

Strategies and Statistical Methods for Linkage Disequilibrium-based Mapping of Complex Traits

by

Tianye Jia

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
The University of Birmingham
Dec 2011

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Nowadays, there are many statistical methods available for genetic association analyses with data various designs. However, it is usually ignored in these analyses that an analytical method must be appropriate for an experimental design from which data is collected. In addition, association study is a population-based analysis and, thus its inference is highly vulnerable to many population-oriented confounding factors. This thesis starts with a comprehensive survey and comparison of those methods commonly used in the literature of genetic association study in order to obtain insights into the statistical aspects and problem of the methods.

On the basis of these reviews, we managed to calculate the optimal trend set for the Armitage's trend test for different penetrance models with a high level of genetic heterogeneity. We introduced two new strategies to adjust for the population stratification in association analyses. We proposed a maximum likelihood estimation method to adjust for biases in statistical inference of linkage disequilibrium (LD) between pairs of polymorphic loci by using non-random samples. In the process of the analysis, we derived a more sophisticated but robust likelihood-based statistical framework, accounting properly for the non-random nature of case and control samples. Finally, we developed a multi-point likelihood-based statistical approach for a genome-wide search for the genetic variants that contribute to phenotypic variation of complex quantitative traits. We tested these methods through intensive simulation studies and demonstrated their application in analyses with large case and control SNP datasets of the Parkinson's disease.

Despite that we have mainly focused on SNP data scored from microarray techniques, the theory and methodology presented here paved a useful stepping stone approach to the

modeling and analysis of data depicting genome structure and function from the new generation sequencing techniques.

ACKNOWLEDGEMENT

First of all, I would like to express my sincere gratitude to Professor Zewei Luo. Without your patient guidance and kind encouragement, this thesis would not be possible to complete. Thank you indeed for instructing me in both my life and academic studies. It is my great honour and fortune to follow your supervision.

I would also like to thank Professor Michael J. Kearsey, my second supervisor. Your encouragement and acknowledgement have brought me the strength to carry on this PhD study as a transferred student from Physics.

Please also allow me to express my warm thanks to all my colleagues who have helped me during my study and writing up the thesis. In no particular order, I would like to thank Lindsey Leach, Minghui Wang, Ning Jiang, Erica Liu, Elena Potokina and Joseph Abraham. I will never forget your help and companion in my study at the University of Birmingham.

I should not forget my friends in daily life in Birmingham as well. Thank you so much for being with me and supporting me. Your names should never be forgotten, and please allow me to thank you altogether: Shutao Chen (Taozi), Yan Wang (Laoda), Chun Liu (Gary), Qiping Yu (Skye), Chung Lo (Clement), Meng (Nicole) & Lyndon Checketts, Xiaoming Ding, Ai Jian (Annie), Hui Jiang, Lin Hu, Xin Wang, Ming Li.

At last but never the least, I would like to especially thank my parents for supporting me, loving me and offering me the chance to study abroad in the UK. I owe you so much and I love you, as always.

Table of Contents

General Introduction.....	1
1.1 Genetic Markers	4
1.2 Strategies for QTL Mappings.....	7
1.2.1 Ideal Population and Random Sampling.....	7
1.2.2 Linkage Based QTL Analysis.....	11
1.2.3 Linkage Disequilibrium Based QTL Analysis	12
1.2.4 Advantages and Challenges of Linkage Disequilibrium Based QTL Analysis.	15
1.3 Structure of the Thesis.....	23
1.4 Preliminaries.....	24
1.4.1 Statistical hypothesis test.....	24
1.4.2 Maximum Likelihood Estimator (MLE) and Likelihood Ratio Test (LRT)	28
1.4.3 Linear Regression Analysis.....	33
Reference	44

Chapter II The Comparison of Linear Models in Association Study and the Optimal Trend Coefficients in Armitage's Trend Test..... 51

Chapter II-1 Introduction of Genetic Models and Statistic Methods of Association Analysis in Randomly Mated Population	52
1.1 Overview	52
1.2 Models of Quantitative Genetic Effects	52
1.2.1 Explicit Model	52
1.2.2 Implicit Model	53
1.3 Association Analysis in Random Samples	53
1.3.1 Simple Linear Regression (SLR).....	56
1.3.2 Fixed Effect Model (ANOVA).....	58
1.3.3 Likelihood Based Approach	62
1.3.4 Under the Scheme of Likelihood.....	66
1.4 Case Control Study	68
1.4.1 Pearson's χ^2 test	69
1.4.2 Allelic Analysis	70
1.4.3 The Armitage's Trend Test.....	71
1.4.4 Liability (Probit) and Logistic Model.....	72
1.5 Rationale of this Study	75

Chapter II-2 Comparison of Statistical Methods and Genetic Models	76
2.1 Overview	76
2.2 Performance of Ordinary Linear Models.....	77
2.2.1 The Optimal Choice of x_i in the SLR for Random Samples.....	77
2.2.2 Comparison of Statistical Power of SLR and ANOVA	82
2.3 The HWE in Case-Control Study	85
2.4 The Equivalence between Models of Quantitative Genetic Effects	87
2.4.1 Under the Generalised SLR.....	87
2.4.2 Under the Generalised Fixed Effect Model.....	90
2.4.3 Under the Scheme of a Likelihood-based Approach.....	91
2.4.4 Conclusion.....	94
2.5 Modified Armitage's Trend Test.....	94
Chapter II-3 A Modified Armitage's Trend Test for Different Penetrance Models and Population Stratification	95
3.1 Overview	95
3.2 The Optimal Trend Set	96
3.2.1 Common Variants or Rare Variants	96
3.2.2 Calculation of the Optimal Trend Set.....	98
3.2.3 The Non-Recessive Model	102
3.2.4 The Recessive Model.....	108
3.2.5 Regarding the False Positive	118
3.2.6 Conclusion.....	119
3.3 Correction for Population Stratification in the Armitage's Trend Test.....	120
3.3.1 Method I: Dummy Variables.....	121
3.3.2 Method II: Non-Central χ^2 Test.....	125
3.3.3 Simulation Study	133
3.3.4 Real Data Analysis	142
3.3.5 Conclusion.....	152
3.4 Conclusion and Discussion.....	153
Reference	156
Chapter III Likelihood-based Methods for Association Studies.....	161
Chapter III-1 Inference of Linkage Disequilibrium in Non-random Sample	162
1.1 Related Publications	162
1.2 Overview	162

1.3	Inferring LD in Random Samples (Hill's Method)	163
1.4	A Likelihood-based Method for Inferring LD from Selected Samples	166
1.5	Simulation Studies	169
1.6	Real Data Analysis	177
1.6.1	β -Thalassemia Dataset	177
1.6.2	Comparison of Three Methods	177
1.7	Conclusion and Discussion	180
Chapter III-2 A Likelihood-based Method for Association Study in Case Control Samples from Multiple Cohorts		182
2.1	Related Publication	182
2.2	Overview	182
2.3	Notations and Models	183
2.4	Allelic Analysis for Multiple Cohorts	193
2.5	Simulation Study	196
2.6	Real Data Analysis	201
2.7	Conclusion and Discussion	211
Chapter III-3 A Composite Likelihood-based Model for Association-based Mapping of Quantitative Trait Loci		213
3.1	Overview	213
3.2	Notations and Models	213
3.3	Simulation Studies	221
3.3.1	Simulation Models	221
3.3.2	The Relationship between T , r and n	229
3.3.3	Detecting Linked QTLs	232
3.4	Conclusion and Discussion	234
Reference		236
Final Conclusion		238

List of Tables

General Introduction

Table I-1. The Settings of Marker and Trait Loci.	13
Table I-2. The Comparison of Linkage Analysis and Association Study.	23
Table I-3. Layouts of a Statistical Hypothesis Test under the Neyman-Pearson Paradigm.	25

Chapter II-1

Table II-1. Joint Distribution of Marker and QTL Genotypes in a Random Mating Population	55
Table II-2. Distribution of Genotypes in Case-Control Study.....	69
Table II-3. Distribution of Allele Frequencies in Case-Control Study.....	70

Chapter II-3

Table II-4. Joint Distribution (Unnormalised) of Marker and QTL Genotypes for Case Control Samples.....	99
Table II-5. Joint Distribution of Marker and QTL Genotypes in a random Population where Case and Control Samples Were Collected.....	99
Table II-6. Simulation Results of Three Trend Sets for Dominant and Additive Model.....	107
Table II-7. Simulation Results of Three Trend Sets for Recessive Model with r Equalling 0.05	116
Table II-8. Simulation Results of Three Trend Sets for Recessive Model with r Equalling 0.10	117
Table II-9. False Positive Rates of Default and Recessive Trend Sets.....	118
Table II-10. The Scheme of Simulation.	135
Table II-11. The Results of Simulation.	136
Table II-12. List of Most Significant Markers with Test Score, of $-\log(\text{P-value})$ from Method I.....	146
Table II-13. Comparison of Three Methods at Significant Loci.	148

Chapter III-1

Table III-1. Frequencies of haplotypes between marker and disease alleles.....	163
Table III-2. Joint distribution of marker and disease genotypes in a randomly mated population	164
Table III-3. Conditional Distribution of Disease Genotypes upon Marker Genotypes.....	166
Table III-4. Summary of Estimates of D for Scheme I.	170
Table III-5. Summary of Estimates of D from Scheme II.	173
Table III-6. Summary of Estimates of D from Scheme III.....	176

Chapter III-2

Table III-7. Conditional Distribution between Marker and QTL Genotypes.....	185
Table III-8. Results for Simulation under Scheme A with Simulation Parameters.....	198
Table III-9. Results for Simulations under Scheme B with Simulation Parameters.	200
Table III-10. Significant Markers Detected by Method 1 from Stage I Data.....	205
Table III-11. P-values of Known Candidate Genes from Three Methods.	207
Table III-12. Significant Markers Detected by Method 1 from Stage II Data.	208
Table III-13. Significant Markers Detected by Method 1 from Combined Stage I & II Data.	209

Chapter III-3

Table III-14. Comparison of Simulation Results with Real and Estimated T.....	223
Table III-15. Simulation Results for QTL Locates at 28.2 with 5cM Mpd.....	224
Table III-16. Simulation Results for QTL Locates at 28.4 with 5cM Mpd.....	225
Table III-17. Simulation Results for 10cM Mpd.....	226
Table III-18. Simulation Results for 50cM Mpd.....	227
Table III-19. Simulation Results for Twin QTLs.....	228

List of Figures

General Introduction

Figure I-1	Network from genotype to phenotype	2
------------	--	---

Chapter II-3

Figure II-1	The relationship between the upper bound of optimal x_2 and marker allele frequency p under a non-recessive model	103
Figure II-2	The relationship between adjusted LD and correlation between x and k with default trend set $\{1, 0.5, 0\}$	105
Figure II-3	The relationship between the value of the optimal x_2 and the adjusted LD	110
Figure II-4	The relationship between adjusted LD and $Cor(x, k)$ with trend set $\{1, 0.5, 0\}$	113
Figure II-5	The relationship between adjusted LD and $Cor(x, k)$ with trend set $\{1, 0.25, 0\}$	114
Figure II-6	The relationship between adjusted LD and $Cor(x, k)$ with trend set $\{1, 0, 0\}$..	114
Figure II-7	Q-Q Plots of Method I (a), Method II (with λ) (b) and Original (c) for Sample 1 in Table II-10 with 10000 replicates	140
Figure II-8	Q-Q Plots of Method I (a), Method II (with λ) (b) and Original (c) for Sample 11 in Table II-10 with 10000 replicates	141
Figure II-9	The Manhattan plots of $-\log(P\text{-value})$ for Method I	143
Figure II-10	The Manhattan plots of $-\log(P\text{-value})$ for Method II	144
Figure II-11	The Manhattan plots of $-\log(P\text{-value})$ for Original	145

Chapter III-1

Figure III-1	Results of analysis for β -thalassemia causing mutation. (a) Estimates of the coefficients of LD from three different methods (b) The LOD score values calculated for the LD estimates from the three methods	179
--------------	---	-----

Chapter III-2

Figure III-2 Genome-wide association results from (a) stage I, (b) stage II and (c) two-stage combined case and control samples..... 204

Figure III-3 means of lod score values of composite likelihood ratio statistics for each simulated population in Table III-19 233

List of Abbreviations

ANOVA	Analysis of variance
BLUE	Best linear unbiased estimation (estimator)
BLUP	Best linear unbiased prediction (predictor)
cM	Centimorgan
df	degree of freedom
ECM algorithm	Expectation-Conditional Maximization algorithm
FDR	False discovery rate
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
LR(T)	Likelihood ratio (test)
MAF	Minor allele frequency
ML(E)	Maximum likelihood (estimation or estimator)
$N(\alpha, \beta)$	Normal distribution with mean α and variance β
PD	Parkinson's disease
QTL	Quantitative trait locus
R	field of the real numbers
REML	Restricted maximum likelihood
SLR	Simple linear regression
Z	Standardized normal distribution, i.e. $N(0, 1)$

CHAPTER I

GENERAL INTRODUCTION

Since genes and chromosomes were found, most traits of human beings or any living things were believed to be controlled by one or more genes as well as a degree of environmental modifications (Kearsey and Pooni, 1996). These traits include not only height, weight, colour of eyes of humans, flowering time of plants and most of other biological characters people are familiar with, but also many genetic-oriented disorders such as Huntingdon's disease (Walker, 2007), Parkinson's disease (Lesage and Brice, 2009), type II diabetes (McCarthy, 2010) and etc.. All those traits show various levels of genetic influence. Naturally, most traits are polygenic or complex and do not follow the Mendelian inheritance patterns, e.g. height of human, as opposed to the simple monogenic traits, e.g. Huntingdon's disease. For complex traits, how many genes are responsible, where and what are they, and how do they act individually and jointly are hence the key initial questions to be answered by geneticists (Mackay, 2001). To answer such questions, a full picture of the genetic architecture between genes and traits has to be outlined, with the most comprehensive and ultimate goal of depicting a completely resolved network among genome, transcriptome, proteome, metabolome and finally the phenotype of traits (Figure I-1). However, people's understandings of such networks are still too limited to fully explain the determination from genes to complex traits, and hence either the direct analysis between gene and trait or certain decompositions of the gene-trait network has to be taken (Schork, 1997).

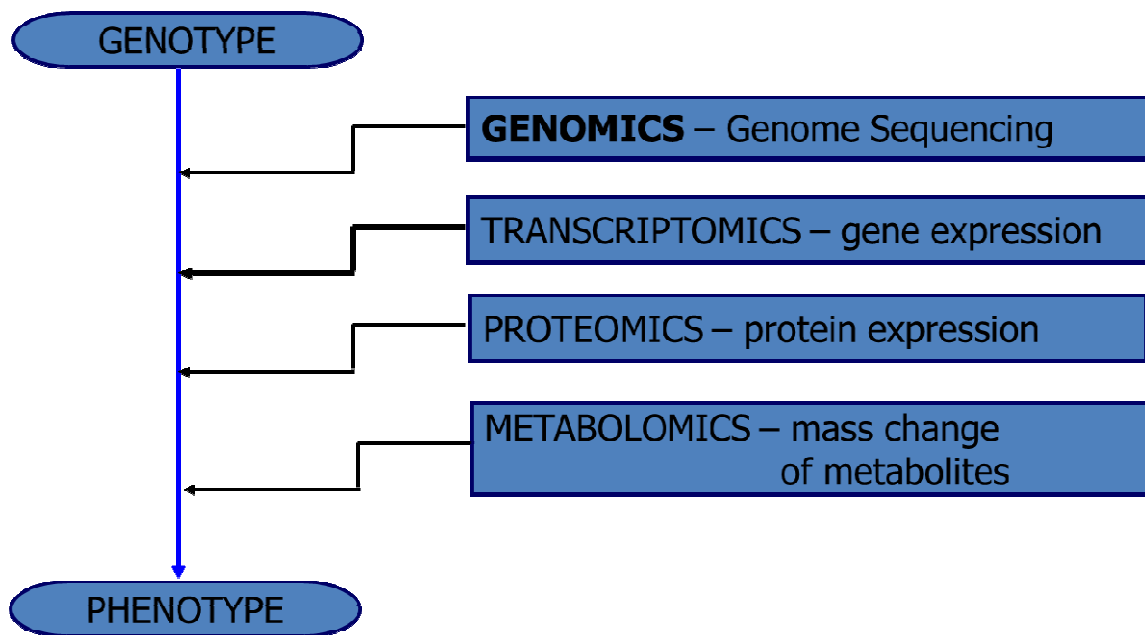


Figure I-1. Network from genotype to phenotype.

Although the pioneer work of mapping causal genes underlying polymorphic traits could be traced back to 1910s (Sturtevant, 1913), it has not yet been largely implemented for human traits until 1980s (Botstein et al., 1980) when restricted fragment length polymorphisms (RFLPs) could be readily verified in the genome (Petes and Botstein, 1977). With an established genetic map of sequence polymorphisms, Botstein and his colleagues (1980) suggested that, instead of mapping causal genes directly, people could map genetic polymorphic loci which are linked to the genes responsible for the trait phenotypes. Such genetic loci identified in mapping studies were later named as quantitative trait loci (QTLs). Literally, ‘quantitative’ might imply continuity, and hence quantitative traits are generally opposite to discrete traits. However, the meaning of quantitative trait is generally broadened to any polygenic traits even while they are not continuous, such as Parkinson’s disease and type II diabetes. Apart from the ultimate phenotypes at the organism level, e.g. the disease symptoms, Schork (1997) suggested mapping genes for intermediate phenotypes, e.g. the variation of gene expression levels or the hormone amounts. It is not only biological because such analysis could help to bridge the gap between genotypes and the ultimate phenotypes, but also statistical because the causal genes are expected to have larger impacts on

intermediate phenotypes than the ultimate one, allowing a higher chance to detect the genotype-intermediate phenotype relationship. Following Schork's suggestion, the concept of 'genetical genomics' was proposed by Jansen & Nap (2001) to map QTLs for mRNA abundances on the genome, namely the expression QTLs (e-QTLs). Later a genome-wide e-QTL analysis conducted by Brem et al. (2002) in the budding yeast proved the gene-gene interaction or epistasis through identifying the trans-acting regulators, which influence the expression levels of one or more other genes. This landmark study illuminated a promising path to construct the gene regulatory network or pathway, and hence e-QTL mapping has become one of the hottest topics in the past decade (Ronald et al., 2005; Kendzierski et al., 2006; Wray, 2007; Potokina et al., 2008; Holloway and Li, 2010; Druka et al., 2010). However, even with a fully constructed gene regulatory network, there is still a long way to go before bridging the gap between the transcriptome and the ultimate phenotype, i.e. the proteome and metabolome. To bridge the remaining gaps, following the same idea as the e-QTL analysis, pioneer work in both the protein QTL (p-QTL) (Foss et al., 2007; Melzer et al., 2008) and the metabolic QTL (m-QTL) (Wentzell et al., 2007; Ferrara et al., 2008) has already been launched, and it is possible that the fast developing technologies could enable the full scale of proteome and metabolome QTL analyses in the coming future and lead to a more comprehensive understanding of the gene-trait network.

The research into the gene-trait network and the QTL analysis is not only for purely scientific purpose, but also for the intensive demands from many areas, such as agriculture, animal husbandry and human public health. For agriculture and animal husbandry, once the causal genes of interest traits are known, the efficiency of selections for plants and animals could be largely improved, especially for some rare traits. For public health, the detection of high risk genes not only helps to warn about a possible disease, but also assists in developing better disease treatment. For example a new concept of pharmacogenomics attempts to localize the

variations in genes that dictate individual specific response to particular drugs, i.e. the efficacy and toxicity, and finally to reach the goal of genetic aided personalized prescriptions (Limdi and Veenstra, 2010). Along with the traditional genetic applications above, genetic engineering such as gene cloning and modification could also benefit from mapping precise gene locations (Kearsey and Pooni, 1996).

1.1 Genetic Markers

Valid QTL mapping requires markers which could identify each individual as well as provide sufficient polymorphism, to detect linkages between the marker and the trait of interest. Before the first generation of genetic markers, i.e. RFLP, was introduced in the late 1970s (Petes and Botstein, 1977), morphological markers, e.g. certain phenotypic variants polymorphisms caused by mutations at a certain loci, were the primary choice since late 1920s (Stadler, 1929). However, although such morphological markers had been proved to be effective for Mendelian traits in certain plants and animals, they are quite vulnerable to pleiotropy and multiple levels of heterogeneities. The polymorphic variation of morphological markers might also be reduced if dominance, deleterious effect and rare variants are present. These drawbacks hinder the application of morphological markers (Worland et al., 1987) and hence a new type of marker, i.e. the genetic marker, with both reliability and consistent high polymorphism, i.e. being consistently co-dominant, is highly desirable (Farooq and Azam, 2002).

A genetic marker is a stretch of polymorphic DNA sequence with a known location in the genome that could be used to identify genes as well as individuals. A genetic marker might be as long as a whole gene or as short as a single nucleotide. Due to its polymorphic nature, the genetic marker is often used interchangeably with the genetic polymorphism and the genetic variant if we are not emphasizing on its location in the genome. Since all heritable

information is believed to be carried by DNA, the genotypic variations at genetic markers are stable and hence represent the fundamental of any phenotypic variations at higher network or pathway levels. On the other hand, most genetic markers are co-dominant and hence are constantly highly polymorphic comparing to the morphological markers introduced above. Following RFLP, a series of succeeded genetic markers have been discovered, such as RAPD (random amplified polymorphism DNA) (Williams et al., 1990), AFLP (amplified fragment length polymorphism) (Vos et al., 1995), mini/microsatellite (Jeffreys et al., 1991), SFP (single feature polymorphism) (Winzeler et al., 1998) and SNP (single nucleotide polymorphism). Among all these genetic markers, SNPs are probably the most popular nowadays. The popularity of SNPs is not only because they are wide spread across the chromosomes (as many as 10 millions in the human genome that may account for approximately 90% of genetic variants in the genome), but also because of their binary property that is very suitable for statistically modelling.

The belief of ‘common disease, common variant’ implies that the wide spread common SNPs, of which the minor allele frequency (MAF) is larger than 5%, could explain most of the common diseases, and such a belief becomes the driven idea of establishing the HapMap project to identify common SNPs in the human genome. Phases I+II of the HapMap project have identified approximately 3 million SNPs (Frazer et al., 2007), and the newly released Phase III released 1.6 million SNPs with a larger sample size and more divergent genetic background than the former two Phases (Altshuler et al., 2010). The growing availability of high density and genome-wide distributed SNP markers facilitates the genome scale fine mapping of QTLs, i.e. the so-called genome-wide association study (GWAS), in hope of identifying, at the molecular level, the causal genes or the quantitative trait nucleotides (QTNs) responsible for the variation of trait of interest. However, although hundreds of GWAS have reported thousands of candidate SNPs associated with complex traits in the past few years

(Donnelly, 2008; Ku et al., 2010), the collective effects of candidate common SNPs could only explain a very small proportion of the heritability for most of the analysed traits. The ‘missing heritability’ failed to be accounted for by common SNPs implies remarkable contributions from less common ($1\% < \text{MAF} < 5\%$) or rare ($\text{MAF} < 1\%$) SNPs, or other structural variants including insertion/deletion and inversion (Goldstein, 2009).

Recently, Yang and his colleagues (2010) revealed that rare variants might be responsible for at least 40% variance of human height that was still missing after calculating the joint contribution of all SNPs with $\text{MAF} > 1\%$, whereas the reported 50 candidate SNPs could only explain about 5% of the whole variance. This result further addresses the urgent demand of QTL analysis with rare variants. Although several pioneer studies of QTL mapping with rare variants, mainly rare SNPs, have been conducted (Bansal et al., 2010; Ku et al., 2010), they are mainly based on targeted candidate genes but are not at a genome-wide scale. It would be an intuitive idea to establish a comprehensive map of all genetic variants including rare variants to perform the genome-wide QTL analyses. However, the bottle-neck of the current genotyping methods impedes such attempts. The difficulty comes down to lacking affordable high throughput genotyping platforms or technologies that are feasible for detecting all type of variants (Scherer et al., 2007). In fact, because of technique limitations, almost all genotyping is partial and even contains errors. Take SNPs for example, the genotyping error rate is typical over 0.1% with the current array-based techniques (Hao et al., 2004; Saunders et al., 2007; Yeung et al., 2008), which is comparable to or even higher than the occurring rate of certain rare SNPs, rendering such genotyping strategies impractical for rare variants. Therefore, a reported singleton or low copy SNP in a sample may be very likely due to an error signal rather than a real hit. Ignoring such errors might cause serious biases for the QTL analyses. By acknowledging this fact, an advanced genotyping technique with reduced error rate is essential for reliable QTL analyses with rare variants.

Recent applications of the next-generation sequencing (NGS) techniques are capable of producing millions of sequence reads in a single run, providing inexpensive genome-wide sequence solutions in a massively parallel manner with an unimaginable speed (Shendure and Ji, 2008; Metzker, 2010). For example, the 1000 Genome Project is utilizing NGS techniques to identify less common SNPs and structural variants as well as their haplotype contexts for HapMap populations (www.1000genomes.org). Compared to the ideal genotyping strategy, the current NGS techniques still have some limitations, e.g. the error rate of single read is relatively high, especially for certain variants ($>1\%$) (Shendure and Ji, 2008). However, such a drawback could be compensated through increasing read coverage. It is believed that the capability for NGS to produce enormous sequencing data inexpensively is extremely attractive, and given the fast development of techniques, it is quite optimistic that above shortages will soon be overcome.

1.2 Strategies for QTL Mappings

1.2.1 Ideal Population and Random Sampling

It would be necessary to clarify several basic concepts and assumptions before the introduction of QTL mapping strategies, as they represent the fundamental factors of establishing reliable statistical approaches for QTL analyses.

Assumption of constant allele frequency

In a natural population, it would be practical to assume that allele frequencies keep approximately constant throughout all generations unless the following disturbances cannot be ignored (Falconer, 1989):

Selection: Selection, either natural or artificial, alters the viability and/or fertility of individuals, and hence the individuals containing certain phenotypes that are favoured by selections could generate more offspring for the next generation than the individuals with less favoured phenotypes. In the presence of selective pressure, the proportion of favoured phenotypes will tend to increase in the population, and the allele frequencies of the genes underlying such phenotypes will hence be shifted. Note here, the selection here is mainly referred to as directional selection, where only one allele is favoured by selection and its frequency will tend to be consistently increased. Other types of selection might not significantly result in the shift of allele frequencies, e.g. disruptive selection and stabilizing selection.

Recurrent Mutation: Unlike a random mutation, which is very likely to get lost during the evolution of a population, the recurrent mutation at the same genetic locus, however, will eventually survive.

Migration: Either into or out of a population, the allele frequencies of the new population might hence change forever if the original population and the migrations have different allele frequencies at certain loci.

Genetic Drift: In a small population, each allele will have its frequency drift randomly generation by generation and eventually get fixed or lost.

It should be noted here that, excluding the genetic drift, the influence of allele frequencies from the other disturbing factors are directional and hence the impact could be exactly evaluated. More importantly, the influences from more than one such directional disturbing factor might finally reach equilibrium such that the allele frequencies will still stay constant henceforward, e.g. balance between mutation and selection. From such a consideration, it would be quite reasonable to assume the allele frequencies keep constant over generations for a large stable population, unless an extraordinary disturbance was introduced recently.

Random mating and Hardy-Weinberg equilibrium

Following the assumption of constant allele frequency, if the genotype frequencies also keep constant over generations, such a population would be referred as following the Hardy-Weinberg Equilibrium (HWE). Although it is not always necessary, random mating, which assumes each pair of gametes from the same gene pool would have the same chance to unite, is the sufficient condition for HWE jointly with the assumption of constant allele frequency, and hence is often implied with the assumption of HWE. For any given genetic marker locus with alleles M and m from a random mating diploid species, frequencies of marker genotypes MM , Mm and mm are expected to be p_M^2 , $2p_M p_m$ and p_m^2 respectively, where p_M is the frequency of allele M and p_m is the frequency of allele m . If the allele frequencies keep constant over generations, the genotype frequencies under random mating will also do so, and hence the HWE is acquired. However, consider another diploid population with the same marker alleles M and m but without any heterozygotes, the frequencies of genotype MM , Mm and mm are hence p_M , 0 and p_m respectively. If all individuals from such a population are strictly selfing with the same fertility, without any of the disturbances as introduced above, both the allele and genotype frequencies are also constant over generations and thus the HWE holds without random mating. Although such a strictly selfing species does not exist in nature because it lacks genetic diversity and hence it is vulnerable to natural selections, similar situations could be found in artificially inbred lines, e.g. barley and Arabidopsis. As a conclusion from above discussions, given the assumption of constant allele frequency, random mating is the necessary and sufficient condition of HWE in an outbred population, otherwise the HWE might not be valid.

Along with the inbreeding mentioned above, non-random mating could also be naturally found with assortative mating, which could be either positive, where individuals would prefer

to mate with other similar individuals, or negative, where individuals would prefer to mate with dissimilar individuals. The positive assortative mating will generally result in the decrease of heterozygotes, similar to the effect of disruptive selection in the absence of over-dominance, where the extreme phenotypes are favoured, and the negative assortative mating will result in the increase of heterozygotes, similar to the effect of stabilizing selections in absence of over-dominance. Although assortative mating will not change the allele frequencies (Falconer, 1989), it does affect the hold of HWE as the frequency of heterozygotes will shift.

A population under HWE and free from all disturbances, e.g. selection, migration, mutation, could be named an ideal population. In many occasions, the population size of such an ideal population is assumed to be infinite, where the probability of inbreeding under random mating could be neglected.

Random and Non-random Samples

The phrase ‘Random Sample’ normally indicates a sample that is collected through simple random sampling without replacement from a given population, where each individual has the same chance to be selected. For convenience, the sampling population could be assumed ideal, which is quite practical for population based association study in humans. If the sample size is large enough, the random sample thus collected could be expected to hold all properties that an ideal population has, i.e. HWE, free from selection, migration, recurrent mutation and genetic drift. Any random sample discussed in this thesis will be referred to as being randomly collected from an ideal population henceforward unless otherwise specified. Note that, even in presence of population stratification, it is still reasonable to assume samples are random in each subpopulation.

1.2.2 Linkage Based QTL Analysis

The genetic linkage describes the tendency of the haplotype of any two or more genetic loci to inherit together during the process of meiosis. Intuitively, such a tendency could be straightforward measured by the recombination fraction, denoted as r , i.e. the chance of two genetic loci to experience a recombination or segregation process during meiosis. It is easy to understand the range of r as $[0, 0.5]$, where the lower bound is acquired if two genetic loci are completely linked to each other and hence are always inherited as a single locus, and the upper bound is acquired if two genetic loci are completely unlinked, either from different chromosomes or distant separated on the same chromosome to allow them to segregate freely.

From the definition of genetic linkage given above, the more tightly linked to each other two genetic loci are, the smaller recombination fraction they will have, and we could hence establish statistical methods to map candidate QTLs of a specific trait through inferring the recombination fractions between a putative QTL and the genetic markers, i.e. statistically testing whether or not $r = 0.5$. In order to obtain an estimate of recombination fraction or its equivalences between any pair of genetic loci, family-based pedigree data are required, which is simply because the recombination event could only be observed during meiosis. Although the linkage analysis has been proved efficient for many Mendelian traits, especially Mendelian diseases caused by high-risk mutations in humans (Cui et al., 2010), the concern of the highly limited resolutions has been repeatedly mentioned and discussed (McMillan and Robertson, 1974; Lander and Botstein, 1989; Boehnke, 1994). It is easy to understand that since the recombination fraction is defined as the chance of segregation between two genetic loci during meiosis, while, without sufficient meiotic events, it would be impossible to accurately estimate r and hence to establish an informative statistical test (Silver, 1985). Hence, in order to improve the mapping resolutions, we may choose either or both of the following solutions to increase the number of recombination events: (1) increase the number

of individuals at each recombination generation, and (2) increase the number of segregating generations.

Due to both the ethical and practical limits, the pedigree data of humans could not be acquired through planned breeding in labs as plants and animals, and hence the collection of pedigree data for humans is far more challenging (Feingold, 2001). Due to the difficulty in collecting large pedigree samples, more and more geneticists turn to the association study as introduced in the following section for better mapping resolutions.

More details about the methods of performing linkage analysis please refer to Risch (1990a, b, c) and Feingold (2001).

1.2.3 Linkage Disequilibrium Based QTL Analysis

Assume that two alleles at a bi-allelic marker locus are denoted as M and m with allele frequencies p_M and p_m respectively, and similarly two alleles at a QTL denoted as A and a with allele frequencies p_A and p_a respectively (Table I-1). The difference between the observed and expected haplotype frequency under random pairing of marker and QTL alleles is defined as the coefficient of linkage disequilibrium (LD) between these two loci, i.e. $D = f_{MA} - p_M p_A$, where f_{MA} is the frequency of the haplotype MA . This simple algebraic equation shows that, if alleles M and A are randomly paired together in the population, it is easy to calculate $f_{MA} = p_M p_A$ and hence $D = 0$, while on the contrary, if alleles at the marker locus and QTL are not randomly segregating but tend to link to each other through, for example, physical linkage, it turns out $D \neq 0$.

Table I-1. The Settings of Marker and Trait Loci.

Locus	Marker		QTL	
Allele	M	m	A	a
Frequencies	p_M	p_m	p_A	p_a

As stated above, the disequilibrium parameter D is defined statistically rather than biologically. Here I would like to clarify the relationship between D and recombination frequency r before presenting a statistical model of LD based QTL mapping. Consider a random mating population free of mutation, immigration or generation overlapping. We can safely assume that the frequencies of marker allele M and QTL allele A at the next generation, say p'_M and p'_A respectively, are equal to those at the current generation. Because the haplotype MA at the next generation is either inherited from haplotype MA of the current generation without recombination or randomly formed by M and A gametes if crossover happens, the frequency of haplotype MA at the next generation could be calculated as $f'_{MA} = (1-r) \times f_{MA} + r \times p_M \times p_A$, where r is the recombination fraction between these two loci, from which the LD coefficient between the marker and trait loci at the next generation could be calculated as

$$D_n = f'_{MA} - p'_M p'_A = (1-r)(f_{MA} - p_M \times p_A) = (1-r)D. \quad (\text{I-1.1})$$

Formula (I-1.1) shows that the LD will decrease with a factor $1-r$ from one generation to the next, and if a population evolves for a certain number of generations, only those strongly linked loci, i.e. with very small r , could remain to be in significant linkage disequilibrium. Given such a feature, LD is hence eligible to detect the genetic linkage other than the recombination fraction, but with the potential of providing higher mapping resolutions. Ideally, the initial LD arises as a mutation was introduced into a population, but other events,

e.g. immigration, might also contribute to LD, the situation of which will be discussed in section I-1.2.4.

Denoting the LD parameter between the marker and QTL loci at generation 0 by D_0 and assuming allele frequencies keep constant across generations, the LD between the two loci at the T -th generation can be expressed as

$$D_T = (1-r)^T D_0. \quad (\text{I-1.2})$$

Formula (I-1.2) shows that the LD will decay with a factor of $1-r$ from one generation to the next. It is also clear that after segregating for a number of generations, only those closely linked loci, i.e. those with very small r , will remain in significant LD. Therefore, estimates of LD in a random mating population can be used to detect the genetic linkage between polymorphic loci. Unlike a family based linkage study which relies on pedigree information, LD analysis can be applied to a wide range of populations, with the potential of providing much higher mapping resolution because it exploits accumulated historical recombination events.

Besides the most prominent LD measure, D as defined above, there are several other measures of LD which are frequently used in the literatures. For example, Lewontin (1964) suggested to scale the standard measure D by its theoretical maximum or minimum, i.e.

$$D' = \begin{cases} \frac{D}{D_{\max}} & \text{if } D \geq 0 \\ \frac{D}{D_{\min}} & \text{if } D < 0 \end{cases}, \quad (\text{I-1.3})$$

where $D_{\max} = \min(p_M p_a, p_m p_A)$ and $D_{\min} = \max(-p_M p_A, p_m p_a)$. D' takes values in the domain $[0, 1]$ irrespective of the allele frequencies. With such a property, it would be more comparable and tractable to measure LD with D' rather than D between any genetic loci pairs.

Another commonly used alternative measurement of LD is defined as the correlation coefficient between alleles at two loci. If we assign alleles M and A with value 1, and alleles m and a with value 0, it could be easily shown that

$$E(M) = p_M, E(M^2) = p_M$$

$$E(A) = p_A, E(A^2) = p_A.$$

As $E(MA) = f_{MA}$, the correlation coefficient (r) between alleles M and A could be derived by

$$r = \frac{\sigma_{MA}}{\sigma_M \sigma_A} = \frac{f_{MA} - p_M p_A}{\sqrt{p_M p_m p_A p_a}} = \frac{D}{\sqrt{p_M p_m p_A p_a}}, \quad (\text{I-1.4})$$

where σ_{MA} is the covariance between alleles M and A , and σ_M^2 and σ_A^2 are the variance of alleles M and A respectively.

1.2.4 Advantages and Challenges of Linkage Disequilibrium Based QTL Analysis

Compared to the linkage based analysis, an association study does not require a known relationship between individuals and hence it could be applied to any population based data, which has the advantages of both easy access and abundant data to be collected (Williams-Blangero and Blangero, 2006). Moreover, the feature of LD as shown in formula (I-1.1) could lead to a much higher mapping resolution than a linkage based analysis, which is because the LD automatically inherits the historical recombination information of the population evolution, even if the data of those previous generations are actually not available, and the recombination fraction could only be evaluated from observed individuals.

Despite the advantages mentioned above, LD-based analysis encounters several new challenges.

Population stratification

The first challenge comes from the population stratification or admixture (Chakraborty and Smouse, 1988). Population stratification means a population of interest is composed of multiple subpopulations which are systematically different in allele frequencies. This phenomenon could naturally be introduced through the migration of genetically isolated populations, and could also be due to the collection of data from different genetic orientations, e.g. different allele frequencies at certain genetic loci. If such a stratified population evolves with interbreeding among its subpopulations, the new population hence formed would be referred to be with population admixture. Note here, an admixed population may also naturally evolve from an ideal population under the pressure of disruptive selections or positive assortative mating, and if such processes are extremely strong or persist long enough, genetically isolated subpopulations may eventually be generated and thus the population stratification remains. For simplicity but without loss of generality, only the case of population with two isolated subpopulation, i.e. with population stratification, will be closely evaluated below to illustrate the impact of the population stratification on association studies. Suppose there are two marker loci with alleles A, a and B, b respectively, and both loci have their allele frequencies denoted as $P_X(A), P_X(B)$ for a random mating population X and $P_Y(A), P_Y(B)$ for a random mating population Y . The coefficients of LDs in populations X and Y respectively could be simply given as $D_X = f_X(AB) - P_X(A)P_X(B)$ and $D_Y = f_Y(AB) - P_Y(A)P_Y(B)$ following the definition, where $f_X(AB)$ and $f_Y(AB)$ are the frequencies of haplotype AB in population X and Y respectively. If the sizes of X and Y are of proportion $m : 1 - m$, the LD coefficient in the stratified population could be calculated as:

$$\begin{aligned} D &= f(AB) - P(A)P(B) \\ &= mf_X(AB) + (1-m)f_Y(AB) - [mP_X(A) + (1-m)P_Y(A)][mP_X(B) + (1-m)P_Y(B)], \quad (I-1.5) \\ &= mD_X + (1-m)D_Y + m(1-m)[P_X(A) - P_Y(A)][P_X(B) - P_Y(B)] \end{aligned}$$

where D is the coefficient of LD in this stratified population and $f(AB)$, $P(A)$, $P(B)$ are the corresponding frequencies of haplotype AB , allele A and allele B respectively. It could be noticed that the first two terms of formula (I-1.5), i.e. mD_X and $(1-m)D_Y$, are due to the LDs in each subpopulation and the third term, i.e. $m(1-m)[P_X(A)-P_Y(A)][P_X(B)-P_Y(B)]$, is due to the population substructure. It is thus clear from formula (I-1.5) that even if there is no real LD between loci A and B in each subpopulation, i.e. $D_X=0$ and $D_Y=0$, a level of association between these two loci in the stratified population might still be observed unless either $P_X(A)=P_Y(A)$ or $P_X(B)=P_Y(B)$. Meanwhile, since the third term of formula (I-1.5) could either be positive or negative, such a structure effect might reduce the observed LD to almost null even with a real LD. It is hence very important to control the influence of population substructure in an accurate and reliable LD-based QTL analysis mapping. For convenience, the LD introduced through population stratification, i.e. the third term in formula (I-1.5), will be referred to as the structural LD henceforward.

