European Journal of Human Genetics (2001) 9, 291 – 300


www.nature.com/ejhg

ARTICLE

Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain?

Joshua Akey[1], Li Jin[1] and Momiao Xiong*[,1]

[1]*Human Genetics Center, University of Texas – Houston, Houston, Texas, USA*

The genetic dissection of complex diseases represents a formidable challenge for modern human genetics. Recently, it has been suggested that linkage disequilibrium (LD) based methods will be a powerful approach for delineating complex disease genes. Most proposed LD test statistics search for association between a single marker and a putative trait locus. However, the power of a single marker association test may suffer because LD information contained in flanking markers is ignored. Intuitively, haplotypes (which can be regarded as a collection of ordered markers) may be more powerful than individual, unorganised markers. In this study, we derive the analytical tools based on standard chi-square statistics to directly investigate and compare the power between multilocus haplotypes and single marker LD tests. More specifically, novel formulas are obtained in order to calculate expected haplotype frequencies of unlimited size. This study demonstrates that the use of haplotypes can significantly improve the power and robustness of mapping disease genes. Additionally, we detail how the power of haplotype based association tests are affected by important population genetic parameters such as the genetic distance between markers and disease locus, mode of disease inheritance, age of trait causing mutation, frequency of associated marker allele, and level of initial LD. Finally, published data from the Hereditary Hemochromatosis disease region is used to illustrate the utility of haplotypes. *European Journal of Human Genetics* (2001) **9**, 291 – 300.

Keywords: linkage disequilibrium; haplotypes; complex disease; power

Introduction

The identification of genes that predispose to complex diseases represents a formidable challenge for modern human genetics. Recently, it has been suggested that the availability of dense single nucleotide polymorphism (SNP) marker maps will make genome wide linkage disequilibrium mapping (LDM) the method of choice for delineating complex trait loci.[1,2] There are two fundamentally different study designs for LDM: (1) family based methods; or (2) case-control studies. The transmission disequilibrium test (TDT) has become a widely used family-based LD method for mapping disease genes.[3] The primary benefit of the TDT is that spurious associations resulting from population stratification are avoided by using heterozygous parents or unaffected sibs of the proband as controls. However, because many complex diseases are late-age of onset it may be difficult to ascertain parents or sibs of probands. Hence, simple case-control study designs are an attractive alternative to family based approaches, although one must again consider the effects of population substructure. Currently, there is considerable debate regarding the optimal study design in genome-wide LDM.

Regardless of which study design will ultimately prove to be most effective, another question remains: Are single markers or haplotypes more powerful in LDM? Intuitively, one would expect that haplotypes would be more powerful due to the simultaneous use of multiple marker information.[4–6] Recently, this heuristic reasoning has prompted several studies which have explored the use of haplotypes, as opposed to individual markers, in LD based association studies. For example, Service *et al*[5] proposed a likelihood ratio based haplotype LD method, which they referred to as

*Correspondence: Dr Momiao Xiong, Human Genetics Center, University of Texas – Houston, P.O. Box 20334, Houston, Texas 77225, USA.
E-mail: mxiong@utsph.sph.uth.tmc.edu
Received 15 August 2000; revised 6 November 2000; accepted 28 November 2000

ancestral haplotype reconstruction (AHR), for initial genome-wide screens. Based on simulation studies they conclude that multi-locus haplotype LD tests may offer greater power in mapping studies than single locus tests. Zollner and von Haeseler[7] used a coalescent approach to study linkage disequilibrium between SNPs and found that haplotypes constructed from two SNPs were much more efficient in detecting associations. However, they did not explore haplotypes comprised of more than two loci. Contrastingly, Long and Langley[8] conclude that marker-based permutation tests are at least as or more powerful than simple haplotype-based tests. It is important to note that all of the preceding theoretical studies involved a simulation approach for assessing power. While simulations are an important and necessary part of theoretical work, derivation of explicit analytical formulas often allows a more holistic and detailed understanding of how parameters affect the power of the process under study.

Moreover, contradictory results have arisen from empirical data as some studies suggest that haplotype based LD methods improve the power over single marker tests,[9,10] while other studies do not.[11,12] Thus, the available literature does not provide a clear answer to the practical utility of multilocus haplotypes in association studies.

In this study, we derive the analytical tools necessary to investigate and compare the power between multilocus haplotypes and single marker LD tests. We formulate our power analysis based on standard chi-square statistics which have been routinely applied to single marker LD tests and are the natural extension to haplotype data. In particular, explicit analytical formulas are derived to calculate expected haplotype frequencies (for haplotypes comprised of any number of markers) and the noncentrality parameter of the test statistics distribution under the alternative hypothesis. We find that haplotypes can significantly improve the power of an association test if analysed appropriately. Furthermore, we explore how the power of haplotype tests are affected by population genetic parameters such as the number of loci generating a haplotype, the genetic distance between marker and disease locus, age of disease mutation, mode of disease inheritance, frequency of associated marker and disease allele, and the level of initial linkage disequilibrium. Finally, published data from the hereditary hemochromatosis disease region is used to illustrate our findings that haplotypes lead to increased power and robustness in association tests.

Methods

Test statistic

We consider a case-control study design for comparing the power between haplotype and single marker LD methods. Furthermore, we formulate our comparisons based on standard chi-square statistics because they are conceptually straightforward and have been widely applied to single marker LD tests (see [13,14]). We first describe the test statistic for the haplotype LD test. Suppose that n affected individuals and n unaffected individuals are sampled. The haplotype frequency data can be arranged in a $2 \times k$ contingency table, where k is the number of haplotypes. The null hypothesis H_0 to be tested is that of equal haplotype frequencies in affected and unaffected individuals. A conventional χ^2 statistic for testing H_0 can be defined as follows:

$$\chi^2_{HT} = 2n \sum_{l=1}^{k} \frac{(\hat{P}_{Al} \quad \hat{P}_{Cl})^2}{\hat{P}_{Al} + \hat{P}_{Cl}}$$

where $\hat{P}_{Al}$ and $\hat{P}_{Cl}$ are the observed frequencies of the l-th haplotype in cases and controls, respectively. Under the null hypothesis of equal haplotype frequencies, χ^2_{HT} is asymptotically distributed as $\chi^2_{k\ 1}$.

The test statistic using individual marker allele data is the same as χ^2_{HT} except $\hat{P}_{Al}$ and $\hat{P}_{Cl}$ are replaced by the observed marker allele frequencies in the cases and controls, respectively.[13] The test statistic using marker allele data will be denoted by χ^2_M.

