Abstract

KIM, YUNJUNG. Analysis of Multilocus Linkage Disequilibrium Structure in the Human Genome. (Under the direction of Dr. Zhao-Bang Zeng.)

The International HapMap Project and high- throughput genotyping technology have generated millions of genome-wide marker data that can be used in genetic studies. Each marker can be analyzed separately. But analyzing multiple markers simultaneously through haplotypes has generated great interest recently. Understanding the haplotype structure in the human genome may provide important information on human evolutionary history and identification of genetic variants responsible for human complex diseases. Since the alleles at closely linked markers on a single chromosome are often in statistical dependence (i.e. linkage disequilibrium (LD)), one crucial aspect of haplotype analysis is to characterize LD patterns in different regions and different populations. To assess the extent of correlation of genetic variation at multiple markers in a given region and a population, pairwise LD measures such as r^2 and D' have been commonly used. However, pairwise LD measures alone may be suboptimal to effectively capture the variability of background levels of disequilibrium since multilocus LD measures can provide information about simultaneous allele associations among multiple loci which pairwise LD measures miss. In addition, in order to fully characterize the haplotype structure and LD pattern at multiple markers, it is necessary to consider high order disequilibria and estimate their values.

In this thesis, multilocus LD structure in human populations of the HapMap project is analyzed. In chapter 2, multilocus LD pattern on a chromosome is

summarized using multiple order Markov Chain (MOMC) models as a statistical framework. In chapter 3, within the same framework of MOMC models, LD among multiple markers are measured and partitioned into various components of lower order disequilibria. In chapter 4, two-and three-locus associations in all combinations of three SNPs are tested and their significances are evaluated via two methods - asymptotic chi-square approximation and empirical distributions of LRT statistics. Three - dimensional LD plot is also constructed to visualize patterns of two-and three-locus associations.

From the analysis of multilocus LD structure in the HapMap data, we observe that most LD is explained by LD between adjacent pairs of markers. LD between adjacent pairs of markers appears to be more significant in high multilocus LD regions than in low multilocus LD regions. High orders of LD such as three or more marker associations become more noticeable in low multilocus LD regions. We also notice that there is significant variation in multilocus LD structure between genomic regions and between populations. Detection of high order disequilibria highly depends on the sample size and requires many observed haplotypes.

Analysis of Multilocus Linkage Disequilibrium Structure In the Human Genome

By

Yunjung Kim

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2008

APPROVED BY:

Zhao-Bang Zeng (Chair of advisory Committee)	Jung-Ying Tzeng
Gregory Gibson	Philip Awadalla

Dedication

To my parents, husband, and sons

Biography

Yunjung Kim was born in Taegu, Korea in 1969. She entered the college of pharmacy at Yeung Nam University in Korea in1988 and earned a B.S. in pharmacy four years later. After graduation, she worked as a pharmacist at a general hospital and a retail drug store for several years until she came to America to accompany her husband. She started her master's studies in statistics at the University of Iowa in 2000 and obtained M.S. in statistics in 2002. Afterwards, she continued her study to pursue a doctoral degree in Bioinformatics at North Carolina State University.

Acknowledgements

Like anyone who earned a doctorate degree, to me it has been a very hard and rocky journey to arrive here. I have struggled and devoted a lot of time, energy, and patience to earning what I longed for. Needless to say, it wouldn't be possible without many people who willingly helped me in the process. Above all, I greatly appreciate my advisor Zhao-Bang Zeng for his patience, encouragement, and constructive critiques. His guidance with great intuitions and sophisticated technical knowledge has been invaluable to me, and it truly has been a pleasure and honor to work with him. I am also very grateful to other committee members – Greg Gibson, Jung-Ying Tzeng, and Philip Awadalla – for the encouragement and support I received. All of them are very knowledgeable in their own areas and have been easily accessible. They provided me with useful and effective guidance. Outside the committee, I owe much thanks to Sheng Feng for his insights and collaboration and providing me with the stepping stones for my projects, to Dahlia Nielsen for answering my questions and helping my presentation, to Jb for administrative help numerous times, and to Dr. Zeng's lab members for their sharp comments and constructive discussions on my presentations.

On my family side, I cannot thank my husband Sangkil enough for his enormous sacrifices, strong support since I started studying as a graduate student. Even though he himself had a lot of work to do, he willingly shared a lot of housework with me and took care of kids to give me more time to study. I also thank the most precious my two kids Noel and Maxwell for understanding and tolerating their mom. I have been guilty to

them all the time. My parents and brothers in Korea have believed in me and prayed for me. Their faith in me kept me going forward to reach the goal. My in-laws, brother-in law, and sister-in-law have also understood and helped me a lot. Without all the help from these people, I would not have made it this far. I really appreciate them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Linkage Disequilibrium (LD)

Assessing the patterns of linkage disequilibrium (LD) has become an important issue in both medical genetics and evolutionary biology since the rapid accumulation of densely spaced DNA sequence variation data in several organisms. LD is defined as the non-random association of alleles at different loci on the same chromosome or on different chromosomes in a population. It is the correlation between polymorphisms that is caused by their shared history of mutation and recombination. In most cases, LD measures the allelic association in the same haplotype and hence can be considered as a measure of chromosomal proximity or linkage of genetic loci. However, physical linkage between loci is neither necessary nor sufficient to generate associations (McVean 2002b). Tight linkage does not necessarily mean high level of LD and alleles at unlinked loci can exhibit high level of LD (Weir 1996).

1.1.1 Factors that influence LD

LD patterns observed in natural populations are the results of a complex interplay between biological factors such as recombination, mutation, and the population's demographic and evolutionary history. The structure and the effective size of populations as well as the force of selection are important factors for the regional LD

patterns. Therefore, it is not surprising that there is a lot of variation in LD among different genomic regions and populations.

Recombination is the main phenomenon that weakens LD. When a new mutation occurs on a chromosome, there is a complete LD between the mutation and alleles at nearby loci. The mutation is co-transmitted with pre-existing variants until they are gradually broken apart by some form of chromosomal rearrangement, such as recombination or gene conversion. Recombination rates are known to vary by more than an order of magnitude across the genome. Since LD is broken down mainly by recombination, the extent of LD is expected to vary in inverse relation to the local recombination rates. Recurrent mutation can also reduce LD, but for SNPs, recurrent mutation is very rare. There is not much evidence to indicate that recurrent mutation contributes significantly to the erosion of LD between SNPs (Ardlie et al., 2002).

Genetic drift, the change in allele and genotype frequency in a finite population every generation owing to the random sampling of gametes, does influence LD. Frequency changes are more fluctuated in small populations. In general, the genetic drift in a finite population increases LD because rare alleles or haplotypes are lost from the population. Population growth decreases LD by reducing the effect of genetic drift. Population admixture or migration (gene flow) can create LD between populations.

Natural selection affects LD by two primary routes. The first way is a hitchhiking effect or selective sweep by which an entire haplotype flanking an advantageous variant can be rapidly swept to high frequency. Selection against deleterious variants can also increase LD because the deleterious haplotypes are swept

from the population. The second way is through epistatic selection for a particular combination of alleles at two or more loci on the same chromosome (Ardlie et al., 2002).

1.1.2 Applications of LD

There are a number of reasons for the study of LD. First of all, LD plays a central role in LD mapping in which unobserved causal variants can be identified through indirect associations between a set of markers and phenotypes. Association or linkage disequilibrium (LD) mapping aims to estimate the locations of genetic variants on chromosome in a much finer scale and predict their genetic effects. We will discuss the applications of LD in association mapping in detail in section 1.3. Second, LD is also of fundamental importance in estimating population recombination rates that are difficult to study experimentally. A number of model based LD measures have been proposed to infer the regional variation of recombination rates in fine scale (McVean et al., 2002; Li and Stephens, 2003). Third, LD is important for untangling the evolutionary history of humans and other organisms, which includes identification of demographic effects such as population growth, bottlenecks and admixture (McKeigue et al., 2000; Wall 2000), and the detection of natural selection (Sabeti et al., 2002).

1.2 Measures of LD

There are a number of LD measures ranging from pairwise LD measures such as D' and r^2 to multilocus LD measures that consider the joint LD among multiple markers,

or model-based LD measures for estimation of population recombination rate, or haplotype-specific LD measure to detect the signatures of positive selection.

1.2.1 Pairwise LD measures

A simple and basic component of many disequilibrium measures is the difference (D) between observed gametic frequency and the expected gametic frequency under the linkage equilibrium. Adopting the standard notation for two adjacent loci (**A** and **B**), with two alleles (*A*, *a* and *B*, *b*) at each locus, the observed haplotype frequency that consists of alleles *A* and *B* is represented by P_{AB} and the expected haplotype frequency is the product of the two allele frequencies (i.e. $P_A \times P_B$). So, the simplest measure of LD is $D = P_{AB} - P_A \times P_B$. In order to compare the magnitude of LD among different pairs of loci with different allele frequencies, several standardization methods have been proposed and their properties are compared in Hedrick (1987), Lewontin (1988), Delvin & Risch (1995) , and Morton et al. (2001).

The two most common pairwise LD measures are D' and r^2. They have very different properties and may be applied for different purposes. D' and its confidence bounds are useful to assess the probability for historical recombination in a given population. D' is defined as

$$D' = D/D_{\max},$$

where $D_{\max} = \min(P_A P_b, P_a P_B)$ if $D > 0$ and $D_{\max} = \max(-P_A P_B, -P_a P_b)$ if $D < 0$. This procedure always makes D' value range between 0 and 1. D' reaches 1 if two or three

gametes out of four possible gametes are observed. When all four gametes between a pair of loci are observed, D' will be less than 1, indicating at least one historical recombination event. However, values of $D' < 1$ have no clear biological interpretation. In addition, D' is strongly inflated in small samples and for SNPs with rare alleles. Therefore, it is suggested that confidence intervals of D' should be used rather than the D' value itself (Gabriel et al. 2002) to measure the extent of LD.

The measure r^2 is defined as

$$r^2 = D^2 \Big/ P_A P_a P_B P_b \ .$$

r^2 reaches 1 when only two gametes are observed. In this case, observation at one marker provides complete information about the other marker. Values less than 1 are easily interpretable as the amount of information provided by one locus about the other. r^2 is a more relevant measure for association studies because there is a simple relationship between r^2 and the sample size required to detect association between a phenotype and marker locus (Pritchard et al. 2001). Suppose an LD of r^2 was measured between a causal locus and a nearby marker locus. Then, to achieve the same power to detect association at the marker locus as we would have at the causal locus, we need to increase our sample size by a factor of $1/r^2$ (Pritchard et al., 2001).

Significance testing for *D* can be conducted as the testing for independence in a 2 $\times 2$ contingency table. The usual methods for this type of test are a chi-square test, likelihood ratio test, or Fisher's exact test. More detailed information about the test can

be found in Weir (1996). Properties of multiple allele LD measures are discussed in Weir and Cockerham (1978) and Weir (1996).

When only genotype data in diploid individuals are available and the gametic phase is unknown, family recruitment or experimental methods may help to estimate the phase in diploid data. However, these methods are currently time consuming and expensive. Under the assumption of random mating in which genotypic frequencies are assumed to be products of gametic frequencies, it is possible to obtain the maximum likelihood estimates of gametic frequencies using EM algorithm (Excoffier et al., 1995). With the estimated gametic frequencies, we can proceed with the standard LD calculations described above. If the assumption of random mating might not be valid for a certain population, alternative Bayesian methods for haplotype construction that are relatively robust to deviations of the random mating assumption (Stephens et al. 2001) or composite genotypic disequilibrium measures (Weir 1996) can be considered. Schaid (2004) extended the composite LD measure to the multiple alleles for two loci.

1.2.2 Multilocus LD measures

Multilocus LD measures can provide information about simultaneous allele associations among multiple loci which pairwise LD measures cannot. To compensate for this limitation, all pairwise LD values between multiple loci can be summarized. For example, Hedrick (1987) proposed a weighted sum of pairwise D' values to describe multilocus LD but this quantity remains difficult to interpret.

Homozygosity of haplotypes (i.e. the probability of selecting two identical haplotypes at random from the population) is suggested as a measure of multilocus disequilibrium (Sabatti et al., 2002). An excess of either homozygosity or heterozygosity signals a departure from the gametic phase equilibrium. Haplotype homozygosity can measure haplotype specific LD. Extension of this method was used as a test for detecting positive selection (Sabeti et al., 2002).

Nothnagel et al. (2002) proposed an entropy-based a multilocus LD measures (ε) which is based on the entropy difference between entropy of haplotype distribution under linkage disequilibrium and that under linkage equilibrium. It is defined as follows. For n bi-allelic loci such as SNPs, let p_j denote the major allele of the jth SNP, j = 1, 2, .., n. Suppose that m haplotype are observed among 2^n possible haplotypes. Then the entropy of the observed haplotype distribution is defined as:

$$H = -\sum_{i=1}^{m} q_i \log q_i \ ,$$

where q_i denotes the frequency of haplotype i.

Under the linkage equilibrium, the frequency of any haplotype k can be calculated using the formula

$$q_k^E = \prod_{j=1}^{n} p_j^{I_k^j} (1-p_j)^{1-I_k^j} \ ,$$

where $I_k^j = 1$ if the allele on haplotype k at the jth SNP is the major allele, otherwise it is 0. The corresponding entropy under the linkage equilibrium is then

$$H_E = -\sum_{k=1}^{2^n} q_k^E \log(q_k^E)$$

The difference between expected and observed entropy, $H_E - H$, is a measure of the sequence's deviation from its linkage equilibrium. In order to scale the multilocus LD measure between 0 and 1, the authors then defined normalized entropy as following:

$$\varepsilon = \frac{H_E - H}{H_E}$$

The authors showed that (H_E - H) is approximately equal to $\frac{1}{2}r^2$ using Taylor expansion to order 2.

The Graphical modeling of the joint distribution of alleles at associated loci proposed by Thomas et al. (2004) is an interesting approach to measure the multilocus linkage disequilibrium. Most approaches to describing LD impose constraints on physical location and characterize LD in terms of contiguous and mutually exclusive genetic regions such as haplotype blocks. Such approaches are effective in describing LD in large genetic regions where recombination and spatial relationships dominate. However, on the fine scale, the mutation process is also important and creates associations between loci that are independent of the physical ordering. The unique idea of this approach is that it allows for non-contiguous and overlapping LD groups.

A Bayesian approach analyzing multilocus data to assess departures from equilibrium is suggested by Ayres et al. (2001). A Markov chain Monte Carlo (MCMC) algorithm is employed to approximate the posterior probability distributions of disequilibrium parameters. Some advantages of their approach are that the uncertainties in parameter estimates can be easily assessed in terms of probabilities and that it can incorporate background knowledge to improve the precision of inferences.

1.2.3 Model-based LD measures and estimation of recombination rates

Pairwise or multilocus LD measures discussed in section 1.2.1 and 1.2.2 can be regarded as summary statistics for the patterns of observed LD. The approximate moments of these summary statistics are known under standard neutral model (Kimura and Ohta 1971; Hill 1974a, 1974b). Although these measures are widely used to describe the patterns of LD, one disadvantage is that there is no direct relationship between them and biological mechanism of interest, such as the underlying recombination rate (Li and Stephens, 2003). However, methods based on explicit evolutionary models can be useful for analyzing patterns of LD across multiple loci and estimating population recombination rates.

The most successful current approaches to fit statistical models to patterns of LD and to estimate recombination rate are based on the framework of coalescent theory (Kingman 1982). In models based on the coalescent theory, the key determinant of the extent of LD is the population recombination rate ($\rho = 4N_e r$) which is the product of the recombination rate per base pair (r) and the effective population size (N_e). Assuming a constant population recombination rate (ρ) from sequence variation data from a region of interest, the ideal approach is to estimate ρ by calculating the full likelihood curve of ρ. However, since full likelihood methods are too computationally intensive and practically impossible for many large datasets, approaches using approximate likelihood methods have been proposed. The assumption of constant recombination rate can be relaxed to adjust for variable recombination rates across the genome.

Hudson (2001) suggested an ad hoc method in which one calculates the likelihood curve for all pairs of segregating sites and then multiplies together all these curves, assuming the independence of pairs. Extension of this method is used for models of gene conversion (McVean et al., 2002), variable recombination rates (McVean et al., 2004), and recurrent mutation (McVean et al., 2002). Fearnhead and Donnelly (2001) is based on dividing the region of interest into sub-regions, calculating the likelihood curve for each sub-region, and then multiplying all the sub-regions together. Li and Stephen (2003) developed a statistical model which directly relate patterns of LD to the underlying recombination process. This method considers each haplotype in an order and attempts to construct it as a mosaic of previously considered haplotypes. It is based on the conditional likelihood of the type of *i*th haplotype given the types of the first *i*-1 haplotypes. A likelihood is then constructed by multiplying these approximate conditional likelihoods together.

1.2.4 Haplotype-specific LD and signatures of positive selection

Most multilocus LD measures discussed in section 1.2.2 describe the general LD for an array of several haplotypes within a chromosomal region and provide a single value for the strength of LD. However, each haplotype may have its own evolutionary history and one may be interested in the LD structure of a specific haplotype. For example, assessing the relative frequency of a particular haplotype and the extent of LD around the variants can provide a signal for positive selection. The logic behind this strategy is following: Under neutral evolution, new mutations require a long time to reach high frequency in the

population, and LD around the mutations will decay substantially during this period because of recombination. As a result, common alleles will typically be old and will be found on smaller haplotye blocks. On the other hand, rare alleles may be either young or old and are associated with either small or large haplotype blocks.

A variety of genetic signatures of positive selection have been described (Bamshad et al. 2003). These include an excess of rare variants (indicating a selective sweep followed by the accumulation of new, rare mutations), large allele frequency difference among populations (indicating differential effects of selection in some but not all of the populations), and an unusually long haplotype given its high frequency. The LRH (long-range haplotype) test is the implementation of third signature (Sabeti et al., 2002). In the LRH test, one measures multilocus LD at a distance *x* from the core region by calculating the extended haplotype homozygosity (EHH). EHH is defined as the probability that two randomly chosen chromosomes carrying the core haplotype are homozygous at all SNPs for the entire interval from the core region to the point *x*. EHH thus detects the transmission of an extended haplotype without recombination. By applying the LRH test, G6PD (Glucose -6-phosphate dehydrogenase) gene and CD40 ligand gene carrying common variants implicated in resistance to malaria were identified as regions that have undergone positive selection or partial selective sweeps. (Sabeti et al., 2002).

1.3 Application of LD in association mapping

In the past decades, mapping disease causing variants using LD between a set of markers and disease phenotype has been popular since this approach has the potential to locate complex disease genes by effectively incorporating the recombination events accumulated throughout many past generations. This enthusiasm has been escalated by the discovery of millions of SNPs. LD mapping using SNPs is now an important strategy for mapping genes involved in complex common diseases.

1.3.1 Complex disease

Although biomedical research has accelerated its pace in the past decades, the root causes of common human diseases remain largely unknown. The clustering of such diseases among family members indicates that common human diseases have some genetic basis to susceptibility. Identifying the causal genes or variants would represent an important step in the path towards improved prevention, diagnosis, and treatment of diseases.

The first genetic diseases to be studied in humans were fairly simple, monogenic, rare at the population level, and highly penetrant disorders that obey the rules of Mendelian inheritance. In contrast, most common complex human diseases such as heart disease, cancers, diabetes are influenced by genetic variations at several loci in the genome, each of low penetrance. In addition, environmental factors, interactions between genes and environments, and interactions among genetic variants at different loci (epistasis) are all likely to be important for complex disease.

The classical method to identify loci contributing to disease is the linkage analysis, in which data are collected from affected families. In linkage analysis, genomic regions that co-segregate with the disease are identified in independent families or an extended pedigree. However, the resolution of linkage analysis is low and typically on the order of a few cM, which may correspond to several Mb of DNA, and 100s or 1000s of genes in terms of human genome.

In order to narrow the interval in which disease gene(s) might lie, other methods were considered, and one of them was by the analysis of linkage disequilibrium (LD). This approach makes use of many opportunities for crossovers between markers and the disease locus over many generations since the first appearance of the mutation. Therefore, it's possible to narrow the interval around the disease locus by detecting an indirect association, through LD, between nearby marker(s) and the disease locus. Even the disease-related variants that are not genotyped would be assayed indirectly through LD with nearby markers. This is known as LD mapping. In order for this strategy to be successful, we need to understand the levels and general patterns of background LD. In particular, we need to gain some insight into how far usable levels of disequilibrium extend in the human genome, and how much variation of LD exists among regions and populations.

1.3.2 Empirical patterns of LD in the human genome

In the past 10 years, there have been a number of studies reporting the empirical patterns of human genome that characterize the extent and range of LD. (Huttley et al., 1999;

Abecasis et al., 2001; Reich et al., 2001; Hinds et al. 2005). In most studies of LD among SNPs, the major conclusions are as follows. First, there's considerable variation in the pattern and extent of LD between different regions and populations and thus it will be difficult to predict LD from one region to another. There are reports of strong LD between markers that are separated by distances of > 100 kb. In contrast, a number of studies reported weak LD at much shorter distances. Abecasis et al. (2001) estimated that physical distance could account for less than 50 % of the variation in LD in their study, suggesting that the remaining variation was probably due to a combination of drift, demographic factors, selection, and variable rates of mutation, recombination, gene conversion, and inherent stochastic nature of LD. Second, non-African populations typically show lower nucleotide diversity and higher LD than African populations.

Despite the apparent complexity of observed patterns of LD, several studies have proposed that the underlying structure of LD in the human genome can be described using a relatively simple framework in which the genomic regions can be divided into a series of discrete haplotype blocks and neighboring blocks are separated by regions of recombination hotspots (Daly et al, 2001; Jeffreys et al., 2001; Gabriel et al., 2002). The haplotype block model has rapidly gained popularity as it offers important implications for association mapping. In theory, if one could identify the haplotype blocks across the genome and carefully choose informative subsets of SNPs (tagSNPs) from each block, the disease association studies could use them without much loss of power. In response to this idea, the international HapMap project was initiated and has already provided information on the patterns of variation and linkage disequilibrium for more than 3

million SNPs and has led to the development of many statistical methods using the SNP data. Results for a similar genome-wide survey of linkage disequilibrium including ~1.5 million SNPs have been reported (Hinds et al., 2005).

1.3.3 Haplotype blocks and some potential problems

The haplotype-block model is expected to capture the underlying structure of LD. Thus, quite a large range of methods for defining haplotype blocks have been proposed so far. These methods can be broadly classified into two main groups: (1) methods that define blocks with limited haplotype diversity (Daly et al. 2001; Patil et al. 2001; Dawson et al. 2002; Zhang et al. 2002a) and (2) methods that utilize pairwise disequilibrium measures and identify recombinational hotspots (Gabriel et al. 2002; Wang et al. 2002).

Daly et al. (2001) compared the observed heterozygosity with expectation under linkage equilibrium in sliding windows in order to search for regions of low haplotype diversity. They also estimated a haplotype transition probability θ ("historical recombination rate") for ancient haplotypes and the frequency of the major haplotypes within each block using a hidden Markov model, where $1-\theta = D'$. Patil et al. (2001) were the first to describe the haplotype block structure of an entire chromosome (chromosome 21). Their method chose blocks by limiting the number of htSNPs via a greedy algorithm. Dawson et al. (2002) used both D' and a reduced haplotype diversity criterion. Zhang et al. (2002a) applied an alternative algorithm called a dynamic programming algorithm to the same data as Patil et al. (2001)'s. Their results were somewhat more parsimonious than Patil et al. (2001)'s. Gabriel et al. (2002) defined

blocks as sets of consecutive sites between which there is little or no evidence of historical recombination. More specifically, their method was based on a composite of local D' values. It generated 95% confidence bounds on D' rather than point estimates of D' because D' values are known to fluctuate upward when sample size is small or rare alleles are examined. Wang et al. (2002) used an adaptation of the four-gamete test (FGT) (Hudson et al. 1985). In their method, the frequencies of the 4 possible haplotypes for each marker pair were computed. If all 4 haplotypes are observed with at least frequency 0.01, a recombination is deemed to have taken place. Blocks were formed by consecutive markers where only less than or equal to 3 haplotypes are observed.

Haplotype blocks (i.e. regions of low haplotye diversity and high LD interspersed by recombination hotspots) have been empirically identified and proposed to constitute a ubiquitous feature of the genome. Characterization of haplotype blocks can provide association studies with a shortcut to screening chromosomal regions to find disease causing variants by using haplotype-tagging SNPs. However, the potential benefits of haplotype blocks may be challenged by concerns regarding their consistency due to the arbitrary choice of block definition, marker density, allele frequency, choice of thresholds in different block definitions, and sample size. Nevertheless, the consensus findings from many groups is that denser marker maps, larger sample sizes, more stringent thresholds , and use of common variants tend to lead to shorter blocks. The size and boundaries of generated blocks could have impact on the downstream analysis of association studies and influence the design of fine mapping of disease causing variants. Thus, the haplotype block model might provide a simple and intuitive way for capturing some of prominent

patterns of LD and the proposal of taking advantage of haplotype blocks in designing genetic association studies is a welcome step forward, but they are not panacea for the field of human complex disease studies (Zeggini eta al., 2005; Lawrence et al., 2005).

1.3.4 Tag SNPs for association studies

The block-like LD pattern and redundancy among SNPs can maximize the efficiency in the laboratory while minimizing loss of information by selecting most informative subset of SNPs (tagSNPs) for genotyping association studies. A number of methods for selecting tagSNPs make explicit use of haplotype blocks, but some methods select tagSNPs without identifying haplotype blocks first. There are two main definitions of tag SNPs based on (1) fraction of common haplotypes distinguished by tag SNPs (Patil et al., 2001; Clayton 2001); (2) pairwise LD measure r^2 (Carlson et al., 2004).

Patil et al. (2001) and Clayton (2001) consider a set of tagSNPs as the minimum set of SNPs that can uniquely distinguish at least α percentage of all the observed haplotypes in each block. In Carlson et al. (2004), all SNPs above the MAF threshold are first selected in a candidate gene region and then the single SNP exceeding r^2 threshold with the maximum number of other SNPs is identified. This maximally informative single SNP and all other SNPs are grouped as a bin of associated SNPs. Because the bin is initially ascertained using a single SNP, all pairwise r^2 within the bin is re-evaluated, and any SNP exceeding threshold r^2 with all other SNPs in the bin is specified as tagSNPs for the bin. Thus, one or more SNPs within a bin are specified as tagSNPs and only one tagSNP would need to be genotyped per bin.

Although the idea of tagSNPs makes it possible to survey substantial fractions of human variation in a cost-effective manner by using a carefully selected subset of all SNPs, there must inevitably be a reduction in the power of tests of association compared to using all SNPs. Several investigators assessed the loss in power when various tagSNP approaches are followed, with different conclusions. Zhang et al. (2004) showed that the use of tagSNP that retain haplotype diversity can result in considerable loss in power of association tests, especially when the causal marker has low frequency. Loss of power is minimized when the difference between minor allele frequencies of the causal SNP and a closely associated marker SNP is small. Phase I HapMap project emphasized the use of common SNPs, and the statistical association between common SNPs and rare disease-causing alleles is usually weaker than that for SNPs whose frequencies closely match the disease allele frequency. Phase II HapMap represents rare variation better than the Phase I HapMap. de Bakker et al. (2005) examined the power under the scenario of rare causal allele and concluded that exhaustive haplotype testing which tests all local haplotypes can improve power for less common causal alleles but are neutral or reduce power when the causal SNP is common. de Bakker et al. (2006) also looked at how well the HapMap tagSNPs can cover common variants in other non-HapMap populations (i.e. tag transferability across populations). Good coverage and high power were achieved using HapMap tagSNPs for non-HapMap populations.

1.4 Consideration of high order linkage disequilibrium

For the assessment of the background levels of LD, measuring multilocus LD may be more advantageous than measuring LD between pairs of markers because the combined analysis of all piarwise LD measures across a region cannot detect simultaneous allele associations among multiple markers. Various multilocus LD measures are reviewed in section 1.2.2. One common property of those methods is to summarize the data with a single multilocus LD measure by first calculating the difference between observed state and the expected one under linkage equilibrium and then normalizing it to allow comparison across different regions and populations. The extent of normalized difference between observed state and the expected state is regarded as the degree of overall departure from linkage equilibrium.

Although those methods overcome the limit of pairwise LD measures by considering joint multilocus LD, they do not distinguish the type and structures of multilocus LD, such as how much multilocus LD is due to LD between marker pairs and how much due to higher orders. For example, if we consider p SNPs simultaneously, there are $2^p - (p+1)$ LD terms of varying orders that need to be considered to fully characterize the LD structure in them. In Chapter 3, we estimate the contribution of various orders of LD to the total multilocus LD in each sliding window of w markers. In Chapter 4, we test the significances of two- and three-locus LD via asymptotic chi-square distribuion and a resampling method for all three combinations of markers. Methods were developed for testing for linkage disequilibria among more than two neutral loci

(Hill 1975, 1976; Smouse 1974; Weir 1996). We first review those methods, and focus the discussion on testing for three-locus disequilibrium.

