Pacific Symposium on Biocomputing 5:672-681 (2000)

THE HAPLOTYPE LINKAGE DISEQUILIBRIUM TEST FOR GENOME-WIDE SCREENS: ITS POWER AND STUDY DESIGN

Momiao Xiong, Joshua Akey, Li Jin
Human Genetics Center, University of Texas - Houston,
P.O. Box 20334, Houston, TX 77030, USA
(mxiong, jakey, ljin)@utsph.sph.uth.tmc.edu

The focus of human genetics continues to shift toward the dissection of complex phenotypes. Integral to these endeavors is the development of powerful analytical tools. To this end, we propose a novel method designated the haplotype linkage disequibrium (LD) test for identifying diseases genes. The basic structure of the haplotype test statistic is a chi-square in which haplotypes, as opposed to individual marker data, are compared between cases and controls. Specifically, we performed power calculations and demonstrate that the use of haplotypes improves the power of mapping disease genes. We show that this approach can be used for initial genome-wide screens in mapping disease genes. Furthermore, we investigated the factors influencing statistical power of the method and discussed basic principals underlying study design. Published data from the Hereditary Hemochromatosis region was used to illustrate the utility of the haplotype test. Also discussed is its relationship with linkage disequibrium.

1 Introduction

Although linkage analysis has been successful in localizing disease genes involved in rare Mandelian diseases, it has been found to be of limited use when applied to complex diseases. It has been suggested that the availability of dense marker maps will make linkage disequilibrium mapping (LDM) the method of choice for mapping complex trait loci [1 2]. The transmission/disequilibrium test (TDT) is a LD based method which has been widely used for mapping disease genes [3]. A simple case-control design may offer an alternative to LDM for populations in the absence of substructure. Single-marker LD methods have been widely applied to fine-scale mapping of disease genes [4 5 6 7]. However, these two point LD methods fail to utilize nearby marker information. To overcome this problem, composite likelihood methods and two- or three-locus haplotype methods for localization of disease genes have been developed [7 8 9 10].

Recently, the proposition of extending the use of LD analysis from fine-scale mapping to genome-wide screens in order to delineate genes underlying complex traits has received considerable attention [26 12 13]. Technological advances and development of statistical methods are the primary impetus driving exploration of LD methods for genome-wide screens. The third generation ge-

netic map comprised of single nucleotide polymorphisms (SNPs) continues to be expeditionsly developed[14]. Their great abundance and accessibility to high-throughput low-cost automated genotyping techniques [15] [16] may lead to very dense genetic maps. With this technology in place LD methods for genome-wide screens may become feasible in young populations where the age of the disease mutation is 10∼50 generations resulting in persistent LD over sizable regions around disease genes. Escamilla et al. examined the feasibility of LD methods for mapping omplex traits and performed an initial scan for Biopolar Disorder oci on chromosome 18, genotyping markers spaced at 6cM intervals across chromosome 18 in 48 patients from Costa Rican families[2]. The results of their study suggest that LD methods will be useful for genome-wide screens.

Chapman and Wijsman investigated single-marker LD tests for genome-wide screens that use a case-control study design[12]. Service et al developed a likelihood ratio based haplotype LD test called the ancestral haplotype reconstruction (AHR) method for initial genome-wide screens[17]. AHR searches for regions in the genome of shared ancestry rather than individual alleles, thus simultaneously utilizing multiple marker information. Unfortunately, the AHR method is a parametric method and requires specification of a genetic model which is generally unknown for complex diseases. Furthermore, the AHR method needs to estimate a number of parameters that will increase exponentially, leading to a decrease in statistical power, as the number of loci generating haplotype increases. In this report, we describe a multi-locus haplotype LD test for genome-wide screen that uses a case-control study design. To evaluate the performance of this simple nonparametric multi-locus haplotype LD test, we develop analytic formula for calculating expected haplotype frequencies and noncentrality parameter of the distribution of the test statistic under the alternative hypothesis. We compare the power of the multi-locus haplotype LD test with that of single marker LD test and demonstrate that the muli-locus haplotype LD test is more powerful. Power calculations will also be performed to investigate how the power is influenced by the level of initial LD, genetic distance between marker and disease locus, age of disease mutation, mode of disease inheritance, and frequencies of associated marker allele and disease locus. Principles underlying study design are also discussed.