Power calculation

The power to detect a disease gene, defined as the probability that the disease susceptibility locus will be detected if it is present, is an important index for evaluating the performance of a gene mapping method. Under the alternative hypothesis, H_a, of unequal haplotype frequencies in cases and controls, χ^2_{HT} is asymptotically distributed as a noncentral $\chi^2_{k\ 1}$ with noncentrality parameter:

$$\lambda = 2n \sum_{l=1}^{k} \frac{(P_{Al} \quad P_{Cl})^2}{P_{Al} + P_{Cl}}$$

where P_{Al} and P_{Cl} are the expected frequencies of the lth haplotype in cases and controls. Suppose that the critical value, α, for a test is $\chi^2_{k\ 1,1\ \alpha}$. The asymptotic power of the test with α-level significance is given by

$$\beta = P_{H_a}(\chi^2_{HT} \geqslant \chi^2_{k\ 1,1\ \alpha})$$

To calculate the power, β, we begin by calculating the noncentrality parameter and expected haplotype frequencies.

We assume that: (1) mating is random in the population; (2) generations are non-overlapping; (3) all alleles at the disease locus are selectively neutral; (4) there are no phenocopies; and (5) the population is isolated. Furthermore, we assume that a recent disease mutation was introduced into the population either by spontaneous mutation or by immigration of individuals carrying the disease mutation t generations ago. The time t is usually referred to as the age of the trait causing mutation.

We consider two alleles at the disease locus: disease allele D with frequency P_D and normal allele d with frequency P_d. We

assume that k markers are located in the region around the disease locus. Let I_j be the number of alleles at the j-th marker (j=1,2, . . . ,k) and $\theta_{i,j}$ be the recombination fraction between the marker M_i and the marker M_j. Let $M_{i_1}\ M_{i_2}\ldots M_{i_k}$ be the haplotype produced by the i_1-th allele at marker M_1, the i_2-th allele at marker M_2, and the i_k-th allele at marker M_k. We denote the frequency of the haplotype $M_{i_1}\ M_{i_2}\ldots M_{i_k}$ by $P_{i_1 i_2}\ldots{}_{i_k}$. Recently, recursive and deterministic formulas have been derived to investigate haplotype frequencies in admixed populations.[15] Here we derive explicit formula for the expected haplotype frequency under a stochastic model[16] which will be useful in our theoretical analysis. Below we give the formula for the expected frequency of three-locus haplotypes whose derivation, along with that for the general k-locus haplotype, can be found in Appendix A. Let $\delta_{i_1 D i_2}(0)$ be a coefficient of initial linkage disequilibrium of the three loci M_1DM_2 at the occurrence of disease mutation. Then

$$\begin{aligned} E\{P_{i_1 D i_2}(t)\} &= \delta_{i_1 D i_2}(0)e^{\ (\theta_{1,D}+\theta_{D,2})t} + P_{i_1}\delta_{D i_2}(0)e^{\ \theta_{D,2}t} \\ &\quad + P_{i_2}\delta_{i_1 D}(0)e^{\ \theta_{1,D}t} + P_{i_1}P_D P_{i_2} \\ E\{P_{i_1 d i_2}(t)\} &= P_{i_1 i_2} \quad E\{P_{i_1 D i_2}(t)\}, \end{aligned} \tag{1}$$

where $\delta_{i_1 D i_2}(0) = P_{i_1}D_{i_2}(0) \quad P_{i_1}\delta_{D_{i_2}}(0) \quad P_{i_2}\delta_{i_1 D}(0) \quad P_{i_1}P_{i_2}P_D$, is the coefficient of initial linkage disequilibrium at the three loci M_1DM_2 [14], $\delta_{Di_2}(0) = P_{Di_2}(0) \quad P_D P_{i_2}$ is the coefficient of initial linkage disequilibrium at the two loci DM_2 and $\delta_{i_1 D}(0) = P_{i_1 D}(0) \quad P_{i_1}P_D$ is the coefficient of initial linkage disequilibrium at the two loci M_1D.

The above formulae have clear biological meaning. For instance at generation t=0, the haplotype frequency $P_{i_1 D i_2}(0)$ consists of four components. The first component involves the coefficient of linkage disequilibrium at three loci: M_1, D, and M_2. The second component is the product of the population frequency of marker allele M_{i_1} and the coefficient of linkage disequilibrium between the disease locus D and marker locus M_2. The third component is the product of the frequency of marker allele M_{i_2} and the coefficient of linkage disequilibrium between the marker locus M_1 and the disease locus D. The fourth component is the equilibrium value of the haplotype frequency $P_{i_1 D i_2}(t)$.

To calculate the noncentrality parameter λ we must obtain the expected haplotype frequencies in affected and unaffected individuals. Let f_{11}, f_{12} and f_{22} be the penetrance of genotypes DD, Dd and dd, respectively, with $f_{11} \geqslant f_{12} \geqslant f_{22} \geqslant 0$. These penetrances can then be used to describe the following disease models: recessive ($f_{11}=x$ and $f_{12}=f_{22}=0$), additive ($f_{11}=x$, $f_{12}=\frac{x}{2}$, and $f_{22}=0$), and dominant ($f_{11}=f_{12}=x$ and $f_{22}=0$), where ($0<x\leqslant 1$). The probability of an individual being affected is given by

$$P(\text{Affected}) = f_{11}P_D^2 + 2f_{12}P_D P_d + f_{22}P_d^2\,.$$

Let us define $a_1 = \frac{f_{11}P_D + f_{12}P_d}{P(\text{Affected})}$, $a_2 = \frac{f_{12}P_D + f_{22}P_d}{P(\text{Affected})}$, $b_1 = \frac{(1\ \ f_{11})P_D + (1\ \ f_{12})P_d}{1\ \ P(\text{Affected})}$, and $b_2 = \frac{(1\ \ f_{12})P_D + (1\ \ f_{22})P_d}{1\ \ P(\text{Affected})}$. Suppose the markers M_i, M_j and M_k are in the order M_i-disease-M_j-M_k. Let H_{ij} and H_{ijk} denote the haplotypes A_iB_j and $A_iB_jC_k$, respectively. We can show that (Appendix B)

$$\begin{aligned} P(H_{ij}|\text{Affected}) &= a_1 P_{iDj} + a_2 P_{idj} \\ P(H_{ij}|\text{Unaffected}) &= b_1 P_{iDj} + b_2 P_{idj} \\ P(H_{ijk}|\text{Affected}) &= a_1 P_{iDjk} + a_2 P_{idjk} \\ P(H_{ijk}|\text{Unaffected}) &= b_1 P_{iDjk} + b_2 P_{idjk}\,. \end{aligned}$$