1.4.1 Two concepts of no three-locus disequilibrium

When considering three biallelic SNPs of M_1, M_2, M_3 with alleles *a*, *A* ; *b*, *B*; and *c*, *C*, there are eight possible haplotypes. Each haplotype frequency can be expressed in terms of three allelic frequencies P_A, P_B, P_C, three two-locus disequilibrium measures (D_{AB}, $D_{BC,}$ D_{AC}) and a measure of additional disequilibrium among three loci (T):

$$P_{ABC} = P_A P_B P_C + P_A D_{BC} + P_B D_{AC} + P_C D_{AB} + T$$

$$P_{Abc} = P_A P_b P_c + P_A D_{BC} - P_b D_{AC} - P_c D_{AB} + T$$

$$P_{aBc} = P_a P_B P_c - P_a D_{BC} + P_B D_{AC} - P_c D_{AB} + T$$

$$P_{abC} = P_a P_b P_C - P_a D_{BC} - P_b D_{AC} + P_C D_{AB} + T$$

$$P_{ABc} = P_A P_B P_c - P_A D_{BC} - P_B D_{AC} + P_c D_{AB} - T$$

$$P_{AbC} = P_A P_b P_C - P_A D_{BC} + P_b D_{AC} - P_C D_{AB} - T$$

$$P_{aBC} = P_a P_B P_C + P_a D_{BC} - P_B D_{AC} - P_C D_{AB} - T$$

$$P_{abc} = P_a P_b P_c + P_a D_{BC} + P_b D_{AC} + P_c D_{AB} - T$$

Two different notions of independence at the level of three loci have been used. The first concept of three-locus equilibrium is expressed by the null hypothesis $H_0 : T = D_{ABC} = 0$. There is a quite large literature on using this three-locus disequilibrium coefficient (D_{ABC}) for formulating multilocus association (Bennett 1954;

Thompson and Baur 1984; Robinson 1991; Long 1995; Weir 1996) and various approaches have been suggested for testing $H_0 : T = D_{ABC} = 0$ (Lancaster 1951; Weir 1996).

Although this concept has been commonly used, specifying the null hypothesis as $H_0 : T = D_{ABC} = 0$ has problems. First, haplotype frequency distribution under the null hypothesis can have negative frequencies. Second, the range of possible values for D_{ABC} may not include zero when some haplotypes are not observed. Thompson and Baur (1984) showed the minimum and maximum values D_{ABC} can take for the given allele frequencies and pairwise disequilibrium coefficients.

$$L \le D_{ABC} \le U$$

where

$$L = \max \{ -P_0(ABC), -P_0(Abc), -P_0(aBc), -P_0(abC) \}$$

$$U = \min \{ P_0(ABc), P_0(AbC), P_0(aBC), P_0(abc) \}$$

and $P_0(ABC), ..., P_0(abc)$ denote haplotype frequencies without the term (T) for three-locus disequilibrium. Both *L* and *U* can be positive or negative. For example, suppose that only two out of eight possible haplotypes are observed with their frequencies of $P_{ABC} = 0.3$ and $P_{abc} = 0.7$. In this case, both min D_{ABC} and max D_{ABC} are 0.084 and zero is not included in the range. Therefore, testing the null as $H_0 : T = D_{ABC} = 0$ may not be appropriate.

The second concept to define no three-locus association was initially proposed by Bartlett (1935). No three-locus association occurs when T (denoted as Z_{ABC}) assumes a value which satisfies the following condition:

$$(P_{ABC}P_{Abc}P_{aBc}P_{abC}) = (P_{ABc}P_{AbC}P_{aBC}P_{abc})$$

This condition means the association between, say, A and B should be the same in the chromosome having C as c. This condition is symmetric among the loci as following.

$$\frac{P_{ABC}P_{Abc}}{P_{ABc}P_{AbC}} = \frac{P_{aBC}P_{abc}}{P_{aBc}P_{abC}} \Leftrightarrow \frac{P_{ABC}P_{aBc}}{P_{ABc}P_{aBC}} = \frac{P_{AbC}P_{abc}}{P_{Abc}P_{abC}} \Leftrightarrow \frac{P_{ABC}P_{abC}}{P_{AbC}P_{aBC}} = \frac{P_{ABc}P_{abc}}{P_{Abc}P_{aBc}}$$

Since there is no explicit solution for Z_{ABC}, it should be obtained using an iterative procedure such as the one suggested by Fienberg (1970). In order to avoid some problems in the first concept, we use the second concept of no three-locus disequilibrium and test for three-locus disequilibrium in Chapter 4.

1.4.2 Several asymptotic tests for three-locus disequilibrium

In this section, we review several asymptotic tests for three-locus disequilibrium in the literature. Lancaster (1951) suggested to use 4 df statistic which asymptotically follows chi-square (4 df), in order to test the hypothesis of global equilibrium ($H_0 : P_{ABC} = P_A P_B P_C$). The 4 df statistic was partitioned into three 1 df chi-square statistic terms for testing two-locus equilibrium ($H_0 : D_{AB} = D_{AC} = D_{BC} = 0$) and one 1 df chi-square statistic term for testing three-locus equilibrium ($H_0 : D_{ABC} = 0$):

$$\chi_L^2 = N\left(\frac{(\hat{P}_{ABC} - \hat{p}_A\hat{p}_B\hat{p}_C)^2}{\hat{p}_A\hat{p}_B\hat{p}_C} + ... + \frac{(\hat{P}_{abc} - \hat{p}_a\hat{p}_b\hat{p}_c)^2}{\hat{p}_a\hat{p}_b\hat{p}_c}\right) = N(r_{AB}^2 + r_{AC}^2 + r_{BC}^2 + r_{ABC}^2)$$

where

$$r_{AB}^2 = \frac{\hat{D}_{AB}^2}{\hat{p}_A \hat{p}_a \hat{p}_B \hat{p}_b}$$
$$r_{AC}^2 = \frac{\hat{D}_{AC}^2}{\hat{p}_A \hat{p}_a \hat{p}_C \hat{p}_c}$$
$$r_{BC}^2 = \frac{\hat{D}_{BC}^2}{\hat{p}_B \hat{p}_b \hat{p}_C \hat{p}_c}$$
$$r_{ABC}^2 = \frac{\hat{D}_{ABC}^2}{\hat{p}_A \hat{p}_a \hat{p}_B \hat{p}_b \hat{p}_C \hat{p}_c}$$

and $\hat{D}_{ABC} = \hat{P}_{ABC} - \hat{p}_A \hat{D}_{BC} - \hat{p}_B \hat{D}_{AC} - \hat{p}_C \hat{D}_{AB} - \hat{p}_A \hat{p}_B \hat{p}_C$. However, the partition due to Lancaster (1951) has been criticized on some grounds. The three two-locus disequilibria are not independent and this is missed by the partitioning. Plackett (1962) argued that Lancaster's criterion was not strictly a test of three-locus association.

Weir (1996) proposed an explicit test statistic for the three-locus disequilibrium:

$$\chi_W^2 = \frac{\hat{D}_{ABC}^2}{\mathrm{Var}(\hat{D}_{ABC})}$$

The approximate variance of $\hat{D}_{ABC}$ was obtained from Fisher's formula:

$$\begin{aligned}\mathrm{Var}(\hat{D}_{ABC}) = \frac{1}{N}\{ & p_A p_a p_B p_b p_C p_c + 6D_{AB}D_{AC}D_{BC} \\ & + p_A p_a[(1-2p_B)(1-2p_C)D_{BC} - D_{BC}^2] \\ & + p_B p_b[(1-2p_A)(1-2p_C)D_{AC} - D_{AC}^2] \\ & + p_C p_c[(1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2] \\ & + D_{ABC}[((1-2p_A)(1-2p_B)(1-2p_C) - 2(1-2p_A)D_{BC} \\ & - 2(1-2p_B)D_{AC} - 2(1-2p_C)D_{AB} - D_{ABC}]\}\end{aligned}$$

To test the null hypothesis $H_0 : T = D_{ABC} = 0$ with chi-square statistic, D_{ABC} is set to zero and all other terms are replaced by their observed values. The problem with Weir (1996) is that if some of the haplotypes are unobserved in the sample, we may get negative three-locus variance under the null hypothesis $H_0 : T = D_{ABC} = 0$.

Brown (1975) suggested a slightly different chi-square test using the second concept of no three-locus disequilibrium. Z_{ABC} satisfying the condition of Bartlett (1935) can be estimated first.

$$\hat{Z}_{ABC} = P^*_{ABC} - p_A D_{BC} - p_B D_{AC} - p_C D_{AB} - p_A p_B p_C \ ,$$

where P^*_{ABC} denote an adjusted haplotype frequency obtained using an iterative procedure (Fienberg 1970). Then, an approximate variance of $\hat{Z}_{ABC}$ can be computed by plugging $\hat{Z}_{ABC}$ into the formula for $\mathrm{Var}(\hat{D}_{ABC})$. A test statistic under the null ($H_0 : T = Z_{ABC}$) is calculated and compared with the chi-square distribution.

$$\chi^2_Z = \frac{\hat{Z}^2_{ABC}}{\mathrm{Var}(\hat{Z}_{ABC})}$$

Contrary to $\mathrm{Var}(\hat{D}_{ABC} \mid T = D_{ABC} = 0)$, $\mathrm{Var}(\hat{Z}_{ABC} \mid T = Z_{ABC})$ is always positive even if some haplotype frequencies are zero.

A likelihood ratio test statistic which asymptotically follows chi-square distribution can also be used to compare two competing models (Hill 1975, 1976).

$$2N \sum_{i,j,k} P_{ijk} \log\left(\frac{P_{ijk}}{P^*_{ijk}}\right)$$

If the number of observed haplotypes is small, a problem arises because some of P_{ijk}^{*} converges to zero using the iterative procedure (Fienberg 1970). To detour this problem, we may define $0/0$ as zero since P_{ijk} is zero if P_{ijk}^{*} converges to zero.

Smouse (1974) proposed a log-linear approach to make inferences about the multilocus linkage disequilibrium. In his log-linear approach, a series of multiplicative models which include different numbers of disequilibrium terms are first constructed. Then, the difference in deviances for two models that differ only by whether a particular disequilibrium term is included provides a chi-square test statistic for that term. The total disequilibrium is partitioned into various components corresponding to two-locus, three-locus, and higher order disequilibria.

1.5 Outline of Research

The preceding sections have shown that characterizing the patterns of LD has been an important issue in mapping complex disease genes since there is a great potential for fine mapping of disease variants using the LD based mapping methods. Various approaches such as pairwise, multilocus, model-based LD measures have been developed to describe the LD pattern. Recently, the concept of haplotype blocks has simplified the complex empirical patterns of LD and carefully selected tagSNPs from the blocks have been used for many genetic association studies. Also, recent evidences of hotspots and cold spots of recombination have suggested that haplotype-based methods may play a key role in the study of common complex diseases. However, not much attention was given to the underlying structure of multilocus LD, such as pairwise LD between marker pairs and higher order LD among more than two markers, when considering haplotype-based methods. For example, if we consider a haplotype consisting of p SNPs, there are $2^p - (p+1)$ LD terms of varying orders that need to be considered to fully characterize the complete LD structure in them. Better understanding of the detailed structure of multilocus LD can be useful in designing powerful statistical methods for disease gene mapping. Therefore, the objectives of this thesis are to propose a multilocus LD measure using the framework of multiple order Markov Chain (MOMC) model which was initially proposed by Feng (2004) and to demonstrate its utility in characterizing the local LD pattern and providing more detailed information about multilocus LD structure in the human genome from the HapMap project.

To this end, the thesis is structured as follows. Chapter 2 introduces a novel approach of characterizing the LD pattern on a chromosome region using multiple order Markov Chain (MOMC) models and a dynamic window algorithm. To illustrate our approach, we detect haplotype blocks and recombination hotspots in the *Drosophila* data of DeLuca et al. (2003) and ENm010 region (chromosome 7p.15) from the HapMap project and compare them with those detected by two other common approaches. In Chapter 3, we employ the same model (MOMC) as a statistical framework to measure the extent of the overall departure from linkage equilibrium, called the total multilocus LD, in a chromosome region and partition it into various orders of LD due to two-, three-locus association, and so on. We also explore the relationship between different orders of LD and different orders of MC models. Since the MOMC models consider consecutive markers in sliding windows, the association among markers distantly spaced cannot be captured. Therefore, in Chapter 4 we consider all combinations of three markers and test for the significances of two-and three-locus associations at each set of three SNPs by adopting the ideas from Hill (1975, 1976) and Long et al. (1995). The results of systemic survey of two- and three-locus disequilibria in one ENCODE region are presented in 3 dimensional LD plots which enable us to identify the regions where 2-locus or 3-locus disequilibria among multiple markers stand out after a FDR adjustment for multiple testing. Finally, Chapter 5 provides general discussion and conclusions of the main results. Advantages as well as limitations of the approaches are discussed and some potential problems for further studies are addressed.

Chapter 2

Modeling the Local Linkage Disequilibrium Pattern by Multiple Order Markov Chains and Dynamic Window Algorithm

2.1 Introduction

Testing for the presence of LD and measuring its value have received a great deal of attention from many research groups studying the human genome. The first studies of LD were mainly in the context of population genetics. For example, the disequilibrium between markers was used to assess the age of various populations. These days, measures of LD have been rediscovered as a tool for disease mapping in which LD between an unknown disease locus and a known set of markers is investigated (Sabatti et al. 2002). The pattern of local LD along a chromosome has been studied extensively to assist in designing genetic association studies.

Recent findings on the pattern of LD reveal a block-like LD structure in which LD decreases very little with distance between markers and the diversity of haplotypes is limited concomitantly. Between these blocks, however, LD is observed to decay rapidly

with physical distance (Daly et al. 2001; Patil et al. 2001; Abecasis et al. 2001; Reich et al. 2001; Gabriel et al. 2002; Dawson et al. 2002; Clark et al. 2003; McVean et al. 2004).

These blocks can be applied to several situations. First, the existence of haplotype block structure has significant implications for association-based methods for mapping disease genes (Wall et al. 2003). Common haplotypes within a block can be used as a multi-allelic marker. They would increase heterozygosity and power for disease association when compared to single SNP markers (Morris et al. 2002; Zhang et al. 2002 a; Klein et al. 2005). Second, if the diversity of haplotypes within blocks is low, these haplotypes can be differentiated by haplotype tagging SNPs (htSNPs). Within a haplotype block, the correlation between neighboring SNPs is sufficiently high to ensure that a few htSNPs can describe the common haplotypes over a substantial range. The use of htSNPs can reduce the number of SNPs that might be tested in either a candidate gene association study or a genome-wide association scan, and thus enables a study of larger populations with reduced genotyping cost (Chapman et al. 2003; Weale et al. 2003; Stram et al. 2003; Carlson et al. 2004). On the other hand, outside of a haplotype block, neighboring SNPs may be essentially uncorrelated and thus study of the region requires nearly all the SNPs in these intervals. Hence, blocks suggest where SNPs should be spaced denser or sparser in genomewide studies. Third, blocks themselves are interesting features of human genomic structure. The discovery of blocks suggests that the recombination events do not occur with uniform probability along the genome (Nothnagel et al. 2002).

A variety of methods to summarize the local LD pattern have been proposed recently and most of them define the block-like structure of LD. In this study, we propose a new haplotype block finding method using multiple order Markov Chain models and a dynamic window algorithm. A slightly modified BIC (i.e. the likelihood of the data minus a penalty for model complexity) is used as a test statistic to evaluate competing models associated with Markov Chain models of different orders. A dynamic window algorithm is designed to compensate for limitations of a fixed window size approach. The extent of LD for consecutive two markers is expressed in terms of the order of Markov Chain, with high MC order indicating high LD and MC 0 indicating linkage equilibrium. The determined MC order for adjacent markers provides convenient profiles to describe the LD pattern along a chromosome. We apply our method to the Drosophila data of DeLuca et al. (2003) and to the phased haplotype data in ENm010 region (chromosome 7p15.2) from the HapMap project (The International HapMap Consortium 2005). The blocks and recombination hotspots found by our method are in general good agreement with D' method (Gabriel et al. 2002) and four-gamete test method (Wang et al. 2002). All three methods provide evidence for a block-like organization of the genetic variation in the region. However, thorough comparison of block characteristics determined by three different methods reveals that there's no consistency among them.

2.2 Methods

2.2.1 Multiple order Markov Chain (MOMC) Models and Likelihood

In this section, we formulate our statistical model and elaborate its theoretical foundation. The model was initially proposed by Feng (2004). We consider a population in Hardy-Weinberg equilibrium. Our current method assumes that phased haplotypes are available and there are no genotyping errors. The phased haplotype data can be obtained either from laboratory techniques such as long-range PCR or chromosomal isolation (Michalatos-Beloin et al., 1996; Douglas et al., 2001) or reconstructed from diplotype data by some statistical or computational methods (Clark et al., 1990; Excoffier et al., 1995; Niu et al., 2002).

A Markov chain is a probabilistic model that can be used to represent dependencies between successive observations of random variables. In this paper, we consider each SNP as a discrete random variable (M_i) taking values in the finite set {0, 1}. For the first-order Markov chain model, denoted as MC1, an observation at a marker depends only on the observation of its adjacent marker on the left or right and is independent of others. Similarly, for the *r*th order Markov chain mode, denoted as MC*r*, an observation at a marker, say M_i, depends on the observations of previous *r* markers from M_{i-1} to M_{i-r} for example. The Markov chain model applied in this study is non-homogenous in that all transition probabilities from one state to another state are locus-specific (or marker-specific).

Consider a dataset containing N independent phased haplotypes of a number of markers from a natural population. Suppose each phased haplotype c (c = 1, 2, . . , N) has L SNP markers, with two alleles at each SNP M_i (i = 1, 2, . . , L) coded as 0 or 1. Let $m_{i,c}$ ($m_{i,c}$ = 0 or 1) denote the observed allele of M_i on the haplotype c. Let $H(i, r)$ denote the haplotype fragment containing r consecutive markers starting from the ith marker (i.e. $M_i, M_{i+1}, ..., M_{i+r-1}$). Let $h_c(i, r)$ denote the observed $H(i, r)$ on a particular chromosome c. Note that $h_c(i, r)$ is a $r \times 1$ vector of 0 and 1. Then, $P[H(i, r) = h_c(i, r)]$ represents the population frequency of the specific haplotype fragment $h_c(i, r)$. Let $P[M_i = m_{i,c} \mid H(i\text{-}r, r) = h_c(i\text{-}r, r)]$ denote the conditional probability that a marker M_i takes value $m_{i,c}$ ($m_{i,c}$ = 0 or 1) given its previous r markers $M_{i\text{-}r}, . . , M_{i\text{-}1}$ taking values $h_c(i\text{-}r, r)$. If we assume that the haplotype counts follow a multinomial distribution, the maximum likelihood estimate (MLE) of $P[H(i\text{-}r, r) = h_c(i\text{-}r, r)]$ is

$$\hat{P}[H(i-r,r) = h_c(i-r,r)] = [\text{count of } h_c(i-r,r)] / N$$

Thus, the conditional probability $P[M_i = m_{i,c} \mid H(i\text{-}r, r) = h_c(i\text{-}r, r)]$ can be estimated as

$$\hat{P}[M_i = m_{i,c} \mid H(i-r,r) = h_c(i-r,r)] = \frac{\hat{P}[H(i-r,r+1) = h_c(i-r,r+1)]}{\hat{P}[H(i-r,r) = h_c(i-r,r)]} .$$

We illustrate our model using a small hypothetical dataset in which each chromosome c consists of only 3 SNP markers of M_1, M_2, M_3. Let $L_c(MCr)$ be the likelihood of observing (M_1, M_2, M_3) assuming MCr on chromosome c.

$$L_c(MC0) = P(M_1) \times P(M_2) \times P(M_3)$$

$$L_c(MC1) = P(M_1) \times P(M_2 \mid M_1) \times P(M_3 \mid M_2) = P(M_1) \times \frac{P(M_1, M_2)}{P(M_1)} \times \frac{P(M_2, M_3)}{P(M_2)}$$

$$L_c(MC2) = P(M_1, M_2) \times P(M_3 \mid M_1, M_2) = P(M_1, M_2) \times \frac{P(M_1, M_2, M_3)}{P(M_1, M_2)}$$

For C independent chromosomes, the likelihood of MCr model is

$$L(\mathrm{MC}r) = \prod_c L_c(\mathrm{MC}r)$$

This likelihood will be the same regardless of the orientation to which a Markov chain moves. That is, L(MCr) calculated from left to right is equal to that calculated from right to left.

There are two types of parameters in the likelihood of MCr model: the (2^r-1) chain initiating probabilities and $2^r \times$(window size - r) transition probabilities. Assuming that all the possible haplotypes are observed, the total number of parameters in MCr model will be the sum of two types of parameters (See Table 2.1). MLEs of both types of parameters can be obtained by assuming that the haplotype counts follow a multinomial distribution. Then, the likelihood $L(MCr)$ is calculated by replacing all parameters with their MLEs. If some of multilocus haplotypes are unobserved due to finite sample size, the associated parameters cannot be estimated and thus 0 will be assigned to them. However, the unobserved haplotypes do not contribute to the likelihood of different orders of MC model by the convention of $0 \times \log(\text{non-negative value}) = 0$ (Liu et al., 2005).

Table 2.1 Number of parameters in MC *r* model with *w* markers

w	MC 0	MC 1	MC 2	MC 3	MC 4	. . .	MC *r*
1	1	-	-	-	-	-	-
2	2	3	-	-	-	-	-
3	3	5	7	-	-	-	-
4	4	7	11	15	-	-	-
5	5	9	15	23	31	-	-
…	-	-	-	-	-	-	-
w	*w*	2*w*-1	4*w*-5	8*w*-17	16*w*-49	. . .	$(2^r-1)+2^r(w-r)$

Assuming that all the possible haplotypes in a window of size *w* are observed, this table shows the total number of parameters in MC *r* model.

2.2.2 Dynamic Window Algorithm

We develop a dynamic window algorithm to find the highest MC order for a genomic region and score for markers in the region. Unlike most sliding window approaches in which the window size is fixed, the dynamic window algorithm dynamically changes the window size when a Markov Chain moves along a chromosome. Specifically, the smallest window size in the algorithm is always two and the largest window size is a pre-defined maximum window size. In this approach, we set the maximum window size to be 15. In theory, an unlimited number of loci can be used as a window size. However, since the haplotype frequencies for long sequences are hard to estimate reliably, the number of loci in a window needs to be limited. Although the number of loci that can be considered at one time is limited by the pre-defined maximum window size, it does not prevent long-range LD from being detected because a long series of the maximum MC order indicates long-range LD region.

Now, we describe the dynamic window algorithm in more detail. For a certain window size of *w*, we compare two models. One is the highest MC order model (MC *w*-1) which we call the full model and the other is the second highest MC order model (MC *w*-2) called the reduced model. If the test statistic (see below) from the reduced model is larger, the algorithm moves to the next marker and the window size is set to be the smallest one. If the test statistic from the full model is larger, the algorithm increases the window size by one by adding one previous marker. If the test statistics from the full model and the reduced model are equivalent, the algorithm chooses the reduced model. By changing the window size dynamically, the algorithm finds the highest MC order that fits the markers in the window and assigns the highest MC order to all markers except the last marker in the window. We intentionally do not assign the highest MC order to the last marker in the window to prevent the assigned MC order from being overridden by the MC order of markers in next window. The window size keeps increasing until MC order does not increase any more or it hits the maximum window size which we set in advance. The final output of the algorithm is the highest MC order that each consecutive two markers can take. The algorithm is summarized in a flow chart in Figure 2.1.

We use a slightly modified BIC as a criterion to compare two competing models. Using the following definition, the MC model with the larger BIC is chosen in a given window. Let $S(r, i, j)$ denote the test statistic associated with a Markov Chain model of order r for markers from i to j.

$$S(r,i,j) = \log(L(MCr_{(i,j)})) - \frac{d_r}{2}\log(n) \quad ,$$

where $L(MCr_{(i,j)})$ is the total likelihood of MC r model computed for markers from i to j; n is the sample size. d_r is minimum of $[(2^r - 1) + 2^r(j - i + 1 - r), n_{h(i,j)} - 1]$, where the first term is the theoretically possible number of parameters for MC r model for markers from i to j and the second term is the observed number of haplotypes in a block of markers from i to j minus 1.

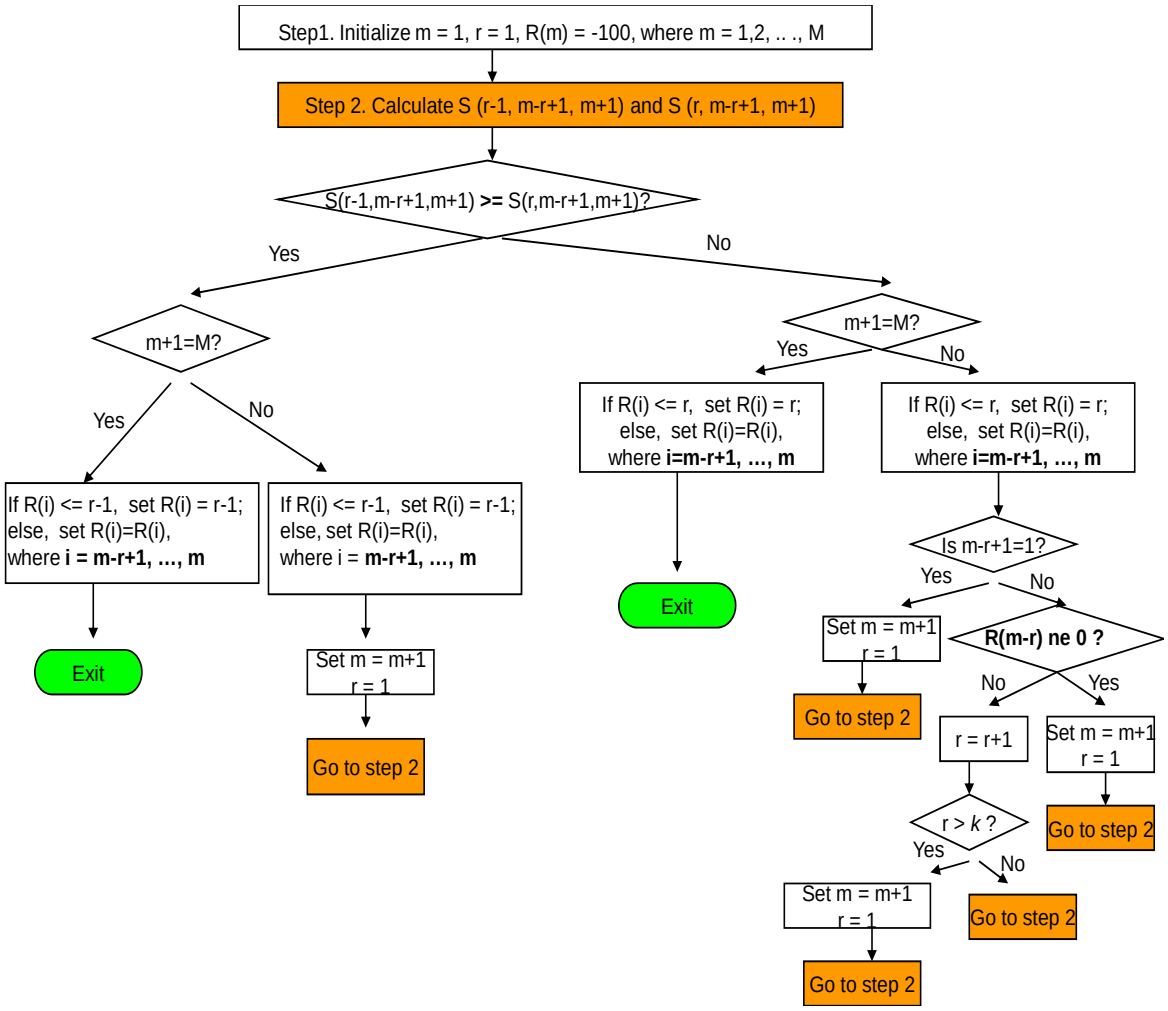


Figure 2.1 Flow chart for a dynamic window algorithm M is the total number of SNPs in each chromosome. S(r, i, j) denotes a test statistic associated with MC model of order r from SNPs from i to j. $R(m)$ denotes the inferred MC order for SNP m, where m = 1, 2, …, M. k denotes a pre-determined maximum window size.

2.3 Results

2.3.1 Analysis of LD pattern in *Drosophila* Data of DeLuca et al. (2003)

We apply the proposed multiple order Markov Chain (MOMC) model and the dynamic window algorithm to the *Drosophila* data of DeLuca et al. (2003) within the 5.5 kb *Ddc* gene region. We observe some distinctive LD patterns between polymorphic sites throughout the 5.5 kb region. *Ddc* encodes Dopa decarboxylase (DDC), which catalyzes the decarboxylation of L-dopamine to dopamine and 5-hydroxytryptophan to serotonin (Blenau et al. 2001). DDC is required for the production of dopamine and serotonin in the central nervous system and in the hypoderm, where it is necessary for sclerotization and melanization of the cuticle (Lunan et al. 1969). Dopamine and serotonin affect mating behavior, fertility, circadian rhythms, endocrine secretion, aggression and learning and memory (Blenau et al. 2001). DeLuca et al. (2003) investigated whether *Ddc* is associated with variation in individual life span. 36 biallelic markers (31 SNPs, 5 insertion / deletion markers) are genotyped. Using cross-over suppressor stocks in *Drosophila melanogaster*, individual chromosomes are made homozygous and their phases are easily determined. The haplotype data are available by DNA sequencing on 173 *Drosophila* lines derived from the Raleigh population. Markers from 1 to 21 are in the promoter region and their physical positions range from 236 bp to 731 bp. In this region, the markers are separated by 2 to 88 bp. The remaining 15 markers are in the coding region with their physical locations from 758 bp to 4694 bp. The distance between markers in the coding region is greater (1 to 943 bp), compared to marker distances in the promoter region.