Consider an admixed population, if generation of structural LD is not recurrent, e.g. due to one event of immigration, after tens of generations, the structural LD could only survive between loci with strong linkage. In this situation, structural LDs may only cause the detection of spurious association if they are newly introduced. On the other hand, as the structural LDs hence introduced will increase the LD between two closely linked loci permanently, it is possible that benefits from analysing an admixed population could be acquired if it has evolved for a sufficient long period free from those confounding factors (Risch, 1992; Zheng and Elston, 1999). However, if the generation of structural LD is recurrent, e.g. multiple immigrations or assortative mating events, the structural LD will be kept among all loci that were influenced by those processes, even if they are on different chromosomes. Since the positively assortative mating is fairly common in humans and many

other species, it would be worthwhile to recognise that loci not under the pressure of such a process are free from the structural LD (Redden and Allison, 2006), which could be easily understood through noticing that the positive assortative mating does not change the allele frequencies of the whole population, and hence even a stratified population was eventually formed, the allele frequency of any a locus free from the assortative mating process should be identical among all subpopulations in the absence of genetic drift. However, Redden and Allison (2006) warned that if multiple traits are jointly or systematically under the positive assortative mating, e.g. the good looking people are more likely to be married to smart and rich people, all related loci under such selections will tend to be associated with each other, which might hence complicate the situation of association studies for certain traits.

To adjust for the effect of population stratification in association studies, there are basically two different strategies. Firstly, assuming such effects are genome-wisely identical or similar, the global inflation factor λ could be computed to justify the structure as the method named as Genomic Control (GC) suggested by Devlin & Kathryn (1999). Secondly, the structure effect for each marker locus respectively could be evaluated given a known population structure which, if unknown, could be acquired through principal component analysis, abbreviated as PCA (Price et al., 2006), or Pritchard's STRUCTURE (Pritchard et al., 2000), and the second strategy is usually known as Structured Association (SA). Among all three methods, GC is the computationally easiest to implement. However, several studies have shown that GC may suffer certain loss of power if the inflation factor can not be assumed constant across the genome (Yu et al., 2006; Price et al., 2006; Rakovski and Stram, 2009), and hence it might be more practical to treat such effects locally. Of the other two, PCA is generally easier and much faster than Pritchard's STRUCTURE, whereas Pritchard's method enjoys the ability of inferring the probability of any individual being related to a particular subpopulation. However, as STRUCTURE assumes random mating in each of the

subpopulations, which might not be valid in inbreeding species, e.g. barley and rice, the application of STRUCTURE are hence limited to outbreeding species. Also, as the number of subpopulation may never be exactly indicated, the question of how to efficiently and properly integrate the information acquired from STRUCTURE is another issue under debate (Balding, 2006). Comparatively, PCA does not have such limitations and could be generally implemented for many datasets.

Genetic heterogeneity and cryptic relatedness

The second challenge mainly comes from the nature of complex traits, the genetic heterogeneity, where multiple genes are responsible for a similar or identical trait. If it is controlled by multiple genes, an ideal statistical model for a complex trait would be expected to integrate all genetic factors together and identify all candidate genes at once. However, as there is still little knowledge about those responsible genes, e.g. number, location, function and interaction, such an ideal statistical model is still a vain hope. One of the simplest solutions is to identify candidate genes one by one and the contributions from the rest, i.e. the background gene effect, could be treated as residual effects to the gene under test. If all individuals are independent of each other and the normality assumption could be applied, these residual effects could be absorbed into the non-biological terms and the statistical model will be fairly simple. However, in the presence of genetically related individuals, where the independent assumption does not stand, a mixed linear model with multivariate normal distribution could be implemented, where the kinship of relatives are introduced through the covariance matrix (Lange, 1978), and this idea was later introduced into the association studies for pedigree data, such as the method Quantitative Transmission Disequilibrium Test (QTDT) (Abecasis et al., 2000; Abecasis et al., 2001). This adoption of multivariate normal distribution enables the association studies in either pedigree data or population based data. However, it should be noted here that the population based data may not be guaranteed free

from kinship between individuals, for example, in many inbreeding species, i.e. barley and rice, individuals are highly related. In such a situation, the proper estimates of the kinship between any pair of individuals are crucial to establish a reliable association study.

The key issue to validate above methods is the assumption of the additive cumulation of effects from all responsible genes. Although this assumption seems quite reasonable if the phenotype is continuous and there is no prior information about the existence of epistasis, it might not stand initially if the phenotype is binary. For a complex trait with a binary phenotype, i.e. the situation of most diseases, the genetic effect from each responsible gene provides extra contribution to the prevalence of the phenotype of interest in the population (Risch, 1990a). It is pretty easy to understand that such contributions are not additive and the inclusion-exclusion principle has to be applied, for example, suppose the contributions of prevalence from two genes are denoted as α and β respectively, their joint contribution would be $\alpha + \beta - \alpha\beta = 1 - (1 - \alpha)(1 - \beta)$. Generally, such interactions, e.g. the term $\alpha\beta$ may not be ignored if the multiple gene effects are explicitly included in the model. One of the possible solutions for binary trait is to use the generalised linear mixed models (GLMM), where the interaction terms could be eliminated through certain transformations, and the kinship information between relatives could be hence implemented. However, as no analytic form of interpreting certain integrations regarding random effects could be acquired for GLMM, certain approximations or numerical integration techniques have to be involved (Breslow and Clayton, 1993). Due to both its mathematical complexity and computational demands, the GLMM has not been widely implemented into association study regarding the kinship of relatives.

Although the GC model actually considered both population stratification and cryptic relatedness between individuals, as has been discussed above, the general control of all

marker loci with the same inflation factor is so restricted that the statistical power might be largely reduced. One of the most comprehensive strategies to overcome both above two obstacles, i.e. the population stratification and the genetic heterogeneity with relatedness, was introduced by Yu, J. and his colleagues in 2006. In their model, the effect of population stratification was considered as a fixed effect and was corrected through the introduction of dummy variables; the background gene effects are treated as random effects, as suggested by Lange (1978) and Abecasis (2000), where each individual is assumed to have its own background gene effect randomly chosen from the same normal distribution, and hence the covariance matrix of this random effects could be written as $2\mathbf{K}\sigma^2$, where \mathbf{K} is the kinship matrix among individuals and σ^2 is the variance of random effects. In case the information about the population stratification and the relatedness among individuals are unknown, the first several principal component vectors could be introduced to replace the dummy variables in order to correct for population stratifications (Price et al., 2006) and the pairwise identity by descent (IBD) among individuals could be estimated from various strategies (Schork, 1993; Pong-Wong et al., 2001). However, although Yu's strategy is generally valid for continuous traits, as has been mentioned above, in the case of binary phenotypes, the additive model of background gene effects will be violated. As the phenotypes of many traits of interest are binary, e.g. most diseases, more reliable and accurate strategies that could properly account for the relatedness among individuals for binary phenotypes are still highly demanded.

Compared to the association study, both above obstacles have no or limited impact on the linkage analysis. For the population stratification, because the recombination fraction does not depend on allele frequencies, whereas LD does, it is normally believed that the linkage analysis is free from the influence of population stratification. However, Wang & Elston (2005) pointed out that population stratification should also have impacts on certain linkage analysis designs if part of the founders' genotypes are missing. For example, if the parental

genotypes of affected sib-pairs are missing, the allele frequencies will be deeply involved in estimating the IBD between sib-pairs, and hence the bias of allele frequencies introduced by population stratification will result in the bias of inferred linkage between markers and the putative QTL. Nevertheless, if all information of founders is available, the linkage analysis will be free of the population stratifications. One such design, transmission disequilibrium test (TDT), integrates both the benefit from linkage analysis and association studies that it is free from the population stratification as the linkage analysis and it confers the high mapping resolution of the association study (Spielman and Ewens, 1993; Ewens and Spielman, 1995). However, as it significantly increases the burden of sampling (Cardon and Palmer, 2003), it will not be so efficient to adopt the TDT unless the population stratification is clearly of concern. On the other hand, since the occurrence of disease in one family is very likely due to only one casualty gene, the influence of genetic heterogeneities in the linkage analysis is similar to that of population stratifications, and hence generally will not cause spuriously detected QTL. However, it has been pointed out that the genetic heterogeneity might result in the loss of power to detect a real QTL in the linkage analysis (Dizier et al., 2000). As the relationship among individuals is usually clear, the modelling of genetic heterogeneity with relatedness could be easily accomplished as suggested by Lange (1978) and Abecasis (2000). As a conclusion, the comparison between linkage analysis and association study is listed in Table I-2. Although the LD-based QTL analysis has above two main obstacles, both its mapping resolution and sampling advantages proclaim its superiority to the linkage-based QTL analysis, and hence the association study becomes more and more a dominant approach in QTL analysis recently.

Table I-2. The Comparison of Linkage Analysis and Association Study.

	Linkage Analysis	Association Study
Type of Data	Pedigree Data	Population and Pedigree Data
Sample Size	Small or Median	Large
Mapping Resolution	Low	High
Confounding Factors	None	Population Stratification Cryptic Relatedness Genetic Heterogeneity

1.3 Structure of the Thesis

This thesis could be divided into two parts, both of which share the same purpose of improving the performance of statistical methods in association studies while dealing with two major confounding factors, i.e. population stratification and non-random sampling.

In Chapter II-1, commonly used statistical methods for various study designs in association analyses are introduced. This is followed by Chapter II-2 which investigates and compares those existing methods in their performance and equivalence. Then a thorough investigation of the choice of trend coefficients in the Armitage's trend test is introduced in Chapter II-3 to optimize its statistical performance, where it will be shown that the genetic heterogeneity could largely influence the 'optimal' choices. Two strategies for controlling population stratification in Armitage's trend test are also introduced to enable joint analysis of data from multiple cohorts. Intensive simulation study and re-analysis of recently published Parkinson's disease case and control data demonstrate that the newly developed method confers significantly improved statistical power for detecting the associations compared to the original trend test method.

Chapter III presents two different likelihood based statistical strategies for association analyses. Chapter III-1 introduces a new strategy of inferring LD between two genetic loci from non-random samples. Such a strategy is later extended into the association analysis between the genetic marker and the QTL for case and control data in Chapter III-2. It will be shown that the above mentioned method for association study confers improved power and flexibility enabling different confounding factors of the disease trait to be tested. Chapter III-3 introduces a likelihood based association analysis which integrates the information from both LD and recombination fraction in order to predict QTLs within marker intervals.

1.4 Preliminaries

1.4.1 Statistical hypothesis test

When referring to a statistical test, people are mainly talking about a statistical hypothesis test, which is designed to distinguish probability distributions from one of which a series of questioned random variables are generated. In order to perform such a test, a paradigm developed by Neyman and Pearson (Rice, 1994) is the most commonly used approach. The probability distributions are divided into two groups, say null hypothesis denoted by H_0 and alternative hypothesis denoted by H_A , and a statistic $T(\mathbf{X})$ of the question sample values \mathbf{X} is obtained to determine whether or not to reject H_0 by comparing $T(\mathbf{X})$ to pre-set acceptance and rejection regions. While applying the Neyman-Pearson paradigm, four kinds of outputs might be yielded as shown in Table I-3. It is common to denote the probability of type I and II errors as α and β respectively. The probability that the null hypothesis (H_0) is rejected when the alternative hypothesis (H_A) is true is referred to as the power of the test, which clearly equals to $1 - \beta$ from Table I-3. A ‘good’ test would have both α and β very small, however, given a fixed sample size, these two probabilities are traded off with each other; as

α decreases, β must increase, and vice versa. In practice, α is fixed to a preset value and a statistic $T(\mathbf{X})$ is then constructed to minimize β , and hence performances of different tests are compared for their power under this scheme.

Table I-3. Layouts of a Statistical Hypothesis Test under the Neyman-Pearson Paradigm.

	Accept Null Hypothesis	Reject Null Hypothesis
Null Hypothesis (H_0) is True	Correct Decision $P=1-\alpha$	Type I Error (False Positive) $P=\alpha$
Alternative Hypothesis (H_A) is True	Type II Error (False Negative) $P=\beta$	Correct Decision $P=1-\beta$

Along with the paradigm, the Neyman-Pearson Lemma was introduced to construct an optimal statistic $T(\mathbf{X})$ to minimize β , and this Lemma states ‘among all tests with a given probability of a type I error, the likelihood ratio test minimizes the probability of a type II error’ (Rice, J.A. 1995), where the definition of likelihood ratio test will be given later. This Lemma provides a clear strategy to construct an optimal statistic, and jointly with the Neyman-Pearson paradigm any statistical hypothesis test could be constructed if only the likelihood functions under both null and alternative hypotheses are available. It could be noted here that likelihood ratio test is an approach or strategy to establish a statistical test rather than a simple statistical test as its name indicated, where there are several other strategies available, such as the Bayesian approach. However, due to the advantages of the likelihood ratio test, where it is ‘almost always applicable and is also optimal in some cases’ (Casella and Berger, 2002), the likelihood ratio test is probably the most commonly used approach.

In many occasions, the statistical hypotheses are given as:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_A : \mu &\neq \mu_0 \end{aligned}$$

and a corresponding test is constructed to examine whether μ is significantly different from μ_0 . With such a concept, the statistical hypothesis test is also named ‘test of significance’ by Fisher (1925a), and the rate of type I error, α , is hence called the ‘significance level’ of the test. Note here, since the null hypothesis and alternative hypothesis are mutually exclusive to each other, for convenience only the null hypothesis will be mentioned afterward unless otherwise necessary.

With the above hypothesis, suppose there are random variables X_1, \dots, X_n following $N(\mu, \sigma^2)$ independently, where μ is unknown, estimated as sample mean \bar{X} , and σ^2 is known, a significant level α means $H_0 : \mu = \mu_0$ would be rejected if $|\bar{X} - \mu_0| > x_0$, where x_0 is determined by equation $P(|\bar{X} - \mu_0| > x_0) = \alpha$. If $H_0 : \mu = \mu_0$ is true, the equation could be solved as $x_0 = \sigma_{\bar{X}} z_{\alpha/2}$, where $\sigma_{\bar{X}}$ is the standard deviation of \bar{X} , i.e. $\sigma_{\bar{X}} = \sigma / \sqrt{n}$, and $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution Z . H_0 is thus accepted if $|\bar{X} - \mu_0| \leq \sigma_{\bar{X}} z_{\alpha/2}$, and hence the $100(1-\alpha)\%$ confidence interval for μ_0 is $[\bar{X} - \sigma_{\bar{X}} z_{\alpha/2}, \bar{X} + \sigma_{\bar{X}} z_{\alpha/2}]$. From the above derivation, $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$, which follows the standard normal distribution if $\bar{X} - \mu$ and $\sigma_{\bar{X}}^2$ are distributed independently, is clearly a statistic test of the given hypotheses with the acceptance region $[-z_{\alpha/2}, z_{\alpha/2}]$, and is hence known as a *Z-statistic* test.

If σ^2 is unknown in the above example, an unbiased estimate of the population variance

could be given as $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$, and by keeping all the other procedures the same, in the

new example, the $100(1-\alpha)\%$ confidence interval for μ_0 is $[\bar{X} - s_{\bar{X}} t_{n-1, \alpha/2}, \bar{X} + s_{\bar{X}} t_{n-1, \alpha/2}]$,

where $s_{\bar{X}}$ is the standard error of \bar{X} that $s_{\bar{X}} = s/\sqrt{n}$ and $t_{n-1, \alpha/2}$ is the upper $\alpha/2$ point of

the Student's t distribution with $n-1$ degrees of freedom. Similarly to the Z -statistic,

$T = \frac{\bar{X} - \mu}{s_{\bar{X}}}$ follows the Student's t distribution with $n-1$ degrees of freedom, and is hence

named the T -statistic test of the hypotheses with the acceptance region $[-t_{n-1, \alpha/2}, t_{n-1, \alpha/2}]$. Note

here, since $\lim_{n \rightarrow \infty} s_{\bar{X}}^2 = \sigma_{\bar{X}}^2$, given a sufficiently large sample size n , the T -statistic and the Z -

statistic are asymptotically equivalent.

In the above statistical hypotheses tests, the null hypothesis $H_0 : \mu = \mu_0$ will be rejected while

μ is either sufficiently large or small, and such a statistical test could be referred as a two-

tailed test. Alternatively, if the null hypothesis changes to $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$, the

corresponding statistical test would be one-tailed, which could be easily derived following a

similar procedure. It could also be noted here that the Z statistic could be alternatively derived

through the likelihood ratio approach (Casella and Berger, 2002).

1.4.2 Maximum Likelihood Estimator (MLE¹) and Likelihood Ratio Test (LRT)

Suppose there are random variables X_1, \dots, X_n with a joint density or frequency function $f(X_1, X_2, \dots, X_n | \theta_1, \theta_2, \dots, \theta_k)$, where $k, n \in \mathbb{Z}^+$, i.e. positive integers, and θ denote unknown parameters, and the corresponding likelihood function is defined as:

$$\begin{aligned} L(\theta | X) &= L(\theta_1, \theta_2, \dots, \theta_k | X_1, X_2, \dots, X_n) \\ &= f(X_1, X_2, \dots, X_n | \theta_1, \theta_2, \dots, \theta_k) \end{aligned}$$

If X_i are assumed to be i.i.d. (independent identically distributed) the likelihood function could be rewritten as:

$$L(\theta | X) = \prod_{i=1}^n f(X_i | \theta_1, \theta_2, \dots, \theta_k). \quad (\text{I-1.6})$$

Because the likelihood function is established as the joint density or frequency function of random variables, it represents the probability of observing such data given a set of unknown parameters. The idea of a maximum likelihood procedure is to inverse the logical dependence between data and parameters by evaluating the most possible or likely values of unknown parameters given the observed data. Note here, the likelihood function (I-1.6) is often referred as the complete likelihood function, where the joint probability of all observations, \mathbf{X} , is adopted. On the other hand, other forms, i.e. prospective and retrospective, of likelihood functions are also available, where a part of the whole observations \mathbf{X} , says \mathbf{X}_1 , could be assumed to be conditioned on the rest observations, says \mathbf{X}_2 , either prospectively or retrospectively, and the conditional density or frequency function, i.e. $f(\mathbf{X}_1 | \mathbf{X}_2)$, is adopted instead of the joint one, i.e. $f(\mathbf{X})$. Despite the equivalence between prospective and retrospective likelihood functions under certain restrictions that have been intensively discussed (Prentice and Pyke, 1979; Weinberg and Wacholder, 1993; Roeder et al., 1996; Murphy and Van der Vaart, 2000), it has been pointed out that if the sufficient statistics of

¹ Without causing any misunderstanding, MLE, as well as later mentioned abbreviations, will be used as the abbreviation for both the estimator and its corresponding value, the estimate.

X_2 could be directly given irrespective of parameters $\boldsymbol{\theta}$, and such statistics are thus ancillary to $\boldsymbol{\theta}$, the conditional likelihood function is statistically equivalent to the complete one (Kalbfleisch and Sprott, 1970; Sprott, 1975; Smyth and Verbyla, 1996).

In the absence of boundaries, if the likelihood function is differentiable on its domain, the candidate MLE, says $\boldsymbol{\theta}_c$ which maximizes $L(\boldsymbol{\theta} | \mathbf{X})$ locally, should meet the necessary condition,

$$\frac{\partial L(\boldsymbol{\theta}_c | \mathbf{X})}{\partial \theta_i} = 0, \quad i = 1, \dots, k. \quad (\text{I-1.7})$$

Note that equations (I-1.7) are not the sufficient condition to maximize $L(\boldsymbol{\theta} | \mathbf{X})$, not even to ensure local maxima/minima (extrema), and hence in order to search the global maximum, the second (partial) derivative equations could be applied in presence of the second-order (partial) derivatives of $L(\boldsymbol{\theta} | \mathbf{X})$ that if $\frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_i \partial \theta_j}$ exists for any $0 < i, j < k$, $i, j \in Z^+$, the Hessian matrix of the second-order partial derivatives could be established as:

$$\mathbf{H}(L(\boldsymbol{\theta} | \mathbf{X})) = \begin{pmatrix} \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_1^2} & \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_2^2} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_k \partial \theta_1} & \dots & \dots & \frac{\partial^2 L(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_k^2} \end{pmatrix}. \quad (\text{I-1.8})$$

If all the eigenvalues of $\mathbf{H}(L(\boldsymbol{\theta}_c | \mathbf{X}))$ are positive, i.e. $\mathbf{H}(L(\boldsymbol{\theta}_c | \mathbf{X}))$ is positively definitive, it could be concluded that $\boldsymbol{\theta}_c$ is the global maximum of $L(\boldsymbol{\theta} | \mathbf{X})$ in the absence of boundary. However, if some eigenvalues of $\mathbf{H}(L(\boldsymbol{\theta}_c | \mathbf{X}))$ equal 0 while the rest are positive, $\mathbf{H}(L(\boldsymbol{\theta}_c | \mathbf{X}))$ is a singular matrix and is inconclusive, so further effort is required to infer whether $\boldsymbol{\theta}_c$ is the MLE or not in this circumstance. In case of boundaries applied, both

boundaries and local maxima of $L(\boldsymbol{\theta} | X)$ have to be checked to determine the true global maximum. In practice, people would prefer to deal with the natural logarithm of $L(\boldsymbol{\theta} | X)$, $l(\boldsymbol{\theta} | X)$, rather than $L(\boldsymbol{\theta} | X)$ itself, where this transformation introduce several advantages. Firstly, $l(\boldsymbol{\theta} | X)$ shares any coincident extrema with $L(\boldsymbol{\theta} | X)$ given that the logarithm function is monotonically increasing on its domain $(0, +\infty)$; Secondly, $l(\boldsymbol{\theta} | X)$ has superb mathematical advantages over $L(\boldsymbol{\theta} | X)$ given that the likelihood function takes the form of function (I-1.6) in most of the time; Thirdly, this transformation enables the use of several important statistical methods, such as Fisher's information matrix (Fisher, 1925b) and Rao's score test (Rao, 1948).

However, equations (I-1.7) may not be directly solved analytically due to the complexity of the likelihood function, and hence either numerical or indirect estimates are required. In such a case, the Expectation-Maximization algorithm (EM algorithm) could be applied to find the MLE iteratively (Dempster et al., 1977). Although EM algorithm was initially designed to find the MLE from incomplete data, e.g. the missing marker genotypes, it could equivalently deal with unobserved data, such as the QTL genotypes. In early 1990s, Meng, X. & Rubin, D. B. (1993) introduced a class of generalised EM algorithm called Expectation Conditional Maximization algorithm (ECM algorithm) that uses a sequence of sub-Conditional Maximization (CM) steps instead of a single M-step, which may be still rather complicated, and hence it could simplify the computation of MLE further. The procedure of performing ECM algorithm to deal with unobserved data could be briefly introduced as follow:

Given observed data X and unobserved data Y , the full data log-likelihood function of parameters $\boldsymbol{\theta}$ could be given as $l(\boldsymbol{\theta} | X, Y) = \log \Pr(X, Y | \boldsymbol{\theta})$. The $t+1$ -st E step of ECM

algorithm calculates the expectation of the full data log-likelihood conditioned on unobserved data Y with parameters θ^t from the t -th CM step, and hence it could be given as:

$$l_c(\theta | X, \theta^t) = \int \log[\Pr(X, Y | \theta)] f(Y | X, \theta^t) dY \text{ if } Y \text{ is continuous}$$

Or (I-1.9)

$$l_c(\theta | X, \theta^t) = \sum_Y \log[\Pr(X, Y | \theta)] \Pr(Y | X, \theta^t) \text{ if } Y \text{ is discrete,}$$

where $f(Y | X, \theta^t)$ and $\Pr(Y | X, \theta^t)$ are the corresponding density or frequency functions of unobserved data Y conditional on observed data X and parameters θ^t . The $t+1$ -st CM step then updates each parameter one by one through maximizing $l_c(\theta | X, \theta^t)$ following the sequence

$$\begin{aligned} l_c(\theta_1^{t+1}, \theta_2^t, \dots, \theta_k^t | X, \theta^t) &\geq l_c(\theta^t | X, \theta^t) \\ l_c(\theta_1^{t+1}, \theta_2^{t+1}, \theta_3^t, \dots, \theta_k^t | X, \theta^t) &\geq l_c(\theta_1^{t+1}, \theta_2^t, \dots, \theta_k^t | X, \theta^t) \\ &\dots \\ l_c(\theta^{t+1} | X, \theta^t) &\geq l_c(\theta_1^{t+1}, \dots, \theta_{k-1}^{t+1}, \theta_k^t | X, \theta^t). \end{aligned}$$

Given a proper initial point θ_0 , the iterative product, θ^t , of t -th ECM-step is promising to converge to the MLE of θ , and hence for a given convergence criterion δ achieved at t_0 -th iteration, i.e. $\delta \geq l(\theta^{t+1} | X) - l(\theta^t | X)$, for any $t > t_0$, the algorithm will yield the approximate MLE of θ such that $\theta^{t_0} \approx \theta_c$.

Obviously, the ECM procedure does not require the Hessian Matrix but only the second-order partial derivative of each sub-CM-step is required to maximize each sub-step, and hence not only the computational requirement is reduced but the issue of singular Hessian Matrix could also be avoided, because it could be much easier to calculate higher-order partial derivatives with respect to a particular parameter and conclude its exact characteristic at a certain point than to evaluate the characteristics of several parameters simultaneously. It could be noticed

that following the procedure of functions (I-1.9) to construct the CM-step at each iteration, the partial derivatives of $l_c(\boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\theta}')$ will keep the same form but with updated $f(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta}')$ (or $\Pr(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta}')$) and $\boldsymbol{\theta}'$, and hence in the following chapters only $f(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta}')$ or $\Pr(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta}')$ will be updated in E-steps instead of evaluating an updated $l_c(\boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\theta}')$.

As mentioned in section 1.4.1, the likelihood ratio approach is a powerful method in statistical inference. In general, the LRT statistic with $H_0 : \boldsymbol{\theta} \in \Theta_0$ and $H_A : \boldsymbol{\theta} \in \Theta$ could be expressed as

$$\lambda(\mathbf{X}) = \frac{\sup_{\Theta_0} L(\boldsymbol{\theta} | \mathbf{X})}{\sup_{\Theta} L(\boldsymbol{\theta} | \mathbf{X})}, \quad (\text{I-1.10})$$

with a rejection region $[0, c]$, where Θ_0 and Θ denote the null and full parameter spaces, and c , with the definition domain $0 \leq c \leq 1$, could be specified given the exact form of $\lambda(\mathbf{X})$ with a certain significant level. Suppose $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ are the MLEs of $\boldsymbol{\theta}$ in parameter space Θ_0 and Θ , respectively, the function (I-1.10) could be rewritten into

$$\lambda(\mathbf{X}) = \frac{L(\hat{\boldsymbol{\theta}}_0 | \mathbf{X})}{L(\hat{\boldsymbol{\theta}} | \mathbf{X})}, \quad (\text{I-1.11})$$

which hence reveals the relationship between MLE and LRT.

With formula (I-1.11) and certain assumptions, the distribution of $\lambda(\mathbf{X})$ might be derived analytically and hence a direct statistical hypothesis test could be performed, e.g. *Z-statistic* given the assumption of normality. However, it is really difficult or even impossible to do so if the likelihood function is very complicated. In such a situation, a widely used asymptotic LRT could be performed by acknowledging that

$$-2 \log \lambda(\mathbf{X}) \sim \chi_{df}^2, \quad (\text{I-1.12})$$

where df is determined by the difference of dimensions between Θ_0 and Θ (Rice, 1994).

Alternative to the LRT, another likelihood-based Rao's score test (Rao, 1948) could be established under the null hypothesis. Assigning

$$\mathbf{U}(\boldsymbol{\theta}) = \left(\frac{\partial l(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_1} \quad \dots \quad \frac{\partial l(\boldsymbol{\theta} | \mathbf{X})}{\partial \theta_k} \right)^T,$$

it could be derived

$$\mathbf{U}^T(\hat{\boldsymbol{\theta}}_0) \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_0) \mathbf{U}(\hat{\boldsymbol{\theta}}_0) \sim \chi_{df}^2,$$

where the df is defined the same way as in formula (I-1.12) of the LRT, \mathbf{U} is often referred as the score matrix, and $\mathbf{I} = -E(\mathbf{H} | \hat{\boldsymbol{\theta}}_0) = E(\mathbf{U}\mathbf{U}^T | \hat{\boldsymbol{\theta}}_0)$ is the Fisher's Information Matrix.

Note here, since χ^2 distribution with 1 degree of freedom is defined as Z^2 , where $Z \sim N(0,1)$, χ^2 test with 1 degree of freedom is hence identical with Z test. For consistence purpose, the χ^2 test with 1 degree of freedom will be mainly used rather than the Z test throughout this thesis.

1.4.3 Linear Regression Analysis

Linear regression model is the most widely used strategy to perform an association-based QTL mapping given its flexibility, manipulability and widely accepted validity. Generally, a linear regression model would have the following form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \text{ with } E(\mathbf{y}) = \mathbf{X}\mathbf{b} \text{ or } E(\boldsymbol{\varepsilon}) = 0, \quad (\text{I-1.13})$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ is the $n \times 1$ column vector of dependent variables, \mathbf{b} is the $m \times 1$ column vector of regression coefficients, \mathbf{X} is the $n \times m$ matrix of independent variables, and $\boldsymbol{\varepsilon}$ is a $n \times 1$ column vector of the deviates defined as $\mathbf{y} - E(\mathbf{y})$, often referred to as the residual term. The variance-covariance matrix, \mathbf{V} , of $\boldsymbol{\varepsilon}$ defines the statistical property of function (I-1.13). If $\mathbf{V} = \sigma^2 \mathbf{I}$, where \mathbf{I} is the identity matrix, function (I-1.13) gives the fixed effect model,

where especially, while $m=2$, function (I-1.13) gives the simple linear regression model.

Alternatively, if \mathbf{V} could be decomposed into $k+1$ partitions that $\mathbf{V} = \sum_{i=1}^k \sigma_i^2 \mathbf{\Omega}_i + \sigma_e^2 \mathbf{I}$, where

each $\mathbf{\Omega}$ is a $n \times n$ symmetric matrix with its elements $|\Omega_{ij}| \leq 1$ and $\Omega_{ii} = 1$ for any

$i, j = 1, 2, \dots, n$, function (I-1.13) gives the mixed model. Another form of function (I-1.13), i.e.

random effect model, is also available while $\mathbf{V} = \sum_{i=1}^k \sigma_i^2 \mathbf{\Omega}_i + \sigma_e^2 \mathbf{I}$ and \mathbf{Xb} is a constant. Since

both fixed and random effect models are special cases of the mixed model, it is convenient to start with introduction of the mixed model.

Alternative to function (I-1.13), the mixed model could be written into the form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \quad (\text{I-1.14})$$

with assumptions:

$$E \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad Var \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \sigma_e^2$$

where \mathbf{Z} could be partitioned into k sub-matrices that $\mathbf{Z} = (\mathbf{Z}_1 \dots \mathbf{Z}_k)$, of which \mathbf{Z}_i is a

$n \times m_k$ design or incidence matrix of a full column rank with its element equalling 1 or 0; \mathbf{u} is

the $m \times 1$ column vector of random effects corresponding to \mathbf{Z} and could hence be partitioned

into $\mathbf{u}^T = (\mathbf{u}_1^T \dots \mathbf{u}_k^T)$, where \mathbf{u}_i is a $m_k \times 1$ column vector corresponding to \mathbf{Z}_i and

$$m = \sum_i m_k; \quad \mathbf{G} = \begin{pmatrix} \gamma_1 \mathbf{G}_1 & & \mathbf{0} \\ & \gamma_2 \mathbf{G}_2 & \\ \mathbf{0} & & \ddots & \\ & & & \gamma_k \mathbf{G}_k \end{pmatrix}, \quad \text{where } \gamma_i = \sigma_e^2 / \sigma_i^2, \text{ and } \mathbf{G}_i = \mathbf{u}_i \mathbf{u}_i^T \text{ is a matrix}$$

need to be specified before use. By noticing that $\mathbf{\Omega}_i = \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T$, it could be shown that

$$\sum_i \mathbf{Z}_i \mathbf{u}_i \mathbf{u}_i^T \mathbf{Z}_i^T = \sigma_i^2 \sum_i \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T = \sigma_e^2 \sum_i \gamma_i \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T \text{ and } \mathbf{V} = \sigma_e^2 (\mathbf{I} + \sum_i \gamma_i \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T). \text{ It is clear}$$

that the variance component set $\Theta : \{\sigma_e^2, \sigma_i^2 \mid i=1, k\}$ is functionally identical to $\Theta' : \{\sigma_e^2, \gamma\}$ to

define the exact form of \mathbf{V} , where $\boldsymbol{\gamma}$ is defined as a $k \times 1$ column vector with its i th element γ_i , and hence for convenience, both $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ will be regarded as the variance components in following discussions.

Before we carry on deep discussion of estimating variance components in mixed model, several simpler cases will be discussed here where the variance components are easy to be handled.

The best linear unbiased estimator (BLUE) of regression coefficients \mathbf{b} of function (I-1.14) is given by:

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \quad (\text{I-1.15})$$

and the variance of this estimate is

$$\text{Var}(\hat{\mathbf{b}}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}, \quad (\text{I-1.16})$$

where \mathbf{V} must not be singular and \mathbf{X} must be of full column rank, or certain kinds of generalised inverse matrices have to be adopted. Searle (1971) emphasized that the estimates acquired by the use of generalised inverse matrix, in case either \mathbf{V} or $\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}$ is singular, cannot be named as estimator given that the corresponding solution from a generalised inverse matrix relies on the exact form of the chosen generalised inverse matrix and hence in later chapters certain modifications will be presented in order to avoid the singularity and retain an unique estimator.

Considering a special case where $\mathbf{V} = \sigma_e^2 \mathbf{I}$, function (I-1.14) is the mixed model and reduced to the form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (\text{I-1.17})$$

with assumptions $E(\mathbf{e})=0$ and $Var(\mathbf{e})=\sigma_e^2\mathbf{I}$. The corresponding BLUE of regression coefficients \mathbf{b} now is

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (\text{I-1.18})$$

with its variance

$$Var(\hat{\mathbf{b}}) = (\mathbf{X}^T \mathbf{X})^{-1} \hat{\sigma}_e^2, \quad (\text{I-1.19})$$

and an unbiased estimator of σ_e^2 could be directly given as

$$\hat{\sigma}_e^2 = \frac{1}{n-m} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}). \quad (\text{I-1.20})$$

With the assumptions that the estimator of \mathbf{b} , $\hat{\mathbf{b}}$, is of normal distribution, i.e. $\hat{\mathbf{b}} \sim N(\mathbf{b}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma_e^2)$, and statistics $\hat{\mathbf{b}}$ and $\hat{\sigma}_e^2$ are independent with each other, a series of t -tests or χ^2 tests against the null hypothesis $H_0: b_i = 0$ could hence be performed for each element of $\hat{\mathbf{b}}$.

In a special case where $m=2$, function (I-1.17) could be reduced further to the simple linear regression (SLR) that:

$$\mathbf{y} = \left(\mathbf{1} \mid \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right) (\alpha \quad \beta) + \mathbf{e} \quad (\text{I-1.21})$$

with assumption $E(\mathbf{e})=0$ and $Var(\mathbf{e})=\sigma_e^2\mathbf{I}$. Formulae (I-1.18), (I-1.19) and (I-1.20) could hence be reduced to

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{\hat{\sigma}_x^2} \begin{pmatrix} \hat{\sigma}_x^2 \bar{y} - \hat{\sigma}_{xy} \bar{x} \\ \hat{\sigma}_{xy} \end{pmatrix}, \quad (\text{I-1.22})$$

$$Var \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{\hat{\sigma}_e^2}{(n-1)\hat{\sigma}_x^2} \begin{pmatrix} E(x^2) & -E(x) \\ -E(x) & 1 \end{pmatrix}, \quad (\text{I-1.23})$$

$$\hat{\sigma}_e^2 = \frac{n-1}{n-2} \hat{\sigma}_y^2 (1 - \hat{\rho}_{xy}^2), \quad (\text{I-1.24})$$

where $\hat{\sigma}_{xy}$ is the estimator of covariance between x and y , and $\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$ is the estimator of correlation between x and y . Given the assumption that $\hat{\beta} \sim N(\beta, \frac{\sigma_e^2}{n\sigma_x^2})$, with a sufficient

large n , a χ^2 test for the significance of the null hypothesis $H_0: \beta = 0$ could be established

as $\frac{\hat{\beta}^2}{\text{Var}(\hat{\beta})} \sim \chi_{df=1}^2$. Note here, since the test is performed under the null hypothesis, where the

null hypothesis $\beta = 0$ implies that \mathbf{y} and \mathbf{X} are uncorrelated, it could hence be derived that

$E_{null}(\hat{\rho}_{xy}) = 0$ and $\hat{\sigma}_{e,null}^2 = \frac{n-1}{n-2} \hat{\sigma}_y^2$, and with a large sample size n , the test statistic could be

further asymptotically written as $\frac{\hat{\beta}^2}{\text{Var}(\hat{\beta})} \approx n \hat{\rho}_{xy}^2 \sim \chi_{df=1}^2$. It should be also noted here that in

certain circumstances, the null hypothesis may be $H_0: \beta = \eta$, where $\eta \neq 0$, as will be

discussed in Chapter II-3, where the corresponding χ^2 test approximately takes the form

$\frac{(\hat{\beta} - \tilde{\eta})^2}{\text{Var}(\hat{\beta})} \sim \chi_{df=1}^2$, where we use the estimation of η , i.e. $\tilde{\eta}$, rather than its estimator, i.e. $\hat{\eta}$, to

indicate that we treat η as a known parameter and its variance will not be included in the

denominator of the χ^2 test.

By now, none of above derivations needs the normality assumption that $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$ and

$\mathbf{u} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{G})$, under which following results could be obtained straightforward by given

definition (I-1.14) (Searle, 1971):

$$\text{a. } \mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \sigma_e^2 (\mathbf{I} + \sum_i^k \gamma_i \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T))$$

b. $\hat{\mathbf{b}} \sim N(\mathbf{b}, (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1})$, $\mathbf{V} = \sigma_e^2 (\mathbf{I} + \sum_i^k \gamma_i \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T)$ and

c. $\hat{\mathbf{b}}$ are distributed independently with $\hat{\sigma}_e^2$.

It could be noticed here that results b and c are coincident with the assumptions being given to validate a χ^2 test against the null hypothesis $H_0: b_i = 0$. Moreover, an *F-statistic* test is also available to test for significance of the null hypothesis $H_0: \mathbf{b} = 0$, e.g. different genotypes are assumed to have different genetic effects.