The above formulae show that each of the haplotype frequencies in affected and unaffected individuals will be a weighted average of P_{iDj} and P_{idj} or P_{iDjk} and P_{idjk}, with the weights determined by the mode of inheritance of the disease. For a recessive disease assuming x = 1, $a_1 = \frac{1}{P_D}$, a_2=0, $b_1 = \frac{1}{1+P_D}$, and $b_2 = \frac{1}{1\ \ P_D^2}$. Thus, $P(H_{ij}|\text{Affected}) = \frac{P_{iDj}}{P_D}$ and $P(H_{ijk}|\text{Affected}) = \frac{P_{iDjK}}{P_D}$. If the recessive disease is rare, then $P(H_{ij}|\text{Unaffected}) \approx \frac{P_{idj}}{P_d}$ and $P(H_{ijk}|\text{Unaffected}) \approx \frac{P_{idjk}}{P_d}$. Details of power calculations involving χ^2_M can be found elsewhere.[13]

Results

Power of χ^2_{HT} and χ^2_M for different disease models

From the previous section we know that the power of χ^2_{HT} depends on a number of parameters such as initial LD, the age of the trait causing mutation, recombination fraction between the marker and disease locus, the mode of disease inheritance, and marker and disease allele frequencies. We now compare the power of the haplotype based statistic, χ^2_{HT}, to that of the single marker statistic, χ^2_M. The markers are assumed to be biallelic (ie, SNPs) and evenly spaced in relation to the disease locus. Specifically, we consider disease carrying haplotypes comprised of two and four marker loci, and to emphasise the number of markers in the haplotype we will refer to them as two and four marker locus haplotypes. To avoid confusion note that two and four marker locus haplotypes are equivalent to three and five locus haplotypes, respectively, when the disease gene is present. For example, the two marker locus haplotype consists of two equally spaced markers with the disease locus located in the middle of marker 1 and 2 (M_1DM_2). Similarly, the four marker locus haplotypes consists of four equally spaced genetic markers with the disease locus in the middle of marker 2 and 3 ($M_1M_2DM_3M_4$). In the following figures, M denotes the single marker LD test, H_2 the two marker locus LD test, and H_4 the four marker locus LD test. Figures 1 and 2 show the power of M, H_2, and H_4 with a significance level α=0.001 as a function of the recombination fraction between markers. For H_2 and H_4 the recombination fraction is between the two markers adjacent to the disease locus. In the case of M, the recombination fraction is between the closest marker and the disease locus. The remaining parameters can be found in the figure legends. Figures 1 and 2 clearly demonstrate that the power of H_4 is much higher than the power of H_2, which in turn is higher than that of M.

Next we compare the performance of χ^2_{HT} and χ^2_M for the genotype relative risk disease model,[1] where the genotype

relative risk for individuals of genotype *Dd* and *DD* is γ and γ^2 times greater than that of individuals with genotype *dd*. The power of *M*, H_2, and H_4 for $\gamma = 4$ and $N = 250$ (where N is the sample size) are shown in Figure 3. Because such a model is more complex than the recessive and dominant disease models, the power of the χ^2_{HT} and χ^2_M are dramatically reduced. Figure 3 also indicates that the conclusions for Figures 1 and 2 are still valid under the genotype relative risk disease model.

Genome wide screens

We now compare the power between χ^2_M and χ^2_{HT} when they are used as tools in genome-wide scans by calculating the sample sizes required to achieve 80% power with a significance level of $\alpha = 0.001$. Table 1 presents the required sample sizes for dominant and recessive diseases. For such simple disease models, when the age of the disease mutation does not exceed 20 generations the samples sizes for both the

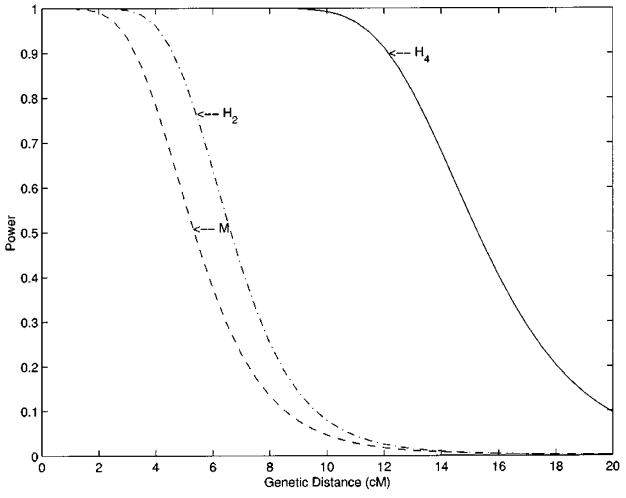

Figure 1 Power of the test statistics χ^2_{HT} and χ^2_M with a significance level $\alpha = 0.001$ for a recessive disease as a function of the genetic distance between markers (see text for details). Two and four marker locus haplotype data (H_2 and H_4, respectively) are used for χ^2_{HT}, and the single marker allele data (*M*) is used for χ^2_M. We assume that $N = 100$, $P_D = 0.1$, equal marker allele frequencies, complete initial linkage disequilibrium, and $t = 15$.

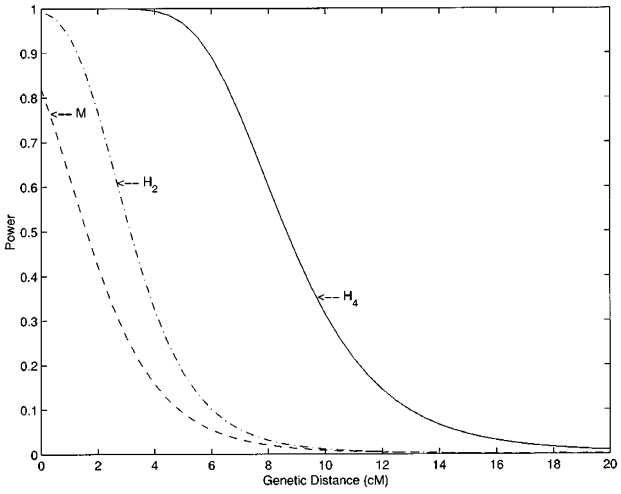

Figure 2 Power of the test statistics χ^2_{HT} and χ^2_M with a significance level $\alpha = 0.001$ for a dominant disease as a function of the genetic distance. The parameters are the same as that of Figure 1.