Figure 2.2 presents the determined MC orders for 36 biallelic markers. In the Figure, the first point at (1.5, 1) indicates that MC 1 is determined between marker 1 and marker 2, meaning that there's one-step dependency between these two markers. MC 0 is determined for markers 2 - 4, indicating that there is no significant dependency among these three markers. Although no association is found for markers 2 - 4, 8 - 9, 10-11, 14 – 16 in the promoter region, overall MC order is higher among markers typed in the promoter region than among markers in the coding region. This result may suggest that the extent of LD is higher in the promoter region. Fewer recombination rates in the region may play a role in causing the high LD. Summarizing the LD patterns in this way provides more direct local information than two dimensional pairwise LD plot reported by DeLuca et al. (2003). In their pairwise LD plot, many significant two-locus LD are detected, but direct information on local regions of high LD cannot be obtained easily.

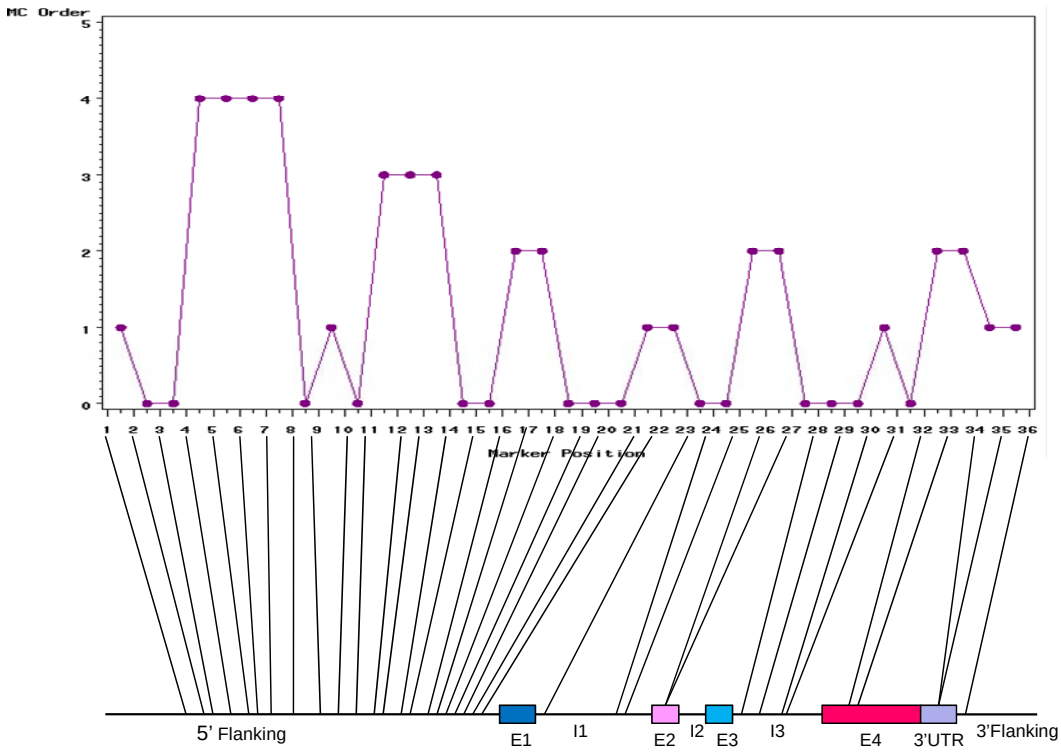


Figure 2.2 LD profile over 36 markers in the *Drosophila* data of DeLuca et al. (2003) in terms of MC order. MC order is determined for markers 1 to 36. Markers from 1 to 22 are in the promoter region and markers from 22 to 36 are in the coding region of Ddc gene. Regions covered by markers 4-8 and markers 11-14 show high MC order, indicating high LD in these regions. Marker positions do not represent their physical distances.

2.3.2 Analysis of one ENCODE region (chromosome 7p15.2) in the HapMap project

We apply the proposed algorithm to the phased haplotype data in ENm010 region (chromosome 7p15.2) downloaded from the HapMap project website. The phased haplotype data were generated using the PHASE v2.1 software and compiled from the genotype data Phase I / rel#16a. The Phase I data include a common SNP (MAF $\geq$ 0.05) every 5 kb across most of the genome in each population. Across the ten ENCODE

regions, the density of SNPs is one per 279bp on average, which is approximately tenfold higher than the Phase I genome-wide data. Phase I / rel#16a data files contain four files of phased haplotypes corresponding to each of the four populations. Table 2.2 compares the datasets of four populations in chromosome 7p15.2. We focus on SNPs with minor allele frequency (MAF) >= 0.05 in each population.

Table 2.2 Number of SNPs and chromosomes in the datasets of each of four populations in ENm010 region

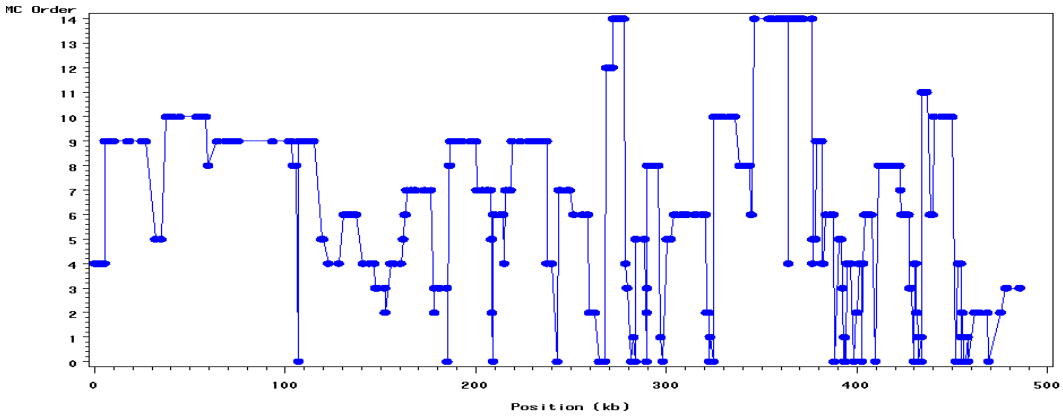
	YRI	CEU	HCB	JPT
# of SNPs in the original dataset	593	618	508	508
# of SNPs with MAF >=0.05	433	471	322	316
# of chromosomes	120	120	90	88

YRI: 90 individuals (30 parent-offspring trios) from Yoruba in Ibadan, Nigeria
CEU: 90 individuals (30 parent-offspring trios) from Utah, U.S.A.
HCB: 45 Han Chinese in Beijing, China
JPT: 44 Japanese in Tokyo, Japan

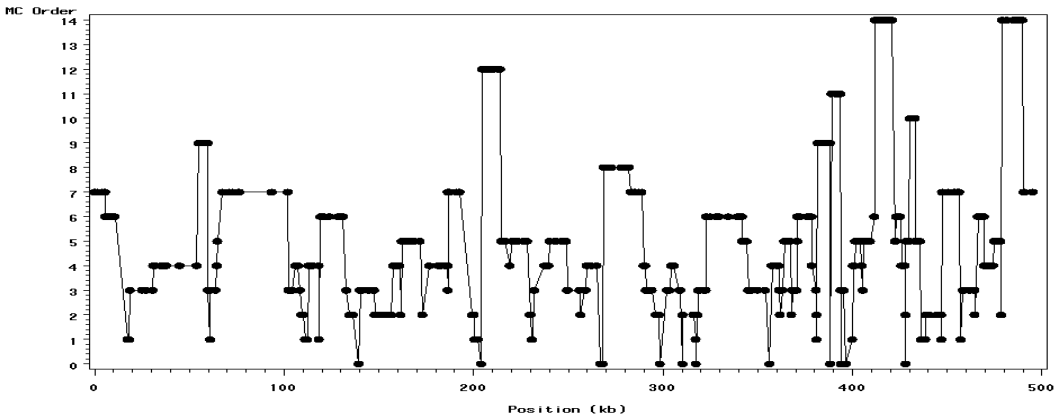
Linkage disequilibrium patterns for each population in ENm010 region are depicted in terms of MC order and compared with a recombination plot in Figure 2.3. The recombination plot in the bottom were reported by the International HapMap Consortium (2005, Supplementary Figure 7). In the top four plots, the maximum window size is set to be 15 for the dynamic window algorithm. So, the highest MC order that can be reached is 14. Positions of MC order 0 in top four plots reasonably match with the recombination hotspots in the bottom plot which shows the estimated recombination rate in ENm010 region. In particular, the highest peak near 400 kb in the bottom plot perfectly matches with the locations of MC order 0 in the top four plots. Besides, locations of many medium or low recombination rates in the bottom plot coincide with those of MC order 0 in the top plots. However, there are some regions (e.g. 0 ~150 kb region) where MC order 0 is determined but the same region appears to be complete recombination cold spots in the bottom plot.

Figure 2.3 Comparison of linkage disequilibrium pattern in terms of MC order and recombination rates for four populations in ENm010 region. Positions of MC order of 0 in the top four plots coincide with the recombination hotspots identified in the bottom plot. The bottom plot which was reported by the International HapMap Consortium (2005, supplementary Figure 7) shows the estimated recombination rate in ENm010 (chromosome 7p.15) region. Recombination hotspots are indicated as red triangles.

YRI population in ENm010 region



CEU population in ENm010 region



HCB population in ENm010 region

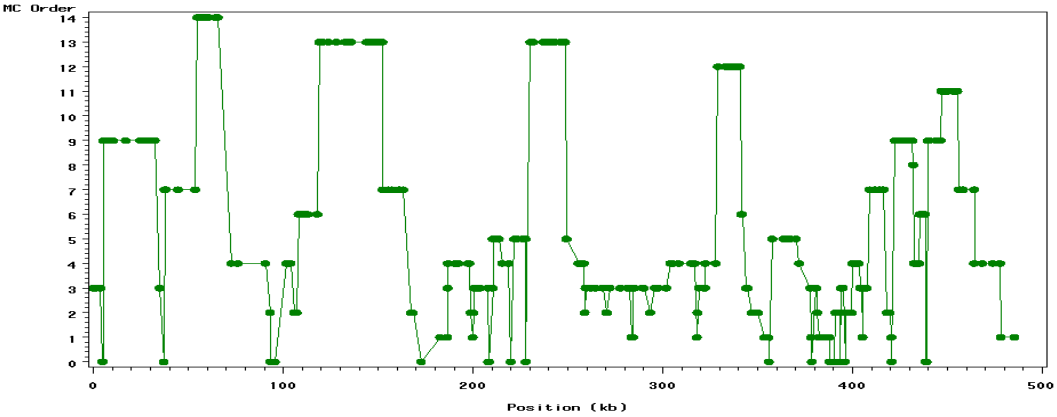
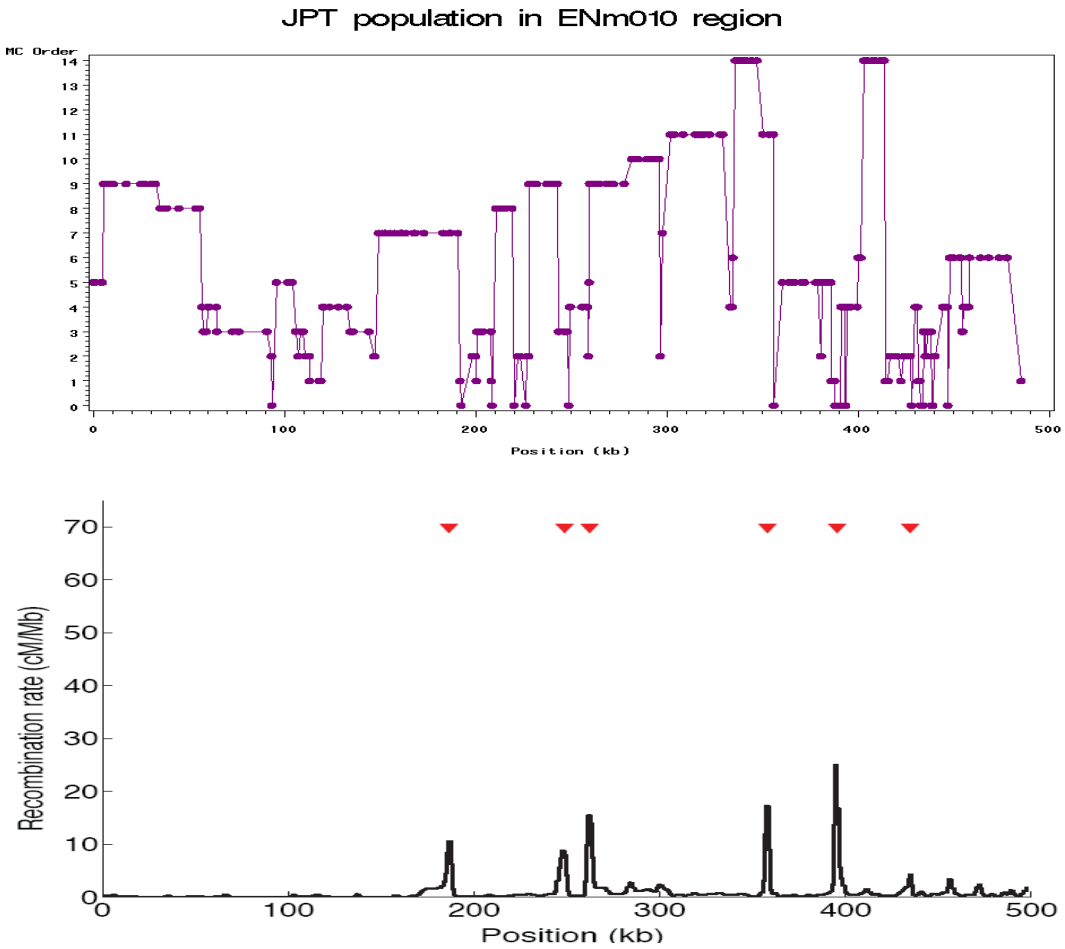


Figure 2.3 (continued)



The block lengths and chromosomal coverage computed in each population reveals that in these very dense and complete data, most of the sequence belongs to a block-like structure of LD that contains many SNPs covering regions of 0.2 kb - 136 kb. However, different definitions of a block using different thresholds (MC order of 1, MC order of 2, and so on) result in various sizes of blocks and chromosomal coverages. For low thresholds, blocks become large and may contain regions of low LD. For more stringent thresholds, coverage decreases but contain regions of higher LD. By increasing

thresholds, the average block length becomes shorter and the chromosomal coverage decreases. So, the optimal criterion for the block definitions is a trade-off between short blocks with purely high LD and longer blocks with mixture of high and low LD. The choice of an optimal threshold in the proposed algorithm needs full evaluation using simulated datasets, hypothesis testing, or cross-validation, but we did not try them in this study. Table 2.3 tabulates the results for different thresholds. Consistent with the results from other groups (Gabriel et al. 2002, Hinds et al. 2005), our result shows that LD extends to a similar and long extent in the Asian samples (JPT and HCB) and the sample from Utah in USA (CEU) and that the block sizes are longer in these three populations (JPT, HCB, and CEU) as compared to those in the sample from an African population (YRI).

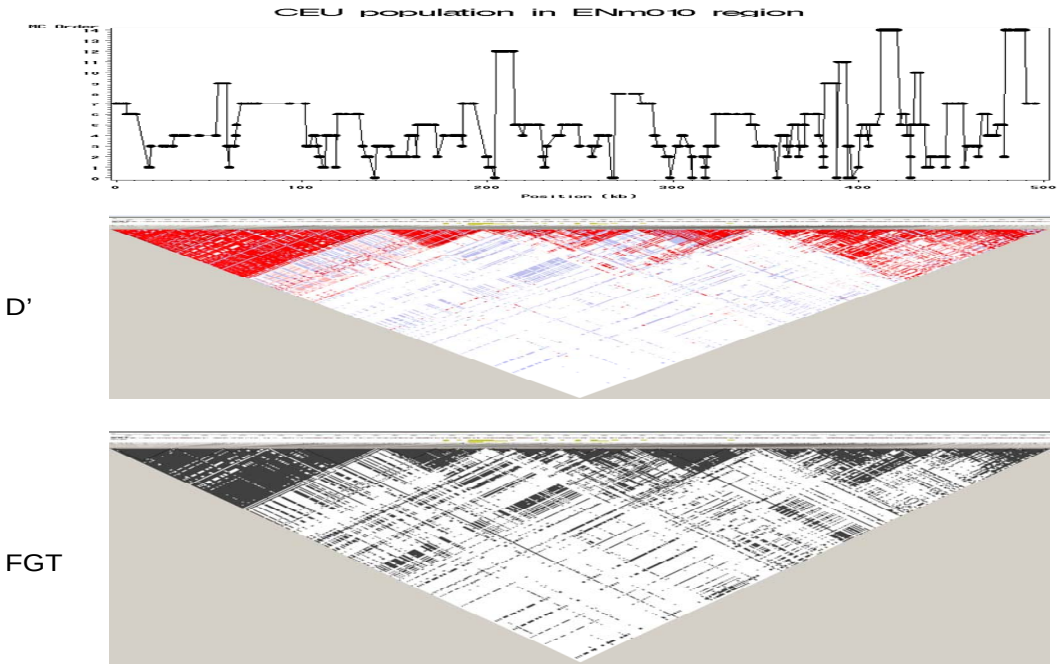
Table 2.3 Summary statistics for block characteristics using two different thresholds

Characteristics	Threshold	YRI	CEU	HCB	JPT
Number of blocks	MC order ≥ 1	23	12	14	14
	MC order ≥ 2	21	21	19	18
Average Number of SNPs in a block (Min, Max)	MC order ≥ 1	18 (2, 75)	38 (3, 90)	22 (3, 82)	21 (2, 65)
	MC order ≥ 2	19 (2, 75)	21 (2, 52)	14 (3, 39)	16 (2, 65)
Average Block size (kb) (Min, Max)	MC order ≥ 1	20.1 (0.2, 106)	39.3 (1.1, 136)	31.6 (1.1, 125)	32.4 (1.5, 107)
	MC order ≥ 2	20.9 (0.2, 106)	21.3 (1.1, 49)	21.4 (1.0, 67)	23.6 (1.1, 107)
Fraction of genome spanned by blocks (%)	MC order ≥ 1	88.8	94.8	90.4	92.8
	MC order ≥ 2	88.4	89.9	83.2	86.8

Next, we compare several block characteristics of each population using the proposed algorithm with those using two other methods. The results are presented in Table 2.4. Two popular methods implemented in Haploview (version 3.2) – one based on a composite of local D' values (Gabriel et. al 2002) and the other based on the four gamete test (Hudson et. al 1985; Wang et. al 2002)) - are used for comparison. Figure 2.4 illustrates the resulting block-like structure of LD from three different block defining methods in ENm010 region. Although high LD regions are discerned from low LD regions at similar points from three methods, the exact boundaries of each block do not

match in three methods. Both D' and our method detect much longer blocks than four gamete test. Regions of high / low LD are clearly differentiated in all three methods and their locations are approximately the same. Four gamete test yields more blocks but most of them have smaller sizes than the other two methods. Our method using a low threshold (i.e. MC order ≥ 1) shows the greatest chromosomal coverage.

Figure 2.4 Comparison of haplotype blocks using three different methods in ENm010 region. For D' plots, different colors indicate the extent of LD: bright red for $D' = 1$ and LOD (the log of the likelihood odds ratio, a measure of confidence in the value of D') ≥ 2; blue for $D' = 1$ and LOD < 2; white for $D' < 1$ and LOD < 2; pink for $D' < 1$ and LOD $\geq$ 2. For the four gamete test plots, white color indicates 4 distinct haplotypes are observed for two markers while the black color indicates less than 4 distinct haplotypes are observed for two markers.


CEU population in ENm010 region
Position (kb)
D'
FGT

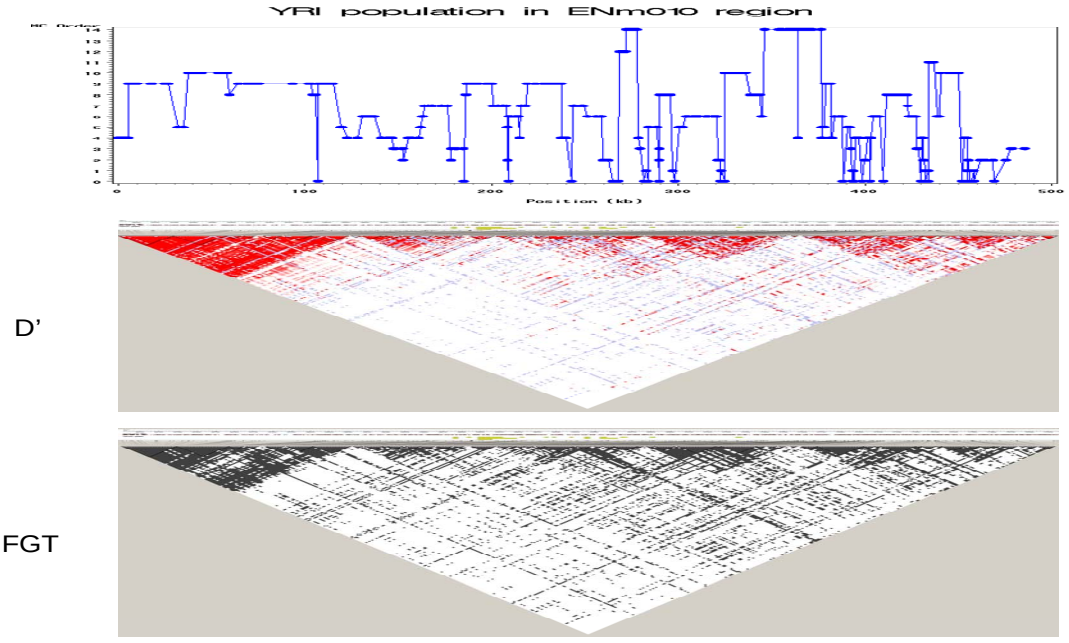
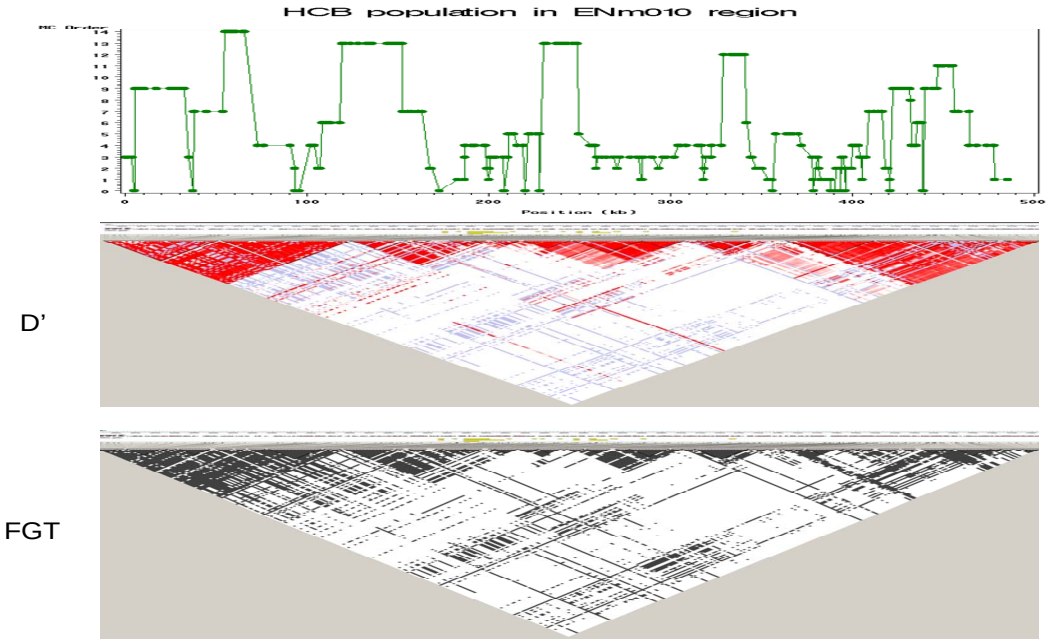

YRI population in ENm010 region
Position (kb)
D'
FGT

Figure 2.4 (continued)



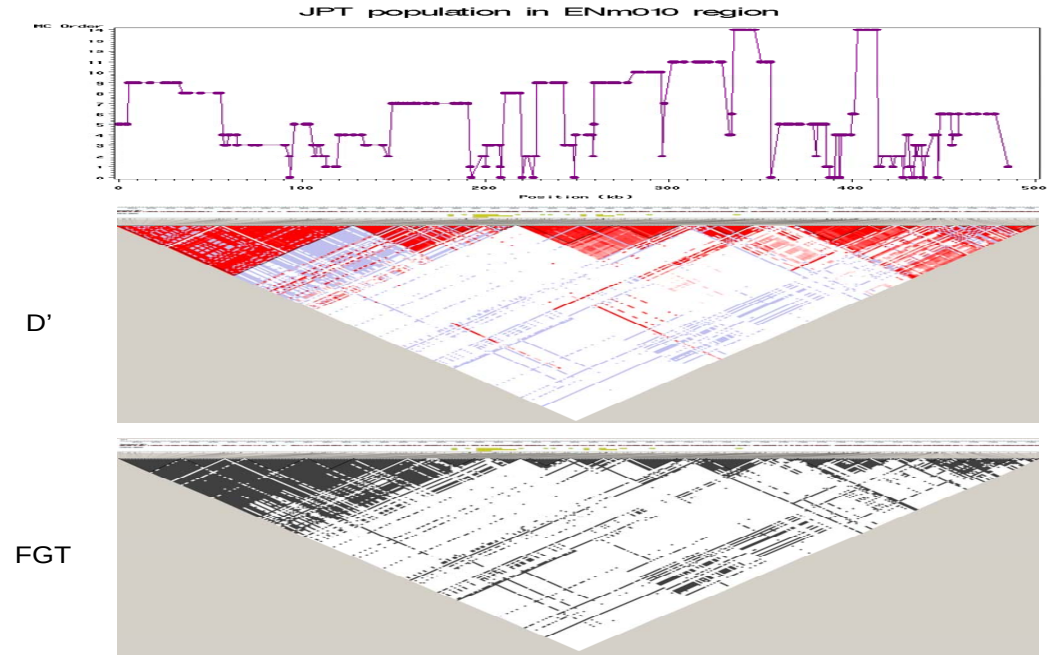


Table 2.4 Haplotype block characteristics according to three different methods

Characteristics	Method	YRI	CEU	HCB	JPT
Number of blocks	Gabriel (2002)	26	24	17	18
	Wang (2002)	52	51	49	41
	Our method (MC order ≥ 1)	23	12	14	14
Average number of SNPs per block (min, max)	Gabriel (2002)	11 (2, 52)	18 (2, 56)	16 (3, 41)	14 (2, 79)
	Wang (2002)	7 (2, 24)	9 (2, 36)	6 (2,17)	7 (2, 18)
	Our method (MC order ≥ 1)	18 (2, 75)	38 (3, 90)	22 (3, 82)	21 (2, 65)
Average block size (kb) (min, max)	Gabriel (2002)	12.3 (0.001, 113)	17.9 (0.021, 102)	24.4 (2.4, 93)	21.9 (0.8, 182)
	Wang (2002)	6.8 (0.001, 56)	7.5 (0.073,45)	7.0 (0.145, 38)	8.5 (0.145, 55)
	Our method (MC order ≥ 1)	20.1 (0.2, 106)	39.3 (1.1, 136)	31.6 (1.1, 125)	32.4 (1.5, 107)
Fraction of genome spanned by blocks (%)	Gabriel (2002)	64.1	86.3	84.9	80.7
	Wang (2002)	70.5	76.7	69.7	71.4
	Our method (MC order ≥ 1)	88.8	94.8	90.4	92.8

2.4 Discussion

A novel method is proposed to describe multilocus LD pattern on densely spaced marker data using multiple order Markov Chain (MOMC) models and a dynamic window algorithm. The magnitudes of LD along a chromosome are represented in terms of the order of Markov Chain (MC), high MC order indicating high LD region. The application of our method to both the *Drosophila* data reported by DeLuca et al. (2003) and the HapMap ENCODE data demonstrate that the MC order reasonably describes LD pattern and block structures. The analysis of the *Drosophila* dataset shows that high MC order is selected in regions of low recombination rate while low MC order is selected in regions of high recombination rate. Overall, the MC order in the *Drosophila* dataset is not very high and diverse haplotypes are observed even for long sequence of markers. In contrast to the *Drosophila* data, we detect strikingly different patterns of LD from the HapMap ENCODE data. The diversity of haplotypes for a long sequence of markers is very limited and a long series of high MC order is often selected. We also observe short regions in which MC order suddenly drops down to MC 0 after a long series of high MC order. Comparison with the estimated recombination rate plot in ENm010 region reveals that the regions of MC order 0 correspond to the recombination hotspots.