2 Test Statistic and Its Statistical Power

Test Statistic. We consider a case-control design for the haplotype test, in which haplotype frequencies of affected individuals (cases) are compared to unaffected individuals (controls). Suppose that the sample size for both cases and controls is n and k is the number of haplotypes. The haplotype frequency

data can be arranged in a $2 \times k$ contingency table. The null hypothesis H_0 to be tested is that the haplotype frequencies in the cases and controls are equal. A conventional χ^2 statistic for testing, H_0, can be defined as follows

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

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

As the number of markers studied increases, the number of possible haplotypes increases. To simplify the analysis and to ensure an appropriate number of counts in each cell in the contingency table, a grouping procedure can be employed. Specifically, the most frequent haplotype in cases is designated haplotype 1 and all others are grouped together as haplotype 2. Thus, the equality of haplotypes between cases and controls can be assessed by χ^2_{HT} with $\chi^2_{(1)}$ distribution. The selection of haplotype 1 requires special caution. One may also group several haplotypes into haplotype 1 and the remainder into haplotype 2. The grouping scheme, which will be studied elsewhere, is beyond the scope of this paper.

Power calculation. The statistical power to detect the disease gene, defined as the probability that a disease susceptibility locus can be detected, is an important index for evaluating the performance of any gene mapping method. Under the alternative hypothesis, H_a, where unequal haplotype frequency distributions in cases and controls are assumed, χ^2_{HT} is asymptotically a noncentral $\chi^2{}_{(k-1)}$ with the noncentrality parameter

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

where P_{Al} and P_{Cl} are the expected frequencies of the lth haplotype in cases and controls. Suppose that the critical value for an α test is $\chi^2_{k-1,1-\alpha}$. The asymptotic power of the test with α-level significance is given by $\beta = P_{H_a}(\chi^2_{HT} \geq \chi^2_{k-1,1-\alpha})$. To calculate the power (β) we need first to calculate the noncentrality parameter and the expected haplotype frequencies based on a multi-locus population genetics model.

For ease of exposition, we only discuss the two-locus case. Extension to multilocus haplotype frequencies are straightfoward but more complicated and will be presented elsewhere. In the population genetics model discussed below, 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; and (4) the population is isolated and homogeneous. A recent disease mutation can be introduced into the population either by spontaneous mutation or by immigration of individuals carrying the mutation t generations ago. The time t is referred to as the age of the mutation.

Now we consider two-locus haplotypes. For the convenience of presentation we only consider the ordering: marker A-disease locus-marker B. The other possible orders can be dealt with in a similar fashion. Two alleles are assumed at the disease locus: disease allele D with frequency P_D and normal allele d with frequency P_d. Let $\theta_{1,2}$ be the recombination fraction between the markers M_1 and M_2. Let $M_{i_1}M_{i_2}$ be the haplotype produced by the i_1-th allele at the marker M_1 and the i_2-th allele at the marker M_2. We denote the frequency of the haplotype $M_{i_1}M_{i_2}$ by $P_{i_1 i_2}$. Recently, Zheng and Elston investigated haplotype frequencies in admixed populations and derived recursive and deterministic formulas for their calculation [18]. Here, we derive explicit formula for the expected two-locus haplotype frequency under a stochastic population genetics model. The proof and the extension to a general k-locus haplotype will be presented elsewhere.

