**[Frontiers in Bioscience 8, a85-93, May 1, 2003]**

# HAPLOTYPE BOLCK LINKAGE DISEQUILIBRIUM MAPPING

**Momiao Xiong [1], Jinying Zhao[1], Eric Boerwinkle [1,2]**

*[1] Human Genetics Center, University of Texas - Houston Health Science Center, Houston, Texas, [2] Institute of Molecular Medicine, University of Texas -Houston Health Science Center, Houston, Texas*

## TABLE OF CONTENTS

## 1. ABSTRACT

Linkage disequilibrium (LD) mapping is emerging as a powerful alternative approach to identifying genes for complex disease. However, the feasibility and success of LD mapping depend largely on the extent and pattern of LD. Erratic pattern of pair-wise LD seriously compromises LD mapping. Recently discovered haplotype block structure dramatically alleviates the irregular pattern of LD and holds the promise for mapping complex disease genes. To facilitate applications of the haplotype block LD mapping, in this report we conduct theoretical analysis for haplotype block LD mapping. We present an overall LD measure of the haplotype to quantify the LD level of the haplotype block, between the haplotype blocks, and between the haplotype block and the marker locus. Most theoretical and empirical studies of the extent of LD and evaluation of the power of LD mapping have focused on pair-wise LD and single marker LD mapping. There is a lack of systematic and integrative analysis for the haplotype block LD mapping. In this report, we develop population genetic models of the haplotype blocks and analytic tools for calculation of noncentrality parameter of the statistic for the haplotype block LD mapping. We evaluate the impact of the population parameters and disease models on the power of the haplotype block LD mapping in the hope to improve its study design. We compare the powers of the single marker LD and haplotype block LD mapping. Haplotype block structure is an important discovery. Our preliminary results of theoretic analysis further demonstrate that the haplotype block LD analysis is a breakthrough in LD mapping and is a promising tool for genome-wide association studies.

## 2. INTRODUCTION

As a dense set of single nucleotide polymorphisms (SNPs) markers becomes increasingly available, linkage disequilibrium (LD) mapping is emerging as a powerful tool for fine mapping of disease susceptibility genes and genome-wide association studies (1). The extent and pattern of LD have been debated for several years (1, 2). Many evolutionary forces such as mutation, genetic drift, selection, recombination, and population bottleneck affect the pattern of LD (3, 4). It is now widely accepted that the pattern of pair-wise LD is erratic (5-7). The relationship between the level of pair-wise LD and distance between two individual markers is not monotonic, which complicates LD mapping.

In the past several years, there have been growing interests in haplotype and haplotype block LD mapping to alleviate the problem of erratic patterns of pair-wise LD (8-18). A haplotype block shows a largely atomistic pattern and island structure of LD, which greatly simplifies association analyses (3).

To facilitate application of haplotype block LD mapping to real data, several issues need to be more-fully addressed. First, there are multiple definitions of haplotype blocks that are not consistent. Second, most investigations of the extent of LD and the power of LD mapping have focused on pair-wise LD and single marker LD mapping. There is a lack of systematic studies of LD mapping that consider haplotypes spanning many marker loci. Third, there is growing consensus (3, 12) that multi-allelic extensions of the usual pair-wise LD measure can be useful for defining blocks and localizing disease susceptibility genes. However, simple multi-allelic extension of pair-wise LD measures cannot completely summarize the LD at multiple loci.

To address the above issues, this report will first present a definition of haplotype blocks which is simple and useful for mapping disease susceptibility genes. Then, we will propose an overall measure of LD for haplotypes, and between a haplotype block and the marker locus. In order to evaluate the power of the haplotype block LD mapping, we will consider population genetic models for haplotype blocks. Finally, we will evaluate the power of the haplotype block LD mapping and the impact of various factors on the power.

## 3. MODELS AND METHODS

### 3.1. Haplotype block characterization

Three methods have been proposed to define haplotype blocks. The first method defines a block as a region in which "LD decays slowly with distance or not all" (3). Unfortunately, the sampling variation of LD shows considerable fluctuation so that analyses of any trends are made difficult. The second method defines blocks through the optimal partition of a chromosome into a minimum number of blocks and minimum number of representative SNPs (18, 19). However, there has been no clear biological interpretation of such partitioned haplotype blocks. The third method for defining haplotype block is based on recombination. A haplotype block is defined as a region in which there are no recombination events evidenced in the study sample (12). As we will discuss below, this definition of haplotype block is useful for association studies and will be the definition used here.

### 3.2. Measure of haplotype block LD

A number of statistics have been proposed to measure pair-wise LD or high order LD (6, 20, 21). One popular pair-wise LD measure is given by

$D_{ij} = h_{ij} - p_i q_j$

where $h_{ij}$ denotes the population frequency of the haplotype $A_i B_j$, while $p_i$ and $q_j$ are population frequencies of the alleles $A_i$ and $B_j$, respectively.

The measure of LD, $D_{ij}$ can be extended to a haplotype block. Suppose that there are $k$ loci within a block. Assume that two alleles $A_1$ and $A_2$ at each locus have frequencies $P_{A_1}$ and $P_{A_2}$, respectively. Consider a $k$ locus haplotype $H_{j_1 j_2 \cdots j_k}$ with a sequence of alleles $A_{j_1}, A_{j_2}, \ldots, A_{j_k}$, where $A_{j_i}$ at the i-th locus is either $A_1$ or $A_2$. Let $P_{H_{j_1 j_2 \cdots j_k}}$ be the population frequency of the haplotype $H_{j_1 j_2 \cdots j_k}$. An overall measure of the haplotype LD at the $k$ loci is defined as

$\delta_{H_{j_1 j_2 \cdots j_k}} = P_{H_{j_1 j_2 \cdots j_k}} - P_{A_{j_1}} P_{A_{j_2}} \ldots P_{A_{jk}}$ .

