This is to certify that the
thesis entitled

STATISTICAL METHODS FOR IDENTIFYING
GENETIC ASSOCIATIONS

presented by

LAN TONG

has been accepted towards fulfillment
of the requirements for the

M. S. degree in Applied Statistics

Marianne Huebner
Major Professor's Signature

4-29-05

Date

# STATISTICAL METHODS FOR IDENTIFYING GENETIC ASSOCIATIONS

By

Lan Tong

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Statistics

2005

# ABSTRACT

# STATISTICAL METHODS FOR IDENTIFYING GENETIC ASSOCIATIONS

By

Lan Tong

This thesis introduces three bio-statistical concepts: Hardy-Weinberg equilibrium, linkage disequilibrium, and haplotype reconstruction and haplotype frequency estimation. Three statistical methods have been discussed and utilized to test these concepts for a population genetics study.

Hardy-Weinberg equilibrium is a basis for genetic inference. It is tested through the exact test implemented by the Arlequin software package. Linkage disequilibrium is an important tool for mapping disease genes. It is tested with a likelihood-ratio test, whose key procedure is the Expectation-Maximum algorithm, implemented by the Arlequin software package. Haplotype information is essential for mapping disease. The haplotype frequencies are estimated through the Bayesian estimation method implemented by the PHASE software package.

The above concepts and tests have been applied to the Isle of Wight cohort study. It has been found that all the loci of interest (*hCV8932056, hCV15862743, hCV8932053,* and *hCV8932052* on the *IL13* gene) are in Hardy-Weinberg equilibrium, and that all pair-wise loci are in linkage disequilibrium. The haplotypes of the most informative SNP pair, *hCV8932056* and *hCV8932052,* have been reconstructed; their frequencies are estimated for eight phenotypes of interests. The contingency tests suggest that there is no association between the haplotype patterns *CA/CG/TA/TG* and *CA/TA/TG* and allergic asthma.

# ACKNOWLEDGMENTS

I would like to express my appreciation to the several people who have been generous with their encouragement and support for this thesis. Most of all, I am grateful to Dr. Marianne Huebner, Department of Statistics, Michigan State University, and Dr. Susan Ewart, Department of Large Animal Clinical Sciences, Michigan State University, for providing me the opportunity of participating the Isle of Wight birth cohort study. Dr. Huebner has mentored me throughout the project and the writing of the thesis. I have had many helpful discussions with Dr. Ewart and Dr. Dennis Gilliland, Department of Statistics, Michigan State University, both of whom provided me invaluable suggestions for the improvement of my thesis. Dr. James Stapleton, Department of Statistics, Michigan State University, has served on my committee and offered me constructive advice about my thesis.

I also want to thank my parents for taking care of my newborn daughter so that I was able to concentrate on my research and thesis writing. I am deeply indebted to my husband whose love, support, and knowledge have made the completion of the project and the thesis possible.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1 INTRODUCTION

## Section 1 Background

The purpose of this thesis is to discuss three bio-statistical topics: Hardy-Weinberg equilibrium, linkage disequilibrium, and haplotype reconstruction and haplotype frequency estimation, and their applications in population genetics.

The Hardy-Weinberg law plays a very important role in the field of population genetics and often serves as a basis for genetic inference (Crow, 1988). This law says that in a large random-mating population with no selection, mutation, or migration, the allele frequencies and the genotype frequencies are constant from generation to generation and that, furthermore, if the alleles $A_1, A_2, \ldots, A_m$ have relative frequencies $f_1, f_2, \ldots, f_m$ respectively, then the relative frequency of homozygous genotypes such as $A_iA_i$ is $g_{ii} = f_i^2$, and the relative frequency of a heterozygous genotype such as $A_iA_j$ $( i \neq j )$ is $g_{ij} = 2f_if_j$. Because of its importance, a lot of effort has been made to test if a population exhibits Hardy-Weinberg equilibrium. This paper discusses, in detail, one of the statistical methods, exact test, used to test Hardy-Weinberg equilibrium.

Linkage disequilibrium is an important tool for mapping disease genes. It describes the nonrandom association of alleles at linked loci. When only genotype and not haplotype frequencies are available, linkage disequilibrium between a pair of loci is tested with a likelihood-ratio test. This paper discusses the likelihood ratio test, with highlight on its key procedure, the Expectation-Maximum algorithm which resolves double heterozygous genotypes into haplotypes when not assuming linkage equilibrium.

Haplotype information is essential for mapping disease. When individuals are homozygous at every locus, haplotypes can be easily determined; however, when individuals are heterozygous at more than one locus, haplotypes cannot be deducted from genotype information. Statistical approaches can be used to reconstruct haplotypes and estimate the relative frequencies of all possible haplotypes. This paper introduces the Bayesian estimation method of reconstructing haplotypes and estimating haplotype frequencies.

All the above three concepts have been applied to the Isle of Wight birth cohort study. This study is committed to identify the genetic susceptibility loci responsible for asthma and allergy by studying a group of children born on the Isle of Wight, United Kingdom. The Hardy-Weinberg equilibrium and linkage disequilibrium for the four SNPs: *hCV8932056, hCV15862743, hCV8932053,* and *hCV8932052* on the *IL13* gene have been tested using the Arlequin software package. Haplotype reconstruction and frequency estimation for SNPs *hCV8932056* and *hCV8932052* have been performed using the PHASE software package.

Both the Arlequin and PHASE software packages are designed for population genetics data analysis. Arlequin implements exact test and likelihood-ratio test to examine the Hardy-Weinberg equilibrium and linkage disequilibrium, respectively. It is available for free at http://lgb.unige.ch/arlequin. PHASE implements the Bayesian algorithm to approximate the posterior distribution of haplotype configurations. It is also free and can be downloaded at http://www.stat.washington.edu/stephens/software.html.

Chapter 1 of this thesis introduces the background, the basic genetic concepts involved in the study, and the data of Isle of Wight birth cohort. Chapter 2 describes Hardy-Weinberg equilibrium, its significance, exact test of Hardy-Weinberg equilibrium, and testing the Isle of Wight birth cohort for Hardy-Weinberg equilibrium using

Arlequin. Chapter 3 discusses linkage disequilibrium, its significance, likelihood ratio test of linkage disequilibrium, and testing the Isle of Wight birth cohort for linkage disequilibrium using Arlequin. Chapter 4 explains the motif of haplotype reconstruction and frequency estimation, Bayesian estimation method, and estimating the Isle of Wight birth cohort for haplotype frequencies using PHASE. Finally, in Chapter 5, the test results in the previous chapters are summarized.

## Section 2 Genetic Concepts

To better understand genetic research and the Isle of Wight birth cohort study, it is helpful to explore some basic genetic definitions and concepts.

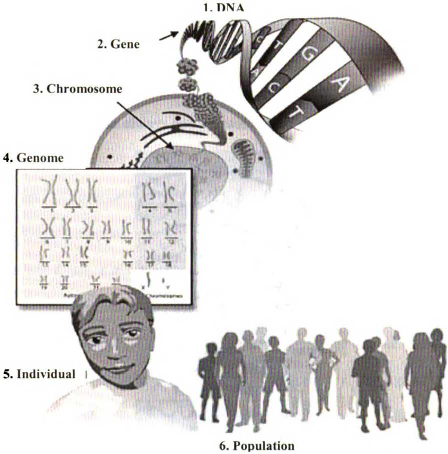


**Figure 1 Levels of Information Transfer[1]**

Genetics is the study of traits passed on from parent to child and variation of those traits within and between individuals. It is about the transfer of information among many different levels (Figure 1). The foundation level is the molecule called **DNA** (1). The information in DNA is organized into **genes** (2). Genes, in turn, make up **chromosomes** (3), which when taken all together form an organism's **genome** (4). Every nucleus in an **individual** (5) contains the genome.

Instructions that provide almost all of the information necessary for a living organism to grow and function are in the nucleus of every cell. The instructions are in the form of a molecule called deoxyribonucleic acid, or **DNA** (Figure 2).

In humans, the DNA molecule consists of two ribbon-like strands that wrap around each other, resembling a twisted ladder. The rungs of the ladder are **nucleotide base pairs**. The bases are called adenine ("*A*"), cytosine ("*C*"), guanine ("*G*") and thymine ("*T*"). These bases always pair up as *A+T* and *C+G.*

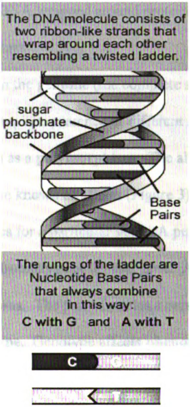

The DNA molecule consists of two ribbon-like strands that wrap around each other resembling a twisted ladder.
sugar phosphate backbone
Base Pairs
The rungs of the ladder are Nucleotide Base Pairs that always combine in this way:
C with G and A with T
C
T

**Figure 2 DNA**[2]

[1] This figure was originally created by GlaxoSmithKline. It is available in the online article 'Genetics at GlaxoSmithKline' at http://genetics.gsk.com/link.htm, 2004.

[2] This figure was originally created by GlaxoSmithKline. It is available in the online article 'Genetics at GlaxoSmithKline' at http://genetics.gsk.com/link.htm, 2004.

Before DNA was discovered, **gene** was defined as a discrete unit which is inherited from parent to offspring and which exerts control on a single character. After DNA was discovered, gene was redefined as a segment of DNA that codes for a protein subunit. The word gene may also be used to refer to a functional DNA segment or a class of functional DNA segments that have the same position, structure and function (Sham, 1998).

DNA is contained in tightly coiled packets called **chromosomes**. Chromosomes consist of the double helix of DNA wrapped around proteins. Each human cell nucleus contains 23 pairs of chromosomes.

A **locus** is a specific position in the **genome** (the complete set of genes). The DNA of most people hare highly similar. The presence of different DNA sequences at the same locus in a population is known as a **polymorphism**. The alternative DNA sequences at a locus on a chromosome pair are known as **alleles** (Figure 3). People can have two identical or two different alleles for a particular gene. A person who has two identical alleles for a gene is said to be **homozygous** for that gene. A person with two different alleles is said to be **heterozygous**. The pair of alleles a person has at a specific location in the genome is called **genotype**. Genotype affects **phenotype**, which is the observable effect of the allele, such as eye color. A combination of alleles is termed **haplotype**. Haplotypic **phase** refers to whether a gametic (a **gamete** is a sperm or an egg that fuse during reproduction) haplotype changes during recombination (GlaxoSmithKline, 2004).

.

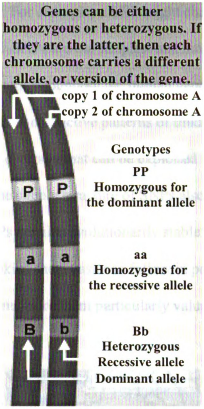

Genes can be either homozygous or heterozygous. If they are the latter, then each chromosome carries a different allele, or version of the gene.
copy 1 of chromosome A
copy 2 of chromosome A
Genotypes
P
P
PP
Homozygous for the dominant allele
a
a
aa
Homozygous for the recessive allele
B
b
Bb
Heterozygous
Recessive allele
Dominant allele

**Figure 3 Allele**[3]

Many diseases are related in some way to genes. Many common diseases result from a change in one or a few susceptibility genes.

To find a gene that is involved in a specific disease, scientists must search for DNA changes that are present more often in people who have a particular disease compared to people who do not have the disease (GlaxoSmithKline, 2004).

