RICE UNIVERSITY

Haplotype Block and Genetic Association

by

Zhaoxia Yu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Rudy Guerra, Chairman
Professor of Statistics

Christopher I. Amos
Professor of Department of Epidemiology,
MD Anderson Cancer Center

Marek Kimmel
Professor of Statistics

Lisa Marie Meffert
Professor of Department of Ecology and
Evolutionary Biology

HOUSTON, TEXAS

AUGUST, 2005
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Abstract

Haplotype Block and Genetic Association

by

Zhaoxia Yu

The recently identified (Daly et al. 2001 and Patil et al. 2001) block-like structure in the human genome has attracted much attention since each haplotype block contains limited sequence variation, which can reduce the complexity in genetic mapping studies. This dissertation focuses on estimating haplotype block structures and their application to genetic mapping using single nucleotide polymorphisms (SNPs) from unrelated individuals. Among other issues, the traditional single marker association study leads to the problem of multiple testing, which is still not well understood in the context of genomewide association studies. The haplotype-based approach is one way to lessen certain problems caused by multiple testing. There is also evidence that haplotype based tests have higher statistical power. We first propose a novel approach to estimate haplotype blocks based on pairwise linkage disequilibrium (LD). The application to simulated data shows that our new approach has higher
power than several existing methods in identifying haplotype blocks. We also examine the impact of marker density and different tagging strategies on the estimation of haplotype blocks. We introduce a new statistic to measure the difference between two different block partitions. Applying the new statistic to real and simulated data we show that a higher marker density is needed than previously expected in order to recover the true block structure over a given region. Finally, we analyzed a real SNP data set. A comparison of the haplotype-SNP based method to the more traditional single-SNP based method shows that the two methods tend to agree more when haplotype block sizes are small. On the other hand, the haplotype-SNP based approach does not always have higher power than the single-SNP based study as is supported by theoretical considerations. Indeed, long haplotype blocks where the LD structure might be very complex can lead to inferior power compared to single-SNP approaches. In practice, it is recommended that single-SNP analyses be run routinely, especially in the presence of moderate to long blocks.
Acknowledgements

First, I want to express my deep and sincere gratitude to my thesis advisor Dr. Rudy Guerra, who admitted me to Rice and made my graduation from Rice University possible. Without his guidance, patience and encouragement, I would not have finished my thesis on time. I thank my other committee Drs. Chris Amos, Marek Kimmel and Lisa Meffert, who have provided great ideas during my thesis proposal and read my manuscript carefully. I also want to thank Drs. David Scott, Javier Rojo and Dennis Cox. Without their excellent lectures, I would not have finished my dissertation in such a statistically significant way. Many thanks to Dr. Peter Olofsson, whose office was on the other side of mine. Peter not only taught me Brownian motion but also helped me open my door hundreds of times using his master key when I locked myself out.

I also want thank my collaborators: Dr. Helen Hobbs and Dr. Jonathan Cohen of University of Texas Southwestern Medical Center at Dallas. With their excellent work and new data, I was motivated to work in this area and finish my thesis.

I cannot omit the other members of Rudy’s group: Blair Christian, Garrett Fox, Jason Gershman and Josue Noyola-Martinez, especially Garrett Fox and Jason Gershman, who helped me on revising my draft. They have been my audience in all of my practical talks in the past three years. I thank them for their
patience, excellent suggestions, and moral support.

I want to mention my appreciation to many staff, students and faculty members in our department, especially my neighbor and classmate Bo Peng, who has helped me greatly in programming. Viachesla (Slava) Fofanov, Talithia Williams, and many others also assisted me very much in revising my draft.

I would like to give my special thanks to Zhongping Jian, who was my fiancé during my writing and now is my husband. He always encourages and supports me. I am devoted to my parents Tongquan Yu and Qingtian Zhao. Like most Chinese parents, they have given me whatever they have, but they never have taken away my own freedom.

Lastly, I cannot forget the financial support. This research was mainly supported by a training fellowship from the W.M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology. Funding from the department of statistics at Rice University and contracts from University of Texas Southwestern Medical Center at Dallas also allowed me to focus on research.
Contents

Abstract ii

Acknowledgements iv

List of Figures ix

List of Tables xv

1 Introduction 1

2 Background 5

2.1 Genotype, Phenotype, HWE, and the Pattern of Inheritance 6

2.2 Genetic Recombination, Haplotype, and Haplotype Reconstruction 8

2.3 Linkage Disequilibrium ........................................... 11

2.3.1 LD Measures ................................................... 14

2.3.2 Sources that Influence Linkage Disequilibrium .......... 17

2.4 The application of LD in Genetic Mapping ................... 18

2.5 Haplotype Block Structure and Haplotype based Association Studies 20
### 3 Haplotype Block Estimation

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Two Main Block Partition Methods</td>
<td>23</td>
</tr>
<tr>
<td>3.2</td>
<td>A New Bivariate Summary Statistic of LD</td>
<td>24</td>
</tr>
<tr>
<td>3.3</td>
<td>Gabriel’s Approach</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>A New Approach to Block Estimation</td>
<td>32</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Comparisons</td>
<td>32</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Real Data</td>
<td>33</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Simulated Data</td>
<td>33</td>
</tr>
</tbody>
</table>

### 4 Effect of Marker Density on Haplotype Block Structure

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Background</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparing Partitions</td>
<td>42</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Similarity Between Two Partitions</td>
<td>42</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Hypothesis Tests for Comparing Two Block Partitions</td>
<td>47</td>
</tr>
<tr>
<td>4.3</td>
<td>Data</td>
<td>50</td>
</tr>
<tr>
<td>4.4</td>
<td>Choices for Marker Selection</td>
<td>52</td>
</tr>
<tr>
<td>4.5</td>
<td>Results</td>
<td>53</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Random Selection</td>
<td>54</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Equally Spaced Selection</td>
<td>57</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Block-Free Tagging</td>
<td>58</td>
</tr>
<tr>
<td>4.5.4</td>
<td>Block-based Tagging</td>
<td>63</td>
</tr>
<tr>
<td>4.5.5</td>
<td>Comparison of the Four Marker Selection Methods</td>
<td>63</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Crossovers. Adapted from biotech-adventure.okstate.edu. ............... 10

2.2 Dynamics of $D$. The subscript "t" indicates the generation. ............ 15

2.3 Heat map of pairwise $LD (|D'|)$ for 126 physically ordered SNPs in the
ABC5G5 and ABCG8 region on chromosome 2. This sample includes more
than 1000 African Americans from Dallas County. The SNPs are not placed
at their relative bp locations but just order index. ......................... 22

3.1 The dynamics of theoretical values of $PLD$ with different recombination
rates between two loci that are positively linked. Assume, at the beginning,
there are only two haplotypes $A_1B_1$ and $A_2B_2$ with equal frequencies. . . . 25
3.2 The dynamics of estimated $PLD$ with a fixed rescaled recombination rate of 0.1 but different population sizes when two loci are positively linked. The number of each haplotype at generation $t$ was simulated assuming random mating and no other evolutionary forces. Assume there are only two haplotypes $A_1B_1$ and $A_2B_2$ at generation 0 with equal proportion of 0.5. The curves of $PLD_t$ show the relation of $PLD_t$ with time (in generations) under different sample sizes. The smooth curve is the theoretical value.

3.3 The 95% upper and 5% low confidence bounds of $D'$, $PLD$, and $r^2$. The x-axis indicates the distances (in bp) between the left-most SNPs versus the next consecutive 50 SNPs.

3.4 Estimated haplotype blocks from three simulated data sets using model A. Hashes indicate SNP locations. See text and the caption of Table 3.3 for details.

3.5 Estimated haplotype blocks from three simulated data sets using model B. Hashes indicate SNP locations. See text and the caption of Table 3.3 for details.

4.1 Three haplotype block partitions of the same area. This figure shows Schwartz's 2003 procedure is too strict. Liu et al. 2004 used a similar graph. The black bars and red boxes indicate positions that SNPs were genotyped and estimated blocks, respectively.
4.2 Similarity measure based on Hamming distance. The similarity between \( P_1 \) and \( P_2 \) is 40%. ........................................... 45

4.3 Similarity measure based on score matrix. ........................................... 45

4.4 The estimated null distribution of similarity scores under the null hypothesis that the two partitions share boundaries by chance. ........................................... 49

4.5 Similarity scores between estimated blocks based on 10, 20, 40, 60, and 80% of SNPs selected randomly and the blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red). ........................................... 55

4.6 P-values for similarity scores between estimated blocks based on 10, 20, 40, 60, and 80% of SNPs selected randomly and blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red). ........................................... 56

4.7 Similarity scores between estimated blocks based on equally spaced SNPs (500bp, 1000bp, 2000bp, and 4000bp) and blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red). ........................................... 59

4.8 P-values for similarity scores between estimated blocks based on equally spaced SNPs (500bp, 1000bp, 2000bp, and 4000bp) and blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red). ........................................... 60
4.9 Similarity scores for estimated blocks based on SNPs selected using Carlson’s method. The x-axis ($r$) indicates the levels of statistical correlation when using Carlson’s method to choose tagging SNPs. 61

4.10 P-values of similarity scores for estimated blocks based on SNPs selected using Carlson’s method. The x-axis ($r$) indicates the levels of statistical correlation when using Carlson’s method to choose tagging SNPs. 62

4.11 Similarity scores for estimated blocks using SNPs selected from block-based tagging method. Haplotype blocks were first estimated using all the samples and markers. Then, Carlson’s method was used to select markers within each haplotype block. The x-axis is the same as that of Figure 4.9. 64

4.12 P-values for the similarity scores using block-based tagging method. The x-axis is the same as that of Figure 4.9 65

4.13 Comparing the performance of random selection, equally-spaced selection and Carlson’s tagging in region ABCG5+8 for the African Americans. 67

4.14 Comparing the performance of random selection, equally-spaced selection and Carlson’s tagging in region ABCG5+8 for the Caucasian Americans. 68

4.15 Comparing the performance of random selection, equally-spaced selection and Carlson’s tagging in region APOAV for the African Americans. 69

4.16 Comparing the performance of random selection, equally-spaced selection, Carlson’s tagging and block-based tagging in region APOAV for the Caucasian Americans. 70
5.1 Pairwise $|D'|$ values in the ABCG5+8, APOAV, and CETP loci. Men and women are combined. A red color indicates higher $LD$ while a blue color indicates lower $LD$. .................................................. 81

5.2 Estimated haplotype blocks at three different gene loci. Each tick mark indicates a marker genotyped in that position. ................. 84

5.3 Association study of sitosterol with ABCG5+8 using the single SNP analysis and the haplotype SNP based analysis. ...................... 88

5.4 Corrected P-values using step-wise procedure by Benjamini and Hochberg 1995 for association between sitosterol and ABCG5+8 using single SNP analysis and haplotype SNP analysis. ...................... 89

5.5 Association study of triglyceride and APOAV using the single SNP analysis and the haplotype-block based analysis. ................. 91

5.6 Corrected P-values using the step-wise procedure by Benjamini and Hochberg 1995 of association study between triglyceride and APOAV with the single SNP analysis the and haplotype SNP analysis. ...................... 92

5.7 Association study of HDL with CETP using the single SNP analysis and the haplotype-block based analysis. ......................... 94

5.8 Corrected P-values using the step-wise procedure by Benjamini and Hochberg (1995). .......................................................... 95
5.9 Simulated statistical power based on single-SNP study and haplotype-SNP study using all 112 Hispanic men that didn’t miss any genotype in one short region at CETP locus. .................................................. 100

5.10 Simulated statistical power based on single-SNP study and haplotype-SNP study using all 330 African American men that didn’t miss any genotype in the APOAV region. .................................................. 102

5.11 Simulated statistical power based on single-SNP study and haplotype-SNP study using all 112 Hispanic American men that didn’t miss any genotype in the CETP region. .................................................. 103

6.1 Single-SNP vs haplotype-SNP association. Each SNP has two alleles represented at black (1) and white (0) in the vertical bars. Haplotypes in the bar-figure and tables are read similarly; for example, 30% of the population has haplotype $AD = 11$. See text for additional explanation. Both this figure and Figure 6.2 have been used by Guerra and Yu 2005 ............. 111

6.2 Single-SNP and haplotype-SNP association. See Figure 6.1 for explanation. 111
List of Tables

2.1 Observed haplotype frequencies. Without loss of generality, we assume that $A_1$ and $B_1$ are the minor alleles and $p_{B_1} \leq p_{A_1}$. ........................................ 12

2.2 LD measures. .................................................... 16

2.3 Association for admixed population. The cell values are observed numbers of individuals in different categories. The odds ratios of population 1 and population 2 are both 1. But the odds ratio of the pooled data is 6.76. ... 18

3.1 Observed haplotype frequencies. Notice that there is only one undetermined variable α if the marginal allele frequencies are fixed. ......................... 27

3.2 Expected haplotype frequencies when independence is assumed. ........ 27

3.3 Results of block estimation based on simulated data sets. In the true model, a 100kb region is separated by three recombination hot spots. Cell values are means over 50 runs, each with 200 chromosomes. Only SNPs with observed MAF>0.1 were used. (A) Hot spot lengths (1kb, 1kb, and 1kb); see Figure 3.4. (B) Hot spot lengths (10kb, 1kb, and 10kb); see Figure 3.5. ...... 38
5.1 Model assumption and hypothesis tests for one-way ANOVA, additive, dominant, and recessive models. ........................................... 77

5.2 Comparison between results from single SNP based association study and haplotype block based association study in ABCG5+8 region. SS indicates the number of SNPs that show single SNP significance and blockwise significance; SN is the number of significant single SNPs but not in significant blocks; NS is the number of SNPs that are not significant in single SNP study but are in significant blocks; NN is the number of SNPs that are neither significant in SNP study nor in haplotype-block based study. The values in the parentheses are the percentages. \( P_{X^2} \) is the P-value calculated from the 2 \( \times \) 2 table based on the four numbers. \( P_{size}, P_{length}, \) and \( P_{density} \) are P-values to test if the agreement is a function of block size, block length, or block density (number of SNPs per kb) using logistic regression. The signs in the parentheses indicate how the block size, block length, and block density affect the agreement. For example, a negative sign indicates that the agreement is negatively associated with size or length, i.e., the results from two approaches tend to agree more in short blocks and agree less in long blocks. Notice that when there only one or two nonzero counts, \( P_{X^2} \) cannot be calculated. ........................................... 87
5.3 Comparison between Benjamini-Hochberg corrected results from the single SNP based association study and the haplotype block based association study in the ABCG5+8 region. ............................ 90

5.4 Comparison between results from the single SNP based association study and the haplotype block based association study in the APOAV region. . 90

5.5 Comparison between Benjamini-Hochberg corrected results from the single SNP based association study and the haplotype block based association study in the APOAV region. ............................ 93

5.6 Comparison between results from the single SNP based association study and the haplotype block based association study in the CETP region. . . 93

5.7 Comparison between Benjamini-Hochberg corrected results from the single SNP based association study and the haplotype block based association study in the CETP region. ............................ 96
Chapter 1

Introduction

Identifying genes that contribute to complex diseases continues to be a challenging problem. Complex diseases such as diabetes and cardiovascular diseases involve multiple interacting genetic determinants. Most complex diseases are also influenced by environmental factors, which makes identifying the genes that contribute to certain complex diseases even harder. Traditionally, many markers are tested independently and multiple testing corrections have to be applied in order to control false positive loci. Due to the very large number of markers, most classical methods such as the Bonferroni correction are too conservative because the dependence among markers is not considered. Although some recent methods (Benjamini and Hochberg 1999; Sabatti et al. 2003; Dudoit 2003 from the microarray literature) can be applied to dependent genetic data, the relative strength of dependence among markers has received limited attention with respect to correct multiple testing. We know that dif-
ferent regions have different levels of dependence, hence a universal correction formula might not be proper. Recently, researchers have started to focus on haplotype-based studies because this approach can reduce the multiple testing problem naturally. An added benefit is that haplotypes may contain interacting loci missed by single marker testing. This dissertation brings together with several issues around haplotype-based genetic mapping.