The normality assumption also enables the use of MLE and LRT for the mixed model, and equation (I-1.14) could be alternatively expressed in the likelihood function under the multivariate normal distribution:

$$L(\mathbf{b}, \gamma, \sigma_e^2 | \mathbf{y}, \mathbf{X}) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})\right). \quad (\text{I-1.25})$$

The MLE of \mathbf{b} from function (I-1.25) could be given by solving:

$$\begin{aligned} \frac{\partial \log L}{\partial \mathbf{b}} &= \frac{\partial (\mathbf{y} - \mathbf{X}\mathbf{b})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})}{\partial \mathbf{b}} \\ &= -2(\mathbf{y} - \mathbf{X}\mathbf{b})^T \mathbf{V}^{-1} \mathbf{X} \\ &= 0 \end{aligned},$$

which yields

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}.$$

This result is identical with the BLUE as given by formula (I-1.15).

Unlike the fixed effect model, where $\mathbf{V} = \sigma_e^2 \mathbf{I}$, the $\hat{\mathbf{b}}$ given as BLUE from formula (I-1.14) or MLE from formula (I-1.25) depends on the exact form of covariance matrix \mathbf{V} , and hence γ , of which the i th element γ_i is defined at function (I-1.14), has to be estimated before or simultaneously with an estimate of \mathbf{b} being properly derived. However, to estimate variance

components in a mixed model may be rather difficult especially while the data is unbalanced, where no close form of estimators is available. The balanced data is defined with an equal sub-class sample size while the unbalanced data is otherwise. Although, there are many different strategies to estimate the variance components, for convenience, two most famous methods will be focused here, i.e. MLE and restricted maximum likelihood (REML) estimator, and some of the other methods will also be mentioned.

Following the maximum likelihood paradigm (I-1.7) and (I-1.8), the MLE of likelihood function given by function (I-1.25) should meet the following requirement:

$$\frac{\partial l}{\partial \mathbf{b}} = \frac{1}{\sigma_e^2} (\mathbf{X}^T \mathbf{H}^{-1} \mathbf{y} - (\mathbf{X}^T \mathbf{H}^{-1} \mathbf{X}) \mathbf{b}) = 0 \quad (\text{I-1.26})$$

$$\frac{\partial l}{\partial \sigma_e^2} = -\frac{n}{2\sigma_e^2} + \frac{1}{2\sigma_e^4} (\mathbf{y} - \mathbf{Xb})^T \mathbf{H}^{-1} (\mathbf{y} - \mathbf{Xb}) = 0 \quad (\text{I-1.27})$$

$$\frac{\partial l}{\partial \gamma_i} = -\frac{1}{2} \text{Tr}(\mathbf{Z}_i^T \mathbf{H}^{-1} \mathbf{Z}_i) + \frac{1}{2\sigma_e^2} (\mathbf{y} - \mathbf{Xb})^T \mathbf{H}^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \mathbf{H}^{-1} (\mathbf{y} - \mathbf{Xb}) = 0, \quad (\text{I-1.28})$$

where $l = \log L$, $\mathbf{H} = \mathbf{V} / \sigma_e^2 = (\mathbf{I} + \sum_i^k \gamma_i \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T)$ and $\mathbf{Z} = (\mathbf{Z}_1 \dots \mathbf{Z}_k)$. Equations (I-1.26)

and (I-1.27) could be solved explicitly to be

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{H}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{H}^{-1} \mathbf{y} \quad (\text{I-1.29})$$

$$\hat{\sigma}_e^2 = \frac{1}{n} (\mathbf{y}^T \mathbf{H}^{-1} \mathbf{y} - \mathbf{y}^T \mathbf{H}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{H}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{H}^{-1} \mathbf{y})). \quad (\text{I-1.30})$$

However, the explicit forms of equations (I-1.28) are available only under certain circumstances even with a balanced design of data (Miller, 1977) and the corresponding exact analytic solutions of variance components have been given by Szatrowski & Miller (1980),

also see Corbeil & Searle (1976a). For unbalanced data, where no analytic MLE of variance components is available, an iterative procedure introduced by Hartley & Rao (1967) could be implemented to calculate the variance components numerically. Later, Hemmerle & Hartley (1973) introduced a computing algorithm, i.e. the W-transformation, which could reduce the dimension of matrices from the sample space into the parameter space and hence largely reduce the tremendous computing load of inverting the covariance matrix \mathbf{V} , where they also noted that their algorithm would not yield negative estimates of variance components. However, the MLEs of variance components for unbalanced data have been proved to be biased in many occasions (Corbeil and Searle, 1976a; Wu et al., 2001). Although the iterative procedure introduced by Hemmerle & Hartley (1973) will not yield negative estimates of variance components, the MLE might eliminate the interaction terms during the iterative process and hence yield a zero estimate for those terms under certain circumstances even while such effect does exist (Corbeil and Searle, 1976a). Note that these interaction terms are not included in the mixed model as introduced at function (I-1.14).

In order to deal with the bias introduced by MLE, the restricted maximum likelihood (REML) estimators of variance components was introduced by Corbeil & Searle (1976b), which are believed to retain less bias than the MLEs (Wu et al., 2001). The idea of REML was adopted from Patterson, H. D. and Thompson, R. (1971) to partition the likelihood function (I-1.25) into two parts, one of which is free from the fixed effect and hence the variance components could be estimated straight without the interference from the fixed effect. Initially, Patterson and Thompson suggested the transformation as $\begin{pmatrix} \mathbf{S} \\ \mathbf{X}^T \mathbf{H}^{-1} \end{pmatrix} \mathbf{y}$, where $\mathbf{S} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is symmetric and idempotent. It is easy to show that $\mathbf{S}\mathbf{X}$ is null and hence $\mathbf{S}\mathbf{y} \sim N(0, \mathbf{S}\mathbf{H}\mathbf{S}\sigma^2)$, which is free from the fixed effect. Similarly it could be shown that

$\mathbf{H}^{-1}\mathbf{X}\mathbf{y} \sim N(\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\mathbf{b}, \mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\sigma_e^2)$ and $Cov(\mathbf{S}\mathbf{y}, \mathbf{H}^{-1}\mathbf{X}\mathbf{y}) = \mathbf{0}$. It could hence be claimed that $\mathbf{S}\mathbf{y}$ and $\mathbf{H}^{-1}\mathbf{X}\mathbf{y}$ are uncorrelated with each other. However, since \mathbf{X} is usually of form

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n_m} \end{pmatrix}, \quad (\text{I-1.31})$$

i.e. \mathbf{X} is defined as a design or incidence matrix similarly as \mathbf{Z} , where $\mathbf{1}_{n_i}$ is a $n_i \times 1$ column vector with all elements equalling 1, Corbeil & Searle (1976b) pointed out that $\mathbf{S}\mathbf{H}^{-1}\mathbf{S}$ might be singular, i.e. one or more of n_i equal 1 that \mathbf{S} is of one or more null columns and thus singular, and hence they suggested use of \mathbf{T} instead, where \mathbf{T} is produced by deleting n_1 th, $(n_1 + n_2)$ th, ..., and $(n_1 + n_2 + \dots + n_m)$ th rows from \mathbf{S} , and hence no null column would be present in \mathbf{T} . It is easy to show that given \mathbf{X} of form (I-1.31), \mathbf{T} retains the same property as \mathbf{S} , i.e. $\mathbf{TX} = \mathbf{0}$, and hence

$$\mathbf{z} = \begin{bmatrix} \mathbf{T} \\ \mathbf{X}^T\mathbf{H}^{-1} \end{bmatrix} \mathbf{y} = \begin{bmatrix} \mathbf{T}\mathbf{y} \\ \mathbf{X}^T\mathbf{H}^{-1}\mathbf{y} \end{bmatrix} \quad (\text{I-1.32})$$

and

$$\mathbf{z} \sim N \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\mathbf{b} \end{pmatrix}, \begin{pmatrix} \mathbf{T}\mathbf{H}\mathbf{T}^T\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\sigma_e^2 \end{pmatrix} \right]. \quad (\text{I-1.33})$$

Since the two partitions of \mathbf{z} are uncorrelated with each other as shown in (I-1.33), we could write their corresponding logarithm likelihood functions separately as:

$$l_1 = -\frac{1}{2}(N-m)\log 2\pi - \frac{1}{2}(N-m)\log \sigma_e^2 - \frac{1}{2}\log |\mathbf{T}\mathbf{H}\mathbf{T}^T| - \frac{1}{2}\mathbf{y}^T\mathbf{T}^T(\mathbf{T}\mathbf{H}\mathbf{T}^T)^{-1}\mathbf{T}\mathbf{y} / \sigma_e^2 \quad (\text{I-1.34})$$

and

$$l_2 = -\frac{1}{2}m \log 2\pi - \frac{1}{2}m \log \sigma_e^2 - \frac{1}{2} \log |\mathbf{X}^T \mathbf{H}^{-1} \mathbf{X}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b})^T \mathbf{H}^{-1} \mathbf{X}(\mathbf{X}^T \mathbf{H}^{-1} \mathbf{X})^{-1} \mathbf{H}^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{b}) / \sigma_e^2 \quad (\text{I-1.35})$$

Likelihood function (I-1.34) is clearly free of \mathbf{b} and the REML estimators of variance components σ_e^2 and γ could be calculated from maximizing function (I-1.34). Equalizing the first partial derivatives of (I-1.34) to 0 gives:

$$\frac{\partial l_1}{\partial \sigma_e^2} = -\frac{1}{2}(N-k) / \sigma_e^2 + \frac{1}{2} \mathbf{y}^T \mathbf{T}^T (\mathbf{T} \mathbf{H} \mathbf{T}^T)^{-1} \mathbf{T} \mathbf{y} / \sigma_e^4 = 0$$

and

$$\begin{aligned} \frac{\partial l_1}{\partial \gamma_i} &= -tr[(\mathbf{T} \mathbf{H} \mathbf{T}^T)^{-1} \cdot \mathbf{T} \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T \mathbf{T}^T] \\ &\quad + \mathbf{y}^T \mathbf{T}^T (\mathbf{T} \mathbf{H} \mathbf{T}^T)^{-1} \mathbf{T} \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T \mathbf{T}^T (\mathbf{T} \mathbf{H} \mathbf{T}^T)^{-1} \mathbf{T} \mathbf{y} / \sigma_e^2, \\ &= 0 \end{aligned}$$

$$i = 1, 2, \dots, k,$$

where the later equation is derived by noting the facts that $\log |\mathbf{M}| = tr(\log(\mathbf{M}))$, $d(tr(\mathbf{M})) = tr(d(\mathbf{M}))$ and $d\mathbf{M}^{-1} = -\mathbf{M}^{-1}d\mathbf{M}\mathbf{M}^{-1}$, where \mathbf{M} is any non-singular square matrix and d is the differential operator.

Again, there is no general analytic solution for these equations although in balanced data the analytic REML estimators of variance components are available (Corbeil and Searle, 1976a; Harville, 1977), which are identical to the analysis of variance (ANOVA) estimators of balanced data (Searle, 1971) and hence are unbiased but have the minimum variances among all possible unbiased estimators. For unbalanced data, either a general iterative procedure or a W-transformation procedure could be applied to give the REML estimators of variance components (Corbeil and Searle, 1976a). Note here, the REML estimator of \mathbf{b} , derived from function (I-1.35), is identical to that as presented by BLUE, i.e. function (I-1.15) and MLE, i.e. function (I-1.29).

Alongside ML and REML, other methods are also available to work out the estimators of variance components, such as the generalised estimating equations (GEE) (Liang and Zeger, 1986), the minimum norm quadratic unbiased estimation (MINQUE) (Rao, 1971a, b) and ANOVA based Henderson's methods for unbalanced data (Searle, 1971).

It could be noted here, if no interaction term is present and the sample size is large enough, MLE and REML estimators would be very close to each other and hence in many occasions it is free to choose either of the methods to derive the estimates of variance components. A comprehensive comparison between MINQUE, Henderson's method, ML and REML was made by Harville, D. A. (1977), and a further work was launched by Wu et al. (2001), which shows that ML, REML, GEE and MINIQUE could similarly be partitioned into mean and covariance functions. These could intuitively reveal the relationship among those methods. From the notation given by both Harville and Wu, since MINQUE does not require the normality assumption and as each iterative process of REML could be recognized as an iterative MINQUE process, both MLE and REML estimators derived under normality assumption might still be valid even if the distributions of the random effects and the residual term are undefined. Note here, as the estimates of variance components from MINIQUE heavily depend on initial values, MINIQUE itself is hence not recommended in practical data analysis for variance estimation.

The statistical preliminaries introduced in this section will be repeatedly referred to in the following chapters, and hence I include them in the general introduction.

Reference

- ABECASIS, G. R., CARDON, L. R. & COOKSON, W. O. C. (2000) A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66, 279-292.
- ABECASIS, G. R., COOKSON, W. O. C. & CARDON, L. R. (2001) The power to detect linkage disequilibrium with quantitative traits in selected samples. *American Journal of Human Genetics*, 68, 1463-1474.
- ALTSHULER, D. M., GIBBS, R. A., PELTONEN, L., DERMITZAKIS, E., SCHAFFNER, S. F., et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 52-58.
- BALDING, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7, 781-791.
- BANSAL, V., LIBIGER, O., TORKAMANI, A. & SCHORK, N. J. (2010) Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11, 773-785.
- BOEHNKE, M. (1994) Limits of resolution of genetic-linkage studies - implications for the positional cloning of human-disease genes. *American Journal of Human Genetics*, 55, 379-390.
- BOTSTEIN, D., WHITE, R. L., SKOLNICK, M. & DAVIS, R. W. (1980) Construction of a genetic-linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32, 314-331.
- BREM, R. B., YVERT, G., CLINTON, R. & KRUGLYAK, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296, 752-755.
- BRESLOW, N. E. & CLAYTON, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- CARDON, L. R. & PALMER, L. J. (2003) Population stratification and spurious allelic association. *Lancet*, 361, 598-604.
- CASELLA, G. & BERGER, R. L. (2002) *Statistical inference*, Pacific Grove, Calif., Duxbury/Thomson Learning.
- CHAKRABORTY, R. & SMOUSE, P. E. (1988) Recombination of haplotypes leads to biased estimates of admixture proportions in human-populations. *Proceedings of the National Academy of Sciences of the United States of America*, 85, 3071-3074.
- CORBEIL, R. R. & SEARLE, S. R. (1976a) Comparison of variance component estimators. *Biometrics*, 32, 779-791.
- CORBEIL, R. R. & SEARLE, S. R. (1976b) Restricted maximum likelihood (REML) estimation of variance components in mixed model. *Technometrics*, 18, 31-38.

- CUI, Y., LI, G., LI, S. & WU, R. (2010) Designs for linkage analysis and association studies of complex diseases. IN BANG, H., ZHOU, X. K., MAZUMDAR, M. & VANEPPS, H. L. (Eds.) *Statistical methods in molecular biology*. Humana Press Inc.
- DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977) Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39, 1-38.
- DEVLIN, B. & ROEDER, K. (1999) Genomic control for association studies. *Biometrics*, 55, 997-1004.
- DIZIER, M. H., QUESNEVILLE, H., PRUM, B., SELINGER-LENEMAN, H. & CLERGET-DARPOUX, F. (2000) The triangle test statistic (TTS): A test of genetic homogeneity using departure from the triangle constraints in IBD distribution among affected sib-pairs. *Annals of Human Genetics*, 64, 433-442.
- DONNELLY, P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, 456, 728-731.
- DRUKA, A., POTOKINA, E., LUO, Z. W., JIANG, N., CHEN, X. W., et al. (2010) Expression quantitative trait loci analysis in plants. *Plant Biotechnology Journal*, 8, 10-27.
- EWENS, W. J. & SPIELMAN, R. S. (1995) The transmission disequilibrium test - history, subdivision and admixture. *American Journal of Human Genetics*, 57, 455-464.
- FALCONER, D. S. (1989) *Introduction to quantitative genetics*, Harlow, Longman Scientific & Technical.
- FAROOQ, S. & AZAM, F. (2002) Molecular markers in plant breeding-i: Concepts and characterization. *Pakistan Journal of Biological Sciences*, 5, 1135-1140.
- FEINGOLD, E. (2001) Methods for linkage analysis of quantitative trait loci in humans. *Theoretical Population Biology*, 60, 167-180.
- FERRARA, C. T., WANG, P., NETO, E. C., STEVENS, R. D., BAIN, J. R., et al. (2008) Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genetics*, 4, 13.
- FISHER, R. A. (1925a) *Statistical methods for research workers*, Edinburgh, Oliver & Boyd.
- FISHER, R. A. (1925b) Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.
- FOSS, E. J., RADULOVIC, D., SHAFFER, S. A., RUDERFER, D. M., BEDALOV, A., et al. (2007) Genetic basis of proteome variation in yeast. *Nature Genetics*, 39, 1369-1375.
- FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449, 851-U3.
- GOLDSTEIN, D. B. (2009) Common genetic variation and human traits. *New England Journal of Medicine*, 360, 1696-1698.

- HAO, K., LI, C., ROSENOW, C. & WONG, W. H. (2004) Estimation of genotype error rate using samples with pedigree information - an application on the genechip mapping 10k array. *Genomics*, 84, 623-630.
- HARTLEY, H. O. & RAO, J. N. K. (1967) Maximum-likelihood estimation for mixed analysis of variance model. *Biometrika*, 54, 93-&.
- HARVILLE, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.
- HEMMERLE, W. J. & HARTLEY, H. O. (1973) Computing maximum likelihood estimates for mixed aov model using w transformation. *Technometrics*, 15, 819-831.
- HOLLOWAY, B. & LI, B. (2010) Expression QTLs: Applications for crop improvement. *Molecular Breeding*, 26, 381-391.
- JANSEN, R. C. & NAP, J. P. (2001) Genetical genomics: The added value from segregation. *Trends in Genetics*, 17, 388-391.
- JEFFREYS, A. J., MACLEOD, A., TAMAKI, K., NEIL, D. L. & MONCKTON, D. G. (1991) Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, 354, 204-209.
- KALBFLEISCH, J. D. & SPROTT, D. A. (1970) Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32, 175-208.
- KEARSEY, M. J. & POONI, H. S. (1996) *The genetical analysis of quantitative traits*, London, Chapman & Hall.
- KENDZIORSKI, C. M., CHEN, M., YUAN, M., LAN, H. & ATTIE, A. D. (2006) Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, 62, 19-27.
- KU, C. S., LOY, E. Y., PAWITAN, Y. & CHIA, K. S. (2010) The pursuit of genome-wide association studies: Where are we now? *Journal of Human Genetics*, 55, 195-206.
- LANDER, E. S. & BOTSTEIN, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185-199.
- LANGE, K. (1978) Central limit-theorems for pedigrees. *Journal of Mathematical Biology*, 6, 59-66.
- LESAGE, S. & BRICE, A. (2009) Parkinson's disease: From monogenic forms to genetic susceptibility factors. *Human Molecular Genetics*, 18, R48-R59.
- LEWONTIN, R. C. (1964) Interaction of selection + linkage .I. General considerations - heterotic models. *Genetics*, 49, 49-&.
- LIANG, K.-Y. & ZEGER, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- LIMDI, N. A. & VEENSTRA, D. L. (2010) Expectations, validity, and reality in pharmacogenetics. *Journal of Clinical Epidemiology*, 63, 960-969.

- MACKAY, T. F. C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics*, 35, 303-339.
- MCCARTHY, M. I. (2010) Genomic medicine genomics, type 2 diabetes, and obesity. *New England Journal of Medicine*, 363, 2339-2350.
- MCMILLAN, I. & ROBERTSON, A. (1974) Power of methods for detection of major genes affecting quantitative characters. *Heredity*, 32, 349-356.
- MELZER, D., PERRY, J. R. B., HERNANDEZ, D., CORSI, A. M., STEVENS, K., et al. (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics*, 4, 10.
- MENG, X. L. & RUBIN, D. B. (1993) Maximum-likelihood-estimation via the ECM algorithm - a general framework. *Biometrika*, 80, 267-278.
- METZKER, M. L. (2010) Applications of next-generation sequencing. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11, 31-46.
- MILLER, J. J. (1977) Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics*, 5, 746-762.
- MURPHY, S. A. & VAN DER VAART, A. W. (2000) On profile likelihood. *Journal of the American Statistical Association*, 95, 449-465.
- PATTERSON, H. & THOMPSON, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-563.
- PETES, T. D. & BOTSTEIN, D. (1977) Simple Mendelian inheritance of reiterated ribosomal DNA of yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5091-5095.
- PONG-WONG, R., GEORGE, A., WOOLLIAMS, J. & HALEY, C. (2001) A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genetics Selection Evolution*, 33, 453 - 471.
- POTOKINA, E., DRUKA, A., LUO, Z. W., WISE, R., WAUGH, R., et al. (2008) Gene expression quantitative trait locus analysis of 16,000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant Journal*, 53, 90-101.
- PRENTICE, R. L. & PYKE, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.
- PRICE, A. L., PATTERSON, N. J., PLENG, R. M., WEINBLATT, M. E., SHADICK, N. A., et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38, 904-909.
- PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.
- RAKOVSKI, C. S. & STRAM, D. O. (2009) A kinship-based modification of the Armitage trend test to address hidden population structure and small differential genotyping errors. *PLoS One*, 4, 10.

- RAO, C. R. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44, 50-57.
- RAO, C. R. (1971a) Estimation of variance and covariance components--MINQUE theory. *Journal of Multivariate Analysis*, 1, 257-275.
- RAO, C. R. (1971b) MINQUE estimation of variance and covariance components. *Annals of Mathematical Statistics*, 42, 1477-&.
- REDDEN, D. & ALLISON, D. (2006) The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture. *Behavior Genetics*, 36, 678-686.
- RICE, J. A. (1994) *Mathematical statistics and data analysis : John a. Rice*, Belmont, Cal., Duxbury Press.
- RISCH, N. (1990a) Linkage strategies for genetically complex traits .1. Multilocus models. *American Journal of Human Genetics*, 46, 222-228.
- RISCH, N. (1990b) Linkage strategies for genetically complex traits .2. The power of affected relative pairs. *American Journal of Human Genetics*, 46, 229-241.
- RISCH, N. (1990c) Linkage strategies for genetically complex traits .3. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics*, 46, 242-253.
- RISCH, N. (1992) Mapping genes for complex diseases using association studies with recently admixed populations. *American Journal of Human Genetics*, 51, A13.
- ROEDER, K., CARROLL, R. J. & LINDSAY, B. G. (1996) A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91, 722-732.
- RONALD, J., BREM, R. B., WHITTLE, J. & KRUGLYAK, L. (2005) Local regulatory variation in *saccharomyces cerevisiae*. *PLoS Genet*, 1, e25.
- SAUNDERS, I. W., HANNAN, G. N., BROHEDE, J., GILES, G. G., JENKINS, M. A., et al. (2007) A range of simple summary genome-wide statistics for detecting genetic linkage using high density marker data. *Genetic Epidemiology*, 31, 565-576.
- SCHERER, S. W., LEE, C., BIRNEY, E., ALTSHULER, D. M., EICHLER, E. E., et al. (2007) Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, 39, S7-S15.
- SCHORK, N. J. (1993) Extended multipoint identity-by-descent analysis of human quantitative traits - efficiency, power, and modeling considerations. *American Journal of Human Genetics*, 53, 1306-1319.
- SCHORK, N. J. (1997) Genetics of complex disease - approaches, problems, and solutions. *American Journal of Respiratory and Critical Care Medicine*, 156, S103-S109.
- SEARLE, S. R. (1971) *Linear models*, New York ; Chichester, Wiley.

- SHENDURE, J. & JI, H. L. (2008) Next-generation DNA sequencing. *Nature Biotechnology*, 26, 1135-1145.
- SILVER, J. (1985) Confidence-limits for estimates of gene linkage based on analysis of recombinant inbred strains. *Journal of Heredity*, 76, 436-440.
- SMYTH, G. K. & VERBYLA, A. P. (1996) A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society Series B-Methodological*, 58, 565-572.
- SPIELMAN, R. S. & EWENS, W. J. (1993) Transmission disequilibrium test (TDT) for linkage and linkage disequilibrium between disease and marker. *American Journal of Human Genetics*, 53, 863-863.
- SPROTT, D. A. (1975) Marginal and conditional sufficiency. *Biometrika*, 62, 599-605.
- STADLER, L. J. (1929) Chromosome number and the mutation rate in avena and triticum. *Proceedings of the National Academy of Sciences of the United States of America*, 15, 876-881.
- STURTEVANT, A. H. (1913) The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14, 43-59.
- SZATROWSKI, T. H. & MILLER, J. J. (1980) Explicit maximum-likelihood estimates from balanced data in the mixed model of the analysis of variance. *Annals of Statistics*, 8, 811-819.
- VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., VANDELEE, T., et al. (1995) AFLP - a new technique for DNA-fingerprinting. *Nucleic Acids Research*, 23, 4407-4414.
- WALKER, F. O. (2007) Huntington's disease. *The Lancet*, 369, 218-228.
- WANG, T. & ELSTON, R. C. (2005) The bias introduced by population stratification in IBD based linkage analysis. *Human Heredity*, 60, 134-142.
- WEINBERG, C. R. & WACHOLDER, S. (1993) Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika*, 80, 461-465.
- WENTZELL, A. M., ROWE, H. C., HANSEN, B. G., TICCONI, C., HALKIER, B. A., et al. (2007) Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genetics*, 3, 1687-1701.
- WILLIAMS-BLANGERO, S. & BLANGERO, J. (2006) Collection of pedigree data for genetic analysis in isolate populations. *Human Biology*, 78, 89-101.
- WILLIAMS, J. G. K., KUBELIK, A. R., LIVAK, K. J., RAFALSKI, J. A. & TINGEY, S. V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic-markers. *Nucleic Acids Research*, 18, 6531-6535.
- WINZELER, E. A., RICHARDS, D. R., CONWAY, A. R., GOLDSTEIN, A. L., KALMAN, S., et al. (1998) Direct allelic variation scanning of the yeast genome. *Science*, 281, 1194-1197.

- WORLAND, A. J., GALE, M. D. & LAW, C. N. (1987) Wheat genetics. IN LUPTON, F. J. H. (Ed.) *Wheat breeding: Its scientific principals*. London, Chapman and Hill.
- WRAY, G. A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, 8, 206-216.
- WU, C. T., GUMPERTZ, M. L. & BOOS, D. D. (2001) Comparison of GEE, MINQUE, ML, and REML estimating equations for normally distributed data. *American Statistician*, 55, 125-130.
- YANG, J. A., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42, 565-U131.
- YEUNG, J. M. Y., SHAM, P. C., CHAN, A. S. W. & CHERNY, S. S. (2008) Openadam: An open source genome-wide association data management system for Affymetrix SNP arrays. *BMC Genomics*, 9, 4.
- YU, J. M., PRESSOIR, G., BRIGGS, W. H., BI, I. V., YAMASAKI, M., et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38, 203-208.
- ZHENG, C. X. & ELSTON, R. C. (1999) Multipoint linkage disequilibrium mapping with particular reference to the African-American population. *Genetic Epidemiology*, 17, 79-101.

CHAPTER II
THE COMPARISON OF LINEAR
MODELS IN ASSOCIATION
STUDY AND THE OPTIMAL
TREND COEFFICIENTS IN
ARMITAGE'S TREND TEST

CHAPTER II-1

INTRODUCTION OF GENETIC MODELS AND STATISTIC METHODS OF ASSOCIATION ANALYSIS IN RANDOMLY MATED POPULATION

1.1 Overview

As has been introduced in General Introduction, the LD-based analysis has two important advantages over the linkage-based analysis in QTL mapping, i.e. the higher mapping resolution and more abundant resources, and hence nowadays, most QTL mappings are based on the inference of LD. In Chapter II-1, I will start with introduction of genetic models that link genetic effects of genes to a quantitative trait, followed by a brief introduction of the commonly used statistical methods in association analysis. Although the association analysis can be implemented for analysing both family and population based samples (Abecasis et al., 2000), as the population based association analysis generally retains higher statistical power as well as the broader genetic background, the population based association analysis is of most interest nowadays (Balding, 2006) and this introduction will focus on population based association analyses only.

1.2 Models of Quantitative Genetic Effects

1.2.1 Explicit Model

An explicit model assumes the observed phenotypes can directly reflect the effects of genotypes (the genotypic values) and environmental factors. Statistically, they are of linear relationship, say $y_{ij} = V_i + \varepsilon_{ij}$ or $E(y_{ij}) = V_i$, where y_{ij} denotes the phenotype of j th

individuals with i th genotype, V_i is the corresponding genotypic value of the causal genes and ε_{ij} denotes the deviation of phenotype y_{ij} from V_i with $E(\varepsilon_{ij})=0$. The explicit model is introduced ideally to deal with continuous phenotypes, but it can also be applied to binary phenotypes, where the genotypic value represents the probability, or equivalent factors, of taking one of the binary phenotypes given a particular genotype in the present sample.

1.2.2 Implicit Model

Implicit model considers that the genotypic values of genotypes can only determine the phenotype through certain threshold(s), and hence in the case of binary phenotype, the observed phenotype can only reflect the genotypic values through the model

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} \geq \phi \\ 0 & \text{if } z_{ij} < \phi \end{cases} \quad \text{and } \phi \in \mathbf{R}, \quad (\text{II-1.1})$$

where $z_{ij} = V_i + \varepsilon_{ij}$, V_i and ε_{ij} have the same definition as that in explicit model, and \mathbf{R} is the field of real numbers. More precisely, the implicit model belongs to the generalised linear model, which will be closely discussed in section II-1.4.4.

1.3 Association Analysis in Random Samples

Taking the same setting of the bi-allelic marker locus and QTL as given in Table I-1, the joint distributions of marker and QTL genotypes can be computed as shown in Table II-1 under the assumption of random mating, where $p = p_M$, $q = p_A$ and $D = f_{AM} - p_A p_M = f_{AM} - pq$. The genotypic values of QTL genotypes are denoted as V_{AA} , V_{Aa} and V_{aa} , or equivalently V_1 , V_2 and V_3 . If we define $\mu = (V_{AA} + V_{aa})/2$, $a = V_{AA} - \mu$ and $d = V_{Aa} - \mu$, the genotypic values can be alternatively expressed as $V_{AA} = \mu + a$, $V_{Aa} = \mu + d$ and $V_{aa} = \mu - a$, where a and d are

commonly denoted as the coefficients of additive and dominance effects. Without loss of generality, one can assume $a > 0$, and in the absence of over-dominance, $|d| \leq a$.

Table II-1. Joint Distribution of Marker and QTL Genotypes in a Random Mating Population.

D is the coefficient of LD between marker locus and QTL; p and q are marker allele (M) and QTL (A) frequencies respectively; a and d are the additive and dominance effects of QTL.

Marker genotype	QTL genotypes		
	AA	Aa	aa
MM	$(D + pq)^2$	$2(D + pq)[p(1 - q) - D]$	$[D - p(1 - q)]^2$
Mm	$2(D + pq)[(1 - p)q - D]$	$2[2D^2 + (1 - 2p)(1 - 2q)D + 2pq(1 - p)(1 - q)]$	$2[D + (1 - p)(1 - q)][p(1 - q) - D]$
mm	$[D - (1 - p)q]^2$	$2[(1 - p)q - D][D + (1 - p)(1 - q)]$	$[D + (1 - p)(1 - q)]^2$
Genotypic value	$\mu + a$	$\mu + d$	$\mu - a$

1.3.1 Simple Linear Regression (SLR)

As suggested by Lande & Thompson (1990), assigning the phenotype to y and the number of marker allele M of i th individual to x_i , function (I-1.21) now takes the form

$$y_{ij} = \alpha + \beta \cdot x_i + e_{ij},$$

or

$$\mathbf{y} = \alpha \cdot \mathbf{1} + \beta \cdot \mathbf{x} + \mathbf{e}$$
(II-1.2)

where \mathbf{y} is a column vector with its element y_{ij} being the phenotype of j th individual with i th genotype; $\mathbf{1}$ is a column with all its element equals 1; \mathbf{x} is the column vector with its element x_i as defined above, and \mathbf{e} is the column vector with its element e_{ij} being the residual term with $E(e_{ij}) = 0$, $Var(e_{ij}) = \sigma^2$ and $Cov(e_{ij}, e_{lk}) = 0$ unless $i = l$ and $j = k$.

Suppose the expectation of phenotype y_{ij} is ideally equal to the genotypic value of its correspondent QTL genotype, the estimate of regression coefficient, i.e. β , can be given as

$$\hat{\beta} = \frac{Cov(xy)}{Var(x)} \text{ from formula (I-1.22). By noticing}$$

$$E(x) = 2f_M^2 + 2f_M f_m = 2f_M,$$

$$E(y) = f_A^2 V_{AA} + 2f_A f_a V_{Aa} + f_a^2 V_{aa} \\ = \mu + (f_A^2 - f_a^2)a + 2f_A f_a d,$$

$$E(xy) = 2(f_{AAMM}V_{AA} + f_{AaMM}V_{Aa} + f_{aaMM}V_{aa}) + (f_{AAMm}V_{AA} + f_{AaMm}V_{Aa} + f_{aaMm}V_{aa}) \\ = 2[f_{AM}f_A V_{AA} + (f_{AM}f_a + f_{aM}f_A)V_{Aa} + f_{aM}f_a V_{aa}] \\ = 2[f_M \mu + (f_{AM}f_A - f_{aM}f_a)a + (f_{AM}f_a + f_{aM}f_A)d]$$

$$E(x^2) = 4f_M^2 + 2f_M f_m,$$

and

$$D = f_{MA} - f_M f_A = -f_{Ma} + f_M f_a = -f_{ma} + f_m f_A = f_{ma} - f_m f_a,$$

it can be hence derived

$$\begin{aligned}
E(\hat{\beta}) &= \frac{E(xy) - E(x)E(y)}{E(x^2) - E^2(x)} \\
&= \frac{\left\{ \begin{aligned} &2[f_M\mu + (f_{AM}f_A - f_{aM}f_a)a + (f_{AM}f_a + f_{aM}f_A)d] \\ &-2f_M \times [\mu + (f_A^2 - f_a^2)a + 2f_Af_ad] \end{aligned} \right\}}{4f_M^2 + 2f_Mf_m - 4f_M^2} \\
&= \frac{D[a + (1-2q)d]}{p(1-p)}
\end{aligned} \tag{II-1.3}$$

It is clear from equation (II-1.3) that the test of null hypothesis: $\beta = 0$ is equivalent to test against $D = 0$, and hence the statistical method such proposed is a LD-based method. Note here, the relationship between β and D as presented above is generally valid for any adequate choice of x_i and will be shown in Chapter II-2.

Under the null hypothesis, where $\beta = 0$, it is clear that $\bar{y}_i - \alpha \sim N(0, \sigma^2 / n_i)$ asymptotically through the central limit theorem irrespective the exact distribution of e_{ij} , where n_i denotes the number of individuals with i th genotype. Because $\hat{\beta}$ is a linear function of $\bar{y}_i - \alpha$, one can hence conclude that $\hat{\beta}$ follows a normal distribution asymptotically under the null hypothesis irrespective of the exact distribution of e_{ij} and the exact values of x_i . Given a sufficiently large sample size, the test statistic can be given as $\frac{\hat{\beta}^2}{Var(\hat{\beta})} \approx n\hat{\rho}_{xy}^2 \sim \chi_{df=1}^2$, where $\hat{\rho}_{xy}$ is the estimator of the correlation coefficient between x and y .

Although it has been mentioned above that any choice of x_i would not violate the normality of $\hat{\beta}$ under null hypothesis, it does affect the statistical power and thus the type II error rate. It would be shown that the suggestion given by Lande & Thompson (1990) is the proper choice only under the additive model, e.g. $d = 0$ in explicit model, and a comprehensive discussion about the choice of x_i will be launched in Chapter II-2.

1.3.2 Fixed Effect Model (ANOVA)

To avoid the uncertain choice of x_i as mentioned in section II-1.3.1, an alternative strategy can be implemented to perform the association analysis through treating the genotypic value of three genotypes independently. The linear regression hence established takes the form of equation (I-1.17) (Knapp and Bridges, 1990)

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

or

(II-1.4)

$$y_{ij} = \beta_i + e_{ij},$$

where $i=1,2,3$ denotes the i th marker genotype with its ‘genetic effect’ β_i on the trait; y_{ij} and e_{ij} denote the phenotype and the residual term of the j th individual with i th marker genotype; \mathbf{X} is the incidence matrix with elements equalling to 1 or 0; \mathbf{y} , \mathbf{b} , \mathbf{e} are corresponding vectors of y_{ij} , β_i and e_{ij} that $Var(\mathbf{e}) = \sigma^2 \mathbf{I}$. Note the constant term, i.e. α in equation (II-1.2), is eliminated here. Such a transformation is to avoid the singularity of \mathbf{X} in the presence of α , and this process will not affect the statistical test hence derived.

From formula (I-1.18), (I-1.19), (I-1.20), we can have:

$$\hat{\beta}_i = \frac{\sum_j y_{ij}}{n_i} = \bar{y}_i, \quad i=1,2,3, \quad (II-1.5)$$

$$Var \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 1/n_1 & & \\ & 1/n_2 & \\ & & 1/n_3 \end{pmatrix} \sigma^2, \quad (II-1.6)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_i)^2}{n-3}, \quad (\text{II-1.7})$$

where \bar{y}_i denotes the sample mean of i th marker genotype with the sample size n_i , and n is the full sample size, where clearly $n = \sum_i n_i$.

Ideally, the expectation of phenotype y_{ij} is supposed to equal the genotypic value of its QTL genotype, and hence one can calculate

$$\begin{aligned} E(y) &= f_A^2 V_{AA} + 2f_A f_a V_{Aa} + f_a^2 V_{aa} \\ &= \mu + (f_A - f_a)a + 2f_A f_a d \end{aligned}$$

$$\begin{aligned} E_1(y) &= \frac{f_{AAMM} V_{AA} + f_{AaMM} V_{Aa} + f_{aaMM} V_{aa}}{f_{MM}} \\ &= \frac{f_M^2 \mu + (f_{AM}^2 - f_{aM}^2)a + 2f_{AM} f_{aM} d}{f_M^2} \\ &= \mu + (f_A - f_a)a + 2f_A f_a d + \frac{2D[f_M a - (D + f_M f_A - f_M f_a)d]}{f_M^2} \\ &= E(y) + \frac{2D[f_M a - (D + f_M f_A - f_M f_a)d]}{f_M^2} \end{aligned}$$

and similarly

$$\begin{aligned} E_2(y) &= \frac{f_{AAMm} V_{AA} + f_{AaMm} V_{Aa} + f_{aaMm} V_{aa}}{f_{Mm}} \\ &= E(y) + \frac{D[-(f_M - f_m)a + 2Dd + (f_M - f_m)(f_A - f_a)d]}{f_M f_m} \end{aligned}$$

$$\begin{aligned} E_3(y) &= \frac{f_{Aamm} V_{AA} + f_{Aamm} V_{Aa} + f_{aamm} V_{aa}}{f_{mm}} \\ &= E(y) + \frac{2D[f_m a + (D + f_m f_a - f_m f_A)d]}{f_m^2} \end{aligned}$$

Clearly, a statistical test for significance of null hypothesis $E_i(y) = E(y) \forall i$ is equivalent to a test of the null hypothesis $D = 0$.