A contemporary kind of genetic map, called a high-density single nucleotide polymorphism ("**SNP**") (Figure 4) map, has the potential to promote this research. SNPs are single-base differences in the DNA sequence that can be observed between different

[3] This figure was originally created by GlaxoSmithKline. It is available in the online article 'Genetics at GlaxoSmithKline' at http://genetics.gsk.com/link.htm, 2004.

individuals in the population. For example, a SNP might change the DNA sequence *AAGGCTAA* to *ATGGCTAA*. SNPs are the simplest and most common forms of DNA polymorphism. They are present throughout the human genome. Groups of neighboring SNPs may have alleles that show distinctive patterns of linkage disequilibrium and as such may create a haplotypic diversity that can be exploited in both genetic linkage and direct association studies. The simple structure of SNPs also allows rapid and efficient genotyping. In addition, SNPs are also evolutionarily stable - not changing much from generation to generation - making them easier to follow in population studies. These features of SNPs in the genome make them particularly valuable as genetic markers (Schork et al, 2000).

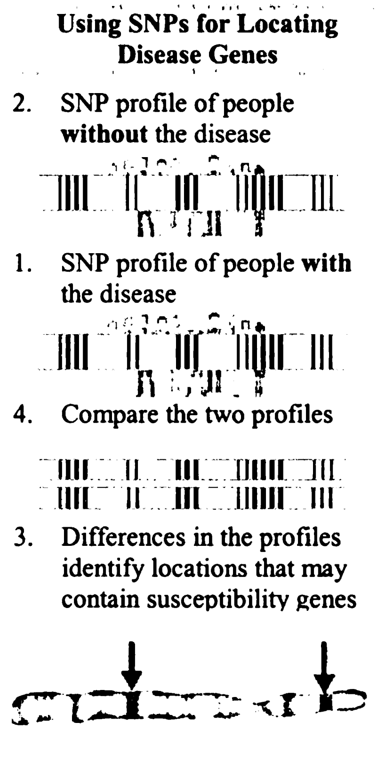

Using SNPs for Locating Disease Genes
2. SNP profile of people without the disease
1. SNP profile of people with the disease
4. Compare the two profiles
3. Differences in the profiles identify locations that may contain susceptibility genes

Figure 4 SNP[4]

[4] This figure was originally created by GlaxoSmithKline. It is available in the online article 'Genetics at GlaxoSmithKline' at http://genetics.gsk.com/link.htm, 2004.

Using the information that SNPs provide, it may be possible to predict people's genetic risk of developing a certain disease, to diagnose a disease more accurately, or to predict how people most likely will respond to a medicine.

## Section 3 Case Study

The Isle of Wight birth cohort study represents an unselected whole population birth cohort based on the Isle of Wight, United Kingdom. The Isle of Wight is a small island (13 x 23 miles) just off the South coast of England with a resident population of 133,000.

The ethnic background of the island residents is mainly Caucasian. While the Isle of Wight population is not genetically homogeneous, it is stable to the extent that the majority of children in the cohort has not moved away and has thus been available for follow up.

Enrollment took place at birth. Of the 1,536 children born on the Isle of Wight between January 1, 1989 and February 28, 1990, informed consent was obtained from the parents of 1,456 children. These children have since been seen at the ages of 1 (n = 1,167; 80.2%), 2 (n = 1,174; 80.6%), 4 (n = 1,218; 83.7%) and 10-years (n = 1,373; 94.3%) (Kurukulaaratchy et al, 2003).

At birth, information on family history of allergy, household pets, parental smoking, socioeconomic status and birth weight were recorded. At every follow-up (1, 2, 4 and 10-years), detailed questionnaires were completed with the parents for each child regarding asthma and allergy prevalence. For each child, the following phenotype information has been recorded: asthma at 1 or 2 years, asthma 4 years, currently diagnosed asthma (CDA) at 10 years, wheezing at 1 or 2 years, wheezing at 4 years, wheezing at 10 years, chronic asthma, and no symptoms. With high cohort retention, this prospectively followed population provides a uniquely characterized resource for ongoing studies (Kurukulaaratchy et al, 2003).

A segment of the data file is presented in Table 1.

**Table 1 Sample of Data Set**

| ID | C8932056_10 | C15862743_10 | C8932053_10 | C8932052_10 | asthma 1,2 | asthma 4 | … | chronic | no symptom |
|---|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | 1 | | | | |
| 2 | CC | CC | CC | GG | 1 | 1 | | 0 | 1 |
| …… | | | | | | | | | |
| 1535 | - | CC | CC | GG | 1 | 1 | | 0 | 1 |
| 1536 | CT | - | - | - | 2 | 1 | | 0 | 0 |

Among the total of 1,536 children, 625 children have complete information about SNPs *hCV8932056-10, hCV15862743-10, hCV8932053-10,* and *hCV8932052-10.* There are two indices under each phenotype with "1" indicating positive result (case) and "0" negative result (control). The population that has complete information about the four SNPs, including cases and controls, is used to test Hardy-Weinberg equilibrium and linkage disequilibrium. The population that has complete information about the most informative SNP pair *hCV8932056-10* and *hCV8932052-10,* but including cases only, is estimated for its haplotype frequency on the basis of phenotypes.

# CHAPTER 2 HARDY-WEINBERG EQUILIBRIUM

## Section 1 Introduction

In a large population, in the absence of natural selection, mutation, or migration, when the mating type frequencies arise from random mating, the ratios of the different genotypes follow a mathematical result established independently by the English mathematician Hardy and the German physician Weinberg. This phenomenon is named "Hardy-Weinberg Equilibrium". We need to examine all polymorphisms for Hardy-Weinberg equilibrium in order to produce valid and significant results in allelic association studies.

Consider a biallelic locus with alleles $A_1$ and $A_2$. Let the relative frequencies of the three genotypes $A_1A_1$, $A_1A_2$ and $A_2A_2$ in a large population be $g_{11}$, $2g_{12}$, and $g_{22}$ such that $g_{11} + 2g_{12} + g_{22} = 1$. Hardy's result was that, if individuals in the population mated with each other at random, these relative frequencies would be such that

$$g_{12}^2 = g_{11}g_{22}.$$

Moreover, if these offspring are mated at random, then the relative frequencies of the genotypes will remain unchanged after a second generation. It follows that these relative frequencies will continue to be in the Hardy-Weinberg ratio as long as mating in the population are random with respect to the locus.

The Hardy-Weinberg equilibrium enables us to relate genotype frequencies to allele frequencies. If the relative frequencies of alleles $A_1$ and $A_2$ are $f_1$ and $f_2$, respectively, then under normal conditions, the relative frequencies of gametes with alleles $A_1$ and $A_2$ will be $f_1$ and $f_2$. According to the Hardy-Weinberg law, under random mating, the relative frequencies of genotype $A_1A_1$, $A_1A_2$ and $A_2A_2$ are $g_{11}=f_1^2$, $g_{12}=2f_1f_2$, and $g_{22}=f_2^2$, regardless of the genotype frequencies in the parental generation. Since the allele frequencies corresponding to these genotype frequencies remain unchanged at $f_1$ and $f_2$, the same genotype frequencies will be maintained in subsequent generations as long as random mating applies in the population.

The same principle can be applied to a locus with more than two alleles. Let the alleles be $A_1, A_2, \ldots, A_m$, with relative frequencies $f_1, f_2, \ldots, f_m$. Under Hardy-Weinberg equilibrium, the relative frequency of a homozygous genotypes such as $A_iA_i$ is $g_{ii} = f_i^2$, and the relative frequency of a heterozygous genotype such as $A_iA_j$ $( i \neq j )$ is $g_{ij} = 2f_if_j$ (Sham, 1998).

Hardy-Weinberg equilibrium has many important applications. The demonstration of the ratio provides strong evidence for a genetic basis for a trait.

## Section 2 Exact Test of Hardy-Weinberg Equilibrium

Due to the importance of the Hardy-Weinberg law in the development of population genetics, testing of the null hypothesis that a population exhibits Hardy-Weinberg equilibrium has drawn a lot of attention during the past decades.

The methods proposed to test Hardy-Weinberg equilibrium can be categorized into two groups. One consists of large-sample goodness-of-fit tests that lean heavily on asymptotic results. However, it has been recognized that such tests can sometimes lead to false rejection or acceptance of Hardy-Weinberg equilibrium when the sample sizes are small and/or some cell frequencies are small or zero. The other approach involves the exact test, which is preferred when the sample size is small and/or some cell frequencies are small or zero (Guo et al, 1992). The exact test is named so because it does not rely on approximations. The Arlequin software package that we used to test Hardy-Weinberg equilibrium for the Isle of Wight birth cohort, implements the exact test. The exact test for Hardy-Weinberg equilibrium for multiple alleles is discussed below.

Consider an autosomal locus that has m alleles $A_1, A_2, \ldots, A_m$. A sample of size $n$ is sampled from a population of interest. Let $c_{ij}$ $(1 \leq j \leq i \leq m)$ denote the observed count of genotype $A_iA_j$. Then the data can be presented as the contingency table

| | | | | |
|---|---|---|---|---|
| $A_1$ | $c_{11}$ | | | |
| $A_2$ | $c_{21}$ | $c_{22}$ | | |
| ... | ... | ... | ... | |
| $A_m$ | $c_{m1}$ | $c_{m2}$ | ... | $c_{mm}$ |
| | $A_1$ | $A_2$ | ... | $A_m$ |

Use $c = (c_{11}, c_{21}, c_{22}, ..., c_{mm})$ to designate this table. Let $c_i$ denote the number of $A_i$ alleles in the sample. Then $c_i = c_{ii} + \sum_{j=1}^{m} c_{ij}$ (where $c_{ij} = c_{ji}$ if $j>i$).

The table of the count of genotype $A_iA_j$ is a random variable. Let $C$ denote this random variable. Then under random mating and the constraint that the number of allele $A_i$ remains unchanged from generation to generation, the distribution of $C$ satisfies multivariate hypergeometric distribution. Thus, the probability of obtaining the sample $c$ is:

$$\Pr(C = c) = \frac{n! \prod_{i=1}^{m} c_i!}{(2n)! \prod_{j>i} c_{ij}!} 2^{\sum\limits_{j>i} c_{ij}} ,$$

where $\sum\limits_{j>i} c_{ij}$ is the number of heterozygous individuals.

The exact test for the Hardy-Weinberg equilibrium given observed sample $c$ has to evaluate

$$P = \sum_{c' \in T} \Pr(c'),$$

where $T = \left\{ c' : \Pr(c') \leq \Pr(c), and \{c_i^{'}\} = \{c_i\}, wherec_i^{'} = \# alleleA_i \right\}$ (Guo et al, 1992).

In order words, the $P$-value of the test is the sum of the probabilities of the tables that have a probability smaller than or equal to the observed contingency table $c$ and have the same allele counts as does $c$. Rejection or acceptance of the null hypothesis depends on whether $P$ is smaller than a pre-specified significance level $\alpha$. A large $P$-value means that the probability of obtaining a sample as extreme or more extreme than the actually observed is large, thus it suggests Hardy-Weinberg equilibrium (Schneider, 2004).