This overall measure of the haplotype LD includes the pair-wise LD measures and higher order measures.

Such an overall measure can be applied to measuring LD between a haplotype block and a marker locus. Consider a haplotype $H_{j_1 j_2 \cdots j_k}$ consisting of alleles $A_{j_1} A_{j_2} \ldots A_{j_k}$ in the block and an allele $M_1$ at the marker locus which is outside the block. The haplotype $H_{j_1 j_2 \cdots j_k}$ and the allele $M_1$ form a ($k$+$1$) locus-haplotype $H_{j_1 j_2 \cdots j_k M_1}$. The overall measure of LD for the haplotype $H_{j_1 j_2 \cdots j_k M_1}$ can be used to measure LD between the haplotype $H_{j_1 j_2 \cdots j_k}$ and the marker allele $M_1$ and will be denoted by $\delta_{HM}$. Some authors suggested that the haplotype block be treated as alleles and multi-allelic analysis for the single marker be applied to the haplotype block analysis (12). Following this line, the measure of LD between a haplotype block and the marker locus, denoted by $D_{HM}$, can be defined as

$$\begin{aligned} D_{HM} &= P_{H_{j_1 j_2 \cdots j_k M_1}} - P_{H_{j_1 j_2 \cdots j_k}} P_{M_1} \\ &= P_{H_{j_1 j_2 \cdots j_k M_1}} - (\delta_{H_{j_1 j_2 \cdots j_k}} + P_{A_{j_1}} P_{A_{j_2}} \ldots P_{A_{jk}}) P_{M_1} \\ &= P_{H_{j_1 j_2 \cdots j_k M_1}} - P_{A_{j_1}} P_{A_{j_2}} \ldots P_{A_{jk}} P_{M_1} - P_{M_1} \delta_{H_{j_1 j_2 \cdots j_k}} \\ &= \delta_{HM} - P_{M_1} \delta_{H_{j_1 j_2 \cdots j_k}} \end{aligned}$$

This measure of LD between a haplotype and a marker is obtained by removing the haplotype block LD from the LD measure between the haplotype and the marker locus.

### 3.3. Test statistic and power calculation

Suppose that $n$ affected individuals and $n$ unaffected individuals are sampled. Assume that there are $l$ haplotypes in the blocks. The haplotype frequency data can be arranged in a 2×$l$ contingency table. The null hypothesis $H_0$ to be tested is that of equal haplotype frequencies in affected and unaffected individuals. A traditional $\chi^2$ statistic for testing $H_0$ is given by

$$\chi^2_{HB} = 2n \sum_{j=1}^{l} \frac{(\hat{h}_{P_j} - \hat{h}_{N_j})^2}{(\hat{h}_{P_j} + \hat{h}_{N_j})} \qquad (1)$$

where $\hat{h}_{P_j}$ and $\hat{h}_{N_j}$ are the observed frequencies of the j-th haplotype in the block in affected and unaffected samples, respectively. It is well known that under the null hypothesis, $\chi^2_{HB}$ is asymptotically distributed as $\chi^2_{l-1}$.

To evaluate its power, we consider the distribution of $\chi^2_{HB}$ under the alternative hypothesis. Under the alternative hypothesis, $H_a$, of unequal haplotype frequencies in the affected and unaffected populations, $\chi^2_{HB}$ is asymptotically distributed as a noncentral $\chi^2_{l-1}$ with noncentrality parameter:

$$\lambda = 2n\sum_{j=1}^{k}\frac{[P(H_j \mid A) - P(H_j \mid N)]^2}{P(H_j \mid A) + P(H_j \mid N)} \qquad (2)$$

where $P(H_j \mid A)$ and $P(H_j \mid N)$ are the expected frequencies of the j-th haplotype in the affected and unaffected populations.

Consider a disease locus having alleles D and d with allele frequencies $P_D$ and $P_d$, respectively. Let $f_{11}$, $f_{12}$ and $f_{22}$ be the penetrance of the genotypes DD, Dd and dd, respectively. With $f_{11} \geq f_{12} \geq f_{22} \geq 0$, the probability of a random individual being affected is given by $P(A) = f_{11}P_D^2 + 2f_{12}P_DP_d + f_{22}P_d^2$.

Consider three relative positions of the disease locus with respect to a haplotype block or blocks: (i) the disease locus is in a block; (ii) the disease locus is outside of a block and (iii) the disease gene is located between two blocks. For convenience, the combined haplotype of the original haplotype $H_j$ and the alleles D or d at the disease locus is denoted by $(H_j, D)$ or $(H_j, d)$ in all three cases without indicating the relative position of the disease locus with respect to the haplotype blocks. Let

$$a_1 = \frac{f_{11}P_D + f_{12}P_d}{P(A)}, \qquad a_2 = \frac{f_{12}P_D + f_{22}P_d}{P(A)}$$

$$b_1 = \frac{(1-f_{11})P_D + (1-f_{12})P_d}{1-P(A)}, \qquad b_2 = \frac{(1-f_{12})P_D + (1-f_{22})P_d}{1-P(A)}$$

It can be shown that (Appendix A) the noncentrality parameter λ given in equation (2) can be rewritten

$$\lambda = 2n\sum_{j=1}^{k}\frac{[(a_1-b_1)\delta_{(H_j,D)} + (a_2-b_2)\delta_{(H_j,d)}]^2}{(a_1+b_1)\delta_{(H_j,D)} + (a_2+b_2)\delta_{(H_j,d)} + 2P(A_{j_1})...P(A_{j_k})} \qquad (3)$$

where $\delta_{(H_j,D)}$ and $\delta_{(H_j,d)}$ denote the overall measure of LD of the joint haplotype $(A_{j_1}...A_{j_k}, D)$ and $(A_{j_1}...A_{j_k}, d)$, respectively.