Following the normality assumption, where $\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I})$, $\hat{\mathbf{b}} \sim N(\mathbf{b}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$ and by

supposing $\hat{\mathbf{b}}$ are distributed independently of $\hat{\sigma}^2$, it is easy to show that $\frac{y_{ij} - \bar{y}_{i.}}{\sigma} \sim Z \forall j$ and

$\sqrt{n_i} \frac{\hat{\beta}_i - \bar{y}}{\sigma} \sim Z \forall i$ under the null hypothesis, where $Z \sim N(0,1)$, $\beta_i = E_i(y) \forall i$ and

$\hat{\beta}_i = \bar{y}_{i.} = \frac{\sum_j y_{ij}}{n_i}$, $\bar{y} = \frac{\sum_{i,j} y_{ij}}{n}$ are the unbiased estimators of $E_i(y)$ and $E(y)$ respectively. One

can hence derive

$$\frac{\frac{\sum_i \left(\sqrt{n_i} \frac{\hat{\beta}_i - \bar{y}}{\sigma} \right)^2}{3-1}}{\frac{\sum_{ij} \left(\frac{y_{ij} - \bar{y}_{i.}}{\sigma} \right)^2}{n-3}} = \frac{\sum_{i=1}^3 n_i (\bar{y}_{i.} - \bar{y})^2}{\sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2} \times \frac{n-3}{2} \sim F(2, n-3)$$

or

$$\begin{aligned} & \frac{\sum_i \left(\sqrt{n_i} \frac{\hat{\beta}_i - \bar{y}}{\sigma} \right)^2}{\frac{3-1}{\frac{\sum_i \left(\sqrt{n_i} \frac{\hat{\beta}_i - \bar{y}}{\sigma} \right)^2 + \sum_{ij} \left(\frac{y_{ij} - \bar{y}_{i.}}{\sigma} \right)^2}{n-1}}} \\ &= \frac{\sum_{i=1}^3 n_i (\bar{y}_{i.} - \bar{y})^2}{\sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2} \times \frac{n-1}{2} \sim F(2, n-1) \end{aligned} \quad (\text{II-1.8})$$

where $\sum_i \left(\sqrt{n_i} \frac{\hat{\beta}_i - \bar{y}}{\sigma} \right)^2 \sim \chi_{df=2}^2$, $\sum_{ij} \left(\frac{y_{ij} - \bar{y}_{i.}}{\sigma} \right)^2 \sim \chi_{df=n-3}^2$ and $F(t_1, t_2)$ denotes the *F-statistic*

with degrees of freedom t_1 in the numerator and degrees of freedom t_2 in the denominator.

Generally, an *F-statistic* can be formed by $\frac{Q_1}{t_1} \bigg/ \frac{Q_2}{t_2}$, where $Q_1 \sim \chi_{df=t_1}^2$ and $Q_2 \sim \chi_{df=t_2}^2$. Since

$\frac{Q_2}{t_2} \rightarrow 1$ while $t_2 \rightarrow \infty$, it can hence be noticed that $F(2, n-3) \approx \frac{1}{2} \chi_{df=2}^2$ asymptotically. Since

$\lim_{n \rightarrow \infty} \frac{\sum_{ij} \left(\frac{y_{ij} - \bar{y}_{i.}}{\sigma} \right)^2}{n-3} = 1$ does not rely on the exact distribution of y_{ij} , if $\sigma^2 = E(\hat{\sigma}_i^2) \forall i$ and

$\hat{\sigma}_i^2 = \frac{\sum_j (y_{ij} - \bar{y}_{i.})^2}{n-1}$, asymptotically one may extend the validation of formulae (II-1.8)

$$\frac{\sum_{i=1}^3 n_i (\bar{y}_{i.} - \bar{y})^2}{\frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n-3}} \sim \chi_{df=2}^2$$

or

(II-1.9)

$$\frac{\sum_{i=1}^3 n_i (\bar{y}_{i.} - \bar{y})^2}{\frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{n-1}} \sim \chi_{df=2}^2$$

as long as $\sqrt{n_i} \frac{\hat{\beta}_i - \bar{y}}{\sigma} \sim Z \forall i$ stands under null hypothesis, which is generally true given that

$\hat{\beta}_i \sim N(\beta_i, \sigma^2 / n_i)$ asymptotically through the central limit theorem and $E(\bar{y}) = \beta_i \forall i$ under the null hypothesis.

It can be noted that above derivations are identical to the analysis of variance (ANOVA) procedure under the one-way fixed effect model with normality assumption. More generally, linear regressions with incidence matrices \mathbf{X} are all equivalent to corresponding ANOVA based methods given the normality assumption (Searle, 1971). Also note here such genetic effects in ANOVA models may be alternatively regarded as random effects (Hill, 1975). However, due to the unbalanced nature of the randomly sampled data, the estimates of such

random effects can be severely biased (Luo, 1993, Knott, 1994), and hence only the fixed model is introduced as above.

1.3.3 Likelihood Based Approach

Taking the joint distribution of marker and QTL genotypes given in Table II-1, the complete likelihood function of a randomly collected sample under the assumption of random mating and independence can be given as

$$\begin{aligned}
 l(\Omega | y, M) &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \log \Pr(y_{ij}, M = i | \Omega) \\
 &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \log \left[\sum_{k=1}^3 \Pr(y_{ij}, G = k | M = i, \Omega) \Pr(M = i | \Omega) \right], \quad (\text{II-1.10}) \\
 &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \log \left[\sum_{k=1}^3 \Pr(y_{ij} | G = k, \Omega) \Pr(G = k, M = i | \Omega) \right]
 \end{aligned}$$

where y_{ij} indicates the observed phenotype of the j th individual with the i th marker genotype, and we suppose this parameter only provides information of phenotype but not its genotype; $M = i$, $G = k$ denote the i th marker genotype and the k th QTL genotype respectively; Ω is the set of parameters, of which the parameters will be defined for a particular model; $\Pr(G = k, M = i | \Omega)$ presents the (i, k) th entry of Table II-1 and will be denoted as h_{ik} in the following discussion; $\Pr(y_{ij} | G = k, \Omega)$ is defined by the exact model of quantitative genetic effects, which can be either explicit or implicit.

However, directly maximizing function (II-1.10) would be really difficult in most situations, since it seems impossible to maximize all parameters at the same time. Alternatively, as introduced at functions (I-1.9), if treating the unknown distribution of QTL genotype as missing data (Luo et al., 2000, Luo and Wu, 2001), equivalently to function (II-1.10), the expected log-likelihood function at the t -th ECM iterative process can be given as

$$\begin{aligned}
l_c(\Omega^t | y, M) &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 \left[\Pr(G = k | y_{ij}, M = i, \Omega^t) \log \Pr(y_{ij}, M = i, G = k | \Omega^t) \right] \\
&= \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 \left[w_{ijk}^t \log h_{ik}^t \Pr(y_{ij} | G = k, \Omega^t) \right]
\end{aligned} \tag{II-1.11}$$

Function (II-1.11) is calculated given the following results

$$\begin{aligned}
\Pr(y_{ij}, M = i, G = k | \Omega^t) &= \Pr(y_{ij} | M = i, G = k, \Omega^t) \Pr(M = i, G = k | \Omega^t) \\
&= h_{ik}^t \Pr(y_{ij} | G = k, \Omega^t)
\end{aligned}$$

$$\Pr(y_{ij} | M = i, G = k, \Omega^t) = \Pr(y_{ij} | G = k, \Omega^t)$$

and

$$\begin{aligned}
w_{ijk}^t &= \Pr(G = k | y_{ij}, M = i, \Omega^t) \\
&= \frac{\Pr(G = k, y_{ij}, M = i | \Omega^t)}{\sum_{k=1}^3 \Pr(G = k, y_{ij}, M = i | \Omega^t)}, \\
&= \frac{h_{ik}^t \Pr(y_{ij} | G = k, \Omega^t)}{\sum_{k=1}^3 h_{ik}^t \Pr(y_{ij} | G = k, \Omega^t)}
\end{aligned} \tag{II-1.12}$$

where the superscript t denotes terms and parameters acquired at the t -th iterative process.

Formula (II-1.12) hence gives the E-step of the t -th iterative algorithm as introduced in section I-1.3.2, and the t -th CM-step can be given by partially maximizing function (II-1.11) conditioned on each parameter. The $t+1$ -st θ_l can hence be given by solving equation

$$\frac{\partial l_c(\Omega_{1,l-1}^{t+1}, \Omega_{l+1,m}^t | y, M)}{\partial \theta_l} = \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 w_{ijk}^t \times \frac{\partial \log [\Pr(y_{ij} | G = k, \Omega_{1,l-1}^{t+1}, \Omega_{l+1,m}^t) h_{ik}^{t+1}]}{\partial \theta_l} \tag{II-1.13}$$

where θ_l denotes the l th parameter, $l = 1, 2, \dots, m$; Ω_{l_1, l_2}^t denotes the set of the l_1 th to the l_2 th parameters at the t -th iterative process, which is null if $l_1 < l_2$.

Note here, as discussed in section I-1.4.2, generally, equations (II-1.13) are only the necessary conditions of maximizing each sub-CM-step, and further efforts of deriving second-order

partial derivatives are required to clarify whether the solutions hence acquired maximize each sub-CM-step. However, in practice, one may ignore the step of checking the sufficient condition of each sub-CM-step if each iterative process does increase the likelihood function (II-1.11) as there might be only one possible root for each sub-CM-step.

As also discussed in section I-1.4.2, given a proper initial point, Ω_0 , the convergence of iteration would be claimed once a certain criteria δ is met at t_0 -th iterative process such that $\delta \geq l(\Omega^{t+1} | y, M) - l(\Omega^t | y, M)$ for any $t > t_0$, as well as the approximate MLEs Ω^{t+1} . A corresponding LRT can hence be performed using

$$-2[l(\Omega_c | y, M)_{D=0} - l(\Omega_c | y, M)] \sim \chi_{df}^2,$$

where df is usually 1, i.e. D is fixed to 0, but if the null hypothesis $D=0$ eliminates other parameters simultaneously, i.e. a parameter is either defined to be a constant value or absolutely excluded from the likelihood function, calculation of df should count in all such parameters along with D . For convenience, in the following discussion, the establishment of LRT might be neglected unless necessary.

Recall the likelihood function (II-1.10), of which the first partial derivatives can be directly given as:

$$\begin{aligned} \frac{\partial l(\Omega | y, M)}{\partial \theta_l} &= \frac{\partial \sum_{i=1}^3 \sum_{j=1}^{n_i} \log \left[\sum_{k=1}^3 \Pr(y_{ij} | G=k, \Omega) \Pr(G=k, M=i | \Omega) \right]}{\partial \theta_l} \\ &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 \frac{1}{\sum_{k=1}^3 \Pr(y_{ij} | G=k, \Omega) h_{ik}} \times \frac{\partial \Pr(y_{ij} | G=k, \Omega) h_{ik}}{\partial \theta_l} \\ &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 w_{ijk} \times \frac{\partial \log [\Pr(y_{ij} | G=k, \Omega) h_{ik}]}{\partial \theta_l} \end{aligned} \quad (\text{II-1.14})$$

where $w_{ijk} = \Pr(G=k | y_{ij}, M=i, \Omega)$. The similarity between formulae (II-1.14) and (II-1.13)

shows that ECM algorithm can be viewed as an iterative strategy of maximizing the

likelihood function (II-1.10) and might be generally adopted without the establishment of the expectation log-likelihood function (II-1.11).

Following the above procedure, the MLEs and LRT can not be evaluated unless the exact form of $\Pr(y_{ij} | G = k, \Omega)$ has been defined. For continuous data, as suggested by Luo et al. (1998), assuming the hold of normality assumption, one may write

$$\Pr(y_{ij} | G = k, \Omega) = \frac{1}{\sqrt{2\pi v}} \times e^{[-(y_{ij} - \mu_k)^2 / 2v]}, \quad (\text{II-1.15})$$

where $\mu_k = \mu + (2 - k)a + \frac{1 - (-1)^k}{2}d$ denotes the genotypic value of k th QTL genotype as given in Table I-1, and v is the variance. Clearly, function (II-1.15) is an explicit model and can be equivalently written into $y_{ij} = \mu_k + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, v)$. For binary data, the intuitional choice of $\Pr(y_{ij} | G = k, \Omega)$ would be

$$\Pr(y_{ij} | G = k, \Omega) = f_k^{y_{ij}} (1 - f_k)^{1 - y_{ij}}, \quad (\text{II-1.16})$$

which is equivalent to

$$\Pr(y_{ij} = 1 | G = k, \Omega) = f_k,$$

where given the k th QTL genotype, y_{ij} is of value 1 or 0 to denote the retaining of a particular phenotype or not with the probabilities f_k and $1 - f_k$ respectively. In an ideal population, f_k is often referred to as the penetrance coefficient of k th QTL genotype, and hence the genetic model given as function (II-1.16) can be named as the penetrance model. It can be also noticed that formula (II-1.16) is an explicit model as $E(y_{ij}) = f_k$ in accordance with the definition from section II-1.2.1.

On the other hand, the implicit models for binary data can be given following function (II-1.1)

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} \geq \phi \\ 0 & \text{if } z_{ij} < \phi \end{cases},$$

and

$$z_{ij} = \mu_k + \varepsilon_{ij}.$$

Under the assumption $\varepsilon_{ij} \sim N(0,1)$, one may write

$$\begin{aligned} f_k &= \Pr(y_{ij} = 1 \mid G = k, \phi) = \Pr(z_{ij} \geq \theta \mid G = k, \phi) \\ &= 1 - \Pr(z_{ij} \leq \theta \mid G = k, \phi) \\ &= 1 - \frac{1}{\sqrt{2\pi v}} \int_{-\infty}^{\phi} \exp \left[-\frac{\left(z - \mu - (2-k)a - \frac{1+(-1)^k}{2}d \right)^2}{2v} \right] dz, \quad (\text{II-1.17}) \\ &= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\phi} \exp \left[-\frac{(z - \mu_k^*)^2}{2} \right] d(z - \mu_k^*) \end{aligned}$$

where $\mu_k^* = \mu_k - \mu = (2-k)a + \frac{1-(-1)^k}{2}d$. Function (II-1.17) is often referred to as the probit

or liability model as firstly introduced by Wright (1934).

1.3.4 Under the Scheme of Likelihood

As has been given by the log-likelihood function (II-1.10), the complete likelihood function takes the form

$$L(\Omega \mid y, M) = \prod_{i,j} \Pr(y_{ij}, M = i \mid \Omega).$$

Under the HWE, the distribution of marker genotypes, i.e. $\Pr(M = i)$, $i = 1, 2, 3$, can only be determined through one parameter p , the frequency of marker allele M , in the form $\Pr(M = 1) = p^2$, $\Pr(M = 2) = 2p(1-p)$ and $\Pr(M = 3) = (1-p)^2$. It is easy to show that n_i , $i = 1, 2, 3$, are the sufficient statistics of p , where n_i denotes the observed number of

individuals with marker genotype $M = i$. Since such sufficient statistics can be directly calculated irrespective of the exact value of the other parameters, the MLE of p given from function (II-1.10) should be identical with the estimate calculated from the sufficient statistics n_i , $i = 1, 2, 3$, and such an estimate can be given as $\tilde{p} = \frac{2n_1 + n_2}{2\sum_{i=1}^3 n_i}$. Incorporating \tilde{p} into the

above likelihood function results in

$$l(\Omega | y, M) = \sum_{ij} \left[\log \Pr(y_{ij} | M = i, \Omega) + \log \Pr(M = i) \right].$$

As the last term, i.e. $\Pr(M = i)$, takes a definite value subjected on i , it will have no impact on further derivations, i.e. MLE and LRT, and hence it can be concluded that the complete likelihood function is equivalent to the conditional likelihood function

$$L(\Omega, M | y) = \prod_{i,j} \Pr(y_{ij} | M = i, \Omega).$$

Although the above discussion explains the equivalence of complete and conditional likelihood function in an ideal population, the difference between the linear models and the likelihood based model arise about the modelling of phenotypic values. The linear models directly model $\Pr(y_{ij} | M = i, \Omega)$, the relationship between phenotype and marker genotype, whereas the likelihood based model models $\Pr(y_{ij} | G = k, \Omega)$, the relationship between phenotype and QTL genotype. Theoretically, it seems more informative to model $\Pr(y_{ij} | G = k, \Omega)$ rather than $\Pr(y_{ij} | M = i, \Omega)$, because the putative QTL is the candidate gene rather than an associated factor. However, several researches have shown that likelihood based model doesn't guarantee a higher statistical power over linear models. For example, with continuous phenotype, simulation analysis of Luo et al. (2000) showed that SLR always presented the highest statistical power even while the optimality of x_i is severely violated, although as indicated by Luo & Wu (2001) likelihood based method has shown an overall

higher statistical power than the SLR while binary phenotype was presented. On the other hand, ANOVA is generally of the least statistical power among all methods (Kendall and Stuart, 1961, Luo et al., 2000).

It should be noted here that the likelihood scheme doesn't guarantee that MLEs of linear models are identical to their BLUEs, although it is always true if the normality assumption of the residual term is held. For binary data, the MLE of ANOVA model, i.e. formula (II-1.4), is identical with its BLUE, but it is not true between the MLE and BLUE of SLR, where, however, it will be shown in section II-2.4.1 that their statistical tests are still asymptotically equivalent.

1.4 Case Control Study

In the previous section, i.e. II-1.3, common statistical methods for association analyses of random samples have been introduced. However, for certain binary traits, e.g. many common genetic diseases, which are of intensive interest but with a low prevalence in the population, it would be highly impractical and useless to perform a simple random sampling as most of the data so collected contain no information of such a trait, and hence alternative sampling strategies are required. In such a situation, Case-Control design is widely adopted in many scientific researches for binary outcomes. Generally, patients with a specific disease or syndrome are collected as case samples and individuals without such outcomes are collected as control samples, where both sampling procedures for case and control can still be assumed random, and further analyses are conducted to measure whether or not the pattern of an exposure, e.g. environment or genetic factors, is associated with the pattern of the Case-Control samples. Statistically, the Case-Control design can significantly increase the statistical power while dealing with rare genetic variants compared to the random sampling

design and hence in allelic association analysis of disease traits, Case-Control is probably the most commonly adopted study design.

Table II-2. Distribution of Genotypes in Case-Control Study.

Where n and N_i represent the number of individuals of each marker genotype in case and the whole data respectively, t and T represent the sample size of case and the whole data respectively, and x_i represents the trend coefficient for each marker genotype.

	MM	Mm	mm	Total
Case	n_1	n_2	n_3	t
Control	$N_1 - n_1$	$N_2 - n_2$	$N_3 - n_3$	$T - t$
Total	N_1	N_2	N_3	T

Suppose the observations of Case-Control samples are presented in a contingency table as Table II-2 with random sampling in both case and control, the distributions of marker genotypes conditioned on case and control individuals should be identical in expectation under the null hypothesis, where there is no LD between the testing marker locus and the disease causal QTL, and hence $E(k_i) = k \forall i$, where $k_i = \frac{n_i}{N_i}$ and $k = \frac{t}{T}$.

1.4.1 Pearson's χ^2 test

Taking the same notations from Table II-2, the Pearson's χ^2 test can be directly applied to calculate

$$\begin{aligned} & \sum_{i=1}^3 \frac{(n_i - tN_i/T)^2}{tN_i/T} + \frac{[N_i - n_i - (T-t)N_i/T]^2}{(T-t)N_i/T} \\ &= \sum_{i=1}^3 \frac{(N_i t - n_i T)^2}{N_i t (T-t)} \sim \chi_{df=2}^2 \end{aligned} \quad (\text{II-1.18})$$

Note here, since $E(k_i) = k \forall i$ under null hypothesis, it is easy to show that the variances of numbers of different genotypes are also identical in expectation, thus $\sigma^2 = E(\hat{\sigma}^2) = E(\hat{\sigma}_i^2) \forall i$,

where $\hat{\sigma}_i^2 = 2k_i(1-k_i)$ and $\hat{\sigma}^2 = 2k(1-k)$ if the binary phenotypes are assigned to be 1 and 0.

We may hence notice formula (II-1.9) in section II-1.3.2 is also valid here and is in proportion to formula (II-1.18) by a factor $\frac{T}{T-1}$, thus asymptotically equivalent to each other.

1.4.2 Allelic Analysis

From the distribution of genotypes shown in Table II-2, the observations of alleles can be calculated as Table II-3.

Table II-3. Distribution of Allele Frequencies in Case-Control Study

The parameters take the same definition as Table I-2.

	M	m	Total
Case	$2n_1 + n_2$	$2n_3 + n_2$	$2t$
Control	$2(N_1 - n_1) + N_2 - n_2$	$2(N_3 - n_3) + N_2 - n_2$	$2(T - t)$
Total	$2N_1 + N_2$	$2N_3 + N_2$	$2T$

Again, a Pearson's χ^2 test can be launched by calculating

$$\chi_A^2 = \frac{2T[T(2n_1 + n_2) - t(2N_1 + N_2)]^2}{t(T-t)[2T(2N_1 + N_2) - (2N_1 + N_2)^2]} \sim \chi_{df=1}^2. \quad (\text{II-1.19})$$

Alternatively, assign $\hat{p}_{M|D} = \frac{2n_1 + n_2}{2t}$, $\hat{p}_{M|C} = \frac{2(N_1 - n_1) + N_2 - n_2}{2(T-t)}$ to represent the frequencies

of allele M in Case and Control respectively. When T is sufficiently large, where both $\hat{p}_{M|D}$

and $\hat{p}_{M|C}$ can be assumed to follow a normal distribution asymptotically, a Z -statistic can then

be used to evaluate the difference between $\hat{p}_{M|D}$ and $\hat{p}_{M|C}$

$$\frac{\hat{p}_{M|D} - \hat{p}_{M|C}}{\sqrt{\text{Var}(\hat{p}_{M|D} - \hat{p}_{M|C})}} = \frac{\hat{p}_{M|D} - \hat{p}_{M|C}}{\sqrt{\frac{T\hat{p}(1-\hat{p})}{2t(T-t)}}} \sim Z, \quad (\text{II-1.20})$$

where $\hat{p} = \frac{2N_1 + N_2}{2T}$ and the variance term is calculated under the null hypothesis, where

$E(\hat{p}_{M|D}) = E(\hat{p}_{M|C}) = p$. As it is easy for us to examine that the statistical test (II-1.20) is identical to the one (II-1.19) given the fact $Z^2 = \chi_{df=1}^2$, only the χ_A^2 test will be mentioned as the allelic analysis henceforward.

Note that similar to formula (II-1.18), formula (II-1.19) from the Pearson's χ^2 test is asymptotically equivalent to the χ^2 test from ANOVA or SLR, where these two are equivalent due to the binary property of β in this situation. The proof is easy to acquire following the same route of formula (I-1.9) and is not repeated here.

1.4.3 The Armitage's Trend Test

As introduced in section II-1.3.1, the χ^2 test based on SLR can be applied irrespective of the exact distribution of phenotypes, and hence one may have

$$\frac{T \left[T \sum_i n_i x_i - t \sum_i N_i x_i \right]^2}{t(T-t) \left[T \sum_i N_i x_i^2 - (\sum_i N_i x_i)^2 \right]} \sim \chi_{df=1}^2. \quad (\text{II-1.21})$$

Assign $x_i = 3 - i \forall i$, formula (II-1.21) gives the most widely used form of Armitage's trend test

$$\chi_G^2 = \frac{T \left[T(2n_1 + n_2) - t(2N_1 + N_2) \right]^2}{t(T-t) \left[T(4N_1 + N_2) - (2N_1 + N_2)^2 \right]} \sim \chi_{df=1}^2. \quad (\text{II-1.22})$$

The idea of analysing a $n \times m$ contingency table using SLR was initially launched by Yates (1948), who also firstly suggested the use of equally spaced trend coefficients, x_i , if no priori

knowledge about such trend is available. Cochran (1954) and Armitage (1955) independently applied Yates' method into $n \times 2$ contingency table and hence formula (II-1.22) is also referred to as the Cochran-Armitage trend test.

Taking the same notation in allelic analysis, a *Z-statistic* equivalent to formula (II-1.22) (Schaid and Jacobsen, 1999) can be established as

$$\frac{\hat{p}_{M|D} - \hat{p}_{M|C}}{\sqrt{[\hat{p}(1 - \hat{p}) + \delta] \frac{T}{2t(T-t)}}} \sim Z, \quad (\text{II-1.23})$$

where $\delta = \hat{p}_{MM} - \hat{p}^2$ adjusts the bias of variance introduced by the deviation from HWE (Weir, 1990) and $\hat{p}_{MM} = \frac{N_1}{T}$ denotes the observed frequency of genotype *MM*.

1.4.4 Liability (Probit) and Logistic Model

As implicit models, both liability (probit) and logistic models can relate binary dependent variables to complex independent variables through introducing a cumulative distribution function with a specific cutting point or threshold in the form of

$$\Pr(y_i = 1 | \mathbf{X}_i) = F^{-1}(\mathbf{X}_i \mathbf{b}),$$

where \mathbf{X}_i denotes the i th column in a matrix of independent variables and \mathbf{b} denotes the vector of regression coefficients. Since $E(y_i | \mathbf{X}_i) = \Pr(y_i = 1 | \mathbf{X}_i)$ for binary data, the above formula indicates a link function

$$F(E(y_i | \mathbf{X}_i)) = \mathbf{X}_i \mathbf{b}, \quad (\text{II-1.24})$$

and hence both liability (probit) and logistic models are members of the generalised linear regression (GLR) model family, which will reduce to the linear regression while $F(E(\mathbf{Y} | \mathbf{X})) = E(\mathbf{Y} | \mathbf{X})$ as given by formula (I-1.13).

For the liability or probit model, the link function is given as

$$F(E(y_i | \mathbf{X}_i)) = \Phi^{-1}(E(y_i | \mathbf{X}_i)), \quad (\text{II-1.25})$$

where $\Phi^{-1}(\cdot)$ is the reverse of cumulative standard normal distribution function, and the expression of $\Phi(x)$ is given as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{t^2}{2}\right] dt. \quad (\text{II-1.26})$$

For a logistic model, the link function is given as

$$F(E(y_i | \mathbf{X}_i)) = \ln\left(\frac{E(y_i | \mathbf{X}_i)}{1 - E(y_i | \mathbf{X}_i)}\right), \quad (\text{II-1.27})$$

where $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$ is named a logit function as the inverse of a logistic function given

by $\text{logistic}(x) = \frac{e^x}{1+e^x}$ and hence a logistic model is often referred as a logit model as well.

The same as the linear model discussed above, both the liability (probit) and logistic models analyse the regression of phenotypes on marker genotypes directly without explicitly modelling the relationship between genotypes at marker loci and QTL. Following the introduction in section II-1.2.2, both the liability (probit) and logistic models assume that genetic effects of different marker genotypes follow an identical distribution and a specified threshold for each marker genotype is then assigned to separate binary phenotypes from this point, where the liability (probit) model assumes a normal distribution and the logistic model assumes a sech-square distribution. Unlike the likelihood based model, where the genotypic values of a putative QTL are more likely to follow a normal distribution, the genetic effects of the testing marker genotypes are related to genotypic values of the putative QTL through complicated functions, and hence the normality assumption may not stand. However, as it will

be shown later, i.e. in section II-2.4.1, in the scheme of score test, any general linear regression models under certain regular conditions are identical to the Armitage's trend test.

Here, we give a proof for the equivalence between a logistic model and the Armitage's trend test. Taking the same setting in Table II-2, let $\mathbf{X}_i \mathbf{b} = \alpha + \beta x_i$, the logistic model (II-1.27) takes the form

$$f_i = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}, \quad (\text{II-1.28})$$

where $f_i = E(y_i | \mathbf{X}_i)$. Since the observations in Table II-2 follow a multinomial distribution, ignoring the constant terms, the log-likelihood function can be given as

$$l(\alpha, \beta | x, y) \propto \sum_i n_i \log f_i + (N_i - n_i) \log(1 - f_i). \quad (\text{II-1.29})$$

It can be easily derived that under the null hypothesis where $\frac{t}{T} = f_i \forall i$

$$\mathbf{U}_0^T = \left(0 \quad \sum_i \frac{(n_i T - N_i t) x_i}{T} \right)$$

and

$$\mathbf{I}_0 = -E(\mathbf{H}_0) = \frac{T(T-t)}{T^2} \begin{pmatrix} T & \sum_i N_i x_i \\ \sum_i N_i x_i & \sum_i N_i x_i^2 \end{pmatrix}.$$

The corresponding score test given at section I-1.4.2 can hence be calculated as

$$\mathbf{U}_0^T \mathbf{I}_0^{-1} \mathbf{U}_0 = \frac{T \left[T \sum_i n_i x_i - t \sum_i N_i x_i \right]^2}{t(T-t) \left[T \sum_i N_i x_i^2 - \left(\sum_i N_i x_i \right)^2 \right]} \sim \chi_{df=1}^2,$$

which is clearly identical to the Armitage's trend test given as formula (II-1.22).

1.5 Rationale of this Study

In above, I have summarized the most basic statistical models for allelic association analyses. From the above introduction, it can be noticed that some of the statistical methods applied to random and non-random samples are highly related. Take the fixed effect (ANOVA) model and the Pearson's chi-square test for example, a fixed effect model initially assumes the normal distribution of the residual term to establish a valid *F*-statistic, however, such an assumption can be relaxed to any distribution if the observed sample size under each marker genotype is sufficiently large, and the Pearson's chi-square test is hence appropriate. Similar situation is also true between the SLR and the Armitage's trend test. However, the likelihood method based on the complete likelihood function is not available for the non-random samples, i.e. case control samples, because the marker allele frequency can not be directly estimated from the data. An alternative strategy to perform the likelihood based association study for case control samples will be introduced in Chapter III-2, but will not be mentioned in the following discussions. Apart from the general validity of both the SLR and the fixed effect (ANOVA) model, the comparison of their performance, e.g. the statistical power, is also of intensive interest. However, as the performance of SLR mainly depends on the choice of x_i as mentioned in section II-1.3.1, it would be necessary to evaluate the best or the optimal choice of x_i which can lead to maximize the statistical power of a SLR and control the rate of false positive. Meanwhile, as has been shown in section II-1.4.4, in the case of binary traits, the explicit and implicit models might result in asymptotically equivalent statistical test, which implies that if the sample size is sufficiently large it might not be necessary to adopt a generalised linear model instead of the ordinary linear models. In the following chapter, more general situation will be considered to acquire detailed statistical insight for analyses based on the explicit and implicit models.

CHAPTER II-2

COMPARISON OF STATISTICAL METHODS AND GENETIC MODELS

2.1 Overview

Although the power functions of both the SLR and the fixed effect model can be found elsewhere (Luo et al., 2000), in Chapter II-2, we will compare their statistical power in a more direct way. However, before a reasonable comparison is launched, the proper choice of x_i for the SLR has to be evaluated through understanding the biological meaning behind the ‘genetic effect’ of each marker genotype. Following such a comparison, we will discuss the relationship between the Armitage’s trend test and the allelic analysis for case control studies as we have noticed that one of the most influential papers (Sasieni, 1997) has certain flaws and insufficiency in its discussion of this topic. We will then show that implementation of the implicit model and the explicit model as introduced in section II-1.2 will generally result in asymptotically identical test statistics, and hence these two models are statistically equivalent. These general equivalences question the attempt to use the generalised linear models, and we might ask that to what extends a generalised linear model can provide more than an ordinary one? All of the three sub-topics we will discuss in Chapter II-2 will reasonably lead us to put more attention to the SLR based methods, and hence the re-analysis of the trend coefficients in the Armitage’s trend test will be launched in Chapter II-3.

2.2 Performance of Ordinary Linear Models

2.2.1 The Optimal Choice of x_i in the SLR for Random Samples

Unlike ANOVA, where the matrix \mathbf{X} in formula (I-1.17) is naturally defined, the x_i of the SLR is chosen arbitrarily, e.g. function (II-1.2) as suggested by Lande, R. & Thompson, R. (1990). In order to understand how the choice of x_i will affect the statistical inference, the relationship between x_i and the phenotypic values of QTL genotypes has to be explicitly characterized.

Suppose the testing marker is associated with a putative QTL, then the distribution of marker genotypes should have impacts on the phenotypes through the LD between the testing marker and the putative QTL. Such impacts are believed to be represented by ‘genetic effects’ for each marker genotypes as V_{MM} , V_{Mm} and V_{mm} , i.e. $E_{MM}(y) = \alpha + V_{MM}$ etc., where $E_{MM}(y)$ denotes the expectation of phenotype given marker genotype MM and α is a constant across different marker genotypes. If the ‘genetic effects’ are known, the regression of phenotypes on marker genotypes can be performed by assigning $x_1 = V_{MM}$, $x_2 = V_{Mm}$ and $x_3 = V_{mm}$. For a convenient way of comparison and calculation, following transformations are applied

$$x_1 = \frac{V_{MM} - V_{mm}}{V_{Mm} - V_{mm}}, \quad x_2 = \frac{V_{Mm} - V_{mm}}{V_{Mm} - V_{mm}} \quad \text{and} \quad x_3 = \frac{V_{mm} - V_{mm}}{V_{Mm} - V_{mm}}, \quad (\text{II-2.1})$$

which can be simply written as ψ , 1, 0 and will produce an identical test statistic to the original setting, i.e. $x_1 = V_{MM}$, $x_2 = V_{Mm}$ and $x_3 = V_{mm}$. It can be proved by noticing that

$$\begin{aligned} \text{Cor}^2(x, y) &= \text{Cor}^2(ax' + b, y) \\ &= \frac{\text{Cov}^2(ax' + b, y)}{\text{Var}(ax' + b)\text{Var}(y)}, \\ &= \frac{\text{Cov}^2(x', y)}{\text{Var}(x')\text{Var}(y)} \\ &= \text{Cor}^2(x', y) \end{aligned} \quad (\text{II-2.2})$$

where $x_i = ax'_i + b$, $a, b \in R$ and $a \neq 0$.

Similar to formula (II-1.3), we can now calculate

$$E(x) = \psi f_M^2 + 2f_M f_m = (\psi - 2)f_M^2 + 2f_M,$$

$$\begin{aligned} E(y) &= f_A^2 V_{AA} + 2f_A f_a V_{Aa} + f_a^2 V_{aa} \\ &= \mu + (f_A^2 - f_a^2)a + 2f_A f_a d, \end{aligned}$$

$$\begin{aligned} E(xy) &= \psi(f_{AAMM}V_{AA} + f_{AaMM}V_{Aa} + f_{aaMM}V_{aa}) + (f_{AAMm}V_{AA} + f_{AaMm}V_{Aa} + f_{aaMm}V_{aa}) \\ &= 2[f_{AM}f_A V_{AA} + (f_{AM}f_a + f_{aM}f_A)V_{Aa} + f_{aM}f_a V_{aa}] + (\psi - 2)(f_{AM}^2 V_{AA} + 2f_{AM}f_{aM}V_{Aa} + f_{aM}^2 V_{aa}) \\ &= 2[f_M \mu + (f_{AM}f_A - f_{aM}f_a)a + (f_{AM}f_a + f_{aM}f_A)d] + (\psi - 2)[f_M^2 \mu + (f_{AM}^2 - f_{aM}^2)a + 2f_{AM}f_{aM}d] \end{aligned}$$

$$E(x^2) = \psi^2 f_M^2 + 2f_M f_m = 4f_M^2 + 2f_M f_m + (\psi^2 - 4)f_M^2,$$

and hence

$$\begin{aligned} E(\hat{\beta}) &= \frac{E(xy) - E(x)E(y)}{E(x^2) - E^2(x)} \\ &= \frac{2D[a + (1 - 2q)d] + (\psi - 2) \left\{ (f_{AM}^2 - f_A^2 f_M^2 - f_{aM}^2 + f_a^2 f_M^2)a + 2(f_{AM}f_{aM} - f_M^2 f_A f_a)d \right\}}{2p(1 - p) + (\psi - 2)f_M^2[(\psi + 2) - 4f_M - (\psi - 2)f_M^2]} \quad (\text{II-2.3}) \\ &= \frac{2D[a + (1 - 2q)d] + 2(\psi - 2)D[pa - (D + 2pq - p)d]}{2p(1 - p) + (\psi - 2)p^2(1 - p)[(\psi - 2)p + (\psi + 2)]} \end{aligned}$$

It can be noticed that $E(\hat{\beta})$ as given in formula (II-2.3) is always proportional to D

irrespective of the value of ψ , e.g. $E(\hat{\beta}) = 0$ under the null hypothesis, and hence from the

discussion given at section II-1.3.1, the statistical test $\frac{\hat{\beta}^2}{\text{Var}(\hat{\beta})} \approx n\hat{\rho}_{xy}^2 \sim \chi_{df=1}^2$ will always be

valid irrespective of the choice of ψ . However, if different ψ are chosen, the test statistics do

vary, and hence the loss of statistical power would be expected once the inappropriate ψ is

implemented due to the lack of prior information about the ‘genetic effects’.

On the above formulation, how do we choose the best or ‘optimal’ ψ to be fitted in the SLR?

With the ‘optimal’ ψ , the corresponding statistical test should have its statistical power maximized and its false positive rate minimized in expectation. Since the complete solution to this question would be really cumbersome, we closely discuss here only the case where $d = 0$, i.e. co-dominance model. A general result will be given later without the details of derivation.

Firstly, since $E(\hat{\beta}) = 0$ under the null hypothesis from formula (II-2.3), the corresponding expectation of the test statistic $\frac{\hat{\beta}^2}{Var(\hat{\beta})}$ clearly equals 0 and thus the minimum. The second requirement for the optimality is automatically fulfilled.

Secondly, since both n and $Var(y)$ are independent of ψ , it is equivalent to maximize the test statistic $n\hat{\rho}_{xy}^2$ or to maximize $\frac{Cov^2(x, y)}{Var(x)}$. When $d = 0$, it can be calculated that

$$\begin{aligned}\frac{Cov^2(x, y)}{Var(x)} &= \frac{[E(xy) - E(x)E(y)]^2}{E(x^2) - E^2(x)} \\ &= \frac{(2Da)^2 [p\psi + (1-2p)]^2}{2p(1-p) + (\psi-2)p^2(1-p)[(\psi-2)p + (\psi+2)]} \\ &= \frac{(2Da)^2}{p(1-p)} \times \frac{p^2\psi^2 + 2p(1-2p)\psi + (1-2p)^2}{p(1+p)\psi^2 - 4p^2\psi + (1-2p)^2 + 1} \\ &= \frac{(2Da)^2}{p(1-p)} \times \frac{1}{1 + \frac{1}{p + (1-p) \times \frac{2p(\psi-2)+1}{xp(\psi-2)+1}}}\end{aligned}$$

It is easy to show that the maximum of $\frac{2p(\psi-2)+1}{\psi p(\psi-2)+1}$ is acquired at $\psi = 2$ and so as the maximum of the test statistic.

From above derivations, the suggestion proposed by Lande & Thompson turns out to be the optimal choice under the co-dominance model. Taking one step further, we can show that

$$\hat{\rho}_{xy}^2 = \lambda^2 = \frac{D^2}{pq(1-p)(1-q)} \text{ given } \psi = 2 \text{ and } d = 0, \text{ where } \lambda \text{ is the correlation coefficient}$$

between marker allele M and A as defined by formula (I-1.4). This result can be explained by realizing that when $d = 0$, the genotypic values of QTL genotypes $V_{AA} = \mu + a$, $V_{Aa} = \mu$ and $V_{aa} = \mu - a$ are equivalent to $V_{AA} = 2$, $V_{Aa} = 1$ and $V_{aa} = 0$ from formula (II-2.2), and hence the correlation between genotypic values of QTL genotypes and the ‘genetic effects’ of marker genotypes under the co-dominance model turns out to be the correlation between the QTL marker allele M and the QTL allele A .