Since there may be innumerable contingency tables having identical marginal counts, simply enumerating such tables is unrealistic. The Arlequin software modifies the Metropolis algorithm (Guo et al, 1992) to construct a Markov chain of contingency tables that have the same allelic counts (the same marginal counts) as the observed table, and a limiting distribution matching Pr(*C*). This approach starts with the observed contingency table. In order to create a new contingency table from an existing one, we randomly select two distinct rows $i_1$, $i_2$ and two distinct columns $j_1, j_2$. Neither the two rows nor the two columns have to be next to each other. The new table is obtained by decreasing the counts of the cells $(i_1, j_1)$ $(i_2, j_2)$ and increasing the counts of the cells $(i_1, j_2)$ $(i_2, j_1)$ by one unit. This leaves the alleles counts $\{c_{ij}\}$ unchanged. For example, by decreasing the counts of the cells $(c_{31}, c_{43})$ and increasing the counts of the cells $(c_{33}, c_{41})$ by one unit in the original sample table *c*, we obtain a candidate for the new table. Note that the marginal counts of rows $c_3$ and $c_4$ and of columns $c_1$ and $c_3$ remain the same as those in table *c*.

| | | | | | | |
|---|---|---|---|---|---|---|
| $A_1$ | $c_{11}$ | | | | | |
| $A_2$ | $c_{21}$ | $c_{22}$ | | | | |
| $A_3$ | $\mathbf{c_{31}-1}$ | $c_{32}$ | $\mathbf{c_{33}+1}$ | | | |
| $A_4$ | $\mathbf{c_{41}+1}$ | $c_{42}$ | $\mathbf{c_{43}-1}$ | $c_{44}$ | | |
| ... | ... | ... | ... | ... | ... | |
| $A_m$ | $c_{m1}$ | $c_{m2}$ | $c_{m3}$ | $c_{m4}$ | ... | $c_{mm}$ |
| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | ... | $A_m$ |

Whether to accept this candidate table or not depends on a probability *R* equal to:

1. $R = \dfrac{\Pr(c^{(k+1)})}{\Pr(c^{(k)})} = \dfrac{c_{i_1 j_1} c_{i_2 j_2}}{(c_{i_1 j_2} + 1)(c_{i_2 j_1} + 1)} \dfrac{(1+\delta_{i_1 j_1})(1+\delta_{i_2 j_2})}{(1+\delta_{i_1 j_2})(1+\delta_{i_2 j_1})}$, if $i_1 \neq j_1$ or $i_2 \neq j_2$

2. $R = \dfrac{\Pr(c^{(k+1)})}{\Pr(c^{(k)})} = \dfrac{c_{i_1 j_1} c_{i_2 j_2}}{(c_{i_1 j_2} + 1)(c_{i_2 j_1} + 2)} \dfrac{4}{1}$, if $i_1 = j_1$ and $i_2 = j_2$

3. $R = \dfrac{\Pr(c^{(k+1)})}{\Pr(c^{(k)})} = \dfrac{c_{i_1 j_1} (c_{i_2 j_2} - 1)}{(c_{i_1 j_2} + 1)(c_{i_2 j_1} + 1)} \dfrac{1}{4}$, if $i_1 = j_2$ and $i_2 = j_1$

where $k$ is the created table number, and $\delta_{ij} = 0$ if $i \neq j$, 1 if $i = j$. $R$ is the ratio of the probabilities of the two tables.

The switch to the new table is accepted if $R$ is larger than 1. The resulted Markov chain has a limiting probability distribution that is the same as the distribution of the contingency tables, $\Pr(C = c) = \dfrac{n! \prod_{i=1}^{m} c_i!}{(2n)! \prod_{j>i} c_{ij}!} 2^{\sum_{j>i} c_{ij}}$.

In practice, the Markov chain starting from the observed contingency table is biased. In order for the chain to be unbiased, it is ideal to start the chain from a random table chosen from the distribution Pr($C$). One solution is to start the chain from the observed table and run for a long time so that the initial table is “forgotten”. This process is referred to as “dememorization”.

After the dememorization, the Markov chain is constructed for a long time, which results in a large number of selected tables. The limiting probability of the Markov chain in a particular table can be interpreted as the long-run proportion of time that the chain stays in that table. Therefore, following the computation of $P$-value in the exact test, the

*P*-value of the test is the proportion of the selected tables that have a probability smaller than or equal to the observed contingency table.

## Section 3 Arlequin Implementation of Exact Test and Case Study

The Arlequin software package is employed to test Hardy-Weinberg equilibrium for the Isle of Wight birth cohort Study. The goal of Arlequin is to provide the users in population genetics with a large set of methods and statistical tests, in order to extract information on genetic and demographic features of a collection of population samples (Schneider, 2004). The software is available on line at http://lgb.unige.ch/arlequin.

Arlequin input file with extension .arp contains both "Profile", descriptions of the properties of the data, as well as "Data", the raw data themselves. The Profile section specifies the title of the project, the number of samples present in the project, the type of data to be analyzed, if the project deals with haplotypic or genotypic data, if the gametic phase of genotypes is known and so on. The sample of our project is defined as DNA multi-locus data with unknown gametic phase. The Data section includes a name of the sample, the size of sample, and the data itself.

A SAS program (Appendix A) has been written to obtain the information required by Data section for our project. The children with complete genotype information at SNPs *hCV8932056-10, hCV15862743-10, hCV8932053-10,* and *hCV8932052-10* are first selected. There are a total of 625 such children (Table 2). Next, the individual genotypes for the four loci are output on two separate lines according to Arlequin's instruction. For instance, sample #1532 has genotype pattern *TT// CC//CT// AG*. A combination of the first allele at each locus, namely *TCCA*, forms the first line for this individual; likely, *TCTG* forms the second line. It should be noted that the sequence of the loci must be kept in the order of their physical positions.

**Table 2 Sample of Data Set with Genotype Information on Four SNPs**

| ID | C8932056_10 | C15862743_10 | C8932053_10 | C8932052_10 | asthma 1,2 | asthma 4 | ... | chronic | no symptom |
|---|---|---|---|---|---|---|---|---|---|
| 2 | CC | CC | CC | GG | 1 | 1 | | 0 | 1 |
| 3 | CC | CC | CC | GG | 1 | 1 | | 0 | 0 |
| ...... | | | | | | | | | |
| 1530 | CC | CC | CC | GG | 1 | 1 | | 0 | 1 |
| 1532 | TT | CC | CT | AG | 2 | 1 | | 0 | 0 |

The input file (with partial dataset) of our project is displayed as follows:

```
[Profile]

   Title="Isle of Wight Asthma HWE test"
   NbSamples= 1
   DataType= DNA
   GenotypicData= 1
   LocusSeparator= WHITESPACE
   GameticPhase= 0
   RecessiveData= 0
   RecessiveAllele= null
   MissingData= '?'
   Frequency= ABS
   CompDistMatrix= 0
   FrequencyThreshold= 1.0e-5
   EpsilonValue= 1.0e-7

[Data]

[[Samples]]

         SampleName="Isle of Wight Asthma HWE test"
                 SampleSize=625
                 SampleData= {
            2 1    C C C G
                   C C C G
         ......

           1532 1  T C C A
                   T C T G
}
```

Once the project file is built, "Calculation Settings" of Arlequin interface is decided. In our project, Hardy-Weinberg equilibrium test with 100,000 steps in Markov chain, 1,000 dememorisation steps, and a significance level of 5% has been set. The number of steps in Markov chain sets the maximum number of alternative tables to explore. A figure of 100,000 is in order. The number of dememorization steps sets the number of steps to perform before beginning to compare the alternative table probabilities to that of the observed table. 1,000 steps are necessary to reach a random starting point corresponding to a table independent from the observed table.

The output of the test is presented below:

```
Hardy-Weinberg equilibrium: (Isle of Wight Asthma HWE test)
===============================
Exact test using a Markov chain (for all Loci):
Forecasted chain length    :100000
Dememorization steps       :1000

Locus  #Genot   Obs.Heter.   Exp.Heter.   P. value   s.d.      Steps done
  1     625     0.30880      0.58841      1.00000    0.00000   100172
  2     625     0.16640      0.49639      0.80856    0.00122   100172
  3     625     0.30720      0.58422      0.79268    0.00123   100172
  4     625     0.28480      0.58663      0.08930    0.00084   100172
```

As shown above, the *P*-values of the test at all loci are larger than 0.05, suggesting acceptance of null hypothesis; namely, all the loci are in Hardy-Weinberg equilibrium. Once Hardy-Weinberg equilibrium is established, we can proceed to test whether two loci are in linkage disequilibrium.

# CHAPTER 3 LINKAGE DISEQUILIBRIUM

## Section 1 Introduction

### *Genetic Linkage*

The aim of ***genetic linkage analysis*** is to infer the relative positions of two or more loci by examining the patterns of allele-transmission from parent to offspring, or the patterns of allele-sharing by relatives.

The genotype of an individual at two loci is formed by the haplotypes of two gametes, each inherited from one parent. A gamete may contain two alleles from the same parental gamete or one allele from each parental gamete. In the first case, the haplotype of the gamete is the same as the haplotype of one of the parental gametes. Such gametes are defined as ***non-recombinants***. In the second case, however, the haplotype of the gamete constitutes a new combination of alleles different from either parental haplotype. Such gametes are defined as ***recombinants***. The recombination fraction, usually denoted as $\theta$, between the two loci on the same chromosome is defined as the probability that a gamete is recombinant. Two loci with a recombination fraction of less than $1/2$ are said to be in ***linkage***. The closer the two loci, the smaller the recombination fraction is, and the more tightly linked are the two loci (Sham, 1998).

### *Allelic Association*

As a result of linkage, some combinations of alleles, i.e. haplotypes, on short chromosomal segments may be preserved over a large number of generations and become quite frequent in the population. The excessive co-occurrence of certain combinations of alleles in the same gamete because of tight linkage is known as ***allelic association***.

Consider two loci $A$ and $B$, with alleles $A_1, A_2, \ldots, A_m$, and $B_1, B_2, \ldots, B_n$. By definition, if the occurrence of allele $A_i$ and allele $B_j$ in a haplotype are independent events, then the relative frequency of the haplotype $A_i B_j$ is equal to the product of the allele frequencies of $A_i$ and $B_j$, i.e. $h_{ij} = f_i(A) f_j(B)$. If $h_{ij} \neq f_i(A) f_j(B)$, then the occurrences of $A_i$ and $B_j$ are not independent, and the two alleles are said to be ***associated*** (Sham, 1998).

***Maintenance of Allelic Associations: Linkage Disequilibrium***

In a large, closed, randomly mating population, let the relative frequency of the haplotype $A_i B_j$ in the current generation be $h_{ij0}$. In the next generation, if the haplotype is a recombinant, then the probability that it is $A_i B_j$ is

$$\Pr(A_i B_j) = f_i(A) f_j(B);$$

If the haplotype is a non-recombinant, then the probability that it is $A_i B_j$ is simply

$$\Pr(A_i B_j) = h_{ij0}.$$

The total probability that a haplotype transmitted to the next generation is $A_i B_j$ is therefore

$$h_{ij1} = \Pr(A_i B_j) = \theta f_i(A) f_j(B) + (1-\theta) h_{ij0}.$$

The change in haplotype frequency from generation 0 to generation 1 is

$$h_{ij1} - h_{ij0} = \theta(f_i(A) f_j(B) - h_{ij0}).$$

When $f_i(A) f_j(B) = h_{ij0}$ for all $i, j$ at the two loci, i.e. if there is no allelic association, there will be no change in haplotype frequencies from generation to generation, and we say that the two loci are in ***linkage equilibrium***. Otherwise the two loci are said to be in ***linkage disequilibrium***.

For most human populations and for most regions of the genome, substantial linkage disequilibrium is only likely to occur between loci with a recombination fraction of less than 1%. This is the rationale behind the use of association analysis as an important tool for mapping susceptibility loci (Sham, 1998).