Haplotypes built from a large set of randomly selected markers are difficult to use in genetic mapping because of the large variation of haplotypes arising from the large number of markers. Consider 10 unlinked biallelic markers. There are $2^{10-1} = 512$ possible haplotypes, many of which will be rare in the population and hence sample. Yet, a relatively small region with 20 “correlated” biallelic markers might only contain 3 or 4 different common haplotypes, which makes haplotype-based genetic mapping feasible. The regions with limited haplotype diversity have been defined as haplotype blocks. Within each block, markers are highly correlated in a genetic sense.

Single nucleotide polymorphisms (SNPs) are the largest class of genetic polymorphism. According to Collins et al. 1998, SNPs explain approximately 90% of DNA variation in humans. It has also been estimated that SNPs with minor allele frequency (MAF) greater than 0.1 occur about every 600-1000 base pairs in the human genome (Wang et al. 1998). The densely spaced SNPs have attracted much attention because of their potential application in fine scale mapping, which can be performed by the study of linkage disequilibrium (LD) between a disease (locus) and nearby markers
using population based data.

Chapter 2 reviews the necessary genetic background and current statistical methods for genetic mapping. Genetic mapping has traditionally focused on linkage analysis using pedigree data. However, without careful attention to study design, linkage analysis may lack efficiency, power, and has limited applicability for complex diseases. Another limitation of linkage analysis is that its resolution is typically on the order of several cM, which corresponds to several Mb of DNA in terms of human genome. Researchers have now turned their interests to the often more efficient and powerful association study design, which can be used for fine mapping of complex diseases.

Chapter 3 discusses the haplotype block phenomenon. We review several commonly used haplotype block partition methods. A new algorithm that we found to be more powerful in identifying haplotype blocks is provided. Haplotype tagging has been widely used in order to reduce genotyping effort. In Chapter 4 we discuss the effect of marker density and tagging strategies on haplotype block estimation.

Using population data from the University of Texas Southwestern Medical Center at Dallas we applied both the traditional single-SNP approach and haplotype-based approach to association. The data, methods, and the comparisons between the results from the two approaches are discussed in Chapter 5.

In the last chapter, we discuss the limitations of current approaches and explore future plans.
Outline:

- Chapter 1: Introduction
- Chapter 2: Background
- Chapter 3: Haplotype Block Estimation
- Chapter 4: Effect of Marker Density on Haplotype Block Structure
- Chapter 5: Single vs. Haplotype SNP Association Study
- Chapter 6: Discussion
Chapter 2

Background

The decade of the 1860's was an exciting one for genetics. In 1865 Gregor Johann Mendel published his now famous genetic work based on a series of pea experiments, while in 1869 Friedrich Miescher successfully isolated deoxyribonucleic acid (DNA). Yet, it was not until the 1940's that scientists begin to realize that DNA is the basic hereditary material. There are around 3,000,000,000 base pairs (bp) of nucleotides in the human genome, which are contained in 23 pairs of chromosomes. Each chromosome is made up of a DNA double helix whose length ranges from 55 Mbp to 250 Mbp. In the normal double-stranded DNA sequence, an adenine (A) in one strand always pairs with a thymine (T) in the complimentary strand; a cytosine (C) in one strand always pairs with a guanine (G) in the complimentary strand.

To locate positions with distinctive features in each chromosome, several types of genetic markers have been used. Aside from genes (the DNA sequences that encode
specific functional products), there are four other major markers: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), insertion/deletion polymorphisms, and single nucleotide polymorphisms (SNPs). In recent years, researchers have been particularly interested in SNPs, which are individual point mutations in the human genome. A simplified but working definition of a SNP is a single biallelic marker. According to Wang et al. 1998, SNPs occur about every 600 base pairs, and cover about 90% of the variation in the human genome. The new DNA microchip technology also makes genotyping the whole human-genome SNP possible, which brings great opportunities to researchers who are searching for the association between diseases and genetic variations. In this dissertation, we will focus on SNP data.

2.1 Genotype, Phenotype, HWE, and the Pattern of Inheritance

*Genotype* refers to paired alleles of fixed loci on homologous chromosomes. For example, at a locus with alleles $A_1$ and $A_2$, there are three possible genotypes. There are two homozygous genotypes, $A_1A_1$ and $A_2A_2$ and one heterozygous genotype $A_1A_2$. In 1908, G.H. Hardy and W. Weinberg independently proposed the HWE law. Assume the allele frequency of $A_1$ is $p_{A_1}$ and the allele frequency of $A_2$ is $p_{A_2}$ (which equals $1 - p_{A_1}$ if there are only two alleles). We say that a locus obeys HWE if the relation
between the genotype frequencies of $A_1 A_1$, $A_1 A_2$, and $A_2 A_2$ are $p_{A_1}^2$, $2p_{A_1}p_{A_2}$ and, $p_{A_2}^2$, respectively. It is important to notice that HWE law holds under a series of assumptions that include non-overlapping generations, identical frequencies in males and females, random mating, large population size, negligible migration and mutation and no effect of natural selection for alleles under study. HWE is a basic principle in population genetics. One application of HWE is to estimate allele frequencies since genotype frequencies rather than allele frequencies are usually observed. Suppose the observed genotype frequencies are $p_{A_1 A_1}$, $p_{A_1 A_2}$ and, $p_{A_2 A_2}$. By HWE, the allele frequency of $A_1$ is $p_{A_1 A_1} + \frac{1}{2}p_{A_1 A_2}$.

Different from genotype, phenotype is the observable expression of a gene. For example, the color of eyes and hair. The pattern of inheritance describes how the genotypes are associated with phenotypes. There are three main patterns in autosomal chromosomes: dominant, recessive, and additive. If an allele of a gene is expressed so long as one copy or more is present, then we say the gene is dominant. If an allele of a gene will be expressed when there are 2 identical copies, then we say the gene is recessive. We say a gene is additive if the combined effects of alleles at different loci are equal to the sum of their individual effects.
2.2 Genetic Recombination, Haplotype, and Haplotype Reconstruction

A haplotype is a combination of genetic markers on the same chromosome. Consider two biallelic markers on the same chromosome. The first marker has alleles $A_1$ and $A_2$ and the second marker has alleles $B_1$ and $B_2$. Assume that there are only two haplotypes, $A_1B_1$ and $A_2B_2$, with equal frequency in the first generation. By Mendel’s second law of heredity, the law of segregation, loci are inherited independently. In the second generation, we would expect that the four different haplotypes: $A_1B_1$, $A_1B_2$, $A_2B_1$ and $A_2B_2$ would occur with equal frequency. However, we often tend to see that two loci on the same chromosome assort together more often than expected by independent segregation. That is, we tend to find more parental haplotypes, $A_1B_1$ and $A_2B_2$, than the recombinant haplotypes, $A_1B_2$ and $A_2B_1$. Now we know that Mendel’s segregation rule applies only when two loci are on separated chromosomes or are far away enough on the same chromosome. Loci close together on the same chromosome tend to cosegregate and are said to be in genetic linkage. In the early 20th century, geneticists believed that genes on the same chromosome are completely linked; i.e., loci on the same chromosome are inherited as one unit. However, in 1909 the Belgian cytologist Janssens found that physical breakage (now called chiasma) and exchange of DNA segments, happens during cell meiosis. Figure 2.1 illustrates how crossovers, which occur at chiasma, can form recombinant haplotypes. During
meiosis, the parental chromosomes are duplicated and aligned, creating a tetrad of chromatids. The chromatids may crossover and as a result exchange DNA. The recombination rate \( r \) between two loci is defined as the probability of an odd number of crossovers between the two loci. When a crossover occurs between two loci, the alleles at the loci can be observed to have undergone recombination. That is:

\[
    r = \frac{\#\text{recombinants}}{\#\text{total progeny}}
\]

The recombination rate ranges from 0 to 0.5. When \( r \) between two loci reaches its maximum of 0.5, we say that the two loci segregate independently or are unlinked. This happens when two loci are far apart from each other physically. When two loci are close to each other physically, the recombination rate \( r \) is less than 0.5, and we say they are linked. In general, the closer the loci are, the larger the chance that they are inherited together from generation to generation and the smaller is the recombination rate. Hence, recombination rate is a statistical measure that can be used to estimate the distance between loci, which is the basis of genetic mapping.

There are different methods to estimate the recombination rate, \( r \). Since the main focus of my thesis is association-based genetic mapping, we will omit the literature concerned with the estimation of \( r \). Instead, we will introduce how to estimate haplotype frequencies since both calculation of linkage disequilibrium (LD, Chapters 3 and 4) and haplotype-SNP association studies (Chapter 5) are based on inferred haplo-
types. Haplotype data can be used directly in many calculations. However, collecting haplotype data directly is both difficult and expensive. Although haplotypes can be obtained by laboratory methods (Saiki et al. 1985; Scharf et al. 1986; Newton et al. 1989; Wu et al. 1989; Ruano et al. 1990) or may be deduced without uncertainty in some cases by genotyping pedigrees (Perlin et al. 1994; Sobel and Lange 1996), the costs of these methods can be prohibitive. Hence, inferring haplotypes from easily obtained genotype data continues to be of interest. In 1990, Clark introduced a statistical based algorithm to infer haplotypes. But, there are several problems with this algorithm, one of which is that the algorithm requires some unambiguous haplotypes in the sample. Excoffier and Slatkin 1995 introduced an expectation-maximization
(EM) algorithm for haplotype estimation. In this EM approach, the phased genotype probabilities are treated as the hidden variable. The haplotype frequencies can be calculated via maximizing the likelihood conditional on the updated phased genotype probabilities. In both algorithms HWE is assumed. Simulations from Excoffier and Slatkin 1995 and later from Fallin and Schork 2000 have shown that the performance of the EM algorithm is quite acceptable. To avoid convergence to a local maximum, Tregouet et al. 2004 introduced a stochastic version of the EM algorithm. In addition to the concern about local maxima, computational speed is also a problem. When the number of loci increases, the time needed to find the maximum likelihood estimate and the space needed to store the variables increase exponentially.

To improve the computational speed of haplotype estimation, there are many further developments that can be applied to solve a large number of markers simultaneously. Details can be found in Stephens et al. 2001a and 2001b; Niu et al. 2002; Lin et al. 2002; and Stephens et al. 2003. An overview of haplotype estimation is given by Guerra and Yu 2005.

2.3 Linkage Disequilibrium

Linkage disequilibrium (LD) or allelic association measures the degree of nonrandom gametic association between loci on the same chromosome. The idea of LD was established and greatly developed many years ago (Cockerham 1954; Kempthorne 1955; Lewontin and Kojima 1960). Consider two loci $A$ and $B$, each having two alleles: $A_1$,
\( A_2 \) and \( B_1, B_2 \). There are four possible haplotypes, as shown in Table 2.1, \( A_1B_1 \), \( A_1B_2 \), \( A_2B_1 \) and \( A_2B_2 \). A natural measure of the strength of the gametic association between two markers is the discrepancy between the observed and the expected haplotype frequencies under assumption of independence between two markers, i.e. \( D = p_{A_1B_1} - p_{A_1}p_{B_1} \). If the two makers are not associated, and we assume that we can ignore evolutionary forces (no drift or selection), then under random mating the population frequency of haplotype \( A_iB_j \) will be \( p_{A_i}p_{B_j} \), hence \( D = 0 \), in which case we say the two loci are in *linkage equilibrium* (LE). Else, when \( D \) does not equal 0, we say the two loci are in *linkage disequilibrium* (LD). Since the first statistic \( D \) was introduced, there have been around 10 different scaled measures of linkage disequilibrium (LD) including the commonly used \( D' \) and \( r^2 \).

<table>
<thead>
<tr>
<th></th>
<th>( B_1 )</th>
<th>( B_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>( p_{A_1B_1} )</td>
<td>( p_{A_1B_2} )</td>
</tr>
<tr>
<td>( A_2 )</td>
<td>( p_{A_2B_1} )</td>
<td>( p_{A_2B_2} )</td>
</tr>
<tr>
<td>( p_{B_1} )</td>
<td>( p_{B_2} )</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.1: Observed haplotype frequencies. Without loss of generality, we assume that \( A_1 \) and \( B_1 \) are the minor alleles and \( p_{B_1} \leq p_{A_1} \).

LD is critical to estimate haplotype blocks, a major topic of this thesis. Therefore, we now consider some technical aspects of LD. \( D \) may be positive, zero or negative. To better understand different LD measures we first define the sign of LD. Without loss of generality, we assume \( A_1 \) and \( B_1 \) correspond to the minor alleles for loci \( A \) and \( B \) and \( p_{B_1} \leq p_{A_1} \). If not, we can just relabel them. This restriction assures the uniqueness of the sign of the LD between two loci. Under this situation, we say two
loci are positively associated if the sign of $D = p_{A_1B_1} - p_{A_1p_{B_1}}$ is positive. Otherwise, we say they are negatively associated.