For a general case, where $d \neq 0$, it should be noticed from formula (II-2.3) D is complexly involved and hence can not be eliminated while maximizing $\hat{\rho}_{xy}^2$. Neglecting the tedious calculations, the optimal ψ can be computed through equalling the first derivative of

$\frac{Cov^2(x, y)}{Var(x)}$ to 0 and by solving the corresponding equation, we can have

$$\psi = 2 - \frac{2Dd}{ap(1-p) + dp[2D + (1-p)(1-2q)]} \quad (\text{II-2.4})$$

It is easy to show that the optimal ψ given as formula (II-2.4) is coincidence to assign x_i with the ‘genetic effects’ as defined above, i.e. $x_1 = \alpha + V_{MM} = E_{MM}(y)$, $x_2 = \alpha + V_{Mm} = E_{Mm}(y)$ and $x_3 = \alpha + V_{mm} = E_{mm}(y)$, and hence taking the ‘genetic effects’ as x_i can automatically maximize the statistical power in expectation. Since the calculation of $E_{MM}(y)$ and etc. is always applicable irrespective of whether the data is random or not, the ‘genetic effects’ can hence be adopted as a general strategy of establishing the optimal choice of ψ as indicated by formulae (II-2.1). However, even knowing this, the optimal ψ can not

be specified without knowledge of D and q , although the information of a , d is given prior to the test and p can be directly estimated from the sample. This limitation definitely obstructs any attempt to derive the optimal ψ ; however, as will be shown in Chapter II-3, such influence from unknown D and q might be largely reduced in a Case-Control study.

It has to be addressed here that the validation of formula (II-2.4) requires $d \neq -a$, i.e. the recessive model, because otherwise the transformation (II-2.1) is not defined. To overcome such a deficiency, an alternative transformation can be applied as

$$x_i = \frac{E_i(y) - E_3(y)}{E_1(y) - E_3(y)} \forall i. \quad (\text{II-2.5})$$

It is easy to prove that $\lim_{E_1(y) \rightarrow E_3(y)} x_1 = 1$, $0 < \lim_{E_1(y) \rightarrow E_3(y)} x_2 < 1$ and $\lim_{E_1(y) \rightarrow E_3(y)} x_3 = 0$ in the absence of over-dominance, and hence transformation (II-2.5) will always have a definition.

Although the exact form of the ‘genetic effects’ can not be specified as indicated above, it is possible to write such ‘genetic effects’ in terms of genotypic values of QTL genotypes as shown in Table II-1

$$\begin{aligned} V_{MM} &= \mu_M + a_M = \mu_M + a + \frac{1}{2}d \times \left(\frac{f_{ma} - f_{mA}}{f_m} + \frac{f_{Ma} - f_{MA}}{f_M} \right), \\ V_{Mm} &= \mu_M + d_M = \mu_M + \frac{1}{2}d \times \left(\frac{f_{ma} - f_{mA}}{f_m} - \frac{f_{Ma} - f_{MA}}{f_M} \right), \\ V_{mm} &= \mu_M - a_M = \mu_M - a - \frac{1}{2}d \times \left(\frac{f_{ma} - f_{mA}}{f_m} + \frac{f_{Ma} - f_{MA}}{f_M} \right), \end{aligned} \quad (\text{II-2.6})$$

where $\mu_M = (V_{MM} + V_{mm}) / 2$, $a_M = V_{MM} - \mu_M = a + \frac{1}{2}d \times \left(\frac{f_{ma} - f_{mA}}{f_m} + \frac{f_{Ma} - f_{MA}}{f_M} \right)$ and

$d_M = V_{Mm} - \mu_M = \frac{1}{2}d \times \left(\frac{f_{ma} - f_{mA}}{f_m} - \frac{f_{Ma} - f_{MA}}{f_M} \right) = \frac{D}{f_M f_m} d$. Formulae (II-2.6) hence present the

relationship between the ‘genetic effects’ of marker genotypes and the genotypic values of

QTL. Without loss of generality, formulae (II-2.6) can be denoted as the definition of the ‘genetic effects’ of marker genotypes. Although those formulae have been written into similar forms, the ‘genetic effects’ defined as formulae (II-2.6) and the genotypic values of QTL given at Table II-1 have different statistical genetic properties. For example, in the dominance model, i.e. $d = a$, it is not guaranteed to claim $V_{MM} = V_{Mm}$, since $V_{MM} - V_{Mm} = 2a \frac{f_{Ma}}{f_M}$ which equals 0 if and only if $f_{Ma} = 0$, which is not generally true unless the marker allele M and the QTL allele A are completely linked as well as $p < q$.

As an alternative to maximizing the expected value of χ^2 , we may be interested in a regression formulation which yields minimum unexplained variance, or equivalently maximum explained variance. Notice from formula (I-1.24), i.e. $\hat{\sigma}_e^2 = \frac{n-1}{n-2} \hat{\sigma}_y^2 (1 - \hat{\rho}_{xy}^2)$, it is equivalent to minimize $\hat{\sigma}_e^2$ and to maximize $\hat{\rho}_{xy}^2$, and hence the ‘genetic effects’ of marker genotypes not only maximize the test statistic in expectation but also minimize the expectation of unexplained genetic variance at the QTL.

2.2.2 Comparison of Statistical Power of SLR and ANOVA

In association studies, we are more interested in the ability of a model to conclude a current or pervious state from the observed data than to predict a further state, and hence it might be of more concern about the probability of false statistical inference than the goodness of fit, i.e. we will be more interested in improving the statistical power rather to acquire a sound goodness of fit statistic, e.g. the coefficient of determination (Steel and Torrie, 1960). As a matter of fact, for most complex traits, a single SNP could only explain very little variance, and the corresponding goodness of fit for such a model will not be impressive at all. With

such a concept, we will focus on the comparison of the statistical powers between the SLR and the ANOVA irrespective their goodness of fit.

Asymptotically, the χ^2 test statistic for ANOVA has been given as formulae (II-1.9), i.e.

$$n-1 \frac{\sum_{i=1}^3 n_i (\bar{y}_{i.} - \bar{y})^2}{\sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2} \sim \chi_{df=2}^2,$$

where $\bar{y}_{i.}$ is the sample mean of phenotype given the i th marker genotype, and \bar{y} is the sample mean of the full sample. We may notice that the $\chi_{df=1}^2$ test statistic for SLR can be explicitly written as:

$$\begin{aligned} n\hat{\rho}_{xy}^2 &= n \frac{\hat{\sigma}_{xy}^2}{\hat{\sigma}_x^2 \hat{\sigma}_y^2} \\ &= n \frac{\left[\sum_i (x_i - \bar{x}) \sum_j (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}) \right]^2}{\sum_{ij} (y_{ij} - \bar{y})^2 \times \sum_i n_i (x_i - \bar{x})^2} \\ &= n \frac{\sum_i n_i (\bar{y}_{i.} - \bar{y})^2}{\sum_{ij} (y_{ij} - \bar{y})^2} \times \frac{\left[\sum_i n_i (x_i - \bar{x}) (\bar{y}_{i.} - \bar{y}) \right]^2}{\sum_i n_i (x_i - \bar{x})^2 \times \sum_i n_i (\bar{y}_{i.} - \bar{y})^2}, \quad (\text{II-2.7}) \\ &= n \frac{\sum_i n_i (\bar{y}_{i.} - \bar{y})^2}{\sum_{ij} (y_{ij} - \bar{y})^2} \times \rho^2(x_i, \bar{y}_{i.}) \end{aligned}$$

where $\rho^2(x_i, \bar{y}_{i.})$ is the correlation between x_i and $\bar{y}_{i.}$ calculated by assigning $\bar{y}_{i.}$ to each corresponding individual. Clearly, formula (II-2.7) is asymptotically proportional to $\rho^2(x_i, \bar{y}_{i.})$ by the second expression of formulae (II-1.9) as given before.

At a significant level α , when the alternative hypothesis is true, i.e. the testing marker locus is associated with a QTL, the statistical test bearing a higher chance of claiming positive has a higher statistical power and hence is preferable to the other. Since the confidence threshold

for a $\chi^2_{df=2}$ test statistic, e.g. formulae (II-1.9), is higher than that of a $\chi^2_{df=1}$ test statistic, e.g. formula (II-1.19), the comparison between the ratio of such confidence thresholds and $\rho^2(x_i, \bar{y}_i)$ is hence crucial to determine the performance between SLR and ANOVA, i.e. the statistical power. Ideally, if the optimal x_i is known, say \hat{x}_i , since $\lim_{n \rightarrow \infty} \bar{y}_i = E_i(y) = \alpha + V_i$, where $E_i(y)$ and V_i are the population mean of phenotype and the ‘genetic effect’ given the i th marker genotype, we can have $\lim_{n \rightarrow \infty} \rho^2(\hat{x}_i, \bar{y}_i) = 1$. In such a circumstance, the statistical power of a SLR will always be higher than that of an ANOVA, because the ratio of confidence thresholds between a $\chi^2_{df=1}$ test statistic and a $\chi^2_{df=2}$ test statistic under the same significant level is always less than 1. If no prior information is available, it would be preferred to choose equally spaced ‘genetic effects’, e.g. $x_i = 3 - i \forall i$, which is the optimal choice given $d = 0$ or co-dominance model as shown above. However, even while $x_i = 3 - i \forall i$ is far from the optimal choice, a SLR might still have a higher statistical power than an ANOVA. For example, suppose $\bar{y}_1 = \bar{y}_2 \neq \bar{y}_3$, taking $x_i = 3 - i \forall i$, we can calculate $\rho^2(x_i, \bar{y}_i) = 1 - p$, and similarly, $\rho^2(x_i, \bar{y}_i) = p$ while $\bar{y}_1 \neq \bar{y}_2 = \bar{y}_3$, where the HWE is assumed. Since the ratio of commonly used confidence thresholds between a $\chi^2_{df=1}$ test statistic and a $\chi^2_{df=2}$ test statistic are $r_{.95} = 0.64$, $r_{.99} = 0.72$ and $r_{.995} = 0.74$, where r_p denotes the ratio of $P\%$ confidence thresholds between a $\chi^2_{df=1}$ test statistic and a $\chi^2_{df=2}$ test statistic, we can notice that a SLR might still have a higher statistical power than an ANOVA for a certain value of p , i.e. the frequency of marker allele M , even while the optimality of chosen x_i is seriously violated.

From the discussions above, we can conclude that it would be recommended to perform a SLR rather than an ANOVA in order to acquire a higher statistical power, especially while the ‘genetic effects’ of maker genotypes can be acquired.

2.3 The HWE in Case-Control Study

Denoting the test statistic (II-1.19) of the allelic analysis as χ_A^2 and the test statistic (II-1.22) of the Armitage's trend test as χ_G^2 , Sasieni, P. D. (1997) compared the χ^2 tests of allelic analysis and the Armitage's trend test by calculating

$$\frac{\chi_A^2}{\chi_G^2} = 1 + \frac{4N_1N_3 - N_2^2}{(2N_1 + N_2)(N_2 + 2N_3)}. \quad (\text{II-2.8})$$

Formula (II-2.8) is definitely a very useful formula and has revealed the relationship between the allelic analysis and the Armitage's trend test, e.g. the test statistics would be identical between these two methods if and only if the HWE holds in the joint samples. In his paper, Sasieni argued that the Armitage's trend test is better than the allelic analysis, and we will later show that this statement is reasonable. However, Sasieni's conclusion is based on several claims that are highly in doubt. Firstly, he claimed that the HWE in both cases and controls does not guarantee the HWE of the joint sample. Since such deviation from the HWE can be properly adjusted by the Armitage's trend test, it is preferable to the allelic analysis; Secondly, he claimed that the observation $2n_1 + n_2$ from Table II-3 can only follow a binomial distribution under the HWE and hence the allelic analysis can only be valid if both cases and control are in HWE. Unfortunately, both the claims are inadequate or even mis-claimed, and such questionable claims have already been adopted by many other researchers, for example, Balding (2006) also mentioned the requirement of HWE for the allelic analysis after the citation of Sasieni's paper. We would like to explain why these claims are questionable as follows. Firstly, a joint population of two HWE subpopulations, say cases and controls, will not follow HWE only if these two subpopulations have different genotype patterns, i.e. different allele frequencies. However, as the validation of a test statistic is generally

considered under the null hypothesis, i.e. $D=0$, which implies $f_i = \frac{n_i}{N_i}$ should equal each other in expectation; in such a circumstance, the genotype distribution should be identical in both the ‘cases’ and the ‘controls’. The first claim is hence inadequate. Secondly, the observation that number $2n_1 + n_2$ follows a binomial distribution only requires the simple random sampling for both cases and controls from the same population, i.e. with the same allele frequency, irrespective of whether or not the HWE holds. Since such a requirement is always true under the null hypothesis, the allelic analysis is generally valid. The second claim given by Sasieni is hence misleading. The true reason to choose the Armitage’s trend test over the allelic analysis lies in the fact that the process of counting allele number in the allelic analysis does not use the information of the genotype distributions. The information includes not only the HWE but also the trend coefficients, and hence although the allelic analysis is generally valid, it is less informative than the Armitage’s trend test. However, if there are too few observations for a certain marker genotype, e.g. due to the low marker allele frequency, the Armitage’s trend test might be highly vulnerable to the genotyping errors and its test statistic might be seriously biased from the $\chi^2_{df=1}$ (Yates, 1934). In such a circumstance, the allelic analysis is more reliable than the Armitage’s trend test.

Note here, the deviation of HWE will only affect the variance parts, or denominator, of formulae (II-1.19) and (II-1.22), and such deviations may be largely attributed to two major sources, the random sampling, i.e. random drift of genotype distribution from its theoretical distribution, and the non-random sampling, e.g. population stratification as introduced in section I-1.2.4. If a deviation is introduced by a random sampling, the Armitage’s trend test can correctly adjust for such a bias of HWE as pointed out by Weir (1990) and hence give a more reliable result, although such an improvement is normally fairly small unless extreme situation happens. Otherwise, if such a deviation is introduced by a population stratification,

as will be shown in section II-3.3, both methods, i.e. the allelic analysis and the Armitage's trend test, will tend to overestimate the variance of the regression coefficients and hence result in a loss of statistical power. Note here, population stratification may also introduce spurious LDs, which may either increase or decrease the numerator of formulae (II-1.19) and (II-1.22), and cause serious false inferences. However, as will be shown in Chapter II-3, the influence from population stratification on the numerators and the denominators of formulae (II-1.19) and (II-1.22) might not be always present together, and we must treat such a situation carefully. A more comprehensive discussion of the influence from population stratification in Case-Control studies will be given in Chapter II-3.

2.4 The Equivalence between Models of Quantitative Genetic Effects

As has been mentioned in section II-1.4.4, the score tests of both logistic and liability (probit) models are asymptotically identical to the Armitage's trend test if a generalised SLR model $F(E(y_i)) = \alpha + \beta x_i$ is adopted. Since identical test statistics will yield identical statistical inferences, these models are asymptotically equivalent in statistical analysis. Notice that the Armitage's trend test is an explicit model and the other two are implicit models, their asymptotical equivalence stimulates the idea that there might be a more general equivalence between explicit and implicit models, and that is what we will show in this section.

2.4.1 Under the Generalised SLR

As the equivalence between the logistic model and the Armitage's trend test under the generalised SLR has been demonstrated in section II-1.4.4, in the following we will show that under the generalised SLR, such equivalence is generally valid.

Consider the log-likelihood function introduced as function (II-1.29)

$$l(\alpha, \beta | x, y) \sim \sum_i n_i \log f_i + (N_i - n_i) \log(1 - f_i).$$

Suppose a link function takes the form of $g^{-1}(f_i) = z_i$, where $z_i = \alpha + \beta x_i$ and $g(\cdot)$ denotes any monotonic, continuous and smooth function defined on \mathbf{R} , from the discussion before and under the null hypothesis, i.e. $D=0$, we can have $z_{i0} = \alpha_0 = z_0 \forall i$, $\beta_0 = 0$ and $\frac{t}{T} = f_i \forall i$,

where α_0 and β_0 are the MLEs of α and β for function (II-1.29) under the null hypothesis.

To establish the score test, we can firstly calculate the first and second-order derivatives of function (II-1.29) as

$$\begin{aligned} \frac{\partial l(\alpha, \beta | x, y)}{\partial \xi} &= \sum_i n_i \frac{1}{f_i} \frac{\partial f_i}{\partial \xi} - \sum_i (N_i - n_i) \frac{1}{1 - f_i} \frac{\partial f_i}{\partial \xi}, \\ &= \sum_i \frac{n_i - N_i f_i}{f_i(1 - f_i)} \frac{\partial g(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \xi}, \end{aligned} \quad (\text{II-2.9})$$

$$\begin{aligned} \frac{\partial^2 l(\alpha, \beta | x, y)}{\partial \xi \partial \psi} &= \sum_i \left(-n_i \frac{1}{f_i^2} \frac{\partial f_i}{\partial \xi} \frac{\partial f_i}{\partial \psi} + n_i \frac{1}{f_i} \frac{\partial^2 f_i}{\partial \xi \partial \psi} \right) + \\ &\quad \sum_i \left[-(N_i - n_i) \frac{1}{(1 - f_i)^2} \frac{\partial f_i}{\partial \xi} \frac{\partial f_i}{\partial \psi} - (N_i - n_i) \frac{1}{1 - f_i} \frac{\partial^2 f_i}{\partial \xi \partial \psi} \right], \\ &= -\sum_i \left[\frac{N_i - n_i}{(1 - f_i)^2} + \frac{n_i}{f_i^2} \right] \left(\frac{\partial g(z_i)}{\partial z_i} \right)^2 \frac{\partial z_i}{\partial \xi} \frac{\partial z_i}{\partial \psi} + \\ &\quad \sum_i \frac{n_i - N_i f_i}{f_i(1 - f_i)} \left[\frac{\partial^2 g(z_i)}{\partial z_i^2} \frac{\partial z_i}{\partial \xi} \frac{\partial z_i}{\partial \psi} + \frac{\partial g(z_i)}{\partial z_i} \frac{\partial^2 z_i}{\partial \xi \partial \psi} \right] \end{aligned} \quad (\text{II-2.10})$$

where either of ξ or ψ represents one of α and β . From the discussion before, under the

null hypothesis, i.e. $D=0$, we have $z_{i0} = \tilde{\alpha}_0 = z_0 \forall i$, $\tilde{\beta}_0 = 0$ and $\frac{t}{T} = f_i \forall i$, where $\tilde{\alpha}_0$ and $\tilde{\beta}_0$

are the MLEs of α and β from function (II-1.29) under the null hypothesis. Incorporating

above results into formulae (II-2.9) and (II-2.10), and by noticing that $\frac{\partial^2 z_i}{\partial \xi \partial \psi} \equiv 0$, we can

have

$$\frac{\partial l(\tilde{\alpha}, \tilde{\beta} | x, y, D=0)}{\partial \xi} = \frac{T}{t(T-t)} g'(z_0) \sum_i (Tn_i - N_i t) \frac{\partial z_i}{\partial \xi}$$

and

$$\begin{aligned} E\left(\frac{\partial^2 l(\tilde{\alpha}, \tilde{\beta} | x, y, D=0)}{\partial \xi \partial \psi}\right) &= -T^2 g''(z_0) \sum_i \frac{\partial z_i}{\partial \xi} \frac{\partial z_i}{\partial \psi} E\left[\frac{N_i - n_i}{(T-t)^2} + \frac{n_i}{t^2}\right] + \\ &\quad \frac{T}{t(T-t)} g''(z_0) \sum_i \frac{\partial z_i}{\partial \xi} \frac{\partial z_i}{\partial \psi} E(Tn_i - N_i t) \quad , \\ &= -\frac{T^2}{t(T-t)} g''(z_0) \sum_i \frac{\partial z_i}{\partial \xi} \frac{\partial z_i}{\partial \psi} N_i \end{aligned}$$

where g' and g'' are the first and second-order partial derivative functions of g respectively,

and $E(Tn_i - N_i t)$ is null by noticing that $E(n_i) = N_i \times \frac{t}{T}$ under the null hypothesis where

$\frac{t}{T} = f_i \forall i$. The corresponding score test can hence be established as

$$\begin{aligned} \mathbf{U}_0^T &= \left(0 \quad \frac{T^2}{t(T-t)} g''(z_0) \sum_i \frac{(n_i T - N_i t) x_i}{T} \right), \\ \mathbf{I}_0 = -E(\mathbf{H}_0) &= g''(z_0) \frac{T^2}{t(T-t)} \begin{pmatrix} T & \sum_i N_i x_i \\ \sum_i N_i x_i & \sum_i N_i x_i^2 \end{pmatrix}, \end{aligned}$$

and

$$\mathbf{U}_0^T \mathbf{I}_0^{-1} \mathbf{U}_0 = \frac{T \left[T \sum_i n_i x_i - t \sum_i N_i x_i \right]^2}{t(T-t) \left[T \sum_i N_i x_i^2 - (\sum_i N_i x_i)^2 \right]} \sim \chi_{df=1}^2,$$

where \mathbf{H}_0 is the Hessian matrix under the null hypothesis, i.e. the matrix of second-order partial derivatives given $D=0$, introduced as formula (I-1.8). It can be easily noticed that the corresponding score test is identical to formula (II-1.21) irrespective of the exact form of function g , if only $\frac{\partial g(z)}{\partial z} \neq 0$ and $\frac{\partial^2 g(z)}{\partial z^2}$ exists, e.g. g is monotonic, continuous and smooth.

Clearly, both liability (probit) and logistic models belong to this category.

Note here, if we take $g(z_i) = z_i$, i.e. an identity function, the above result will still hold, as the identity function is clearly monotonic, continuous and smooth. However, it can be noticed that given the same link function $E_i(y) = f_i = \alpha + \beta x_i$, the MLEs of function (II-1.29) for α and β are different from their corresponding BLUEs despite the asymptotically equivalent test statistics.

2.4.2 Under the Generalised Fixed Effect Model

Take the same log-likelihood function (II-1.29) as section II-2.4.1, but consider a generalised fixed effect model as $g^{-1}(E(y_i | \mathbf{X})) = g^{-1}(f_i) = \beta_i$ instead of the generalised SLR, we can calculate the first and second derivatives of function (II-1.29) under the scheme of generalised fixed effect model as

$$\begin{aligned} \frac{\partial l(\beta_k, k=1,2,3 | x, y)}{\partial \beta_i} &= \sum_k n_k \frac{1}{f_k} \frac{\partial f_k}{\partial \beta_i} - \sum_k (N_k - n_k) \frac{1}{1-f_k} \frac{\partial f_k}{\partial \beta_i}, \\ &= \frac{n_i - N_i f_i}{f_i(1-f_i)} g'(\beta_i) \end{aligned} \quad (\text{II-2.11})$$

$$\frac{\partial^2(\beta_k, k=1,2,3 | x, y)}{\partial \beta_i \partial \beta_j} = \frac{n_i - N_i f_i}{f_i(1-f_i)} g''(\beta_i) - \left(\frac{n_i}{f_i^2} + \frac{N_i - n_i}{(1-f_i)^2} \right) g''(\beta_i) \quad \text{if } i = j$$

and (II-2.12)

$$\frac{\partial^2(\beta_k, k=1,2,3 | x, y)}{\partial \beta_i \partial \beta_j} = 0 \quad \text{if } i \neq j,$$

where $i, j = 1, 2, 3$.

Under the null hypothesis, with the same results introduced in section II-2.4.1 that $\frac{t}{T} = f_i \forall i$

and $E(n_i) = N_i \times \frac{t}{T}$, we can hence have

$$U_i = \frac{\partial l(\tilde{\beta}_k, k=1,2,3 | x, y, D=0)}{\partial \beta_i} = T \frac{Tn_i - N_i t}{t(T-t)} g'(\tilde{\beta}_i),$$

$$I_{ij} = -E\left(\frac{\partial^2(\tilde{\beta}_k, k=1,2,3 | x, y, D=0)}{\partial \beta_i \partial \beta_j}\right) = 0 \quad \text{if } i \neq j$$

and

$$I_{ij} = -E\left(\frac{\partial^2(\tilde{\beta}_k, k=1,2,3 | x, y, D=0)}{\partial \beta_i \partial \beta_j}\right) = \frac{T^2 N_i}{t(T-t)} g''(\tilde{\beta}_i) \quad \text{if } i = j.$$

By noticing $\bar{y} = \frac{t}{T}$ and $\bar{y}_i = \frac{n_i}{N_i}$, we can establish the score test for the generalised fixed

effect model as

$$\mathbf{U}_0 = (U_1 \quad U_2 \quad U_3)^T,$$

$$\mathbf{I}_0 = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix}$$

and hence

$$\begin{aligned} \mathbf{U}_0^T \mathbf{I}_0^{-1} \mathbf{U}_0 &= \sum_i \frac{(N_i t - n_i T)^2}{N_i t (T-t)} \\ &= \sum_i \frac{N_i (\bar{y}_i - \bar{y})^2}{\bar{y}(1-\bar{y})} \sim \chi_{df=2}^2, \end{aligned}$$

which is clearly identical to the Pearson's chi-square test (II-1.19) and hence is asymptotically equivalent to formulae (II-1.9), i.e. the *F-test* statistic of the fixed effect model.

2.4.3 Under the Scheme of a Likelihood-based Approach

For the likelihood based method of a random sample as introduced in section II-1.3.3, the score test is impractical due to a complicated information matrix hence derived, and we will prove the statistical equivalence between the implicit and explicit models under the LRT.

Under the penetrance model as introduced at formula (II-1.16), i.e. the explicit model of a binary trait for a random sample, the ECM algorithm of binary data at t -th iterative step can be given as:

E-step:

Formula (II-1.12) now takes the form of

$$w_{ijk}^t = \frac{h_{ik}^t (f_k^t)^{y_{ij}} (1 - f_k^t)^{1-y_{ij}}}{\sum_{k=1}^3 h_{ik}^t (f_k^t)^{y_{ij}} (1 - f_k^t)^{1-y_{ij}}}.$$

CM-step:

It is clear that penetrance coefficients are independent of each other if w_{ijk}^t are treated as constant, and hence each penetrance coefficient can be updated independently as

$$f_k^{t+1} = \frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} w_{ijk}^t y_{ij}}{\sum_{i=1}^3 \sum_{j=1}^{n_i} w_{ijk}^t}, \quad (\text{II-2.13})$$

where the other parameters, i.e. D , p and q , can be updated independently of f_k (Luo and Wu, 2001).

For implicit models, taking $f_k = g(\mu_k)$, where $g(\cdot)$ denotes any monotonic, smooth and continuous function defined on \mathbf{R} and $\mu_k = \mu + (2-k)a + \frac{1-(-1)^k}{2}d$, we can update new

parameters at the t -th CM-step through equating following formulae to 0

$$\begin{aligned}
\frac{\partial l_c(\Omega^t | y, M)}{\partial \xi} &= \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 w_{ijk}^t \left(\frac{y_{ij}}{f_k} \frac{\partial f_k}{\partial \xi} - \frac{1-y_{ij}}{1-f_k} \frac{\partial f_k}{\partial \xi} \right) \\
&= \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 w_{ijk}^t \left[\frac{y_{ij} - f_k}{f_k(1-f_k)} \frac{\partial f_k}{\partial \mu_k} \frac{\partial \mu_k}{\partial \xi} \right], \\
&= \sum_{k=1}^3 \left\{ \frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} w_{ijk}^t (y_{ij} - f_k)}{f_k(1-f_k)} \frac{\partial f_k}{\partial \mu_k} \frac{\partial \mu_k}{\partial \xi} \right\}
\end{aligned} \tag{II-2.14}$$

where ξ denotes either of a , d or μ . Denoting $\frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} w_{ijk}^t (y_{ij} - f_k)}{f_k(1-f_k)} \frac{\partial f_k}{\partial \mu_k} = P_k^t$, we can notice

that equating formulae (II-2.14) to 0 is equivalent to the following equations

$$\begin{aligned}
P_1^t + P_2^t + P_3^t &= 0, \\
P_1^t - P_3^t &= 0
\end{aligned} \tag{II-2.15}$$

and

$$P_2^t = 0,$$

of which the only solution is $P_1^t = P_2^t = P_3^t = 0$. This result implies that

$$\sum_{i=1}^3 \sum_{j=1}^{n_i} w_{ijk}^t (y_{ij} - f_k) = 0 \forall k, \text{ which is identical to formulae (II-2.13) from the explicit model,}$$

and hence the sub-CM-steps of updating parameters are related to the penetrance coefficients for explicit and implicit models, i.e. formulae (II-2.13) and (II-2.15) respectively, are mutually deductive. Since such a mutual derivation will always hold during each CM step, we can conclude that the MLEs of explicit and implicit models hence derived are mutually deductive. Notice such a mutual derivation does not rely on any hypothesis, i.e. null or alternative, the test statistics of LRTs derived through formulae (II-2.13) and (II-2.15) are hence identical, and the explicit and implicit models are statistically equivalent for a random sample under the scheme of a likelihood-based approach.

2.4.4 Conclusion

Above derivations show that explicit and implicit models are statistically equivalent to each other under a quite general situation, and due to the simplicity of manipulating an explicit model, it seems rather rational for us to implement an explicit model rather than an implicit one unless we can prove that the introduction of an implicit model does significantly improve the statistical inferences, e.g. the sample size is fairly small and the asymptotical equivalence does not stand. It can also be noted here that even in the presence of covariates, e.g. the dummy variables that are introduced to correct the population stratification, such statistical equivalence can also be proved following a similar procedure as introduced in section II-2.4.1 and II-2.4.2.

2.5 Modified Armitage's Trend Test

From the above discussions, the general statistical equivalence between explicit and implicit models results in the preference to the explicit model due to its mathematical simplicity, and we have also shown that an SLR generally has a higher statistical power than that of a mixed effect model especially if proper x_i , i.e. the genetic effects, are chosen. As both above results generally stand irrespective of the exact distribution of the phenotype, we would prefer to deal with the Case-Control samples under an SLR, i.e. the Armitage's trend test or the allelic analysis. Meanwhile, we also argued that the Armitage's trend test integrates more information than the allelic analysis as both the trend and HWE information have been removed during the procedure of performing the later test. With such an understanding, we will then focus on the modification of Armitage's trend test in the next chapter, as it is probably the best choice among the existing methods of association study for Case-Control samples.

CHAPTER II-3

A MODIFIED ARMITAGE'S TREND TEST FOR DIFFERENT PENETRANCE MODELS AND POPULATION STRATIFICATION

3.1 Overview

As has been introduced in section II-2.2.1, if an SLR is implemented to perform the association study of a quantitative trait in a random sample, the optimal choice of x_i can both maximize the statistical power and explain the largest proportion of variance for SLR, where, however, such an optimal choice may not be available due to the unknown parameter D and q . For a Case-Control study, the optimal x_i can be similarly defined, however, unlike the situation of a quantitative trait, another factor, i.e. background genes, will also influence the optimal choice of x_i as mentioned in section I-1.2.4. In this chapter, following Armitage's denotation, x_i will be referred as the coefficient of trend or trend coefficient when we are dealing with Case-Control studies.

In a paper Sasieni published in 1997, he suggested the use of trend sets $\{1, 1, 0\}$ and $\{1, 0, 0\}$ for dominance and recessive models respectively. However, as has been shown in section II-2.2.1, even for a random sample, Sasieni's suggestion might not be the optimal choice unless certain restraints are met, which are hardly true in most situations. By knowing this, we will surely ask how far away Sasieni's suggestion is from the optimal choices. To answer this question, we will firstly theoretically calculate the appropriate choices of the trend coefficients for different penetrance models, e.g. the co-dominant model, and we will

introduce two different strategies to correct the population stratification while applying the Armitage's trend test. At last, the hence modified Armitage's trend tests will be implemented for analyzing both simulation data and a real dataset.

3.2 The Optimal Trend Set

3.2.1 Common Variants or Rare Variants

Before computing the optimal trend set, two points have to be clarified in advance. Firstly, for many diseases, even a homozygote of disease causal alleles will not ensure a 100% chance of expressing disease phenotype and such an incidence risk might increase as the increase of age. If only one of alleles at the putative QTL is responsible for the disease of interest, which is generally true for SNP markers, a disease causal allele with its penetrance coefficient of the homozygote less than 1 is practically equivalent to a virtual allele completely linked with the causal one but with a lower allele frequency and the penetrance coefficient of its homozygote equalling 1. For example, a disease causal allele γ with its allele frequency P and penetrance coefficients $0 < f_1 < 1$, $0 < f_2 < f_1$ and $f_3 = 0$ are practically equivalent to a virtual allele γ^* with its allele frequency $P^* = P\sqrt{f_1}$ and penetrance coefficient $f_1^* = 1$, $f_2^* = (f_2 - Pf_2)/(\sqrt{f_1} - Pf_1)$ and $f_3^* = 0$, because it is easy to examine that they will cause the same disease incidence rate, i.e. $P^2 f_1 + 2P(1-P)f_2$, in a random population. It is hence reasonable to assume $f_1 = 1$, $f_2 = f$ and $f_3 = 0$ without loss of generality, and f is the only questioned variable left. Secondly, 'common disease, common variant', as a widely adopted hypothesis, is the proposition behind the HapMap project. However, this concept has been challenged as only a small proportion of phenotypic variance can be explained by claimed QTLs for most common diseases (Moore et al., 2010, Cantor et al., 2010). For example, only less than 10% of Parkinson's disease is due to the known candidate genes (Lesage and Brice,

2009). Behind such a low discovery rate, it can be either common variants with small genetic risk, e.g. small penetrance coefficients, or rare variants with high genetic risk, both of which are hard to detect through a statistical test because of their small contribution to case individuals. As from the first part of this discussion, both above situations would be statistically attributed to rare variants with an extremely high incidence risk, i.e. the penetrance coefficient of homozygote always equals 1. With a similar consideration, Yang et al. (2010) found that the contribution from rare variants can explain most of the remaining heritability that is missing from the current GWASs for human height. Meanwhile, the small contribution from a single candidate gene implies that there might be a considerable level of genetic heterogeneity, and such an effect will be included in our following derivations.

Note here, complex traits controlled by multiple genes might be grouped into two categories based on whether all multiple genes affect each individual or not. Firstly, several genes of each individual might contribute to the phenotype of interest aggregately, and further complexities may be introduced if these genes are interactive, i.e. epistasis, or/and individuals are highly related (Abecasis et al., 2000, Yu et al., 2006); Alternatively, a single or a few genes might solely be responsible for the phenotype of interest for a group of individuals, and other groups of people may have different causal genes, i.e. the genetic heterogeneity. As suggested by Yang et al. (2010), we may reasonably believe that large parts of missing variance or heritability are due to the rare variants, which further implies that most of these rare variants might be functional individually rather than aggregately in a particular individual due to their low frequencies in the population. Nevertheless, although we can not completely rule out the possibility that epistasis might play an important role in the unexplained variance in case multiple genes are functional aggregately, because it is extremely impractical to detect every pairwise gene interaction simply using phenotype-genotype data (Cantor et al., 2010). Also such analyses of epistasis are mainly performed after candidate QTLs having been

detected. So, we will only consider the situation of genetic heterogeneity in our following analysis but excluding epistasis.

3.2.2 Calculation of the Optimal Trend Set

In section II-2.2.1, we have shown that the optimal choice of x_i for a random sample is $x_1 = E_1(y)$, $x_2 = E_2(y)$ and $x_3 = E_3(y)$, and it will be now demonstrated that it is also true for a Case-Control study. Using the same notations as in Table II-2 and denoting $k_i = \frac{n_i}{N_i}$, we can calculate that

$$\begin{aligned} Cov(x, y) &= \frac{\sum_{i=1}^3 N_i k_i (x_i - \bar{x})}{T} \\ &= \frac{\sum_{i=1}^3 N_i k_i (x_i - \bar{x})}{T} - \frac{\sum_{i=1}^3 N_i \bar{k} (x_i - \bar{x})}{T}, \\ &= Cov(x, k) \end{aligned}$$

and hence the Armitage's trend test statistic, i.e. $n\hat{\rho}_{xy}^2$, can be rewritten as

$$\begin{aligned} \chi^2 &= \left(\frac{Var(k)}{Var(y)} \right) \times \left(\frac{TCov^2(x, k)}{Var(x)Var(k)} \right) \\ &= \left(\frac{Var(k)}{Var(y)} \right) \times TCor^2(x, k) \end{aligned}$$

Since $Var(k)$ and $Var(y)$ can be calculated directly and independently, one of the optimal choices of x_i , the trend coefficients, is clearly $x_i = E(k_i)$. By noticing that $k_i = \bar{y}_i$ and therefore $E(k_i) = E_i(y)$, we have proved the claim at the beginning of this section. For practical and convenience to make comparison, we will take the equivalent transformation (II-2.5) as

$$x_1 = 1, \quad x_2 = \frac{E(k_2) - E(k_3)}{E(k_1) - E(k_3)} \quad \text{and} \quad x_3 = 0,$$

and hence x_2 is of the key interest in the following discussion.

To derive the exact form of x_2 , a genetic model for Case-Control samples has to be characterized. Assign f_i to denote the penetrance coefficient of the i th QTL genotype and g_{ij} to denote the joint distribution of the i th marker genotype and the j th QTL genotype in a random mating population from which case and control samples were collected. We can hence have the joint distribution of marker and QTL genotypes as shown in Table II-4 and the value of g_{ij} as given in Table II-5. This is an alternative expression of Table II-1, with the transformations $Q = p + D/q$ and $R = p - D/(1-q)$, where p, q are the frequencies of alleles M and A respectively, and D is the coefficient of LD with the disease causal allele A positively associated with M , i.e. $D > 0$.

Table II-4. Joint Distribution (Unnormalised) of Marker and QTL Genotypes for Case Control Samples.

f_i denotes the penetrance coefficient of i th QTL genotypes; g_{ij} denotes the joint probability of the i th marker genotype and the j th QTL genotype in a random population.

	Cases			Controls		
	MM	Mm	Mm	MM	Mm	mm
AA	$f_1 \times g_{11}$	$f_1 \times g_{21}$	$f_1 \times g_{31}$	$(1-f_1) \times g_{11}$	$(1-f_1) \times g_{21}$	$(1-f_1) \times g_{31}$
Aa	$f_2 \times g_{12}$	$f_2 \times g_{22}$	$f_2 \times g_{32}$	$(1-f_2) \times g_{12}$	$(1-f_2) \times g_{22}$	$(1-f_2) \times g_{32}$
aa	$f_3 \times g_{13}$	$f_3 \times g_{23}$	$f_3 \times g_{33}$	$(1-f_3) \times g_{13}$	$(1-f_3) \times g_{23}$	$(1-f_3) \times g_{33}$

Table II-5. Joint Distribution of Marker and QTL Genotypes in a random Population where Case and Control Samples Were Collected.

$Q = p + D/q$ and $R = p - D/(1-q)$, where p, q denotes the allele frequencies of marker allele M and QTL allele M respectively and D is the coefficient of linkage disequilibrium between the marker locus and the QTL.