Let $\delta_{i_1 D i_2}(0)$ be the coefficient of initial LD at the three loci $M_1 D M_2$ at the occurrence of disease mutation. Then, the expectation of the frequency of haplotype $M_{i_1}DM_{i_2}$ is

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

where $\delta_{i_1,i_2}(t)$ and $\delta_{i_1 D i_2}$ are the two-locus and three-locus coefficients of LD defined as [19] $\delta_{i_1 D i_2}(t) = P_{i_1 D i_2}(t) - P_{i_1}\delta_{D i_2}(t) - P_{i_2}\delta_{i_1 D}(t) - P_{i_1}P_D P_{i_2}$, $\delta_{i_1 i_2}(t) = P_{i_1 i_2}(t) - P_{i_1}P_{i_2},$, and P_i is the frequecy of allele i which is assumed to be a constant over time. This fomula is equivalent to Equation (2.1) in Service et al. [17]. It should be noted that as the age of the mutation (t) increases the haplotype frequencies will converge to their equilibrium frequency due to the attenuation of LD.

To calculate the noncentrality parameter λ we must obtain the expected frequencies of the haplotype in affected and unaffected individuals. Let f_{11}, f_{12} and f_{22} be the penetrance of genotypes DD, Dd and dd, respectively, with $f_{11} \geq f_{12} \geq f_{22} \geq 0$ for 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 $f_{22} = 0$) cases ($0 < x \leq 1$), respectively. 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 $a_1 = \frac{f_{11}P_D + f_{12}P_d}{P(\text{affected})}$, $a_2 = \frac{f_{12}P_D + f_{22}P_d}{P(\text{Affected})}$, $b_1 = \frac{(1-f_{11})P_D + (1-f_{12}P_d}{1-P(\text{Affected})}$, and $b_2 = \frac{(1-f_{12})P_D + (1-f_{22})P_d}{1-P(\text{Affected})}$. Let H_{ij} denote the haplotypes $M_{i_1}M_{j_2}$. We can show that

$$P(H_{ij}|\text{Affected}) = a_1 P_{iDj} + a_2 P_{idj}$$

$$P(H_{ij}|\text{Unaffected}) = b_1 P_{iDj} + b_2 P_{idj}$$

From the above formula we can see that each of the haplotype frequencies in affected and unaffected individuals will be a weighted average of P_{iDj} and P_{idj} with the weights determined by the mode of disease inheritance. For a single locus, the test statistic using marker allele data in case-control designs, χ^2_M, is the same as χ^2_{HT} except $\hat{P}_{Al}$ and $\hat{P}_{Cl}$ are replaced by observed marker allele frequencies in the cases and controls, respectively [12].

3 Factors Influencing Statistical Power

From the previous section we show that the power of the χ^2_{HT} depends on a number of parameters such as the initial LD, the age of mutation, recombination fraction between the marker and disease loci, the mode of disease inheritance, haplotype frequencies and disease allele frequency. In this section, we study the impact of those parameters on the statistical power of the haplotype LD test.

Initial LD. We refer to the haplotype in which the disease causing mutation(s) occured as the associated haplotype (or associated allele if one single marker is used), and we will designate it haplotype 1 or allele 1 hereafter. It should be noted that the initial level of LD is determined by the difference of two parameters: the frequency of the associated haplotype in the population ($P_1(0)$), and that in the affected population ($P_{1D}(0)$).

If there is only one disease causing mutation in the population introduced by mutation or immigration, then $P_1(0) = P_1$ and $P_{1D}(0) = 1$. However, multiple disease causing mutations may exist in the population, which can result from mutations and/or immigation of individuals carrying mutations. We will focus on the most recent mutation (MRM) for LD based mapping since only the MRM will maintain relatively strong LD with nearby markers. Therefore, $P_1(0)$ will remain to be P_1, but $P_{1D}(0) \leq 1$. A reduction of intial LD will lead to a reduction of the statistical power of the LD methods.