### 3.4. Population genetic model of haplotype block

The pattern of haplotype block LD involves history of populations. It will be useful to study the population genetic models of haplotype block LD. For the convenience of presentation, we assume that: (1) mating is random in the population; (2) generations are non-overlapping; (3) there are no phenocopies; and (4) the population is isolated. We assume that the population was formed t generations ago. For the simplicity of presentation, we further assume that the mutations within the block can be neglected. Therefore, the frequencies of the haplotypes within the blocks are assumed to be constant.

Consider two haplotype blocks, and the haplotype A in the first block and the haplotype B in the second block. Let θ be the recombination fraction between two blocks. Let $P_A$ and $P_B$ be the frequency of the haplotypes *A* and *B*, respectively. Due to the recombination, the frequencies of the joint haplotype *AB* changes over generations and hence is a function of the time. By the same argument, the measure of LD between the haplotype A and the haplotype B is also a function of the time. Therefore, the frequency of the joint haplotype AB and the measure of LD between the haplotypes A and B are denoted by $P_{AB}(t)$ and $\delta_{H_{AB}}(t)$, respectively. Since the overall LD measures of the haplotypes *A* and *B* involve only mutations and we assume that the mutation within the blocks are neglected, the overall LD measures of the haplotypes A and B are constant and denoted by $\delta_{H_A}$ and $\delta_{H_B}$, respectively. It can be shown that, on the average, the LD measure $\delta_{H_{AB}(t)}$ between the haplotypes *A* and *B* is given by (Appendix B)

$$\delta_{H_{AB}}(t) = (1-\theta)^t\delta_{H_{AB}}(0) + [1-(1-\theta)^t](\delta_{H_A}\delta_{H_B} + P_{A_{j1}...A_{jk}}\delta_{H_B} + P_{B_{j1}}...P_{B_{jk}}\delta_{H_A})$$

where $\delta_{H_{AB}}(0)$ is the initial measure of LD between the haplotypes *A* and *B*.
If the haplotype *B* is degenerated into the marker allele *B*, then $\delta_{H_{AB}}(t)$ is reduced to

$$\delta_{H_{AB}}(t) = (1-\theta)^t\delta_{H_{AB}}(0) + [1-(1-\theta)^t]P_B\delta_{H_A}.$$

If both the haplotypes *A* and *B* are degenerated into the marker alleles *A* and *B*, then

$$\delta_{H_{AB}}(t) = (1-\theta)^t\delta_{AB}(0) = (1-\theta)^t D_{AB}(0),$$

where $D_{AB}(0)$ denotes the traditional initial measure of LD between two marker loci. Therefore, when two haplotypes are degenerated into the markers, the measure of LD between two haplotype blocks based on the overall LD measure of the haplotype is reduced to the traditional measure of LD between two markers. It shows that the traditional measure of LD, *D,* is a special case of the proposed measure of LD between the blocks.

Now we study how to calculate the expectation of the joint haplotypes across several blocks. Suppose that we have *k* haplotype blocks. Let $I_j$ be the number of haplotypes at the j-th block (j=1, 2,…, k) and $\theta_{i,j}$ be the recombination fraction between the i-th block and the j-th block. Let $H_{i_1}H_{i_2}...H_{i_k}$ be the joint haplotype produced

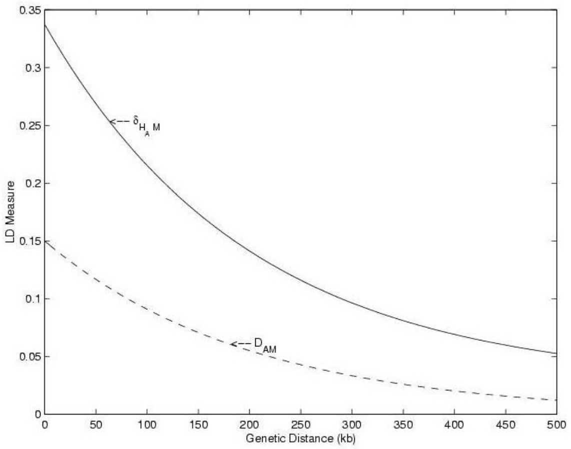

**Figure 1**. New measure of LD, $\delta_{H_A M}$ , between a haplotype and a marker, and the traditional measure of LD, $D_{AM}$ , as a function of the genetic distance between the haplotype block and the marker locus, assuming each haplotype frequency and each allele frequency to be equal to 0.5, the size of the block is equal to 26 kb, and the distance between the blocks is equal to 5kb, sample size n=500, and t = 500 generations.

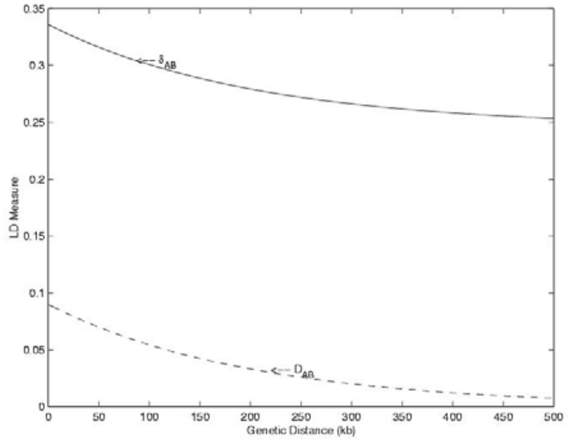

**Figure 2**. The LD measures $\delta_{AB}$ and D$_{AB}$ between the haplotype blocks as a function of the genetic distance between the haplotype blocks, assuming that the four initial joint haplotype frequencies are 0.4, 0.1, 0.4 and 0.1, respectively. Other parameters are assumed to be the same as that of Figure 1.