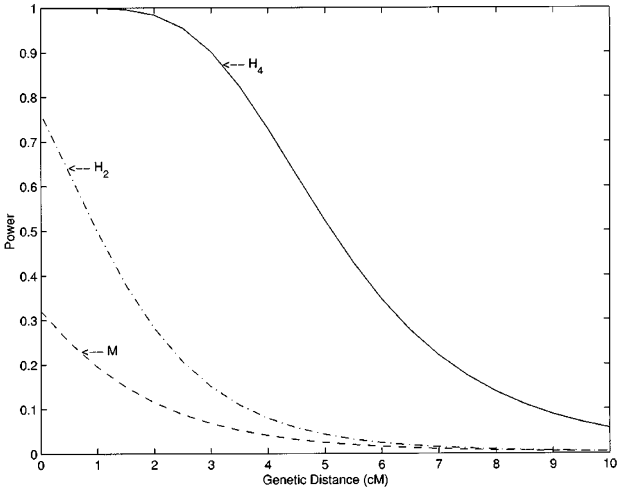

Figure 3 Power of the test statistics χ^2_{HT} and χ^2_M with a significance level $\alpha = 0.001$ for a genotypic relative risk $\gamma = 4$. A sample size of $N = 250$ is assumed. Other parameters are the same as that of Figure 1.

Table 1 Number of cases required to achieve 80% power with a significance level α=0.001 for a recessive and dominant disease

	4 cM			*5 cM*			*10 cM*		
	t=*10*	t=*20*	t=*50*	t=*10*	t=*20*	t=*50*	t=*10*	t=*20*	t=*50*
Recessive									
H_2	15	23	77	17	29	127	29	77	1544
M	22	34	122	24	43	204	43	122	2530
Dominant									
H_2	42	66	227	47	81	376	81	227	4602
M	95	144	485	106	176	802	176	485	9812

single marker LD and haplotype LD tests are realistic even if the genetic distance between the adjacent markers is 10 cM. It is interesting to note that the reduction of sample sizes by H_2 for a dominant disease gene is larger than that for a recessive disease.

Table 2 shows that the sample sizes required for M, H_2, and H_4 under the genotype relative risk models. The markers are assumed to be equally spaced with a genetic distance between adjacent markers of 1 cM, and the frequencies of the two alleles at each marker are assumed to be equal. We can see from Table 2 that the required sample size for H_4 is much smaller than that required by M. When the frequency of the disease allele is very small, neither χ^2_M nor χ^2_{HT} has the power to detect the disease even when the markers are spaced every 1 cM. However, when the frequency of the disease allele is larger than 0.05 the sample sizes required by the four-locus haplotype LD test are within reach.

Population genetic parameters

We now systematically study the effect of various population genetic parameters on the power of χ^2_{HT} to more fully understand what populations are most appropriate for LD approaches and to provide guidance in study designs. Figure 4 shows the power of the χ^2_{HT} with four marker locus haplotypes for a recessive disease under four different ages of trait-causing mutations. We can see from Figure 4 that the power of χ^2_{HT} decreases as the age of the mutation increases. As the generations increase, much of the allele sharing identical by decent (IBD) with the original founder chromosome decays through recombination. Hence, linkage disequilibrium also decreases between surrounding markers and the disease locus itself. Figure 5 displays how the frequency of the associated marker allele impacts the power of the χ^2_{HT} with four marker-locus haplotypes. For the convenience of presentation, we assume that the alleles at the four marker loci M_1, M_2, M_3 and M_4 have the same frequency distributions. Figure 5 shows that the power of χ^2_{HT} decreases as the frequency of the associated marker allele increases. Smaller associated marker allele frequencies lead to stronger initial linkage disequilibrium and hence higher power. Initial linkage disequilibrium is largely determined by the initial

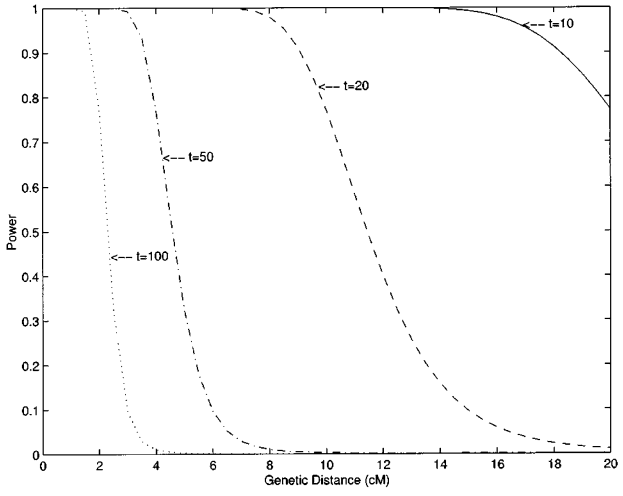

Figure 4 The effect of the age of the mutation on power of the test statistic χ^2_{HT}. The parameters are set as follows: $N=100$, $P_D=0.1$, complete initial linkage disequilibrium, dominant disease, and equal marker allele frequencies.

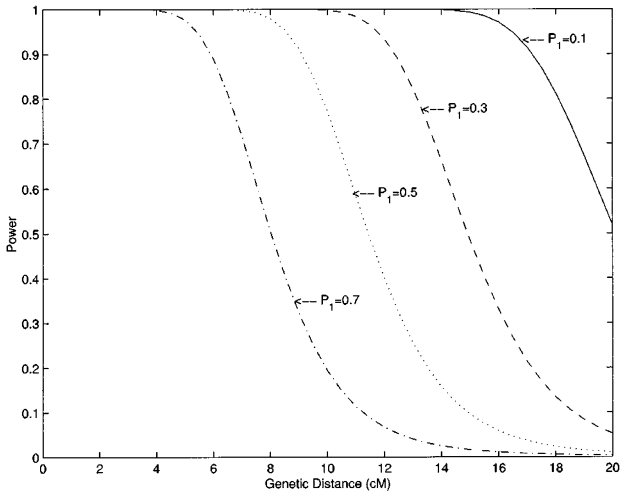

Figure 5 The effect of the frequency of the associated marker allele on the power of the test statistic χ^2_{HT}. We assume that $t=20$. Remaining parameters are as that in Figure 4.

Table 2 Number of cases required to achieve 80% power with a significance level α=0.001 for a complex disease

Test	P_D=*0.01*		P_D=*0.05*		P_D=*0.1*		P_D=*0.5*	
statistic	t=*10*	t=*20*	t=*10*	t=*20*	t=*10*	t=*20*	t=*10*	t=*20*
γ=4								
H_2	8,797	10,003	460	521	150	170	14	16
H_4	3,112	3,320	214	242	84	87	11	14
M	20,010	22,120	894	900	297	328	26	28
γ=2								
H_2	60,564	68,950	2,617	2,973	712	807	39	44
H_4	19,120	19,430	923	953	284	293	22	24
M	140,194	154,990	5,440	5,550	1,558	1,722	78	86

frequency of the associated marker allele and haplotype frequency in the disease mutation-carrying chromosomes.