Our approach characterizes multilocus LD patterns using the haplotype frequencies observed in the sample and compares models via the BIC criterion. We do not make any assumption on the modeling of the origin of LD such as the population history and the recombination rates since it might be inappropriate to model LD with certain simplified biological assumptions in a natural population without a clear understanding of the history of the populations for the genomic region. Therefore, our

results represented by the order of MC along the chromosome can be regarded as just statistical description of correlations among multiple markers. In that sense, this approach is similar to D-based pairwise measures, but in contrast to some parametric LD measures (Morton et al. 2001; Pritchard et al. 2001; Kristin et al. 2002).

A dynamic window algorithm is designed to compensate for limitations of a fixed window size approach. In a fixed window size approach, the results are highly dependent on the window size (w) and this fact confounds with the true extent of LD when the LD measures are expressed in terms of MC order. Moreover, the optimal window size is hard to choose and it should be influenced by the underlying LD pattern. Since LD patterns vary substantially from region to region, use of a fixed window size becomes more problematic when a large genomic region or the whole genome is analyzed. Instead of fixing a window size, the dynamic window algorithm starts with the smallest window size and keeps increasing the window size until MC order doesn't increase any more or the pre-defined maximum window size has been reached. In theory, the window size can be unlimited, but we set the maximum window size because the number of haplotypes increases exponentially as the number of loci increases. Since we are interested in finding the highest order of LD in a local region, we only compare two models (one model with the highest MC order and the other model with the second highest MC order) in a window of size w, not checking all possible models in a window (i.e. from MC 0 to MC w-1) exhaustively. However, since we always start the window size from the smallest one and increase it by one by adding one previous marker whenever the test statistic of full model is larger, it is essentially all possible models in each window are checked indirectly.

The concept of haplotype blocks has received a lot of attention due to its possibility as multiallelic markers with increased heterozygosity in association studies and to describe human genetic variation in an efficient and simple way. However, the observed inconsistency in block characteristics among different methods illustrates the subjectivity of haplotype block definition and prevents the conclusive characterization of the region's block structure. Haplotype block assignment is found to change not only due to different methods but also as a result of altering key parameters such as thresholds within the same set of definition criterion. Studies using dense datasets have suggested that while the haplotype block model provides a useful view of broad patterns of LD, specific attributes of blocks, particularly block boundaries, can be strongly influenced by the markers and samples chosen for analysis (Phillips et al., 2003; Ke et al., 2004; Sun et al., 2004). There are also multiple statistical definitions of blocks which yield similar, but non-identical patterns. Thus, some concerns about both the suitability of ad hoc approaches for the crude identification of block structure and the validity of the notion of haplotype blocks as a genomic feature have been raised (Schulze et al., 2004; Zeggini et al., 2005; Ding et al., 2005; Nothnagel et al., 2005).

Chapter 3

Measuring and Partitioning the High Order Linkage Disequilibrium by Multiple Order Markov Chains

Yunjung Kim, Sheng Feng, Zhao-Bang Zeng (2008).

Accepted in *Genetic Epidemiology*

3.1 Abstract

A map of the background levels of disequilibrium between nearby markers can be useful for association mapping studies. In order to assess the background levels of linkage disequilibrium (LD), multilocus LD measures are more advantageous than pairwise LD measures because the combined analysis of pairwise LD measures is not adequate to detect simultaneous allele associations among multiple markers. Various multilocus LD measures based on haplotypes have been proposed. However, most of these measures provide a single index of association among multiple markers and does not reveal the complex patterns and different levels of LD structure. In this paper, we employ non-homogenous, multiple order Markov Chain (MOMC) models as a statistical framework to measure and partition the LD among multiple markers into components due to different orders of marker associations. Using a sliding window of multiple markers on phased haplotype data, we compute corresponding likelihoods for different MC orders in each window. The log-likelihood difference between the lowest MC order model (MC0) and the highest MC order model in each window is used as a measure of the total LD or the overall deviation from the gametic equilibrium for the window. Then, we partition the total LD into lower order disequilibria and estimate the effects from two-, three-, and higher order disequilibria. The relationship between different orders of LD and the log-likelihood difference involving two different orders of MC models are explored. By applying our method to the phased haplotype data in the ENCODE regions of the HapMap project, we are able to identify high / low multilocus LD regions. Our results reveal that the most LD in the HapMap data is attributed to the LD between adjacent pairs of markers across the whole region. LD between adjacent pairs of markers appears

to be more significant in high multilocus LD regions than in low multilocus LD regions. We also find that as the multilocus total LD increases, the effects of high order LD tends to get weaker due to the lack of observed multilocus haplotypes. The overall estimates of first, second, third, and fourth order LD across the ENCODE regions are 64%, 23%, 9%, and 3%.

3.2 Introduction

Assessing the patterns of linkage disequilibrium along a chromosome has been an important issue in disease mapping studies and in studies of history of humans and other species. In particular, in an association mapping of disease genes, the inference is usually based on a linkage disequilibrium test for association between genetic variation at a known set of markers and disease phenotypes. If such an association is detected between a particular marker locus and the phenotype, it suggests that either the variation at that marker locus affects the phenotype of interest, or that the variation of that marker locus is in LD with the true phenotype-related locus, which was not genotyped. The association signal and the LD pattern in the region provide important information about candidate location of disease genes. Many studies have indicated that the levels of disequilibrium vary a lot across genomic regions and populations. To design and interpret disease mapping studies, one needs to refer to a map of the background levels of disequilibrium that can be expected in a given region and in a given population. To construct such a LD map, the levels and patterns of disequilibrium between close-by markers need to be measured effectively (Sabatti et al., 2002; Muller, 2004; Greenspan et al., 2006).

In association mapping, many investigators have studied the question whether haplotype-based association tests are more powerful than single-locus tests. Akey et al. (2000) found that haplotype-based tests can improve the power of association mapping. In contrast, Long et al. (1999) and Kaplan et al. (2001) found that single-locus tests are as much powerful as haplotype-based tests. Nielsen et al. (2004) compared the power of haplotype and single-marker tests under different patterns of pairwise and three-locus LD. They found that haplotype-based tests tend to be more powerful when moderate to high levels of three-locus LD exist and single-marker tests tend to prevail when pairwise LD between the markers and the functional site is high.

Most previous studies on LD patterns and LD blocks have focused on using pairwise LD measures and have not examined and utilized joint multilocus LD (Nielsen et al. 2004). It is more advantageous to use multilocus LD to assess the background levels of LD because the combined analysis of all pairwise LD measures across a region is insufficient to detect simultaneous allele associations among multiple loci. An illustrative example can be found in Nielsen et al. (2004) and Muller (2004). Muller (2004) reviewed various multilocus LD measures. One common property of these methods is to summarize the data with a single multilocus LD measure by first calculating the difference between the observed state and the expected one under linkage equilibrium and then normalizing it to allow comparison across different regions and populations. The extent of normalized difference between the observed state and the expected state can be regarded as the degree of overall departure from linkage equilibrium. Although these methods overcome the limit of pairwise LD measure by considering joint multilocus LD, they do not distinguish the type and structures of

multilocus LD, such as how much multilocus LD is due to LD between marker pairs and how much due to higher orders. For example, if we consider p SNPs simultaneously, there are $2^p - (p+1)$ LD terms of varying orders that need to be considered to fully characterize the LD structure in them. In this paper, we develop a non-homogeneous, multiple order Markov Chain model for characterizing LD structure and use it to estimate the contributions of hierarchical structure of multiple markers to multilocus LD. This partitioning of LD will help us to better understand the LD structure and provide more useful information in designing appropriate methods for fine mapping of disease genes.

For this study, we are interested in measuring the extent of the overall departure from linkage disequilibrium, called the total multilocus LD, in a chromosome region and partitioning the total multilocus LD into various orders of LD due to two-, three-locus association, and so on. For this purpose, we employ a non-homogenous, multiple order Markov Chain (MOMC) model as a statistical framework. Our measure of the multilocus LD based on the MOMC model is very similar to that of Nothnagel et al. (2002) which was based on the concept of entropy. Both methods can describe the general LD patterns along a chromosome and detect haplotype blocks as one of applications. However, our MOMC model can provide more detailed information about the structure and detailed patterns of LD. To our knowledge, this is the first time that a model is constructed to partition the total multilocus LD into LD components of lower orders. We apply our method to the phased haplotype data in the ENCODE regions of the HapMap project. By partitioning the multilocus LD into different components, we observe that a great proportion of the total LD can be explained by LD between adjacent marker pairs. We also observe significant variation in the paritions of LD between chromosomal regions

and between populations. YRI population has more high order LD as compared to the other three populations (CEU, HCB, and JPT).

3.3 Methods

3.3.1 Multilocus LD measure

We now discuss our multilocus LD measure in this section within the framework of multiple order Markov Chain (MOMC) models which was elaborated in section 2.2.1. A sliding-window approach is used to compute the magnitude of multilocus LD along the region of interest and to partition the overall departure from linkage equilibrium (LE) into lower order disequilibria. Within each window of size *w*, Markov chain models for different orders from *0* to *w*-1 are applied to fit the haplotype data. Then, the log-likelihoods corresponding to different MC orders are computed. In each window, the MC0 model represents the random association of alleles from different marker loci (i.e. linkage equilibrium) while the MC*w*-1 model represents the full haplotype model with non-random association of alleles among all markers within a window. In between, the other MC models involve different levels of Markov properties (conditional independence). Note that the log-likelihood of a lower MC order is always smaller or equal to that of a higher MC order. That is, $LL(MC0) \leq LL(MC1) \leq \ldots \leq LL(MCw\text{-}1)$.

Using the log-likelihoods from different MC orders in each window, we define $\delta_1, \delta_2, ..., \delta_{w-1}$ as following.

$$\delta_1 = \frac{LL(MC0) - LL(MC1)}{LL(MC0)} \times \frac{w}{w-1}$$

$$\delta_2 = \frac{LL(MC0) - LL(MC2)}{LL(MC0)} \times \frac{w}{w-1}$$

...

$$\delta_{w-1} = \frac{LL(MC0) - LL(MCw-1)}{LL(MC0)} \times \frac{w}{w-1} \ ,$$

where $LL(MC0)$, $LL(MC1)$, …, $LL(MCw\text{-}1)$ denote the log-likelihoods computed from MC0, MC1, …, MCw-1 model, respectively and w denotes a window size. They measure the deviation from LE when the loci are modeled by different MC orders. $w/(w-1)$ is multiplied to make δ's range between 0 and 1. Before the normalization, δ's depend on the window size (w) and $(w-1)/w$ is the maximum value that they can reach. By multiplying $w/(w-1)$, the parameters are bounded as $0 \le \delta_1 \le \delta_2 \le ... \le \delta_{w-1} \le 1$ regardless of window size w, with 0 indicating linkage equilibrium and 1 indicating the complete linkage disequilibrium. It is interesting to note that δ_{w-1} coincides with the normalized entropy difference (ε') developed by Nothnagel et al. (2002).

3.3.2 Partition of the multilocus LD

The overall deviation from the linkage equilibrium (δ_{w-1}) in each window can be partitioned into the contribution from each lower order MC model in the following way:

$$\phi_1 = \frac{\delta_1}{\delta_{w-1}} = \frac{LL(MC1) - LL(MC0)}{LL(MCw-1) - LL(MC0)}$$

$$\phi_2 = \frac{\delta_2 - \delta_1}{\delta_{w-1}} = \frac{LL(MC2) - LL(MC1)}{LL(MCw-1) - LL(MC0)} \quad (1)$$

....

$$\phi_{w-1} = \frac{\delta_{w-1} - \delta_{w-2}}{\delta_{w-1}} = \frac{LL(MCw-1) - LL(MCw-2)}{LL(MCw-1) - LL(MC0)}$$

To show the parametric compositions of ϕ_i's in terms of marker LD measures, we can approximate the expected values of numerators and denominators of different ϕ_i ($i = 1,2,..,w-1$) in equations (1) by using a Taylor series expansion of order 2. This approximation can help us to understand and interpret the meanings of different quantities. It turns out that the numerators and denominators of different ϕ_i ($i = 1,2,..,w-1$) are related to the sum of the corresponding squared correlation coefficients of different markers at different orders. The detailed derivation of the approximate expectation of the log-likelihood difference between different MC order models (MC1, MC2, …, MC*w*-1) and the independent model (MC0) is given in Appendix A.

For example, for a window of size 2 consisting of two markers (M_i, M_j), the expected difference between log-likelihoods of MC1 and MC0 is related to a half of the square of the correlation coefficient between the two markers.

$$\frac{1}{N} E[LL(MC1) - LL(MC0)] = \sum_i \sum_j P_{ij} \log(\frac{P_{ij}}{P_{i.} \times P_{.j}}) \cong \frac{1}{2} r_{ij}^2 , \quad (2)$$

where N denotes the sample size; E denotes expectation; $r_{ij}^2 = \frac{(D_{ij})^2}{P_{i.} \times P_{.j} \times (1-P_{i.}) \times (1-P_{.j})}$; the two-locus linkage disequilibrium coefficient (D_{ij}) is defined as $P_{ij} - P_{i.}P_{.j}$. Similar approximation results for a pair of markers are found in two papers (Nothnagel et al., 2002; Liu et al., 2005). The second term in equation (2) is the same as the mutual information between two systems in information theory (Kullback, 1978; Goebel et al., 2005). Note that if we multiply $2N$ to equation (2), the expressions in (2) become the standard test of independence in a 2×2 contingency table, χ_1^2 test. Using the moment generating function, we can easily prove that the expressions in (2) follow a gamma distribution with two parameters ($\alpha = 1/2$, $\beta = 1/N$).

For a window of size 3 consisting of three markers (M_i, M_j, M_k) in this order, we have the following approximations.

$$\frac{1}{N} E[LL(MC1) - LL(MC0)] \cong \frac{1}{2} r_{ij}^2 + \frac{1}{2} r_{jk}^2 \quad ,$$

where $r_{ij}^2 = \frac{(P_{ij.} - P_{i..} \times P_{.j.})^2}{P_{i..} \times P_{.j.} \times (1-P_{i..}) \times (1-P_{.j.})}$ and $r_{jk}^2 = \frac{(P_{.jk} - P_{.j.} \times P_{..k})^2}{P_{.j.} \times P_{..k} \times (1-P_{.j.}) \times (1-P_{..k})}$.

Similarly,

$$\frac{1}{N} E[LL(MC2) - LL(MC0)] \cong \frac{1}{2} r_{ijk}^2 + \frac{1}{2} r_{ij}^2 + \frac{1}{2} r_{jk}^2 + \frac{1}{2} r_{ik}^2 \quad ,$$

where $r_{ijk}^2 = \frac{(D_{ijk})^2}{P_{i..} \times P_{.j.} \times P_{..k} \times (1-P_{i..}) \times (1-P_{.j.}) \times (1-P_{..k})}$ and the three-locus linkage disequilibrium coefficient (D_{ijk}) is defined as $D_{ijk} = P_{ijk} - P_{i..}D_{jk} - P_{.j.}D_{ik} - P_{..k}D_{ij} - P_{i..}P_{.j.}P_{..k}$.

Thus, $\frac{1}{N} E[LL(MC2) - LL(MC1)]$ is approximately equal to $\frac{1}{2} r_{ijk}^2 + \frac{1}{2} r_{ik}^2$.

The generalization of approximation of the expected difference between *LL*(MC*l*) $(1 \le l \le w-1)$ and *LL*(MC0) in an arbitrary window of size *w* can be found in Appendix B.

Based on the approximation of expected values of numerators and denominators of different ϕ_i ($i = 1, 2, .., w-1$), we can interpret them in the following way. ϕ_1 measures the relative contribution from MC1 to the total disequilibrium (δ_{w-1}) and estimates the proportional effect from the adjacent pairs of markers to the total LD. ϕ_2 measures the residual contribution from MC2 after subtracting the relative contribution from MC1 and estimates the proportional effect from the adjacent three markers and non-adjacent two markers to the total LD. Finally, ϕ_{w-1} measures the residual contribution from MC*w*-1 after subtracting all the lower MC orders and estimates the proportional effect from the leftover which were not included in previous ϕ's. The sum of all ϕ_i ($i = 1, 2, .., w-1$) is 1.

3.4 Results

3.4.1 Data

We apply our method to the phased haplotype data of four populations in ten ENCODE regions from the HapMap project. The phased haplotype data in the ENCODE regions are downloaded from the HapMap project website (**http://www.hapmap.org/downloads/phasing/2005-03_phaseI/ENCODE/**). The data are generated using the PHASE v2.1 software (**http://www.stat.washington.edu/stephens/software.html**) and compiled from the

genotype data Phase I / rel#16a. Phase I data include a common SNP with Minor Allele Frequency $\geq$ 0.05 in every 5 kb across most of the genome in each population (The International HapMap Consortium, 2005). To produce more complete genotype data, the HapMap ENCODE project resequenced and genotyped a representative collection of ten regions, each 500 kb in length. Each 500 kb region was resequenced in 48 unrelated individuals (16 Yoruba, 8 Japanese, 8 Han Chinese, and 16 CEPH). All SNPs identified, along with SNPs in dbSNP, were genotyped in the 269 HapMap DNA samples: (1) 90 individuals (30 parent-offspring trios) from the Yoruba in Ibadan, Nigeria (abbreviation YRI); (2) 90 individuals (30 trios) from Utah, from the Centre d'Etude du Polymorphisme Humain collection (abbreviation CEU); (3) 45 Han Chinese in Beijing, China (abbreviation HCB); (4) 44 Japanese in Tokyo, Japan (abbreviation JPT). Across the ten ENCODE regions, the density of SNPs is one per 279bp on average, which is approximately tenfold higher than the Phase I genome-wide data. Table 3.1 summarizes the datasets of each population in the ten ENCODE regions. We focus on SNPs with MAF ≥ 0.05 in all populations.

Table 3.1 Summary of datasets in HapMap ENCODE regions.

Region	(Position)	YRI	CEU	HCB	JPT
ENr112	(2p16.3)	1157 (0.27, 0.43 ± 0.58)	875 (0.33, 0.57 ± 0.76)	872 (0.34, 0.57 ± 0.77)	861 (0.34, 0.58 ± 0.78)
ENr131	(2q37.1)	1070 (0.27, 0.47 ± 0.59)	984 (0.29, 0.51 ± 0.66)	938 (0.31, 0.53 ± 0.72)	909 (0.31, 0.55 ± 0.73)
ENr113	(4q26)	1218 (0.24, 0.41 ± 0.50)	1069 (0.25, 0.47 ± 0.63)	948 (0.28, 0.53 ± 0.72)	872 (0.30, 0.57 ± 0.77)
ENm010	(7p15.2)	433 (0.62, 1.15 ± 1.66)	471 (0.61, 1.06 ± 1.47)	322 (0.87, 1.52 ± 1.88)	316 (0.87, 1.55 ± 1.90)
ENm013	(7q21.13)	850 (0.36, 0.59 ± 0.73)	747 (0.44, 0.67 ± 0.78)	650 (0.48, 0.77 ± 0.87)	643 (0.49, 0.77 ± 0.88)
ENm014	(7q31.33)	997 (0.33, 0.5 ± 0.56)	908 (0.34, 0.55 ± 0.63)	674 (0.45, 0.74 ± 0.83)	582 (0.49,0.85 ± 1.22)
ENr321	(8q24.11)	851 (0.33, 0.59 ± 0.77)	537 (0.48,0.93 ± 1.24)	612 (0.45, 0.82 ± 1.11)	604 (0.45, 0.83 ± 1.12)
ENr232	(9q34.11)	594 (0.34, 0.84 ± 1.56)	463 (0.46, 1.08 ± 2.01)	498 (0.43, 1.0 ± 1.9)	476 (0.41, 1.05 ± 2.7)
ENr123	(12q12)	479 (0.65, 1.0 ± 1.1)	744 (0.4, 0.67 ± 0.84)	450 (0.59, 1.1 ± 1.3)	442 (0.67, 1.1 ± 1.2)
ENr213	(18q12.1)	740 (0.39, 0.67 ± 0.82)	608 (0.53, 0.82 ± 1.0)	466 (0.64, 1.1 ± 1.4)	511 (0.58, 0.98 ± 1.2)

The length of each ENCODE region is 500 kb. In each population of each region, # of SNPs with MAF>=0.05 and distance between adjacent SNPs with (median, mean ± sd) in kb are listed. The regions have comparable SNP densities among the four populations.

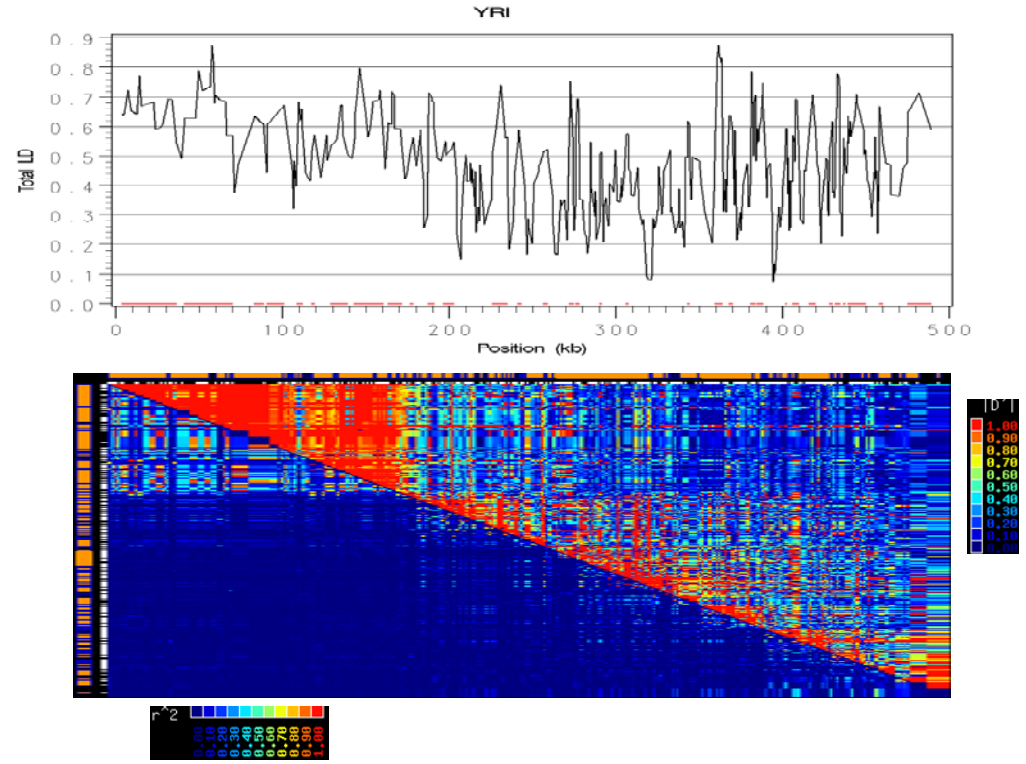
3.4.2 Profiles of LD pattern

Total multilocus LD values computed in each sliding window of size 5 (i.e. 5 consecutive SNPs) are used to create local LD profiles. Ad-hoc haplotype blocks are defined using a threshold of 0.5. The choice of these two control parameters is made on the recommendation of Nothnagel (2004, 2005). Nothnagel (2004) proposed the normalized entropy difference as a measure of multilocus linkage disequilibrium and defined haplotype blocks as an application. Nothnagel (2004) thoroughly investigated the influence of window sizes and various thresholds for the block definition using simulated datasets and real dataset and recommended medium window sizes such as 4-6 and medium thresholds between 0.4 and 0.6. For small window sizes 2-3, multilocus LD profiles vary widely. For large window sizes 7-9, the distinction between low and high LD regions is obscured a lot due to the smoothing effect. For low thresholds, blocks become unreasonably large and contain regions of low LD. For more stringent thresholds, chromosomal coverage decreases a lot and only a few regions of very high LD are defined as blocks.

We compare our multilocus LD profiles with the LD pattern from other pairwise LD measures. We use a software package called HaploBlockFinder (Zhang et al., 2003) version 0.7 to perform pairwise LD analysis. For each of four populations, Figure 1 compares the magnitudes of total LD (δ_{w-1}) for window size 5 along the marker sequence in the top plots with the pairwise LD measures of $|D'|$ and r^2 in the bottom plots. We define ad-hoc haplotype blocks as a union of windows over which the level of total LD is 0.5 or above. The blocks identified from our approach is depicted by red bold lines in the top plots and compared with those from pairwise LD measures (see Figure 3.1). In the

bottom plots, the blocks are defined as a consecutive set of markers in which minimal $|D'|$ is 0.95 or above and high LD regions appear as different sizes of red triangles. Using the threshold of 0.5, we identified many short blocks (35 blocks) in YRI, fewer but longer blocks in the other three populations - 21 blocks in CEU, 11 blocks in HCB, and 15 blocks in JPT. The average block lengths are 5.5 kb in YRI, 17.6 kb in CEU, 36.3 kb in HCB, and 26.6 kb in JPT. The size of each block varies a lot, from 0.01 to 32 kb in YRI, 0.1 to 168 kb in CEU, 0.5 to 94 kb in HCB, and 0.3 to 94 kb in JPT. The large scale LD patterns show that high LD regions in the top plots are similar to the ones specified in the bottom plots and the sites of recombination hotspots around high LD regions in the bottom plots are also similar to the regions where total LD drops down sharply in the top plots. Both top and bottom plots agree that the LD pattern in YRI population is very different from those observed in the other three populations in which LD extends to a similar and long extent. This fact is consistent with the results discovered by other research groups (Gabriel et al., 2002; Hinds et al., 2005).

Figure 3.1 Comparison of LD patterns obtained from our multilocus LD measure and pairwise LD measures using HaploBlockFinder version 0.7. In the top diagrams, multilocus total LD in ENm010 region is plotted against the physical location of the central marker in each window for four different populations (YRI, CEU, HCB, JPT). Window size 5 is used for this analysis. The locations of blocks using a threshold of 0.5 are depicted by red bold lines below the graph. In the bottom diagrams, the LD values are plotted using two pairwise LD measures, $|D'|$ (top right) and r^2 (bottom left) and the haplotype blocks are shown on the top side and the left side with orange color. Short little white lines on the top and left in the bottom diagram represent the position of SNPs.