by the i$_1$-th haplotype in the block 1, the i$_2$-th haplotype in the block 2, and the i$_k$-th haplotype in the block *k*. The frequency of the haplotype $H_{i_1}H_{i_2}...H_{i_k}$ is given by $P_{H_{i_1}\cdots i_k}$ and the frequency of the haplotype $H_{i_j}$ is denoted by $P_{H_{i_j}}$ . The expected frequency of the joint haplotype is given in Appendix C. The general formula for the joint haplotype frequencies is complicated. However, as we can see, the formula for the frequency of the joint haplotype in two blocks is quite simple. Let $P_{H_{i_1 i_2}(0)}$ be the initial frequency of the haplotype $H_{i_1}H_{i_2}$ at *t=0* generation and $\delta_{H_{i_1 i_2}}(0) = P_{H_{i_1 i_2}}(0) - P_{H_{i_1}}P_{H_{i_2}}$ . Then, on the average, the frequency $P_{H_{i_1 i_2}}$ of the haplotype $H_{i_1}H_{i_2}$ is given by

$$P_{H_{i_1 i_2}}(t) = D_{H_{i_1 i_2}}(0)e^{-\theta_1 t} + P_{H_{i_1}}P_{H_{i_2}}$$

# 4. RESULTS

## 4.1. Haplotype block LD measure

In this report we propose an overall measure of LD for the haplotypes defining block. We now compare the LD measure $\delta_{H_A M}$ between a haplotype block and a marker locus based on the overall LD measure with the traditional LD measure $D_{H_A M}$ between a haplotype block and a marker locus. Figure 1 shows the measures $\delta_{H_A M}$ and $D_{H_A M}$ as a function of the genetic distance between the haplotype block and the marker locus. For the simplicity of the presentation, we assume there are two haplotypes in the block. Figure 1 clearly demonstrates that the new LD measure, $\delta_{H_A M}$ , is much larger than the traditional measure $D_{H_A M}$ .

Next we compute the new measure of LD, $\delta_{AB}$ , and the traditional measure of LD, D$_{AB,}$ between the two haplotype blocks. The traditional measure of LD between the two blocks is defined as the pair-wise measure of LD between two loci if the multiple haplotypes are treated as multiple alleles. Let $H_{A_j}$ and $H_{B_k}$ be the j-th haplotype in the first block and the k-th haplotype in the second block, respectively. Then, the measures of LD, $\delta_{AB}$ and D$_{AB,}$ are defined as

$$\delta_{AB} = \sum_j \sum_k P(H_{A_j}H_{B_k})\delta_{H_{A_j}H_{B_k}}$$

$$D_{AB} = \sum_j \sum_k P(H_{A_j}H_{B_k})D_{H_{A_j}H_{B_k}}$$

where $P(H_{A_j}H_{B_k})$ is the frequency of the joint haplotype.

Figure 2 shows the measures of LD, $\delta_{AB}$ and $D_{AB}$ , between the two haplotype blocks as a function of the genetic distance between the two haplotype blocks. We can clearly see that the measure of LD, $\delta_{AB}$ , is much larger than the measure of LD, $D_{AB}$ , although both measures decrease as the genetic distance between the two haplotype blocks increases.

## 4.2. Power of the haplotype block LD mapping

Interest in LD is due largely to the role which genome-wide association studies may play in mapping complex disease genes. However, the prospect of genome association mapping depends on the pattern of LD and the density of the markers. Figure 3 shows the power of haplotype block LD mapping and single marker LD

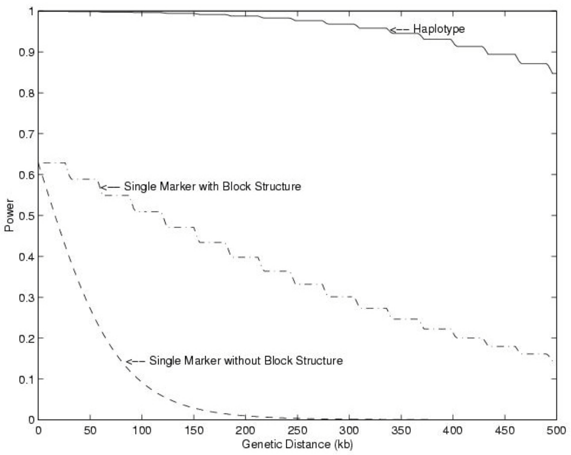

**Figure 3**. Power of haplotype block LD mapping and single marker LD mapping in the presence or absence of haplotype block structure as a function of the genetic distance between the region of the interest and disease locus, assuming the distance between blocks = 5 kb, two haplotypes with equal frequencies, two alleles at the marker locus with equal frequencies, the disease allele frequency $P_D$ =0.1, n=500, t=500 generations.

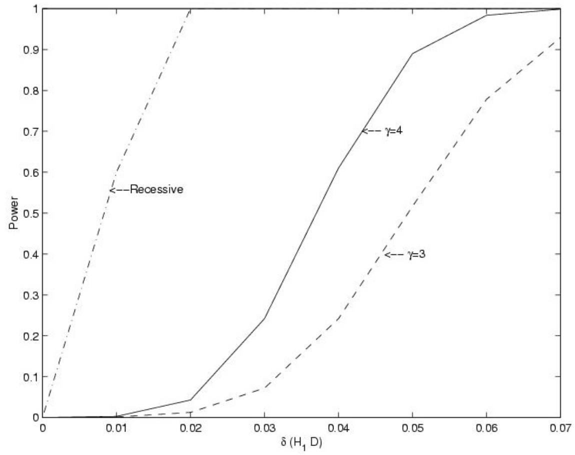

**Figure 4**. Power of haplotype block LD mapping as a function of the LD measure between the haplotype block and the disease locus under three disease models: recessive disease, genotype risk disease models with γ =3, and γ = 4. Assume that two marker loci having two alleles with equal frequencies span four haplotypes with $P(H_2) = P(H_3) = P(H_4) = 0.025$, the disease allele frequency is equal to 0.1, n=500, t=500 generations.