For clarity of presentation, we now consider the two marker locus haplotype LD test. Figure 6 shows how the initial haplotype frequencies in the disease mutation-carrying chromosomes affects the power of the haplotype LD test. The initial haplotype frequency in other chromosomes are assumed to be equal. The larger the initial haplotype frequency in disease mutation-carrying chromosomes, the larger the initial linkage disequilibrium between markers and disease loci. We can see from Figure 6 that the power of χ^2_{HT} increases as the initial associated haplotype frequency increases.

An example: hereditary hemochromatosis

As a practical example, we present the test statistic values (χ^2_{HT} and χ^2_M) of various sized haplotypes and single marker tests for data surrounding the HFE gene, which when mutated leads to hereditary hemochromatosis (HH).[17] As the number of markers generating the haplotype increases the number of possible haplotypes and degrees of freedom increases exponentially. This is particularly problematic if mulitallelic markers are used to generate multilocus haplotypes, as is the case in this example. To simplify the analysis as well as ensure an appropriate number of counts in each cell of the contingency table, a grouping procedure can be employed. Specifically, we designate the most frequent haplotype in cases as haplotype 1 and all others are grouped together as haplotype 2. The equality of haplotypes between cases and controls can be assessed by χ^2_{HT} with χ^2_1 distribution. It is important to note that in the theoretical power comparisons of χ^2_{HT} and χ^2_M no grouping procedure was used and all possible haplotype combinations were analysed.

Markers D6S265, HLA-A, HLA-F, D6S258, D6S105, D6S424, and D6S1260 lie approximately 3.9, 3.8, 3.7, 1.9, 1.8, 1.7 and 1 cM, respectively, centromeric to the HFE gene. Tables 3 and 4 present the LD test statistic results for individual marker data and haplotypes. We observe several remarkable features. First, although markers D6S1260 and D6S424 are closer to the HFE gene than the other markers, the smallest *P* value occurs at marker D6S105, which is 1.8 cM away from the HFE gene. Second, while the *P* values of the single-marker LD test at the markers D6S1260 and D6S424 are 0.017 and 0.002, respectively, the *P* value of the two-locus haplotype LD test for these markers is 0.000027; thus the haplotype LD test offers considerable improvement over the single-marker LD test. Third, the values of the single-marker LD test statistic at different markers oscillates erratically, whereas the values of the haplotype LD test statistic follow a simpler pattern and change more smoothly, as shown in Figure 7. To interrogate the performance of the grouping procedure for the haplotype

Table 3 Single marker LD test

Marker	*Allele*	*Frequency cases*	*Frequency control*	χ^2	P-*value*
D6S265	1	25	4	16.8	4.2×10^{-5}
HLA-A	3	26	3	19.1	1.24×10^{-5}
HLA-F	2	32	11	10.9	9.6×10^{-4}
D6S258	4	42	16	16.3	5.4×10^{-5}
D6S306	3	46	26	6.9	0.009
D6S105	8	37	8	21.4	3.7×10^{-6}
D6S424	6	50	25	9.5	0.002
D6S1260	4	45	25	5.7	0.017

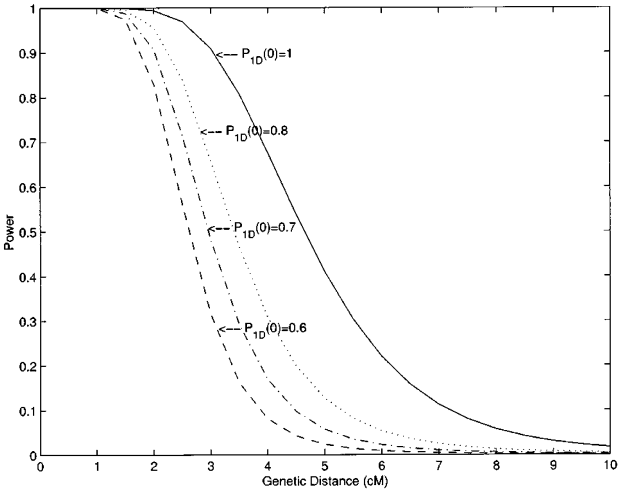

Figure 6 The effect of the initial associated marker allele frequency and haplotype frequency in the disease mutation-carrying chromosomes. We assume equal marker allele frequencies. Remaining parameters are as that in Figure 5.

Table 4 Haplotype LD test results for haplotypes of various size

Haplotype	*D6S265*	*HLA-A*	*HLA-F*	*D6S258*	*D6S306*	*D6S105*	*D6S424*	*D6S1260*	*Frequency cases*	*Frequency controls*	χ^2	P-*value*
A	1*	3	2	4	3	8	6	4	17	0	17.1	3.55×10^{-5}
B	–	3	2	4	3	8	6	4	17	0	17.1	3.55×10^{-5}
C	–	–	2	4	3	8	6	4	20	0	20.2	7.00×10^{-6}
D	–	–	–	4	3	8	6	4	21	0	21.8	3.00×10^{-6}
E	–	–	–	–	3	8	6	4	24	0	24.9	3.60×10^{-7}
F	–	–	–	–	–	8	6	4	26	0	28.6	8.90×10^{-8}
G	–	–	–	–	–	–	6	4	39	12	17.6	2.70×10^{-5}

*Denotes the most frequently occurring allele in cases.

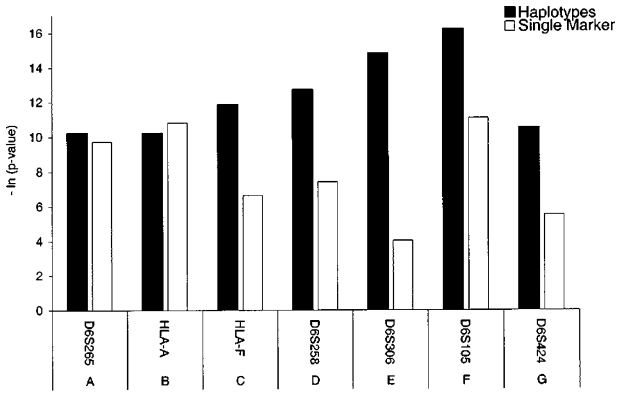

Figure 7 A comparison of the *P*-values (reported as -ln *P* value) between single marker LD tests and various sized haplotype LD tests for data surrounding the HFE gene. The *P* values are from Tables 3 and 4.

LD test we consider the 8-locus haplotype (D6S265-D6S1260) LD test and use two statistics. One statistic uses the grouping procedure described above and the other statistic analyses all possible observed haplotypes. The *P* value of the haplotype LD test using the grouping procedure is 0.0025, whereas the *P* value of the haplotype LD test using all observed haplotypes is 0.3597.