YRI
Total LD
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
Position (kb)
D'
r^2

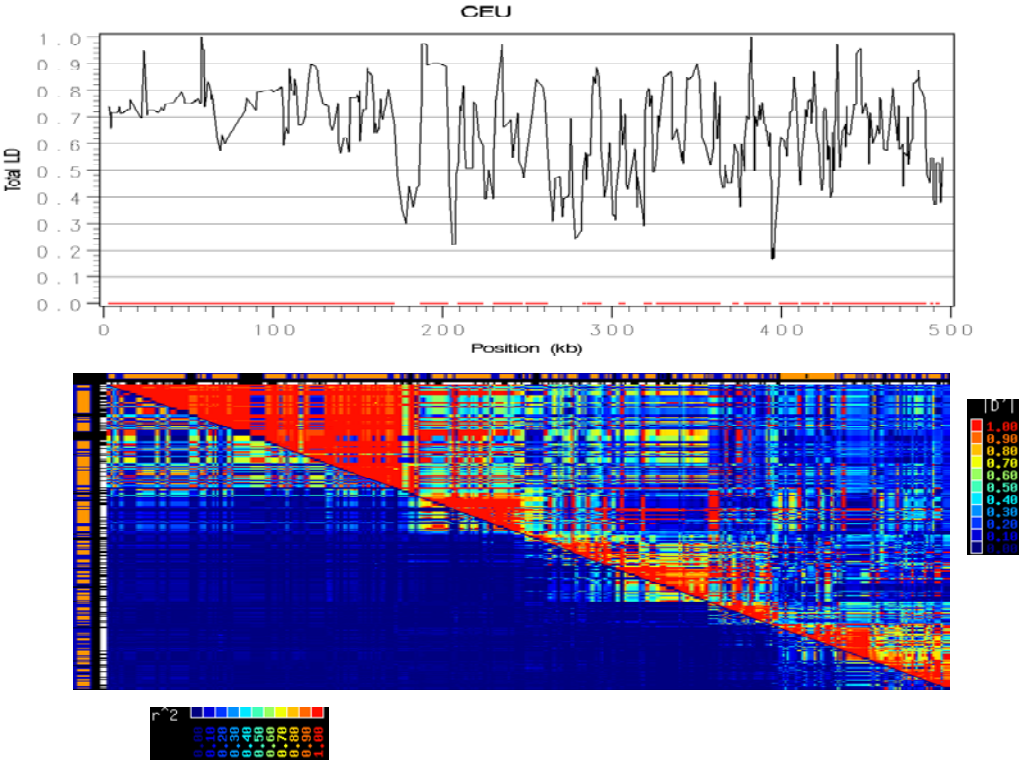
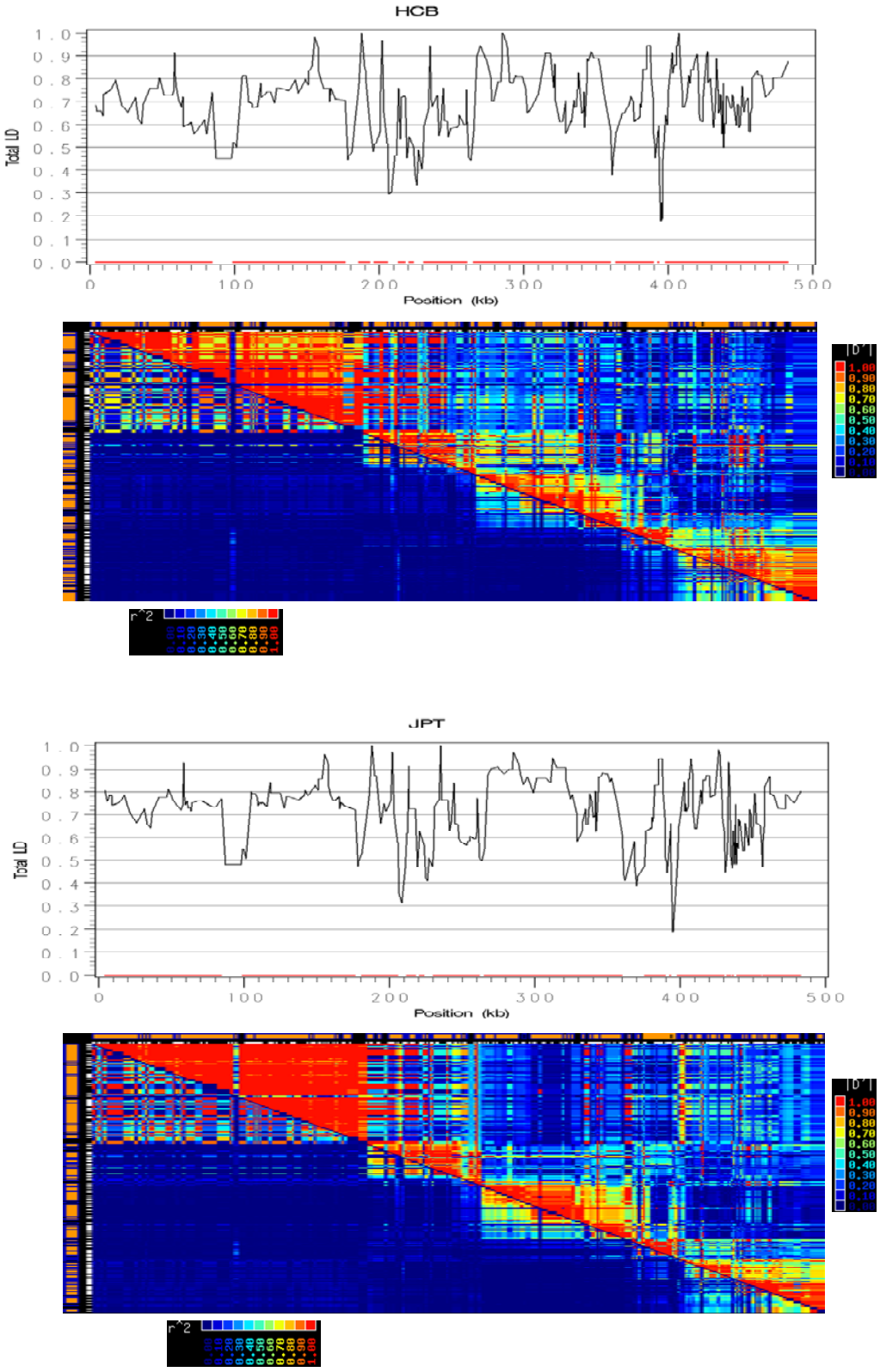

CEU
Total LD
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
Position (kb)
D'
r^2

Figure 3.1 (continued)



3.4.3 Partitioning of total LD

So far, we have demonstrated that the LD profile created by our total LD measure reasonably agree with that created by pairwise LD measures. Nothnagel (2002, 2004) claimed that the normalized entropy difference could be considered as a generalization of r^2 to haplotypes of more than two bi-allelic loci becasue the normalized entropy difference and r^2 share a number of similarities. However, the reason why there is a good agreement between pairwise and multilocus approaches is not discussed in those papers. By decomposing the total LD into various lower orders of disequilibria, we could understand the reason better. Using the log-likelihoods from different MC orders for a fixed window size *w*, we partition the overall departure from LE (δ_{w-1}) into effects of lower order disequilibria such as two-, three- , up to *w* - locus disequilibrium. Figure 3.2 illustrates an example of partitioning when a window size 5 is applied to the JPT population in ENm010 region. The average percentage contribution of ϕ_1, ϕ_2, ϕ_3, ϕ_4 across the region is 67%, 23%, 8%, and 2%, respectively. Table 3.2 shows a comprehensive analysis of average percentage contribution of various orders of LD in the ten ENCODE regions and four populations. Regardless of genome regions and populations, it is obvious from the analysis that the great bulk of the total disequilibrium is from the contribution of MC1 (ϕ_1) which estimates the association between adjacent two markers. Compared to the other three populations, YRI population consistently shows smaller effect of pairwise association in all regions, i.e. having proportionally more high order LD.

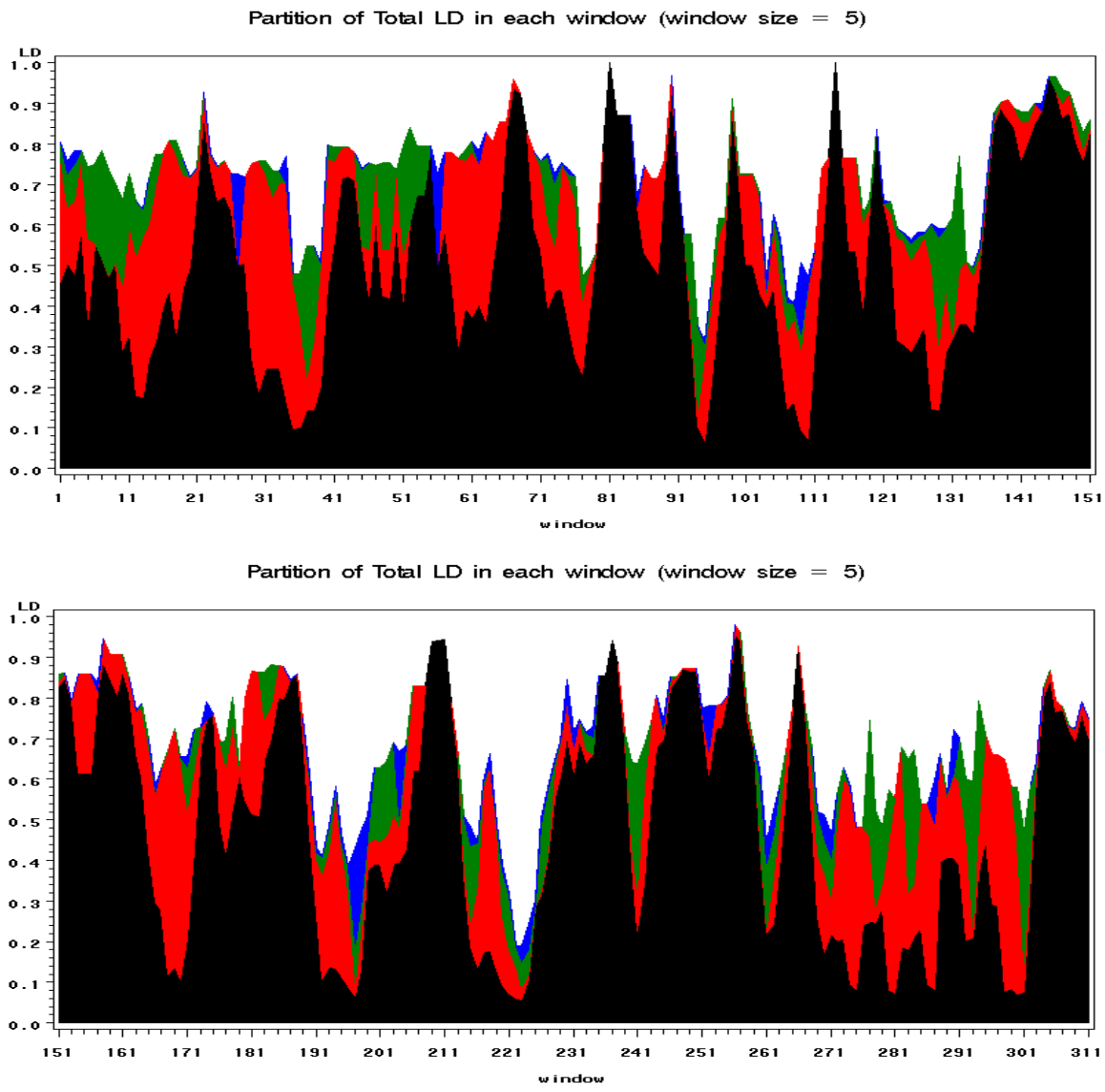


Figure 3.2 Partitioning of the total LD by the contribution from ϕ_1, ϕ_2, ϕ_3, ϕ_4 using a window size 5 in ENm010 region and JPT population.

Black represents the contribution from ϕ_1 (effect from adjacent pairs of markers); red represents the contribution from ϕ_2 (effect from adjacent three markers and non-adjacent two markers); green is from ϕ_3 (effect from the adjacent four markers, non-adjacent three markers, and non-adjacent two markers); blue is from the contribution from ϕ_4 (the residual effect which were not included in previous ϕ's).

Table 3.2 Estimation of proportional effects of lower order disequilibrium in the ten ENCODE regions of HapMap data.

Region	Relative percentage contribution from each lower order disequilibrium to the total LD															
	YRI				CEU				HCB				JPT			
	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4
ENr112	53	29	13	6	71	21	7	2	67	24	7	2	66	24	7	2
	(25)	(22)	(16)	(10)	(24)	(21)	(12)	(5)	(24)	(21)	(12)	(6)	(24)	(21)	(12)	(7)
ENr131	56	26	13	5	65	23	9	3	62	24	9	4	63	25	9	3
	(26)	(22)	(17)	(11)	(24)	(22)	(13)	(7)	(26)	(22)	(14)	(9)	(25)	(22)	(13)	(9)
ENr113	61	25	10	4	68	21	8	3	71	21	6	2	78	17	4	1
	(26)	(22)	(14)	(9)	(25)	(20)	(14)	(8)	(25)	(21)	(12)	(6)	(24)	(20)	(9)	(6)
ENm010	57	26	13	4	68	21	9	2	65	26	7	1	67	23	8	2
	(25)	(21)	(15)	(9)	(23)	(20)	(13)	(6)	(27)	(24)	(13)	(5)	(27)	(23)	(13)	(6)
ENm013	51	29	15	5	57	27	12	4	63	24	9	3	63	25	9	3
	(26)	(24)	(18)	(11)	(26)	(24)	(15)	(9)	(24)	(21)	(14)	(8)	(24)	(22)	(14)	(8)
ENm014	55	26	14	5	61	25	11	3	72	19	7	2	74	19	6	2
	(27)	(23)	(17)	(11)	(26)	(22)	(16)	(9)	(26)	(21)	(14)	(7)	(23)	(20)	(10)	(6)
ENr321	58	26	11	4	76	17	5	1	71	20	7	2	72	19	6	2
	(26)	(22)	(15)	(9)	(22)	(20)	(11)	(5)	(25)	(21)	(14)	(6)	(24)	(21)	(13)	(7)
ENr232	55	26	13	5	80	15	4	1	67	21	9	3	66	22	6	3
	(26)	(22)	(17)	(10)	(21)	(16)	(10)	(5)	(23)	(19)	(13)	(7)	(23)	(20)	(13)	(7)
ENr123	55	28	12	5	59	26	10	4	60	25	11	4	57	27	11	5
	(27)	(23)	(15)	(11)	(26)	(22)	(14)	(10)	(27)	(21)	(16)	(11)	(27)	(23)	(17)	(12)
ENr213	55	27	12	5	63	24	9	3	73	19	6	1	68	23	8	1
	(27)	(25)	(17)	(11)	(24)	(22)	(14)	(8)	(23)	(20)	(11)	(5)	(24)	(22)	(12)	(5)

Table 3.2 (continued)

The average proportional effects from adjacent pairs of markers (ϕ_1), from adjacent three and non-adjacent two markers (ϕ_2), from adjacent four, non-adjacent three, and non-adjacent two markers (ϕ_3), from adjacent five, non-adjacent four, non-adjacent three, and non-adjacent two markers (ϕ_4) to the total LD are estimated across each ENCODE region in each population. The numbers in parentheses are standard deviation.

3.4.4 Multilocus total LD and haplotype diversity

Within high LD regions or haplotype blocks, the diversity of haplotypes is very limited and only a few haplotypes are observed (Daly et al., 2002). Thus, it is expected that there should be a strong negative correlation between magnitude of total LD (δ_{w-1}) and the number of haplotypes. To find out the relationship between the multilocus total LD and haplotype diversity, we count the total number of distinct haplotypes in each window for a fixed window of size *w*. For moderate window sizes 4-7, mean of total LD is computed for different numbers of haplotypes. We observe that mean of total LD decreases linearly as the number of haplotypes increases. Therefore, there is a clear inverse relationship between magnitude of total LD and number of haplotypes. Figure 3.3 illustrates this relationship after all ENCODE regions are combined in each population. Window size 5 is used for this analysis. We get very similar results for other window sizes (data not shown).

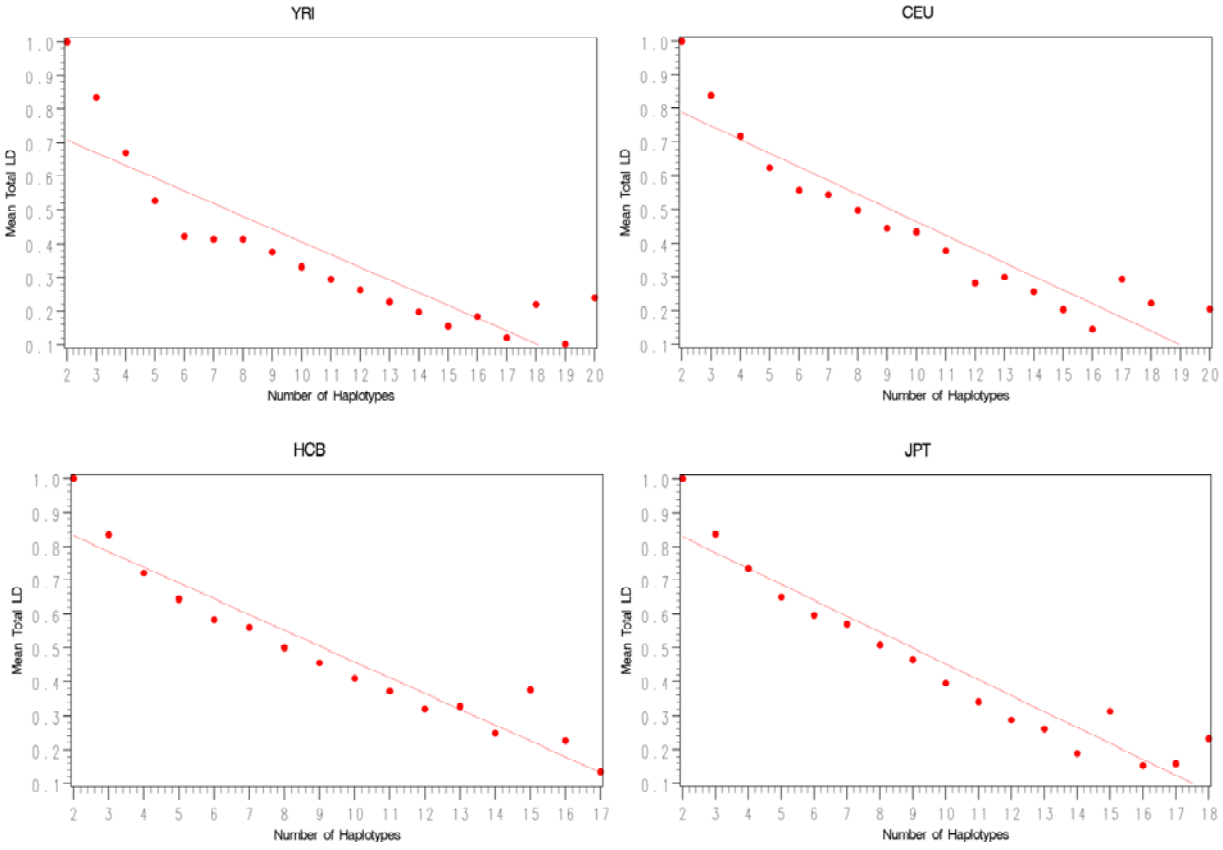


Figure 3.3 Scatterplots of mean total LD vs. number of haplotypes.
For each population, all ENCODE regions are combined and mean total LD is computed for different numbers of haplotypes. Window size 5 is used for this analysis. The correlation between two variables is -0.88 ($p < 0.0001$) in YRI, -0.93 ($p < 0.0001$) in CEU, -0.95 ($p<0.0001$) in HCB, -0.95 ($p < 0.0001$) in JPT. For other window sizes, very strong negative relationship between mean total LD and number of observed haplotypes is also observed (data not shown).

3.4.5 High level of LD vs. high order of LD

One misunderstanding about LD measures is that high level of LD implies high order of LD which describes the disequilibrium among alleles at more than two loci. We were interested in determining whether high total LD regions necessarily have high order of LD and found that this is not always the case. Methods based on the pairwise LD measures can identify regions with high level of LD, but not high order of LD because

LD is measured only between pairs of loci. From Figure 3.2, it was noticed that the contribution of MC1(ϕ_1) tends to be larger in high total LD regions as opposed to the contribution of higher MC orders (ϕ_2,..,ϕ_{w-1}) in those regions. To see if we can detect this pattern in other regions and populations, we divide the whole region into three groups (low, medium, and high haplotype diversity) depending on the number of distinct haplotypes. For a window of size 5, if the number of distinct haplotypes is 2-4, the window is categorized as a low haplotype diversity or high LD region; if the number of distinct haplotypes is 5-7, as a medium haplotype diversity or medium LD region; if the number of distinct haplotypes is ≥ 8, as a high haplotype diversity or low LD region. Table 3.3 shows that in all categories of haplotype diversity, the effect of ϕ_1 or adjacent pairwise marker association accounts for the most variation in the total LD followed by ϕ_2, ϕ_3, ϕ_4 in this order. As haplotype diversity increases, however, the relative contribution of adjacent pairwise association reduces significantly while the relative contribution of higher MC orders increases. The explanation is that detection of higher MC orders requires many distinct haplotypes. A large number of distinct haplotypes observed in low LD regions can increase the likelihoods of higher MC orders. Similarly, detection of high order of LD requires many different haplotypes and large sample size.

Table 3.3 Estimation of proportional effects of lower order disequilibrium grouped by different haplotype diversity in the ten ENCODE regions of HapMap data.

Region	YRI												CEU											
	Low diversity				Medium diversity				High diversity				Low diversity				Medium diversity				High diversity			
	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4
ENr112	71	23	4	2	52	29	14	6	48	32	13	7	81	14	4	1	63	26	9	2	58	27	11	4
ENr131	69	22	7	2	52	27	14	6	54	26	15	5	72	20	6	1	59	26	11	4	64	23	10	4
ENr113	79	16	4	1	54	28	12	5	54	27	14	5	77	17	5	1	58	26	12	4	54	30	10	6
ENm010	66	23	7	3	58	25	13	4	52	28	15	6	77	16	6	1	64	22	11	3	50	32	14	4
ENm013	65	22	10	2	49	30	16	6	46	32	17	6	63	25	10	2	52	30	13	5	65	18	12	5
ENm014	72	20	6	2	51	28	16	5	54	24	15	7	67	23	8	2	55	26	15	5	50	32	13	5
ENr321	69	22	7	2	56	28	12	4	49	28	16	7	82	14	3	1	71	20	8	1	67	22	8	3
ENr232	73	19	7	1	53	28	14	5	55	26	14	6	87	10	2	1	74	19	5	2	67	21	8	4
ENr123	74	21	4	1	53	28	13	6	46	33	14	7	74	20	3	2	53	29	13	6	45	33	17	5
ENr213	72	21	6	1	53	28	13	6	48	31	15	6	68	21	9	2	57	29	10	4	61	21	12	6

Table 3.3 (continued)

	HCB												JPT											
	Low diversity				Medium diversity				High diversity				Low diversity				Medium diversity				High diversity			
	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4
ENr112	72	22	5	1	60	28	9	4	59	21	16	4	71	22	5	2	62	26	9	3	56	24	14	6
ENr131	71	21	5	2	55	27	13	5	54	29	12	5	71	21	6	2	55	29	12	5	57	25	14	5
ENr113	75	20	5	1	60	26	10	3	64	19	14	3	81	15	3	1	66	23	7	3	62	25	8	6
ENm010	73	21	5	1	64	26	8	2	54	32	11	3	73	20	6	1	68	23	8	1	53	30	14	4
ENm013	69	21	7	3	57	27	12	3	57	25	11	6	69	23	6	2	55	29	12	4	55	27	16	3
ENm014	79	15	5	1	61	26	10	3	65	20	7	7	80	15	4	1	65	24	8	3	69	23	7	1
ENr321	79	14	5	1	60	26	10	3	64	26	6	3	79	14	5	1	63	26	8	3	64	25	6	5
ENr232	74	18	7	2	64	22	10	3	61	23	12	4	72	20	6	1	63	22	11	3	57	26	13	4
ENr123	83	14	2	1	60	26	10	4	44	31	17	8	81	14	3	2	61	28	7	3	45	29	18	8
ENr213	78	18	3	1	67	21	10	2	68	18	11	3	71	22	6	1	63	24	11	3	61	25	11	3

The numbers in each cell represents an average percentage contribution from each $\phi_i (i = 1, 2, 3, 4)$ across different haplotype diversity groups. $\phi_i (i = 1, 2, 3, 4)$ have the same meanings as Table 1. For each window of size 5, if the number of distinct haplotypes is 2-4, the window is categorized as a low haplotype diversity or high LD region; if the number of distinct haplotypes is 5-7, as a medium haplotype diversity or medium LD region; if the number of distinct haplotypes is ≥ 8, as a high haplotype diversity or low LD region.

3.5 Discussion

Linkage disequilibrium (LD) between loci is defined as the deviation of the haplotype frequencies from their expectation under independence. Our approach characterizes multilocus LD using a multiple order Markov Chain model in a sliding window. Different orders of Markov Chain model in a window produce different log-likelihoods, with the smallest log-likelihood from the independent (MC0) model to the largest log-likelihood from the full haplotype (MC*w*-1) model. Our multilocus LD measure (δ_{w-1}) is based on the log-likelihood difference between the full haplotype model (MC*w*-1) and the independent model (MC0) for a certain window size *w*. If the log-likelihood of the full model is more deviated from the independent model in one region than other regions, it indicates that there is higher LD in the region. That is, δ_{w-1} can measure the extent of the total local LD for multiple markers.

In addition to measuring the total LD in a local region, our method further partitions the total disequilibrium (δ_{w-1}) to lower order disequilibria of two-, three-, …, and all- marker association in each window. From the approximation of $\frac{1}{N}E[LL(MCw-1)-LL(MC0)]$, we can see that the total multilocus disequilibrium can be decomposed into all combinations of lower order disequilibria. However, the relative contribution from each of lower order disequilibria is quite different depending on the extent of total multilocus LD. We observe that in high total LD regions, a great proportion of the total multilocus disequilibrium is accounted for by adjacent two marker associations. Contrarily, in low total LD regions, the contribution due to adjacent two marker association is reduced a lot, and high orders of LD such as three or more marker

associations become more noticeable. This fact implies that the extent of total multilocus disequilibrium does not necessarily indicate the orders of LD. Paradoxically, high level of LD in a region does not mean a high proportion of high order LD. The reason is that high level LD is usually associated with low number of haplotypes and that detection of high order LD requires a large sample size and many diverse haplotypes. Thus, in small samples, the estimates of high order LD are not quite stable. We should use caution when interpreting the estimates of high order LD, particularly in small samples.

The log-likelihood difference between MC0 model and MC *w*-1 model is partitioned into components for the log-likelihood difference between two consecutive MC order models. We estimated the proportional effect of each component, but did not explicitly test the significance of it. The results of statistical test for each component will be published elsewhere. When the log-likelihood difference between MC0 model and MC w-1 model is completely partitioned, there are $2^w - (w+1)$ LD coefficients with varying orders. We can test a hypothesis whether *w* markers are in linkage equilibrium using a chi-square distribution with $2^w - (w+1)$ degree of freedom *if all hyplotypes are observed*. If the hypothesis is rejected, our next step is to find which LD coefficients are significant and which are not. However, due to finite sample size, many hyplotypes for multiple markers are usually unobserved. This imposes significant problems for the test of high order LD coefficients. In general the chi-square approximation performs reasonably well for the statistical test of some coefficients, such as pairwise LD, in a sample with reasonable size, such as HapMap samples. But, our preliminary result indicates that it is not appropriate to use chi-square approximation for the test of high

order LD coefficients unless sample size is very large. There is a need for an alternative approach for the statistical test.

Our idea of partitioning the total LD into various components of lower order disequilibria is related to Smouse's method (1974) but overcomes a critical computational problem associated with his method. In his log-linear approach, a series of multiplicative models with different numbers of disequilibrium terms are first constructed. Then, the difference in deviances for two models that differ only by whether a particular disequilibrium term is included provides a chi-square test statistic for that term. However, a critical computational problem arises when one or more hyplotypes are unobserved (with zero observed frequency for the haplotypes). In this case, one is faced with the problem of $\log(0) = -\infty$ and the subsequent tests can not be performed. In contrast, our approach does not need to observe all the possible multilocus haplotypes to measure the multilocus total LD and estimate the effects from lower order disequilibria because the likelihoods of different orders of MC model are computed using the observed haplotypes. Needless to say, this property is very useful in small samples.

Large scale LD profiles show reasonable agreement between our multilocus method and traditional pairwise LD measures. However, we expect that fine scale LD patterns can be somewhat different between the two methods. For example, the number of blocks and block boundaries can be different from method to method. Ke et al. (2004) compared different methods of block definition with respect to number of blocks, average block length, and proportion of sequence contained within blocks. They found that there does not appear to be a strong convergence of block-detection methods. Since the true underlying block structure is unknown, it is difficult to compare which method is more

appropriate than the other. Blocks are heavily dependent on the factors related to the sample - SNP density, sample size, and marker selection (Ke et al. 2004, Sun et al. 2004, Nothnagel et al. 2005).

A large number of markers can be used in a sliding window theoretically, but the number of markers in a window should be limited in practice. As the number of markers in a window increases, the number of possible haplotypes can increase dramatically, but we cannot observe many of them in small samples. Rare long haplotypes which are present in the population but are unobserved in small samples can cause the estimates of LD to be inflated since low numbers of haplotypes falsely indicate high LD. This problem becomes more severe for large windows.

Like most sliding window approaches, finding an optimal window size to capture the LD pattern accurately over a large region is quite challenging. Window size is negatively correlated with the variability in LD trend. If too small window size is used throughout the whole region, the values of multilocus LD measures are so fluctuated that it is hard to separate high LD regions from low LD regions. Haplotype blocks are also fragmented with only a few SNPs in each block. Applying the definition of haplotype-tagging SNPs within these blocks may be useless. On the other hand, too large window size can introduce other problems such as excessive noise, computer memory problem, and smoothing effect. Nothnagel (2004) recommended medium window sizes such as 4-6 after extensive investigation on this matter.

3.6 Acknowledgements

This work was partially supported by NIH GM45344, UL1 RR024128, and by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number 2005-00754. YK is supported by a university genomics science fellowship and SF is supported by NIH grant UL1 RR024128.

Chapter 4

Tests for two- and three-locus disequilibria and construction of 3D LD plot

4.1 Introduction

With the availability of very large numbers of dense SNP data, there has been of great interest utilizing multiple closely linked markers simultaneously to find the SNP(s) that are close to the causal variants responsible for complex human diseases in genetic association studies. Analyzing SNPs individually can lose information in their joint distribution unless SNPs are widely spaced so that there is little LD between them or all SNPs are typed so that any causal variant is likely to be typed in the study. In reality, SNP densities in most studies are between the two extremes (Balding 2006). Under such circumstances, multi-SNP analyses may have potential advantages over single-SNP analysis although the power comparisons between two approaches have revealed somewhat mixed results (Akey et al 200; Long et al., 1999; Kaplan et al., 2001; Fallin et al., 2001).

In single- SNP or multi-SNP association analyses, LD between markers and a putative causal variant cannot be directly estimated because only phenotypes, not genotypes, are observed. Nonetheless, detailed understanding of the structure and pattern

of LD, facilitated by testing for various orders of disequilibria, among SNPs in a chromosome region of interest can be useful for developing more efficient statistical methods for disease gene mapping. Moreover, such knowledge may provide valuable information on inferring human evolutionary history.

Statistical tests for two-locus disequilibrium have been well-studied and widely used (Weir 1979; Weir 1996). However, methods to test for the significance of three-locus disequilibrium or higher-order disequilibria are far less well studied. A three locus disequilibrium, D_{ABC}, defined by Bennett (1954) has been commonly used. It describes the quantity that arises naturally in disease association test statistics (Nielsen et al., 2004). Various properties of this measure have been examined (Hill 1976; Thompson et al., 1984) and several methods have been suggested for testing $H_0 : D_{ABC} = 0$ (Hill 1976; Weir 1996). However, specifying the null hypothesis as $H_0 : D_{ABC} = 0$ has problems in some situations. Haplotype frequencies under the null can have negative values and the range of possible values for D_{ABC} may not include zero for some haplotype distributions of three diallelic markers.