mapping in the presence and absence of haplotype block structure. For the simplicity, the size of each block is assumed to be equal to the average size 26kb of the block on chromosome 21 (19) and the distance between the neighboring blocks is assumed to be equal to 5kb (12). We assume the genotype relative risk disease model (22), where the genotype relative risk for individuals of genotype Dd and DD is γ and $\gamma^2$ times greater than that of individuals with genotype dd. Figure 3 shows that power of the haplotype block LD mapping is much higher than that of single marker LD mapping in the presence and the absence of any haplotype block structure. Figure 3 also indicates that the power of single maker LD mapping in the presence of block structure decreases much slower than that of single marker LD mapping in the absence of block structures as the genetic distance increases.

The LD between a haplotype block and a disease locus apparently influences the power of haplotype block LD mapping. In figure 4, we consider four haplotypes in each block. We assume that the frequency of the haplotype associated with the disease allele is a function of the LD between the haplotype and the disease susceptibility allele and the frequencies of the remaining three haplotypes are assumed to be equal and that the LD measures between the three haplotypes and disease locus are equal to zero. We consider three disease models: a recessive disease mode and two genotype relative risk models with γ = 3 and γ =4. We can see from figure 4 that the power of the haplotype block LD mapping depends on the extent of LD between the haplotype and the disease locus, and the disease models.

### 4.3. Mapping

Next we examine how haplotype block mapping can accurately localize the disease locus. We assume that the haplotypes are structured into blocks and the disease gene is located within the third block. A statistic testing individual SNPs and a statistic testing haplotypes are used to localize the disease gene. For convenience of presentation, the below results are the expected theoretic values of the test statistics. These expected statistics were then converted into p-values. Figure 5 shows the profiles of p-values of the single marker LD test and haplotype LD test. Two features are evident from the data. First, both profiles of the p-values of the single marker LD test and haplotype LD test have block structures, i. e., p-values of both tests are constant within block because there is no recombination within the blocks. Second, both the single marker LD test and haplotype LD test can only identify the block in which the disease gene is located, but the haplotype LD test has smaller p-values than the single marker LD test.

## 5. DISCUSSION

Whole genome LD mapping is emerging as an alternative to genetic linkage analysis for the identification of susceptibility genes influencing complex diseases. Its feasibility depends on the extent and pattern of LD, which have been under debate for the past several years (4). Recent discovery of a haplotype block structure may extend the persistence of LD, increase the power of LD mapping and simplify association studies. To help realize applications of haplotype block LD mapping, it is timely to conduct theoretical analyses of haplotype block LD patterns.

A key for haplotype block LD analysis is how to define the block structures. Goldstein (3) proposed to use islands of LD as one way to characterize the block structure of the haplotype. However, the pair-wise LD among SNPs within a large haplotype often shows irregular pattern.

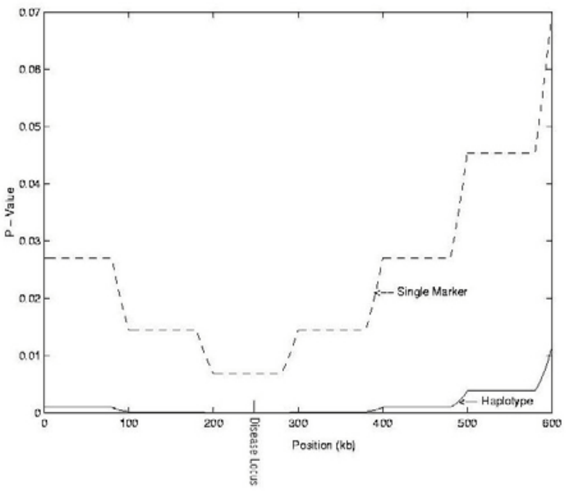

**Figure 5**. P-values for the single marker LD test and haplotype LD test. Assume that the length of each haplotype block is equal to 80 kb, and the distances between the neighboring blocks are equal to 20 kb, the genotype risk disease model with γ = 2.5, four haplotypes with equal frequencies, the disease gene is located at 250 kb from one end of the chromosome, n=500, t=500 generations.

Many evolutionary forces such as selection, local mutation, genetic drift, population bottleneck and recombination affect LD. It is difficult to know which evolutionary force influenced LD among particular sites within a given sample and to uniquely detect the boundaries of the blocks. Patil *et al*. (18) and Zhang *et al*. (19) proposed to use the criterion of minimizing the number of representative SNPs to partition chromosomes into haplotype blocks. However, such identified blocks may contain recombination events, and the pattern of haplotype LD may be complex. Daly *et al*. (12) noted that recombination events are clustered between blocks, with little or no evidence of recombination within blocks. Therefore, in this report, we define a block as a region within which there are no recombination events. In practice, such requirement may not be completely satisfied. However, this characterization of the haplotype block has clear biological meaning and can allow us to observe more regular pattern of LD. Therefore, using such defined haplotype block will increase the power to identify disease gene and simplify LD analyses.

To measure departures from linkage equilibrium may provide important information on the location of disease gene. A number of statistics have been proposed to measure LD levels. However, most measures have focused on pair-wise LD, which quantifies the degree of nonrandom association between pairs of markers (1, 6, 12, 20). Several authors (12) proposed to use multi-allelic extensions of pair-wise LD measures for quantifying the LD of haplotypes. Unfortunately, it is difficult for traditional multi-allelic extensions of pair-wise LD measure to include high order disequilibrium at multiple loci. In this report, we presented an overall LD measure of a haplotype to quantify the degree of all possible pair-wise LD and high order LD in a haplotype. On the basis of overall LD measure of the haplotype, we define the LD measures between the haplotype blocks and between a haplotype block and a marker locus. We showed that the LD measure between haplotype blocks and between the haplotype block and the marker locus based on this overall measure is, in general, larger than that based on the usual multi-allelic extension of pair-wise LD.

