the results from our procedure, the generalized TDT, and the conventional TDT, respectively.
TABLE 4
EM-derived estimated numbers of haplotypes of the first five selected SNPs at 189, 557, 1505, 2732, and 3653 bp of candidate gene 1 and significance levels for the three samples studied

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Sample 1 (95 trios)</th>
<th>Sample 2 (97 trios)</th>
<th>Sample 3 (192 trios)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11111</td>
<td>96.0</td>
<td>122.0</td>
<td>99.0</td>
</tr>
<tr>
<td>11112</td>
<td>—</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>11122</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>11221</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>12122</td>
<td>1.0</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>21122</td>
<td>9.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>22111</td>
<td>9.0</td>
<td>3.0</td>
<td>5.0</td>
</tr>
<tr>
<td>22112</td>
<td>1.0</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>22122</td>
<td>67.0</td>
<td>48.0</td>
<td>74.0</td>
</tr>
<tr>
<td>Total</td>
<td>190.0</td>
<td>190.0</td>
<td>194.0</td>
</tr>
</tbody>
</table>

$S_{HL} = \{0.05, 0.01, 0.001\}^3$ |

<table>
<thead>
<tr>
<th>P value at second locus</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.078</td>
<td>0.001</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Trans., transmitted.

*No haplotype in the sample.

Empirical $P$ values $S^*$ obtained from 10,000 randomized permutations for 0.05, 0.01, and 0.001 significance levels, respectively.

Corresponding statistic $S$ at second locus; see Table 3.

and to estimate transmission haplotype frequencies for independent case-parent trio data as done in Zhao et al. (2000), together with the procedure reported herein. Indeed, this idea has been extended for case-control association analysis with multiple tightly linked SNP markers by using EM-derived haplotype frequencies (Fallin et al. 2001). For example, if the haplotype frequencies or numbers of multiple SNPs are estimated by the PL-EM algorithm (Qin et al. 2002) or the algorithm proposed by Niu et al. (2002), we can calculate the maximum statistic $S^*$ for each site with a wider moving window to find associated functional sites or regions, using simulation permutation to obtain $P$ values as indicated here.

In the current version of our program, the missing genotype of a parent and/or the affected offspring at any SNP site is assigned a code that differs from the codes assigned to the two possible SNPs at the position. As expected, this approach will yield rare haplotypes, which may eventually affect the power of the test to detect functional sites.

The chi-square statistic is highly sensitive to small cell counts. Recently, Seltman et al. (2001) pointed out that although haplotypes are more informative than alleles of individual SNPs in the transmission disequilibrium-based approach, the larger number of haplotypes relative to alleles at individual loci might decrease the statistical power of haplotype-based TDT. The reason for this may be the larger number of degrees of freedom required. As indicated earlier, our program automatically clusters haplotypes into fewer groups, and eventually a maximum $S(x)$ with the smallest $P$ value is chosen as $S^*(x)$ at locus $x$. By applying the method to the simulated data sets, we found that the shorter the window width, the smaller the number of degrees of freedom. This suggests that the clustering of haplotypes into fewer groups may, in fact, improve the power of the test.

The number of possible haplotypes increases rapidly with the size of the searching window. To avoid very rare haplotypes, we compared sample sizes from 100 to 1000 and found that the maximum searching window width should be set in the range approximately four to eight tightly linked SNPs (equivalent to $\sim$16–256 possible haplotypes). By empirical study, Niu et al. (2002) and Qin et al. (2002) also suggested that five to eight SNPs appeared to be a good choice for the atomistic unit size before the ligation step. Specifically, the maximum width was set at 5 in all the LD analyses reported in this study. If the sample size is large enough, it is sufficient to use a single initial condition for the haplotype frequencies (e.g., all haplotypes are set equally frequent) to start the EM estimation (Hawley and Kidd 1995; Slatkin and Excoffier 1996). Given the sample size in the GAW 12 data set, we used this single initial condition. On the other hand, if the sample size is not large enough, then multiple initial conditions should be used for EM estimation to avoid going to a local optimum.

An example has been given for the first five selected SNP loci of candidate gene 1 to illustrate our procedure. Table 4 shows the estimated numbers of haplotypes...
found by using the EM algorithm for the first five selected SNPs of candidate gene 1 and P values for the maximum statistic $S^*$ obtained from 10,000 randomized permutations in all samples. The nonancestral and ancestral sequence variants of candidate gene 1 were coded as alleles 1 and 2, respectively. On the basis of the estimated haplotype numbers, it is easy to see evidence of ancestral haplotypes that are shared more often among the haplotypes transmitted than among those not transmitted to affected offspring. In this case, allele 2 at base pair 557 is associated with the disease.

Recently, Zhang et al. (2002b) applied the haplotype block partition and “tag SNPs,” distinguishing algorithms to association studies. After haplotype block partition, tag SNPs were further selected within each haplotype block and used for association analysis. Their simulation results indicated that the genotyping efforts can be significantly reduced by using the tag SNPs without much loss of power, and haplotype-based LD analysis can be much more powerful than marker-by-marker analysis. It seems that a more effective and powerful searching strategy should combine the procedures of identifying haplotype block structure and the corresponding tag SNPs with the PL-EM (Qin et al. 2002) and/or PL-Gibbs sampling (Niu et al. 2002) algorithms as well as our moving-window statistic $S(x)$ and permutation methods.