Discussion

Most LD mapping methods use a single marker to assess LD between a marker and disease locus. However, several markers within small regions may be in strong LD with both each other and the disease locus. Single marker based LD methods may not capture all of the available LD information, which is contained in multi-locus haplotypes. In this study, simple chi-square statistics, which have been the foundation for single marker LD tests, were extended to the analysis of haplotypes. Furthermore, using explicit analytical formula, we show that haplotypes result in increased power for detecting associations. More specifically, the power of single marker, two marker locus, and four marker locus haplotypes were investigated. In general, two and four marker locus haplotypes offer higher power than the corresponding single marker test. For instance, when the genetic distance between adjacent markers is 1 cM and the disease locus is in the middle of the markers, sample sizes required by the two and four marker locus haplotype LD test are roughly one-half and one-quarter, respectively, of the sample size required for the corresponding single marker LD test.

Moreover, another advantage of haplotypes is increased robustness compared to single marker LD tests. Evolutionary forces such as random drift, mutation at the marker locus, and varying degrees of initial LD tend to increase the variability of the observed magnitude of LD between any single marker and disease gene leading to complex patterns of association, even with tightly linked markers. However, simultaneous analysis of multiple markers in the form of haplotypes results in comparatively simpler patterns of LD. The potential smoothing effect that χ^2_{HT} has on *P* values is demonstrated in Figure 7. Furthermore, it is not necessary to analyse every observed haplotype, which could actually decrease the power to observe an association (see the HH example). Therefore, it is important to develop methods for identifying and prioritising the most informative haplotypes. Several approaches have already been proposed[5,6,18,19] and it is likely that this will be an important area of future research.

We chose to explore the power of haplotypes in a case-control study design rather than a TDT approach for several reasons. First, case-control studies are more economically efficient because they do not require parents or sibs to be genotyped. Second, several studies suggest that case-control studies are at least as powerful as the TDT and often more efficient.[20] Third, the primary criticism of case-control studies is the possibility of spurious associations due to population substructure. However, recent reports demonstrate simple and efficient approaches to assessing whether a significant association is due to 'true' LD or if it is likely attributable to substructure.[21] Finally, the use of extended marker haplotypes for the TDT has been explored.[22]

Like single marker LD tests, the age of the disease mutation has a large impact on the power of a haplotype LD test. Isolated populations where the age of disease mutation is less than 20 generations are most suitable for LD methods. Populations with past gene flow or population stratification will have an increased probability of false positives. Although the population genetic model that we assumed in our power analysis is simplistic, populations with young disease mutations and little gene flow do exist and include French Canadians,[23] Ashkenazi Jews,[24] Costa Ricans,[10] and the Amish.[25]

It is important to note that haplotype data is not readily available for genome-wide LD screens. Thus, it is very important to establish the advantages of haplotypes before embarking on the arduous task of reconstructing them from genotype data or laborious experimental methods. The aim of this study was to provide a detailed proof of the principle demonstrating that haplotypes should be considered as an important resource in mapping genes underlying complex diseases. New analytical and experimental approaches have been and continue to be developed, making haplotype reconstruction in the absence of family data a surmountable problem.[26–28]

In conclusion, LD is a complex phenomenon that is subject to the demographic history of a population, which itself is a stochastic process of evolution and cannot be observed. Population genetic models involving a number of parameters have attempted to describe the dynamics of demographic history. However, the complicated nature of these models has made the interpretation of LD results very difficult. Although the haplotype LD test that we describe has higher power and

is more robust than the corresponding single-marker LD tests, caution still needs to be taken in study design and data interpretation. Future studies will focus on expanding the general theory presented here to multilocus disease models and more complicated population genetic models.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516 – 1517.
2 Camp NJ: Genomewide transmission/disequilibrium testing-consideration of the genotype relative risks at disease loci. *Am J Hum Genet* 1997; **61**: 1424 – 1430.
3 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506 – 513.
4 Barton NH: Estimating multilocus linkage disequilibria. *Heredity* 2000; **84**: 373 – 389.
5 Service SK, Lang DW, Freimer NB, Sandkuijl LA: Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet* 1999; **64**: 1728 – 1738.
6 Maclean CJ, Martin RB, Sham PC, Wang H, Straub RE, Kendler KS: The trimmed-haplotype test for linkage disequilibrium. *Am J Hum Genet* 2000; **66**: 1062 – 1075.
7 Zollner S, von Haeseler A: Coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet* 2000; **66**: 615 – 628.
8 Long AD, Langley CH: The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999; **9**: 720 – 731.
9 Martin ER, Lai EH, Gilbert JR *et al*: SNPing away at complex diseases. *Am J Hum Genet* 2000; **67**: 383 – 394.
10 Escamilla MA, Mcinnes LA, Spesny M *et al*: Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. *Am J Hum Genet* 1999; **64**: 1670 – 1678.
11 Clark AG, Weiss KM, Nickerson DA *et al*: Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 1998; **63**: 595 – 612.
12 Terwilliger JD, Weiss KW: Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* 1998; **9**: 578 – 594.
13 Chapman NH, Wijsman EM: Genome screens using linkage disequilibrium tests: optimal marker – characteristics and feasibility. *Am J Hum Genet* 1998; **63**: 1872 – 1885.
14 Weir BS: *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates Inc. Publishers, Sunderland, Massachusetts, 1996.
15 Zheng C, Elston RC: Multipoint linkage disequilibrium mapping with particular reference to the African-American population. *Genet Epidemiol* 1999; **17**: 79 – 101.
16 Xiong MM, Guo SW: Fine-Scale genetic mapping based on linkage disequilibrium: theory and application. *Am J Hum Genet* 1997; **60**: 1513 – 1531.
17 Jazwinksa, EC, Cullen LM, Busfield F *et al*: Haemochromatosis and HLA-H. *Nat. Genet* 1996; **14**: 249 – 251.
18 McPeek MS, Strahs A: Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* 1999; **65**: 858 – 875.
19 Toivonen HT, Onkamo P, Vasko K *et al*: Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet* 2000; **67**: 133 – 145.
20 Morton NE, Collins A: Tests and estimates of allelic association in complex inheritance. *PNAS* 1998; **95**: 11389 – 11393.
21 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220 – 228.
22 Clayton D, Jones H: Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 1999; **65**: 1161 – 1169.
23 Casaubon LK, Melanson M, Lopes-Cendes I *et al*: The gene responsible for a severe form of peripheral neuropathy and agenesis of the corpus callosum maps to chromosome 15q. *Am J Hum Genet* 1996; **58**: 28 – 34.
24 Risch N, De Leon D, Ozelius L *et al*: Genetic analysis of idiopathnic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet* 1995; **9**: 152 – 159.
25 Sulisalo T, Klockars JM, Akitie Ö *et al*: High-resolution linkage disequilibrium mapping of the cartilage-hair hypoplasia gene. *Am J Hum Genet* 1994; **55**: 937 – 945.
26 Martin RB, Alda M, Maclean CJ: Parental genotype reconstruction: applications of haplotype relative risk to incomplete parental data. *Genet Epidemiol* 1998; **15**: 471 – 490.
27 Woolley AT, Guillemette C, Li Cheung C, Housman DE, Lieber CM: Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat Biotechnol* 2000; **18**: 760 – 763.
28 Gentalen E, Chee M: A novel method for determining linkage between DNA sequences: hybridization to paired probe arrays. *Nucleic Acids Res* 1999; **15**: 1485 – 1491.