Another definition of no three-locus association initially proposed by Bartlett (1935) has been considered. No three-locus association occurs when the following condition is satisfied:

$$(P_{ABC}P_{Abc}P_{aBc}P_{abC}) = (P_{ABc}P_{AbC}P_{aBC}P_{abc})$$

This condition means the association between, say, A and B should be the same in the chromosome having C as c. This condition is symmetric among the loci as following:

$$\frac{P_{ABC}P_{Abc}}{P_{ABc}P_{AbC}} = \frac{P_{aBC}P_{abc}}{P_{aBc}P_{abC}} \Leftrightarrow \frac{P_{ABC}P_{aBc}}{P_{ABc}P_{aBC}} = \frac{P_{AbC}P_{abc}}{P_{Abc}P_{abC}} \Leftrightarrow \frac{P_{ABC}P_{abC}}{P_{AbC}P_{aBC}} = \frac{P_{ABc}P_{abc}}{P_{Abc}P_{aBc}}$$

When some of possible haplotypes for three diallelic markers are not observed, the first definition has problems as discussed above but the second definition is still suitable for testing for the three-locus association. Based on the second definition, several methods to test for the three-locus association have been developed (Hill 1975, 1976; Smouse 1974; Brown 1975). The methods in Hill (1975, 1976) were based on the hierarchical models of dependence of frequencies used for tests of association in multi-dimensional contingency tables (Goodman 1969; Fienberg 1970). For the test for three-locus disequilibrium, two models with the null model including all pairwise associations and the alternative model including all pairwise and three-locus associations were compared using a likelihood ratio test (LRT) statistic which asymptotically follows the chi-square distribution with 1 degree of freedom.

A big obstacle facing the hypothesis tests for various orders of LD is that many haplotypes for multiple markers are usually unobserved due to finite sample size. This imposes significant problems, especially for the test of high order LD coefficients. Brown (1975) argued that relatively large samples are required to detect high order LD, unless allele frequencies are intermediate and disequilibrium is relatively intense. With large samples, the distribution of test statistics under the null hypothesis of no three-locus disequilibrium can be reasonably well approximated by χ_1^2. If many haplotypes are unobserved, however, it is questionable to test the hypothesis using the chi-square approximation because there might be no degree of freedom left for the test. Long et al. (1995) suggested an alternative approach in which empirical distributions of statistics by

resampling the observed data are used to test hypotheses about different orders of disequilibrium. Such empirical distributions avoid the large sample assumption at the expense of more computing time. This resampling method may provide more reliable p-values for the test of high order LD even when there are many missing haplotypes.

In this study, we are interested in understanding the detailed multilocus LD structure among SNPs in the human genome. Specifically, we focus on testing for two- and three-locus disequilibria among SNPs in an ENCODE region of the HapMap data and constructing 3 dimensional LD plots to identify regions where different orders of LD stand out. For our purposes, we borrow the ideas from Hill (1975, 1976) and Long et al. (1995). The significance of three-locus disequilibrium is evaluated via χ_1^2 and a resampling method. To adjust for multiple testing problem, we apply a FDR procedure (Storey 2002).

4.2 Method

4.2.1 Models & Hypothesis Testing

A series of hierarchical models are considered. The constraints for each model are specified using the following four terms:

$$D_{AB} = P_{AB} - P_A P_B$$
$$D_{BC} = P_{BC} - P_B P_C$$
$$D_{AC|B} = D_{AC} - \frac{D_{AB} D_{BC}}{P_B(1 - P_B)}$$
$$W = (P_{ABC} P_{Abc} P_{aBc} P_{abC}) - (P_{ABc} P_{AbC} P_{aBC} P_{abc})$$

D_{AB} denotes the common measure of two-locus marginal disequilibrium between the first marker (M_1) and the second marker (M_2). Similarly, D_{BC} denotes the measure of two-

locus marginal disequilibrium between the second marker (M_2) and the third marker (M_3). $D_{AC|B}$ denotes the measure of two-locus partial disequilibrium between the first marker (M_1) and the third marker (M_3). The last term W denotes three-locus disequilibrium measure.

- Model 0: The alleles at three markers are completely independent.

  $(D_{AB} = 0,\ D_{BC} = 0,\ D_{AC|B} = 0,\ W = 0)$

- Model 1: Marginal association between alleles at M_1 and M_2.

  $(D_{AB} \neq 0,\ D_{BC} = 0,\ D_{AC|B} = 0,\ W = 0)$

- Model 2: Independence of M_1 and M_3 conditional on M_2.

  $(D_{AB} \neq 0,\ D_{BC} \neq 0,\ D_{AC|B} = 0,\ W = 0)$

- Model 3: All pairwise association but no 3-locus association

  $(D_{AB} \neq 0,\ D_{BC} \neq 0,\ D_{AC|B} \neq 0,\ W = 0)$

- Model 4: All pairwise and 3-locus association

  $(D_{AB} \neq 0,\ D_{BC} \neq 0,\ D_{AC|B} \neq 0,\ W \neq 0)$

The maximized log-likelihoods for each model are computed in the following way:

$$LL(Model\ 0) = \sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C} n_{ijk} \times \log(\hat{P}_{i..} \times \hat{P}_{.j.} \times \hat{P}_{..k})$$

$$LL(Model\ 1) = \sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C} n_{ijk} \times \log(\hat{P}_{ij.} \times \hat{P}_{..k})$$

$$LL(Model\ 2) = \sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C} n_{ijk} \times \log(\hat{P}_{i..} \times \frac{\hat{P}_{ij.}}{\hat{P}_{i..}} \times \frac{\hat{P}_{.jk}}{\hat{P}_{.j.}})$$

$$LL(Model\ 3) = \sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C} n_{ijk} \times \log(\frac{\hat{e}_{ijk}}{N})$$

$$LL(Model\ 4) = \sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C} n_{ijk} \times \log(\frac{n_{ijk}}{N}) \quad ,$$

where n_{ijk} denotes the number of observed haplotypes; $\hat{e}_{ijk}$ denotes the converged number of haplotypes using an iterative procedure (see below).

It is interesting to note that the hypothesis test of global equilibrium ($H_0 : P_{ABC} = P_A P_B P_C$) can be decomposed into four mutually exclusive terms: two terms for testing two-locus marginal equilibrium, one term for the conditional independence between M_1 and M_3 , and one term for additional three-locus equilibrium. That is,

$$2N\sum_{i,j,k} P_{ijk} \ln\left(\frac{P_{ijk}}{p_i p_j p_k}\right)$$

$$= 2N\left(\sum_{i,j} P_{ij} \ln\left(\frac{P_{ij}}{p_i p_j}\right) + \sum_{j,k} P_{jk} \ln\left(\frac{P_{jk}}{p_j p_k}\right) + \sum_{i,j,k} P_{ijk} \ln\left(\frac{P^*_{ijk}}{\frac{P_{ij}P_{jk}}{p_j}}\right) + \sum_{i,j,k} P_{ijk} \ln\left(\frac{P_{ijk}}{P^*_{ijk}}\right)\right),$$

where P^*_{ijk} denotes the haplotype frequency distribution under Model 3. This partition is different from Lancaster (1951) in which the test of global equilibrium was partitioned into three terms for testing two-locus marginal equilibrium and one term for additional three-locus equilibrium as following:

$$\chi^2_L = N\left(\frac{(\hat{P}_{ABC} - \hat{p}_A\hat{p}_B\hat{p}_C)^2}{\hat{p}_A\hat{p}_B\hat{p}_C} + ... + \frac{(\hat{P}_{abc} - \hat{p}_a\hat{p}_b\hat{p}_c)^2}{\hat{p}_a\hat{p}_b\hat{p}_c}\right) = N(r^2_{AB} + r^2_{AC} + r^2_{BC} + r^2_{ABC}),$$

where

$$r_{AB}^2 = \frac{\hat{D}_{AB}^2}{\hat{p}_A \hat{p}_a \hat{p}_B \hat{p}_b}$$

$$r_{AC}^2 = \frac{\hat{D}_{AC}^2}{\hat{p}_A \hat{p}_a \hat{p}_C \hat{p}_c}$$

$$r_{BC}^2 = \frac{\hat{D}_{BC}^2}{\hat{p}_B \hat{p}_b \hat{p}_C \hat{p}_c}$$

$$r_{ABC}^2 = \frac{\hat{D}_{ABC}^2}{\hat{p}_A \hat{p}_a \hat{p}_B \hat{p}_b \hat{p}_C \hat{p}_c}$$

and $\hat{D}_{ABC} = \hat{P}_{ABC} - \hat{p}_A \hat{D}_{BC} - \hat{p}_B \hat{D}_{AC} - \hat{p}_C \hat{D}_{AB} - \hat{p}_A \hat{p}_B \hat{p}_C$. It is easy to show that the test of global equilibrium is not exactly but approximately partitioned into the above four terms suggested by Lancaster (1951).

4.2.2 Iterative Proportional Fitting

The haplotype frequencies under Model 3 cannot be estimated directly because there is no closed-form of solution to estimate them. The following iterative proportioning fitting procedure provides a way of computing nonnegative three-locus haplotype frequencies (Goodman 1969; Fienberg 1970; Gange 1995).

For a 3 marker system (M_i, M_j, M_k) with *i*=a or A; *j*=b or B; *k*=c or C, we obtain the observed numbers of two-locus haplotypes with alleles *i*, *j*, and *k* (i.e. $n_{ij.}$, $n_{i.k}$, $n_{.jk}$). For the initial step, set $e_{ijk}^{(0)}$ =1, where $e_{ijk}^{(0)}$ denotes the initial value for each three-locus haplotype with alleles *i*, *j*, and *k*. For the first cycle,

$$e_{ijk}^{(1)} = e_{ijk}^{(0)} \times \left(\frac{n_{ij.}}{e_{ij.}^{(0)}} \right) \quad \text{for all } i, j, \text{ and } k$$

$$e_{ijk}^{(2)} = e_{ijk}^{(1)} \times \left(\frac{n_{i.k}}{e_{i.k}^{(1)}} \right) \quad \text{for all } i, j, \text{ and } k$$

$$e_{ijk}^{(3)} = e_{ijk}^{(2)} \times \left(\frac{n_{.jk}}{e_{.jk}^{(2)}} \right) \quad \text{for all } i, j, \text{ and } k$$

For the second and subsequent cycles, the numbers of three-locus haplotype are adjusted so that they add to the numbers of two-locus haplotypes. At the end of each cycle, the amount of change introduced by the last cycle can be assessed. If the change is sufficiently small, the procedure is terminated. Otherwise, another cycle is performed. Among several possible criteria which can be used to terminate the iterative procedure, the criterion we used is to terminate when $| e_{ijk}^{(3u)} - e_{ijk}^{(3u-3)} | \leq 0.01,$ where $u = 1, 2, \ldots$.

It's interesting to note that the iterative procedure can be used to estimate the haplotype frequencies for all models we discussed in the previous section. For Model 3, the iterative procedure needs many cycles of iteration. For the other models (Model 0, Model 1, and Model 2) which we can compute the haplotype frequencies directly, only one cycle of iteration is needed to converge.

4.2.3 Problems associated with missing haplotypes

Due to the common problem of small sample size, it is quite often that we cannot observe all eight possible haplotypes for three SNPs. If an observed number of two-locus haplotype (i.e. $n_{ij.}$, $n_{i.k}$, or $n_{.jk}$) is zero, it is reasonable to speculate that the estimated numbers of constituent three-locus haplotypes are zero. Otherwise, if one of the constituent haplotypes is estimated to be a positive number, the other constituent haplotypes will be a negative number since the observed number of two-locus haplotype is zero. To compromise this problem, the iterative procedure converges to 0 for both constituent three-locus haplotypes. For example, if $n_{AB+} = 0$, both $\hat{e}_{ABC}$ and

$\hat{e}_{ABc}$ converge to zero. This can be done with a minor change in the iterative procedure as suggested by Fienberg (1970). The only change required in the procedure is to define zero divided by zero as zero. Thus, if an observed number of two-locus haplotype is zero, three-locus haplotypes constituting the two-locus haplotype will remain zero during the iterative procedure.

4.2.4 Construction of an empirical distribution of LRT statistic

In this section, we describe a resampling method which we use to evaluate the significances of likelihood ratio test statistics. We consider a 3 marker system (M_1, M_2, M_3) in this order with $M_1 = a$ or A, $M_2 = b$ or B, $M_3 = c$ or C. Suppose that haplotype frequencies of the 3 markers are estimated from N phased haplotypes assuming multinomial distribution. From the observed haplotype frequency distribution, we can obtain a haplotype frequency distribution under a particular null hypothesis. Depending on the null hypothesis, however, we may not have a closed form of solution for haplotype frequencies. In that case, we need an iterative procedure in order to produce the haplotype frequency distribution.

Given the haplotype probability distribution specified by the null hypothesis, we draw N haplotypes using a standard algorithm generating multinomial random variates and repeat the procedure a large number of times (say, 1000 times). In the end, R resampled samples can be generated. After fitting null and alternative models to the rth resampled sample, we compute $G[r]$ and evaluate the empirical distribution of G.

$$G[r] = 2 \times \{ (LL[r] \mid H_a) - (LL[r] \mid H_0) \}, \text{ where } r = 1,2,...,R$$

P-value is computed as the number of $G[r]$ that is more than or as extreme as the observed LRT statistic divided by the total number of resampling.

4.3 Results

4.3.1 Comparison of χ^2 distribution with empirical distributions of LRT statistics

To evaluate the significances of LRT statistic testing for various LD orders, it may be more reasonable to use empirical distributions of LRT statistic by a resampling method than to rely on chi-square distribution since the latter requires large sample assumption. To determine under what conditions the two methods agree or disagree each other, we compare chi-square distribution with empirical distributions of LRT statistic for different hypothesis tests and different MAF thresholds. For this purpose, we create two datasets from the phased haplotype data of the YRI population in ENm010 region from the HapMap project. One dataset consists of 433 SNPs with MAF>= 0.05 and the other consists of 140 SNPs with MAF>=0.3. Then, we randomly select 15 SNPs from each dataset and consider all possible three-SNP combinations (i.e. 455 triples).

Figure 4.1 compares the critical values of chi-square distribution with the averaged critical values of empirical distributions of LRT statistic at different tail probabilities across 455 triples. In the likelihood ratio tests for two-locus associations (i.e. M0 vs. M1, M1 vs. M2, and M2 vs. M3), the critical values of χ_1^2 distribution and those from the resampled distributions agree very well, regardless of the MAF threshold. On the other hand, in the likelihood ratio test for three-locus association (M3 vs. M4), they reveal some discrepancies. The critical values of χ_1^2 distribution are larger than those of

empirical distributions of LRT statistic at different tail probabilities, making chi-square approximation more conservative. If we use chi-square approximation for the test, type I error will be smaller than it should be. We also notice that the departure from χ_1^2 distribution is larger in the dataset of random SNPs with MAF >=0.05. The reason is that with lower MAF threshold, the mean number of observed haplotypes is reduced. If a set of three SNPs has many missing haplotypes, the haplotype frequency distribution under the null model (Model 3) is often the same as the alternative model (Model 4) and the observed LRT statistic comparing the two models is zero. In this case, the LRT statistics computed from the resampled data are near zero, too. Therefore, the averaged critical values of empirical LRT statistics at different tail probabilities across all the examined triples tend to be smaller than the theoretical critical values of χ_1^2. In summary, if all eight possible haplotypes are observed, χ_1^2 is a good approximation of the distribution of observed LRT statistic to test for the three-locus disequilibrium. If many haplotypes are unobserved, however, the inference based on χ_1^2 distribution is too stringent and thus the resampling method may be preferred.

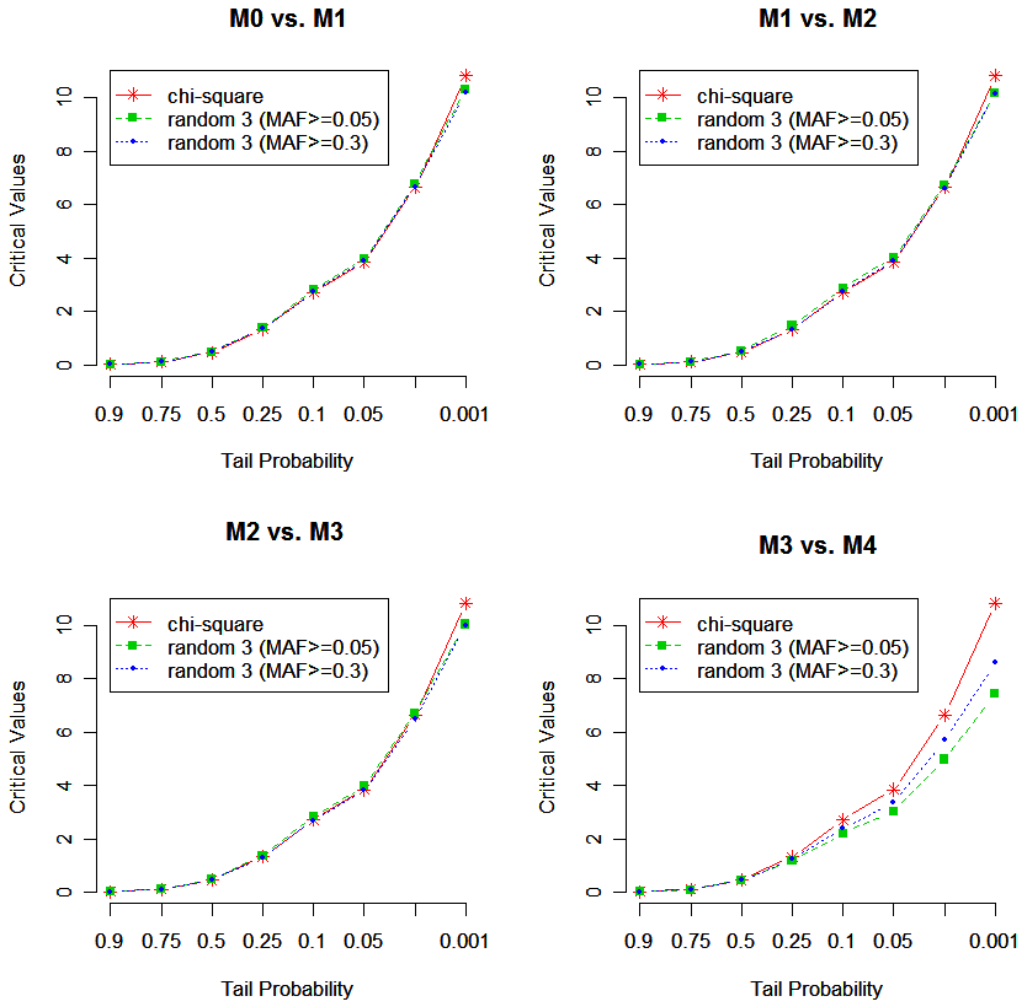


Figure 4.1 Comparison between chi-square distribution and resampled distributions of LRT statistics for different hypothesis tests. For the hypothesis tests of 2-locus LD (M0 vs. M1, M1 vs. M2, M2 vs. M3), chi-square distribution with 1 degree of freedom agrees well with the empirical distributions of LRT statistics regardless of the minor allele frequency threshold. However, for the hypothesis test of three-locus LD (M3 vs. M4), the chi-square distribution with 1 degree of freedom is more conservative than the resampled distribution of LRT statistic. When the allele frequencies of markers are closer to 0.5, the resampled distribution of LRT statistics is closer to the chi-square distribution.

4.3.2 Tests for two- and three-locus disequilibria

We present systematic survey results for two- and three-locus disequilibria in four populations in ENm010 region. From each population, we select SNPs with MAF >=0.3 (140 SNPs from YRI; 144 from CEU; 165 from HCB; 170 from JPT). Such a high MAF threshold is chosen due to a couple of reasons. First, our primary interest is in detecting triples or regions showing significant three-locus disequilibria. Our previous result indicates that three-locus associations cannot be detected in a set of three SNPs with small MAF due to many missing haplotypes. As the MAF threshold increases, the test for three-locus association will be more accurate with many observed haplotypes. In that case, chi-square approximation and the resampling method can be more comparable. Second, when the total number of 3-SNP sets is huge, it is prohibitively intensive to apply the resampling method to all triples.

For all three combinations of SNPs, we compute LRT statistics to test for the global disequilibrium, two-locus disequilibria, and three-locus disequilibrium. For global disequilibrium and two-locus disequilibria tests, p-values are estimated based on only chi-square approximation since chi-square approximation and resampling method yield very similar results. For the test for three-locus disequilibrium, p-values are estimated via the chi-square approximation and resampling method. To adjust multiple comparisons problems, we apply a FDR procedure (Storey 2002) to the raw p-values from different hypothesis tests. We compute the FDR adjusted p-values, called q-values using the "qvalue" function in R package. A null hypothesis is rejected if the corresponding q-value is less than or equal to a significance level α. We control the expected FDR at 5%.

After performing a multiple test adjustment on the p-values estimated from the chi-square approximation, we obtain quite a contrasting result across different populations for the three-locus disequilibrium test. In YRI, we detect 4,510 triples showing significant three-locus association, which takes about 1% of the total number of triples in the population. In CEU, 6,841triples (1.4%) showing significant three-locus associations are detected. However, in HCB and JPT, no significant three-locus associations are detected after adjusting for multiple tests. Even the smallest p-value has the corresponding q-value a lot greater than 0.05, indicating that the false discovery rates are much greater than 5%. We have to tolerate at least 30% of FDR in order to detect some three-locus associations in HCB and JPT. By comparing the average LRT statistics for overall disequilibrium (Model 0 vs. Model 4), all pairwise disequilibria (Model 0 vs. Model 3), and three-locus disequilibrium (Model 3 vs. Model 4) across total triples in each population, we also see that the statistics for three-locus associations are on average much lower in JPT and HCB in comparison with YRI and CEU although the average three-locus LD takes only a small proportion of the overall LD for all populations (see Table 4.1). One possible reason for failure of detecting three-locus LD may be due to the limitation of haplotype phasing since unrelated individuals rather than trios are genotyped in HCB and JPT.

Table 4.1 The average likelihood ratio statistics of total pairwise LD, three-locus LD, and overall LD across all triples

	Total pairwise LD	Three-locus LD	Overall LD
YRI	38.65	1.43	40.08
CEU	58.35	1.5	59.85
HCB	54.10	0.89	55.0
JPT	49.58	0.97	50.55

4.3.3 Three dimensional LD plots

We depict the test results of two- and three-locus LD in 3 dimensional LD plots. Specifically, we divide triples showing either 2-locus or 3-locus disequilibria in each population into three different categories and plot them (see Figures 4.2 – 4.5). From the 3 dimensional LD plots, we indentify an interesting pattern particularly in 3-SNP sets (i.e. triples) of no or little LD between pairs yet strong three-locus LD. Those triples are more dispersed outside of strong pairwise LD regions in YRI while they are more clustered together in CEU. The reason for this observation might be that YRI population has more genetic diversity and is older than CEU. In contrast to YRI and CEU, we do not observe any such triples in JPT and HCB.

Table 4.2 summarizes several properties of three categories in CEU population. We observe very similar characteristics for each category in the other three populations (data not shown). Triples in the first category represented by red color show significant all three pairwise disequilibria yet insignificant three-locus disequilibrium. Regions in the first category are similar to haplotype block regions which appear as different sizes of

triangles in typical 2 dimensional LD plots. The overall LD is quite large, but most of the overall LD is contributed by pairwise LD. The second category represented by green color is consisted of triples showing both significant all pairwise and three-locus disequilibria. Not many triples belong to the second category, but three markers of each triple in the category are very highly correlated and physically close. Triples in the third category represented by blue color have only significant three-locus disequilibrium. For the triples in the third category, the genomic distance between markers is greater than that of the other two categories. All triples in the third category have 7 or 8 observed haplotypes. The overall LD in the third category is smaller than that in the first or second category, but the contribution of three-locus disequilibrium to the overall LD is much greater.

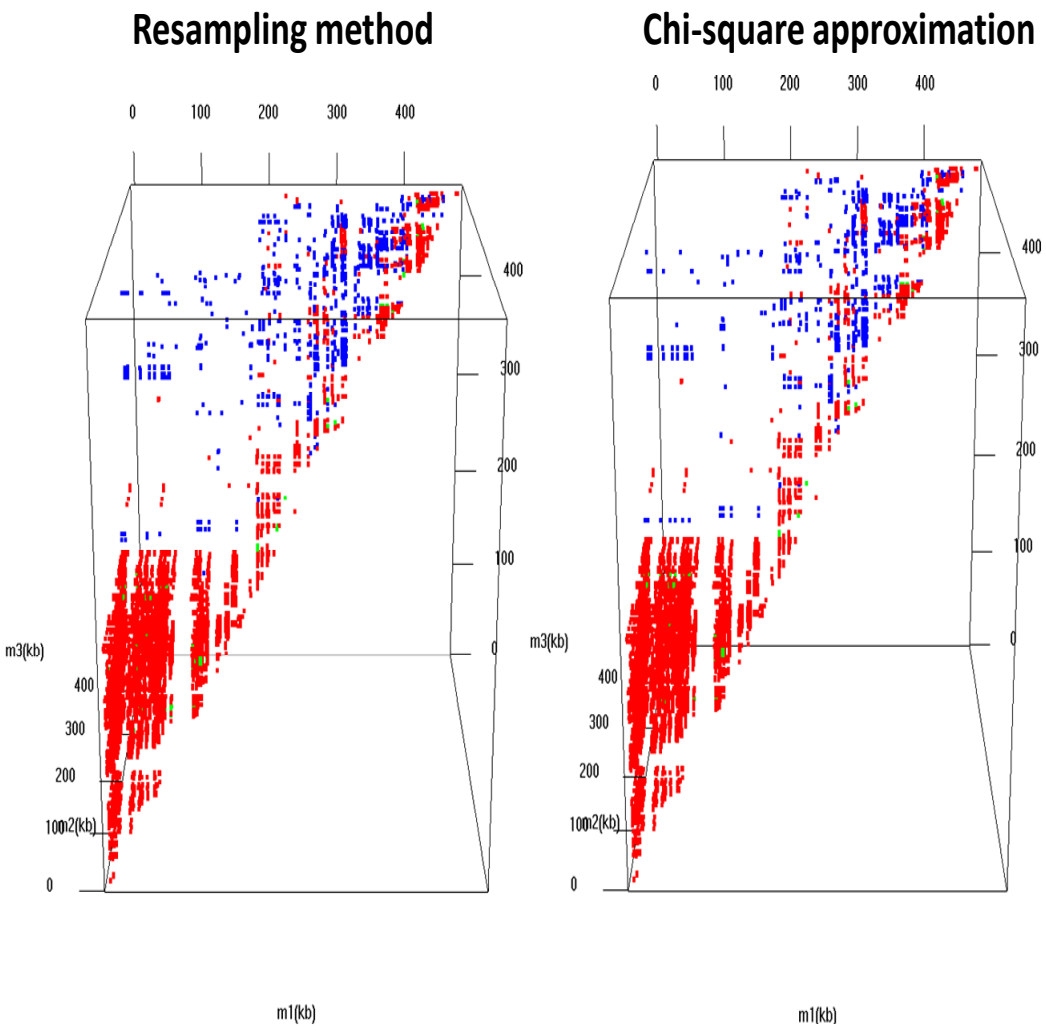


Figure 4.2 Regions showing significant 2-locus or 3-locus disequilibria in <u>YRI</u>
Red, green, and blue represent sets of 3-SNP (i.e. triples) showing significant all pairwise disequilibria but insignificant three-locus disequilibrium, triples showing both significant all pairwise disequilibria and significant three-locus disequilibrium, and triples showing only significant three-locus disequilibrium, repectively.