$LD$ describes the phenomenon that loci close to each other tend to be inherited together in meiosis. Initial disequilibrium may be introduced into a population by mutation or migration. Then, over generations, the initial strong association can be eroded by crossovers during meiosis. In order to understand how $LD$ can be eroded by crossovers during meiosis, consider two loci, 1 and 2. At the beginning, locus 1 was biallelic with alleles $A_1$ and $A_2$, while locus 2 had a unique allele, say $B_1$. Hence, there were only two possible haplotypes: $A_1B_1$ and $A_2B_1$, with frequencies $p_{A_1}$ and $p_{A_2}$ since locus 2 has only one type of allele $B_1$. Then a mutation in the second locus happened on all the individuals who had allele $A_2$ at the first locus. Now, loci 1 and 2 are in perfect linkage disequilibrium: the information from locus 1 gives complete information of locus 2, and vice-versa. Denote this current difference between the observed frequency of haplotype $A_1B_1$ and the expected value under no association as $D_0$, the initial value of $LD$. Crossovers can occur during meiosis. In a simplified version, assuming an infinite population size without drift, migration, new mutations or admixture, eventually, the two loci will reach linkage equilibrium so long as the recombination rate between them is greater than 0. Suppose the effective population size is $N$, then the population recombination parameter $\theta$ equals $4Nr$. It is easy to show that at generation $t$ the deviation $D_t$ of observed haplotype frequency from expected is weakened by an exponential factor: $D_t = D_0(1 - \theta)^t$, as illustrated in
2.3.1 LD Measures

$|D|$ can reach its maximum value of 0.25 when the two SNPs are completely associated and the minor allele frequencies both equal 0.5. The range of $D$ is highly dependent on the marginal allele frequencies $p_{A_1}$ and $p_{B_1}$. Hence, this raw difference, $D$, is not suitable for comparing the strength of association among different pairs of loci. In 1964, Lewontin developed a standardized version of $D'$:

$$D' = \begin{cases} \frac{D}{\min(p_{A_1}(1-p_{B_1}), p_{B_1}(1-p_{A_1}))} & \text{if } D > 0 \\ \frac{D}{\min(p_{A_1}p_{B_1}, (1-p_{A_1})(1-p_{B_1}))} & \text{if } D < 0 \end{cases} \quad (2.1)$$

Under complete independence, $D'$ is 0, just as $D$ is; when two SNPs are completely associated, $|D'|$ reaches a maximum value of 1, regardless of the values of the marginal allele frequencies. Hence, $|D'|$ has been widely used in evolutionary genetics. However, values between 0 and 1 do not have a clear and direct interpretation. As pointed out by Lewontin in 1988, there is still one degree of freedom: two populations may have the same $|D'|$ values but the recombination fraction can be very different. Also, estimation of $D'$ shows high variation - $D'$ of two loci can equal 1 even if they are physically far away from each other (Hudson 1985). The estimation of $D'$ is also biased when the sample size is small. The lack of a meaningful biological interpretation motivated the development of several other LD measures for different purposes.
Figure 2.2: Dynamics of $D$. The subscript "t" indicates the generation.
One commonly used $LD$ measure is defined by statistical correlation, $r$, or its square $r^2$, given by Hill and Weir 1994. $r$ is the correlation between two loci with the two alleles on each locus labeled by 0 and 1. It can be shown that $r$ is a function of $D$:

$$r = \frac{D}{\sqrt{p_{A_1}(1 - p_{A_2})p_{B_1}(1 - p_{B_1})}}$$ (2.2)

$|r|$ or $r^2$ reaches its minimum 0 when two markers are independent. Whether $|r|$ or $r^2$ achieves the maximum value 1 depends on whether the allele frequencies are the same. Although $r$ has been also criticized because of its large variation, it is preferred over $D'$ for association studies because it is directly related to the sample size required to detect a disease locus with a certain level of statistical power.

Besides these two most popular measures ($D'$ and $r$), Devlin and Risch 1995 summarize several other widely used measures listed in Table 2.2.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Formula</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$p_{A_1B_1} - p_{A_1}p_{B_1}$</td>
<td>Lewontin and Kojima (1960)</td>
</tr>
<tr>
<td>$D'$</td>
<td>See formula 2.1</td>
<td>Lewontin (1964)</td>
</tr>
<tr>
<td>$r$</td>
<td>$\frac{D}{\sqrt{p_{A_1}(1 - p_{A_2})p_{B_1}(1 - p_{B_1})}}$</td>
<td>Hill and Weir (1994)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$\frac{D}{p_{A_1}(1 - p_{B_1})p_{A_2}B_2}$</td>
<td>Lewin and Bertell (1978)</td>
</tr>
<tr>
<td>$Q$</td>
<td>$\frac{D}{p_{A_1B_1}p_{A_2B_2} + p_{A_1}B_1p_{A_2}B_2}$</td>
<td>Olson and Wijsman (1994)</td>
</tr>
<tr>
<td>$d$</td>
<td>$\frac{D}{(1-p_{A_1})(1-p_{B_1})}$</td>
<td>Nei and Li (1980)</td>
</tr>
</tbody>
</table>

Table 2.2: LD measures.

Both Fisher’s exact test and Pearson’s $\chi^2$ test can be applied to evaluate $LD$ between two loci for phase-known data. Under the null hypothesis (linkage equilibrium),
the variance of $D$ is $\frac{1}{2n} p_A p_A p_B p_B$, and the corresponding statistic

$$X^2 = \frac{2nD^2}{p_A p_A p_B p_B}, \text{ where } n \text{ is the number of subjects,}$$ (2.3)

asymptotically follows a $\chi^2$ distribution with one degree of freedom. $X^2$ also equals $2nr^2$, which indicates that the measure $r^2$ determines the power of detecting association for a given sample size. However, when phase is unknown, this test for $LD$ can be biased (Schaid 2004) because the usual maximum likelihood estimator of haplotype frequencies is under an assumption of Hardy-Weinberg equilibrium (Fallin and Schork 2000), which may or may not hold true.

### 2.3.2 Sources that Influence Linkage Disequilibrium

Linkage disequilibrium measures the association among loci at a population level. Although the age of the loci and recombination are the two most important forces of $LD$, it is also subject to other evolutionary factors. Haplotype diversity might be lost as genetic drift tends to fix certain alleles in each locus. Hence high LD can be created by genetic drift. One reason that LD can extend over long regions is the bottleneck effect. In a recently expanded population with a small number of founders, individuals in the current generation have low heterozygosity, on average, thus high $LD$. There are many ways that selection can affect $LD$. One of them is the so-called “hitchhiking” effect. Natural selection can also increase $LD$ through the hitchhiking effect (Huttley et al. 1999; Parsch et al. 2001; Wang et al. 2002)
and epistasis (Wade 2000). It is well known that population admixture can create artificial LD (Weir 1996), even when the two loci are completely unlinked. Suppose there are two populations with the same population size, 36, as shown in Table 2.3. We are studying if a certain disease (d for diseased and n for normal) is associated with a SNP with two alleles A and a. There is no association between marker and disease in both populations (the odds ratio (OR) equals 1). However, if we pool the data together, the pooled OR is 6.76. Thus, spurious association may present when there is undetected population structure.

<table>
<thead>
<tr>
<th></th>
<th>Pop1</th>
<th></th>
<th>Pop2</th>
<th></th>
<th>Pooled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d</td>
<td>n</td>
<td>d</td>
<td>n</td>
<td>d</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>5</td>
<td>25</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>a</td>
<td>5</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.3: Association for admixed population. The cell values are observed numbers of individuals in different categories. The odds ratios of population 1 and population 2 are both 1. But the odds ratio of the pooled data is 6.76.

Hence, when studying association between markers and diseases, we should always be aware of possible pseudo association due to pooling groups with different genetic backgrounds together.

2.4 The application of LD in Genetic Mapping

Linkage analysis is one of the most important methods in gene mapping (Ott 1999). For humans, recombination frequencies between two loci can be estimated from family data. By using proper mapping functions, linkage analysis has successfully lo-
cated many genes that are associated with Mendelian disorders such as hemophilia (McKusic 1969) and Huntington’s disease (Huntington’s Disease Collaborative Research Group 1993).

However, the pedigree data required for linkage analysis are expensive and often hard to collect, especially when identifying mutations related to late-onset diseases. Locating genes using linkage analysis is limited to about 1 cM (Risch and Merikangas 1996), which corresponds to approximately 1000kb, on average, since recombination events are rare between nearby loci. It is very hard to identify genes that contribute to complex traits, such as diabetes and cholesterol, using linkage analysis because complex diseases are usually caused by effects from variation in multiple genes and each gene often only has a small effect. Identifying such genes requires very large number of families and big pedigrees. It is believed that most complex traits result from interactions between genetic factors and environmental factors, which are not easily accommodated in linkage analysis models.

All of the limitations of traditional linkage analysis have led to an alternative approach for fine-scale gene mapping: association study using unrelated individuals. For continuous phenotypes, single marker association can be analyzed using a one-way ANOVA test which tests if all means are the same across different genotype categories. A $\chi^2$ test can be used for discrete traits. The association study can map suspected gene markers to a small region with length on the order of a few kb with higher power but lower cost than traditional linkage analysis (Risch and Merikangas
2.5 Haplotype Block Structure and Haplotype based Association Studies

The traditional association test using one SNP (or more generally, one marker) at a time has some weaknesses independent of the approach, linkage or association. With hundreds of markers, the multiple testing problem is not easy to overcome. Since SNPs close to each other are not independent, classical corrections cannot be applied or are too conservative. Though newer methods like false discovery rate (FDR), have been introduced and are used, the different levels of $LD$ in different regions are not accounted for because of using a uniform correction formula, thus wasting information. Several mutations in the same chromosome can interact to create a "super allele" which represents effects from several loci on a haplotype. Haplotype-SNP based association studies can simultaneously account for both the dependence and interaction effects among SNPs. In addition to these benefits, there is some evidence that haplotype-SNP based association studies can have higher statistical power than single SNPs. Several papers (e.g. Akey et al. 2001) have shown that under certain conditions, the haplotype-based approach is more powerful than the traditional single-SNP study. Yet, it may not be practical to consider a large number of SNPs together. In a region with $S$ SNPs, the expected number of haplotypes
will be $2^{s-1}$. Although large, this number might be reduced to only 3 to 4 common haplotypes if all SNPs in a region show strong $LD$.

Hence, we can divide a region with many SNPs into short pieces based on how strongly the SNPs are associated with each other. It is generally true that markers far away from each other show less $LD$, compared with markers relatively close together. However, crossovers are not uniformly distributed on human chromosomes. Some chromosomes have higher recombination rates than others. The X chromosome usually has a higher recombination rate. On a chromosome, the two ends (telomeres) tend to have higher recombination rates than the center (centromere). Even within a gene, there may be regions that have gone through more crossovers than other regions. Regions that have dramatically higher recombination rates are called recombination hot spots, while regions between two hotspots are called haplotype blocks. One can draw pairwise $LD$ measures (e.g. $|D'|$) in a heat map and code the high values with warm colors like red and low values with cold colors like blue; Figure 2.3 shows the values of $|D'|$ in two consecutive genes, ABCG5 and ABCG8, on chromosome 2 from more than 1000 African Americans from the Dallas Heart Study. It is clear that there are several red “squares”, which indicate that there are regions within which the pairwise $LD$s are all very high. However, it is not straightforward to specify block boundaries.

Several block partition algorithms based on different definitions have been suggested. In Chapter 3, we discuss some existing block definitions together with our
Figure 2.3: Heat map of pairwise LD ($|D'|$) for 126 physically ordered SNPs in the ABCG5 and ABCG8 region on chromosome 2. This sample includes more than 1000 African Americans from Dallas County. The SNPs are not placed at their relative bp locations but just order index.

A new approach based on a new summary statistic of pairwise linkage disequilibrium. The impact of marker density and tagging strategies on estimated block partitions have not been fully explored. They are studied in Chapter 4.

SNPs within haplotype blocks are expected to show strong LD. As a result, we expect limited haplotype variations in each haplotype block, which makes haplotype based association feasible. By applying both single-SNP based and haplotype-SNP based association studies, I will illustrate their advantages and disadvantages in Chapter 5.
Chapter 3

Haplotype Block Estimation

3.1 Two Main Block Partition Methods

The haplotype block phenomenon was first described by Daly et al. 2001. In the same year, another paper by Patil et al. 2001 described the same phenomenon. Since then, much effort has been put into studying the haplotype block structure because its great potential value for studying human evolutionary history and detecting disease variants for complex diseases. There have been numerous versions of block estimation methods. Most of them can be classified into one of two major types: haplotype diversity and pairwise LD.

Patil et al. 2001 focused on maintaining maximal information within a block by using only a few tagging SNPs. In their method, blocks are defined as physically consecutive SNPs such that at least 80% of the observed haplotypes are represented
more than once. A greedy search method was used by the authors. Later, Zhang et al. 2002 introduced a dynamic algorithm that can estimate the haplotype blocks and the corresponding tagging SNPs much more efficiently.

In this thesis, we will focus on the second major block partition method based on pairwise LD. The first paper to introduce the idea was Gabriel et al. 2001. Using the same idea but a novel bivariate summary statistic for $LD$, we provide a new partition method that provides better statistical power in detecting haplotype blocks.

### 3.2 A New Bivariate Summary Statistic of $LD$

In most situations, the sign of $D'$ is ignored. However, as pointed out by Thompson et al. 1988, larger sample size is needed in order to identify negative association than positive association. Motivated by this observation, we propose a new summary statistic $LDB = (PLD, minp)$. $PLD$ is the abbreviation for proportional linkage disequilibrium and $minp$ is the minimum of the four haplotype frequencies. They are defined by:

$$PLD = \max(p_{A_1B_1} + p_{A_2B_2}, p_{A_1B_2} + p_{A_2B_1})$$

$$minp = \min(p_{A_1B_1}, p_{A_2B_2}, p_{A_1B_2}, p_{A_2B_1}).$$

Without loss of generality, assume $p_{A_1B_1} + p_{A_2B_2} \geq p_{A_1B_2} + p_{A_2B_1}$, in which case $PLD$ equals $p_{A_1B_1} + p_{A_2B_2}$. By this definition, the value of $PLD$ is between 0.5 and 1. Suppose the observed association between the two loci $A$ and $B$ are positive
Figure 3.1: The dynamics of theoretical values of PLD with different recombination rates between two loci that are positively linked. Assume, at the beginning, there are only two haplotypes $A_1B_1$ and $A_2B_2$ with equal frequencies.

(defined in Section 2.3.1). With restriction that $0 < \alpha \leq p_{B_1} \leq 0.5$ and $\alpha > p_{A_1}p_{B_1}$ (positive associated), it is not hard to prove that both PLD and $r$ take their respective maxima $1 - p_{A_2} + p_{B_1}$ and $\sqrt{\frac{p_{B_1}(1-p_{A_1})}{p_{B_2}(1-p_{A_2})}}$. Their respective minima are $p_{A_1}p_{B_1} + (1-p_{A_1})(1-p_{B_1})$ and 0. The dynamic evolution of PLD is similar to other LD measures: PLD decreases with time and will reach the minimum value eventually, as shown by Figure 3.1.