The extent of LD and the feasibility of LD mapping for complex disease have been under debate over the past several years. Block structure of haplotypes, if any, will greatly affect the extent of LD and the power of LD mapping. Systematically investigating the extent of LD due to block structures and evaluating the power of haplotype block LD mapping will be useful in the design and practical application of haplotype LD mapping. Therefore, in this report, we presented a simple population genetic model of the haplotype blocks and a formula for calculation of the haplotype LD, which is the basis of power evaluation of the haplotype block LD mapping. Comparing frequencies of the haplotypes in cases and controls is simple and widely used method for association studies. We extended this straight-forward statistical method to haplotype block LD mapping. Finally, we developed analytic formula for the calculation of the noncentrality parameters of the proposed haplotype block LD test.

We can expect that the power of haplotype LD mapping in the presence of block structure will be higher than that of classical single LD mapping that does not make use of blocks. For convenience, we assume an isolated population, with a history of 500 generations. We used the average block length of *25kb* based on the blocks found on chromosome 21 (18). We studied three cases: single marker LD mapping in the absence of haplotype block structure, single marker LD mapping in the presence of haplotype block structure and haplotype block LD mapping in which four haplotypes were assumed. For simplicity, we considered the recombination events as the major evolutionary force, which is an ideal case. We also assumed that there are no phenocopies. However, the presence of phenocopies will not change the major conclusions of the report. We demonstrated that in the absence of haplotype block structure the power of single marker LD mapping is low and rapidly approaches zero as a function of distance. In the presence of block structure, the power of the single marker LD mapping is still low, but it attenuates more slowly. In the presence of block structure, the power of the haplotype LD mapping is high. It is well recognized that the pattern of pair-wise LD is erratic and the level of pair-wise LD dramatically varies over distance between two markers. Any mapping methods based on pair-wise LD will inevitably show inconsistent results over a set of nearby markers and are not effective for association studies.

The cause of haplotype block structures is unclear. It may be due to possible hot and cold spots of recombination, or it may be due to genetic drift and population bottlenecks. Subrahmanyan *et al*. (23) reported that we can observe block structure even when recombination rate is uniformly distributed across the genome. However, even if the block structure is caused by genetic drift, if the number of founders in the population is not large, the number of haplotypes may still be small, which leads to small number of degrees of freedom and high power of the haplotype test.

Haplotype block structure is an important realization in both theory and practice in LD mapping. Haplotype LD mapping has several advantages over single marker LD mapping. First, the pattern of overall haplotype LD is more regular than that of pair-wise LD. Second, the LD level between the haplotype and the disease locus is stronger than the average pair-wise LD level. Third, the extent of haplotype LD is larger than that of pair-wise LD because we assume that there are no recombination events within the block and that the overall LD of the haplotype block is stronger than the pair-wise LD. Fourth, in the presence of block structure, the number of the haplotypes is limited. The increased LD level of the haplotype will not be completely balanced by the increased number of degrees of freedom of the haplotype LD test statistic. Therefore, the power of the haplotype LD mapping is higher than that of the single marker LD mapping.

There is debate over the universal presence of any haplotype block structure in the whole human genome and populations. It is possible that haplotype block structure in a region may exist in some populations, but may not exist in other populations. The total length of the genome showing haplotype block structure in some populations may be larger than in other populations. Studying haplotype blocks structure in particular populations of interest is useful for efficient study design for haplotype block LD mapping. We can expect that the genome regions showing the haplotype block structure vary from population to population. Therefore, we suggest that the criterion for selecting populations which are best suited for LD mapping should be the existence of well-organized haplotype block structure across the genome.

Although the extent and pattern of LD plays a crucial role in the identification of disease genes, we showed that the power of LD mapping also depends on the penetrance and disease model. The relationship between the phenotypes and genotypes is complicated. Recently, it was reported that the human disease phenotypes are influenced not only by the DNA variations, but also by self-organizing networks and system dynamics (24). Although progress in mapping disease genes has been made in the past decades, identifying genes influencing complex diseases is a complex and difficult task.

Some researchers are charging ahead to identify SNPs that define haplotype blocks in one or two populations that may, or may not be applicable to other populations. Others are simply attempting association mapping with available markers, largely ignoring this population genetic characteristics. We believe that a more prudent approach would be to invest more effort to understand the characteristics and utility of haplotype blocks within and among populations. Finally, we believe that "pilot studies" or "demonstration projects" should be carried out to shed light on appropriate approaches for large genome-wide haplotype block mapping.

## 6. APPENDIX

### 6.1. Appendix A

Note from Akey et al. (8) that

$P(H_j \mid A) = a_1 P(H_j, D) + a_2 P(H_j, d)$ and

$P(H_j \mid N) = b_1 P(H_j, D) + b_2 P(H_j, d)$. But,

$P(H_j, D) = \delta_{(H_j, D)} + P(A_{j_1}) \ldots P(A_{j_k}) P(D)$ and

$P(H_j, d) = \delta_{(H_j, d)} + P(A_{j_1}) \ldots P(A_{j_k}) P(d)$

After some calculations we can show that

$$a_1 P(H_j, D) + a_2 P(H_j, d) = a_1 \delta_{(H_j, D)} + a_2 \delta_{(H_j, d)} + P(A_{j_1}) \ldots P(A_{j_k})$$

$$b_1 P(H_j, D) + b_2 P(H_j, d) = b_1 \delta_{(H_j, D)} + b_2 \delta_{(H_j, d)} + P(A_{j_1}) \ldots P(A_{j_k})$$

Substituting the above equations into (2) will result in the equation (3).