	MM	Mm	mm
AA	$q^2 Q^2$	$2q^2 Q(1-Q)$	$q^2 (1-Q)^2$
Aa	$2q(1-q)QR$	$2q(1-q)(Q+R-2QR)$	$2q(1-q)(1-Q)(1-R)$
aa	$(1-q)^2 R^2$	$2(1-q)^2 R(1-R)$	$(1-q)^2 (1-R)^2$

In the presence of genetic heterogeneity, where there are multiple independent genes at different loci to cause a similar phenotype or symptom, the case sample can be divided into two parts, of size N_1 and N_2 , where the former is contributed from the putative QTL, while the later one, due to the heterogeneity, has the same distribution of marker genotypes as that of the control sample with a sample size N_3 . From the discussion in section II-3.2.1, it is reasonable to assume q , the frequency of disease causal allele, to be very small, and we can have $f_1 = 1$, $f_2 = f$ and $f_3 = 0$. Following Table II-4 and Table II-5, we can calculate

$$\begin{aligned}
f_{AA} &= g_{11} + g_{21} + g_{31} = q^2 \\
f_{Aa} &= g_{12} + g_{22} + g_{32} = 2q(1-q) \\
f_{aa} &= g_{13} + g_{23} + g_{33} = (1-q)^2 \\
f_{MM}^{case} &= (g_{11} + f \times g_{12}) / (f_{AA} + f \times f_{Aa}), \\
f_{Mm}^{case} &= (g_{21} + f \times g_{22}) / (f_{AA} + f \times f_{Aa}), \\
f_{mm}^{case} &= (g_{31} + f \times g_{32}) / (f_{AA} + f \times f_{Aa}), \\
f_{MM}^{control} &= [g_{13} + (1-f) \times g_{12}] / [f_{aa} + (1-f) \times f_{Aa}], \\
f_{Mm}^{control} &= [g_{23} + (1-f) \times g_{22}] / [f_{aa} + (1-f) \times f_{Aa}], \\
f_{mm}^{control} &= [g_{33} + (1-f) \times g_{32}] / [f_{aa} + (1-f) \times f_{Aa}], \\
k_1 &= (N_1 f_{MM}^{case} + N_2 f_{MM}^{control}) / [N_1 f_{MM}^{case} + (N_2 + N_3) f_{MM}^{control}], \\
k_2 &= (N_1 f_{Mm}^{case} + N_2 f_{Mm}^{control}) / [N_1 f_{Mm}^{case} + (N_2 + N_3) f_{Mm}^{control}], \\
k_3 &= (N_1 f_{mm}^{case} + N_2 f_{mm}^{control}) / [N_1 f_{mm}^{case} + (N_2 + N_3) f_{mm}^{control}],
\end{aligned}$$

where f_G , f_G^{case} and $f_G^{control}$ denote the frequencies of genotype G in the population, case and control samples, respectively, and k_i is the expected case proportion given the i th marker genotype. One may hence derive

$$\begin{aligned}
x_2 &= \frac{k_2 - k_3}{k_1 - k_3} \\
&= \frac{\left\{ (1-p) \left[\frac{(N_2 + N_3)(2f(1-q) + q)R(2(1-f)qQ + (1-q)R) +}{N_1(2(1-f)q + (1-q))Q(2f(1-q)R + qQ)} \right] \times [q(1-f)(1-Q) + f(1-q)(1-R)] \right\}}{\left[2f((1-q)^2 R(1-R) - q^2 Q(1-Q)) + q((1-q)(Q + R - 2QR) + 2qQ(1-Q)) \right] \times \\
&\quad \left[2f^2(N_1 + N_2 + N_3)(1-q)q(2QR - Q - R) + \right. \\
&\quad \left. q(N_1(1+q)(1-Q)Q + (N_2 + N_3)(q(Q + R - 2QR) + (1-q)R(1-R))) + \right. \\
&\quad \left. f(N_1 Q((Q + R - 2QR)(1-q^2) - 2q^2 Q(1-Q)) + (N_2 + N_3)((Q + R - 2QR)(1 - (1-q)^2 - 2q^2))) \right] \Bigg\}}
\end{aligned}$$

where

$$\text{if } f \neq 0 \text{ and } q \rightarrow 0, x_2 = \frac{(1-R)[N_1Q + (N_2 + N_3)R]}{N_1(Q + R - 2QR) + (N_2 + N_3)(2R - 2R^2)} \quad (\text{II-3.1})$$

and

$$\text{if } f = 0 \text{ and } q \rightarrow 0, x_2 = \frac{(1-Q)(1-R)[N_1Q^2 + (N_2 + N_3)R^2]}{(Q + R - 2QR)[N_1(1-Q)Q + (N_2 + N_3)(1-R)R]} \quad (\text{II-3.2})$$

We can immediately notice that the theoretical value of x_2 is free from the penetrance coefficient f , unless in the recessive model where $f = 0$. This result shows that in the Armitage's trend test, only two models need to be considered, says $f = 0$ and $f \neq 0$, and let's call them the 'Recessive Model' and the 'Non-Recessive Model' respectively. Notice that both formulae (II-3.1) and (II-3.2) automatically regroup the case individuals that are contributed from alternative genes or genetic heterogeneity, i.e. N_2 , to the control individuals, we may hence name the proportion $\frac{N_1}{N_1 + N_2 + N_3}$ as the effective case proportion denoted as r .

Compared to the original Armitage trend test, which does not assume the HWE as we have discussed in section II-2.3, we calculated formulae (II-3.1) and (II-3.2), i.e. the optimal trend coefficient x_2 , based on a population in HWE. However, an assumption of HWE does not limit the application of our optimal trend coefficient x_2 , because in the absence of confounding factors, e.g. population stratification etc., it would be quite reasonable to assume HWE in an out-bred species. Moreover, even in the presence of population stratification, each subpopulation could still be assumed as following HWE. We will deal with population stratification extensively in section II-3.3.

3.2.3 The Non-Recessive Model

Prior to any further calculations, it would be very useful to evaluate the range of x_2 in these two models and then interpret the genetic significance behind it to outline some basic pictures. Here, we would like to start with a briefly discussion of the non-recessive model first.

As q can be assumed very small, we can accept $R = p$, and by replacing Q as $p + D'(1 - p)$ where $D' = D/q(1 - p)$ is the standardised coefficient of LD as defined by formula (I-1.3), the optimal x_2 in the non-recessive model as defined by formula (II-3.1) can hence be rewritten as

$$\begin{aligned} x_2 &= \frac{p(1 - D'r) + D'r}{2p(1 - D'r) + D'r} \\ &= \frac{1}{2 - \frac{1}{\frac{p}{D'r} - p + 1}}, \end{aligned} \quad (\text{II-3.3})$$

where the use of D' rather than D is mainly because D' still has a definition even while $q \rightarrow 0$ and r is the effective case proportion as defined above. It is easy to understand that formula (II-3.3) is a monotonic increasing function of D' since both p and r fall within their domain $(0, 1)$, and the boundary of x_2 can hence be computed as $[\frac{1}{2}, \frac{p(1 - r) + r}{2p(1 - r) + r}]$, where the lower and upper bounds can be achieved by setting D' equal to 0 and 1 respectively. It can be noticed that the lower bound implies the default trend set $\{1, 0.5, 0\}$ is the optimal choice under the null hypothesis where $D' = 0$. However, this result is due to the transformation (II-2.5), and theoretically any trend sets should have asymptotically the same false inference rate when $D' = 0$. On the other hand, the upper bound is still a function of the effective case proportion r and the marker allele frequency p , both of which have their domains $(0, 1)$. We may notice that the optimal x_2 can only approach 1 while $p \rightarrow 0$ or $r \rightarrow 1$, where, however,

both situations are far from the reality in practice: Firstly, r has to be smaller than 1, as there are always control samples, i.e. N_3 cannot be null, and a considerable proportion of case individuals may come from the genetic heterogeneity, say N_2 might not be ignored; Secondly, a genetic variant would rarely be chosen as a common marker to perform GWAS if its minor allele frequency (MAF) is lower than 0.05 (Tabangin et al., 2009), which is mainly because of both the belief of ‘common disease, common variant’ and the fear of spurious association due to both genotyping and sampling errors. It hence turns out that the upper bound of x_2 will not be too much different from 0.5 as shown in Figure II-1, which illustrates the relationship between the upper bound of the optimal x_2 (vertical axis) and the allele frequency p (horizontal axis) from formula (II-3.3) with the effective case proportion r equalling to 0.05 (blue), 0.1 (green) and 0.2 (pink) and the equal sample sizes in case and control.

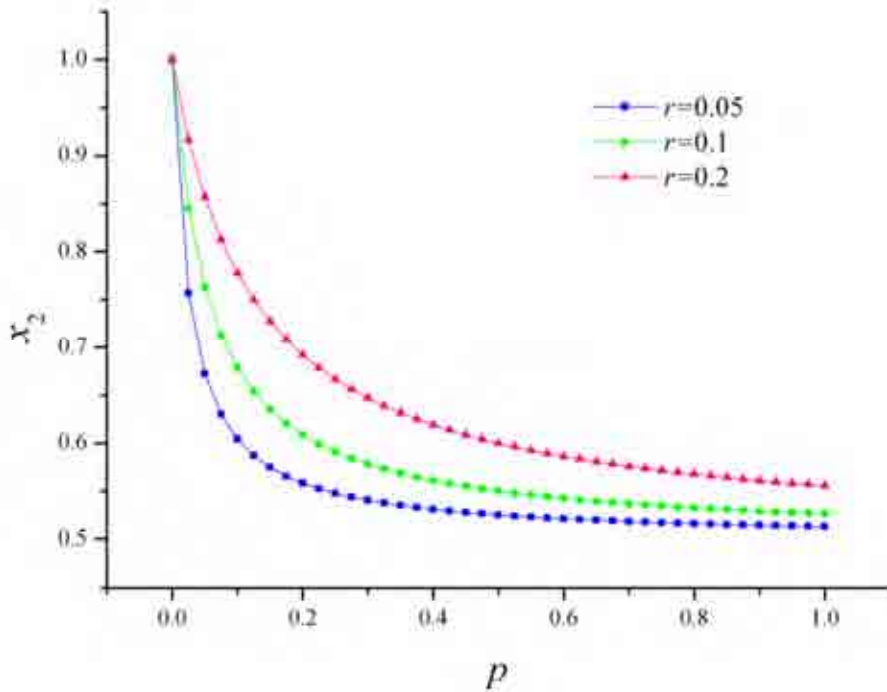


Figure II-1. Figure II-1 presents the relationship between the upper bound of optimal x_2 (vertical axis) and marker allele frequency p (horizontal axis) under a non-recessive model, where the blue, green and pink curves are drawn with r equalling to 0.2, 0.1 and 0.05 respectively, which are equivalent to heterogeneity level 60%, 80% and 90% given the equal sample sizes in case and control.

Since the correlation $Cor^2(x, k)$ as indicated above is the key factor that affects the statistical power as mentioned above, the less it will deviate from 1, the better the corresponding trend set will be. The theoretical performance of default trend set $\{1, 0.5, 0\}$ under the non-recessive model are hence illustrated in Figure II-2 through calculating $Cor^2(x, k)$ as the increase of D' , where k are assigned with the value of an optimal trend set as such an assignment is equivalent to assign $k_i = E_i(y)$; the case and control sample sizes are assumed to be equal and r is assumed to be 0.05, both of which together indicate that the putative QTL of interest explains 10% of the whole case sample; p is assigned to be 0.05 (blue), 0.1 (green), 0.2 (pink), 0.5 (yellow). It can be noticed from Figure II-2 that even with the largest loss of statistical power, i.e. $p = 0.05$, $D' = 1$ and the corresponding optimal x_2 should be around 0.65 as indicated in Figure II-1, the value of $Cor^2(x, k)$ only reduces by less than 1% of its maximum value 1. For a higher marker allele frequencies, says $p = 0.5$, such deviation is even smaller than 0.001. Since the optimal trend set is identical among any non-recessive models, we may hence conclude that the default trend set $\{1, 0.5, 0\}$ can be generally adopted for non-recessive models. Such a statement will be further evaluated through a simulation study in the following discussion.

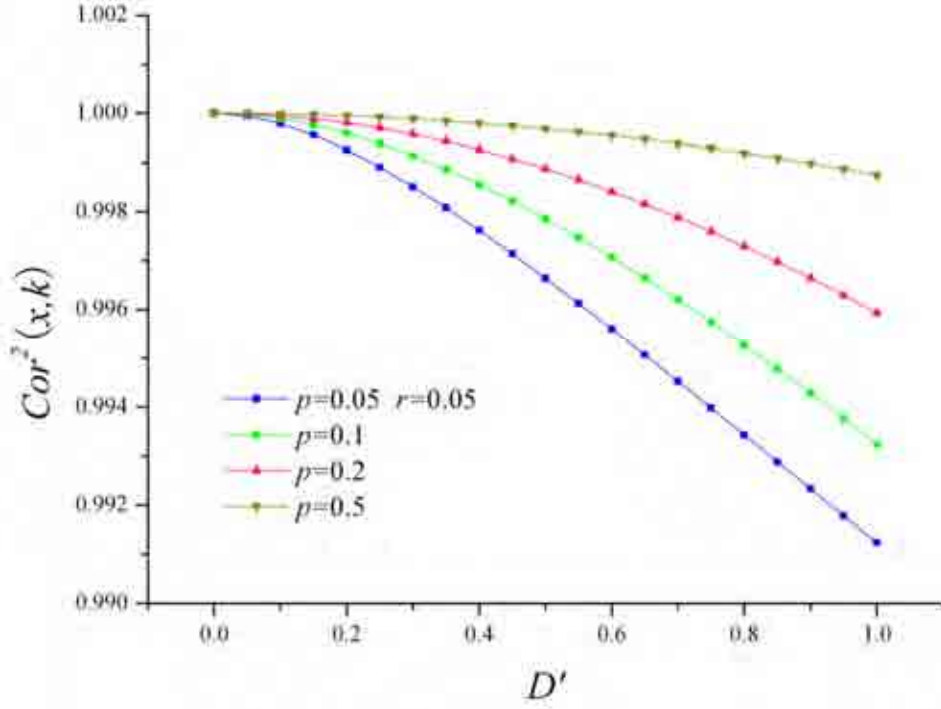


Figure II-2. In a non-recessive model, the relationship between adjusted LD D' and correlation between x and k with default trend set $\{1, 0.5, 0\}$, given marker allele frequency p and the exact case proportion r .

Consider a Case-Control sample with equal sample size in case and control, i.e. $N_1 + N_2 = N_3$, is selected from an infinite population with penetrance coefficients $f_1 = 1$, $f_2 = f$ and $f_3 = 0$, where N_1 is the sample size of case individuals contributed from the putative QTL, N_2 is the sample size of case individuals due to heterogeneities, e.g. background gene effects and environmental effects, and N_3 is the control sample size. Since it is the effective case proportion, $r = \frac{N_1}{N_1 + N_2 + N_3}$, that matters as indicated at formula (II-3.3), only the value of r will be updated in the following simulations with constant 1000 cases and 1000 controls. In our simulation studies, N_1 case individuals are generated following the joint distribution for case individuals as indicated in Table II-4, where the sample size of individuals with i th marker genotype is counted as n_{1i} ; N_2 case individuals and N_3 control individuals are generated following the joint distribution for control individuals as indicated in Table II-4 as well, where the sample sizes of individuals with i th marker genotype from such case and

control samples are counted as n_{2i} and n_{3i} respectively. Finally, we denote $n_{1i} + n_{2i} = n_{Di}$ and n_{Ci} to be the observed sample size of individuals with i th marker genotypes in case and control respectively. The parameters and the results of such simulations are both listed in Table II-6, where each test is replicated 1000 times and the number of tests which pass the 0.1% level of significance after the Bonferroni's correction is reported as the result. As both the optimal and default trend sets have been defined before, we only note here the dominant trend set takes the form $\{1, 1, 0\}$ as suggested by Sasieni (1997). Also note here, we implement the Bonferroni's correction for two reasons: Firstly, in order to compare the statistical powers, instead of comparing the average test-statistics or p-values, we can set up a restrictive threshold and measure the hits of each test that pass it. Secondly, as the Bonferroni's correction is widely used to establish a threshold that identifies a genome-wide significance, we would like to adopt it for the sake of practicality.

It can be noticed from the upper part of Table II-6, where the penetrance coefficient $f = 0.5$, i.e. the co-dominance model, although the simulation results do show certain improvement of the statistic power by using the optimal trend set, shown as column Optimal, than the default one, shown as column Default, such an improvement is quite small, and hence can be neglected. As a similar conclusion can also be acquired from the simulation results of the dominance model, where $f = 1.0$ as shown at the lower part of Table II-6, and hence we can conclude that the optimal trend set and the default trend set have very similar statistical powers while dealing with case-control samples under a non-recessive model. Such a result is coincident with our previous analyses illustrated at Figure II-2, where we have suggested the general validity and robustness of the default trend set.

Table II-6. Simulation Results of Three Trend Sets for Dominant and Additive Model

For each simulation, 1000 cases and 1000 controls were generated, and the remaining parameters are listed in the first 5 columns, where f denotes the penetrance coefficient of heterozygote at the disease locus, i.e. $f=0.5$ denotes co-dominant and $f=1.0$ denotes dominant; r denotes the effective case proportion; p and q are the frequencies of marker and disease allele which are positively associated with the adjusted linkage disequilibrium D' . Each simulation was repeated by 1000 times, and the corresponding empirical statistical powers given the 0.1% significant level after Bonferroni's correction were listed in the last three columns with trend sets 'Optimal', i.e. calculated from formula II-3.1, 'Default', i.e. $\{1, 0.5, 0\}$, and 'Dominant', i.e. $\{1, 1, 0\}$, respectively.

f	r	p	q	D'	Proportion of 1000 replicates surpassed 0.1% Significant Level after Bonferroni's Correction		
					Optimal	Default	Dominant
0.5	0.10	0.01	0.001	0.2	35.8	35.8	NA
0.5	0.10	0.01	0.001	0.5	100	100	NA
0.5	0.10	0.05	0.001	0.2	0.8	0.7	NA
0.5	0.10	0.05	0.001	0.5	82.9	81.8	NA
0.5	0.10	0.05	0.001	0.8	100	100	NA
0.5	0.10	0.10	0.001	0.2	0.1	0.1	NA
0.5	0.10	0.10	0.001	0.5	30	29.8	NA
0.5	0.10	0.10	0.001	0.8	97.1	96.8	NA
0.5	0.10	0.20	0.001	0.2	0	0	NA
0.5	0.10	0.20	0.001	0.5	3	3.1	NA
0.5	0.10	0.20	0.001	0.8	47.8	48.1	NA
0.5	0.10	0.20	0.001	1.0	87.1	86.6	NA
1.0	0.10	0.01	0.001	0.2	36.7	36.6	36.6
1.0	0.10	0.01	0.001	0.5	100	100	100
1.0	0.10	0.05	0.001	0.2	0.7	0.8	0.4
1.0	0.10	0.05	0.001	0.5	83.1	82.4	82.4
1.0	0.10	0.05	0.001	0.8	100	100	100
1.0	0.10	0.10	0.001	0.2	0.1	0.1	0.1
1.0	0.10	0.10	0.001	0.5	30.6	30.1	28.5
1.0	0.10	0.10	0.001	0.8	97.1	97.2	96
1.0	0.10	0.20	0.001	0.2	0	0	0
1.0	0.10	0.20	0.001	0.5	3.1	3.1	2.8
1.0	0.10	0.20	0.001	0.8	48.4	47.8	39.5
1.0	0.10	0.20	0.001	1.0	87.4	86.8	83.3

Meanwhile, we can also notice that the dominance trend set $\{1, 1, 0\}$, shown in the most right column 'Dominant' in Table II-6, has a generally lower statistical power than the other two. It can hence be concluded that the trend set $\{1, 1, 0\}$ is not recommended even under the dominance model, and the suggestion from Sasieni is inadequate. It can be noticed that such a difference will become more significant as p deviates away from 0, which is mainly because while p is small, i.e. $p = 0.01$, the observed individuals with marker genotype MM in both Case and Control are extremely rare or even missing due to the limited sample size, and in such a situation, there is fundamentally no difference among these three trend sets. More

precisely, if one of the observed genotypes is missing from either or both case and control samples, the Armitage's trend test is not recommended due to two reasons: Firstly, the case proportion of such a marker genotype is either 0 or undefined, and hence the trend coefficient for such a genotype is non-informative; Secondly, the missing genotype also hinders the information about the HWE, which is not necessarily true of the joint samples to valid an Armitage's trend test though. A good example of such an influence can be examined by denoting $N_1 = 0$ in formula (II-2.8), i.e. a genotype is missing from both case and control samples, and the ratio hence derived is always less than 1 as the second term in formula (II-2.8) is negative. Such a result means the Armitage's trend test hence derived will always underestimate a test statistic than the allelic analysis, and because of the very general validation of the allelic analysis, the Armitage's trend test is not favoured under such a situation.

Finally, it can be concluded that under a non-recessive model, although the default trend set $\{1, 0.5, 0\}$ is actually the optimal choice under the null hypothesis as mentioned above, it is still a very powerful trend set if not the most. Thus, under a non-recessive model, we can feel free to use the default trend set $\{1, 0.5, 0\}$ without the risk of losing the statistical power. Plus, another advantage of using the default trend set $\{1, 0.5, 0\}$ lies in the fact that there is no need to determine which marker allele is positively associated with the disease causal allele prior to launch an Armitage's trend test as the corresponding test statistic is symmetric.

3.2.4 The Recessive Model

In the previous section, we have shown that the default trend set $\{1, 0.5, 0\}$ can be generally adopted to conduct an Armitage's trend test under a non-recessive model without the risk of

losing statistical powers. However, as will be shown in this section, the choice of an adequate trend set under a recessive model is not as straightforward as that under a non-recessive model.

To understand the optimal x_2 under the recessive model as given by formula (II-3.2), we can take the same transformation $\{Q, R, N_1, N_2, N_3\} \rightarrow \{q, r, D'\}$ as launched in the previous section. However, the formula of the optimal x_2 turns out to be much more complicated than that has been derived for the non-recessive model, and it can be expressed as

$$x_2 = \frac{(1-D')\{p^2 + D'[D'(1-p) + 2p](1-p)r\}}{[D'(1-p) + 2p]\{p + D'[1-D'(1-p) - 2p]r\}} \quad (\text{II-3.4})$$

Unlike function (II-3.3), function (II-3.4) is not generally monotonic, and hence the range of the optimal x_2 is highly subject to the parameters. In order to understand how these parameters can affect the value of the optimal x_2 , we illustrate the relationship between the value of the optimal x_2 and the adjusted LD coefficient, i.e. D' , subject to various r and p as shown in Figure II-3 (a)-(c).

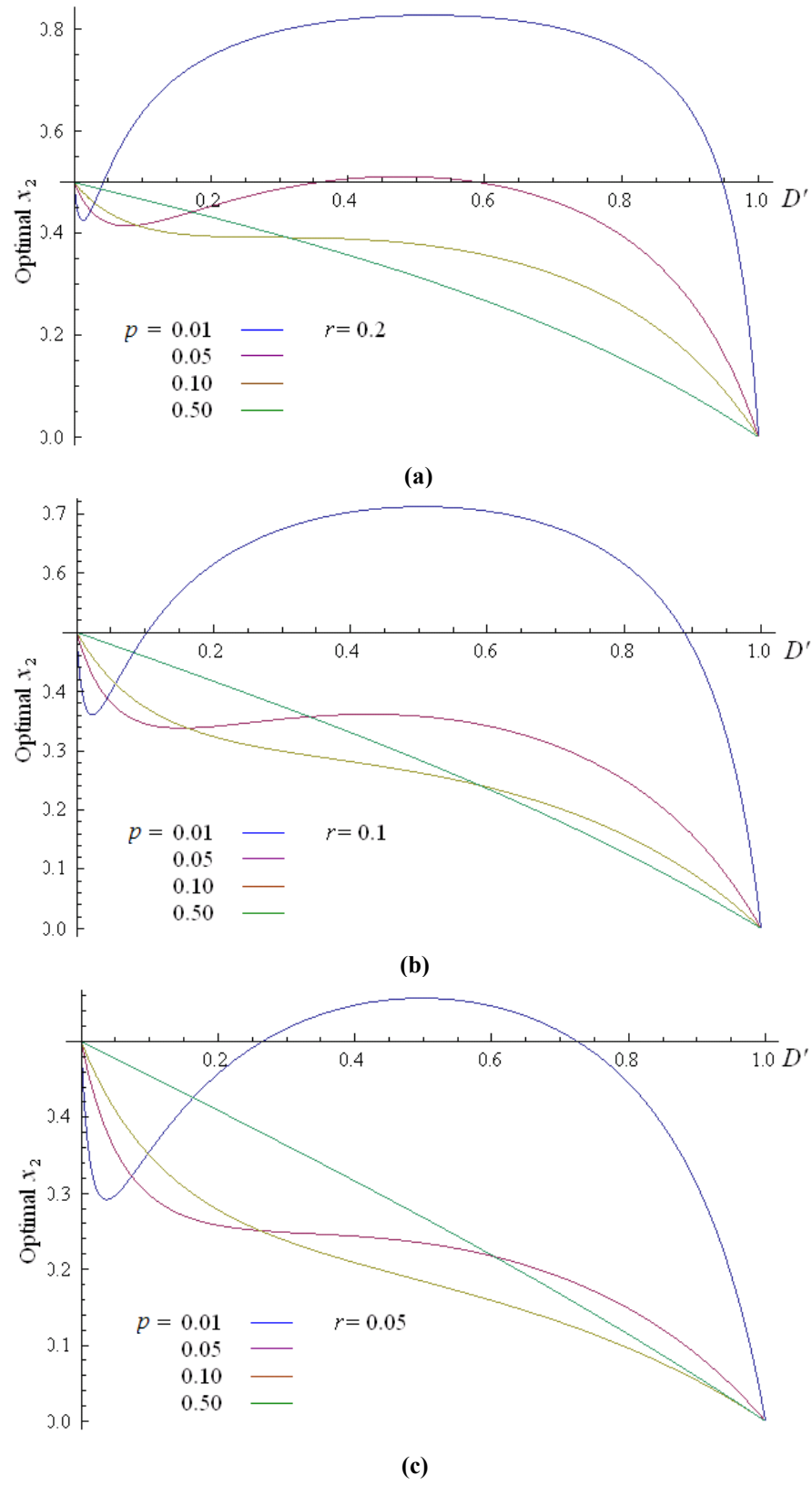


Figure II-3. The relationship between the value of the optimal x_2 and the adjusted LD coefficient D' from formula II-3.2, given the effective case proportion $r = 0.2$ (a), $r = 0.1$ (b) and $r = 0.05$ (c). For each figure, four curves are presented based on different marker allele frequencies p , i.e. 0.01 blue, 0.05 pink, 0.10 orange, 0.50 green.

From Figure II-3 (a)-(c), we can notice that the optimal x_2 is not always monotonic any more, especially while p is small, e.g. $p = 0.01$. Meanwhile, unlike under the non-recessive model, where the optimal x_2 will not be too much different from 0.5 even while $D' = 1$ unless p is sufficiently small (Figure II-1), we can notice that under the recessive model the optimal x_2 will eventually reach 0 as $D' \rightarrow 1$ irrespective of the other parameters. Such a non-monotonic property and a wide range of possible values simply hinder any attempt to find a general valid trend set for the recessive model. However, we can notice that in the condition of a high level of heterogeneity, e.g. $r \leq 0.1$, and not too small a marker allele frequency, e.g. $q \geq 0.05$, the optimal x_2 is merely monotonic within the interval $[0, 0.5]$, and by noticing that the corresponding curve of the optimal x_2 is rather flat around $x_2 = 0.25$ for $D' \in [0.2, 0.8]$, we can propose the trend set $\{1, 0.25, 0\}$ for the recessive model if no information about D' and r is available. This approximation is reasonable by noticing that one single gene normally contributes only a very small fraction of the whole case sample and the normally implemented genetic markers have an MAF over 0.01 or 0.05.

Note here, it is true that the information of D' and r might be acquired from the genotype data directly. To investigate a possible D' , we can calculate the average D' among genetic markers within the interval between two recombination hotspots, as it is reasonable to believe the putative QTL should have a much stronger association with a marker within the same recombination hotspot interval than with one outside. It is also possible to estimate the level of genetic heterogeneity if the causal gene is recessive. However, due to the concern of its accuracy, the estimate of r might not be practical for a complex disease where a single gene may only contribute a fairly small fraction of the whole case sample. Nevertheless, if we can assume the level of heterogeneity is high, the inference of D' solely is sufficient to give an adequate estimate of the optimal x_2 as $\frac{1-D'}{2}$. However, there are several concerns about

estimating D' between genetic markers: Firstly, a Case-Control sample is not a random sample, and hence the marker allele frequency estimated from such a sample might be biased. Such a bias will surely cause a more severe bias to the estimate of D' as will be shown in Chapter III-1. In that chapter, we will introduce a new strategy to estimate the LD coefficient in non-random samples and hence we can still estimate a reliable D' with the new strategy. Secondly, as such an estimate will be repeated at the level of n^2 , where n is the number of markers in the same recombination hotspot interval, even if such intervals are known, the computational burden would be quite considerable. Although both concerns can be solved as discussed above, we will show later at the end of this section that a higher D' does not guarantee a higher statistical power, so we will not implement such a strategy to calculate the optimal x_2 . On the other hand, we will also show that the trend set $\{1, 0.25, 0\}$ as proposed above performs reasonably well.

Similar to Figure II-2, we can evaluate the performance of different trend sets in comparison to the optimal one under the recessive model by calculating their corresponding $Cor^2(x, k)$ as the increase of D' , because $Cor^2(x, k)$ of the optimal trend set will always equal 1. The results of such analyses are illustrated in Figure II-4 to Figure II-6, where the corresponding trend sets are $\{1, 0.5, 0\}$, $\{1, 0.25, 0\}$ and $\{1, 0, 0\}$, respectively, and in each comparison, we always assume $r = 0.05$ and p takes value of 0.05 (Blue Square), 0.1 (Green Dot), 0.2 (Pink Upward Triangle) and 0.5 (Brown Downward Triangle). It can hence be noticed that both trend sets $\{1, 0.5, 0\}$ and $\{1, 0, 0\}$ have poor performance at certain regions, for instance, $\{1, 0.5, 0\}$ may suffer from a serious loss of statistical power if D' is large than 0.2 and $\{1, 0, 0\}$ may suffer from an even more serious loss if D' is smaller than 0.8. Such a gap, i.e. between 0.2 and 0.8, can be efficiently filled up by the trend set $\{1, 0.25, 0\}$ as illustrated at Figure II-5. We can notice that the trend set $\{1, 0.25, 0\}$ will suffer a certain loss of statistical power at two extremes as this trend set is initially designed to represent the optimal trend set within

$0.2 < D' < 0.8$ as indicated in the discussion for Figure II-3. Because the statistical power only matters in the presence of a true LD, i.e. D' is sufficiently large, the default trend set $\{1, 0.5, 0\}$ is not suitable for the recessive model, and the remaining two can be jointly applied instead of the optimal trend set, which normally might not be available.

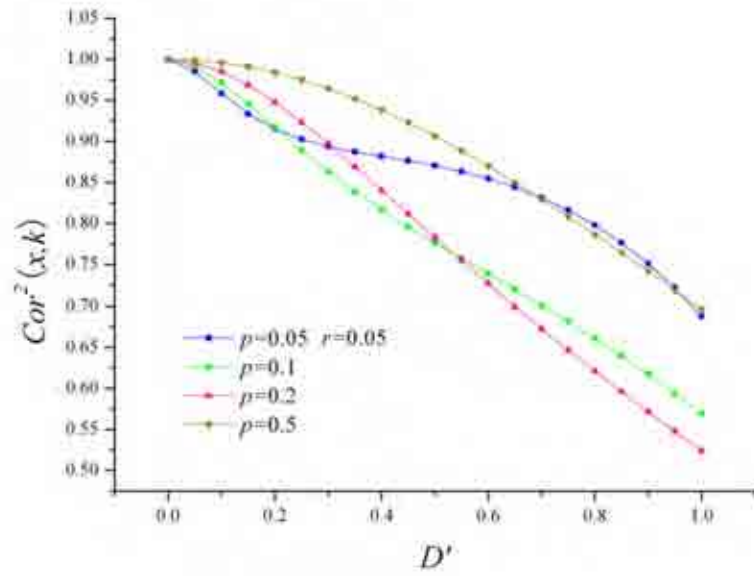


Figure II-4. Under a recessive model, the relationship between D' and $Cor^2(x, k)$ with trend set $\{1, 0.5, 0\}$, i.e. the default trend set, in terms of marker allele frequency p and the effective case proportion r .

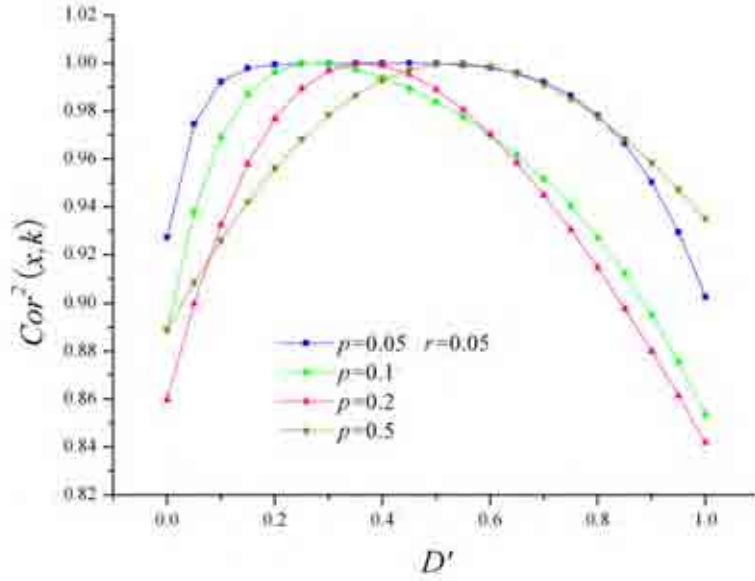


Figure II-5. Under a recessive model, the relationship between D' and $Cor^2(x, k)$ with trend set $\{1, 0.25, 0\}$, i.e. the recessive trend set, in terms of marker allele frequency p and the effective case proportion r .

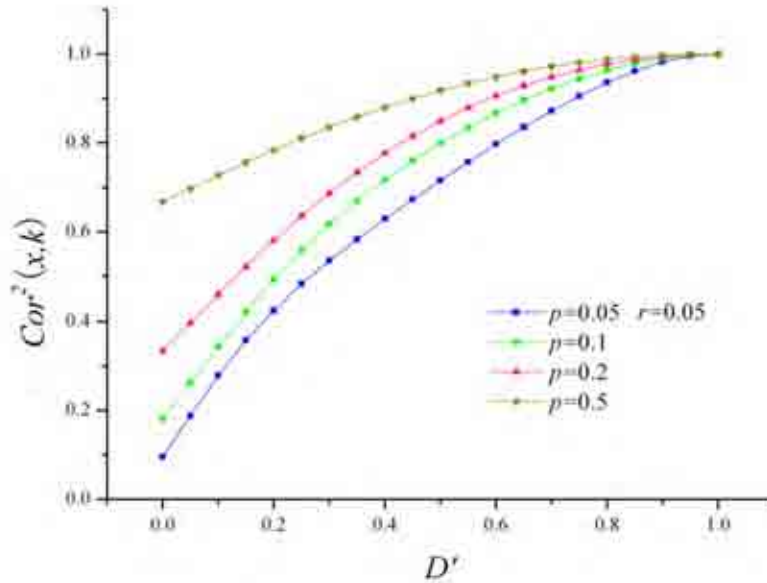


Figure II-6. Under a recessive model, the relationship between D' and $Cor^2(x, k)$ with trend set $\{1, 0, 0\}$, in terms of marker allele frequency p and the effective case proportion r .

In order to evaluate our suggestion above, simulation studies are carried out following the almost identical procedure as introduced in the previous section II-3.2.3, except the recessive model is implemented, i.e. $f = 0$. The remaining parameters and the simulation results are all listed in Table II-7 and Table II-8 subject to different r , i.e. the effective case proportion, as r

is important to determine the optimal trend set under the recessive model as indicated by Figure II-3. Because the performance of trend set $\{1, 0, 0\}$ is too poor to be mentioned unless D' is fairly close to 1, in the simulations we only include the optimal trend set, the default trend set $\{1, 0.5, 0\}$ and the recessive trend set $\{1, 0.25, 0\}$, the results of which are represented in columns Optimal, Default and Recessive respectively. In Table II-7, r equals 0.05, which means the putative QTL is responsible for 10% of the whole case individuals given the equal case and control sample size. Despite the best performance of the optimal trend set throughout, we can notice that unless D' is small, e.g. $D' < 0.2$, the recessive trend set has shown much stronger statistical power than the default one, and the advantage of the recessive trend set is amplified with increase of D' , which is exactly what we have predicted from Figure II-4 and Figure II-5. A similar result can be observed in Table II-8, where $r = 0.1$, i.e. the putative QTL is responsible for 20% case individuals, but because the overall statistical power increases tremendously due to the increase of r , the advantage of the recessive trend set over the default one is not so obvious. Compared to the optimal trend set, the recessive one can provide similar statistical powers especially for $D' = 0.5$ and $D' = 0.8$. However, it does suffer a certain loss of statistical power especially while $D' = 1$ as shown in Table II-7, but such a loss of statistical power has been alleviated with the increase of r as can be observed in Table II-8, where both the optimal and recessive trend set share similar statistical power except for $D' = 0.2$, because almost all replicates for both trend sets are significant while $D' = 1$. The loss of statistical power for the recessive trend set while D' approaches its extremes is exactly what we have predicted in Figure II-5, and hence we would again suggest the use of trend set $\{1, 0, 0\}$ as a complementary to the recessive trend set $\{1, 0.25, 0\}$ in cases when D' are extremely high. However, as the performance of trend set $\{1, 0, 0\}$ is heavily dependent on the genotype assigned with trend coefficient 1, if this particular genotype is extremely rare or even missing in case or/and control samples, this trend set should not be implemented as we have discussed in section II-2.3.

Table II-7. Simulation Results of Three Trend Sets for Recessive Model with r Equalling 0.05

For each simulation, 1000 cases and 1000 controls were generated, and the remaining parameters are listed in the first 4 columns with the penetrance coefficient f of heterozygote at the disease locus equalling 0 constantly, i.e. recessive, where r denotes the effective case proportion; p and q are the frequencies of marker and disease allele which are positively associated with the adjusted linkage disequilibrium D' . Each simulation was repeated by 1000 times, and the corresponding empirical statistical powers given the 0.1% significant level after Bonferroni's correction were listed in the last three columns with trend sets 'Optimal', i.e. calculated from formula II-3.2, 'Default', i.e. {1, 0.5, 0}, and 'Recessive', i.e. {1, 0.25, 0}, respectively.

r	p	q	D'	Proportion of 1000 replicates surpassed 0.1% Significant Level after Bonferroni's Correction		
				Optimal	Default	Recessive
0.05	0.01	0.001	0.2	45.5	45.4	36.1
0.05	0.01	0.001	0.5	100	100	100
0.05	0.01	0.001	0.8	100	100	100
0.05	0.05	0.001	0.2	1.8	1.4	1.5
0.05	0.05	0.001	0.5	95.3	85.4	95.2
0.05	0.05	0.001	0.8	100	99.8	100
0.05	0.10	0.001	0.2	0.3	0.3	0.3
0.05	0.10	0.001	0.5	64.3	42.9	63.4
0.05	0.10	0.001	0.8	100	95.3	99.9
0.05	0.20	0.001	0.2	0	0	0
0.05	0.20	0.001	0.5	13.8	7.6	13.6
0.05	0.20	0.001	0.8	91.2	56.9	87.4
0.05	0.20	0.001	1.0	100	86.2	99.5
0.05	0.35	0.001	0.2	0	0	0
0.05	0.35	0.001	0.5	1.8	1.2	1.8
0.05	0.35	0.001	0.8	29.8	13.6	28.3
0.05	0.35	0.001	1.0	80.2	35	71.2
0.05	0.50	0.001	0.2	0	0	0
0.05	0.50	0.001	0.5	0.4	0.4	0.5
0.05	0.50	0.001	0.8	5.2	3.2	4.5
0.05	0.50	0.001	1.0	22	8.6	18.4

Table II-8. Simulation Results of Three Trend Sets for Recessive Model with r Equalling 0.10

For each simulation, 1000 cases and 1000 controls were generated, and the remaining parameters are listed in the first 4 columns with the penetrance coefficient f of heterozygote at the disease locus equalling 0 constantly, i.e. recessive, where r denotes the effective case proportion; p and q are the frequencies of marker and disease allele which are positively associated with the adjusted linkage disequilibrium D' . Each simulation was repeated by 1000 times, and the corresponding empirical statistical powers given the 0.1% significant level after Bonferroni's correction were listed in the last three columns with trend sets 'Optimal', i.e. calculated from formula II-3.2, 'Default', i.e. $\{1, 0.5, 0\}$, and 'Recessive', i.e. $\{1, 0.25, 0\}$, respectively.

r	p	q	D'	Proportion of 1000 replicates surpassed 0.1% Significant Level after Bonferroni's Correction		
				Optimal	Default	Recessive
0.10	0.01	0.001	0.2	100	100	100
0.10	0.05	0.001	0.2	62.1	59.5	59.9
0.10	0.05	0.001	0.5	100	100	100
0.10	0.10	0.001	0.2	18.2	16.6	16.7
0.10	0.10	0.001	0.5	100	100	100
0.10	0.20	0.001	0.2	2.6	1.7	2.4
0.10	0.20	0.001	0.5	96.9	91.5	96.9
0.10	0.20	0.001	0.8	100	100	100
0.10	0.50	0.001	0.2	0	0.1	0
0.10	0.50	0.001	0.5	13.1	10.8	13
0.10	0.50	0.001	0.8	83.7	68.6	82.5
0.10	0.50	0.001	1.0	99.7	95	99.5

Note here, it seems quite reasonable to assume that $D' \approx 1$ if the density of markers is sufficiently high, and in such a circumstance only the trend set $\{1, 0, 0\}$ is required for the recessive model because normally we believe that only a marker closely linked to the putative QTL matters. However, although a closely linked marker can preserve a higher LD than a marker less closely linked does if no confounding effect involved, e.g. the immigration, as indicated in Table II-6, Table II-7 and Table II-8, the marker allele frequency also plays an important role in determining the test statistical power and hence we can not simply assume a marker with lower LD is less important. Take the first and last rows in Table II-7 for example, it is clear that the statistical power of a sample with $p = 0.01$ and $D' = 0.2$ is twice as that of a sample with $p = 0.50$ and $D' = 1.0$ if either an optimal trend set or a recessive trend set is implemented. With this understanding, we should always implement the recessive trend set, i.e. $\{1, 0.25, 0\}$, because the linkage disequilibrium is not the only factor that influences the association between a marker locus and a QTL.