The procedure proposed in this study is nonparametric and robust. For a complex disease controlled by multiple genes, the effect of a functional site accounting for as little as 4.6% of the variance in liability (e.g., candidate gene 6) could be detected with the sample size studied. Covariate factors such as gender, age, environment, the disease mode (dominant, recessive, additive, or multiplicative), and penetrance, were not considered in the proposed method. Also, the power of a haplotype-based test study depends on the data structure, the trait of interest, the polymorphism information content of marker loci, the degree of LD among the markers, the allele frequencies, the amount of population stratification, and the marker density. Our approach makes the implicit assumption that the underlying population is homogeneous. In a non-admixed population, if linkage disequilibrium exists across a region, then recombination must be quite infrequent and probably can be safely ignored. If population migration, admixture, or stratification is present, this should affect the estimates of allele or haplotype frequencies and decrease the power of LD detection. A model-based procedure implementing some appropriate parameters may then improve the power of the method. Also, it remains to be determined in a future study whether our idea can be extended to other data structures, such as nuclear families with multiple sibs and/or one or both parents missing, or to extended pedigrees.

The authors thank the editor, anonymous reviewers, and Dr. Hongyu Zhao for their valuable comments and suggestions on the manuscript. Also, we thank Dr. David Clayton for allowing us to use his TRANSMIT program and providing advice on using it. The use of the GAW 12 simulated data was permitted by the Southwest Foundation for Biomedical Research, which was supported by a GAW grant, GM31575 from National Institute of General Medical Sciences. We are grateful to Dr. Jean MacCher and Ms. Vanessa Olmo for handling our request for the GAW 12 data. This research was supported in part by National Institutes of Health (NIH) grant DA12844 (to M.D.L.); GM28356, RR03655, and DK57292 (to R.C.E.); GM58934 (to F.A.W.); National Science Foundation grant DMS-9971770 (to S.L.); and a General Clinical Research Center grant from the NIH (M01-RR02111) awarded to the University of Tennessee Health Science Center.

LITERATURE CITED


Kruglyak, L., 1999 Prospects for whole-genome linkage disequi-
Lazzeroni, L. C., and K. Lange, 1998 A conditional inference
Li, J., D. Wang, J. Dong, R. Jiang, K. Zhang et al., 2001 The power of trans-
Martin, E. R., N. L. Kaplan and B. S. Weir, 1997 Tests for linkage and
Martin, E. R., S. A. Monks, L. W. Warren and N. L. Kaplan, 2000 A test for linkage and association in general pedigrees: the pedi-
Niu, T., Z. S. Qin, X. Xu and J. S. Liu, 2002 Bayesian haplotype
Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi et al., 2001 Blocks of limited haplotype diversity revealed by high-
Qin, Z. S., T. Niu and J. S. Liu, 2002 Partition-ligation-expectation-
maximization algorithm for haplotype inference with single-
Risch, N. J., 2000 Searching for genetic determinants in the new
Selke, H., K. Roeder and B. Devlin, 2001 Transmission/disequili-
brium test meets measured haplotype analysis: family-based asso-
Schaid, D. J., 1996 General score tests for associations of genetic
markers with disease using cases and their parents. Genet. Epide-
Poland, 2002 Score tests for association between traits and
Sham, P., 1997 The transmission/disequilibrium tests for multial-
Sham, P. C., and D. Curtis, 1995 An extended transmission/disequilib-
Slagkin, M., and L. Excoffier, 1996 Testing for linkage disequilib-
Spiegelman, R. S., and W. J. EWens, 1996 The TDT and other family-
Spiegelman, R. S., and W. J. EWens, 1998 A sibship test for linkage in the presence of association: the sib transmission/disequilib-
Spiegelman, R. S., R. E. McGinnis and W. J. EWens, 1993 The transmis-
sion test for linkage disequilibrium: the insulin gene and insulin-
Sun, F., W. D. Flanders, Q. Yang and M. J. Khoury, 1999 Transmis-
sion disequilibrium test (TDT) when only one parent is available: the 1-TDT. Am. J. Epidemiol. 150: 97–104.
Teng, J., and N. Risch, 1999 The relative power of family-based
and case-control designs for linkage disequilibrium studies of
Terwilliger, J. D., and J. Ott, 1992 A haplotype-based “haplotype
Thomson, G., 1995 Mapping disease genes: family-based association
Toivonen, H. T., P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon et al., 2000 Data mining applied to linkage disequilibrium map-
Wilson, S. R., 1997 On extending the transmission/disequilibrium
Xiong, M., and S. W. Guo, 1997 Fine-scale genetic mapping based
Zhang, K., M. Deng, T. Chen, M. S. Waterman and F. Sun, 2002a A dynamic programming algorithm for haplotype block parti-
Zhang, K., P. Calabrese, M. Nordborg and F. Sun, 2002b Haplo-
Zhang, S. L., A. J. Pakstis, K. K. Kidd and H. Y. Zhao, 2001 Compar-
Zheng, C., and R. C. Elston, 1999 Multilocus linkage disequilib-
rum mapping with particular reference to the African-American
Zhu, X., and R. C. Elston, 2000 Power comparison of regression
Zhu, X., and R. C. Elston, 2001 Transmission/disequilibrium tests