In Figure 1(a) and 1(b), we show the effect of $P_1(0)$ and $P_{1D}(0)$ on the statistical power of the test statistic χ^2_{HT}. We assume that $t = 15$ in the calculation. Without loss of generality, the following parameters were assumed: the sample size (n) is 200, the disease allele frequency (P_D) is 0.1, and D is

dominant over d. In Figure 1(a), $P_{1D}(0) = 1$. In Figure 1(b), $P_1(0) = 0.5$. From Figure 1, one can conclude that a smaller $P_1(0)$ and/or a greater $P_{1D}(0)$ confer higher statistical power. This is consistent with our previous observation made for single marker based LD analysis[20]. This conclusion is robust for other parameters as well (data not shown).

Age of the mutation (t). In Eq 2, we have shown that a smaller t results in stronger LD. It should be noted that t refers to the age of the MRM. In Figure 2, we show the effect of t on the power of χ^2_{HT}. Again we assume that $n = 200$, $P_D = 0.1$, $P_1(0) = 1$, $P_{1D}(0) = 1$, and D is dominant over d. From Figure 2, one can conclude that a younger mutation is associated with higher statistical power, which is consistent with the observation we previously made for single marker LD analysis[20].

Genetic distance between the markers and trait locus. The genetic distance between the markers and trait locus is measured by the recombination fraction (θ). The level of LD as well as the power of the LD based statistics decrease when θ increases. This property is manifested in Figure 1 and Figure 2. The length of the distance (or the size of the fragment) where a certain significant power can be reached, say 80%, will be simply referred to as the extent of detectable LD. The extent of detectable LD is therefore statistic-dependent. A more powerful method results in a larger extent given everything else being identical. It should be indicated that for the current case-control study, θ and t are presented as a composite parameter θt in power determination.

Mode of inheritance and penetrance. For single gene disorders, a recessive trait renders a higher statistical power than a dominant trait in LD mapping. Table 1 and Figure 1 provide evidence for this assertion. This is consistent with the previously made observation for single marker based LD analysis[20]. For complex traits such as the one described in Risch and Merikangas[1] where the complexity is reflected by the level of penetrance, a drastic reduction of statistical power is expected (see Table 2).

Frequency of disease allele P_D. The statistical power to detect disease mutations increase with the frequency of disease allele for LD based mapping approaches. Table 2 shows such a relationship between the P_D and the statistical power of χ^2_{HT} for complex traits following Risch and Merikangas[1]. This is again in concordance with the result for single marker based LD analysis[20].

Comparison of haplotype based methods and single marker based methods. Haplotype methods can be contrasted with single marker methods by comparing the power of χ^2_{HT} with that of χ^2_M under various parameters. We assume that the disease gene is located in the mid point between the two markers M_1 and M_2. Figure 3 shows the power of the χ^2_{HT} for two-locus haplotype (H_2) and χ^2_M for single marker (M) with $\alpha = 0.001$, for recessive, dominant, and

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

	4cM			5cM			10cM		
	t=10	t=20	t=50	t=10	t=20	t=50	t=10	t=20	t=50
Recessive									
$\chi^2_{HT(2)}$	15	23	77	17	29	127	29	77	1544
χ^2_M	22	34	122	24	43	204	43	122	2530
Dominent									
$\chi^2_{HT(2)}$	42	66	227	47	81	376	81	227	4602
χ^2_M	95	144	485	106	176	802	176	485	9812

Table 2: Number of cases required to achieve 80% power with a significance level $\alpha = 0.001$ for complex traits

test	$P_D = 0.01$		$P_D = 0.1$		$P_D = 0.5$	
statistic	t=10	t=20	t=10	t=20	t=10	t=20
$\gamma^2 = 4$						
χ^2_M	20,010	22,120	297	328	26	28
$\chi^2_{HT}(2)$	8,797	10,003	150	170	14	16
$\chi^2_{HT}(3)$	5,591	7,020	117	142	12	15
$\gamma^2 = 2$						
χ^2_M	140,194	154,990	1,558	1,722	78	86
$\chi^2_{HT}(2)$	60,564	68,950	712	807	39	44
$\chi^2_{HT}(3)$	37,340	47,159	481	597	29	36