### 6.2. Appendix B

Note that two events produce the haplotype *AB* at the next generation. One is that if the haplotype is *AB* and there is no recombination between the two blocks, the haplotype will remain. Another event is that if we have the haplotype A and B and we assume that there is recombination between the two haplotypes, then the haplotypes *A* and *B* will produce the haplotype *AB* at the next generation. Therefore, we have that

$P_{AB}(t+1) = (1-\theta) P_{AB}(t) + \theta P_A P_B$. But,

$P_{AB}(t) = \delta_{H_{AB}}(t) + P_{A_{j1}} \ldots P_{A_{jk}} P_{B_{j1}} \ldots P_{B_{jl}}$ and

$P_A = \delta_{H_A} + P_{A_{j1}} \ldots P_{A_{jk}}$, $P_B = \delta_{H_B} + P_{B_{j1}} \ldots P_{B_{jl}}$, which implies that

$$P_{AB}(t+1) = (1-\theta) P_{AB}(t) + \theta(\delta_{H_A} + P_{A_{j1}} \ldots P_{A_{jk}})(\delta_{H_B} + P_{B_{j1}} \ldots P_{B_{jl}})$$

Thus,

$$\begin{aligned} & P_{AB}(t+1) - P_{A_{j1}} \ldots P_{A_{jk}} P_{B_{j1}} \ldots P_{B_{jl}} \\ & = (1-\theta)[P_{AB}(t) - P_{A_{j1}} \ldots P_{A_{jk}} P_{B_{j1}} \ldots P_{B_{jl}}] \\ & + \theta(\delta_{H_A} \delta_{H_B} + P_{A_{j1}} \ldots P_{A_{jk}} \delta_{H_B} + P_{B_{j1}} \ldots P_{B_{jl}} \delta_{H_A}) \end{aligned}$$

The above equation can be simplified to

$$\begin{aligned} & \delta_{H_{AB}}(t+1) \\ & = (1-\theta) \delta_{H_{AB}}(t) + \theta(\delta_{H_A} \delta_{H_B} + P_{A_{j1}} \ldots P_{A_{jk}} + P_{B_{j1}} \ldots P_{B_{jl}}) \\ & = (1-\theta)^2 [\delta_{H_{AB}}(t-1) + \theta(1-\theta) + \theta][\delta_{H_A} \delta_{H_B} \\ & + P_{A_{j1}} \ldots P_{A_{jk}} \delta_{H_B} + P_{B_{j1}} \ldots P_{B_{jl}} \delta_{H_A}] \\ & = (1-\theta)^{t+1} \delta_{H_{AB}}(0) + [1-(1-\theta)^{t+1}][\delta_{H_A} \delta_{H_B} \\ & + P_{A_{j1}} \ldots P_{A_{jk}} \delta_{H_B} + P_{B_{j1}} \ldots P_{B_{jl}} \delta_{H_A}] \end{aligned}$$

### 6.3. Appendix C

By the same argument as that in Akey *et al* (8), we obtain the following recursive formula for the frequency of the joint haplotype

$$P_{H_{i_1 \ldots i_k}}(t+1) = (1 - \sum_{j=1}^{k-1} \theta_{(j,j+1)}) P_{H_{i_1} \ldots H_{i_k}}(t) + \sum_{j=1}^{k-1} \theta_{(j,j+1)} P_{H_{i_1} \ldots H_{i_j}}(t) P_{H_{i_{j+1} \cdots i_k}}(t),$$

which implies that

$$E\{P_{H_{i_1 \ldots i_k}}(t+1) - P_{H_{i_1 \ldots i_k}}(t)\} \approx -\sum_{j=1}^{k-1} \theta_{(j,j+1)} E[P_{H_{i_1 \ldots i_k}}(t)] + \sum_{j=1}^{k-1} \theta_{(j,j+1)} E[P_{H_{i_1 \ldots i_j}}(t)] E[P_{H_{i_{j+1} \ldots i_k}}(t)]$$

It follows from the above equation that

$$\frac{dE\{P_{H_{i_1 \ldots i_k}}(t)\}}{dt} = -\sum_{k=1}^{k-1} \theta_{(j,j+1)} E[P_{H_{i_1 \ldots i_k}}(t)] + \sum_{j=1}^{k-1} \theta_{(j,j+1)} E[P_{H_{i_1 \ldots i_j}}(t)] E[P_{H_{i_{j+1} \ldots i_k}}(t)] \qquad (C1)$$

This equation can be solved recursively.

If we assume that there is no mutation at the marker locus, the frequency of the haplotype in each block is, in general, assumed to be a constant. We first consider the joint haplotype across the two blocks. It follows from (*C1*) that

$$\frac{dE\{P_{H_{i_1 i_2}}(t)\}}{dt} = -\theta_{12} E[P_{H_{i_1 i_2}}(t)] + \theta_{12} P_{H_{i_1}} P_{H_{i_2}}$$

Solving the above equation yields,

$$E[P_{H_{i_1 i_2}}(t)] = D_{H_{i_1} H_{i_2}}(0) e^{-\theta_{12} t} + P_{H_{i_1}} P_{H_{i_2}},$$ where

$$D_{H_{i_1} H_{i_2}}(0) = P_{H_{i_1 i_2}}(0) - P_{H_{i_1}} P_{H_{i_2}}$$

For the joint haplotype across three blocks, we have

$$\frac{dE[P_{H_{i_1 i_2 i_3}}(t)]}{dt} = -(\theta_{12} + \theta_{23}) E[P_{H_{i_1 i_2 i_3}}(t)] + \theta_{12} P_{H_{i_1}} E[P_{H_{i_2 i_3}}(t)] + \theta_{23} P_{H_{i_3}} E[P_{H_{i_1 i_2}}(t)]$$