## Section 2 Likelihood Ratio Test of Linkage Disequilibrium

For genotypic data where the haplotypic phase is unknown, linkage disequilibrium between a pair of loci is tested for genotypic data using a likelihood-ratio test. The likelihood ratio is between the likelihood of the data assuming linkage equilibrium (denoted $L_0$) and the likelihood of the data not assuming linkage equilibrium (denoted $L_1$). $L_0$ is computed by using the fact that, under the hypothesis of linkage equilibrium, the haplotype frequencies are obtained as the product of the allele frequencies. $L_1$ is obtained by applying the Expectation-Maximum algorithm to estimate haplotype frequencies. Both $L_0$ and $L_1$ assume Hardy-Weinberg equilibrium (random mating). The ratio of $L_0$ and $L_1$ suggests the degree of deviation from linkage equilibrium (Schneider et al, 2000).

Suppose that two loci $A$ and $B$, with distinctive alleles $A_1, A_2, ..., A_m$ and $B_1, B_2, ..., B_n$, have been genotyped in a random sample of the population. Each individual has a genotype of the form $A_i A_j \, // \, B_k B_l$. There are $m(m+1)/2$ possible genotypes at locus $A$, and $n(n+1)/2$ possible genotypes at locus $B$, so that the total number of joint genotypes is $(m(m+1)/2)(n(n+1)/2)$. The aim is to test the null hypothesis of linkage equilibrium between $A$ and $B$.

### *Computation of $L_0$*

Under the assumption of linkage equilibrium, the relative frequency of a haplotype is equal to the products of the allele frequencies:

$$h_{ik}=f_i(A)f_k(B)$$

$$h_{il}=f_i(A)f_l(B)$$

$$h_{jk}=f_j(A)f_k(B)$$

$$h_{jl}=f_j(A)f_l(B),$$

where the allele frequencies can be obtained by simple counting on the basis of given genotype information.

Under the assumption of random mating, the genotype frequencies can be expressed as follows:

$$g_{iikk} = h_{ik}h_{ik} = h_{ik}{}^2 = f_i(A)^2 f_k(B)^2$$

$$g_{iikl} = h_{ik}h_{il} + h_{il}h_{ik} = 2h_{il}h_{ik} = 2f_i(A)^2 f_l(B)\, f_k(B)$$

$$g_{ijkk} = h_{ik}h_{jk} + h_{jk}h_{ik} = 2h_{ik}h_{jk} = 2f_i(A)f_j(A)f_k(B)^2$$

$$g_{ijkl} = h_{ik}h_{jl} + h_{jl}h_{ik} + h_{il}h_{jk} + h_{jk}h_{il} = 4f_i(A)f_j(A)f_k(B)f_l(B).$$

The observed genotypic counts $c_{1111}$, $c_{1112}$ , ..., $c_{mmnn}$ follow a multinomial distribution with parameters $c_{....}$ (the sample size) and the genotypic frequencies $g_{1111}$, $g_{1112}$, ..., $g_{mmnn}$. The log-likelihood of observing $c_{ijkl}$ is therefore

$$\ln L_0 = \sum_{\substack{i,j=1,\ldots,m \\ k,l=1,\ldots,n}} c_{ijkl} \ln(g_{ijkl}),$$

where the summation is taken over all $(m(m+1)/2)(n(n+1)/2)$ possible genotypes. The maximum likelihood estimate of a population genotype frequency is simply the sample genotype frequency (Sham, 1998).

***Computation of $L_1$***

When not assuming linkage equilibrium, in the case that an individual is heterozygous for both loci, the haplotypes cannot be deduced from the genotype. For example, the genotype $A_iA_j \,//\, B_kB_l$ can be made up of the haplotype pairs $A_iB_k \,/\, A_jB_l$ or $A_jB_k \,/\, A_iB_l$. In order to estimate the haplotype frequencies, we use the criterion of maximizing the likelihood of observing genotype data, $L_1$. From the discussions in the preceding subsection, we know

$$\ln L_1 = \sum_{\substack{i,j=1,\ldots,m \\ k,l=1,\ldots,n}} c_{ijkl} \ln(g_{ijkl}),$$

where $c_{ijkl}$ are the observed genotypic counts, $g_{ijkl}$ are the estimated genotype frequencies, which are calculated from the estimated haplotype frequencies under the assumption of random mating as follows:

$$g_{iikk} = h_{ik} h_{ik} = h_{ik}{}^2$$

$$g_{iikl} = h_{ik} h_{il} + h_{il} h_{ik} = 2h_{il} h_{ik}$$

$$g_{ijkk} = h_{ik} h_{jk} + h_{jk} h_{ik} = 2h_{ik} h_{jk}$$

$$g_{ijkl} = h_{ik} h_{jl} + h_{jl} h_{ik} + h_{il} h_{jk} + h_{jk} h_{il} = 2(h_{ik} h_{jl} + h_{jk} h_{il}).$$

In the following, we use an iterative method of counting based on the Expectation-Maximum algorithm to obtain the maximum likelihood estimation of haplotype frequencies.

The EM algorithm is a numerical method of finding maximum likelihood estimates for parameters given incomplete data. It begins by setting the initial haplotype frequencies as $h_{uv,0}$, $u = 1,\ldots,m, v = 1,\ldots,n$. It is reasonable to set these initial haplotype frequencies as the products of the allele frequencies, just as the haplotype frequencies under the assumption of Hardy-Weinberg equilibrium and linkage equilibrium.

The count of genotype $A_i A_j \,//\, B_k B_l$ can be considered as the sum of 2 unobserved counts: $c_{ijkl} = c_{ik,jl} + c_{il,jk}$, where $c_{ijkl}$ is the count of genotype $A_i A_j \,//\, B_k B_l$, $c_{ik,jl}$ is the number of individuals with haplotype pair $A_i B_k \,/\, A_j B_l$, and

$c_{il,jk}$ is the number of individuals with haplotype pair $A_iB_l\,/\,A_jB_k$. For any heterozygous genotype $A_iA_j\,//\,B_kB_l$, the initial expected values of the unobserved counts of haplotype pairs are calculated as follows:

Under Hardy-Weinberg equilibrium, $g_{ik,jl} = h_{ik}h_{jl}$ and $g_{il,jk} = h_{il}h_{jk}$, where $g_{ik,jl}$ and $g_{il,jk}$ are the relative frequencies of the genotypes obtained from the haplotype pairs $A_iB_k\,/\,A_jB_l$ and $A_iB_l\,/\,A_jB_k$ respectively. Therefore, the fraction of genotype $A_iA_j\,//\,B_kB_l$ that is obtained by the haplotype pair $A_iB_k\,/\,A_jB_l$ is $\dfrac{h_{ik,0}h_{jl,0}}{h_{ik,0}h_{jl,0}+h_{il,0}h_{jk,0}}$, and the count of haplotype pair $A_iB_k\,/\,A_jB_l$ is

$$c_{ik,jl,0} = \frac{c_{ijkl}(h_{ik,0}h_{jl,0})}{h_{ik,0}h_{jl,0}+h_{il,0}h_{jk,0}}.$$

Similarly, the count of haplotype pair $A_iB_l\,/\,A_jB_k$ is

$$c_{il,jk,0} = \frac{c_{ijkl}(h_{il,0}h_{jk,0})}{h_{ik,0}h_{jl,0}+h_{il,0}h_{jk,0}}.$$

This step of computing the expected counts of haplotype pairs is called expectation step.

Using the initial expected values as real data, we can compute the number of heterozygous haplotypes and get a set of revised haplotype counts as follows:

$$c_{pq,1} = \sum_{\substack{s=1,\ldots,m\\ t=1,\ldots,n}} c_{pq,st,0}\,,$$

where $p = 1,\ldots,m, q = 1,\ldots,n$. The revised haplotype frequencies are, therefore,

$$h_{uv,1} = \frac{c_{uv,1}}{\sum\limits_{\substack{u=1,\ldots,,m \\ v=1,\ldots,n}} c_{uv,1}},$$

where $u = 1,\ldots,m, v = 1,\ldots,n$.

This step of reevaluating the haplotype frequencies from the relative probabilities of the possible haplotype pairs is called maximization step. The maximum likelihood estimate of population haplotype frequency is simply the sample haplotype frequency.

In the next iteration, this set of revised haplotype frequencies are used to obtain a set of revised expected values of the unobserved counts, which are denoted as $c_{ik,jl,1}$ and $c_{il,jk,1}$, etc.

$$c_{ik,jl,1} = \frac{c_{ijkl}(h_{ik,1}h_{jl,1})}{h_{ik,1}h_{jl,1} + h_{il,1}h_{jk,1}}$$

$$c_{il,jk,1} = \frac{c_{ijkl}(h_{il,1}h_{jk,1})}{h_{ik,1}h_{jl,1} + h_{il,1}h_{jk,1}}$$

The cycle of revising the haplotype frequencies, revising the expected values of the unobserved counts, and counting the haplotypes is repeated until the changes in haplotype frequencies from one iteration to the next become negligible, i.e. convergence is reached. These are then the local maximum likelihood estimates of the haplotype frequencies (Sham, 1998).

While the EM algorithm guarantees convergence, it is not guaranteed to converge to the global maximum when there are multiple local maximums. To increase the chance of obtaining the global maximum, it is best to try numerous initial values for the haplotype frequencies (Long et al, 1995). In Arlequin, initial values of 100 and more are in order (Schneider et al, 2004). The set of haplotype frequencies with the highest local maximum likelihood is then used as the final estimation.

After the haplotype frequencies are estimated, the genotype frequencies and ln$L_1$ are computed in the same way as in the computation of $L_0$.

***Likelihood-ratio test***

$\ln L_0$ and $\ln L_1$ have $(m-1)+(n-1)$ and $mn-1$ estimated parameters, respectively. The likelihood ratio statistic given by

$$S = -2(\ln L_0 - \ln L_1)$$

has an asymptotical Chi-square distribution with $(mn-1)-((m-1)+(n-1)) =$ $(m-1)(n-1)$ degrees of freedom. A statistically significant $P$-value suggests rejection of null hypothesis.

In the case of small samples with large number of alleles per locus, the Chi-square distribution does not apply to the likelihood ratio distribution. In order to better approximate the underlying distribution of the likelihood ratio statistic, we perform a randomization test. Such test is non-parametric, not based on asymptotic approximation, and applicable to context with few data sets. The procedure is as follows:

1. Permute the alleles between individuals at one locus only.
2. Re-estimate the likelihood of the data $L_1$ by the EM algorithm. $L_0$ is unaffected by the permutation procedure.
3. Repeat steps 1-2 a large number of times to get the null distribution of $L_1$, and therefore the null distribution of $S$.

The $P$-value is calculated by

$$P - value = \frac{\sum_{n(s_i) \le n(s_1)} n(S_i)}{\sum n(S_i)} ,$$

where $S_1$ is the log likelihood ratio of the observed data (Schneider, 2004).

## Section 3 Arlequin Implementation of Likelihood Ratio Test and Case Study

The Arlequin software package is used to test linkage disequilibrium for the Isle of Wight birth cohort. Exactly the same project file is prepared as the test of Hardy-Weinberg equilibrium. Under "Calculation Settings", pair-wise linkage disequilibrium test with 16,000 permutations, 100 initial conditions, and a significance level of 5% has been set. Number of permutations sets the number of random permuted samples to generate. 16,000 permutations guarantee to have less than 1% difference with the exact probability in 99% of the cases. Number of initial conditions sets the number of random initial conditions from which the EM is started to repeatedly estimate the sample likelihood. The haplotype frequencies globally maximizing the sample likelihood will be eventually kept. In our project, 100 initial conditions are used.