complex disease, as a function of recombination fraction (θ_1) between the two markers. We also assumed that $n = 200$, $P_D = 0.1$, $P_1(0) = 0.5$, $P_{1D}(0) = 1$, and $t = 15$. For a complex trait, we assumed $\gamma = 4$, where the genotypic relative risk for individuals of genotype Dd and $DD$is γ and γ^2 times greater than that for individuals with genotype dd[1]. In all three models, the statistical power of χ^2_{HT} is consistently higher than that of χ^2_M, indicating that haplotype based LD analyses is generally superior to single marker based LD analyses.

The statistical power of the two approaches was further investigated by examining various combinations of parameters including t, P_D, θ, γ, and the mode of inheritance (Table 1 for single gene diseases, and Table 2 for complex traits). In the Tables, the statistical power is presented as the number of samples required to achieve 80% power with a significance level of $\alpha = 0.0001$. The reduction of sample size from a single marker approach to a two-locus

haplotype approach varies with the set of parameters chosen. However, a reduction of at least 30% in sample size is generally encountered.
Number of markers. Increasing the number of markers in a haplotype based LD analysis does not guarantee an increase in power (data not shown). The grouping strategy is beyond the scope of this article and will be presented somewhere else.

4 Study Design

For population based mapping projects, three items can be manipulated by the researcher: selection of population, density of the markers, and sample size. Mode of inheritance and penetrance have to be accepted *a priori* and are generally unknown. Initial LD, age of mutations, and frequency of disease alleles are generally population specific. However, a judicious choice of populations will result in a significant gain of statistical power. For example, a heterogeneous population such as the US population is not an ideal choice for LD mapping[13]. The best scenario is to study an isolated population that was recently founded by a very small number of individuals as a mutation recently introduced into this population generates strong initial LD over a large chromosomal region. Table 1 and 2 showed that a reasonable power can be achieved to detect the LD within 4 - 5 cM with relatively small sample sizes in isolated populations with relatively young mutations. Populations with young disease mutations and without significant gene flow from other populations exist, for example, French Canadians[23], Ashkenazi Jews[24], Costa Rica[2], certain Chinese populations[22], and the Amish[25]. Moreover, the density of the markers is crucial to ensure a reasonable power by the extent of detectable LD, which itself is determined by the demographic history of the population of interest (i.e. initial LD, age of mutations, frequency of disease alleles, and the penetrance of the disease causing loci). Sample size can then be determined when information of regarding the population and marker density are given.

5 An Example

The haplotype LD test was applied to map the HFE gene[7]. Markers D6S265(M1), HLA-A (M2), HLA-F (M3), D6S258 (M4), D6S306 (M5), D6S105 (M6), D6S464 (M7), and D6S1260 (M8) lie approximatelly 3.9, 3.8, 3.7, 1.9, 1.7, 1.8, 1.7 and 1cM, respectively, centromeric to the HFE gene. The χ^2 values for those markers are 16.1, 18.2, 10.3, 11.7, 5.6, 18.7, 8.3, and 5.7, respectively. Table 3 presents the results for multi-locus haplotype LD test. We observe several remarkable features. First, the markers D6S1260 and D6S464 are closer to the

Table 3: Haplotype Test results for individual marker alleles.

H	M1	M2	M3	M4	M5	M6	M7	M8	P_{cases}	P_{cont}	χ^2
A	1[a]	3	2	4	3	8	6	4	17	0	17.1
B	-	3	2	4	3	8	6	4	17	0	17.1
C	-	-	2	4	3	8	6	4	20	0	20.2
D	-	-	-	4	3	8	6	4	21	0	21.8
E	-	-	-	-	3	8	6	4	24	0	24.9
F	-	-	-	-	-	8	6	4	26	0	28.6
G	-	-	-	-	-	-	6	4	39	12	17.6

a: represents the most frequently occuring marker allele in cases.