As a brief conclusion here, the recessive trend set $\{1, 0.25, 0\}$ should be mainly used if a recessive model is under test, but the trend set $\{1, 0, 0\}$ can be implemented as a complementary analysis in case D' between the testing marker and the QTL is extremely high providing the genotype assigned with trend coefficient 1 is not too rare.

3.2.5 Regarding the False Positive

Above, we have analysed the statistical power of the Armitage's trend test with different trend sets as shown in Table II-6, Table II-7 and Table II-8. However, another issue remains; do the alternative trend sets other than the default one increase the chance of claiming a false positive? We would like to assess this question empirically. As we have mentioned before, any trend sets should have asymptotically identical false inference rates under the null hypothesis, i.e. $D = 0$, where the false negative rate is always null and hence is obviously identical among different trend sets. To evaluate the false positive rates among different trend sets, the default trend set $\{1, 0.5, 0\}$ and the recessive trend set $\{1, 0.25, 0\}$ are compared through a simulation study under the null hypothesis, where p , the frequency of marker allele M , increase from 0.01 to 0.90 with step 0.01 and the simulation for each p was replicated 1000 times under the null hypothesis with 1000 case individuals and 1000 control individuals. The corresponding simulation results are given in Table II-9, and we can notice that the difference between these two trend sets is not significant (P-value = 0.204).

Table II-9. False Positive Rates of Default and Recessive Trend Sets

The false positive rates of default $\{1, 0.5, 0\}$ and recessive $\{1, 0, 0\}$ trend sets are analysed through simulations under the null hypothesis. 1000 cases and 1000 controls are simulated for each of 1000 replicates given a particular marker allele frequency p , which increase from 0.01 to 0.90 with step 0.01. The number of claimed positive from 1000 replicates at the 0.01 significance level for each step was recorded., of which the means and standard deviations of these two trend sets are listed as below. A t -test was carried out between these two trend sets with 178 degrees of freedom.

		Trend Set		t -Test (df)
		Default	Recessive	
False positive rate per 1000 replicates at 99% confidential interval	Mean	10.438	9.854	1.275 ($df=178$)
	Standard Deviation	3.093	3.051	

Although above empirical analyses show that the alternative trend sets are unlikely to cause a higher false positive rate, we should again note here that in practice, the genotyping errors

might largely influence the test statistics from an alternative trend sets, especially the one $\{1, 0, 0\}$. This is mainly because such an alternative trend set heavily relies on one of the two homozygotes, and if one of the homozygote genotypes is comparatively rare, even a single individual being wrongly genotyped would dramatically shift the test statistics. Comparatively, the default trend set is more robust under such a situation. Of course, as we have suggested above, if some of the genotypes are missing for either case or control sample due to too low an allele frequency, the allelic analysis is recommended.

3.2.6 Conclusion

In above discussions, we have shown that only two situations need to be considered in a Case-Control study, i.e. the non-recessive model where $f \neq 0$ and the recessive model where $f = 0$. Because of their different features, the performance of adopting the optimal trend set for these two models are quite different. In a non-recessive model, because the optimal x_2 will not significantly differ from 0.5 due to the high heterogeneity level in practice, the improved statistical power from the use of optimal trend set is very limited, and hence the default trend set $\{1, 0.5, 0\}$ can be a both conventional and appropriate choice instead of the optimal one. However, in a recessive model, because the optimal x_2 highly depends on the parameters, i.e. r and D' , the performance of the default trend can be much worse than an alternative one, e.g. the optimal trend set and the recessive trend set. As the optimal trend set under such a situation might not be available, it is hence recommended to adopt the recessive trend set $\{1, 0.25, 0\}$ instead and the trend set $\{1, 0, 0\}$ can be implemented as a complementary as the simple recessive trend set might suffer a certain loss of statistical power if the testing marker is in extremely strong association with a QTL.

As the choice of an optimal trend set has become clear, in the next stage, the information of population stratification will be combined into the Armitage's trend test in order to analyse jointly datasets from different resources.

3.3 Correction for Population Stratification in the Armitage's Trend Test

As has been discussed in section I-1.1 and II-3.2.1, the low discovery rate of causal genes urges the need to increase the statistical power of GWAS. Other than the improvement of statistical methods, the increase of sample size would be a more direct and convenient way given the explosion of available GWAS data in the past few years (Ku et al., 2010). To integrate results from different GWAS, meta-analysis is probably the most common way to analyse and integrate information from those previous results. However, as an 'analysis of analyses', meta-analysis will surely lose some information because much of the genetic information is not included in a result output. If the original genotype-phenotype information of all integrated data is available, the implementation of an association study on these integrated data would be an optimal choice, which would definitely lead to the increase of the corresponding statistical powers if genuine genetic association does exist in many of those resources. However, to integrate data from different resources will almost surely encounter the issue of population stratification, because many of these data are collected from different genetic cohorts. Although most of these cohorts are European oriented in the major reference dataset, we can not exclude the possibility that minor differences might exist for certain regions in the genome, and as we will show later, the influence from even a minor difference in marker allele frequency can be dramatically amplified by the differences between case-control ratios from different sources. In order to cope with the issue of population stratification raised by the combination of data from multiple cohorts, two approaches, i.e.

Method I (Dummy Variables) and Method II (Non-central χ^2 test), are introduced in the following sections and evaluated by both simulation studies and real data analyses.

3.3.1 Method I: Dummy Variables

To correct the effect of population stratification in a linear model, our intuitional idea would be the introduction of an extra parameter vector λ , a dummy variable, to adjust this substructure effect. Suppose there are n individuals sampled from m subpopulations, we can write a linear model in the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\lambda$, where \mathbf{y} is a $n \times 1$ column vector of phenotypes with the i th phenotype being denoted as y_i ; \mathbf{X} is a $n \times 1$ column vector of trend coefficients with its i th element denoted as x_i ; \mathbf{Z} is a $n \times m$ matrix with its element z_{ij} equalling 1 only if the corresponding i th individual belongs to the j th subpopulation or otherwise 0; β is the regression coefficient for the trend coefficients and λ is an $m \times 1$ column vector with its j th element representing the substructure effect of the j th subpopulation with $i \leq n$, $j \leq m$ and $i, j \in \mathbb{Z}^+$.

The formula $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\lambda$ can be alternatively expressed as

$$\mathbf{y} = (\mathbf{X} \mid \mathbf{Z}) \begin{pmatrix} \beta \\ \lambda \end{pmatrix}. \quad (\text{II-3.5})$$

We can hence easily compute the estimator of $\begin{pmatrix} \beta \\ \lambda \end{pmatrix}$ as $\left(\begin{array}{c|c} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \hline \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} \end{array} \right)^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{y}$ from formula

(I-1.18). The non-singularity of $\left(\begin{array}{c|c} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \hline \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} \end{array} \right)$ can be understood by noticing that the

columns of \mathbf{X} and \mathbf{Z} are linearly independent with each other, which is generally true as the columns of \mathbf{Z} are obviously linearly independent and \mathbf{X} can not be represented by any a linear combination of such columns unless rare occasions happens, e.g. one of the marker genotypes

are missing from one subpopulation and the other two genotypes share the identical trend coefficients, and hence matrices $(\mathbf{X} \mid \mathbf{Z})$ and $\begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix}(\mathbf{X} \mid \mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} \end{pmatrix}$ are all of full-rank of $m+1$. Since $\mathbf{X}^T \mathbf{X}$ is a real number and $\mathbf{Z}^T \mathbf{Z}$ is a non-singular diagonal matrix, we can work out

$$\begin{pmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1} \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \\ -(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X} \mathbf{A}^{-1} & (\mathbf{Z}^T \mathbf{Z})^{-1} + (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \end{pmatrix},$$

where $\mathbf{A} = \mathbf{X}^T \mathbf{X} - \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X} = \sum_{j=1}^m n_j \text{Var}_j(x)$ is a real number. The estimator of β can

hence be calculated as:

$$\begin{aligned} \hat{\beta} &= \frac{\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}}{\mathbf{X}^T \mathbf{X} - \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X}} \\ &= \frac{\sum_{j=1}^m n_j \text{Cov}_j(x, y)}{\sum_{j=1}^m n_j \text{Var}_j(x)}, \end{aligned} \quad (\text{II-3.6})$$

where $\text{Cov}_j(x, y)$ denotes the covariance between trend coefficients and phenotypes of the j th subpopulation and $\text{Var}_j(x)$ denotes the variance of trend coefficients of the j th subpopulation.

It is easy to examine that under the null hypothesis, where $\text{Cov}_j(y, x) = 0$ for any $j \leq m$ and $j \in Z^+$, estimator (II-3.6) gives $\hat{\beta}_0 = 0$ and hence $\beta_0 = E(\hat{\beta}_0) = 0$, where $\hat{\beta}_0$ is the estimator of β under the null hypothesis. In order to establish a statistical test for $\hat{\beta}$, we now proceed to calculate the variance of $\hat{\beta}$ under the null hypothesis.

Since the variance-covariance matrix of regression coefficient $\begin{pmatrix} \beta \\ \lambda \end{pmatrix}$ can be estimated as

$\hat{\sigma}^2 \begin{pmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} \end{pmatrix}^{-1}$ as introduced at formula (I-1.19), where $\hat{\sigma}^2$ is the unbiased estimator of

variance of residual term which can be given following (Searle, 1971)

$$\hat{\sigma}^2 = \frac{1}{n-m-1} (\mathbf{y} - (\mathbf{X} \mid \mathbf{Z}) \begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix})^T (\mathbf{y} - (\mathbf{X} \mid \mathbf{Z}) \begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix}). \quad (\text{II-3.7})$$

By noticing that

$$\begin{aligned} \mathbf{y} - (\mathbf{X} \mid \mathbf{Z}) \begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix} &= \mathbf{y} - (\mathbf{X} \mid \mathbf{Z}) \left(\begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} (\mathbf{X} \mid \mathbf{Z}) \right)^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{y}, \\ &= \mathbf{M} \mathbf{y} \end{aligned}$$

where $\mathbf{M} = \mathbf{I} - (\mathbf{X} \mid \mathbf{Z}) \left(\begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} (\mathbf{X} \mid \mathbf{Z}) \right)^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix}$ is an idempotent matrix with the property

$\mathbf{M}\mathbf{M} = \mathbf{M}$ and $\mathbf{M}^T = \mathbf{M}$, formula (II-3.7) can hence be alternatively written as

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-m-1} \mathbf{y}^T \mathbf{M} \mathbf{y} \\ &= \frac{1}{n-m-1} \mathbf{y}^T \begin{pmatrix} \mathbf{I} - \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T + \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T + \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \\ -\mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T - \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \end{pmatrix} \mathbf{y} \\ &= \frac{1}{n-m-1} \left(\mathbf{y}^T (\mathbf{I} - \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T) \mathbf{y} - \mathbf{A}^{-1} \left(\mathbf{X}^T (\mathbf{I} - \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T) \mathbf{y} \right)^T \left(\mathbf{X}^T (\mathbf{I} - \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T) \mathbf{y} \right) \right) \\ &= \frac{1}{n-m-1} \sum_{j=1}^m n_j \text{Var}_j(y) \left(1 - \frac{\left(\sum_{j=1}^m n_j \text{Cov}_j(x, y) \right)^2}{\sum_{j=1}^m n_j \text{Var}_j(y) \times \sum_{j=1}^m n_j \text{Var}_j(x)} \right) \end{aligned}$$

Given the expression of $\hat{\sigma}^2$ above, under the null hypothesis, where $\text{Cov}_j(x, y) = 0 \forall j$, the variance of β_0 , $\text{Var}(\beta_0)$, can hence be estimated as

$$\begin{aligned}
Var(\hat{\beta}_0) &= \frac{\hat{\sigma}_0^2}{\mathbf{X}^T \mathbf{X} - \mathbf{X}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X}} \\
&= \frac{1}{n-m-1} \frac{\sum_{j=1}^m n_j Var_j(y)}{\sum_{j=1}^m n_j Var_j(x)} \quad , \tag{II-3.8}
\end{aligned}$$

where $\hat{\sigma}_0^2$ is an unbiased estimator of the residual variance under the null hypothesis.

After deriving $\hat{\beta}$ and $Var(\hat{\beta}_0)$, we can perform a chi-square test against the null hypothesis that there is no association between the testing marker and the disease as

$$\begin{aligned}
\chi_1^2 &= \frac{(\hat{\beta} - \hat{\beta}_0)^2}{Var(\hat{\beta}_0)} \\
&= (n-m-1) \times \frac{\left[\sum_{j=1}^m n_j Cov_j(x, y) \right]^2}{\sum_{j=1}^m n_j Var_j(y) \times \sum_{j=1}^m n_j Var_j(x)} \tag{II-3.9}
\end{aligned}$$

with 1 degree of freedom. The normality of $\hat{\beta}$ can be understood through a similar discussion given in section II-1.3.1.

Note here, normally, while performing a linear regression with dummy variables, it would be recommended to perform a test of whether the coefficients of dummy variables, i.e. λ , are significantly different from each other. However, in our situation, under the null hypothesis, λ represent the ratios of case individuals in each subpopulation, and inequalities among case ratios can be observed prior to a test, so we may feel comfortable to adopt formula (II-3.5). What's more, as will be shown in the next section II-3.3.2, even when the ratios of case individuals in each subpopulation are identical, it is still reasonable to keep the dummy variables in order to correct the overestimate of $Var(\hat{\beta}_0)$ introduced by population stratifications.

3.3.2 Method II: Non-Central χ^2 Test

Other than correcting population stratification with dummy variables, i.e. Method I, we can treat the influence from population stratification as an impact on the β_0 , i.e. the expected regression coefficient β of the SLR under the null hypothesis, which will deviate from 0 due to the effect. Our aim in this section is to evaluate and remove such a deviation. At the same time, we will discuss how to properly estimate the variance of β in the presence of population stratification.

In order to evaluate the deviation of β from 0 under the null hypothesis, suppose there are m subpopulations with n_i individuals each, where in each subpopulation the proportion of case individuals are denoted as θ_i , and the testing marker allele frequency of each subpopulation is denoted by p_i . Notice that under the null hypothesis, the case and control samples from the same subpopulation should share the same distribution, and if the HWE is assumed we can have the following results theoretically

$$N'_1 = \sum_{i=1}^m n_i p_i^2, \quad N'_2 = \sum_{i=1}^m 2n_i p_i (1 - p_i), \quad N'_3 = \sum_{i=1}^m n_i (1 - p_i)^2$$

and

$$k'_1 = \frac{\sum_{i=1}^m n_i \theta_i p_i^2}{\sum_{i=1}^m n_i p_i^2}, \quad k'_2 = \frac{\sum_{i=1}^m n_i \theta_i p_i (1 - p_i)}{\sum_{i=1}^m n_i p_i (1 - p_i)}, \quad k'_3 = \frac{\sum_{i=1}^m n_i \theta_i (1 - p_i)^2}{\sum_{i=1}^m n_i (1 - p_i)^2},$$

where N'_1 , N'_2 and N'_3 represent the number of individuals of three genotypes in the combined population respectively, and k'_1 , k'_2 and k'_3 represent the corresponding case proportion of each genotype.

Let the trend set take the form $x_1 = 1$, $x_2 = t$ and $x_3 = 0$, and we may hence calculate the variance of x_i under the null hypothesis as:

$$\begin{aligned} Var'(x) &= E(x^2) - E^2(x) \\ &= \frac{N'_1 + N'_2 t^2}{n} - \left(\frac{N'_1 + N'_2 t}{n} \right)^2 \\ &= \frac{n(N'_1 + N'_2 t^2) - (N'_1 + N'_2 t)^2}{n^2} \end{aligned}$$

Note that, as it is easy to see that $E(N_i) = N'_i$ under the null hypothesis and HWE, where N_i denotes the observed number of individuals with the i th genotypes, we can approximately have $Var'(x) \approx Var(x)$. On the other hand, it is clear that $Var'(y) = Var(y)$ as $Var(y)$ is only related to the case proportion in the combined Case-Control sample. As $\beta_0 = \frac{Cov'(x, y)}{Var'(x)}$, where $Var'(x)$ can be approximately given as above, what matters now is the numerator $Cov'(x, y)$. As has been proved at section II-3.2.2 that $Cov'(x, y) = Cov'(x, k)$, the covariance between x and y under the null hypothesis can be theoretically calculated as

$$\begin{aligned} Cov'(x, y) &= Cov'(x, k') \\ &= \frac{\sum_{j=1}^3 N'_j k'_j (x_j - \bar{x})}{n} \\ &= \frac{\left[\left(\sum_{i=1}^m n_i \theta_i p_i^2 \right) \left(1 - \frac{N'_1 + N'_2 t}{n} \right) + \left(\sum_{i=1}^m n_i \theta_i 2 p_i (1 - p_i) \right) \left(t - \frac{N'_1 + N'_2 t}{n} \right) \right. \\ &\quad \left. + \left(\sum_{i=1}^m n_i \theta_i (1 - p_i)^2 \right) \left(0 - \frac{N'_1 + N'_2 t}{n} \right) \right]}{n} \\ &= \frac{\left[\left(\sum_{i=1}^m n_i \right) \left(\sum_{i=1}^m n_i \theta_i p_i^2 \right) + t \left(\sum_{i=1}^m n_i \right) \left(\sum_{i=1}^m n_i \theta_i 2 p_i (1 - p_i) \right) \right. \\ &\quad \left. - \left(\sum_{i=1}^m n_i p_i^2 + t \sum_{i=1}^m n_i 2 p_i (1 - p_i) \right) \left(\sum_{i=1}^m n_i \theta_i \right) \right]}{n^2} \\ &= \frac{\sum_{i=1}^m \sum_{i' < i \leq m} n_i n_{i'} (\theta_i - \theta_{i'}) (p_i - p_{i'}) (2t + (p_i + p_{i'}) (1 - 2t))}{n^2} \end{aligned}$$

For a real sample, as n , n_i and θ_i are observed and $E(\hat{p}_i) = p_i$, we can give β_0 approximately as

$$\beta_0 \approx \frac{\sum_{i=1}^{m-1} \sum_{i'=i+1}^m n_i n_{i'} (\theta_i - \theta_{i'}) (\tilde{p}_i - \tilde{p}_{i'}) (2t + (\tilde{p}_i + \tilde{p}_{i'}) (1 - 2t))}{n(N_1 + N_2 t^2) - (N_1 + N_2 t)^2}, \quad (\text{II-3.10})$$

where \hat{p}_i and \tilde{p}_i denote the estimator and estimate of the testing marker allele frequency of samples selected from the i th subpopulation under the null hypothesis, the later of which can be directly given as $\frac{2N_{i1} + N_{i2}}{n_i}$, where N_{ij} denotes the observed number of individuals with the j th marker genotype from the i th subpopulation. It is clear that β_0 calculated as equation (II-3.10) is not generally equal to 0.

If $\beta_0 \neq 0$, we can have $E_j(y) = \alpha_0 + x_j \beta_0$ under the null hypothesis, where $E_j(y)$ denotes the expectation of a phenotype with the j th marker genotype, and we may hence claim that $\bar{y}_j - \alpha_0 - x_j \beta_0$ asymptotically follows a normal distribution with mean 0, where \bar{y}_j denotes the sample mean of phenotypes with the j th marker genotype. From such a result, we can derive

$$\begin{aligned} \hat{\beta} &= \frac{\hat{\sigma}_{xy}^2}{\hat{\sigma}_x^2} = \frac{1}{\hat{\sigma}_x^2} \times \frac{\sum_{j=1}^3 N_j (x_j - \bar{x})(\bar{y}_j - \bar{y})}{\sum_{j=1}^3 N_j - 1} \\ &= \frac{1}{\hat{\sigma}_x^2} \times \left[\frac{\sum_{j=1}^3 N_j (x_j - \bar{x})(\bar{y}_j - \tilde{\alpha}_0 - x_j \tilde{\beta}_0)}{\sum_{j=1}^3 N_j - 1} + \frac{\sum_{j=1}^3 N_j (x_j - \bar{x})^2}{\sum_{j=1}^3 N_j - 1} \tilde{\beta}_0 \right] \\ &= \frac{\sum_{j=1}^3 N_j (x_j - \bar{x})(\bar{y}_j - \tilde{\alpha}_0 - x_j \tilde{\beta}_0)}{\hat{\sigma}_x^2 \left(\sum_{j=1}^3 N_j - 1 \right)} + \tilde{\beta}_0 \end{aligned}$$

As the first term of the above formula asymptotically follows a normal distribution with mean 0, we may claim that $\hat{\beta}$ asymptotically follows a normal distribution as well, with mean $\tilde{\beta}_0$

though. Therefore we have $\frac{\hat{\beta} - \tilde{\beta}_0}{\sqrt{Var(\hat{\beta})}} \sim N[0,1]$ or $\frac{(\hat{\beta} - \tilde{\beta}_0)^2}{Var(\hat{\beta})} \sim \chi^2_{df=1}$ when the samples size is

sufficiently large that the central limit theorem could be applied.

With above derivations, our next aim is to properly estimate the variance of $\hat{\beta}$. Actually, such a topic has already been considered by Devlin & Roeder in their famous paper ‘Genomic Control for Association Studies’ published in 1999, and we will follow a similar procedure to their derivation. However, we will also show that Devlin & Roeder’s suggestion to use an inflation factor to reduce the false positive rate in the presence of population stratification may not be valid because such an inflation of variance will only reduce the test statistic of the corresponding Chi-square test rather than increase it.

Following Devlin & Roeder’s idea, we adopt Wright’s coefficient of inbreeding F , to generally represent any influence that results in the reduction of heterozygotes. Because population stratification would decrease the proportion of heterozygotes in the combined population, and we can hence consider equations

$$\Pr(MM) = \frac{N_1}{T} = Fp + (1-F)p^2,$$

$$\Pr(Mm) = \frac{N_2}{T} = 2(1-F)p(1-p)$$

and

$$\Pr(mm) = \frac{N_3}{T} = F(1-p) + (1-F)(1-p)^2.$$

Denoting the solutions of F for above equations are F_1 , F_2 and F_3 , respectively, and by noticing that $F_1 + F_3 = 2F_2$, we can accept F_2 as an appropriate estimate of F ,

thus $\tilde{F} = 1 - \frac{N_2}{2p(1-p)T}$. Note here, such an estimate would also minimize the sum of mean square errors for above three equations, i.e. the estimate of linear least squares.

In the following, we will evaluate the influence of F on the variance of $\hat{\beta}$. For convenience, only the non-dominance situation will be considered, where equally spaced trend sets are implemented. For instance, let G denote the number of allele M for a single individual, we can calculate $E(G) = 2p$ and $Var(G) = 2(1+F)p(1-p)$. For reference, we would like to cite the Armitage's trend test here as given at formula (II-1.22)

$$\chi_G^2 = \frac{T[T(2r_1 + r_2) - R(2N_1 + N_2)]^2}{R(T-R)[T(4N_1 + N_2) - (2N_1 + N_2)^2]} \sim \chi_{df=1}^2,$$

where r_i and s_i represent the sample size of i th genotype in case and control respectively; R and S are the sample size of case and control respectively; t is the trend coefficient from trend set $\{1, t, 0\}$. Recall that a $\chi_{df=1}^2$ distribution is formed by $\frac{\varphi^2}{Var(\varphi)}$, where $\varphi \sim N(\varphi_0, V(\varphi))$, as introduced in section I-1.4.2, and we can hence rewrite formula (II-1.22) into

$$\chi_G^2 = \frac{K^2}{V(K)} \sim \chi_{df=1}^2,$$

where $K = S \sum_{i=1}^R G_i - R \sum_{j=1}^S H_j$; $V(K) = \frac{t(T-t)}{T} [T(4N_1 + N_2) - (2N_1 + N_2)^2]$ is the variance of K calculated from formula (II-1.22); G_i and H_j denote the number of marker M of the i th case individual and the j th control individual respectively. Note here, although such a chi-square test given by formula (II-1.22) is general non-central as $E(K) \neq 0$ under the null hypothesis in the presence of population stratification as indicated by (II-3.10), the derivation of $Var(K)$ will not be influenced. From Devlin & Roeder, we can calculate the variance of K as

$$\begin{aligned}
Var(K) &= Var\left(S \sum_{i=1}^R G_i - R \sum_{j=1}^S H_j\right) \\
&= S^2 \sum_{i=1}^R Var(G_i) + R^2 \sum_{j=1}^S Var(H_j) + 2S^2 \sum_{i < l} Cov(G_i, G_l) \\
&\quad + 2R^2 \sum_{j < l} Cov(H_j, H_l) - 2SR \sum_{i=1}^R \sum_{j=1}^S Cov(G_i, H_j)
\end{aligned} \tag{II-3.11}$$

In their paper, Devlin & Roeder argued that $Cov(G_i, G_l)$, $Cov(H_j, H_l)$ and $Cov(G_i, H_j)$ equal $4Fp(1-p)$ if they are from the same subpopulation. However, recalling that F is defined in the combined population, we cannot simply assume F is identical throughout all subpopulations. More precisely, under the null hypothesis, where case and control samples of any a subpopulation are generated with an identical distribution for marker genotypes, the population stratification is the only source of such a generalised F if each subpopulation is ideal, e.g. without inbreeding and under HWE, and hence it is not only unnecessary but also a mistake to calculate the variance as suggested by Devlin & Roeder. The correct way to calculate formula (II-3.11) is to accept $Cov(G_i, G_l) = Cov(H_j, H_l) = Cov(G_i, H_j) = 4Fp(1-p)$ in the full sample, and hence we can have

$$\begin{aligned}
Var(K) &= 2SRTp(1-p)(1+F) \\
&\quad + 4Fp(1-p) \left[S^2 R(R-1) + R^2 S(S-1) - 2R^2 S^2 \right] \\
&= 2SRTp(1-p)(1-F)
\end{aligned} \tag{II-3.12}$$

The above variance of K is the proper estimate which counts in the influence of F introduced by population stratification. Recall the alternative variance of K , i.e. $V(K)$, calculated by formula (II-1.22), we can have

$$\begin{aligned}
V(K) &= \frac{t(T-t)}{T} \left[T(4N_1 + N_2) - (2N_1 + N_2)^2 \right] \\
&= \frac{RS}{T} (4N_1N_3 + N_2N_3 + N_1N_2) \\
&= \frac{RS}{T} [(2N_1 + N_2)(2N_3 + N_2) - N_2T] \quad . \quad (II-3.13) \\
&= RST \left[4p(1-p) - 2(1-\tilde{F})p(1-p) \right] \\
&= 2RSTp(1-p)(1+\tilde{F})
\end{aligned}$$

Comparing formulae (II-3.12) and (II-3.13), we can notice that the inflation of variance caused by population stratification is fixed as $\lambda = \frac{1-\tilde{F}}{1+\tilde{F}}$ when a default trend set is implemented, and hence such an inflation factor will always reduce the test statistic of an Armitage's trend test, which will also be shown in the simulation studies.

By noticing that under the null hypothesis and theoretically

$$\begin{aligned}
K_0 &= 2S \sum_i^m n_i \theta_i p_i - 2R \sum_i^m n_i (1-\theta_i) p_i \\
&= 2T \sum_i^m n_i p_i \left(\theta_i - \frac{R}{T} \right) ,
\end{aligned}$$

where clearly $K_0 = 0$ if $p_i = p \forall i$ or $\theta_i = \frac{R}{T} \forall i$, we can establish a corresponding Chi-square test as

$$\frac{(K - K_0)^2}{Var(K)} = \frac{4 \left[SR(p_{Case} - p_{Control}) - 2T \sum_i^m R_i (p_i - p) \right]^2}{SRN_2} \sim \chi_{df=1}^2, \quad (II-3.14)$$

where n_i , R_i and S_i are the full, case and control sample size collected from the i th subpopulation, respectively; N_{ij} is the corresponding sample size for the j th marker genotype in the i th subpopulation; p_i is the allele frequency of marker M in the i th subpopulation and can be estimated as $\frac{2N_{i1} + N_{i2}}{2n_i}$; $p = \frac{2N_1 + N_2}{2T}$, $p_{case} = \frac{2r_1 + r_2}{2R}$ and $p_{control} = \frac{2s_1 + s_2}{2S}$. Formula

(II-3.14) is hence the modified Armitage's trend test for the non-recessive model in the

presence of population stratification. However, as above derivations are not always valid for a trend set other than the default one, we can simply assume the inflation factor $\lambda = \frac{1-F}{1+F}$ is generally valid for any trend set, and hence we can write the general test statistic of our Method II as

$$\begin{aligned}\chi_{II}^2 &= T \frac{[Cov(x, y) - Cov'(x, y)]^2}{\lambda Var(x) Var(y)} \\ &= T \frac{[S(r_1 + r_2 t) - R(s_1 + s_2 t) - \eta]^2}{\lambda R(T - R)[T(N_1 + N_2 t^2) - (N_1 + N_2 t)^2]}\end{aligned}\quad (II-3.15)$$

which follows the χ^2 with the degree of freedom 1, where

$$\eta = \frac{\sum_{i=1}^{m-1} \sum_{i'=i+1}^m n_i n_{i'} (\theta_i - \theta_{i'}) (\tilde{p}_i - \tilde{p}_{i'}) (2t + (\tilde{p}_i + \tilde{p}_{i'}) (1 - 2t))}{n^2}$$

and $p_i = \frac{2N_{i1} + N_{i2}}{2n_i}$. Clearly, a sufficient condition for $\eta = 0$ is $\theta_i = \theta_j$ or $p_i = p_j$ for all $i \neq j$. Thus either different Case-Control ratios or different marker allele frequencies in different subpopulations might contribute to spurious associations. However, unlike in a population-based sample as introduced at section I-1.2.4, whether causal allele frequencies between subpopulations are equal or not is not an essential issue any more.

From our above derivations, the population stratification will have impacts on both the numerator and denominator of an Armitage's trend test, i.e. formula (II-3.15). However, such two causes of influence have different impacts on the test statistic. For the numerator, the impact is defined by an extra term η , i.e. the deviation of β_0 from 0. Because η can be either positive or negative, the presence of population stratification might result in an increase of false positive rate. On the contrary, the impact on the denominator is defined by the inverse of inflation factor $\lambda = \frac{1-F}{1+F}$. Since $0 < \lambda < 1$, the presence of population stratification will

increase the variance calculated from an original Armitage's trend test, i.e. formula (II-1.21) and (II-1.22), and hence such an effect would only reduce the test statistic rather to increase. As the second impact is only related to the difference in marker allele frequencies irrespective of the case-control ratio of samples selected from each subpopulation, in order to properly correct for such an effect, the dummy variables introduced in Method I should always be kept. Also because the second effect has rarely been properly addressed before, we will address its influence particularly in the simulations study demonstrated in the next section.

3.3.3 Simulation Study

In order to evaluate the efficiency of above two methods in adjusting the population stratification, we implement simulation studies to investigate their performance under a variety of parameter set ups. In the following simulations, we generate 25 Case-Control samples, where each of them is formed by a combination of Case-Control samples randomly collected from two subpopulations in HWE. These 25 samples can be categorised into five populations with 5 each as shown in Table II-10: There is no association between the testing marker and the disease in Population I, II and III, i.e. $D = 0$, but such an association exists in Population IV and V, i.e. $D = 0.0003$; The marker allele frequencies are identical between subpopulations in Population III and V, i.e. $p_1 = p_2 = 0.55$, but are different in the other populations, where $p_1 = 0.60$ and $p_2 = 0.30$ in Population I, and $p_1 = 0.60$ and $p_2 = 0.55$ in Population II and IV. Note here, because the disease causal allele is assumed to be rare, i.e. $q = 0.001$, even though the value of D (0.0003) is fairly small in Population IV and V, in terms of the adjusted LD (i.e. D' as introduced at section I-1.2.4), we could realise that $D = 0.0003$ is adequate. For example, the corresponding D' equals 0.4 in Population V, and the corresponding D' equal 0.5 or 0.4 in Population IV depending on p_1 or p_2 respectively. For the sake of comparison and consistency, the sample size of control collected from each

subpopulation is fixed as 5000, i.e. 10000 in total, and the combined sample size of case from both subpopulation is fixed as 6000. However, the sample size of case for each subpopulation varies in order to evaluate the performances of different methods under different data structures, where the cases collected from each subpopulation are shown in column Case I and II. The level of heterogeneity (LoH) is set to 0.95 for Population III and IV, i.e. the putative QTL is responsible for 5% case individuals. For convenience, only the co-dominance situation is considered, i.e. $f = 0.5$, and hence the default trend set is implemented as we have suggested in the section II-3.2.

To analyse simulated data as introduced at Table II-10, we will implement three methods, i.e. Method I as introduced at section II-3.3.1, Method II as introduced at section II-3.3.2 and the original Armitage's trend test with default trend set without any modification as introduced as formula (II-1.22). In order to demonstrate the overestimate of $Var(\hat{\beta})$ if ignoring the inflation factor λ as indicated in section II-3.3.2, we will list two results for Method II, i.e. one for χ_{II}^2 given as formula (II-3.15) and shown in Column 'Method II (with λ)' and the other for $\lambda \times \chi_{II}^2$, i.e. without the correction of variance as shown in Column 'Method II (without λ)'. For each Case-Control sample, simulations are replicated for 1000 times, and the mean and variance of the corresponding test statistics are summarized in Table II-11.

Table II-10. The Scheme of Simulation.

Columns Case I and II give the sample size of case in subpopulation I and II; columns p_1 and p_2 list allele frequencies of testing marker in subpopulation I and II; q is the disease causal allele frequency, D is the coefficient of LD between marker alleles and disease causal allele, and LoH is the abbreviate for the level of heterogeneity which represents the proportion of case individuals caused by the putative QTL associated with the testing marker.

Sample		Case I	Case II	p_1	p_2	q	D	LoH
1	Population I	3000	3000	0.60	0.30	0.001	0.00	/
2		3500	2500	0.60	0.30	0.001	0.00	/
3		4000	2000	0.60	0.30	0.001	0.00	/
4		4500	1500	0.60	0.30	0.001	0.00	/
5		5000	1000	0.60	0.30	0.001	0.00	/
6	Population II	3000	3000	0.60	0.55	0.001	0.00	/
7		3500	2500	0.60	0.55	0.001	0.00	/
8		4000	2000	0.60	0.55	0.001	0.00	/
9		4500	1500	0.60	0.55	0.001	0.00	/
10		5000	1000	0.60	0.55	0.001	0.00	/
11	Population III	3000	3000	0.55	0.55	0.001	0.00	/
12		3500	2500	0.55	0.55	0.001	0.00	/
13		4000	2000	0.55	0.55	0.001	0.00	/
14		4500	1500	0.55	0.55	0.001	0.00	/
15		5000	1000	0.55	0.55	0.001	0.00	/
16	Population IV	3000	3000	0.60	0.55	0.001	0.0003	0.95
17		3500	2500	0.60	0.55	0.001	0.0003	0.95
18		4000	2000	0.60	0.55	0.001	0.0003	0.95
19		4500	1500	0.60	0.55	0.001	0.0003	0.95
20		5000	1000	0.55	0.55	0.001	0.0003	0.95
21	Population V	3000	3000	0.55	0.55	0.001	0.0003	0.95
22		3500	2500	0.55	0.55	0.001	0.0003	0.95
23		4000	2000	0.55	0.55	0.001	0.0003	0.95
24		4500	1500	0.55	0.55	0.001	0.0003	0.95
25		5000	1000	0.55	0.55	0.001	0.0003	0.95

Table II-11. The Results of Simulation.

Means and variances of test statistics for Non-recessive model from populations given in Table II-10, by comparing Method I, Method II (with and without λ) and the non-adjusted Armitage Trend Test. A random variable with chi-square distribution with degree freedom 1 should have mean 1.0 and variance 2.0. Results are based on 1000 replicates for each population.