The output of the test is presented below:

```
Pairwise linkage disequilibrium: (Isle of Wight Asthma LD test)
==============================

Permutation test using the EM algorithm
Number of permutations: 16000
Number of initial conditions for EM: 100

Pair(0, 1)
      LnLHood LD : -788.23438         LnLHood LE : -790.29395
      Exact P=  0.04306  +-   0.00149 (16002 permutations done)
      Chi-square test value= 4.11914 (P =  0.04240,  1 d.f.)
Pair(0, 2)
      LnLHood LD : -839.46460         LnLHood LE : -942.81079
      Exact P=  0.00000  +-   0.00000 (16002 permutations done)
      Chi-square test value=206.69238 (P =  0.00000,  1 d.f.)
```

```
Pair(1, 2)
        LnLHood LD : -761.75385         LnLHood LE : -780.76373
        Exact P=  0.00000  +-   0.00000 (16002 permutations done)
        Chi-square test value=38.01978 (P =  0.00000,  1 d.f.)
Pair(0, 3)
        LnLHood LD : -861.22888         LnLHood LE : -958.38812
        Exact P=  0.00000  +-   0.00000 (16002 permutations done)
        Chi-square test value=194.31848 (P =  0.00000,  1 d.f.)
Pair(1, 3)
        LnLHood LD : -777.89398         LnLHood LE : -796.34106
        Exact P=  0.00000  +-   0.00000 (16002 permutations done)
        Chi-square test value=36.89417 (P =  0.00000,  1 d.f.)

Pair(2, 3)
        LnLHood LD : -564.50177         LnLHood LE : -948.85785
        Exact P=  0.00000  +-   0.00000 (16002 permutations done)
        Chi-square test value=768.71216 (P =  0.00000,  1 d.f.)

Table of significant linkage disequilibrium (significance level=0.0500):

Locus  # |  0|  1|  2|  3|
        0 |  *   +   +   +
        1 |  +   *   +   +
        2 |  +   +   *   +
        3 |  +   +   +   *
```

Loci 0, 1, 2, and 3 refer to SNPs *hCV8932056-10, hCV15862743-10, hCV8932053-10,* and *hCV8932052-10* respectively.

As shown above, the $P$-values of all the pair-wise tests are less than 0.05, suggesting rejection of null hypothesis; namely, all the pairs of loci are in linkage disequilibrium.

# CHAPTER 4 HAPLOTYPE FREQUENCY ESTIMATION

## Section 1 Introduction

Haplotype information is an essential ingredient in many analyses of fine-scale molecular genetics data. For example, haplotype analysis is an important tool for linkage disequilibrium assessment, disease-gene discovery, genetic demography, and chromosomal-evolution studies. However, many haplotype analysis methods rely on phase information from the individuals under study. As mentioned in Section 2 of Chapter 3, for autosomal loci, when only the multi-locus genotypes for each individual are provided, the phase information for those individuals with multiple heterozygous phenotypes is inherently ambiguous. Phase can be established by genotyping family members of each study subject to infer parental chromosomes, but this requires recruitment and genotyping of relatives, which is expensive and may not be realistic. Laboratory techniques have also been used to determine haplotypes, but these approaches are technologically demanding and often cost-prohibitive. Alternatively, statistical methods can be used to infer phase at linked loci from genotypes and to estimate frequency of all possible haplotypes (Fallin et al, 2000).

The problem of unknown phase and one of its possible solutions can be explained in the example of Clark's algorithm. In Clark's algorithm, when a homozygote is found, a haplotype is unambiguously identified. When a single-locus heterozygote is found, a possible haplotype pair is inferred. For each of the remaining multi-locus heterozygotes, we need to determine whether it can produce a haplotype that has been established. If it can, identify the complementary haplotype by using the established haplotype as one of the actual haplotypes that it implies. Continue this process until the phase information for all individuals is either resolved or identified as unresolved. This algorithm is intuitively appealing and effective in resolving haplotypes when the dataset contains a sufficient number of homozygous individuals. It also performs well for relatively small sample sizes. However, three problems can arise with this procedure. It may not be possible to start the iterative algorithm if there is no unambiguous or single-locus heterozygous individuals in the sample. There may be unresolved haplotypes left at the end. In addition, haplotypes may be erroneously inferred if a crossover product of two actual haplotypes is identical to another true haplotype (Clark, 1990).

Compared with Clark's algorithm, the Bayesian method is more accurate in inferring haplotype information and can handle more loci. For the Isle of Wight birth cohort, we estimate the haplotype frequencies using the PHASE software package that implements the Bayesian algorithm. The Bayesian algorithm will be introduced in the following section.

## Section 2 Bayesian Estimation Method for Haplotype Frequency

Suppose there is a sample of *n* diploid individuals from a population. Let $G = (G_1,...,G_n)$ denote the (known) genotypes for the individuals, and let $H = (H_1,...,H_n)$ denote the (unknown) corresponding haplotype pairs.

The Bayesian algorithm regards the unknown haplotypes as unobserved random quantities and aims to evaluate their conditional distribution given the genotype data. Gibbs sampling, a type of Bayesian approach, is used to obtain an approximate sample from the posterior distribution of *H* given *G*, $\Pr(H \mid G)$. Informally, the algorithm starts with an initial guess $H^{(0)}$ for *H*, repeatedly chooses an individual at random, and estimates that individual's haplotypes under the assumption that all the other haplotypes are correctly reconstructed. Repeating this process enough times results in an approximate sample from $\Pr(H \mid G)$. Formally, this method involves constructing a Markov chain $H^{(0)}$, $H^{(1)}$, $H^{(2)}$, ..., with stationary distribution $\Pr(H \mid G)$, on the space of possible haplotype reconstructions, using an algorithm of the following form.

Start with some initial haplotype reconstruction $H^{(0)}$. For $t = 0, 1, 2, ...$, obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:

1. Choose an individual, *i*, uniformly and at random from all ambiguous individuals (i.e., individuals who are heterozygous at more than one loci).
2. Sample $H_i^{(t+1)}$ from $\Pr(H_i \mid G, H_{-i}^{(t)})$, where $H_{-i}$ is the set of haplotypes excluding individual *i*. The conditional probability $\Pr(H_i \mid G, H_{-i}^{(t)})$ depends on the genetic and demographic models.
3. Set $H_j^{(t+1)} = H_j^{(t)}$ for $j = 1, ..., n, j \neq i$.

It has been proved that this process produces a Markov chain (Gilks et al, 1996).

To help illustrate Gibbs sampling, let us consider a simple example. Suppose that a random sample of two individuals from a population have been genotyped for two loci $A$ and $B$. We have $G = \{G_1, G_2\}$, where $G_1 = \{A_1A_2 // B_1B_2\}$ and $G_2 = \{A_3A_4 // B_3B_4\}$. The corresponding haplotype pair $H_1$ is either $\{A_1B_1 / A_2B_2\}$ or $\{A_1B_2 / A_2B_1\}$. Similarly, the corresponding haplotype pair $H_2$ is either $\{A_3B_3 / A_4B_4\}$ or $\{A_3B_4 / A_4B_3\}$. The purpose is to find the distribution of $H_1$ and $H_2$ conditional on $G$, i.e.

$$\Pr(H_1 = \{A_1B_1 / A_2B_2\} \mid G) \text{ and } \Pr(H_1 = \{A_1B_2 / A_2B_1\} \mid G)$$

and

$$\Pr(H_2 = \{A_3B_3 / A_4B_4\} \mid G) \text{ and } \Pr(H_2 = \{A_3B_4 / A_4B_3\} \mid G).$$

From a certain genetic model we know the conditional probabilities $\Pr(H_1 \mid G, H_{-1})$ and $\Pr(H_2 \mid G, H_{-2})$, in this case, i.e. $\Pr(H_1 \mid G, H_2)$and $\Pr(H_2 \mid G, H_1)$.

Starting with an initial guess $H^{(0)}$ for $H$ , say

$$H^{(0)} = (H_1^{(0)}, H_2^{(0)}) = (\{A_1B_2 / A_2B_1\}, \{A_3B_3 / A_4B_4\}),$$

we obtain $H^{(1)}$ from $H^{(0)}$ as follows:

1. Choose an individual uniformly and at random from $G$. Say we get individual #1, i.e. $H_1^{(0)}$.

2. Since the genetic model gives $\Pr(H_1 \mid G, H_2^{(0)})$ and we have $H_2^{(0)} = \{A_3B_3 / A_4B_4\}$, thus we know the distribution of $H_1$ conditional on $H_2^{(0)}$, say

$$\Pr(H_1 = \{A_1B_2 / A_2B_1\} \mid G, H_2^{(0)} = \{A_3B_3 / A_4B_4\}) = 0.2$$

and

$$\Pr(H_1 = \{A_1B_1 / A_2B_2\} \mid G, H_2^{(0)} = \{A_3B_3 / A_4B_4\}) = 0.8 .$$

Randomly drawing $H_1^{(1)}$ on the basis of this distribution, we obtain, say, $H_1^{(1)} = A_1B_1 / A_2B_2$.

3. Set $H_2^{(1)} = H_2^{(0)}$.

So far we obtain a new sample $H^{(1)} = (H_1^{(1)}, H_2^{(1)}) = (\{A_1B_1 / A_2B_2\}, \{A_3B_3 / A_4B_4\})$.

Next, by repeating the above steps, we obtain $H^{(2)}$ from $H^{(1)}$:

1. Choose an individual uniformly and at random from $G$. Say we get individual #2 this time.

2. Since the genetic model gives $\Pr(H_2 \mid G, H_1^1)$, we know the distribution of $H_2$ conditioned on $H_1^{(1)}$, say

$$\Pr(H_2 = \{A_3B_4 / A_4B_3\} \mid G, H_1^{(1)} = \{A_1B_2 / A_2B_1\}) = 0.4$$

and

$$\Pr(H_2 = \{A_3B_3 / A_4B_4\} \mid G, H_1^{(1)} = \{A_1B_2 / A_2B_1\}) = 0.6 .$$

Again, we randomly draw $H_2^{(2)}$ on the basis of this distribution, say we obtain

$H_2^{(2)} = A_3B_4 / A_4B_3$.

3. Set $H_1^{(2)} = H_1^{(1)}$.

Now we obtain another sample $H^{(2)} = (H_1^{(2)}, H_2^{(2)}) = (\{A_1B_2 / A_2B_1\}, \{A_3B_4 / A_4B_3\})$.

Let this process continue for a large number of times, say 10,000 times. We obtain a sample of 10,000 $H's$, from which the distribution of $H_1$ and $H_2$ conditional on $G$, $\Pr(H_1 \mid G)$ and $\Pr(H_2 \mid G)$ can be calculated.

As shown by Stephens et al., the Bayesian method has three major advantages over Clark's method and the EM algorithm: increased accuracy, wider applicability (for instance, it can handle a large number of loci), and the facility to assess accurately the uncertainty associated with each phase call (Stephens et al, 2001).

## Section 3 PHASE Implementation of Bayesian Estimation and Case Study

The PHASE software package is used to estimate the haplotype frequencies for the children of Isle of Wight. PHASE implements the Bayesian statistical method for reconstructing haplotypes and estimating haplotype frequencies from genotype data, and for estimating recombination rates and identifying recombination hotspots (Stephens et al, 2003). The program is available on line at http://www.stat.washington.edu/stephens/software.html.

We want to choose the SNP pairs that are in linkage equilibrium to conduct haplotype analysis. Although all of the SNP pairs are in linkage disequilibrium, since the combination of SNPs *hCV8932056* and *hCV8932052* is found to be most informative by the preliminary study, we choose these two SNPs to perform haplotype frequency estimation on the basis of eight phenotypes: asthma at 1 or 2 years, asthma at 4 years, currently diagnosed asthma at 10 years, wheezing at 1 or 2 years, wheeze at 4 years, wheeze at 10 years, chronic asthma, and no symptom at all ages.