The dynamic figures illustrate that the maximum proportion that two haplotypes can cover decreases from generation to generation. When the association is negative (defined in Section 2.3.1), a similar property can be found for $minp$: the value will
Figure 3.2: The dynamics of estimated $PLD$ with a fixed rescaled recombination rate of 0.1 but different population sizes when two loci are positively linked. The number of each haplotype at generation $t$ was simulated assuming random mating and no other evolutionary forces. Assume there are only two haplotypes $A_1B_1$ and $A_2B_2$ at generation 0 with equal proportion of 0.5. The curves of $PLD_t$ show the relation of $PLD_t$ with time (in generations) under different sample sizes. The smooth curve is the theoretical value.
increase until the equilibrium value is reached.

It is interesting that both $PLD$ and $r^2$ reach the maximum, 1, under the same conditions. While $r^2$ can reach 0 as long as two markers are independent, $PLD$ may not reach 0.5 even if the two markers are not associated. However, this does not prevent $PLD$ from being a useful statistic especially when we are interested in interpreting the strength of the association between two markers. We mentioned in Chapter 2 that one defect of $D'$ is that any value between 0 and 1 lacks a meaningful interpretation. Different from $D'$, this unscaled measure, $PLD$, gives the proportion of individuals that have two complimentary haplotypes by a pair of SNPs. Indeed, the statistical correlation $r$ is not a completely standardized measure either, although it has been rescaled by the four marginal allele frequencies. The following proof shows that if we rescale $PLD$ and $r^2$, the same formula will be obtained.

For simplicity, we use the following simple notations:

<table>
<thead>
<tr>
<th></th>
<th>$B_1$</th>
<th>$B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$\alpha$</td>
<td>$p_{A_1} - \alpha$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$p_{B_1} - \alpha$</td>
<td>$1 - p_{A_1} - p_{B_1} + \alpha$</td>
</tr>
<tr>
<td></td>
<td>$p_{B_1}$</td>
<td>$1 - p_{B_1}$</td>
</tr>
</tbody>
</table>

Table 3.1: Observed haplotype frequencies. Notice that there is only one undetermined variable $\alpha$ if the marginal allele frequencies are fixed.

<table>
<thead>
<tr>
<th></th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$p_{A_1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$p_{A_1}p_{B_1}$</td>
<td>$p_{A_1}(1 - p_{B_1})$</td>
<td>$p_{A_1}$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$p_{B_1}(1 - p_{A_1})$</td>
<td>$(1 - p_{A_1})(1 - p_{B_1})$</td>
<td>$1 - p_{A_1}$</td>
</tr>
<tr>
<td></td>
<td>$p_{B_1}$</td>
<td>$1 - p_{B_1}$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

Table 3.2: Expected haplotype frequencies when independence is assumed.
In this proof, we assume that the marginal allele frequencies are fixed for both loci. When two loci are positively associated, $PLD$ and $r$ reach their strongest association, $1-p_A + p_B$ and $\sqrt{\frac{p_B(1-p_A)}{p_A p_B (1-p_A)(1-p_B)}}$, respectively when $\alpha = p_B$. Their minimums $p_A p_B + (1-p_A)(1-p_B)$ and 0 are reached when $\alpha = p_A p_B$. The standardized versions are hence

\[
PLD' = \frac{(\alpha + 1 - p_A - p_B + \alpha) - (p_A p_B + (1-p_A)(1-p_B))}{(p_B + 1 - p_A) - (p_A p_B + (1-p_A)(1-p_B))}
\]

\[
= \frac{\alpha - p_A p_B}{p_B(1-p_A)}
\]

\[
= D'
\]

\[
r' = \frac{(\alpha - p_A p_B) / \sqrt{p_A p_B (1-p_A)(1-p_B)}}{(p_B - p_A p_B) / \sqrt{p_A p_B (1-p_A)(1-p_B)}} - 0
\]

\[
= D'
\]

Similarly, the standardized $PLD$ and $r$ can also be proved to be equivalent to $D'$ when the two loci are negatively associated.

### 3.3 Gabriel’s Approach

The pairwise-based approach proposed by Gabriel et al. 2002 partitions a sequence into physically disjoint regions (also called blocks) such that the pairwise $LD$s within a same region are high while $LD$s between SNPs within different blocks are very low. In their approach, each pair of SNPs is classified into one of three different categories
based on confidence bounds of $D'$. A pair of SNPs is defined as being in “strong LD” if the upper 95% confidence bound is greater than 0.98 and the lower 95% confidence bound is greater than 0.7, i.e.,

$$\begin{align*}
\text{strong LD} & \quad \text{if } D'_u > 0.98 \text{ and } D'_l > 0.7 \\
\text{weak LD} & \quad \text{if } D'_u < 0.9
\end{align*}$$

(3.1)

A pair of SNPs is said to have “strong evidence for historical recombination” if the upper 95% bound is less than 0.9. A contiguous region of SNPs is defined as a **haplotype block** if

$$\frac{\#\text{strong LD}}{\#\text{strong LD} + \#\text{weak LD}} > 95\%$$

The remaining 5% may include pairs that are classified as showing strong evidence for historical recombination yet actually were caused by evolutionary forces other than recombination. This could be due to selection, genetic drift, mutation, or gene conversion. After classifying all of the pairs on a given chromosome, the process can start from the left-most SNP and one finds the longest haplotype block satisfying the haplotype block criterion. Then, from the first SNP that follows the previously identified block, the process starts again and continues until the entire sequence of SNPs has been analyzed. Notice that the process can also be started from any SNP in the sequence. The authors found that there is not much difference. In this dissertation, we always start the process from the left end. The reason that the
confidence bound is used instead of the point estimate is that the point estimate is unreliable when either the sample size or the marginal allele frequencies are small.

Note that, aside from the two defined categories, there is a third category: pairs that do not belong to either of these two classes. We label those pairs as “unknown LD”, which means that there is not enough evidence to determine if these pairs are in “strong LD” or “weak LD”. The pairs in the unknown class are not used in the block partition algorithm in order to reduce noises from the “unknown LD”.

The confidence boundaries of $D'$ are very wide in many situations: even when the lower bound is close to 0, both the point estimation and the upper bound can be near the maximum value of 1. Hence, many pairs are classified into the unknown class, which may bias the results. The new measure, $PLD$, has confidence intervals that tend to be much narrower. Figure 3.3 shows point estimates and confidence bounds of $D'$, $r^2$, and $PLD$ based on the left-most SNP versus the next consecutive 50 SNPs in a region of chromosome 20 based on 48 unrelated individuals from Gabriel et al. 2002. Both the point estimates and the confidence bounds were calculated using maximum likelihood method. We also used resampling techniques (resample 48 individuals with replication) and obtained similar results. Figure 3.3 shows that the confidence interval of $PLD$ is much smaller than those of other two measures.
Figure 3.3: The 95% upper and 5% low confidence bounds of $D'$, PLD, and $r^2$. The x-axis indicates the distances (in bp) between the left-most SNPs versus the next consecutive 50 SNPs.
3.4 A New Approach to Block Estimation

Using the same pairwise LD approach as Gabriel et al. 2002, we used $LDB = c(PLD, minp)$ instead of $D'$ as a way to judge the degree of LD between two SNPs. We say that a pair of loci is in “strong LD” if either $PLD > 0.95$ or $minp < 0.015$; in “strong evidence of historical recombination” if $PLD < 0.9$ and $minp > 0.03$. This definition more completely captures the expects of LD for the specified purpose of haplotype block partition: minimize the number of haplotypes within a haplotype block. $PLD$ and $minp$ complement each other in identifying pair of SNPs that are in LD. Because of smaller variance, about 30% of less pairs were classified to “unknown LD” when using the summary statistic than Gabriel’s approach.

3.4.1 Comparisons

When comparing results from different methods, we use the three criteria from Wall and Pritchard 2003. Wall and Pritchard suggested three criteria to test if a haplotype block model is reasonable for a given segment of DNA: (1) coverage, (2) overlapping, and (3) holes. Coverage is the proportion of a region that is covered by haplotype blocks. We say that two blocks within the same partition overlap if at least one locus is shared between them. If two loci within a block are in strong LD and a SNP in between is not in strong LD with both of them, we say there is a hole in the block. Ideally, there should be no overlapping or holes in estimated haplotype blocks. In practice, the coverages of two approaches are not comparable since we do
not know the true coverage of the real data. However, simulated data allow coverage comparison. Following other authors studying block partitioning, we restrict attention to SNPs with minor allele frequency (MAF) > 0.1 in order to avoid small sample size problems.

3.4.2 Real Data

We first applied both Gabriel’s method and our new approach to data used in Gabriel et al. 2002. There are four populations: CEPH, African American, East Asians and Nigerians. When there were trios or families, only unrelated individuals were used. The percentages of overlap and holes were similar using both methods. Using Gabriel’s approach the mean coverage of 50 regions for the four populations is 46%, 24%, 41%, and 25%, respectively. The coverage increases to 58%, 41%, 54%, and 45%, which brings relative increase of 26%, 71%, 32%, and 80%, respectively, using the new approach.

3.4.3 Simulated Data

To more properly evaluate Gabriel’s method and our new method, we applied the two approaches to simulated data. In the appendix of Li and Stephens 2003, the authors proposed a coalescent based approach to simulate sequence data with a single recombination hotspot. Assume a sequence with $S$ segregating sites is aligned in a rescaled region with a physical length of 1. The background recombination rate is $\rho$. 
Also assume a recombination hotspot has recombination rate \( \lambda \rho \) (\( \lambda > 1 \)) which starts at position \( a \) and ends at position \( b \). The sequences with one single recombination hotspot can be generated using the following steps:

- Simulate \( S' = S(1+w(\lambda - 1)) \) sites with recombination rate \( \rho' = \rho(1+w(\lambda - 1)) \) using Hudson's coalescence based algorithm. Here \( w = b - a \).

- Multiply each position by \( (1+w(\lambda - 1)) \) so that the total length is enlarged from 1 to \( (1+w(\lambda - 1)) \) and the recombination rate is decreased to \( \rho \).

- With probability \( 1 - \frac{1}{\lambda} \), randomly delete sites within \( a \) and \( a + w\lambda \).

- For the sites left between \( a \) and \( a + w\lambda \), shrink the adjacent distances by factor \( \lambda \).

- Relabel the positions of sites on the right side of \( a \) while keeping the distance between adjacent sites the same.

After all these steps, the physical length returns to 1. The mutation rate has not changed since parts of sites are randomly deleted. The shrinkage of distance in the designed region has elevated the recombination rate in the recombination hotspot from \( \rho \) to \( \lambda \rho \).

We followed the parameters used by Wall and Pritchard 2003, which were chosen to match empirical data. We assume a constant population size, \( N = 10000 \), and a rescaled mutation rate of \( \theta = 7.836 \times 10^{-5}/bp \). Based on these parameters,
we used two block models (model A and model B). Model A assumed the block length has an exponential distribution with mean 30kb, separated by three recombination hot spots each with length 1kb. The simulated data from Wall and Pritchard 2003 show that this model matches the empirical data rather well. To test the performance of different partition algorithms in different scenarios, we then simulated data using a block model that is quite different from the estimated block structure from empirical data. In model B, we assumed there were two long recombination hotspots of length 10kb and one short hotspot of length 1kb, although 10kb may be too large for recombination hotspots. For each haplotype block model, we ran 50 data sets each sampling 200 chromosomes. The 200 chromosomes in each data set were then randomly paired to form 100 individuals. To mimic population genotype data we ignored phase information and used an EM algorithm (Excoffier and Slatkin 1995) to infer the phase information as is typical in practice. Besides the three criteria from Wall and Pritchard 2003, we also considered type I and type II errors since the true block structure of the simulated data is known. The type I error is defined as the probability of classifying a SNP in a recombination hotspot as a SNP in a haplotype block; and the type II error is the probability of being unable to identify SNPs that are in haplotype blocks. The cell values in Table 3.3 are means over 50 runs.

Table 3.3 shows that there might be some but not dramatic improvement in
Figure 3.4: Estimated haplotype blocks from three simulated data sets using model A. Hashes indicate SNP locations. See text and the caption of Table 3.3 for details.
Figure 3.5: Estimated haplotype blocks from three simulated data sets using model B. Hashes indicate SNP locations. See text and the caption of Table 3.3 for details.
<table>
<thead>
<tr>
<th>Simulation</th>
<th>Coverage</th>
<th>Holes</th>
<th>Overlap</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabriel et al.</td>
<td>33.3%</td>
<td>23.5</td>
<td>17.3</td>
<td>0.47</td>
<td>21.77</td>
</tr>
<tr>
<td>Yu and Guerra</td>
<td>50.0</td>
<td>16.4</td>
<td>16.3</td>
<td>1.83</td>
<td>6.78</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabriel et al.</td>
<td>24.6%</td>
<td>25.9</td>
<td>15.3</td>
<td>0.57</td>
<td>22.03</td>
</tr>
<tr>
<td>Yu and Guerra</td>
<td>41.8</td>
<td>22.6</td>
<td>19.1</td>
<td>6.37</td>
<td>7.86</td>
</tr>
</tbody>
</table>

Table 3.3: Results of block estimation based on simulated data sets. In the true model, a 100kb region is separated by three recombination hot spots. Cell values are means over 50 runs, each with 200 chromosomes. Only SNPs with observed MAF>0.1 were used. (A) Hot spot lengths (1kb, 1kb, and 1kb); see Figure 3.4. (B) Hot spot lengths (10kb, 1kb, and 10kb); see Figure 3.5.

holes, no improvement in overlappings for both haplotype block models; similar type I error in model A and higher type I error in model B. However, our new approach has much higher coverage and power to identify haplotype blocks for both block models. This advantage of our new approach is also illustrated in Figure 3.4 and Figure 3.5. Our new approach identifies more and longer blocks, resulting haplotype block partition that is more close to the assumed model.
Chapter 4

Effect of Marker Density on Haplotype Block Structure

In Chapter 3 we proposed a new block partition method and compared it with a popular approach. In that chapter, haplotype block estimation was based on all SNPs with $MAF > 0.1$. The haplotype block structure has been studied by many other groups because of its potential application in both genetic mapping and human evolutionary history. However, there has not been a systematic study on the impact of marker density and marker selection methods on the estimation of haplotype blocks. Different marker selection strategies have been discussed in the context of reducing genotyping expenses. In this chapter we consider some of these maker selection strategies in the context of haplotype block estimation.
The data from the Dallas Heart Study (DHS) contains genes in which the entire DNA sequence has been obtained and all SNPs identified, which allows for a comprehensive study of the effect of marker density and marker selection strategies on haplotype block estimation. In this chapter, using this data set, we will show the effect of marker density and different tagging strategies on haplotype block estimation. Haplotype block structure may be lost even when using tagging SNPs that purportedly keep a large portion of "information" (Johnson et al. 2001; Carlson et al. 2003).

4.1 Background

Haplotype block structure characterizes the high correlation among markers that are physically close on a chromosome. It is with the hope that blocks can better help us understand human evolutionary history and genetic mapping that more and more researchers have accepted this concept, although more efforts need to be put toward the biological meaning and usefulness. Haplotype blocks are usually estimated from marker genotype data based on user-defined models, which have a similar purpose but different results.