Let

$$D_{H_{i_1 i_2}}(0) = P_{H_{i_1 i_2}}(0) - P_{H_{i_1}} P_{H_{i_2}},$$

$$D_{H_{i_2 i_3}}(0) = P_{H_{i_2 i_3}}(0) - P_{H_{i_2}} P_{H_{i_3}}$$ and

$$D_{H_{i_1} H_{i_2} H_{i_3}}(0) = P_{H_{i_1 i_2 i_3}}(0) - P_{H_{i_1}} D_{H_{i_2} H_{i_3}}(0) - P_{H_{i3}} D_{H_{i_1} H_{i_2}}(0) - P_{H_{i_1}} P_{H_{i_2}} P_{H_{i_3}}$$

Then, solving the above equations for $E[P_{H_{i_1 i_2 i_3}}(t)]$ yields,

$$E[P_{H_{i_1 i_2 i_3}}(t)] = D_{H_{i_1} H_{i_2} H_{i_3}}(0) e^{-(\theta_{12}+\theta_{23})t} + P_{H_{i_3}} D_{H_{i_1} H_{i_2}}(0) e^{-\theta_{12} t}$$
$$+ P_{H_{i_1}} D_{H_{i_2} H_{i_3}}(0) e^{-\theta_{23} t} + P_{H_{i_1}} P_{H_{i_2}} P_{H_{i_3}}$$

The expected frequency of the joint haplotype across more blocks can be similarly obtained by recursively solving equation (*C1*)

## 7. REFERENCES

1. Pritchard J.K. & M. Pezeworki: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69, 1-14 (2001)

2. Ardlie K.G., L. Kruglyak & M. Seilstad: Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* 3, 299-309 (2002)

3. Goldstein D.B.: Islands of linkage disequilibrium. *Nat Genet* 29, 109-111 (2001)

4. Nakajima T., L.B. Jorde, T. Ishigami, S. Umemura, M. Emi, J-M. Lalouel & I. Inoue: Nucleotide diversity and haplotype structure of the human angiotensinogen gene in two populations. *Am J Hum Genet* 70, 108-123 (2002)

5. Reich D.E., M. Cargill, S. Bolk, J. Ireland, B. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward & E.S. Lander: Linkage disequilibrium in the human genome. *Nature* 411, 199-204 (2001)

6. Jorde L.B.: Linkage disequilibrium and the search for complex disease genes. *Genome Res* 10, 1435-1444 (2000)

7. Weiss K.M. & A. G. Clark: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 18, 19-24 (2002)

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

9. Service S.K., D.W. Lang, N.B. Freimer & L.A. Sandkuijl: Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet* 64, 1728-1738 (1999)

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

11. MacLean C.J., R.B. Martin, P.C. Sham, H. Wang, R.E. Straub & K.S. Kendler: The trimmed-haplotype test for linkage disequilibrium. *Am J Hum Genet* 66, 1062-1075 (2000)

12. Daly M.J., J.D. Rioux, S.F. Schaffner, T.J. Hudson & E.S. Lander: High-resolution haplotype structure in the human genome. *Nat Genet* 29, 229-232 (2001)

13. Jeffreys A.J., L. Kauppi & R. Neumann: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29, 217- 222 (2001)

14. Johnson G.C., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. DiGenova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C. Gough, D.G. Layton & J. Todd: Haplotype tagging for the identification of common disease genes. *Nat Genet* 29, 233-237 (2001)

15. Rioux D.J., M.J. Daly, M.S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E. J. Kulbokas, S. O'Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S.B. Bull, R.S. McLeod, A. M. Griffiths, A. Bitton, G.R. Greenberg, E.S. Lander, K.A. Siminovitch & T.J. Hudson: Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 29, 223-228 (2001)

16. Stephens J.C., J.A. Schneider, D.A. Tanguay, J. Choi, T. Acharya, S.E. Stanley, R. Jiang, C.J. Messer, A.C. Chew, J.H. Han, J. Duan, J.L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W.R. Newell, A. Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C. M. Drysdale, K. Nandabalan, R. S. Judson, G. Ruaño & G. F. Vovis: Haplotype variation and linkage

disequilibrium in 313 human genes. *Science* 293, 489-493 (2001)

17. Tishkoff S.A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E.H. Touma, S.M. Williams & A.G. Clark: Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293, 455-462 (2001)

18. Patil N., A.J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. N. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. A. Fodor & D. R. Cox: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 1719- 1723 (2001)

19. Zhang K., M. Deng, T. Chen, M.S. Waterman & F. Sun: A dynamic programming algorithm for haplotype partitioning. *Proc Natl Acad Sci USA* 99, 7335-7339 (2002)

20. Devlin B. & N. Risch: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311-322 (1995)

21. Ayres K.L. & D. J. Balding: Measuring gametic disequilibrium from multilocus data. *Genetics* 157, 413-423 (2001)

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

23. Subrahmanyan L., M. A. Eberle, A. G. Clark, L. Kruglyak & D. A. Nickerson: Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am J Hum Genet* 69, 381-395 (2001)

24. Strohman R.: Maneuvering in the complex path from genotype to phenotype. *Science* 296, 701-703 (2002)

**Key Words:** Haplotype Block, Linkage Disequilibrium Mapping, Complex Diseases, Association Study, Population Genetics

**Send correspondence to:** Dr. Momiao Xiong, Human Genetics Center, University of Texas - Houston, P.O. Box 20334, Houston, Texas 77225, Tel: 713-500-9894, Fax: 713-500-0900, E-mail: mxiong@sph.uth.tmc.edu