Appendix A

The haplotype $M_{i_1}, \ldots, M_{i_k}$ may contain the disease locus. If the haplotype contains the disease locus, the disease locus will be explicitly expressed in the haplotype, for example, $M_{i_1}, \ldots, M_{i_l}$, D, $M_{i_{l+1}}, \ldots, M_{i_k}$. Recall that $\theta_{j,j+1}$ is the recombination fraction between the marker M_{i_j} and the marker $M_{i_{j+1}}$. Let $P_{i_1,\ldots,i_k}(t)$ be the frequency of the haplotype $M_{i_1}, \ldots, M_{i_k}$ in generation t. By the same argument as that in Zheng and Elston,[15] we obtain the following recursive formula for the frequency of the haplotype:

$$\begin{aligned}P_{i_1,\ldots,D,\ldots,i_k}(t+1) = &(1 - \sum_{j=1}^{l-1}\theta_{j,j+1} - \theta_{l,D} - \theta_{D,l+1} \\ &- \sum_{j=l+1}^{k-1}\theta_{j,j+1})P_{i_1,\ldots,D,\ldots,i_k}(t) \\ &+ \theta_{l,D}P_{i_1,\ldots,i_l}P_{D,i_{l+1},\ldots,i_k}(t) \\ &+ \theta_{D,l+1}P_{i_{l+1},\ldots,i_k}P_{i_1,\ldots,i_l,D}(t) \\ &+ \sum_{j=1}^{l-1}\theta_{j,j+1}P_{i_1,\ldots,i_j}P_{i_{j+1},\ldots,D,\ldots,i_k}(t) \\ &+ \sum_{j=l+1}^{k-1}\theta_{j,j+1}P_{i_{j+1},\ldots,i_k}P_{i_1,\ldots,D,\ldots,i_j}(t)\end{aligned}$$

If we assume that there is no mutation at the marker locus, the frequency of the haplotype that contains no disease locus is, in general, assumed constant. It follows from the above recursive formula that:

$$\begin{aligned}&E\{P_{i_1,\ldots,D,\ldots,i_k}(t+1) - P_{i_1,\ldots,D,\ldots,i_k}(t)\} = \\ &\left[\sum_{j=1}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-1}\theta_{j,j+1}\right]E\{P_{i_1,\ldots,D,\ldots,i_k}(t)\} \\ &+ \sum_{j=1}^{l-1}\theta_{j,j+1}P_{i_1,\ldots,i_j}E\{P_{i_{j+1},\ldots,D,\ldots,i_k}(t)\} \\ &+ \theta_{l,D}P_{i_1,\ldots,i_l}E\{P_{D,i_{l+1},\ldots,i_k}(t)\} \\ &+ \theta_{D,l+1}P_{i_{l+1},\ldots,i_k}E\{P_{i_1,\ldots,i_l,D}(t)\} \\ &+ \sum_{j=l+1}^{k-1}\theta_{j,j+1}P_{i_{j+1},\ldots,i_k}E\{P_{i_1,\ldots,D,\ldots,i_j}(t)\}\end{aligned}$$

Thus, the above equation leads to:

$$\begin{aligned}&\frac{dE\{P_{i_1,\ldots,D,\ldots,i_k}(t)\}}{dt} = \\ &\left[\sum_{j=1}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-1}\theta_{j,j+1}\right]E\{P_{i_1,\ldots,D,\ldots,i_k}(t)\} \\ &+ \sum_{j=1}^{l-1}\theta_{j,j+1}P_{i_1,\ldots,i_j}E\{P_{i_{j+1},\ldots,D,\ldots,i_k}(t)\} \\ &+ \theta_{l,D}P_{i_1,\ldots,i_l}E\{P_{D,i_{l+1},\ldots,i_k}(t)\} \\ &+ \theta_{D,l+1}P_{i_{l+1},\ldots,i_k}E\{P_{i_1,\ldots,i_l,D}(t)\} \\ &+ \sum_{j=l+1}^{k-1}\theta_{j,j+1}P_{i_{j+1},\ldots,i_k}E\{P_{i_1,\ldots,D,\ldots,i_j}(t)\} \qquad (2)\end{aligned}$$

Solving this equation recursively, we can obtain the expected haplotype frequency. For the ease of exposition, we first consider three locus haplotypes and then extend the results to the general k-locus haplotype.

For the three locus haplotype M_iDM_j, let $P_{iDj}(t)$ be the frequency of the haplotype M_iDM_j. Let $\theta_{i,D}$ and $\theta_{D,j}$ be the recombination fraction between the marker M_i and disease locus, and marker M_j and disease locus, respectively. P_j is the frequency of the marker allele M_j. P_i and P_D are defined as described above. It follows from (2) that:

$$\begin{aligned}&\frac{dE\{P_{iDj}(t)\}}{dt} = \\ &(\theta_{i,D} + \theta_{D,j})E\{P_{iDj}(t)\} + \theta_{i,D}P_iE\{P_{Dj}(t)\} + \theta_{D,j}P_jE\{P_{iD}(t)\}\end{aligned}$$

where $P_{Dj}(t)$ and $P_{iD}(t)$ are the frequencies of the haplotype DM_j and M_iD, respectively. Solving the above equation for $E\{P_{iDj}(t)\}$ leads to:

$$\begin{aligned}&E\{P_{iDj}(t)\} = \\ &\delta_{iDj}(0)e^{-(\theta_{i,D}+\theta_{D,j})t} + P_i\delta_{Dj}(0)e^{-\theta_{Dj}t} + P_j\delta_{iD}(0)e^{-\theta_{iD}t} + P_iP_jP_D\end{aligned}$$

where $\delta_{iDj}(0) = P_{iDj}(0) - P_i\delta_{Dj}(0) - P_j\delta_{iD}(0) - P_iP_jP_D$, a coefficient of initial linkage disequilibrium at three loci M_iDM_j,[14] $\delta_{Dj}(0) = P_{Dj}(0) - P_DP_j$, and $\delta_{iD}(0) = P_{iD}(0) - P_iP_D$.