Several millions of SNPs with minor allele frequency bigger than 10% have been estimated to exist in the human genome (Kruglyak and Nickerson 2001; Carlson et al. 2003). To reduce the genotyping effort, different selection methods such
as random selection, equally spaced selection, and statistical tagging have been considered. Several tagging strategies were developed to keep a certain amount of information while reducing the number of SNPs needed to be genotyped. There are two major types of statistical tagging methods. The first type focuses on identifying a minimum number of SNPs that can capture the most common haplotypes (Johnson et al. 2001) and thus reflect a large proportion of haplotype diversity (PDE, Clayton 2002). The idea was also introduced by Patil et al. 2001 and then further developed by Zhang et al. 2004. This method is bound with the concept of haplotype blocks. Halld’orsson et al. 2004 gave a block-free tagging algorithm whose purpose is also to select a minimum number of SNPs while maintaining a certain amount of haplotype diversity. The second type of statistical tagging approach is based on pairwise LD $(r)$ (Carlson et al. 2003). While both methods try to find a minimum number of tagging SNPs, the latter one tries to pick up the minimum number tagging SNPs that all the SNPs are either directly assayed or exceed some threshold with the assayed SNPs.

Though the proportion of SNPs that are identified as tagging SNPs varies by method, population, and region, studies indicate that more SNPs are needed than previously thought in order to keep certain level of power for association testing (Carlson et al. 2003; Wang et al. 2003; Sun et al. 2004).

In this thesis we consider these issues of marker density and marker selection strategies in the context of pairwise LD approaches to haplotype block estima-
tion. Indeed, the computation involved in estimating haplotype blocks with pairwise LD is relatively small when haplotype phase information is missing. In this chapter, we will report on the impact of marker density and various tagging SNP strategies on the estimation of haplotype block structure using Gabriel's method since these have been used by many researchers (for example, Malik et al. 2005; Florez et al. 2005; and Pearce et al. 2005).

4.2 Comparing Partitions

In order to compare two different estimated partitions it is necessary to introduce both measures and tests that can describe how similar two partitions are.

4.2.1 Similarity Between Two Partitions

Schwartz et al. 2003 addressed how to test if two partitions are the same. They proposed the number of shared boundaries as the test statistic to measure the concordance of two different partitions. Suppose $B_1$ and $B_2$ are the number of boundaries in two partitions and $S$ is the number of SNPs. The probability that two partitions share exactly $m$ boundaries under the null hypothesis that the number of shared boundaries have occurred by chance among $S$ SNPs can
be calculated as

\[ P(M = m) = \frac{\binom{B_1}{m} \binom{S-1-B_1}{B_2-m}}{\binom{S-1}{B_2}} \]

since \( M \) follows a hypergeometric distribution. The corresponding \( P \) value is

\[ P-value = P(M \geq m) = \sum_{i=m}^{\min(B_1,B_2)} \frac{\binom{B_1}{i} \binom{S-1-B_1}{B_2-i}}{\binom{S-1}{B_2}} \]

But there are several shortcomings that can be remedied. First, they did not provide a measure of similarity. Second, their test statistic is based on the number of exact shared boundaries, which is too strict for comparing the similarity between different haplotype block estimations. For example, Figure 4.1 shows three haplotype block partitions on the same area. Using Schwartz's method, partition \( P_1 \) and \( P_2 \) share 0 boundaries; partition \( P_1 \) and \( P_3 \) also share 0 boundaries. Hence, by Schwartz's test, both pairs disagree at the same level (the same \( P \)-value). However, the figure shows that the blocks estimated from \( P_1 \) and \( P_2 \) are much more similar than the blocks estimated from \( P_1 \) and \( P_3 \).

Liu et al. 2004 introduced a projection-based measure that uses the physical lengths of estimated haplotype blocks as weights when evaluating the disagreement between two partitions. Using this method, a good match does not require the boundaries are exactly the same. However, the measure is not symmetric.
Figure 4.1: Three haplotype block partitions of the same area. This figure shows Schwartz's 2003 procedure is too strict. Liu et al. 2004 used a similar graph. The black bars and red boxes indicate positions that SNPs were genotyped and estimated blocks, respectively.

The measure calculated based projecting partition $P_1$ to $P_2$ usually is not the same to that based on projecting partition $P_2$ to $P_1$. In this chapter, we introduce two measures based on the percentage of matching. In the first measure, we label all SNPs that are estimated in blocks (including both the boundaries and interior SNPs) as 1 and those SNPs out of blocks (recombination hot spots) as 0. For example, partition $P_1$ in Figure 4.2 contains two blocks (the green bars) and two (single) recombination hotspots (red bars). We label the SNPs as 1111100111. Similarly, $P_2$ is labeled as 0000001111. A natural measure of similarity based on the labeling is how much they agree, defined as 1 minus the Hamming distance. Hence the similarity between the two partitions is $1 - 6/10 = 40%$. 
Figure 4.2: Similarity measure based on Hamming distance. The similarity between $P_1$ and $P_2$ is 40%.

Figure 4.3: Similarity measure based on score matrix.
Although the Hamming similarity measure does take the number of SNPs into account, it doesn't differentiate between SNPs at boundaries and interior SNPs. Borrowing some ideas from DNA sequence alignment (Durbin et al. 1998), we generalized the Hamming similarity measure by applying a score matrix. For a particular partition, each SNP is classified into one of four categories: interior, left boundary, right boundary and non-block. The categories define a weight/score matrix; for example, see Figure 4.3. The actual weight values can be re-defined according to user's interests. In this analysis, we assign the highest score of 1 to perfect matches within each category (main diagonal of Figure 4.3). A score of 1 is also given if interior SNPs in one partition match with boundary SNPs of the other partition. When a left boundary in one partition is the right boundary of another partition, we give the lowest score of 0. Asides from matching scores, we apply a penalty if in a given region one partition shows a single block, while the other partition shows multiple blocks. In Figure 4.3 this is indicated by and "X". When this happens, we say there is at least one gap that separating the partitions, and 1 point is deducted from the similarity calculation. To calculate the similarity score, add up the scores from all the comparisons and apply penalties. The matching score (or similarity measure) is defined as the sum divided by the total number of SNPs, S. In Figure 4.3 we give an example. These are $S = 12$ SNPs. The scores of weights are $(1,1,1,1,0,1,1,1,1,1,1,1)$, which gives a sum of 11. Since there is one
gap between the two partitions, the final weight score is $\frac{11-1}{12} = 83.3\%$. This
definition of similarity also allows user-adaptation via changing the score matrix
if there is any preference or any prior information about the relative importance
among different matches. In Section 4.2.2 we discuss statistical significance of
the proposed measure.

Compared with the Hamming score, the weight score based on score matrix is
more reasonable. We will this definition in the left part of this section.

4.2.2 Hypothesis Tests for Comparing Two Block Partitions

In Schwartz et al. 2003, the null hypothesis is that the shared boundaries
have occurred by chance. Assume that in a region with $S$ SNPs, there are $B_1$
boundaries in partition $P_1$ and $B_2$ boundaries in partition $P_2$. For fixed $B_1$ and
$B_2$, Schwartz et al. 2003 show that under the null hypothesis the number ($M$) of
shared boundaries follows a hypergeometric distribution. Therefore, $P$-value can
be calculated and evaluated for significance. Liu et al. 2004 examined statistical
significance based on a null distribution obtained through simulations based on
regenerated pairs of partitions with $B_1$ and $B_2$ boundaries randomly selected at
SNP loci. Both Schwartz et al. 2003 and Liu et al. 2004 assume $B_1$ and $B_2$ are
fixed. However, the number of boundaries is part of the unknown information in
the inferential process, hence, they cannot be assumed to be constants. Also, the
number of boundaries is not the most important parameter in the estimation of a block partition. Two partitions can have exactly the same number of boundaries but very different structures. Hence, we relax these constraints and allow the number of boundaries to follow a Poisson distribution with the observed number of boundaries as the mean. A simulation based hypothesis test for two partitions is described below:

1. Calculate the weight score of estimated partitions $P_1$ and $P_2$ and denote it by $score_0$.

2. For $i = 1$ to $i = B$.

   (a) Generate a random number $m_{1i}$ from a Poisson distribution with mean $B_1$. Randomly pick $m_{1i}$ boundaries and denote the corresponding partition as $P'_{1i}$;

   (b) Similarly, generate a random number $m_{2i}$ first from a Poisson distribution with mean $B_2$ and then generate a random partition $P'_{2i}$ with $m_{2i}$ boundaries;

   (c) Set $s_i$ to be the similarity measure between $P'_{1i}$ and $P'_{2i}$;

3. The $P$ values can be calculated as the percent of similarity measures in $score$ that exceed $score_0$.

We simulated matching $score_3$ using 50 SNPs and different $B_1$ and $B_2$. For each pair of $(B_1, B_2)$, 5000 matching scores were generated using the procedure.
Figure 4.4: The estimated null distribution of similarity scores under the null hypothesis that the two partitions share boundaries by chance.

above in order to approximate the null distribution. The histograms of the matching scores were drawn in Figure 4.4.

When $S = 50$ and $B_1 = B_2$, the center of the simulated scores shifts to the left as the number of boundaries $B_1$ and $B_2$ increase from 5 to 15, as shown in Figure 4.4. The lower right subplot also describes the approximated null distribution when $B_1 = 15$ and $B_2 = 10$, which is left skewed and centered at around 0.68 (the median). When the simulated P-value is less than 0.05, we say that the similarity between two partitions are not due to chance. In model A of Section 3.4.3, 50 data sets were generated. Using this simulation procedure and 0.05 as the threshold to test if the estimated partitions for simulated data sets are the same as the assumed model, 45 were rejected, i.e., 45 comparisons showed the similarity is NOT due to chance. In this situation, the power to
detect similarity not due to chance is $\frac{45}{50} = 0.9$.

4.3 Data

To access the impact of different tagging strategies on block partition, we used the genotype data from the Dallas Heart Study (DHS) conducted by University of Texas Southwestern Medical Center. More than 3000 individuals were sampled from Dallas County. After excluding those who were on treatment or missed most of the marker information, 3126 individuals were included in the final analysis. Out of the 3126 individuals, there were 1638 African Americans (Black), 927 Caucasian Americans (White) and 561 Hispanics. Two genetic loci were considered in this study: ABCG5+8 and APOAV. Although we will not use the data from CETP gene and the Hispanic sample in this chapter, due to the low marker density in CETP and the possible population admixture in the Hispanics for the sake of completeness, we will still describe the data here since they will be studied in next chapter.

Region ABCG5+8 is a 96kb locus that contains two genes, ABCG5 and ABCG8, which are in a head-to-head orientation in chromosome 2. 184 SNPs were genotyped over this 96kb region. Removing one SNP that was extremely far away from the others and another 13 SNPs that violated Hardy Weinberg Equilibrium, we analyzed 170 SNPs remained, which is about one marker every
565 base pairs. Different ethnic groups have different numbers of informative (MAF > 10%) SNPs. In the DHS sample, Blacks, Hispanics and Whites had 124, 126, and 113 informative SNPs, respectively, at the ABCG5+8 region. The corresponding marker densities are one SNP every 773bp, 761bp, and 849bp on average. The mean and median MAF in the Blacks are 0.236 and 0.257, respectively; the mean and the median MAF of the Hispanics are 0.258 and 0.308, respectively. The Whites have mean 0.207 and median 0.192, which are much lower than the other two groups.

Gene cluster APOAI/CIII/AIV is located on 11q23 and covers 67.5kb. 109 SNPs were genotyped in this region, which is about one SNP every 620bp. The Blacks, Hispanics, and Whites have 61, 59, and 42 informative SNPs, respectively. The corresponding marker densities are one SNP every 1107bp, 1145bp, and 1608bp on average. Similar to the region ABCG5+8, the Hispanics have the highest MAF in APOAV region, with a mean of 0.162 and median of 0.125; the blacks have a mean of 0.16 and median of 0.113; while the MAF of the whites have a mean of 0.124 and median of 0.066, which are much lower than the other two groups.

The cholesteryl ester transfer protein (CETP) gene locus is 40kb long and located on 16q21. CETP catalyzes the transfer of cholesteryl ester from HDL to VLDL and LDL. It plays a key role in HDL metabolism. 52 SNPs were genotyped at the CETP locus, and the marker density is about one SNP every
773 bp on average. The Blacks, Hispanics, and Whites have 33, 31 and 35 informative SNPs, respectively. The corresponding marker densities are one SNP every 1217bp, 1296bp and 1148bp. The whites have mean and median MAF 0.213 and 0.200; the Hispanics have mean and median MAF 0.209 and 0.175; while the blacks have mean and median MAF 0.189 and 0.150. Different from the other two regions, the Whites have both higher marker density and MAF as compared to the other two groups.

Through sequencing, all SNPs were identified at each locus. Therefore, we treated the block structure estimated from all informative SNPs and all subjects as the “true” block structure. To investigate the impact of sample size, marker density and selection we compared block partition estimated from various subjects and marker subsets to the “true” partition as described above.

4.4 Choices for Marker Selection

Genotyping is expensive and time consuming. Investigators would like to have their association studies on a smaller subset of SNPs while still keep a higher power in association studies. Different marker selection strategies have been proposed (Johnson et al. 2001; Zhang et al. 2002; Carlson et al. 2003; Chapman et al. 2003; Meng et al. 2003; Halld’orsson et al. 2004). We used four different marker selection methods: random selection, equally spaced selection, block-free
tagging, and block-based tagging. In random marker selection, we randomly selected 10, 20, 40, 60, and 80 percent of the markers from the original set of markers. In equally spaced marker selection, we selected one marker every 0.5kb, 1kb, 2kb, and 4kb. If there was more than one marker in each interval, we selected the first one. The block-free SNP tagging method is based on Carlson’s method (Carlson et al. 2003). This method identifies minimum number of SNPs that all known common SNPs either are directly assayed or exceed a threshold of $r$. $r$ was chosen at values of 0.6, 0.7, 0.8, and 0.9. The fourth selection method we considered is the block based method. In this method, markers were selected using Carlson’s method in each “true” block. The methods were compared at sample sizes: 25, 50, 100, and 400.