HFE gene than other markers. However, the p-value of the single-marker LD test is the smallest at marker D6S265, which is 3.9cM away from the HFE gene. Second, the p-value of the two-locus haplotype LD test is 0.000027, much smaller than the two smallest p-values observed at D6S265 (0.017) and D6S424 (0.004) using single-marker LD test. The haplotype LD test offered a significant improvement over the single-marker LD test. Third, the values of the single-marker LD test statistic at different markers oscillate in punctuate waves. Even if markers are clustered within small regions some may not show strong LD due to uncertainty of initial LD between the markers and disease loci, random drift, and mutation at the marker locus. The haplotype LD test has the valuable property of a "smoothing effect" such that often-incongruent patterns of single-marker LD test become more interpretable (see Table 3).

6 Conclusion

In this article, we presented a case-control design for a haplotype LD test. We stuided the factors affecting the statistical power of the method in detecting LD with genes underlying simple and complex diseases. Finally, we showed that this method can be used for genome-wide screens in young isolated populations. It should be indicated that the rate-limiting step in implementing the haplotype LD test is construction of haplotypes when parental genotypes are unavailable. New analytical approaches have been and continue to be developed making this a potentially surmountable challenge in the future [27].

References

1. N. Risch and K. Merikangas, *Science* **273**, 1516 (1996).

2. M.A. Escamilla *et al*, *Am. J. Hum. Genet.* **64**, 1670 (1999).
3. R.S. Spielman *et al*, *Am. J. Hum. Genet.* **52**, 506 (1993).
4. T.K. Cox *et al*, *Am. J. Hum. Genet.* **45**, A13 (1989).
5. J. Hästbacka *et al*, *Nat. Genet.* **2**, 204 (1992).
6. N.L. Kaplan *et al*, *Am. J. Hum. Genet.* **56**, 18 (1995).
7. B. Devlin and N. Risch, *Genomics* **29**, 311 (1995).
8. J.D. Terwilliger, *Am. J. Hum. Genet.* **56**, 777 (1995).
9. M. Xiong and S. Guo, *Am. J. Hum. Genet.* **60**, 1513 (1997).
10. L.C. Lazzeroni, *Am. J. Hum. Genet.* **62**, 159 (1998).
11. J.D. Terwilliger and K.M. Weiss, *Cur. Opin. Biotechnol.* **9**, 578 (1998).
12. N.H. Chapman and E.M. Wijsman, *Am. J. Hum. Genet.* **63**, 1872 (1998).
13. L. Kruglyak, *Nat. Genet.* **22**, 139 (1999).
14. D.G. Wang *et al*, *Science* **280**, 1077 (1998).
15. G. Ramsay, *Nat. Biotech.* **16**, 40 (1998).
16. T.J. Griffiin *et al Proc. Natl. Acad. Sci. USA* **96**, 6301 (1990).
17. S.K. Service *et al*, *Am. J. Hum. Genet.* **64**, 1728 (1999).
18. Zheng and Elston, *Genet. Epidemiol.* , (1999).
19. B. Weir, Genetic Data Analysis II , Sinauer, Massachusetts (1996).
20. M. Xiong and L. Jin, *Am. J. Hum. Genet.* , (1999).
21. A. de la Chapelle and F.A. Wright, *Proc. Natl. Acad. Sci. USA* **95**, 12416 (1998).
22. J. Xiao *et al*, *Proc. Natl. Acad. Sci. USA* , (In Press) (1999).
23. Casaubon *et al*, *Am. J. Hum. Genet.* **58**, 28 (1996).
24. N. Risch *et al*, *Nat. Genet.* **9**, 152 (1995).
25. Sulisalo *et al*, *Am. J. Hum. Genet.* **55**, 937 (1994).
26. W. Thomas *et al*, *Hum. Genet.* **102**, 517 (1998).
27. R.B. Martin *et al*, *Genet. Epidemiol.* **15**, 471 (1998).