Finally, we provide the formula for k-locus haplotype frequencies. Suppose that we have the haplotype $M_{i_1}, \ldots, M_{i_l}$, D, $M_{i_{l+1}}, \ldots, M_{i_k}$. Let $\theta_{l,D}$ and $\theta_{D,l+1}$ be the recombination fraction between the marker M_{i_l} and disease locus, and the marker $M_{i_{l+1}}$ and disease locus, respectively. Let $\theta_{j,j+1}$ be the recombination fraction between marker M_j and the marker M_{j+1}. $\delta_{i_1,\ldots,i_j}(0)$ denotes a coefficient of initial linkage disequilibrium at the loci $M_{i_1,\ldots,}M_{i_j}$. Then, the expected frequency of haplotype $M_{i_1,\ldots,i_k}$ is given by:

$$\begin{aligned}&E\{P_{i_1,\ldots,D,\ldots,i_k}\} = \\ &\delta_{i_1,\ldots,i_k}(0)e^{-\left(\sum_{j=1}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-1}\theta_{j,j+1}\right)t} \\ &+ P_{i_1}\delta_{i_2,\ldots,i_k}(0)e^{-\left(\sum_{j=2}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-1}\theta_{j,j+1}\right)t} \\ &+ P_{i_k}\delta_{i_1,\ldots,i_{k-1}}(0)e^{-\left(\sum_{j=1}^{l-1}\theta_{j,j-1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-2}\theta_{j,j+1}\right)t} \\ &+ P_{i_1i_2}\delta_{i_3,\ldots,i_k}(0)e^{-\left(\sum_{j=3}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-1}\theta_{j,j+1}\right)t} \\ &+ P_{i_{k-1}i_{k-2}}\delta_{i_1,\ldots,i_{k-3}}(0)e^{-\left(\sum_{j=1}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-3}\theta_{j,j+1}\right)t} \\ &+ P_{i_1}P_{i_k}\delta_{i_2,\ldots,i_{k-1}}e^{-\left(\sum_{j=2}^{l-1}\theta_{j,j+1} + \theta_{l,D} + \theta_{D,l+1} + \sum_{j=l+1}^{k-2}\theta_{j,j+1}\right)t} \\ &+ \cdots + P_{i_1,\ldots,i_{l-2}}P_{i_{l+1},\ldots,i_k}\delta_{l-1,l,D}(0)e^{-(\theta_{l-1,l}+\theta_{l,D})t} \\ &+ P_{i_1,\ldots,i_{l-1}}P_{i_{l+2},\ldots,i_k}\delta_{il,D,i_{l+1}}(0)e^{-(\theta_{l,D}+\theta_{D,l+1})t} \\ &+ P_{i_1,\ldots,i_l}P_{i_{l+3},\ldots,i_k}\delta_{D,i_{l+1},i_{l+2}}(0)e^{-(\theta_{D,l+1}+\theta_{l+1,l+2})t} \\ &+ P_{i_1,\ldots,i_{l-1}}P_{i_{l+1},\ldots,i_k}\delta_{i_l,D}(0)e^{-\theta_{l,D}t} + P_{i_1,\ldots,i_l}P_{i_{l+2},\ldots,i_k}\delta_{D,i_{l+1}}(0)e^{-\theta_{D,l+1}t} \\ &+ P_{i_1\ldots i_k} \qquad (3)\end{aligned}$$

where

$$\begin{aligned}
\delta_{i_l,D}(0) &= P_{i_l D}(0) \quad P_{i_l} P_D \\
\delta_{i_l D,i_{l+1}}(0) &= P_{i_l} D_{l+1}(0) \quad P_{i_l}\delta_{D,i_{l+1}}(0) \quad P_{i_{l+1}}\delta_{i_l,D}(0) \quad P_{i_l} P_D P_{i_{l+1}} \\
\vdots \delta_{i,\ldots,i_k}(0) &= P_{i_l,\ldots,i_k}(0) \quad P_{i_1}\delta_{i_2,\ldots,i_k}(0) \quad P_{i_k}\delta_{i_1,\ldots,i_{k\ 1}}(0) \\
&\quad P_{i_1 i_2}\delta_{i_3,\ldots,i_k}(0) \quad P_{i_{k\ 1} i_k}\delta_{i_1,\ldots,i_{k\ 2}}(0) \quad P_{i_1} P_{i_k}\delta_{i_2,\ldots,i_{k\ 1}}(0) \\
&\quad \cdots \quad P_{i_1\ldots i_{l\ 1}} P_{i_{l+1},\ldots,i_k}\delta_{i_l,D}(0) \quad P_{i_1,\ldots,i_l} P_{i_{l+2},\ldots,i_k}\delta_{D,i_{l+1}}(0) \\
&\quad P_{i_1,\ldots,i_k}
\end{aligned}$$

Appendix B

Suppose that the disease locus is located in the interval flanked by the markers M_i and M_j and that H_{ij} denotes haplotype M_iM_j. Let A denote 'affected'. Then:

$$\begin{aligned}
P(H_{ij}, A) &= P(H_{ij}, D, A) + P(H_{ij}, d, A) \\
&= P_{iDj} P(A|D) + P_{idj} P(A|d) \\
&= P_{iDj}(f_{11}P_D + f_{12}P_d) + P_{idj}(f_{12}P_D + f_{22}P_d)
\end{aligned}$$

Thus, by Bayes formula, we have:

$$\begin{aligned}
P(H_{ij}|A) &= \frac{P(H_{ij}, A)}{P(A)} \\
&= \frac{1}{P(A)}[(f_{11}P_D + f_{12}P_d)P_{iDj} + (f_{12}P_D + f_{22}P_d)P_{idj}] \\
&= a_1 P_{iDj} + b_1 P_{idj}
\end{aligned}$$

where $a_1 = \frac{f_{11}P_D + f_{12}P_d}{P(A)}$ and $a_2 = \frac{f_{12}P_D + f_{22}P_d}{P(A)}$. Let N denote 'unaffected' or 'normal individual'. Similarly, we obtain:

$$P(H_{ij}|N) = b_1 P_{iDj} + b_2 P_{idj} \tag{4}$$

where $b_1 = \frac{(1\ \ f_{11})P_D + (1\ \ f_{12})P_d}{1\ \ P(A)}$ and $b_2 = \frac{(1\ \ f_{12})P_D + (1\ \ f_{22})P_d}{1\ \ P(A)}$. By the same argument as above, $P(H_{ijk}|A)$ can be derived.