4.5 Results

In this part, we first present the impacts of sample size and the thresholds for marker selection (percent of SNPs used in random selection; space length in equally spaced selection; $r$ for block-free and block-based tagging) for each marker selection method. We then compare the efficiencies of keeping block structure of the four selection methods for fixed sample sizes.
4.5.1 Random Selection

For random selection, we estimated the block structure based on 10%, 20%, 40%, 60%, and 80% of the SNPs and 25, 50, 100, and 400 individuals randomly selected from the original data set. All 20 combinations were repeated 50 times. In each replication, both the markers and the individuals were selected randomly from the full data set. The mean similarity scores between the estimated block structure from the selected SNPs and the block structure based on the full set are reported in graphical form (Figure 4.5). In random selection, we found both the percent of SNPs used and the sample size affect the block partition, however, the effect of percent of SNPs used (equivalent to marker density in a given region) is much higher than that of sample size since when sample size is larger than 50, the effect of sample size is not very dramatic. Hence, sample sizes of 50 or larger are recommended. Figure 4.5 shows that for a given percent of SNPs used the similarity scores of Whites are higher than that of Blacks. When using 100 samples, in order to attain 80% matching, 80% of SNPs are needed for African Americans but only 60% of SNPs are required for Whites in both the ABCG5+8 region and the APOAV region. Hence larger percent of SNPs is required for the black group. This result is consistent with the well known fact that there is generally lower LD in African Americans than in Caucasian Americans due to African Americans being an older population.

In Section 4.2.2 we described a procedure to test for similarity between two
Figure 4.5: Similarity scores between estimated blocks based on 10, 20, 40, 60, and 80% of SNPs selected randomly and the blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red).
Figure 4.6: P-values for similarity scores between estimated blocks based on 10, 20, 40, 60, and 80% of SNPs selected randomly and blocks estimated using the full set. Four different sample sizes were chosen: 25 (orange), 50 (green), 100 (black), and 400 (red).
partitions. The P-values for similarities between blocks estimated from subsets of SNPs and blocks estimated using the full set are presented in Figure 4.6. To enlarge the difference of P-values, we rescaled the P-values to $-\log_{10}(P)$. In the figure, P-values are truncated to 0.00001. When a large proportion of SNPs are chosen, the agreement is so significant that the computer will report the P-value to be 0. To avoid $\log_{10}(0)$, we added 0.00001 to every P-value. Hence, $p = 0.05$ is equivalent to $y = 1.3$ and $p = 0.01$ is equivalent to $y = 2$. Figure 4.6 shows that the similarity scores start to show significance when at least 40% SNPs were used in the region ABCG5+8 if the sample size is not too small (bigger than 25). But 60% of SNPs are needed in order to show strong matches in the APOAV region. This agrees with the trend of scores in Figure 4.5. For each fixed sample size and percentage, the P-values in the White group is larger, which also agrees with larger of percent of SNPs being required for African Americans as compared to Caucasian Americans.

4.5.2 Equally Spaced Selection

In addition to selecting a certain percentage of SNPs randomly, we considered equally spaced selection. We first divide a given region into segments of equal length and then the first SNP of each segment is selected within each segment. In this study, we tried four different space lengths: 0.5kb, 1kb, 2kb, and 4kb in the ABCG5+8 region (96kb) and the APOAV region (67.5kb). The similarity
score based on different spacing intervals and sample size are plotted in Figure 4.7. The trends and patterns of matching scores using equally space selection are very similar to that of random selection except high matching scores are obtained in Whites even when SNPs were sampled at a density of one marker every 4kb. The P-values in Figure 4.8 gave similar conclusions about the effect of space lengths and sample size.

### 4.5.3 Block-Free Tagging

We used Carlson’s method to select tagging SNPs from the original marker list. To study the impact of different thresholds, we chose the statistical correlation (defined in Section 2.3.1) $r = 0.6, 0.7, 0.8, \text{ and } 0.9$. Both the matching scores in Figure 4.9 and the simulated P-values in Figure 4.10 reveal that higher similarity was obtained if either the sample size or the threshold of $r$ is increased. Different from random selection and equally spaced selection, it seems that for each fixed $r$, region APOAV has higher similarity than region ABCG5+8 for Blacks. This is because a higher percentage of SNPs was tagged in the APOAV region. The number of tagged SNPs increases with the increase in $r$. For Blacks, 42 (33%), 65 (52%), 66 (52%), and 76 (60%) SNPs out of 126 SNPs in region ABCG5+8 were tagged but 38 (62%), 55 (90%), 61 (100%), and 61 (100%) SNPs out of 61 SNPs were tagged in the APOAV region.
Figure 4.7: Similarity scores between estimated blocks based on equally spaced SNPs (500bp, 1000bp, 2000bp, and 4000bp) and blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red).
Figure 4.8: P-values for similarity scores between estimated blocks based on Equally space SNPs (500bp, 1000bp, 2000bp, and 4000bp) and blocks estimated using the full set. Four different sample sizes were chosen: 25(orange), 50(green), 100(black), and 400(red).
Figure 4.9: Similarity scores for estimated blocks based on SNPs selected using Carlson’s method. The x-axis (r) indicates the levels of statistical correlation when using Carlson’s method to choose tagging SNPs.
Figure 4.10: P-values of similarity scores for estimated blocks based on SNPs selected using Carlson’s method. The x-axis (r) indicates the levels of statistical correlation when using Carlson’s method to choose tagging SNPs.
4.5.4 Block-based Tagging

The last strategy we tried was block-based tagging. In block-based SNP selection, we first use all the SNPs and individuals in the APOAV region from the Caucasian Americans to estimate the blocks. Of course, in reality this does not make sense since we usually have a smaller testing data set before we obtain a data set with larger sample size. Here we just want to know if selected SNPs based on the known block structure can keep the block structure. In each estimated block, we use Carlson’s tagging method with criteria $r = 0.6, 0.7, 0.8,$ and 0.9. The singletons (SNPs that are not estimated into any block) enter the tagging list automatically. When $r = 0.6$, 19 out of 42 SNPs were tagged. Compared with 11 tagged SNPs using block-free tagging at the same threshold, the block-based based tagging is less efficient in terms of keeping statistical power in association studies, as discussed by Halldórsson et al. 2004. Similar with the other three marker selecting strategies, the matching scores and the similarities increase with an increase in the number of SNPs tagged (Figure 4.11 and Figure 4.12). A comparison between the block-free approach and the block-based approach is discussed in Section 4.5.5.

4.5.5 Comparison of the Four Marker Selection Methods

We studied the effect of sample size and the thresholds for each marker selection in Sections 4.5.1, 4.5.2, 4.5.3, and 4.5.3 on the accuracy of block partitioning.
Figure 4.11: Similarity scores for estimated blocks using SNPs selected from block-based tagging method. Haplotype blocks were first estimated using all the samples and markers. Then, Carlson's method was used to select markers within each haplotype block. The x-axis is the same as that of Figure 4.9.
Figure 4.12: P-values for the similarity scores using block-based tagging method. The x-axis is the same as that of Figure 4.9.
In order to study which marker selecting strategy is more efficient, we plotted the matching score versus the percentage of markers used instead of different threshold values for the different strategies. For example, we plotted the matching scores versus percentage of markers for African Americans in ABCG5+8 in Figure 4.13, in each subplot the sample size is fixed. Matching scores versus percentage of SNPs for different strategies were plotted using different colors. Due to the large number of selected SNPs using block-free tagging methods for the blacks in APOAV region, it is not obvious which is the best. But all the other three figures (Figures 4.13, 4.14, 4.16) show that the performances of random selection and equally spaced selection are better than that of block-free tagging since higher matching scores were obtained at similar percent of SNPs at each fixed sample size. The data from Caucasian Americans in both ABCG5+8 region and APOAV region illustrate that the performance of equally spaced selection is superior to that of random selection, while the data from African Americans show a similar performance between two methods. Figure 4.16 also shows that block-based tagging is not more efficient than block-free tagging in terms of keeping haplotype block structure.
Figure 4.13: Comparing the performance of random selection, equally-spaced selection and Carlson’s tagging in region ABCG5+8 for the African Americans.
Figure 4.14: Comparing the performance of random selection, equally-spaced selection and Carlson's tagging in region ABCG5+8 for the Caucasian Americans.
Figure 4.15: Comparing the performance of random selection, equally-spaced selection and Carlson's tagging in region APOAV for the African Americans.
Figure 4.16: Comparing the performance of random selection, equally-spaced selection, Carlson's tagging and block-based tagging in region APOAV for the Caucasian Americans.
4.6 Conclusion

In this chapter, we studied three factors that might affect the haplotype block estimation. Sample size clearly has an effect; however, as long as the sample size is not too small, the effect is limited compared with percent of SNPs used. Usually sample sizes of 50 or larger are recommended since most of the figures show that the similarity scores become stable when sample size is larger than 50. The similarity score increases with the increase of percent of SNPs used in random selection, the decrease of the space length in equally spaced selection, and the increase of $r$ in block-free and block-based tagging. No matter what strategy is used, the larger the percent of SNPs used, which is equivalent to the higher the marker density in a given region, the higher the matching scores. Equally spaced tagging might be generally better than random selection (Figure 4.14 and Figure 4.16) but this studies cannot conclusively determine their relative performances (Figure 4.13 and Figure 4.15). Indeed, makers are not evenly distributed across the chromosomes and so equally spaced sampling may not be obtainable for arbitrary spacing. Both methods appear much better than the other two statistical tagging methods, which were designed to maintain observed haplotype variation with a minimum number of SNPs. Tagging SNPs have also been advocated for association studies. Halld'orsson et al. 2004 claimed that block-based tagging is less efficient than block-free tagging in maintaining statistical power for association studies. In this study, we found that
the block-based tagging is worse than block-free tagging in keeping haplotype block structure.
Chapter 5

Single vs. Haplotype SNP

Association Studies: A Case Study

In the last two chapters, we focused on haplotype block estimation and the factors that affect estimation. The estimated haplotype blocks can partition long regions into short segments, which may make haplotype-SNP association more feasible than single-SNP association studies. In this chapter, we investigate and compare single-SNP association and haplotype-SNP association to identify possible mutations in real data collected by the Dallas Heart Study group. The haplotype based approach is compared with the traditional single-SNP approach in six race-sex groups. The comparisons show that single-SNP
significance can be identified in most cases by a haplotype-SNP based study and thus the haplotype-SNP based study might have higher power. However, one must show caution when a large number of SNPs are considered simultaneously even when they are in a haplotype block.

5.1 Background of Cardiovascular Diseases and their Risk Factors

Multiple metabolic risk factors contribute to the leading cause of death of the United States, namely, cardiovascular diseases. Although many environmental factors are involved, such as diet and exercise, many cardiovascular diseases show familial aggregation, which motivates researchers to look for genetic causes of these abnormalities. Previous studies have reported associations between the ABCG5+8 locus and plant sterols, triglyceride and the gene cluster APOAV, and HDL and gene CETP. High triglyceride and low HDL are two very important metabolic risk factors affecting the development of cardiovascular diseases. In a sample Caucasian males, those who have low HDL (in the lower 30 percent tail) have a 62.5% higher chance to develop coronary artery calcification compared with those who have high HDL (in the upper 30 percent tail). Caucasian women who have high triglycerides have a 55.6% higher chance of developing coronary artery calcification compared with those who have low triglycerides.
(Guerra et al. 2005).

5.2 Data

These data are part of the Dallas Heart Study (DHS) which sampled three major ethnicities: African American, Hispanic and Caucasian American. We have already described the genotype data in Chapter 4. In this section we summarize phenotypes associated with heart diseases and the subjects that were sampled.

5.2.1 Subjects

The three ethnicity groups in our analysis include 3551 individuals altogether. After excluding those who were on hypercholesterolemia treatment or missing all marker information, we were left with 3126 individuals for analysis. Out of the 3126 individuals, there are 953 black women, 685 black men, 482 white women, 445 white men, 324 Hispanic women and 237 Hispanic men.

5.2.2 Metabolic Risk Factors

In this study, two plant sterols (sitosterol and campesterol), triglycerides and high-density lipoprotein (HDL) were measured. Since all of them are skewed to some degree, the log transformation was applied. The two plant sterols,
sitosterol and campesterol, are highly correlated. Pearson’s correlation ranges from 0.85 to 0.90 for the log-scaled versions of sitosterol and campesterol, among the different ethnic and sex groups. The ratio between sitosterol and cholesterol is of great interest because the amount of sterols depends on the amount of cholesterol (Miettinen et al. 1994). However, the results from this ratio and sitosterol are very similar and so only the results based on sitosterol will be reported.

5.3 Statistical Methods and Results

To locate mutations associated with traits, we first applied the traditional single-SNP study, i.e., all the markers were tested one by one. Due to the large number of SNPs, P-values with corrections based on the false discovery rate (Benjamini and Hochberg 1999) are also reported. To take account of LD between SNPs, we performed haplotype-SNP based studies. The haplotype-SNP based studies confirm the results from single-SNP based studies in most situations and also detect some interesting regions where one or several specific haplotype groups show significantly different trait values.
5.3.1 Single-SNP Based Method

Different ethnicity groups vary both genotypically and phenotypically. The genetic effect on women and men can also be quite different. To protect from getting pseudo association or losing true association, we classified subjects by ethnicity and sex. Thus, there are six ethnic-sex combinations. Four models were applied to each SNP based on the mode of inheritance: the one-way analysis of variance (ANOVA); the additive model, the dominant model, and the recessive model. Consider a SNP with major allele A (more frequent) and minor allele T (less frequent). There are three possible genotypes: AA, AT, and TT. The one-way ANOVA model makes no restrictions about the traits for these three genotype groups; the additive model assumes that the average trait value from the heterozygote group is between the trait values from the two homozygous groups; the dominant or recessive models assume that the heterozygote group has the same trait values as one of the homozygote groups.

The constraints and hypotheses for the different models are listed in Table 5.1.

<table>
<thead>
<tr>
<th></th>
<th>$\mu_{AA}$</th>
<th>$\mu_{AT}$</th>
<th>$\mu_{TT}$</th>
<th>$H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way ANOVA</td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>$\mu_3$</td>
<td>$\mu_1 = \mu_2 = \mu_3$</td>
</tr>
<tr>
<td>Additive</td>
<td>$\mu_1$</td>
<td>$(\mu_1 + \mu_2)/2$</td>
<td>$\mu_2$</td>
<td>$\mu_1 = \mu_2$</td>
</tr>
<tr>
<td>Dominant</td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>$\mu_2$</td>
<td>$\mu_1 = \mu_2$</td>
</tr>
<tr>
<td>Recessive</td>
<td>$\mu_1$</td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>$\mu_1 = \mu_2$</td>
</tr>
</tbody>
</table>

Table 5.1: Model assumption and hypothesis tests for one-way ANOVA, additive, dominant, and recessive models.

Since no prior information is available about how each SNP controls the phe-
notype, we calculated P-values using all four different models in Table 5.1. In this study a SNP is said to be associated with the trait if at least one of the four P values is less than 0.01. In the three gene-trait systems considered, we found several significant SNPs. In the ABCG5+8 locus we found SNPs X44040782(D19H), X44048416, X44048966, X44073141 are statistically associated with sitosterol. SNP X44040782 was also reported by Berge et al 2002. In the APOAV locus, three SNPs, X116195213, X116200059, and X116201692 are found to be related with triglyceride. SNPs X56766760, X56768721, X56771555, X56781775, and X56790488 in the CETP locus are associated with HDL in most of race-sex groups. To correct for multiple comparisons, we applied the stepwise approach that controls the false discovery rate (FDR) advised by Benjamini and Hochberg 1999. This approach has been shown to be proper for dependent problems (Benjamini and Yekutieli 2001), especially for genome screens of complex disorders (Sabatti et al. 2003). Ultimately, define a SNP to be associated with a trait if the adjusted P-value is less than 0.05. The significant SNPs we found based on FDR correction using 0.05 as the threshold are similar to what we found using the raw P-values and 0.01 as the threshold.