PHASE input file specifies how many individuals there are to be analyzed, how many loci each individual has been typed at, what physical positions the loci are, what sort of loci these are (SNP or microsatellite), and the ID and the genotypes for each individual.

A SAS program (Appendix B) has been written to obtain the information required by PHASE input file. The children who have complete information on SNPs *hCV8932056* and *hCV8932052* and are positive (i.e. "2" if the phenotype has indices "1" and "2", and "1" if the phenotype has indices "0" and "1") within each phenotype are first selected. There are a total of 164 such children in the group "asthma at 1 or 2 years", 99 in "asthma at 4 years", 92 in "currently diagnosed asthma at 10 years", 48 in "wheeze at 1 or 2 years", 145 in "wheeze at 4 years", 228 in "wheeze at 10 years", 34 in "chronic asthma", and 308 in "no symptoms at all age". Next, the individual genotypes for the two loci are output on two separate lines according to PHASE's instruction. In fact, it is the same way as Arlequin. It should also be noted that the sequence of the loci must be kept in the order of their physical positions.

As an example, the input file (with partial raw data) for the phenotype group "asthma at 1 or 2 years" of our project is displayed as follows:

```
164
2
P 128122939 128126575
SS
8
CG
CG
16
CA
TG
...
1510
CG
CG
1524
CG
TG
```

This presentation says that there are 164 children (who have asthma at 1 or 2 years) typed at two loci (*hCV8932056* and *hCV8932052*) whose relative positions along the chromosome are *128122939* and *128126575*, and which are bi-allelic. The genotype information then follows, with the ID and two lines for each child. For example, the individual #8 has the genotype pattern *CC//GG.*

PHASE produces a number of output files: a summary report and additional reports whose suffixes indicate the contents of the file. "_freqs" estimates the sample haplotype frequencies; "_pairs" lists the most likely pairs of haplotypes for each individual, together with their probability; "_recom" contains estimates of recombination parameters across the region; and "_monitor" measures the goodness of fit of the estimated haplotypes to the underlying model. Among them the output file "_freq" directly pertains to the research purpose of the Isle of Wight Birth Cohort Study. The output files "_freq" of the eight groups of children are consolidated and presented in Table 3.

Figure 5 produced by a Matlab program (Appendix C) visualizes the haplotype frequency estimation in Table 3. In the graph, a12 stands for the phenotype of "asthma at 1 or 2 years", a4 for "asthma at 4 years", cda10 for "currently diagnosed asthma at 10 years", w12 for "wheeze at 1 or 2 years", w4 for "wheeze at 4 years", w10 for "wheeze at 10 years", chronic for "chronic asthma", and control for "control".
It can be clearly seen that haplotype *CG* is found in all of the eight phenotypes with high probabilities ranging from 0.663015 to 0.790994. The other haplotypes are also found in all the other phenotypes, but with very low probabilities.

In order to view better the relationship between each haplotype and phenotype, graphs of the relative frequencies of each haplotype vs. each phenotype are shown in Figure 6, which is produced by Matlab (Appendix C).

**Table 3 Haplotype Frequency Estimation for SNPs *hCV8932056* and *hCV8932052* by Phenotypes**

| Phenotypes | Haplotype | E (freq) | SE |
|---|---|---|---|
| Asthma at 1 or 2 years<br>(n=164) | CA | 0.080212 | 0.004531 |
| | CG | 0.706373 | 0.004531 |
| | TA | 0.124056 | 0.004531 |
| | TG | 0.089359 | 0.004531 |
| Asthma at 4 years<br>(n=99) | CA | 0.070573 | 0.005364 |
| | CG | 0.742559 | 0.005364 |
| | TA | 0.126397 | 0.005364 |
| | TG | 0.060472 | 0.005364 |
| Currently diagnosed asthma at 10 years<br>(n=92) | CA | 0.040528 | 0.003686 |
| | CG | 0.790994 | 0.003686 |
| | TA | 0.117081 | 0.003686 |
| | TG | 0.051398 | 0.003686 |
| Wheeze at 1 or 2 years<br>(n=48) | CA | 0.060431 | 0.009539 |
| | CG | 0.679152 | 0.009539 |
| | TA | 0.158319 | 0.009539 |
| | TG | 0.102098 | 0.009539 |
| Wheeze at 4 years<br>(n=145) | CA | 0.074331 | 0.004659 |
| | CG | 0.763600 | 0.004659 |
| | TA | 0.104979 | 0.004659 |
| | TG | 0.057090 | 0.004659 |
| Wheeze at 10 years<br>(n=228) | CA | 0.066354 | 0.003643 |
| | CG | 0.753822 | 0.003643 |
| | TA | 0.117857 | 0.003643 |
| | TG | 0.061968 | 0.003643 |
| Chronic asthma<br>(n=34) | CA | 0.116397 | 0.014663 |
| | CG | 0.663015 | 0.014663 |
| | TA | 0.148309 | 0.014663 |
| | TG | 0.072279 | 0.014663 |
| Control<br>(n=308) | CA | 0.071836 | 0.003412 |
| | CG | 0.730112 | 0.003412 |
| | TA | 0.124593 | 0.003412 |
| | TG | 0.073459 | 0.003412 |

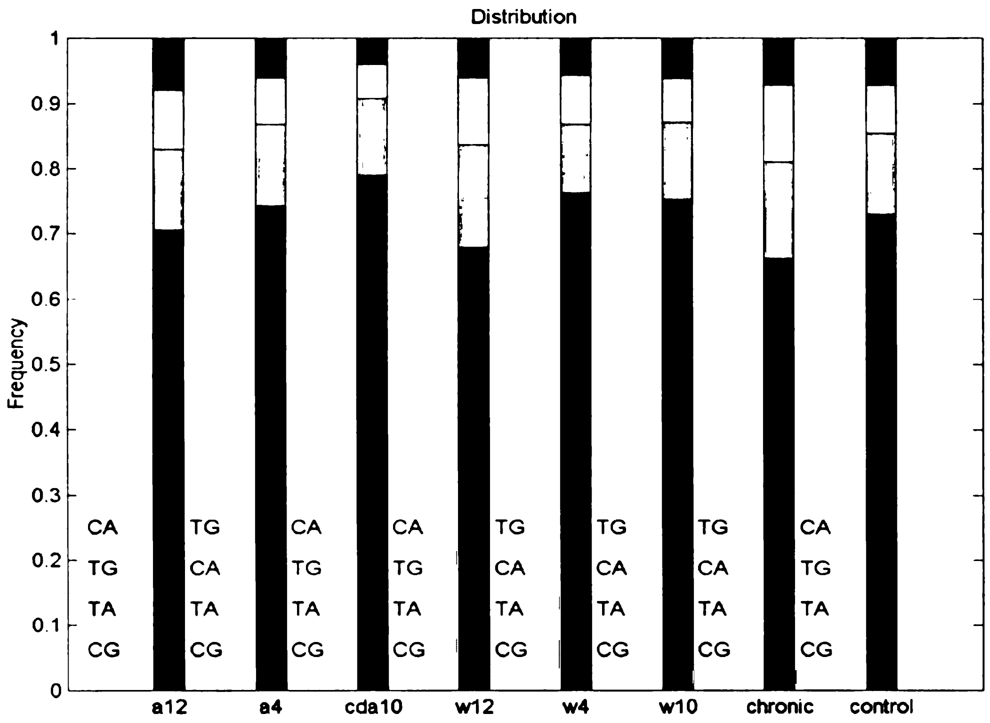
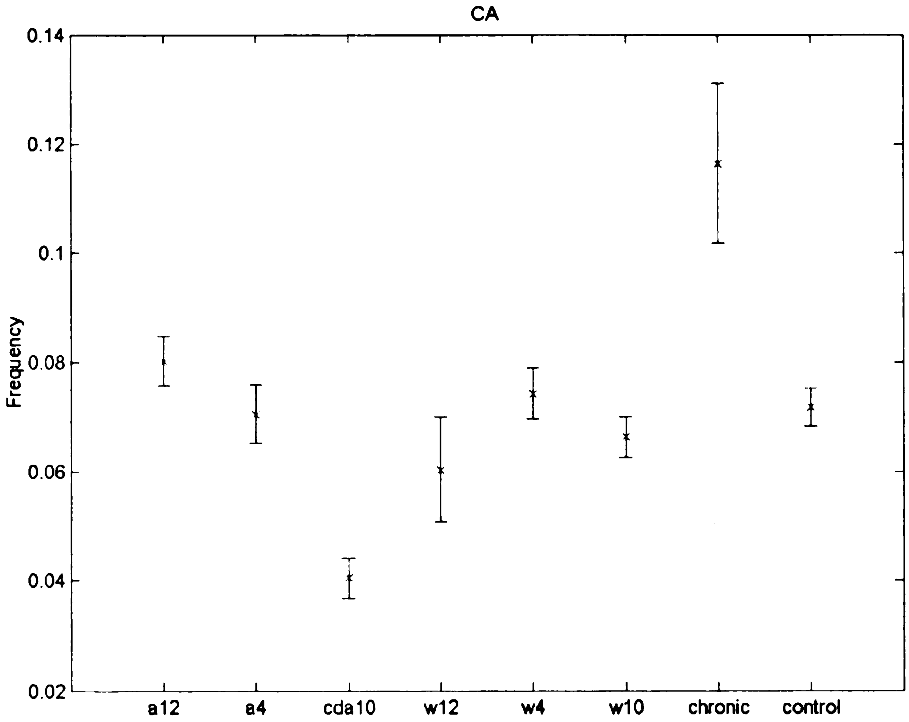
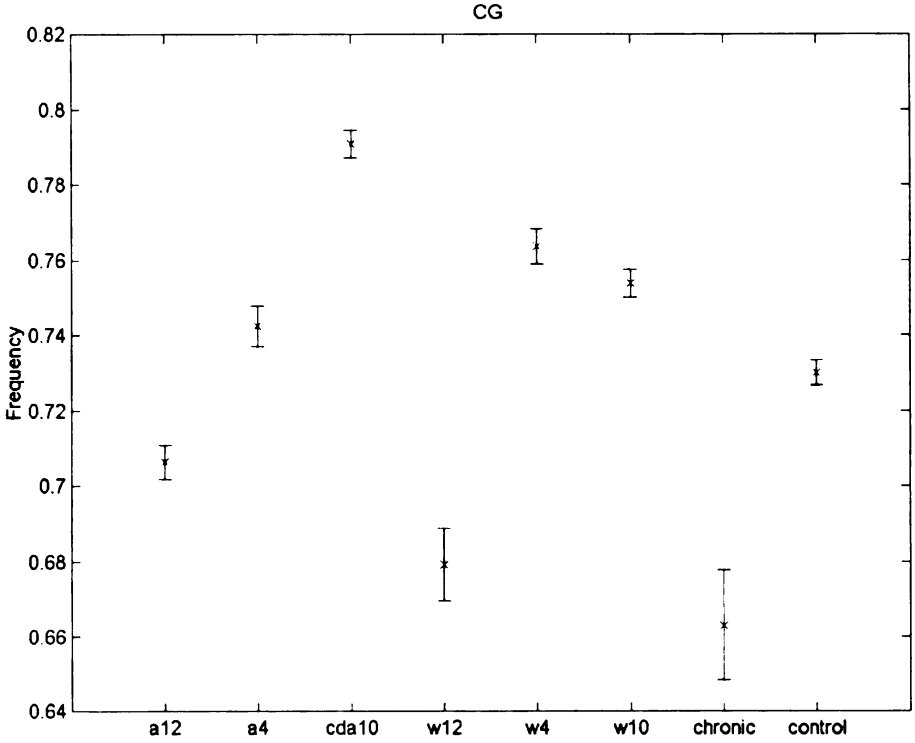


**Figure 5 Haplotype Frequency Estimation for SNPs *hCV8932056* and *hCV8932052* by Phenotypes**

As shown in Figure 6, haplotype *CG* is prominently related with the phenotype of "currently diagnosed asthma at 10 years" with relative frequency of 0.791. It has the least relationship with the phenotype of "chronic asthma" with relative frequency of 0.663. Haplotype *TA* is found in "wheeze at 1 or 2 years" with the highest relative frequency of 0.158 and in "wheeze at 4 years" with the lowest relative frequency of 0.106. Haplotype *CA* is found in "chronic asthma" with the highest relative frequency of 0.116 and in "currently diagnosed asthma at 10 years" with the lowest of relative frequency of 0.04. Haplotype *TG* is found in "wheeze at 1 or 2 years" with the highest relative frequency of 0.102 and in "currently diagnosed asthma at 10 years" with the lowest relative frequency of 0.05.