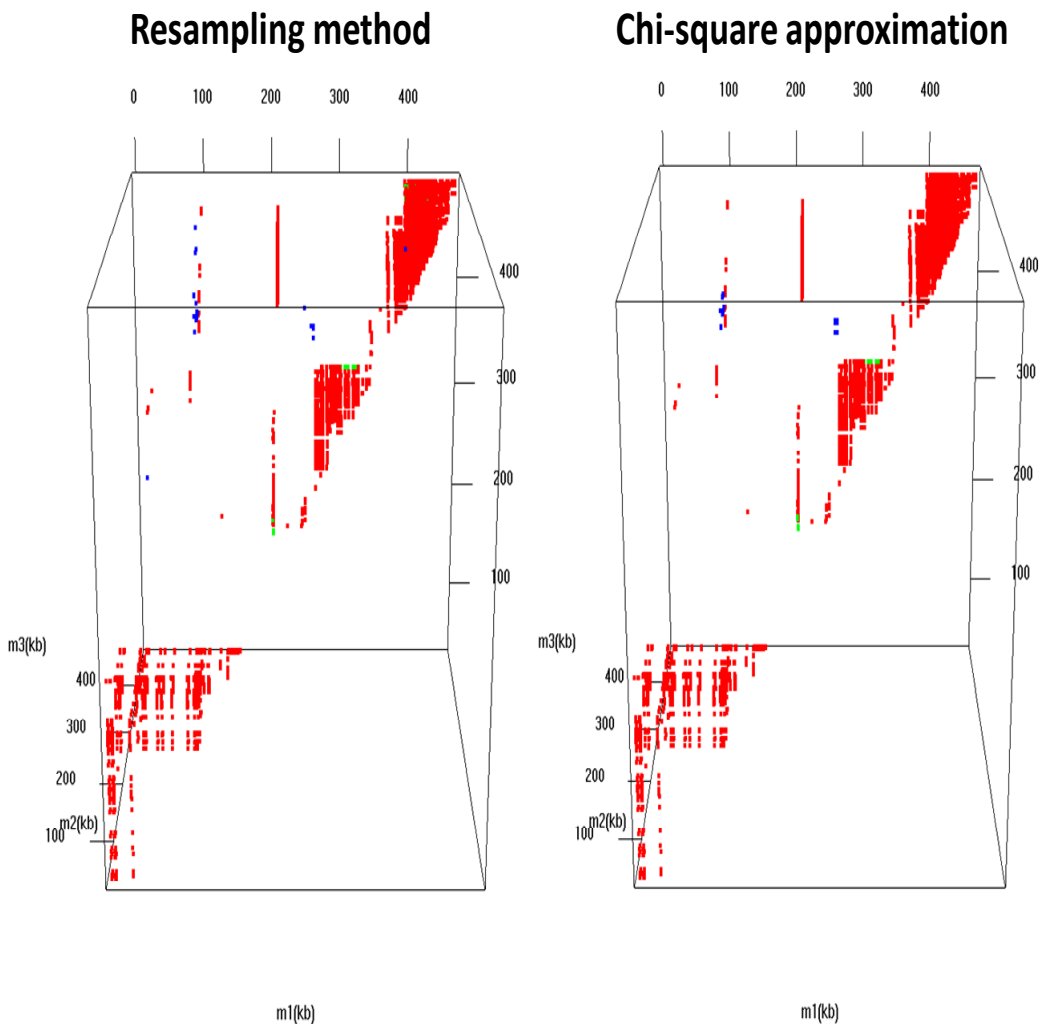


Figure 4.3 Regions showing significant 2-locus or 3-locus disequilibria in HCB
The color coding is same as Figure 4.2.

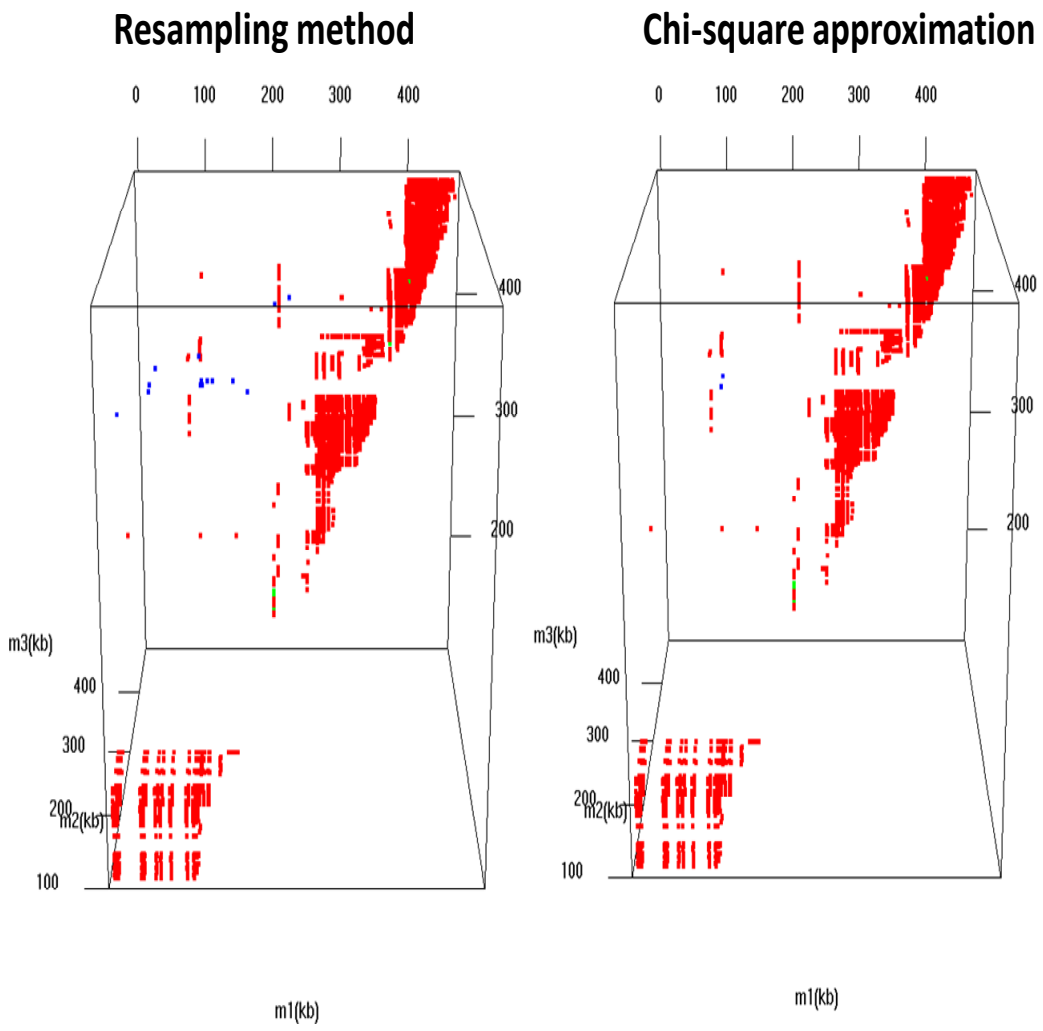


Figure 4.4 Regions showing significant 2-locus or 3-locus disequilibria in JPT
The color coding is same as Figure 4.2.

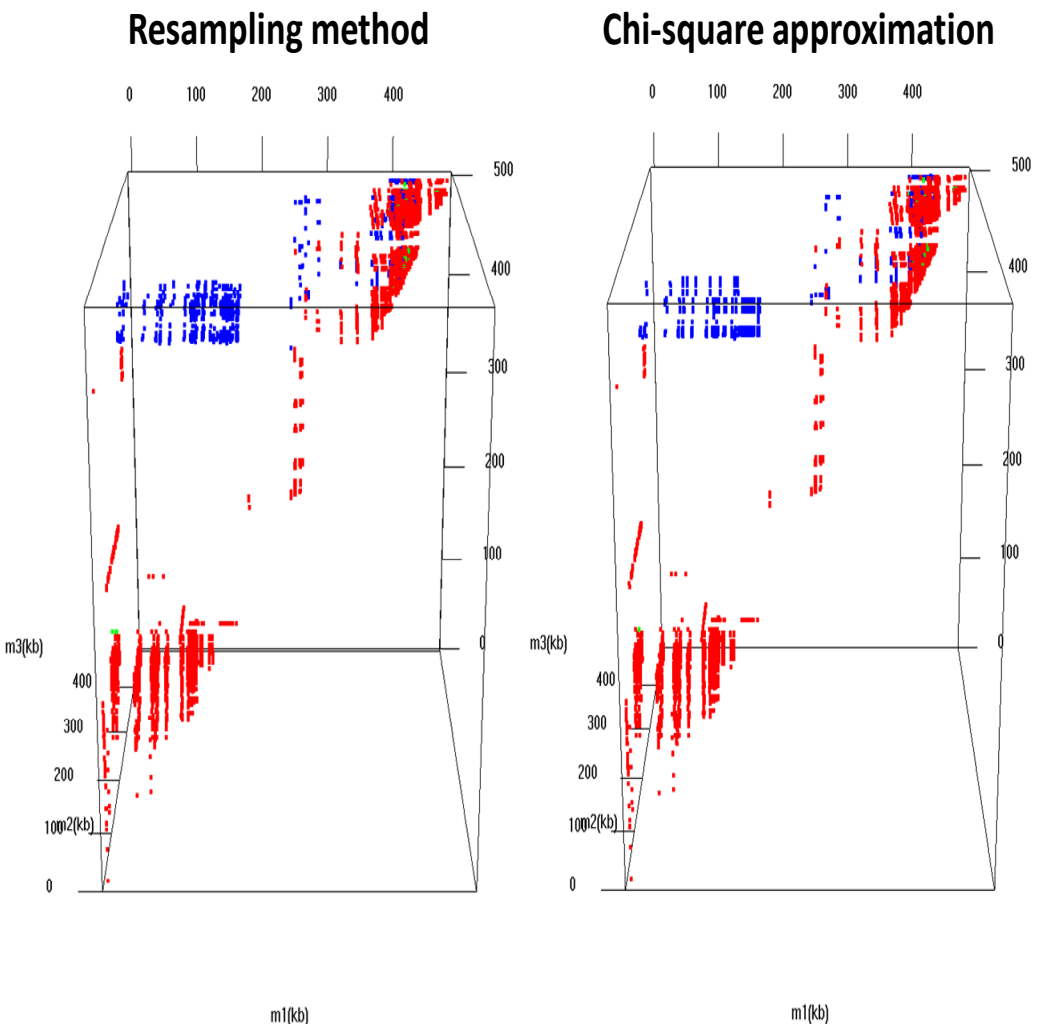


Figure 4.5 Regions showing significant 2-locus or 3-locus disequilibria in CEU
The color coding is same as Figure 4.2.

Table 4.2 Summary of triples showing significant 2-locus or 3-locus disequilibria in CEU

	Significant three pairwise disequilibria and insignificant 3-locus disequilibrium		Significant three pairwise disequilibria and significant 3-locus disequilibrium		Insignificant three pairwise disequilibria and significant 3-locus disequilibrium	
	Resampling	Chi-square	Resampling	Chi-square	Resampling	Chi-square
Number of triples[a] (Proportion of total triples)	11106 (2.3%)	11126 (2.3%)	399 (0.08%)	379 (0.08%)	1147 (0.2%)	1108 (0.2%)
Mean overall LD[b]	122.2	122.2	83.5	82.0	16.67	17.28
Mean three-locus LD[c]	0.007	0.007	15.94	16.20	12.92	13.59
% Contribution of three-locus LD to overall LD (mean)	0.0074	0.0074	21	21	82	79
Number of observed haplotypes	3~8	3~8	5~8	5~8	7~8	7~8
Mean distance between pairs of markers (kb)	20.38	20.35	19.83	20.36	111	101.3

Table 4.2 (continued)

a: Using chi-square approximation, raw p-values <= 0.04 (Model 0 vs. Model 1), 0.05 (Model 1 vs. Model 2), 0.0028 (Model 2 vs. Model 3), 0.0007 (Model 3 vs. Model 4) are used as cutoffs to control expected FDR at 5%. Using 1,000 times of resampling, p-values of 0 is used as a cutoff to declare the significance of three-locus association.
b: Overall total LD is measured by the log-likelihood difference between Model 0 and Model 4.
c: Three-locus LD is measured by the log-likelihood difference between Model 3 and Model 4.

4.3.4 Determinants of three-locus disequilibrium

To determine conditions in which a triple is more likely to have significant three-locus association, we checked a couple of possible factors. First, we compute the correlation between observed LRT statistic testing for three-locus disequilibrium and $|W|$, where W is defined as

$$W = (P_{ABC}P_{Abc}P_{aBc}P_{abC}) - (P_{ABc}P_{AbC}P_{aBC}P_{abc}).$$

The maximum absolute value W can take is 0.00391 when the following haplotype frequencies are observed: ($P_{ABC} = P_{Abc} = P_{aBc} = P_{abC} = 0.25,\ P_{ABc} = P_{AbC} = P_{aBC} = P_{abc} = 0$) or ($P_{ABC} = P_{Abc} = P_{aBc} = P_{abC} = 0,\ P_{ABc} = P_{AbC} = P_{aBC} = P_{abc} = 0.25$). As the observed LRT statistic testing for three-locus disequilibrium increases, the absolute value of W also increases. The correlation between two variables is 0.79 in YRI. The other three populations show similar magnitudes of correlation. If at least one two-locus marginal haplotype frequencies is zero, W is zero and thus it is less likely to detect significant three-locus association. In triples having three or less haplotypes, W is always zero and three-locus association cannot be detected. In triples having four or more haplotypes, it is possible to detect three-locus association since W can take nonzero values. Second, we checked whether number of observed haplotypes in a triple is an important determinant for detection of three-locus association. We group total triples by their observed number of haplotypes and count the number of triples whose three-locus associations are declared to be significant. Table 4.3 shows the result in YRI population using p-values estimated from chi-square approximation. In triples with four or less haplotypes, no significant three-locus associations are detected. Their overall disequilibria are fully explained by

the associations of pairs of markers. In triples with five or more haplotypes, we can detect significant three-locus association. Using a resampling method and other populations, we obtain very similar results (data not shown).

Table 4.3 Triples showing significant 2-locus or 3-locus disequilibria detected by chi-square approximation in YRI

Observed number of haplotypes in a triple	Number of triples	Triples in global disequilibrium	Triples showing significant 3-locus LD			Total triples showing significant 3-locus LD
			No significant 2-locus LD	One or two significant 2-locus LD	All three significant 2-locus LD	
2	69	69	0	0	0	0
3	816	816	0	0	0	0
4	11153	11153	0	0	0	0
5	16123	16123	0	30	38	68
6	57761	57761	0	104	89	193
7	50789	50789	64	941	51	1056
8	310869	209835	1588	1567	38	3193
Total	447580	346546	1652	2642	216	4510

Note: To control expected FDR at 5%, raw p-values <= 0.212 (Model 0 vs. Model 4), 0.07 (Model 0 vs. Model 1), 0.03 (Model 1 vs. Model 2), 0.003 (Model 2 vs. Model 3), 0.0005 (Model 3 vs. Model 4) are used as cutoffs.

4.4 Discussion

From the tests for two- and three-locus LD using the data in ENm010 region, we observed quite different patterns in regions of little pairwise LD yet strong three-locus LD across different populations. In HCB and JPT, no such regions are detected. In YRI, those regions are more scattered while in CEU they are more clustered together into a few groups. We also evaluated the adequacy of asymptotic chi-square distribution for hypothesis tests of different orders of LD by comparing chi-square distribution with the resampled distribution of LRT statistics. Our analysis shows that chi-square approximation tends to be more conservative than a resampling method for the test for three-locus disequilibrium. The distributional discrepancy between two methods is greater when minor allele frequency of each marker is low.

Different population demographic history may be responsible for quite different pattern in regions of high order LD across four populations. YRI has longer demographic history and has gone through many recombination events and typically shows higher level of haplotype diversity than non-African populations. Since the number of observed haplotypes is an important factor in whether we can detect three-locus disequilibrium, high level of haplotype diversity throughout the whole region in YRI may suggest one reason for scattered three-locus LD regions. However, no detection of three-locus disequilibrium in JPT and HCB may be caused by some limitation in inference of haplotype phases from unrelated individuals' genotypes.

Incorporating the knowledge of high order LD between markers and a putative causal variant in multilocus association mapping is potentially more informative and more powerful than utilizing only pairwise LD between each marker and the causal

variant. For example, there are situations where no or little LD between any pairs of markers and a functional site but three-locus LD among them can be strong. Thus, the functional site can be identified by consideration of high order LD. Nielsen et al. (2004) discussed that when moderate to high levels of high order LD exist, haplotype tests tend to be more powerful whereas single marker tests tend to prevail when pairwise LD is high. Although there are many factors to affect the power of single-SNP and haplotype-based analysis including specific LD pattern in a region of interest, measuring the magnitudes of various LD orders and testing for their significances may be useful in designing more effective strategies for identification of causal variant(s) responsible for complex diseases. Moreover, selecting sets of SNPs that might have high order LD may be utilized for potential screening of interacting genes and investigating epistatic effects on complex traits.

A resampling method used to evaluate the significance of three-locus LD has advantages and disadvantages over chi-square approximation. When all eight possible haplotypes are observed and marker allele frequencies are not far from 0.5, chi-square approximation and resampling method produce similar results. However, when many haplotypes are unobserved, p-values estimated from the chi-square approximation and those from the resampling method are very different. Since the resampling method does not depend on the large sample requirement, p-values estimated from the resampling method may be more reliable. However, one drawback of the resampling method compared to the chi-square approximation is the computational burden especially when we consider all three combinations of many markers in a large region. Thus, for the test of three-locus disequilibrium, it may be reasonable to use the chi-square approximation

as an initial step and then to apply the resampling method to confirm the suggestive findings.

A large sample size is definitely required to detect three-locus LD reliably because more diverse haplotypes can be observed in a large sample. Detection of higher-order LD requires much larger sample size. However, even if the sample size is sufficiently large enough, there is still possibility that some haplotypes cannot be observed in some regions such as recombination cold spots. In such regions, it will be more difficult to detect significant high order LD because the LD between pairs of markers is almost complete (i.e. equal to the overall LD).

Chapter 5

General Discussion and Conclusions

With growing belief that haplotypes may hold the key to better understand human evolutionary history and to identify genetic variants responsible for complex diseases, characterization of haplotype and LD patterns throughout the human genome has been of great interest in the past ten years. Due to the concept of haplotype blocks, the chromosomes can be relatively simply structured such that each chromosome is divided into many blocks, within which the haplotype diversity is quite limited. Characterization of haplotype blocks has been popular because it can provide association studies with a useful view of broad patterns of LD and an intermediate step to select informative haplotype-tagging SNPs. However, there is no universally accepted definition of haplotype blocks and the block structures are strongly affected by the block definition as well as the markers and samples. Moreover, since block-like patterns have been determined empirically, their biological relevance still needs to be investigated. Although there are usually recombination events between blocks, substantial LD is also found between loci in different blocks and blocks cannot be regarded as independent units. Therefore, using each block as a unit in designing genetic association studies may not be the most efficient strategy (Zhao et al., 2003; Li et al, 2007).

When multiple closely linked markers, often in LD, in a chromosome region are studied to assess the association between the markers and traits of interest, there is a general belief that analysis based on haplotypes may be more powerful than individual

marker analysis. However, the power comparison between these two types of analysis revealed some contradictory findings. Some studies support the single marker analysis (Long et al., 1999) and others favor the haplotype-based analysis (Akey et al., 2001). It is likely that one method may outperform the other depending on certain disease models and certain LD patterns. By comparing the power between single marker and haplotype-based case- control tests of three loci including one putative causal locus, Nielsen et al. (2004) demonstrated that these contradicting findings could be explained by three-locus LD among the two markers and one putative causal variant. Haplotype-based analysis performed better when there were moderate to high levels of three-locus LD while single marker analysis was better when pairwise LD between each marker and the causal variant was high. Therefore, the power of detecting associations between markers and phenotypes was closely related with the patterns of two-locus and higher order LD.

Strategies for performing haplotype-based analyses have been the subject of open debate and active research. Some important practical issues in haplotype-based association tests that are still under active research are how to handle large number of haplotypes and reduce the number of degree of freedom without much loss of power, how to treat the missing phase information, how to adjust for the multiple testing problem, and how to choose the number of adjacent SNPs that should be considered simultaneously. This thesis does not directly address any of the above issues which limit haplotype-based association studies in practice. However, the results in this thesis have implications for disease gene mapping, suggesting a possible strategy for making disease gene mapping more powerful. Utilizing the knowledge of multilocus LD structure in a region of interest may make it possible to choose between single marker analysis and

haplotype-based analysis. For example, a LD map that shows the magnitudes of various LD orders may be useful for predicting statistical power of association studies. However, due to two issues - finite sample size and a large number of LD coefficients - exhaustive estimation and testing of all types of LD coefficients for long haplotypes seems to be very difficult. Under such circumstances, our approach using MOMC models provides an easy and sound solution by describing the general LD patterns and estimating the relative contributions of various LD orders in neighboring markers along a region of interest. When haplotypes are relatively short such as 3-SNP and the region of interest is small such as a candidate gene region, we can exhaustively estimate and test all types of LD to dissect the LD structure as we demonstrated in Chapter 4.

5.1 Summary of Main Results

In Chapter 2, we summarize multilocus LD pattern in a region using multiple order Markov Chain (MOMC) models and a dynamic window algorithm. High / low LD regions are identified and represented in terms of Markov Chain (MC) order along a chromosome. In a LD profile created by our method, high MC order indicates high LD regions such as haplotype blocks while low MC order indicates low LD regions such as recombination hotspots. However, different definitions of a block using different thresholds (MC order 1, MC order 2, and so on) result in various sizes of blocks and chromosomal coverages. By increasing thresholds, the average block length becomes shorter and chromosomal coverage decreases. From the analysis of four populations in ENm010 region, we observe that Asian samples (JPT and HCB) and the sample from Utah in USA (CEU) have a similar long extent of LD and longer block sizes compared to

an African population (YRI), which agrees with the findings from many other groups. Comparison of typical block characteristics using our method with those using two other methods (Gabriel et al., 2002; Wang et al. 2002) reveals that both D' and our method detect much longer blocks than four gamete test (Wang et al. 2002) and our method shows the greatest chromosomal coverage. The general LD patterns detected by our method are similar to those obtained by the other two methods (Gabriel et al. 2002; Wang et al. 2002), but specific attributes of blocks do not agree in all three methods. Even within the same method, selection of different thresholds changes many attributes of blocks. Especially, the block boundaries are very sensitive to the choice of threshold. These facts illustrate the subjectivity of haplotype block definition and prevent the conclusive characterization of the region's block structure.

We focus on two goals in Chapter 3. One is to characterize general multilocus LD pattern among multiple markers using multiple order Markov chain (MOMC) models as a statistical framework. The other is to reveal the complex LD structure among multiple neighboring markers by estimating the percentage contributions of various lower order LD to the total multilocus LD. Our multilocus LD measure based on the log-likelihood difference between Markov chain order 0 model and Markov chain order w-1 model for w consecutive markers is very similar to the entropy based multi-marker LD measure (Nothnagel et al. 2002), due to the connections between the entropy concept and likelihood theory. Most multilocus LD measures in the literature including Nothnagel's entropy-based multilocus LD measure provide a single index of association among multiple markers, but they do not reveal the complex patterns and different levels of LD structure. In contrast to those multilocus LD measures, our method can provide not only

the overall measure of association among multiple markers but also more detailed information about the LD structure. By applying our method to the ten ENCODE regions of the HapMap project, we observe that (1) Most LD in the ENCODE regions is attributed to the LD between adjacent pairs of markers across the whole region; (2) LD between adjacent pairs of markers appears to be more significant in high multilocus LD regions than in low multilocus LD regions. High orders of LD such as three or more marker associations become more noticeable in low multilocus LD regions. These facts lead us to conclude that high level of LD in a region does not necessarily indicate high proportion of high order LD because high level of LD is usually associated with low number of haplotypes and detection of high order LD requires a large sample size and many diverse haplotypes.

Chapter 4 presents the results of statistical tests for two- and three-locus LD in sets of three SNPs from the data of four different populations in ENm010 region. The motivation for detection of high order LD is that increased understanding of multilocus LD structure in a genomic region of interest can be helpful in designing more powerful multi-SNP association studies because the power of multi-SNP analysis appears to increase as the magnitude of high order LD increases. The significance of the likelihood ratio test statistic for three-locus disequilibrium is evaluated via asymptotic chi-square approximation and empirical distributions of LRT statistic (i.e. resampling method). We also apply a FDR procedure to account for the multiple testing problems. The key observations from our analysis of HapMap data are as follows. First, for the test for three-locus disequilibrium, chi-square approximation tends to be more conservative than the resampling method, thus less number of triples showing significant three-locus

disequilibrium are detected via chi-square approximation. The distributional discrepancy between two methods is greater when the minor allele frequency of each marker is far from 0.5. Second, we notice that detection of significant three-locus disequilibrium highly depends on the sample size and requires high haplotype diversity and uneven distribution of haplotype frequencies. Finally, after controlling FDR at 5%, no triples showing significant three-locus associations are detected in HCB and JPT, whereas triples showing significant three-locus association are more dispersed in YRI and they are clustered together in CEU.

5.2 Advantages and Limitations

The International HapMap project has generated genome-wide and densely spaced sequence variation data in several human populations from Asia, Africa, and Europe. For this type of data, traditional pairwise LD measures alone are suboptimal to assess the background levels of LD and multilocus LD measures such as ours will certainly provide more information to assess the variability of background correlation across genomic regions.

Since multiple order Markov Chain (MOMC) models can represent dependencies between alleles of successive markers and incorporate various LD orders among multiple markers, the idea of employing MOMC models to describe the local LD pattern is interesting. However, our first attempt to summarize the local LD pattern in terms of Markov chain order turns out to be problematic because the MC order does not directly indicate the magnitude of LD. Rather, it indicates the number of associated markers in a region. For example, MC order 7 implies 7-step dependency among markers. To

overcome this problem, we use the normalized log-likelihood difference between MC 0 and MC *w*-1 as a measure of multilocus LD. Since the normalized values range between 0 (linkage equilibrium) and 1(complete LD), they can indicate the magnitude of LD directly. When a very long sequence of markers is considered, pairwise LD measures have to deal with exponentially large number of pairs. However, our approach calculates multilocus LD for each sliding window and thus it can provide convenient profiles of LD pattern even for a very long sequence. The idea of sliding window is quite useful for long sequences, but one downside is that LD between markers that are more distant than the specified window cannot be captured.

In addition to measuring LD among multiple markers, our approach provide more detailed information about the underlying structure of LD by partitioning the total multilocs LD into components of lower order disequilibria. Partitioning demonstrates that the total multilocus LD is mainly accounted for by adjacent pairs of markers and this fact explains why there is a good agreement between the LD profile using our multilocus LD measure and those using pairwise LD measures.

In section 1.2.3, we have discussed several model-based LD measures which provide powerful tools for estimating population recombination rates. In contrast to those methods based on explicit evolutionary models, our approach does not make model assumptions about population history nor incorporates the recombination rates. Thus, our approach can be regarded as non-parametric summary statistics for the LD pattern. When the assumptions on evolutionary models are violated, our non-parametric approach might provide more robust assessment for the levels and patterns of LD. However, since our approach does not directly relate LD patterns to some biological mechanisms which

create or maintain LD in a population, it may be difficult to apply our approach to detect the biological mechanisms such as the underlying recombination or natural selection.

Our multilocus LD measure computes the level of association among multiple markers in each window of size *w*, but physical distance between markers is not incorporated into the calculation. A model adapted from Malecot equation has been used to describe the decline of LD with increasing physical distance *d* between markers (Maniatis et al., 2002). This method is used to fit the Malecot parameter ε (the exponential decline of association between two loci) and construct LD maps with a map location in εd LD units. This LD map can describe the patterns of LD in the form of a metric map. However, LD is known to be highly variable over distance depending on regions and populations. For example, Abecasis et al. (2001) estimated that physical distance could account for less than 50 % of the variation in LD in their study. Thus, it is not clear how helpful such explicit incorporation of physical distance will be in describing the extent of LD.

Another limitation of our method is the input of only haplotype data, assuming that they are directly observed or the phases are inferred accurately. Unless the haplotype data are available, the inference of phases using statistical or computational approaches such as EM algorithm (Excoffier et al., 1995), the coalescence-based algorithm (Stephens et al., 2001), and the partition-ligation algorithm (Niu et al., 2002) is required as an intermediate step. Since most available genetic data are unphased genotype data and many existing phasing algorithms have some limitations, further study needs to be done in order to use the genotype data directly as an input for our method.

Bibliography

Abecasis GR, Norguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 68:191-197

Akey J, Jin L, Xiong M (2000) Haplotypes vs. single marker linkage disequilibrium tests: What do we gain? *Eur J Hum Genet* 9:291-300.

Ardlie KG, Kruglyak L, Seislstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nature Review Genetics* 3: 299-309

Ayres KL, Balding DJ (2001) Measuring gametic disequilibrium from multilocus data. *Genetics* 157: 413-423

Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nature Review Genetics* 7: 781-791

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nature Review Genetics* 4: 99-111

Bartlett MS (1935) Contingency table interactions. *J R Statist Soc Suppl* 2: 248-252

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate-a practical and powerful approach to multiple testing. *J Roy Stat Sco Ser B Meth* 57:289-300

Bennett JH (1954) On the theory of random mating. *Ann Eugen* 184:311-317

Berchtold A, Raftery AE (2002) The mixture transition distribution model for high-order Markov Chains and non-Gaussian time series. *Statistical Science* 17 (3): 328-356

Blenau W, Baumann A (2001) Molecular and pharmacological properties of insect bioamine receptors: lessons from *Drosophila melanogaster* and *Apis mellifera*. *Arch. Insect Biochem. Phys.* 48:13-38

Brown AHD (1975) Sample sizes required to detect linkage disequilibrium between two or three loci. *Theor. Popul. Biol*. 8:184-201

Carlson CS et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74: 106-120

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56: 18-31

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111-122.

Clark AG, Nielsen R, Singnorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E (2003) Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. Am J Hum Genet 73: 285-300

Clark AG (2004) The role of haplotypes in candidate gene studies. *Genet Epid* 27:321-333

Clayton D (2001) http://www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf

Czika WA (2003) Accounting for within- and between- locus dependencies in Marker association tests. PhD thesis, North Carolina State University

Daly MJ, Rioux JD, Schaffner SF, Hudson TH, Lander ES (2001) High resolution haplotype structure in the human genome. *Nature Genet* 29:229-232

Dawson E, Abecasis GR et al. (2002) A first generation linkage disequilibrium map of human chromosome 22. *Nature* 418: 544-548

de Bakker PI et al. (2005) Efficiency and power in genetic association studies. *Nat Genet* 37:1217-1223

de Bakker PI et al. (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* 38:1298-1303

DeLuca M, Roshina NV, Geiger-Thornsberry GL, Lyman RF, Pasyukova EG, Mackay TF (2003) Dopa decarboxylase (Ddc) affects variation in Drosophila longevity. *Nature Genetics* 34: 429-433

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

Ding K et al. (2005) The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. *Mol Biol Evol* 22: 148-159

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet* 28:361-364.