5.3.2 Haplotype-SNP Based Study

Researchers have been focusing on haplotype-SNP based association studies in recent years. Haplotype-SNP studies are preferred over single-SNP studies for
several reasons. First, it is hard to solve the multiple testing problem in single-
SNP based studies both because the number of SNPs are usually very large,
and SNPs are not independent. This makes most multiple testing corrections
improper. Second, in haplotype-SNP based studies, SNPs within the same
chromosome are treated together, which is more biologically meaningful than
some randomly picked sets since each amino acid is decided by three contiguous
nucleotides. Third, in recent studies, haplotype-SNP based association studies
have been shown higher power than single-SNP studies. The comparisons have
been based on theory (Akey and Xiong 2001), real data (Martin et al. 2000) and
simulated data (Morris and Kaplan 2002; Zaykin et al. 2002; Hao et al. 2004),
which makes haplotype-SNP based studies very promising and appealing.
However, usually it is impractical to study haplotypes constructed from all
SNPs with a long region due to the large variation of haplotypes arising from
the large number of markers. In 2001, several groups (Daly et al. 2001; Jeffrey et al.
2001; Patil et al. 2001) reported that a long chromosome region can be
partitioned into disjoint blocks that are separated by recombination hotspots.
Each block contains some highly correlated and physically contiguous SNPs.
Recombination events are very rare within these blocks compared to hotspots
that separate the blocks. The most appealing property of a haplotype block
model is that within each block there is limited haplotype variation/diversity,
which enables us to study SNPs within each block simultaneously. Another
reason that we prefer haplotype blocks over randomly picked sets or sliding windows is that some particular haplotype groups within blocks might be passed from the same diseased ancestor, which might increase the power to detect association. Still, despite all these apparent advantage of haplotype blocks over single SNP approaches, we do not find overwhelming evidence of block superiority in the context of real data as discussed below.

5.3.2.1 Haplotype Block Structure

To show linkage disequilibrium (LD) levels over chromosomal regions we use the absolute value of $D'$. Since women and men in the same ethnic group do not differ very much on these three autosomal regions, they are combined to investigate the LD structure. The LD matrix plots for the three ethnic groups in the loci ABCG5+8, APOAV, and CETP are shown in Figure 5.1. Warmer colors, like red, indicate high LD value while cooler colors, like dark blue, indicate low LD. All images show some "block-like" structure: there are several big red squares with high LD in each pixel. LD values between two red squares usually are very low. The heat maps and all the block based results are constructed using only informative SNPs (MAF > 0.1) since SNPs with small heterogeneity (very low minor allele frequency) might be caused by recent mutations and they are more susceptible to small size variation, both of which could bias the estimation of LD.
Figure 5.1: Pairwise $|D'|$ values in the ABCG5+8, APOAV, and CETP loci. Men and women are combined. A red color indicates higher $LD$ while a blue color indicates lower $LD$. 
Using our new block partition method (Ch3), Blacks have more but shorter blocks, while Hispanics and Whites have fewer but longer blocks (Figure 5.2). This is to be expected since Blacks are an older population and hence have higher recombination rate and lower \( LD \). In the ABCG5+8 region, 23 blocks were estimated in Blacks with a mean size (number of SNPs) of 4.6 (median 3) and a mean length of 1.48kb (median 1.04kb). Some of the blocks are difficult to visualize in Figure 5.2 due to SNPs being so close together. Thirteen blocks were estimated from the Hispanics with a mean size of 8.77 (median 4) and a mean length of 5.97kb (median 2.50kb); 15 blocks were estimated from the Whites with a mean size of 6.6 (median 4) and a mean length of 3.89kb (2.16kb). In the APOAV region, 13 blocks were estimated from the blacks with a mean number of SNPs of 4.62 (median 4), a mean length of 3.98kb (median 2.5kb); 8 blocks from the Hispanics with a mean number of SNPs of 7.13 (median 5), a mean length of 4.41kb (median 2.31kb); 7 blocks from the whites with a mean number of SNPs 5.57 (median 4) and a mean length 4.18kb (median 4.19kb). In the CETP region, 10 blocks were estimated from the blacks with a mean number of SNPs 3.1 (median 3) and a mean length 2.76kb (median 1.94kb); 6 blocks from the Hispanics, with a mean size of 4.67 (median 3.5) and a mean length of 4.89kb (median 5.00kb); 6 blocks from the whites, with a mean size of 5.5 (median 5.5) and a mean length of 5.29kb (median 6.76kb).

Though the estimated haplotype blocks at each of the three loci from the three
ethnic groups show some differences, they share many common gaps (or so-called “recombination hotspots”). For example, there is a less than 200bp long gap starting at position 56770435 and ending at 56770633 in CETP, just after the 56770 kb location (see Figure 5.2). The estimated block structure from all three ethnicity groups shows that there is a recombination hotspot in that region. The similarity might indicate a true recombination hotspot since it is strongly consistent in all three groups.

In quantitative association studies, additive model seems to be more commonly applied than other penetrance models. From the parameterization point of view, the additive model is more efficient since it estimates fewer parameters. Also, most SNPs show a strong additive effect in single SNP based studies. Motivated by these facts, we modeled the association between haplotypes and traits by an additive model in each block. Since the linkage phase information is unknown, haplotypes were first estimated using the EM algorithm (Excoffier and Slatkin 1995). Suppose there are H different haplotypes from a population in a block. Denote them by \( h_1, \ldots, h_H \). Assume haplotype \( h_i \) has effect \( b_i \) on a trait. Suppose the \( i^{th} \) person has two haplotypes in the block, \( h_{i_1} \) and \( h_{i_2} \). His trait \( y_i \) will be modeled by:

\[
y_i = \mu + b_{i_1} + b_{i_2} + \text{error}
\]
Figure 5.2: Estimated haplotype blocks at three different gene loci. Each tick mark indicates a marker genotyped in that position.
Using the **haplo.glm** program from Schaid et al. 2002 and Lake et al. 2003, we carried out haplotype-block based association between sitosterol and ABCG5+8, triglyceride and CETP, and HDL and APOAV. Like the single-SNP study, the false discovery rate corrected version of the analysis is also presented. The global P-values (which test the null hypothesis that none haplotype has an effect) for each block, together with the single SNP P-values, are illustrated in Figure 5.3, 5.5 and 5.7.

### 5.4 Comparison between Single-SNP Based and Haplotype-block Based Methods

While there have been several papers showing successes in real data via haplotype based association, they were either working on case-control studies or in a very short region with relatively few SNPs (e.g., less than 5 SNPs). It is therefore unknown how haplotype association works in the context of many more SNPs. It maybe that more SNPs introduce more noise. Martin et al. 2000 used sliding windows with 2 or 3 SNPs around the targeted disease causal marker and found that higher power was gained by using haplotype-based analysis. Epstein and Satten 2003 considered a region with 5 SNPs. In our study, the “size” of haplotype blocks we considered were as large as 33 SNPs. Also, since the phase information is unknown, analyzing and explaining the results
is more difficult with more SNPs. We compared the single-SNP based study and haplotype-SNP based study in this way: we first studied how much the blockwise P-values agree with single SNP P-values within each block; then we tested if the agreement (or disagreement) is a function of block size (the number of SNPs), length of blocks (the physical length in kb), or maker density within each block.

For both the single-SNP based study and the haplotype-SNP based study we used a significance threshold of 0.05 to decide if a single SNP or a whole block is associated with a certain trait. A χ² test with one degree of freedom is used to test the null hypothesis that there is no association between the results from the haplotype based study and the significance from the single SNP study. The counts are summarized in Tables 5.2, 5.4, and 5.6. In region ABCG5+8, the Blacks have smaller and shorter blocks compared with the other two ethnic groups. The P-value from the χ² test is less than 0.0001, which indicates that the results of association test from the haplotype based study strongly agree with the results of association tests from the single SNP study. The results from the two approaches in the other two groups are either weakly associated or there is lack of evidence. SNPs would seem to express their effects much better in shorter blocks than longer blocks, which may explain why Blacks show a strong association between single-SNP and haplotype-SNP results.

In the APOAV region, no significant block is detected from Blacks (Figure 5.5).
However, since there is only one significant single SNP, the single SNP analysis also agrees with haplotype based analysis very well. Results from the other two ethnicities do not agree, where long blocks were estimated.

The results from the CETP region are very similar with region ABCG5+8 (Figure 5.7).

<table>
<thead>
<tr>
<th></th>
<th>African American</th>
<th>Hispanic</th>
<th>Caucasian American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>women/men</td>
<td>women/men</td>
<td>women/men</td>
</tr>
<tr>
<td>SS</td>
<td>23(22.5)/25(24.5)</td>
<td>3(2.6)/8(7.0)</td>
<td>5(5.1)/0(0)</td>
</tr>
<tr>
<td>SN</td>
<td>4(3.9)/4(3.9)</td>
<td>25(21.9)/3(2.6)</td>
<td>15(15.2)/12(12.1)</td>
</tr>
<tr>
<td>NS</td>
<td>25(24.5)/27(26.5)</td>
<td>30(26.3)/52(45.6)</td>
<td>21(21.2)/0(0)</td>
</tr>
<tr>
<td>NN</td>
<td>50(49.0)/46(45.0)</td>
<td>56(49.1)/51(46.7)</td>
<td>58(58.5)/87(87.9)</td>
</tr>
<tr>
<td>$P_{\chi^2}$</td>
<td>0/0</td>
<td>.027/.277</td>
<td>.888/-</td>
</tr>
<tr>
<td>$P_{size}$</td>
<td>.784(+)/.435(+)</td>
<td>&lt;.0001(-)/.452(+)</td>
<td>.448(-)/-</td>
</tr>
<tr>
<td>$P_{length}$</td>
<td>.306(+)/.917(+)</td>
<td>&lt;.0001(-)/.082(+)</td>
<td>.719(-)/-</td>
</tr>
<tr>
<td>$P_{density}$</td>
<td>.920(+)/.284(+)</td>
<td>1/.317(+)</td>
<td>.510(+)/-</td>
</tr>
</tbody>
</table>

Table 5.2: Comparison between results from single SNP based association study and haplotype block based association study in ABCG5+8 region. SS indicates the number of SNPs that show single SNP significance and blockwise significance; SN is the number of significant single SNPs but not in significant blocks; NS is the number of SNPs that are not significant in single SNP study but are in significant blocks; NN is the number of SNPs that are neither significant in SNP study nor in haplotype-block based study. The values in the parentheses are the percentages. $P_{\chi^2}$ is the P-value calculated from the 2 x 2 table based on the four numbers. $P_{size}$, $P_{length}$, and $P_{density}$ are P-values to test if the agreement is a function of block size, block length, or block density (number of SNPs per kb) using logistic regression. The signs in the parentheses indicate how the block size, block length, and block density affect the agreement. For example, a negative sign indicates that the agreement is negatively associated with size or length, i.e., the results from two approaches tend to agree more in short blocks and agree less in long blocks. Notice that when there only one or two nonzero counts, $P_{\chi^2}$ cannot be calculated.

We also tested if the agreement (or disagreement) between the two approaches is a function of block size, block length, or marker density of blocks. Smaller
Figure 5.3: Association study of sitosterol with ABCG5+8 using the single SNP analysis and the haplotype SNP based analysis.
Figure 5.4: Corrected P-values using step-wise procedure by Benjamini and Hochberg 1995 for association between sitosterol and ABCG5+8 using single SNP analysis and haplotype SNP analysis.
<table>
<thead>
<tr>
<th></th>
<th>African American</th>
<th>Hispanic</th>
<th>Caucasian American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>women</td>
<td>men</td>
<td>women</td>
</tr>
<tr>
<td>SS</td>
<td>10</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>SN</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>NS</td>
<td>26</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>NN</td>
<td>63</td>
<td>57</td>
<td>114</td>
</tr>
<tr>
<td>$P_{x^2}$</td>
<td>.002</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$P_{size}$</td>
<td>.996(-)</td>
<td>.650(-)</td>
<td>-</td>
</tr>
<tr>
<td>$P_{length}$</td>
<td>.012(+)</td>
<td>.187(-)</td>
<td>-</td>
</tr>
<tr>
<td>$P_{density}$</td>
<td>.920(+)</td>
<td>.284(+)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.3: Comparison between Benjamini-Hochberg corrected results from the single SNP based association study and the haplotype block based association study in the ABCG5+8 region.

<table>
<thead>
<tr>
<th></th>
<th>African American</th>
<th>Hispanic</th>
<th>Caucasian American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>women</td>
<td>men</td>
<td>women</td>
</tr>
<tr>
<td>SS</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>SN</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>NS</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>NN</td>
<td>52</td>
<td>52</td>
<td>36</td>
</tr>
<tr>
<td>$P_{x^2}$</td>
<td>-</td>
<td>-</td>
<td>.004</td>
</tr>
<tr>
<td>$P_{size}$</td>
<td>-</td>
<td>-</td>
<td>.282(+)</td>
</tr>
<tr>
<td>$P_{length}$</td>
<td>-</td>
<td>-</td>
<td>.555(+)</td>
</tr>
<tr>
<td>$P_{density}$</td>
<td>-</td>
<td>-</td>
<td>.819(+)</td>
</tr>
</tbody>
</table>

Table 5.4: Comparison between results from the single SNP based association study and the haplotype block based association study in the APOAV region.
Figure 5.5: Association study of triglyceride and APOAV using the single SNP analysis and the haplotype-block based analysis.
Figure 5.6: Corrected P-values using the step-wise procedure by Benjamini and Hochberg 1995 of association study between triglyceride and APOAV with the single SNP analysis the and haplotype SNP analysis.
<table>
<thead>
<tr>
<th></th>
<th>African American</th>
<th>Hispanic</th>
<th>Caucasian American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>women</td>
<td>men</td>
<td>women</td>
</tr>
<tr>
<td>SS</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SN</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>NS</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NN</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$P_{\chi^2}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$P_{size}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$P_{length}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$P_{density}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.5: Comparison between Benjamini-Hochberg corrected results from the single SNP based association study and the haplotype block based association study in the APOAV region.