CA
Frequency
0.14
0.12
0.1
0.08
0.06
0.04
0.02
a12
a4
cda10
w12
w4
w10
chronic
control


CG
Frequency
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
a12
a4
cda10
w12
w4
w10
chronic
control

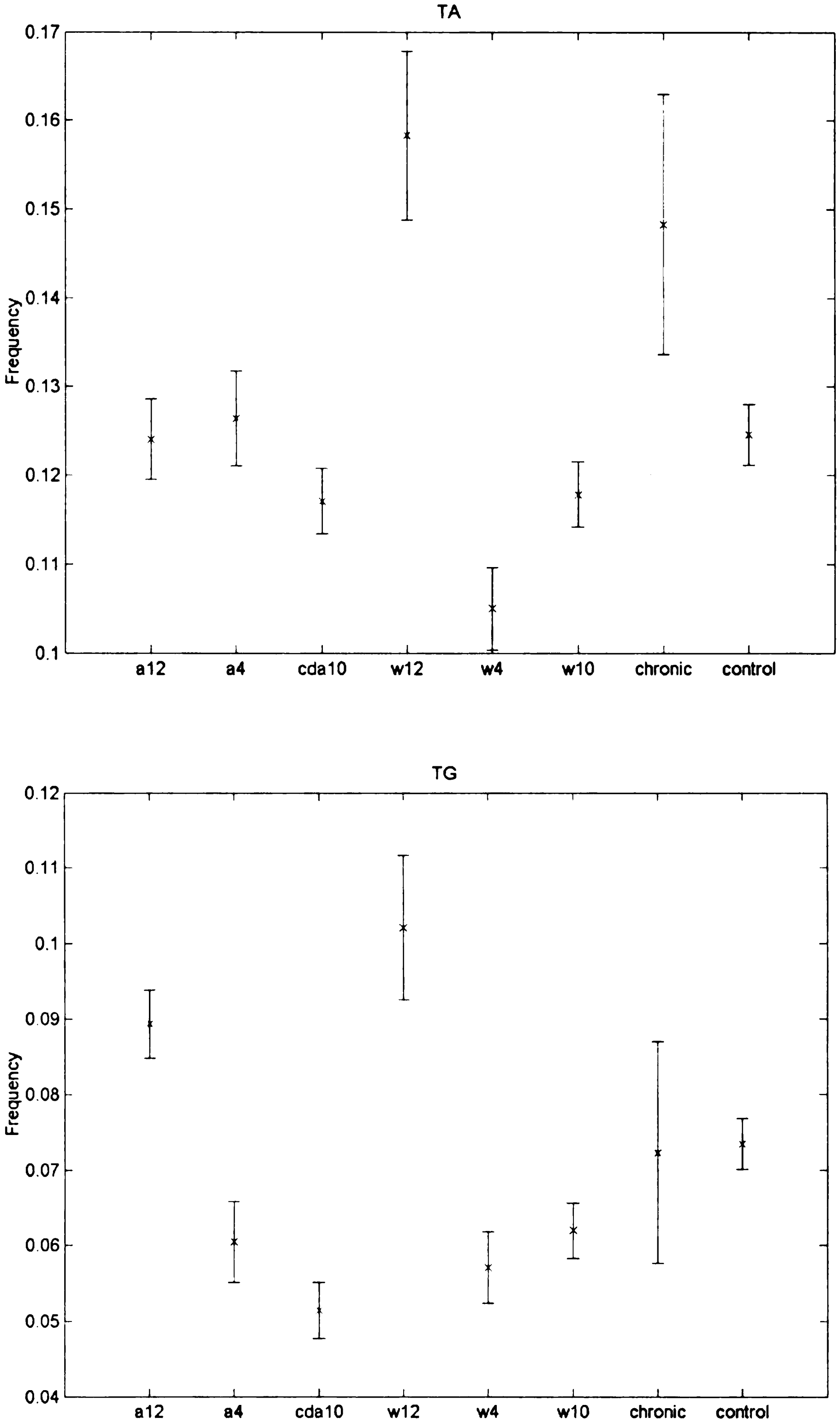


**Figure 6 Relationship between Haplotypes and Phenotypes**

In order to examine whether there exist any statistically significant associations between the haplotypes and the phenotypes, we conducted a contingency test. The contingency tables of the haplotype counts of each infected group vs. the haplotype counts of the control group are first formed, which are shown in Table 4-Table 10.

**Table 4 Two-way Contingency Table of Asthma 1 or 2 vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| A12 | 13 | 116 | 20 | 15 |
| control | 22 | 225 | 38 | 23 |

**Table 5 Two-way Contingency Table of Asthma 4 vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| a4 | 7 | 74 | 13 | 6 |
| control | 22 | 225 | 38 | 23 |

**Table 6 Two-way Contingency Table of CDA10 vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| CDA10 | 4 | 73 | 11 | 5 |
| control | 22 | 225 | 38 | 23 |

**Table 7 Two-way Contingency Table of Wheeze 1 or 2 vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| w12 | 3 | 33 | 8 | 5 |
| control | 22 | 225 | 38 | 23 |

**Table 8 Two-way Contingency Table of Wheeze 4 vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| W4 | 11 | 111 | 15 | 8 |
| control | 22 | 225 | 38 | 23 |

**Table 9 Two-way Contingency Table of Wheeze 10 vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| w10 | 15 | 172 | 27 | 14 |
| control | 22 | 225 | 38 | 23 |

**Table 10 Two-way Contingency Table of Chronic vs. Control**

| | CA | CG | TA | TG |
|---|---|---|---|---|
| chronic | 4 | 23 | 5 | 2 |
| control | 22 | 225 | 38 | 23 |

A SAS program has been written to perform the contingency tests. The SAS code has performed the contingency test between the haplotype CA/CG/TA/TG of asthma at 1 or 2 and control is shown in Appendix D. Below is the test report for the group of asthma at 1 or 2 years vs. the control group:

```
Contingency Test between Asthma at 1 or 2 and Control (Haplotypes: CA/CG/TA/TG)

                Statistics for Table of group by haplotype

          Statistic                       DF     Value    Prob
          Chi-Square                       3     0.5449   0.9089
          Likelihood Ratio Chi-Square      3     0.5371   0.9107
          Mantel-Haenszel Chi-Square       1     0.1329   0.7155
          Phi Coefficient                        0.0340
          Contingency Coefficient                0.0340
          Cramer's V                             0.0340

                         Fisher's Exact Test

          Table Probability (P)                  0.0018
          Pr <= P                                0.8957

          Sample Size = 472
```

Although SAS gives the *P*- values of several kinds of Chi-square tests, it should be noted that since some of the cells have expected counts less than 5, Chi-square may not be a valid test. Instead, Fisher's exact test's result is used when deciding if there is any association between the haplotypes and the phenotypes. As shown above, the *P*-value of 0.8957 suggests that the probabilities for the group of asthma at 1 or 2 and the control group are independent of the haplotypes *CA/CG/TA/TG*.

Table 11 summarizes the seven contingency tests. As we can see, the *P*-values range from 0.1599 to 0.9914, suggesting that haplotypes *CA/CG/TA/TG* are not associated with asthma and wheeze symptoms.

**Table 11 Summary of Contingency Tests (Haplotype CA/CG/TA/TG)**

Fisher's Exact Test

| Comparison Groups | *P*-value | Sample Size |
|---|---|---|
| a12 vs. control | 0.8947 | 472 |
| a4 vs. control | 0.9914 | 408 |
| CDA10 vs. control | 0.7222 | 401 |
| w12 vs. control | 0.6792 | 357 |
| w4 vs. control | 0.8143 | 453 |
| w10 vs. control | 0.9200 | 536 |
| chronic vs. control | 0.6813 | 342 |

Since *CG* is the most predominant haplotype, it is interesting to know if the less common haplotypes affect asthma and allergy. Thus, similar contingency tests are performed to examine the association between each infected group and the control group for the haplotypes *CA/TA/TG* only. The test results shown in Table 12 suggest that haplotypes *CA/TA/TG* are not associated with asthma and wheeze symptoms either.

**Table 12 Summary of Contingency Tests (Haplotype CA/TA/TG)**

Fisher's Exact Test

| Comparison Groups | *P*-value | Sample Size |
|---|---|---|
| a12 vs. control | 0.8953 | 131 |
| a4 vs. control | 0.9574 | 109 |
| CDA10 vs. control | 0.8092 | 103 |
| w12 vs. control | 0.8814 | 99 |
| w4 vs. control | 0.8109 | 117 |
| w10 vs. control | 0.9744 | 139 |
| chronic vs. control | 0.7817 | 94 |

# CHAPTER 5 CONCLUSIONS

This thesis has discussed three bio-statistical concepts: Hardy-Weinberg equilibrium, linkage disequilibrium, and haplotype reconstruction and haplotype frequency estimation, and their applications in the Isle of Wight cohort study which aims to identify genetic susceptibility loci for allergic asthma.

Hardy-Weinberg equilibrium has been tested through the exact test implemented by the Arlequin software package. It is found that all the loci of interest (*hCV8932056, hCV15862743, hCV8932053,* and *hCV8932052* on the *IL13* gene) are in Hardy-Weinberg equilibrium. This test result provides a valid assumption for the test of linkage disequilibrium.

Linkage disequilibrium between pair-wise loci is tested with a likelihood-ratio test, whose key procedure is the Expectation-Maximum algorithm, implemented by the Arlequin software package. It is found that all the pair-wise loci are in linkage disequilibrium.

The PHASE software package, which implements the Bayesian estimation method, is used to reconstruct haplotypes and estimate haplotype frequencies of the most informative SNP pair, *hCV8932056* and *hCV893205.* The subsequent contingency test suggests that there is no association between the haplotype patterns *CA/CG/TA/TG* and *CA/TA/TG* and allergic asthma.