Excoffier L, Slatkin M (1995) Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921-927

Fallin DA, Cohen L et al. (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* 11:143-151

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics* 159: 1299-1318

Feng S. (2004) Statistical studies of genomics data. PhD thesis. North Carolina State University.

Fienberg SE (1970) The analysis of multidimensional contingency tables. *Ecology* 51 (3): 419-433

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229

Gange SJ (1995) Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 49(2): 134-138

Goebel B, Dawy Z, Hagennauer J, Mueller J. (2005) An approximation to the distribution of finite sample size mutual information estimates. *IEEE Proc* 2:1102-1106

Goodman LA (1969) On partitioning chi-square and detecting partial association in three-way contingency tables. *J Roy Statist Soc B* 31: 486-498

Greenspan G, Geiger D (2006) Modeling haplotype blocks using Markov Chains. *Genetics* 172:2583-2599

Hedrick PW (1987) Gametic disequilibrium measures: Proceed with caution. *Genetics* 117: 331-341

Hill WG (1974a) Disequilibrium among several linked neutral genes in finite population. I. Mean changes in disequilibrium. *Theor Popu Biol* 5: 366-392

Hill WG (1974b) Disequilibrium among several linked neutral genes in finite population. I. Variances and covariances of disequilibria. *Theor Popu Biol* 6: 184-198

Hill WG (1975) Tests for association of gene frequencies at several loci in random mating diploid populations. *Biometrics* 31:881-888.

Hill WG (1976) Chapter Non-random association of neutral linked genes in finite populations. *Population Genetics and Ecology* pp. 339-376, New York: Academic Press.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genoem patterns of common DNA variation in three human populations. *Science* 307: 1072-1079

Hudson R (2001) Two-locus sampling distributions and their application. *Genetics* 159: 1805-1817

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147-164

Huttley G, Smith M, Carrington M, O'Brien S (1999) A scan for linkage disequilibrium across the human genome. *Genetics* 152: 1711-1722

Jeffreys AJ, Kauppi L, Neuman R (2001) Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet* 29(2): 217-222

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

Ke X et al. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* 13: 577-588

Kimura M, Otha T (1971) *Theoretical aspects of population genetics*. Princeton, NJ: Princeton University Press

Kingman JFC (1982) The coalescent. *Stoch. Proc. Appl.* 13:235-248

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308: 385-389

Kristin GA, Kruglyak L, Seielstad, M (2002) Patterns of linkage disequilbrium in the human genome. *Nat Rev Genet* 3(4):299-309

Kullback S. 1978. Information theory and Statistics. New York: John Wiley & Sons.

Lancaster HO (1951) Complex contingency tables treated by partition of χ^2. *J Roy Stat Soc Series B* 13: 242-249

Lawrence RW, Evans DM, Cardon LR (2005) Prospects and pitfalls in whole genome association studies. *Phil Trans R Soc B* 360:1589-1595

Lewontin RC (1988) On measures of gametic disequilibrium. *Genetics* 120: 849-852

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165: 2213-2233

Li Y, Sung WK, Liu JJ (2007) Association mapping via regularized regression analysis of single-nucelotide polymorphism haplotypes in variable-sized sliding windows. *Am J Hum Genet* 80:705-715

Liu Z, Lin S (2005) Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet Epidemiol* 29: 353-364.

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

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* 56:799-810

Lunan KD, Mitchell HK (1969) The metabolism of tyrosine O phosphate in *Drosophila*. *Arch. Biochem. Phys.* 132: 450-456

May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ (2002) Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nature Genet* 31(3): 272-275

McKeigue PM, Carpenter JR, Parra EJ, Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann Hum Genet* 64:171-186

McVean G, Awadalla P, Fearnhead P (2002a) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160: 1231-1241

McVean GA (2002b) http://www.stats.ox.ac.uk/~mcvean/L8notes.pdf

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304: 581-584

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res* 24: 4841-4843.

Morris RW (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiology* 23(3): 221-233

Morton NE, Zhang W, Taillon-Miller P et al. (2001) The optimal measure of allelic association. *Proc natl Acad Sci USA* 98: 5217-5221

Muller JC (2004) Linkage disequilibrium for different scales and applications. *Briefings in bioinformatics* 5 (4) : 355-364

Nielsen DM, Ehm MG, Zaykin DV, Weir BS (2004) Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics* 168:1029-1040.

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70:157-169.

Nothnagel M (2004) The definition of multilocus haplotype blocks and common diseases. PhD thesis, Humboldt University, Berlin, Germany.

Nothnagel M, Frst R, Rohde K (2002) Entropy as a measure for linkage disequilibrium over multiple haplotype blocks. *Hum Hered* 54: 186-198

Nothnagel M, Rohde K (2005) The Effect of Single-Nucleotide Polymorphism Marker Selection on Patterns of Haplotype Blocks and Haplotype Frequency Estimates. *Am J Hum Genet* 77:988-998.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, LEE DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294: 1719-1723

Phillips MS et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hotspots. *Nature Genet* 33: 382-387

Plackett RL (1962) A note on interactions in contingency tables. *J R Statist Soc B* 24:162-166

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

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. *Nature* 411:199-204

Rinaldo A, Bacanu SA, Delvin B, Sonpar V, Wasserman L, Roeder K (2005) Characterization of multilocus linkage disequilibrium. *Genet Epidem* 28:193-206

Robinson WP, Asmussen MA, Thomson G (1991) Three-locus systems impose additional constraints on pairwise disequilibrium. *Genetics* 129: 925-930

Sabatti C, Risch N (2002) Homozygosity and linkage disequilibrium. *Genetics* 160:1707-1719

Sabeti PC et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-837

Schaid DJ (2004) Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 166:5 05-512

Schaid DJ (2004) Evaluating associations of haplotypes with traits. *Genet Epid* 27:348-364

Schulze TG, Zhang K et al. (2004) Defining haplotype blocks and tag SNPs in the human genome. *Hum Mol Genet* 13:335-342

Smouse PE (1974) Likelihood analysis of recombinational disequilibrium in multiple-locus gametic frequencies. *Genetics* 76: 557-565

Stephens JC, Rogers J, Ruano G (1990) Theoretical underpinning of the single-molecule-dilution (SMD) method of direct haplotype resolution. *Am J Hum Genet* 46:1149-1155

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978-989.

Storey JD (2002) A direct approach to false discovery rates. *J Roy Stat Soc Ser B Meth* 64: 479-498

Stram DO et al. (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from Multiethnic Cohort Study. *Hum Hered* 55: 27-36

Sun X, Stephens JC, Zhao H (2004) The impact of sample size and marker selection on the study of haplotype structures. *Hum Genomics* 1: 179-193

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437: 1299-1320

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-862

Thomas A, Camp NJ (2004) Graphical modeling of the joint distribution of alleles at associated loci. *Am J Hum Genet* 74:1088-1101

Thomson G, Baur M (1984) Third order linkage disequilibrium. *Tissue Antigens* 24:250-255

Wall JD (2000) Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154:1271-1279

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Review Genetics* 4:587-597

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 71(5): 1227-1234

Weale ME et al. (2003) Selection and evaluation of tagging SNPs in the neuronalsodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* 73: 551-565

Weir BS (1996) *Genetic Data Analysis* II. Sunderland, MA: Sinauer Associates, Inc

Weir BS, Cockerham CC (1978) Testing hypotheses about linkage disequilibrium with multiple alleles. *Genetics* 88: 633-642

Weir (1979) Inferences about linkage disequilibrium. *Biometrics* 35: 235-254

Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S (2005) Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genetics* 6 (Suppl I):S134

Zeggini E et al. (2005) Characterization of the genomic architecture of human chromosome 17q and evaluation of different methods for haplotype block definition. *BMC Genet* 6: 21

Zhang K, Calbrese P, Nordborg M, Sun F (2002a) Haplotype block structure and its application to association studies: power and study designs. *Am J Hum Genet* 71: 1386-1394

Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. *Bioinformatics* 19 (10): 1300-1301.

Zhang W, Collins A, Morton NE (2004) Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum Genet* 115 (2):157-164

Zhao H, Pfeiffer R, Gail MH (2003) Haplotype analysis in population genetics and association studies. *Pharmacogenomics* 4(2): 171-178

Appendices

Appendix A

1. For a 2 marker system (M_i, M_j) with $i = a$ or A; $j = b$ or B

Consider two biallelic markers with the alleles a, A; b, B. The frequencies of the four haplotypes at the two markers can be uniquely expressed in terms of the allele frequencies, P_A and P_B, and the pairwise linkage disequilibrium, denoted as D_{AB} as following:

$$P_{AB} = P_A P_B + D_{AB}$$

$$P_{Ab} = P_A P_b - D_{AB}$$

$$P_{aB} = P_a P_B - D_{AB}$$

$$P_{ab} = P_a P_b + D_{AB} \quad .$$

With $N = n_{ab} + n_{aB} + n_{Ab} + n_{AB}$ (N is the sample size), the likelihood is

$$L = \frac{N!}{n_{ab}!n_{aB}!n_{Ab}!n_{AB}!}(P_{ab})^{n_{ab}}(P_{aB})^{n_{aB}}(P_{Ab})^{n_{Ab}}(P_{AB})^{n_{AB}}$$

Then, if the constant term in the likelihood is ignored, the log-likelihood can be expressed as

$$\log L \propto n_{ab}\log(P_{ab}) + n_{aB}\log(P_{aB}) + n_{Ab}\log(P_{Ab}) + n_{AB}\log(P_{AB}) \quad ,$$

where $\propto$ denotes "proportional to". And the log-likelihoods of $MC0$ and $MC1$ can be written as

$$LL(MC0) = \sum_i \sum_j n_{ij} \log(P_{i.} \times P_{.j})$$

$$LL(MC1) = \sum_i \sum_j n_{ij} \log(P_{ij}) \quad .$$

Now, the expected difference between the log-likelihood of *MC*1 and the log-likelihood of *MC*0 is

$$\frac{1}{N} E[LL(MC1) - LL(MC0)] = \sum_i \sum_j E(\frac{n_{ij}}{N}) \times \log\left(\frac{P_{ij}}{P_{i.} \times P_{.j}}\right)$$

$$= \sum_i \sum_j P_{ij} \log\left(\frac{P_{ij}}{P_{i.} \times P_{.j}}\right) = \sum_i \sum_j (D_{ij} + P_{i.} \times P_{.j}) \log\left(\frac{D_{ij} + P_{i.} \times P_{.j}}{P_{i.} \times P_{.j}}\right)$$

$$= \sum_i \sum_j (D_{ij} + P_{i.} \times P_{.j}) \log\left(1 + \frac{D_{ij}}{P_{i.} \times P_{.j}}\right)$$

Using $\log(1+x) \cong x - \frac{x^2}{2}$ by a Taylor series expansion of order 2, the above expressions can be approximated as

$$\cong \sum_{i=a}^{A} \sum_{j=b}^{B} (D_{ij} + P_{i.} \times P_{.j}) \times (\frac{D_{ij}}{P_{i.} \times P_{.j}} - \frac{1}{2} \times \frac{D_{ij}^2}{(P_{i.} \times P_{.j})^2}) = \frac{1}{2} \times \frac{D_{AB}^2}{P_a P_A P_b P_B}$$

$$= \frac{1}{2} r_{AB}^2$$

2. For a 3 marker system (M_i, M_j, M_k) in this order with *i* = *a* or *A*, *j* = *b* or *B*, *k* = *c* or *C*

When considering three biallelic markers with the alleles *a*, *A*; *b*, *B*; and *c*, *C*, there are eight possible haplotypes. The seven independent haplotype frequencies can be expressed in terms of the three allelic frequencies P_A, P_B, P_C, the three pairwise linkage

disequilibria (D_{AB}, $D_{BC,}$ D_{AC}) and the three-locus linkage disequilibrium (D_{ABC}) as following [Robinson et al., 1991]:

$$P_{ABC} = P_A P_B P_C + P_A D_{BC} + P_B D_{AC} + P_C D_{AB} + D_{ABC}$$

$$P_{Abc} = P_A P_b P_c + P_A D_{BC} - P_b D_{AC} - P_c D_{AB} + D_{ABC}$$

$$P_{aBc} = P_a P_B P_c - P_a D_{BC} + P_B D_{AC} - P_c D_{AB} + D_{ABC}$$

$$P_{abC} = P_a P_b P_C - P_a D_{BC} - P_b D_{AC} + P_C D_{AB} + D_{ABC}$$

$$P_{ABc} = P_A P_B P_c - P_A D_{BC} - P_B D_{AC} + P_c D_{AB} - D_{ABC}$$

$$P_{AbC} = P_A P_b P_C - P_A D_{BC} + P_b D_{AC} - P_C D_{AB} - D_{ABC}$$

$$P_{aBC} = P_a P_B P_C + P_a D_{BC} - P_B D_{AC} - P_C D_{AB} - D_{ABC}$$

$$P_{abc} = P_a P_b P_c + P_a D_{BC} + P_b D_{AC} + P_c D_{AB} - D_{ABC}$$

With $N = n_{abc} + n_{abC} + n_{aBc} + n_{aBC} + n_{Abc} + n_{AbC} + n_{ABc} + n_{ABC}$ (N denotes the sample size), the log-likelihoods are proportional to

$$\log L \propto n_{abc} \log(P_{abc}) + n_{abC} \log(P_{abC}) + ... + n_{ABc} \log(P_{ABc}) + n_{ABC} \log(P_{ABC}) .$$

The log-likelihoods of $MC0$, $MC1$, and $MC2$ are

$$LL(MC0) = \sum_{i=a}^{A} \sum_{j=b}^{B} \sum_{k=c}^{C} n_{ijk} \log(P_{i..} \times P_{.j.} \times P_{..k})$$

$$LL(MC1) = \sum_{i=a}^{A} \sum_{j=b}^{B} \sum_{k=c}^{C} n_{ijk} \log(P_{i..} \times \frac{P_{ij.}}{P_{i..}} \times \frac{P_{.jk}}{P_{.j.}})$$

$$LL(MC2) = \sum_{i=a}^{A} \sum_{j=b}^{B} \sum_{k=c}^{C} n_{ijk} \log(P_{ijk}) \quad .$$

(2A) Approximation of $\frac{1}{N}E[LL(MC1) - LL(MC0)]$:

$$\frac{1}{N}E[LL(MC1)-LL(MC0)]$$

$$=\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}E\left(\frac{n_{ijk}}{N}\right)\left[\log\left(P_{i..}\times\frac{P_{ij.}}{P_{i..}}\times\frac{P_{.jk}}{P_{.j.}}\right)-\log(P_{i..}\times P_{.j.}\times P_{..k})\right]$$

$$=\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\log\left(\frac{P_{i..}\times\frac{P_{ij.}}{P_{i..}}\times\frac{P_{.jk}}{P_{.j.}}}{P_{i..}\times P_{.j.}\times P_{..k}}\right)=\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\log\left(\frac{P_{ij.}\times P_{.jk}}{P_{i..}\times P_{.j.}^{2}\times P_{..k}}\right)$$

$$=\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\log\left(\frac{P_{ij.}}{P_{i..}\times P_{.j.}}\times\frac{P_{.jk}}{P_{.j.}\times P_{..k}}\right)$$

$$=\underbrace{\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\log\left(\frac{P_{ij.}}{P_{i..}\times P_{.j.}}\right)}_{(1)}+\underbrace{\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\log\left(\frac{P_{.jk}}{P_{.j.}\times P_{..k}}\right)}_{(2)}$$

By expanding (1),

$$\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\log\left(\frac{P_{ij.}}{P_{i..}\times P_{.j.}}\right)=P_{abc}\log\left(\frac{P_{ab}}{P_{a}\times P_{b}}\right)+P_{abC}\log\left(\frac{P_{ab}}{P_{a}\times P_{b}}\right)$$
$$+P_{aBc}\log\left(\frac{P_{aB}}{P_{a}\times P_{B}}\right)+P_{aBC}\log\left(\frac{P_{aB}}{P_{a}\times P_{B}}\right)+P_{Abc}\log\left(\frac{P_{Ab}}{P_{A}\times P_{b}}\right)+P_{AbC}\log\left(\frac{P_{Ab}}{P_{A}\times P_{b}}\right)$$
$$+P_{ABc}\log\left(\frac{P_{AB}}{P_{A}\times P_{B}}\right)+P_{ABC}\log\left(\frac{P_{AB}}{P_{A}\times P_{B}}\right)$$

By collecting common terms, we have

$$= (P_{abc} + P_{abC}) \times \log\left(\frac{P_{ab}}{P_a \times P_b}\right) + (P_{aBc} + P_{aBC}) \times \log\left(\frac{P_{aB}}{P_a \times P_B}\right)$$
$$+ (P_{Abc} + P_{AbC}) \times \log\left(\frac{P_{Ab}}{P_A \times P_b}\right) + (P_{ABc} + P_{ABC}) \times \log\left(\frac{P_{AB}}{P_A \times P_B}\right) .$$

Using $\log(1+x) \cong x - \frac{x^2}{2}$ by a Taylor series expansion of order 2, (1) can be approximated as following:

$$\cong (P_a P_b + D_{AB}) \times \left(\frac{D_{AB}}{P_a P_b} - \frac{1}{2} \times \frac{D_{AB}^2}{(P_a P_b)^2}\right) + (P_a P_B - D_{AB}) \times \left(\frac{-D_{AB}}{P_a P_B} - \frac{1}{2} \times \frac{D_{AB}^2}{(P_a P_B)^2}\right)$$
$$+ (P_A P_b - D_{AB}) \times \left(\frac{-D_{AB}}{P_A P_b} - \frac{1}{2} \times \frac{D_{AB}^2}{(P_A P_b)^2}\right) + (P_A P_B + D_{AB}) \times \left(\frac{D_{AB}}{P_A P_B} - \frac{1}{2} \times \frac{D_{AB}^2}{(P_A P_B)^2}\right)$$
$$= \frac{1}{2} \times \frac{D_{AB}^2}{P_A P_a P_B P_b}$$
$$= \frac{1}{2} \times r_{AB}^2$$

From (2), we also get $\frac{1}{2} \times r_{BC}^2$.

Therefore, for 3 markers with the ordering of M_i, M_j, and M_k, the expected difference between the log-likelihood of $MC1$ and the log-likelihood of $MC0$ is approximately equal to

$$\frac{1}{N} E[LL(MC1) - LL(MC0)] \cong \frac{1}{2} r_{ij}^2 + \frac{1}{2} r_{jk}^2 \quad ,$$

where $r_{ij}^2 = \frac{(P_{ij.} - P_{i..} \times P_{.j.})^2}{P_{i..} \times P_{.j.} \times (1 - P_{i..}) \times (1 - P_{.j.})}$ and $r_{jk}^2 = \frac{(P_{.jk} - P_{.j.} \times P_{..k})^2}{P_{.j.} \times P_{..k} \times (1 - P_{.j.}) \times (1 - P_{..k})}$.

(2B) Approximation of $\frac{1}{N} E[LL(MC2) - LL(MC0)]$:

$$\frac{1}{N}E[LL(MC2)-LL(MC0)]=\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\log\left(\frac{P_{ijk}}{P_{i..}\times P_{.j.}\times P_{..k}}\right)$$
$$=\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\log\left(1+\frac{D_{ijk}+P_{i..}D_{jk}+P_{.j.}D_{ik}+P_{..k}D_{ij}}{P_{i..}\times P_{.j.}\times P_{..k}}\right)$$

Using $\log(1+x)\cong x-\frac{x^2}{2}$ by a Taylor series expansion of order 2, the above expressions can be approximated as following:

$$\cong\sum_{i=a}^{A}\sum_{j=b}^{B}\sum_{k=c}^{C}P_{ijk}\times\left(\frac{D_{ijk}+P_{i..}D_{jk}+P_{.j.}D_{ik}+P_{..k}D_{ij}}{P_{i..}\times P_{.j.}\times P_{..k}}-\frac{1}{2}\times\left(\frac{D_{ijk}+P_{i..}D_{jk}+P_{.j.}D_{ik}+P_{..k}D_{ij}}{P_{i..}\times P_{.j.}\times P_{..k}}\right)^2\right).$$

By expanding the summations, we have

$$=P_{abc}\times\left(\frac{(-D_{ABC})+P_aD_{BC}+P_bD_{AC}+P_cD_{AB}}{P_aP_bP_c}-\frac{1}{2}\times\left(\frac{(-D_{ABC})+P_aD_{BC}+P_bD_{AC}+P_cD_{AB}}{P_aP_bP_c}\right)^2\right)$$
$$+P_{abC}\times\left(\frac{D_{ABC}+P_a(-D_{BC})+P_b(-D_{AC})+P_CD_{AB}}{P_aP_bP_C}-\frac{1}{2}\times\left(\frac{D_{ABC}+P_a(-D_{BC})+P_b(-D_{AC})+P_CD_{AB}}{P_aP_bP_C}\right)^2\right)$$
$$+P_{aBc}\times\left(\frac{D_{ABC}+P_a(-D_{BC})+P_BD_{AC}+P_c(-D_{AB})}{P_aP_BP_c}-\frac{1}{2}\times\left(\frac{D_{ABC}+P_a(-D_{BC})+P_BD_{AC}+P_c(-D_{AB})}{P_aP_BP_c}\right)^2\right)$$
$$+P_{aBC}\times\left(\frac{(-D_{ABC})+P_aD_{BC}+P_B(-D_{AC})+P_C(-D_{AB})}{P_aP_BP_C}-\frac{1}{2}\times\left(\frac{(-D_{ABC})+P_aD_{BC}+P_B(-D_{AC})+P_C(-D_{AB})}{P_aP_BP_C}\right)^2\right)$$
$$+P_{Abc}\times\left(\frac{D_{ABC}+P_AD_{BC}+P_b(-D_{AC})+P_c(-D_{AB})}{P_AP_bP_c}-\frac{1}{2}\times\left(\frac{D_{ABC}+P_AD_{BC}+P_b(-D_{AC})+P_c(-D_{AB})}{P_AP_bP_c}\right)^2\right)$$
$$+P_{AbC}\times\left(\frac{(-D_{ABC})+P_A(-D_{BC})+P_bD_{AC}+P_C(-D_{AB})}{P_AP_bP_C}-\frac{1}{2}\times\left(\frac{(-D_{ABC})+P_A(-D_{BC})+P_bD_{AC}+P_C(-D_{AB})}{P_AP_bP_C}\right)^2\right)$$

$$+P_{ABc}\times\left(\frac{(-D_{ABC})+P_A(-D_{BC})+P_B(-D_{AC})+P_cD_{AB}}{P_AP_BP_c}-\frac{1}{2}\times\left(\frac{(-D_{ABC})+P_A(-D_{BC})+P_B(-D_{AC})+P_cD_{AB}}{P_AP_BP_c}\right)^2\right)$$

$$+P_{ABC}\times\left(\frac{D_{ABC}+P_AD_{BC}+P_BD_{AC}+P_CD_{AB}}{P_AP_BP_C}-\frac{1}{2}\times\left(\frac{D_{ABC}+P_AD_{BC}+P_BD_{AC}+P_CD_{AB}}{P_AP_BP_C}\right)^2\right) \quad .$$

With some algebra and substitution of eight haplotype frequencies with their corresponding expressions in terms of allelic frequencies, pairwise linkage disequilibrium coefficients, and three-locus linkage disequilibrium coefficient, we have the following approximation.

$$\cong\frac{1}{2}\times\frac{D_{ABC}^2}{P_AP_BP_CP_aP_bP_c}+\frac{1}{2}\times\frac{D_{AB}^2}{P_AP_BP_aP_b}+\frac{1}{2}\times\frac{D_{BC}^2}{P_BP_CP_bP_c}+\frac{1}{2}\times\frac{D_{AC}^2}{P_AP_CP_aP_c}$$

Or it can be rewritten as

$$\frac{1}{2}\times r_{ABC}^2+\frac{1}{2}\times r_{AB}^2+\frac{1}{2}\times r_{BC}^2+\frac{1}{2}\times r_{AC}^2 \quad ,$$

where $r_{ABC}^2=\frac{D_{ABC}^2}{P_AP_BP_CP_aP_bP_c}$; $r_{AB}^2=\frac{D_{AB}^2}{P_AP_BP_aP_b}$; $r_{BC}^2=\frac{D_{BC}^2}{P_BP_CP_bP_c}$; $r_{AC}^2=\frac{D_{AC}^2}{P_AP_CP_aP_c}$.

Therefore, for 3 markers with the ordering of M_i, M_j, and M_k, the expected difference between the log-likelihood of $MC2$ and the log-likelihood of $MC0$ is approximately equal to

$$\frac{1}{N}E\left[LL(MC2)-LL(MC0)\right]\cong\frac{1}{2}r_{ijk}^2+\frac{1}{2}r_{ij}^2+\frac{1}{2}r_{jk}^2+\frac{1}{2}r_{ik}^2 \quad ,$$

where $r^2_{ijk} = \frac{(D_{ijk})^2}{P_{i..} \times P_{.j.} \times P_{..k} \times (1-P_{i..}) \times (1-P_{.j.}) \times (1-P_{..k})}$;

$r^2_{ik} = \frac{(D_{ik})^2}{P_{i..} \times P_{..k} \times (1-P_{i..}) \times (1-P_{..k})}$; three-locus disequilibrium (D_{ijk}) is defined as $P_{ijk} - P_{i..}D_{jk} - P_{.j.}D_{ik} - P_{..k}D_{ij} - P_{i..}P_{.j.}P_{..k}$.

.

Appendix B

1. Generalization of the approximation of expected difference between *LL*(*MC l*)

$(1 \le l \le w-1)$ **and *LL*(*MC* 0):**

$$\frac{1}{N} E[LL(MC1) - LL(MC0)] \propto \sum_{i=1}^{w-1} r^2_{i(i+1)} \ , \tag{B1}$$

where $\propto$ denotes "proportional to".

$$\frac{1}{N} E[LL(MC2) - LL(MC0)] \propto \sum_{i=1}^{w-2} r^2_{i(i+2)} + \sum_{i=1}^{w-2} r^2_{i(i+1)(i+2)} + \text{all terms in (B1)} \tag{B2}$$

$$\frac{1}{N} E[LL(MC3) - LL(MC0)] \propto \sum_{i=1}^{w-3} r^2_{i(i+3)} + \sum_{i=1}^{w-3} \sum_{j=i+1}^{i+2} r^2_{ij(i+3)} + \sum_{i=1}^{w-3} r^2_{i(i+1)(i+2)(i+3)} + \text{all terms in (B2)} \tag{B3}$$

$$\frac{1}{N} E[LL(MC4) - LL(MC0)] \propto \sum_{i=1}^{w-4} r^2_{i(i+4)} + \sum_{i=1}^{w-4} \sum_{j=i+1}^{i+3} r^2_{ij(i+4)} + \sum_{i=1}^{w-4} \sum_{\substack{j_1 \ne j_2 \\ \in [i+1,\, i+3]}} r^2_{ij_1j_2(i+4)} + \sum_{i=1}^{w-4} r^2_{i(i+1)(i+2)(i+3)} + \text{all terms in (B3)} \ , \tag{B4}$$

where $\sum_{\substack{j_1 \neq j_2 \\ \in [i+1,\ i+3]}}$ denotes all possible combination of two different markers between M_{i+1} and M_{i+3}. Finally, the expected difference between $LL(MC\ l)$ $(5 \leq l \leq w-1)$ and $LL(MC\ 0)$ can be written as

$$\begin{aligned}\frac{1}{N}E[LL(MC\ l) - LL(MC0)] \propto & \sum_{i=1}^{w-l} r^2_{i(i+l)} + \sum_{i=1}^{w-l}\sum_{j=i+1}^{i+l-1} r^2_{ij(i+l)} \\ & + \sum_{i=1}^{w-l} \sum_{\substack{j_1 \neq j_2 \\ \in [i+1,\ i+l-1]}} r^2_{ij_1j_2(i+l)} \\ & + \sum_{i=1}^{w-l} \sum_{\substack{j_1 \neq j_2 \neq j_3 \\ \in [i+1,\ i+l-1]}} r^2_{ij_1j_2j_3(i+l)} \\ & \cdots \\ & + \sum_{i=1}^{w-l} \sum_{\substack{j_1 \neq j_2 \neq \ldots \neq j_{l-1} \\ \in [i+1,\ i+l-1]}} r^2_{ij_1j_2\cdots j_{l-1}(i+l)} \\ & + \text{all terms in } \frac{1}{N}E[LL(MC\ l-1) - LL(MC0)]\ ,\end{aligned}$$

where $\sum_{\substack{j_1 \neq j_2 \neq j_3 \\ \in [i+1,\ i+l-1]}}$ and $\sum_{\substack{j_1 \neq j_2 \neq \ldots \neq j_{l-1} \\ \in [i+1,\ i+l-1]}}$ denote all possible combination of three different markers and (l-1) different markers between M_{i+1} and M_{i+l-1}, respectively.