<table>
<thead>
<tr>
<th></th>
<th>African American</th>
<th>Hispanic</th>
<th>Caucasian American</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>women</td>
<td>men</td>
<td>women</td>
</tr>
<tr>
<td>SS</td>
<td>8</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>SN</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>NS</td>
<td>7</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>NN</td>
<td>12</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>$P_{\chi^2}$</td>
<td>.03</td>
<td>.005</td>
<td>.806</td>
</tr>
<tr>
<td>$P_{size}$</td>
<td>.058(-)</td>
<td>1</td>
<td>.705(+)</td>
</tr>
<tr>
<td>$P_{length}$</td>
<td>.702(+)</td>
<td>.706(+)</td>
<td>.432(+)</td>
</tr>
<tr>
<td>$P_{density}$</td>
<td>.147(+)</td>
<td>.988(-)</td>
<td>.963(+)</td>
</tr>
</tbody>
</table>

Table 5.6: Comparison between results from the single SNP based association study and the haplotype block based association study in the CETP region.
Figure 5.7: Association study of HDL with CETP using the single SNP analysis and the haplotype-block based analysis.
Figure 5.8: Corrected P-values using the step-wise procedure by Benjamini and Hochberg (1995).
and shorter blocks tend to have much less noise, which as stated above would be expected to have more agreement. For each SNP in a block, if the results from the single-SNP approach and the haplotype-SNP approach agree, we will say there is an agreement; otherwise we call it disagreement. The results agree if both are significant or both are not significant. Thus for each SNP in a block we define a binary variable: 1=agree, 0=disagree. To test if the binary results are associated with the block size, block length or marker density, we use the binary results as the dependent variable and the block size, length, or marker density as a covariate, respectively, in a logistic regression. The associations between block size, length, and the agreement are weak except in the Hispanics group in the ABCG5+8 region, Hispanic women in APOAV region, and White men in CETP region. We also realize that all of the significant associations are negative. One possible explanation is that when considering all the SNPs
within a long chromosome region, even if there is no recombination event in the history, a mutation itself can create different haplotypes. As the number of haplotypes increases, the degrees of freedom are enlarged, hence power will be lost. The tables all show that the agreement is not associated with marker density in any study.

To correct for multiple comparisons, we did the same study using the step-wise correcting procedure provided by Benjamini and Hochberg 1999. A P-value of 0.05 was used as the threshold to separate evidence of significance and a lack of evidence to reject \( H_0 \). Since the corrected P-values are more conservative, fewer single SNPs and blocks show significance. Despite the difficulty of applying \( \chi^2 \) tests due to sparse cells in the ABCG5+8 region and the CETP region, we again found very significant negative association with the size of the blocks. The details are reported in Figures 5.4, 5.6, 5.8 and Tables 5.3, 5.5, 5.7. The X-axis of figures shows the relative physical positions of the genotyped SNPs. The y-axis illustrates the -log10(P). The solid horizontal line is corresponding to the P-value at 0.05 level. Rescaled P-values from the single-SNP study were drawn using circles that are connected by lines. The height of each shaded rectangle indicates the global P-value from the haplotype based study within each block.

Comparison of the two approaches in this section indicate that they tend to agree more when the estimated haplotype blocks are shorter or smaller in size.
5.5 Simulation

Section 5.4 discussed differences between results from single-SNP study and haplotype-SNP study and the factors that affect the differences. Since the data were real we have an opportunity to assess relative differences between single and haplotype based methods. However, we are unable to evaluate power since we do not know the causal loci for the various traits. For this we need simulation. Most of current simulation work that studies the statistical power simulates both genotype and phenotype. When they simulate genotype, their assumptions are usually much simpler than reality. For example, Akey et al. 2001 assume that the markers around the disease causal locus have equal allele frequencies and are symmetrically arranged on both sides of the disease causal locus; Morris and Kaplan 2002 assume both equal allele frequencies and complete association between susceptibility alleles and marker alleles. Simulations based on these simple assumptions may favor haplotype-SNP approaches in terms of statistical power.

To study the relative power of the two approaches in a more realistic scenario, we used the real DHS genotype data from the APOAV region for all 330 African American men that didn’t have any missing genotype and the CETP region for all 112 Hispanic men that didn’t miss any genotype. The trait, however, was simulated based on single marker using an additive model. For example, for a locus with genotype AA, AT, and TT, we assume the trait follows normal
distributions with means 0, 5, and 10, respectively. We also used two different standard deviations, 10 and 15. For example, consider the 6th (also the last) haplotype block at the CETP locus for Hispanic men. Five SNPs were genotyped in this block. We first treated the third SNP (marked by red star) as the causal SNP. For each standard deviation, we simulated trait values 100 times in order to estimate the power to detect association with the trait. The power from every other SNP based on single-SNP approach is presented using triangle in Figure 5.9. The red horizontal line indicates the global P-value from haplotype-SNP approach using all the other SNPs besides the disease causal SNP. Figure 5.9 shows the haplotype approach has higher power than any single SNP. However, the result is reversed if we use the fourth SNP (marked by blue star) as the causal SNP.

We also studied the whole CETP locus for all 112 Hispanic men and the whole APOAV locus for the 330 Black men. For each SNP in a block, we used this SNP as the disease causal SNP and used the additive model described in last paragraph to simulate the trait values. In the single-SNP study we tested the association between the simulated trait and all the other SNPs within the same block one at a time assuming an additive effect of the markers. If at least one of the P-values (uncorrected) is less than 0.05, we say the single-SNP study “hit” the disease causal SNP. The haplotype-SNP study was also performed by testing the association between the simulated trait and the haplotypes constructed by
Figure 5.9: Simulated statistical power based on single-SNP study and haplotype-SNP study using all 112 Hispanic men that didn’t miss any genotype in one short region at CETP locus.
all the other SNPs besides the causal SNP. We then used haplo.score (Schaid 2002) to perform the haplotype association study. If the global P-value is less than 0.05, we say the haplotype-SNP study “hit” the disease causal locus. For each SNP in each block, the procedure of simulating a trait and testing association by the two approaches was repeated 100 times. The statistical power of capturing the disease causal SNP was then defined as the total number of hits divided by the number of runs (100 here). In Figures 5.10 and 5.11, letter ‘s’ and ‘b’ indicate results from single-SNP study and haplotype-SNP study within a block, respectively. The x-axis indicate the relative physical positions of SNPs. The height of a letter ‘s’ is the power of capturing the disease causal SNP by using single-SNP study based on all the other SNPs within the same block; the height of a letter ‘b’ is the power of capturing the disease causal SNP by haplotype-SNP approach based on all the other SNPs within the same block. In blocks with two SNPs, the powers from the two approaches are the same because there is only one SNP left once the SNP that simulated the trait was removed.

In the APOAV region based on 330 African American men, single-SNP approach has higher power than haplotype-SNP approach in most blocks except block 4 (B4 in Figure 5.10), and block 10 (B10). In the CETP region (Figure 5.11) based on 112 Hispanic men, haplotype-SNP study shows higher power only for traits simulated based on 3 SNPs out of 6 SNPs in block 1 when sd=10.
Figure 5.10: Simulated statistical power based on single-SNP study and haplotype-SNP study using all 330 African American men that didn’t miss any genotype in the APOAV region.
Figure 5.11: Simulated statistical power based on single-SNP study and haplotype-SNP study using all 112 Hispanic American men that didn’t miss any genotype in the CETP region.
Haplotype-SNP study doesn't have higher statistical power than that of single-SNP study in all the other blocks. Based on this simulation, we lack evidence to conclude that haplotype-SNP approach is always more advantageous.

5.6 Conclusion

In this chapter, we studied single-SNP and haplotype-SNP approaches to association with real data from the DHS: plant sterols and SNPs in the ABCG5+8 locus, triglyceride and the APOA5 locus, and HDL and the CETP locus. By using traditional single-SNP studies, we confirmed known significant associations between some single SNPs and traits. We then estimated the haplotype block structure at each of the three loci. Within the same locus, different ethnicity groups have different block structures, however, they also share some common hotspots of recombination. The two approaches agree very well in the African Americans, which have smaller and shorter blocks, but not the Hispanics and the Caucasian Americans, which show larger and longer blocks. Also, the discordance seems to occur in long blocks where haplotype diversity is relatively higher and more noise exists apparently from evolutionary forces such as mutation, compared to shorter blocks. Several publications report that the haplotype-SNP study is more powerful than the single-SNP studies. However, the assumptions appear to be simplified (Martin et al. 2000; Akey et al. 2001;
Morris and Kaplan 2002; and Epstein et al. 2003). To study which approach is more powerful in a more realistic context, we used real genotype data sets from the Dallas Heart Study with simulated traits based on single markers. The result shows that the haplotype-SNP study does not consistently dominate single-SNP study. Detailed study about the subtle LD structure within each block might be necessary in order to fully understand and interpret the differences.
Chapter 6

Discussion

This thesis centers on current issues and some solutions for haplotype block structure estimation and its applications in genetic association mapping using SNP data. Haplotype block structure in the human genome was proposed in 2001 by Daly et al. 2001 and Patil et al. 2001. Our contribution involves three different but related problems: haplotype block estimation (Chapter 3); the impact of marker density, sample size and selection strategy on haplotype block estimation (Chapter 4); comparison between single-SNP and haplotype-SNP based association studies (Chapter 5).

In Chapter 3, we proposed a new bivariate summary statistic for LD: $BLD = (PD, minp)$. As mentioned by Hedrick 1987, a measure of LD should have (1) biological interpretation, (2) statistical tests available or easily developed, (3) be directly related mathematically to evolutionary factors such as recombination,
selection, genetic drift, gene flow, among other factors. and (4) be standardized to allow comparisons across loci or populations. We verified (1), (2), (4) and part of (3) in Chapter 3.

The SNP data from Dallas Heart Study (DHS) directed by Dr. Helen Hobbs of the University of Texas Southwestern Medical Center are fully genotyped in over 300 genes in more than 3000 unrelated individuals. This unique data set enabled us to study the impact of marker density, sample size and SNP selection strategies on haplotype block estimation. We studied four different SNP selection methods, namely, random selection, equally spaced selection, \( r \)-based tagging and block-based tagging (Chapter 3). Different thresholds and sample sizes were used. The random selection and equally spaced selection procedures have similar effects on estimation of haplotype block structure. To our surprise, we found that the \( r \)-based tagging and block-based tagging are generally less efficient than the other two methods. Both sample size and marker density affect the accuracy of estimated block structure. Sample size greater than 50 seems to be necessary while a small sample size of approximately 25 is not enough for estimating haplotype blocks. Compared with sample size, the percent of SNPs used (marker density) seems to affect accuracy more strongly. Based on the tagging strategies we studied, increasing percent of SNPs used (equivalent to marker density in a given region) might be the best way to accurately estimate block structure.
Besides Carlson's method, there are other statistical tagging methods. The reason we chose Carlson's method is because it does not require haplotype phase information. Even though we can estimate haplotype pairs, most haplotypes are rare (<1%) due to the large number of SNPs considered in our study, which makes methods based on haplotype diversity very inefficient. Other tagging methods that can better utilize genotype data would be worth investigation. Carlson et al. 2004 claim that their new approach has better power than many haplotype diversity based approaches. Having compared Carlson's to random or equal spaced selection, we may infer that the other methods for tagging may also be less accurate. Based on our study, we suggest that marker density might be the last source that we should consider to cut operating budgets.

The target of statistical genetics is to identify disease causal markers. The traditional approach is testing markers one by one. With modern developments in biotechnology high-throughput assays enable whole genome scan. However, the large number of SNPs leads to multiple testing problems that do not have satisfactory solution. Traditional corrections such as Bonferroni correction cannot be applied to such large and correlated data. False discovery rate (FDR) correction has claim to be suitable for correlated data like SNPs (Benjamini and Hochberg 1999, Sabatti et al. 2003), but, the LD structure is not specifically used in FDR procedure. As an alternative, and with the hope of reducing problems caused by multiple testing, efforts have also been put on multiloci association studies:
logistic regression by Cruickshanks et al. 1992; Bayesian genomic control by Devlin and Roeder 1999; decay of haplotype sharing by McPeek et al. 1999; logic regression by Kooperberg et al. 2001; sums of single-SNP statistics by Hoh and Ott 2003 and Hao et al. Hao 2004; Hotelling's $T^2$ by Xiong et al. 2002 and Fan and Knapp 2003, cladistic association by Templeton et al. 1987. There are also theoretical considerations (Akey et al. 2001), simulation work (Zollner and von Haeseler 2000), and evidence from real data (Morris and Kaplan 2002) that haplotype based association studies have higher power. However, most of these studies are limited to ultimately a few SNPs ($S < 10$). Studies involving many SNPs are needed and this thesis is a contribution in this direction.

In our haplotype-SNP study, we first partition a long region to short regions using our newly proposed haplotype block estimation method. Then, association studies within each haplotype block were performed using the \texttt{haplo.score} program developed by Schaid et al. 2002 and Lake et al. 2003. There are several benefits in using haplotype-block based haplotype-SNP methods. First, at least in our real data, the genes usually cover 40-100 kb. It is computationally impractical to study all of the SNPs together in such long and dense areas. Second, since SNPs within each haplotype block are highly correlated, only a limited amount of haplotype diversity is observed.

Using the genotype data and several complex traits from DHS, we investigated genetic association using both single-SNP and haplotype-SNP approaches.
Comparison between these two approaches show that the two methods tend to agree more when haplotype block size is small. Association tests using our simulated traits based on read genotype data also show that haplotype-SNP based study does not always have higher power than the single-SNP based study, especially in long haplotype blocks, where the \( LD \) structure might be very complex and more SNPs introduce more noise. Even when higher power is gained, the biological explanation must be considered. For example, suppose there are two SNPs A and B and a disease locus D. In Figure 6.1, both pairs (A,D) and (B,D) have very high \( LD \) (\( D' = 1 \)). Using either an asymptotic \( \chi^2 \) test or Fisher's exact test, neither of them will show significance based on a sample size of 10. But if we study the association between the combination of A,B and D, both tests show significance (p-value< 0.01). The biological explanation can be that only individuals who have a ‘0’ allele at SNP A and a ‘0’ allele at SNP B are not diseased. In Figure 6.2, with sample size 50, the association between A and D is marginal with asymptotic \( P=0.012 \) and Fisher exact \( P=0.025 \); there is no evidence of association between B and D or between A and B. But the combined table shows that multiloci (A, B) can bring higher power. However, it is difficult to explain since the two haplotypes ‘11’ and ‘00’ are associated with diseases while the two homozygotes are not. Hence, attention must be paid to discern between statistical significance and biological significance.
Figure 6.1: Single-SNP vs haplotype-SNP association. Each SNP has two alleles represented at black (1) and white (0) in the vertical bars. Haplotypes in the bar-figure and tables are read similarly; for example, 30% of the population has haplotype $AD = 11$. See text for additional explanation. Both this figure and Figure 6.2 have been used by Guerra and Yu 2005.

Figure 6.2: Single-SNP and haplotype-SNP association. See Figure 6.1 for explanation.
Bibliography


112