# APPENDICES

## APPENDIX A

## SAS CODE TO CREATE INPUT FILES FOR ARLEQUIN

```
data thesis1.hweld1;
set thesis.asthmagroups;
ct_dash=0;
if C8932056_10="-" then ct_dash=ct_dash+1;
if C8932056_10="-" then C8932056_10="??";
if C15862743_10="-" then ct_dash=ct_dash+1;
if C15862743_10="-" then C15862743_10="??";
if C8932053_10="-" then ct_dash=ct_dash+1;
if C8932053_10="-" then C8932053_10="??";
if C8932052_10="-" then ct_dash=ct_dash+1;
if C8932052_10="-" then C8932052_10="??";
if ct_dash>=1 then delete;
freq=1;
if
CDA10="0"|CDA10="1"|asthma12="1"|asthma12="2"|asthma4="1"|asthma4="2"|wheez
e12="1"|wheeze12="2"|wheeze4="1"|wheeze4="2"|wheeze10="1"|wheeze10="2"|nosym
ptoms="0"|nosymptoms="1" then output;
run;

data thesis.hweld2;
set thesis.hweld1;
c11=substr(C8932056_10,1,1);
c21=substr(C15862743_10,1,1);
c31=substr(C8932053_10,1,1);
c41=substr(C8932052_10,1,1);
c12=substr(C8932056_10,2,1);
c22=substr(C15862743_10,2,1);
c32=substr(C8932053_10,2,1);
c42=substr(C8932052_10,2,1);
SNP1=trim(c11) || trim (c21) || trim (c31) || trim(c41);
SNP2=trim(c12) || trim (c22) || trim (c32) || trim(c42);
run;

data thesis.hweld3;
set thesis.hweld2 nobs=nobs;
filename fn 'C:\ hweld.txt';
file fn;
put @20 ID freq @30 c11 ' ' c21 ' ' c31 ' ' c41 / @30 c12 ' ' c22 ' ' c32 ' ' c42;
run;
```

# APPENDIX B

## SAS CODE TO CREATE INPUT FILES FOR PHASE

```
data thesis2.a12_1;
set thesis2.asthmagroups;
ct_dash=0;
if C8932056_10="-" then ct_dash=ct_dash+1;
if C8932056_10="-" then C8932056_10="??";
if C8932052_10="-" then ct_dash=ct_dash+1;
if C8932052_10="-" then C8932052_10="??";
if ct_dash>=1 then delete;
if asthma12="2" then output;
run;

data thesis2.a12_2;
set thesis2.a12_1;
c11=substr(C8932056_10,1,1);
c21=substr(C8932052_10,1,1);
c12=substr(C8932056_10,2,1);
c22=substr(C8932052_10,2,1);
SNP1=trim(c11) || trim(c21);
SNP2=trim(c12) || trim(c22);
run;

data thesis2.a12_3;
set thesis2.a12_2 nobs=nobs;
filename fn 'C:\snp56_52_a12_2_no_mis.txt';
file fn;
put ID / SNP1 / SNP2;
run;
```

# APPENDIX C

## MATLAB CODE TO CREATE FIGURES OF HAPLOTYPE FREQUENCY

```
% For each haplotype combination, plot freq vs. subject group

clear all;
close all;

freq_fn_label={
   'ast_grp_a12_no_mis.out_freqs',     'a12';
   'ast_grp_a4_no_mis.out_freqs',      'a4';
   'ast_grp_CDA10_no_mis.out_freqs',  'CDA10';
   'ast_grp_w10_no_mis.out_freqs',     'w10';
   'ast_grp_w12_no_mis.out_freqs',     'w12';
   'ast_grp_w4_no_mis.out_freqs',      'w4';
   'ast_grp_chron_no_mis.out_freqs',   'chron';
   'ast_grp_control_no_mis.out_freqs', 'control'
};
[no_fn tmp]=size(freq_fn_label);


for fn_ct=1 : no_fn
   %check the number of lines in the file
   fid=fopen(char(freq_fn_label(fn_ct,1)));
   data_ct=0;
   while ~feof(fid)
      data_line=fgetl(fid);
      data_ct=data_ct+1;
   end
   fclose(fid);
   %Do not count the first head line
   freq_data(fn_ct).no_data=data_ct-1;

   fid=fopen(char(freq_fn_label(fn_ct,1)));
   % Get rid of the first line
   fgetl(fid);
   % Read data
   for data_ct=1 : freq_data(fn_ct).no_data;
      freq_data(fn_ct).idx(data_ct)=fscanf(fid, '%i',1);
      freq_data(fn_ct).haplotype{data_ct}=fscanf(fid, '%10s',1);
      freq_data(fn_ct).freq_ave(data_ct)=fscanf(fid, '%f',1);
      freq_data(fn_ct).freq_se(data_ct)=fscanf(fid, '%f',1);
   end
   fclose(fid);

   %Sort the data w.r.t. freq ave
   [sort_data sort_idx]=sort(-freq_data(fn_ct).freq_ave);
   freq_data(fn_ct).idx=freq_data(fn_ct).idx(sort_idx);
   freq_data(fn_ct).haplotype=freq_data(fn_ct).haplotype(sort_idx);
   freq_data(fn_ct).freq_ave=freq_data(fn_ct).freq_ave(sort_idx);
   freq_data(fn_ct).freq_se=freq_data(fn_ct).freq_se(sort_idx);
end

%Enumerate all haplotypes
haplotype_set=freq_data(1).haplotype;
no_haplotype=freq_data(1).no_data;
for i=2 : no_fn
   for j=1 : freq_data(i).no_data
      b_find=0;
      for k=1 : no_haplotype
         if strcmp( char(haplotype_set(k)),  char(freq_data(i).haplotype(j)) )
            b_find=1;
            break;
         end
      end
      % Add to haplotype_set if the same haplotype is not found in haplotype_set
      if ~ b_find
         no_haplotype=no_haplotype+1;
         haplotype_set(no_haplotype)=freq_data(i).haplotype(j);
      end
   end
end

%Construct the structure to hold the freq_ave and freq_se for each haplotype in each file
for fn_ct=1 : no_fn
   haplotype(fn_ct).freq_ave=zeros(1,no_haplotype);
   haplotype(fn_ct).freq_se=zeros(1,no_haplotype);
   for i=1 : freq_data(fn_ct).no_data
      for j=1 : no_haplotype
         if strcmp( char(haplotype_set(j)),  char(freq_data(fn_ct).haplotype(i)) )
            haplotype(fn_ct).freq_ave(j)=freq_data(fn_ct).freq_ave(i);
            haplotype(fn_ct).freq_se(j)=freq_data(fn_ct).freq_se(i);
            break;
         end
      end
   end
end

%plot 1
```

```
bar_freq=zeros(no_fn,no_haplotype);
for fn_ct=1 : no_fn
   for j=1 : freq_data(fn_ct).no_data
      bar_freq(fn_ct,j)=freq_data(fn_ct).freq_ave(j);
      bar_label_pos(fn_ct,j)=sum(freq_data(fn_ct).freq_ave(1:j));
   end
end
figure(1);
%Draw the bars
bar(bar_freq,.3,'stack');
%Draw the labels
for fn_ct=1 : no_fn
   for j=1 : freq_data(fn_ct).no_data
      text(fn_ct-.8,j/16,char(freq_data(fn_ct).haplotype(j)));
   end
end

%hold on;
title('Distribution');
ylabel('Frequency (%)');
fn_fig=['distrib.jpg'];
set(gca,'XTick',[1:no_fn],'XTickLabel',freq_fn_label(:,2)');
print(fn_fig,'-djpeg');

%plot 2
for i=1 : no_haplotype
   figure(2);
   for j=1 : no_fn
      y(j)=haplotype(j).freq_ave(i);
      e(j)=haplotype(j).freq_se(i);
   end
   errorbar([1:no_fn],y,e,'x');
   title(char(haplotype_set(i)));
   ylabel('Frequency (%)');
   set(gca,'XTick',[1:no_fn],'XTickLabel',freq_fn_label(:,2)');

   fn_fig=[char(haplotype_set(i)) '.jpg'];
   print(fn_fig,'-djpeg');
end
```

# APPENDIX D

## SAS CODE TO CONDUCT CONTINGENCY TESTS

```
data thesis3.contingency;
        do group=1 to 2;
            do haplotype=1 to 4;
                 input count @@;
                 output;
            end;
        end;
cards;
13 116 20 15
22 225 38 23
;

proc freq data=thesis1.contingency;
   weight count;
   tables group*haplotype / exact;
   title 'Contingency Test between Asthma at 1 or 2 and Control (CA/CG/TA/TG)';
run;
```

## APPENDIX E

## NOTATIONS

| Symbols | Notations |
|---|---|
| $A, B, \ldots$ | Loci |
| $A_{i,}\ i=1,\ \ldots,\ m$ | Alleles at locus A with *m* alleles |
| $f_i\ ,\ i = 1,\ \ldots,\ m$ | Relative frequencies of alleles at a locus with *m* alleles |
| $c_{i,}\ i=1,\ \ldots,\ m$ | Counts of alleles at a locus with *m* alleles |
| $g_{ij}\ ,\ i, j = 1,\ \ldots,\ m$ | Relative Frequencies of genotypes at a locus with *m* alleles |
| $g_{ijkl}\ ,\ i, j = 1,\ \ldots,\ m,$<br>$k, j = 1,\ \ldots,\ n,$ | Relative frequencies of genotypes at two loci which have *m* alleles and *n* alleles respectively. |
| $g_{ik,jl}$ | Relative frequencies of genotypes obtained from haplotype pair $A_iB_k\ /\ A_jB_l$ at loci *A* and *B*. |
| $c_{ij,}\ i, j =1,\ \ldots,\ m$ | Counts of genotypes at a locus with *m* alleles |
| $c_{ijkl,}\ i, j =1,\ \ldots,\ m,$<br>$k, j = 1,\ \ldots,\ n,$ | Counts of genotypes at two loci which have *m* alleles and *n* alleles respectively. |
| $c_{ik,jl}$ | Counts of genotypes obtained from haplotype pair $A_iB_k\ /\ A_jB_l$ at loci *A* and *B*. |
| $h_{ij}\ ,\ i, j = 1,\ \ldots,\ m$ | Relative frequencies of haplotypes at a locus with *m* alleles |

# BIBLIOGRAPHY

# BIBLIOGRAPHY

Clark, A. G. Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology Evolution. 1990: 7(2): 111-122.

Crow, J. F. Eighty years ago: the beginnings of population genetics. Genetics 1988 Jul: 119 (3): 473-6.

Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology Evolution. 1995: 12(5): 921-927.

Fallin, D. & Schork, J. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. American Journal of Human Genetics 2000: 67: 947-959.

Gilks, W. R., Richardson S., & Spiegelhalter, D. J. Markov Chain Monte Carlo in Practice. London: Chapman & Hall, 1996.

GlaxoSmithKline. Genetics at GlaxoSmithKline. [On-Line]. Available: http://genetics.gsk.com/link.htm, 2004.

Guo, S. W., & Thompson, E. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 1992: 48: 361-372.

Kurukulaaratchy, R. J., Fenn, M. H., Waterhouse L. M., Matthews, S. M., Holgate, S. T., Arshad, S. H. Characterization of wheezing phenotypes in the first 10 years of life. Clinical and Experimental Allergy 2003: 33: 573-578.

Kurukulaaratchy, R. J., Matthews, S., Holgate, S. T., Arshad, S. H. Predicting persistent disease among children who wheeze during early life. European Journal of Respiratory Diseases 2003: 22: 719-720.

Kurukulaaratchy, R. J., Matthews, S., Waterhouse, L., Arshad, S. H. Factors influencing symptom expression in children with bronchial hyperresponsiveness at 10 years of age. The Journal of Allergy and Clinical Immunology 2003: 112: 311-316.

Long, J. C., Williams, R. C. & Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. American Journal of Human Genetics 1995: 56:799-810.

Schneider, S., Roessli, D., & Excoffier, L. Arlequin ver. 2.000: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.

Schork, N. J., Fallin, D., & Lanchbury, S. Single nucleotide polymorphisms and the future of genetic epidemiology. Clinical Genetics 2000: 58:250-264.

Sham, P. Statistics in Human Genetics. New York: Oxford University Press, 1998.
Stephens, M., Smith, N. J., & Donnelly, P. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 2001: 68: 978-989.

Stephens, M., Smith, N. J., & Donnelly, P. Documentatin for PHASE ver.2.02. Department of Statistics, University of Washington, Seattle, WA, 2003.