Sample	Method I		Method II (without λ)		Original		Method II (with λ)	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
1	1.051	2.151	0.876	1.493	0.876	1.493	1.050	2.139
2	1.000	1.688	0.829	1.162	18.22	59.02	0.994	1.666
3	1.061	2.353	0.865	1.563	69.30	243.7	1.032	2.204
4	0.997	2.061	0.787	1.285	157.2	510.1	0.937	1.829
5	1.037	2.181	0.779	1.230	276.4	923.5	0.922	1.727
6	0.944	1.928	0.940	1.908	0.940	1.908	0.944	1.942
7	0.939	1.667	0.929	1.630	1.473	3.809	0.934	1.653
8	1.011	2.077	0.979	1.950	3.179	11.78	0.984	1.980
9	0.977	1.891	0.914	1.652	5.757	20.55	0.918	1.666
10	1.037	2.008	0.917	1.573	9.761	37.85	0.920	1.578
11	0.999	1.833	0.999	1.834	0.999	1.834	1.001	1.847
12	0.982	1.823	0.975	1.800	0.985	1.833	0.975	1.795
13	1.083	2.326	1.055	2.205	1.078	2.309	1.055	2.207
14	0.991	1.919	0.931	1.694	0.985	1.928	0.933	1.718
15	1.158	2.629	1.029	2.078	1.117	2.425	1.029	2.080
16	7.903	28.87	7.864	28.58	7.863	28.58	8.010	29.66
17	8.143	31.18	8.048	30.45	12.56	49.32	8.207	31.85
18	7.632	28.91	7.391	27.11	17.45	68.13	7.537	28.23
19	7.674	29.41	7.169	25.66	24.29	97.22	7.314	26.75
20	7.277	26.66	6.432	20.83	32.09	118.6	6.559	21.73
21	7.641	29.60	7.642	29.61	7.642	29.60	7.745	30.51
22	7.953	29.62	7.902	29.25	7.995	30.01	8.011	30.09
23	7.755	29.01	7.551	27.50	7.963	29.37	7.651	28.19
24	7.186	26.15	6.751	23.08	7.649	28.46	6.854	23.98
25	6.823	24.38	6.065	19.27	7.622	26.80	6.150	19.86

Method I: Linear Model with Dummy Variables (Section II-3.3.1)

Method II: Non-central Chi-square Test (Section II-3.3.2)

Original: Armitage's Trend Test (Section II-1.4.3)

In Population I, we set the difference of allele frequencies between two subpopulations fairly large, i.e. $p_1 - p_2 = 0.3$, to demonstrate the influence of ignoring λ as we have thoroughly discussed above. For Sample I, where the case control ratios between two sub-samples are identical, we have $\eta = 0$ and hence Method II (without λ) is identical to Original as indicated by (II-3.15). We may easily notice that both Method II (without λ) and Original suffer significant reduction of test statistics comparing to Method I and Method II (with λ) (P-value = 0.004) as shown in Table II-11. Similar reductions can also be observed for the results of Method II (without λ) in the other samples, although the introduction of η has effectively eliminated the influence from the population stratification for the numerator part. We can hence conclude that λ must be included in Method II to establish a reliable test without loss of statistical power. Of course, as we can notice from the results of the Original, the ignorance of η will cause much more serious biases. Even if the allele frequencies are quite similar in two subpopulations as shown in Population II, i.e. $p_1 - p_2 = 0.05$, a large difference between case control ratios, i.e. Sample 8, 9 and 10, may still result in a significant increase of false positive rates under the null hypothesis, i.e. the means of test statistics from the Original dramatically increase in these samples where no true LD is present. A similar dramatic increase of test statistics for Original can also be observed in Population IV, where the alternative hypothesis is true, i.e. $D \neq 0$. Obviously, in these situations, the implementation of a modified Armitage's trend test is highly recommended, and as we can observe in Populations I, II and IV in Table II-11, both Method I and Method II (with) λ can properly adjust the influence from population stratification, although Method I shows a higher statistical power if the difference of allele frequencies between sub-samples are large, i.e. Sample 18, 19, 20, while Method II is better if otherwise, i.e. Samples 16, 17. Note here, in the presence of LD, the effect of population stratification can either increase or decrease a test statistic, where, however, only the case of increasing can be observed in our simulation as Population IV.

In Population III and V, where subpopulations are identical as $p_1 = p_2$, the combined Case-Control samples are equivalent to being collected from a single population. In such a circumstance, the adjustment for population stratification is not necessary, and both Methods I and II may suffer certain bias as unnecessary parameters are implemented. Such a bias can be observed from the results for Population V as shown in Table II-11, where we can notice that as the difference between Case I and II increase, both Methods I and II will gradually suffer a larger loss of statistical power if compared to the Original. Note here, for Samples 21 and 22, Method II (with λ) has shown even higher test statistics than that of the Original, which is mainly because the inflation factor λ will always less than 1 in the presence of LD even if there is no population stratification as we will discuss later. We might suspect that the increase in test statistic will also increase the rate of false positive. However, as shown in the results for Population II, both Methods I and II will not cause any increase of false positive rate under the null hypothesis, i.e. $D=0$, and we can hence conclude that the increased statistical power from Method II (with λ) is not because of a skewed distribution. Actually, such an increase of statistical power can be expected. In the presence of LD, i.e. the alternative hypothesis is true, the case individuals contributed from the putative QTL do have different allele frequencies in associated markers from the control individuals, and as the a difference will surely lead to a reduction of heterozygotes, the generalised F hence exists in such a case-control sample. A similar discussion can be conducted between the original Armitage's trend test and the allelic analysis that in the presence of LD, as the combined case and control samples will tend to be biased downwards when the population deviates from the HWE, we might more likely to have $4N_1N_3 > N_2^2$ and hence $\chi_A^2 > \chi_G^2$ as indicated by formula (II-2.8). More precisely, we can calculate that $\frac{\chi_A^2}{\chi_G^2} = 1 + F(1 + F) < \frac{1}{\lambda}$, and hence it can be

concluded that $\chi_H^2 > \chi_A^2 > \chi_G^2$ given $t=1/2$, i.e. x_2 the trend coefficient for heterozygote, if the alternative hypothesis is true, i.e. $D \neq 0$.

In order to demonstrate both Method I and Method II (with λ) can properly adjust for the population stratification without increase in false positive rate, we draw Q-Q Plots of Method I, Method II (with λ) and Original with their theoretical distribution, i.e. chi-square distribution with 1 degree of freedom, for Sample 1 and 11 with 10000 replicates each. The Q-Q plots for Sample 1 are illustrated as Figure II-7, where the case-control ratios of sub-samples are identical but the marker allele frequencies are different between subpopulations. We can notice that the Original, shown as Figure II-7 (c), are severely biased downwards from its theoretical distribution due to the overestimate of variance, but such an influence from the population stratification can be properly adjusted by both Method I, Figure II-7 (a), and Method II (with λ), Figure II-7 (b), where both of which suffer slightly downward bias if the test statistics are large, i.e. over 10.0, which implies that both modified methods might suffer a certain loss of statistical power if the test statistic is large. As 10.8 is the threshold at significance level 0.001 in a single test, such a bias should not be a problem for a single test, but if a family-based significance level is applied for multiple tests, e.g. the Bonferroni's correction or FDR, such a bias might result in the loss of statistical power. The Q-Q plots of Method I, Method II (with λ) and Original for Sample 11 are illustrated as Figure II-8 (a), (b) and (c), respectively, where the case control ratios of sub-samples are identical and the marker allele frequencies are identical in two subpopulations. It is easy to notice that all three methods fit the theoretical chi-square distribution with 1 degree of freedom almost perfectly, and hence we can implement both Method I and Method II (with λ) without the risk of causing false positive even in the absence of population stratification. From the discussion for Figure II-8 and the simulation results of Sample 21 in Table II-11 we can conclude that Method II (with λ) would be the most powerful test among these three if the two sub-

samples are actually sampled from the same population with equal case control ratios. Such a conclusion also implies that by letting $\eta=0$, Method II (with λ) should have a higher statistical power than the Original if both are applied on a case-control sample without population stratifications. Unfortunately, as has been indicated in the results of Population IV and V in Table II-11, as the difference of case control ratio between sub-samples increases, the estimate of η will tend to over-correct the population stratification and the corresponding test statistics of Method II (with λ) will reduce so fast that it will become less favoured than Method I.

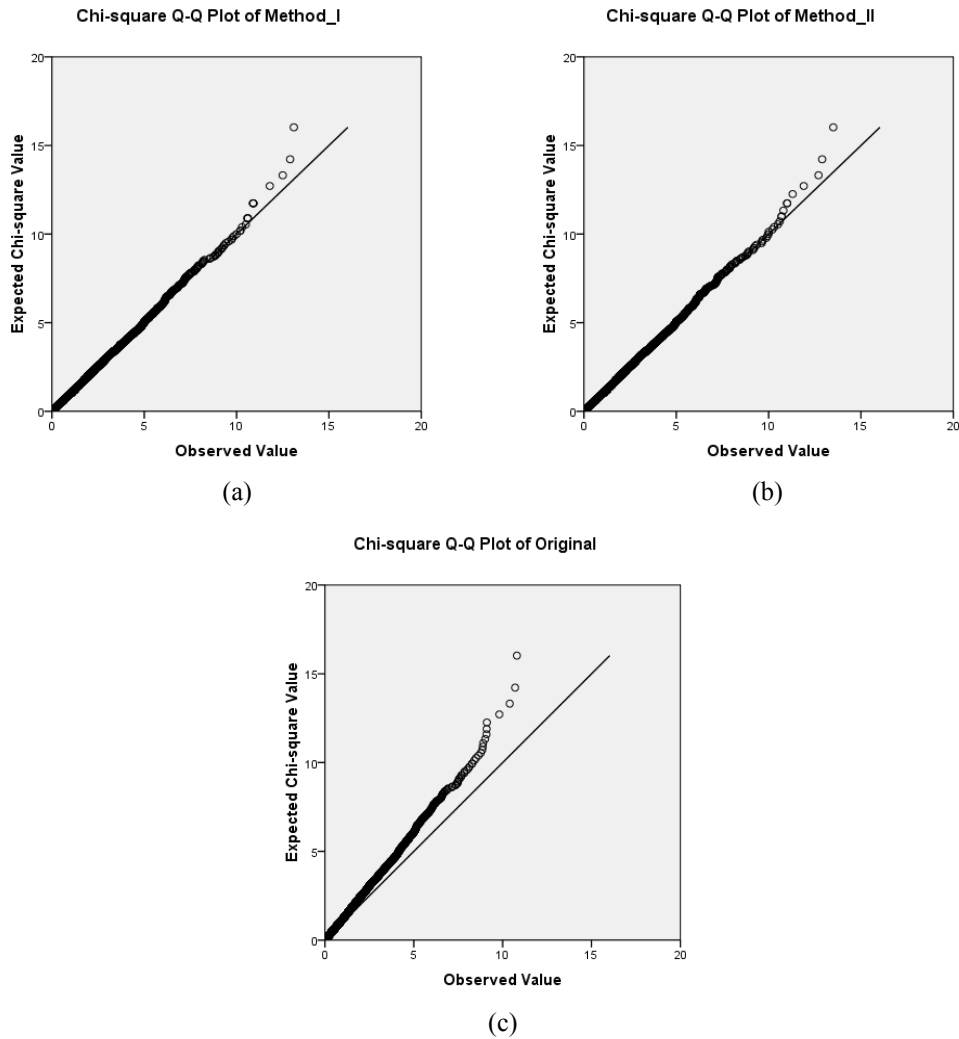


Figure II-7. Q-Q Plots of Method I (a), Method II (with λ) (b) and Original (c) for Sample 1 in Table II-10 with 10000 replicates.

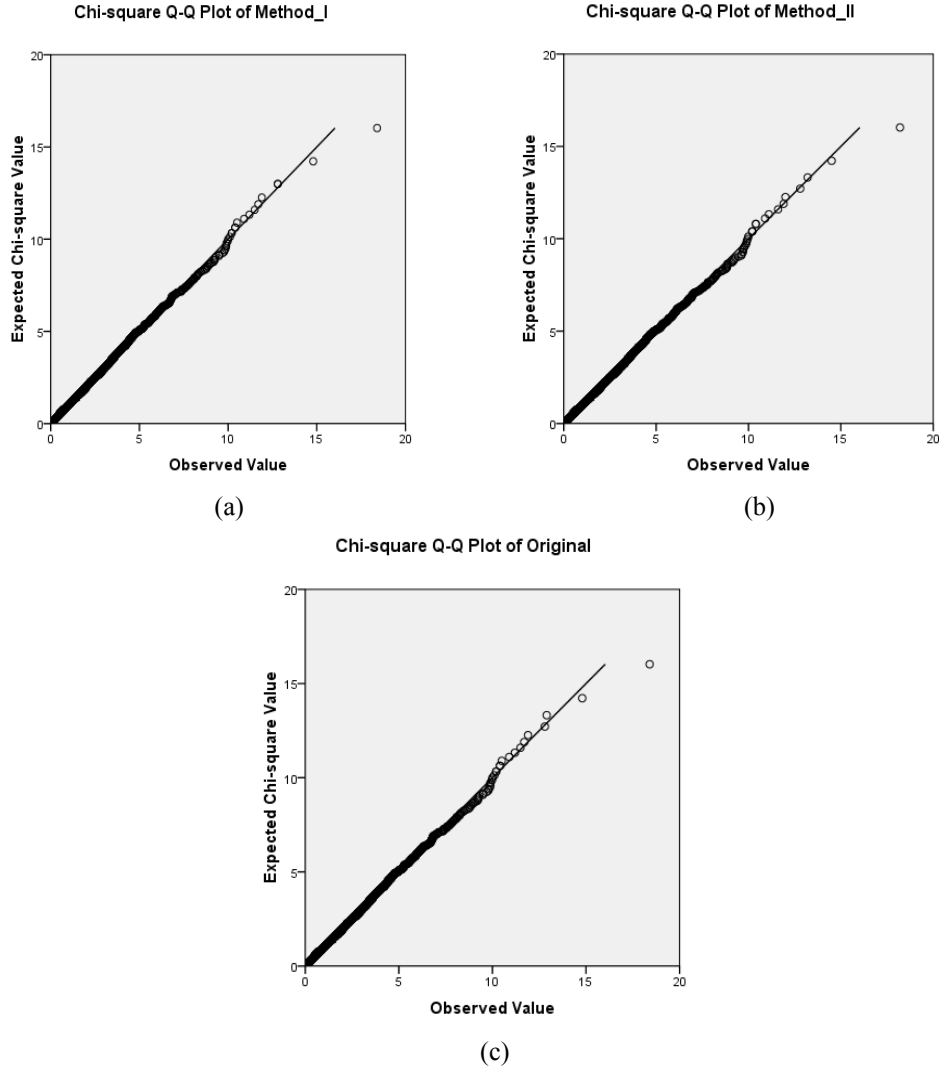


Figure II-8. Q-Q Plots of Method I (a), Method II (with λ) (b) and Original (c) for Sample 11 in Table II-10 with 10000 replicates.

In the above results and discussions, we have demonstrated that integrating samples with population stratification might cause serious false inferences, which, however, can be properly adjusted through both Method I and Method II (with λ) as introduced in section II-3.3.1 and II-3.3.2 respectively. We have also demonstrated that the implementation of Method I and Method II (with λ) will not cause any increase of false positive rate, although their statistical powers might be reduced if the difference of case control ratios between sub-

samples is sufficiently large. Of these two, the modified Armitage's trend test, Method I is favoured if the case control ratios are largely different between sub-samples and Method II (with λ) is favoured otherwise. It is also worth noting that by assigning $\eta = 0$, Method II (with λ) is more powerful than the original Armitage's trend test for an ideal population.

3.3.4 Real Data Analysis

In order to further evaluate both Method I and Method II (with λ), we re-analyse the Stage I data of the Parkinson's disease from Simon-Sanchez et al. (2009) with both our new methods. For Stage I data, there are 4005 individuals collected from the United States and 1686 individuals from Germany. Of these 5691 individuals, there are 1713 cases and 3978 controls, where the sample from the United States consists of 971 cases and 3034 controls and the sample from Germany consists of 742 cases and 944 controls. After quality control, 453,585 SNPs are left to analyse. Although both samples are European Cohort, as we have discussed above, this does not guarantee that both samples will have identical allele frequencies at each marker locus, and hence either Method I or II should be implemented rather than Original. As it is obvious that the case-control ratios are largely different between samples from the United States and Germany, i.e. approximately 2:5, we would recommend Method I rather than Method II (with λ) as we have concluded in simulation studies in section II-3.3.3. Note here, as Method II (without λ) will not be included in this section, we will simply use 'Method II' instead of 'Method II (with λ)' for convenience. As usual, the test results from Methods I and II are compared with Original, and their corresponding Manhattan plots, of $-\log(\text{P-value})$, are illustrated in Figure II-9, Figure II-10 and Figure II-11, respectively. We have implemented three trend sets with all of three methods, i.e. the default $\{1, 0.5, 0\}$, the recessive $\{1, 0.25, 0\}$ and the extremely recessive $\{1, 0, 0\}$, but only the results of default and the recessive are illustrated in each of below figures as (a) and (b) respectively. Although we

illustrate the threshold of 0.05 significant level after Bonferroni's correction as the blue line in each of the figures introduced below, we have also calculated the Q-value for each method with a particular trend set, and the significant markers with a Q-value less than 0.05 in Method I are listed in Table II-12.

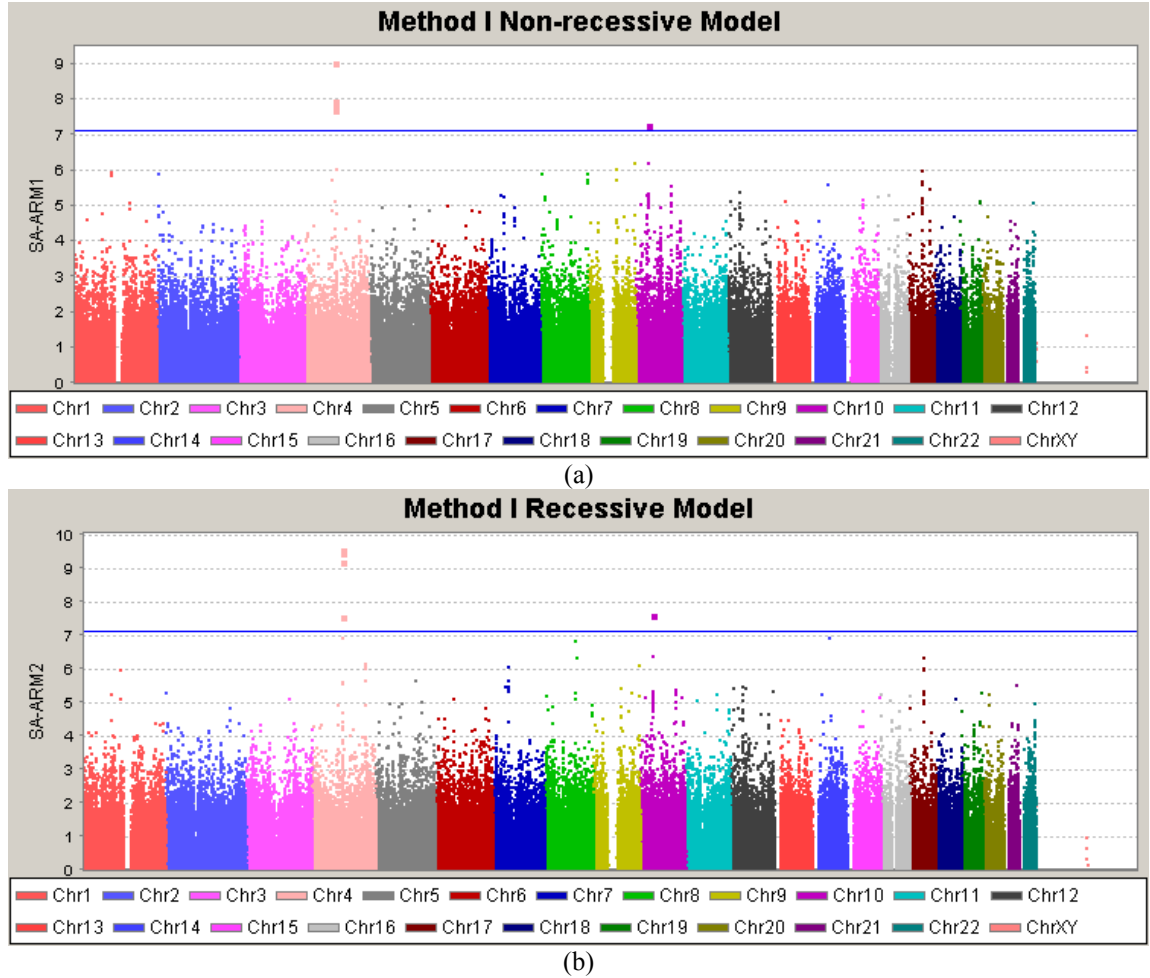


Figure II-9. The Manhattan plots of $-\log(\text{P-value})$ for Method I with the default trend set (a), i.e. $\{1, 0.5, 0\}$, and the recessive trend set (b), i.e. $\{1, 0.25, 0\}$. The blue line is the threshold of 0.05 significant level after Bonferroni correction.

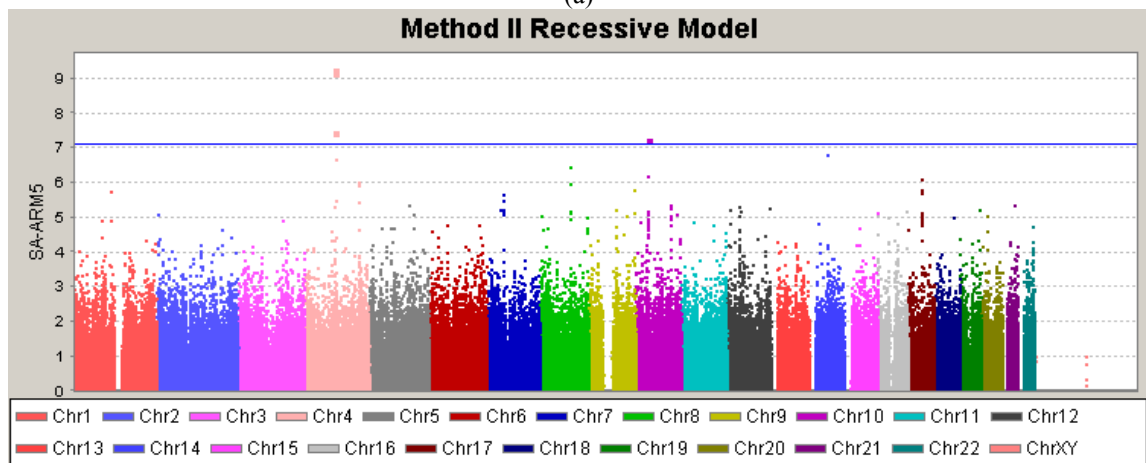
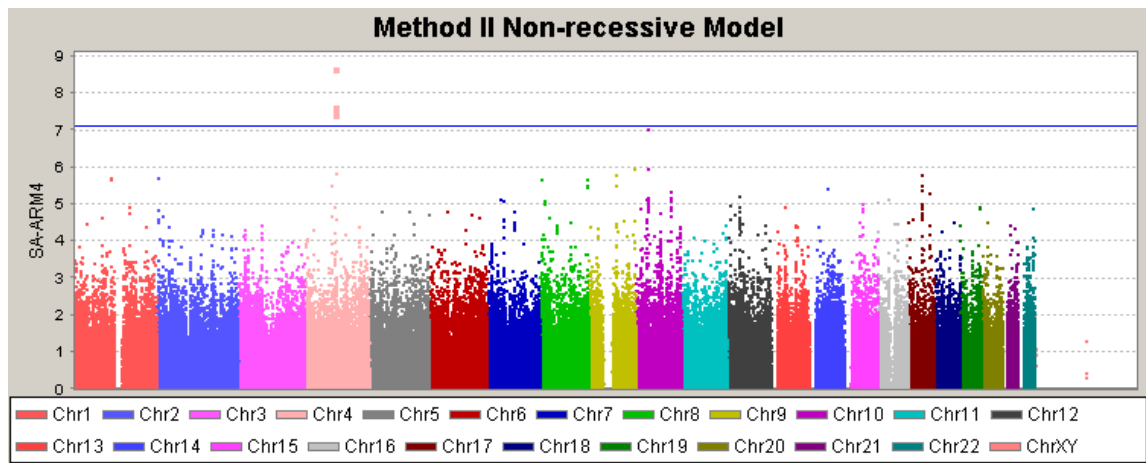
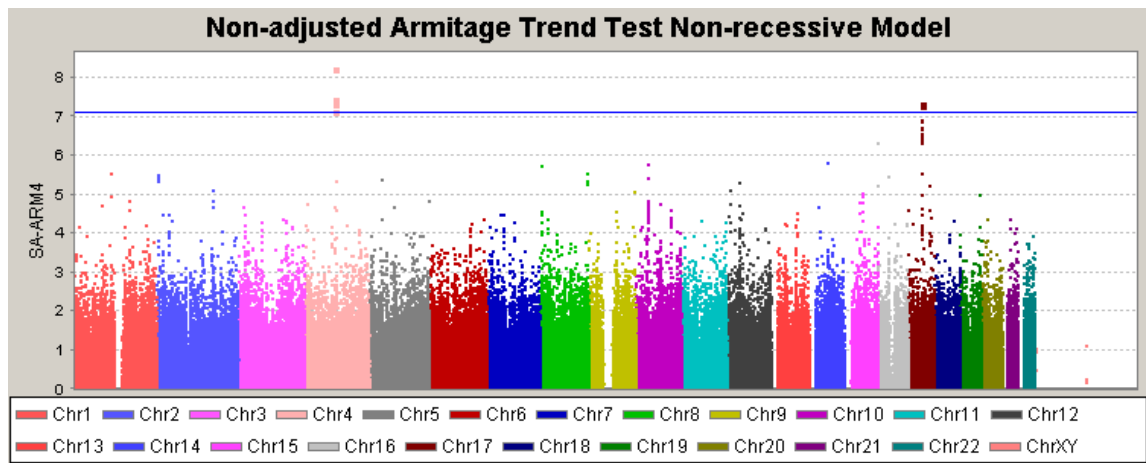
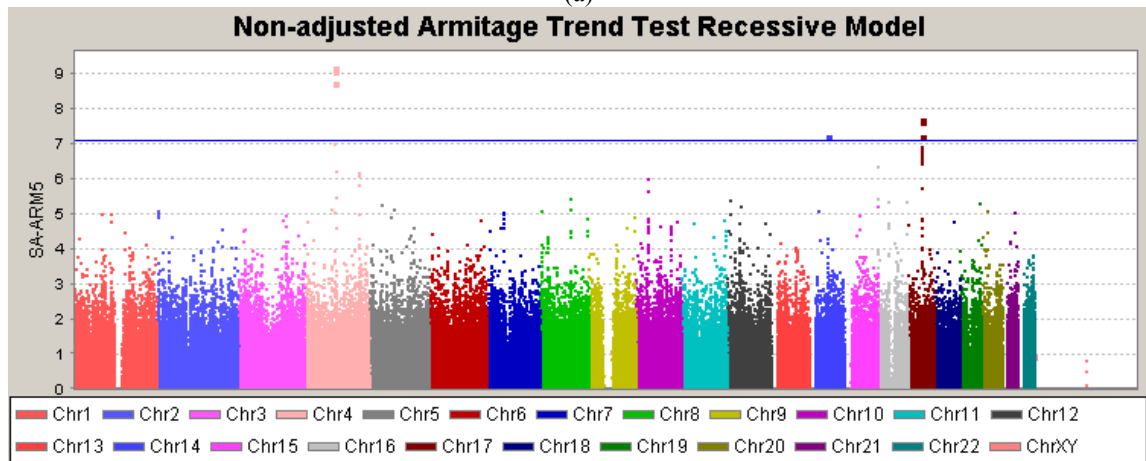


Figure II-10. The Manhattan plots of $-\log(\text{P-value})$ for Method II with the default trend set (a), i.e. $\{1, 0.5, 0\}$, and the recessive trend set (b), i.e. $\{1, 0.25, 0\}$. The blue line is the threshold of 0.05 significant level after Bonferroni correction.



(a)



(b)

Figure II-11. The Manhattan plots of $-\log(\text{P-value})$ for Original, i.e. the non-adjusted Armitage's trend test, with the default trend set (a), i.e. $\{1, 0.5, 0\}$, and the recessive trend set (b), i.e. $\{1, 0.25, 0\}$. The blue line is the threshold of 0.05 significant level after Bonferroni correction.

Table II-12. List of Most Significant Markers with Test Score, of $-\log(P\text{-value})$ from Method I

* rs12431733 is very likely to be a false recorded marker since there is no any other marker within the $\pm 1\text{Mb}$ region with a $-\log(P\text{-value})$ over 4.0. Besides the nearest known gene BMP4 is over 0.1 Mb away from this locus.

Marker name	Chromosome	Location (Mb)	Candidate Gene	Test Score ($-\log P$) and Trend Set	Q-value in Method I
rs3857059	4	90.675	SNCA	9.57 {1, 0.25, 0}	0.0000
rs11591754	10	35.219	CREM or CUL2	7.60 {1, 0.25, 0}	0.0022
rs12431733*	14	54.291	BMP4	6.93 {1, 0, 0}	0.0060
rs2616510	8	89.018	MMP16	6.91 {1, 0, 0}	0.0065
rs7678831	4	159.441	RXFP1 (LGR7)	6.68 {1, 0, 0}	0.0129
rs199533	17	44.829	NSF or MAPT	6.34 {1, 0.25, 0}	0.0166
rs2708909	7	48.052	SUNC1	6.04 {1, 0.25, 0}	0.0224
rs6542651	2	3.760	LOC728597	5.90 {1, 0.5, 0}	0.0364
rs7004938	8	140.259	HS. 8610 (UniGene)	5.89 {1, 0.5, 0}	0.0364
rs7013027	8	4.924	CSMD1 or PLEKHA3	5.88 {1, 0.5, 0}	0.0364
rs10857899	1	112.128	RAP1A	5.87 {1, 0.5, 0}	0.0364

Comparing Figure II-9, Figure II-10 and Figure II-11, we observed that the Manhattan plots of Methods I and II are quite similar to each other but they are both different from Original at several important marker loci, i.e. rs11591754, rs2616510 and rs199533, which are reported

as significant markers in Table II-12. As these differences may be largely due to the population stratification, i.e. different allele frequencies in sub-samples, we evaluate the difference of allele frequencies between two sub-samples at each significant marker loci with a t -test, and the corresponding test results are listed in the column ‘P-value of t -test’ in Table II-13. We can notice that the three significant markers rs11591754, rs2616510 and rs199533, do have distinguishable allele frequencies in two sub-samples, and their corresponding t -test results are highlighted in red as shown in Table II-13. These results hence demonstrate that both Methods I and II have properly adjust the population stratification in the combined data. In Simon-Sanchez’s paper (2009), rs11591754 and rs2616510 were missed mainly because of the ignorance of population stratification introduced by combining samples, and for the same reason the test result of rs199533 was over estimated, although the correction for population stratification does not affect its significance under the scheme of Q-values. Among these three methods, we can see from Table II-13 that Method I generally yields the highest test statistics, and also because of the implementation of Method I will not cause any increase of false positive as we have proved through comprehensive simulation studies in section II-3.3.3, we can conclude that Method I is preferable than Method II for this particular data as we have predicted before. In the following discussions, only the test results from Method I will be reported unless otherwise specified.

Table II-13. Comparison of Three Methods at Significant Loci.

-log(P-value) of significant markers in three methods with different trend sets are shown in Table II-12. The highest -log(P-value) of significant markers under the specific trend set of each method is highlighted in Green. The genetic markers with significant (or almost significant) different allele frequencies in two subpopulations are highlighted in Red for P-value column in order to show the performances of our SA-Methods and Yellow indicates non-significant but still possible different marker allele frequencies.

* significant at level 0.05

Marker Name	Method I			Method II			Method III			P-values of <i>t</i> -test
	{1,0.5,0}	{1,0.25,0}	{1,0,0}	{1,0.5,0}	{1,0.25,0}	{1,0,0}	{1,0.5,0}	{1,0.25,0}	{1,0,0}	
rs3857059	7.82	9.57	6.36	7.80	9.63	6.56	7.44	9.20	6.30	0.923067
rs11591754	7.28	7.60	7.53	6.77	6.98	6.88	5.78	6.01	5.95	0.060454
rs2616510	4.70	6.82	6.91	4.53	6.47	6.45	3.60	5.42	5.79	0.063007
rs7678831	4.00	6.15	6.68	3.86	5.92	6.42	3.76	5.80	6.35	0.840235
rs199533	6.00	6.34	6.32	5.61	5.93	5.92	7.30	7.63	7.56	0.014987*
rs2708909	5.27	6.04	5.79	5.28	5.90	5.54	4.46	5.05	4.80	0.260811
rs6542651	5.90	5.27	1.31	5.82	5.19	1.27	5.48	4.88	1.21	0.727532
rs7004938	5.89	4.91	3.77	5.76	5.05	4.05	5.53	4.84	3.88	0.807172
rs7013027	5.88	5.20	3.88	5.45	4.84	3.63	5.73	5.09	3.81	0.903645
rs10857899	5.87	5.10	3.98	5.76	4.98	3.88	5.51	4.77	3.71	0.815306

Method I: Linear Model with Dummy Variables (Section II-3.3.1)

Method II: Non-central Chi-square Test (Section II-3.3.2)

Method III: The Original Armitage's Trend Test (Section II-1.4.3)

Other than the comparison between statistical methods, we can also compare the performances of different trend sets as introduced in section II-3.2. In Table II-13, the highest test statistic among three trend sets in each method is highlighted in Green. From these results, we noticed that of the top 6 significant SNPs, the 4 most significant results are detected under trend set $\{1, 0.25, 0\}$ and 2 are detected under trend set $\{1, 0, 0\}$, and such a result might support the implementation of trend set $\{1, 0.25, 0\}$. Although because of sampling errors, a marker having its peak value with trend set $\{1, 0.25, 0\}$ does not guarantee the corresponding gene is recessive. However, we are quite convinced to believe that the associated genes with rs2616510 and rs7678831 are recessive due to their significant increases of test statistics with trend set $\{1, 0, 0\}$ compared to that with trend set $\{1, 0.5, 0\}$.

After discussing the GWAS results statistically, we would like to look at what is behind these significant markers. Although some of them have already been repetitiously reported, e.g. rs3857059 in gene SNCA and rs199533 in gene MAPT, some of the rest have rarely been mentioned before, e.g. rs2616510 for gene MMP16 and rs11591754 for CUL2 or CREM. As Parkinson's disease is the second largest neurodegeneration disease after Alzheimer's disease, its candidate genes are very likely to be involved in the neurodegeneration process, and we would like to give a brief review of candidate genes which are functional in such a process as following:

1. rs3857059 is located at 90.675Mb on the human Chromosome 4, within the range of the candidate gene SNCA, also known as PARK1, which is one of the most widely reported genes to be associated with the Parkinson's disease (Kruger et al., 1999, Farrer et al., 2001, Mamah et al., 2005, Maraganore et al., 2005, Mueller et al., 2005, McCulloch et al., 2008, Myhre et al., 2008, Sutherland et al., 2009). There are also several other markers in support of

this candidate gene, such as rs11931074 ($-\log(\text{P-value})=9.47$), rs2736990 ($-\log(\text{P-value})=9.18$) and etc..

2. rs11591754 is located at 35.219Mb on human Chromosome 10. After correction for population stratification, both Methods I and II have indicated a strong association between disease and this locus. Although rs11591754 itself is not in any known gene region, around its locus, there are over 10 supportive markers with $-\log(\text{P-value})$ over 5.0 in the surrounding 0.3Mb, and there are two genes which fall within this region: CUL2 and CREM. Although there is no previous report about either of these two genes associated with the Parkinson's disease, some researches have shown that neurodegenerative disease, such as the Parkinson's disease, is related with ubiquitination (Shimura et al., 2000, Kahle and Haass, 2004) and since CUL2's production Cullin has been reported to play a role in ubiquitination (Marin and Ferrus, 2002), it may be reasonable for us to consider CUL2 as a candidate gene for the Parkinson's disease. Recently, a newly published PCR Array 'The Human Parkinson's Disease RT² Profiler™' by QIAGEN has included CUL2 as one of its 84 candidate genes (http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-124A.html). On the other hand, CREM is intensively reported to be highly involved with neurodegeneration (Mantamadiotis et al., 2002, Klejman and Kaczmarek, 2006, Valor et al., 2010), and hence it is also an important candidate gene of the Parkinson's disease. Both these two candidate genes have not been reported in Simon-Sanchez's paper (1999) probably due to the ignorance of population stratification.

3. rs2616510 is located at 89.018Mb on human Chromosome 8. As it was only observed as significant or close to in trend sets $\{1, 0.25, 0\}$ and $\{1, 0, 0\}$, we could suggest that its associated putative gene may be recessive. This marker is 30Kb away from the initial of gene MMP16 and with its supportive markers, such as rs278891 ($-\log(\text{P-value}) = 5.30$) and

rs2664363 ($-\log(\text{P-value}) = 5.12$), in the same region. Although lack of report before, recently, Edwards et al. (2001) reports a detected association between the MMP16 and the Parkinson's disease from a different data with a P-value at level 10^{-5} . Along with Edwards' reports, several papers in Neurology have predicted the relationship between the MMP family and the Parkinson's disease (Lorenzl et al., 2002, Kim et al., 2009), and we might hence be supportive in that gene MMP16 plays an important role in the Parkinson's disease. In Simon-Sanchez, J. et al. (2009), this candidate gene has not been reported due to both the population stratification as well as the improper choice of trend set, where only $\{1, 0.5, 0\}$ were implemented.

4. rs7678831 is located at 159.441 on human Chromosome 4, just located in the gene RXFP1/LGR7 according to the UniGene. Although lack of direct evidence, Piccenna, L. et al. (2005) pointed out that 'the strong expression of LGR7 in the claustrum is of interest', given that the claustrum is implicated to be functional in the pathology of neurodegenerative disease, e.g. the Alzheimer's disease and the Parkinson's disease.

5. rs199533 is located at 44.829Mb on human Chromosome 17, just within the region of the gene NSF. Several other supportive markers with $-\log(\text{P-value})$ over 5.0 were in the range from 43.72Mb to 44.82Mb. The genes CRHR1, IMP5, MAPT, STH, KIAA1267, LRRC37A and NSF locate within this region, which have been repeatedly reported in association with Parkinson's disease (Martin et al., 2001, Scott et al., 2001, Zappia et al., 2003, Healy et al., 2004, Kwok et al., 2004, Skipper et al., 2004, Levecque et al., 2004, Mamah et al., 2005, Fidani et al., 2006, Fung et al., 2006, Goris et al., 2007, Vandrovcova et al., 2007, Winkler et al., 2007, Zabetian et al., 2007). Although Original has shown a higher test statistic than both Methods I and II, it might be explained by the chance that the presence of population

stratification increases the testing value instead of decreasing it. Nevertheless, after correction, the association between this locus and disease is still significant.

Since no experimental or independent evidence is available to support the relationship between other candidate genes detected here and the Parkinson's disease, no comments on the other candidates will be given.

3.3.5 Conclusion

In section II-3.3, we have introduced two strategies to adjust for population stratification if samples are combined from multiple resources. Method I is established by the implementation of dummy variables to denote subpopulations in a linear model, and Method II is established by correcting the influence from population stratification for both the numerator and denominator of formula (II-1.21). We have evaluated the performance of both methods in comprehensive simulation studies, and the results indicate that both methods can properly adjust for population stratification and will not cause any increase of false positive rate. Our simulation studies also show that Method I is preferable if the case-control ratios are largely different among sub-samples while Method II is preferable otherwise. An incidental result from the simulation studies indicates that in the absence of population stratification, where $\eta = 0$, Method II is more powerful than the original Armitage's trend test if the HWE in each subpopulation holds. We also implemented both Methods I and II into an analysis of the Parkinson's disease data, and several new candidate genes have been revealed by our methods, e.g. MMP16, CREM, CUL2, all of which have long been proposed as candidates by neurobiologist but are still missing from the detected associations in the current literature of the GWAS with the genetic disorder.

3.4 Conclusion and Discussion

In Chapter II-3, we have firstly discussed the optimal trend set for the Armitage's trend test under different penetrance models as well as the presence of genetic heterogeneity. It turns out that a dominant QTL, i.e. $(f_1 = f_2 = 1, f_3 = 0)$, or an additive QTL, i.e. $(f_1 = 1, f_1 > f_2 \neq 0, f_3 = 0)$, will result in an almost identical optimal trend set if a putative QTL can only explain a small proportion of case individuals, and the corresponding optimal trend set can be reasonably represented as $\{1, 0.5, 0\}$. For the recessive model, as the optimal trend set varies in term of D' , we hence suggested using both $\{1, 0.25, 0\}$ and $\{1, 0, 0\}$ to detect recessive QTLs. These results do not favour the suggestion given by Sasieni (1997), where he suggested to use trend set $\{1, 1, 0\}$ for dominant QTLs and $\{1, 0, 0\}$ for recessive ones. However, the results from simulation studies for dominant, additive and recessive QTLs do support our suggestion. Secondly, we introduced two strategies to correct for population stratification when multiple samples are combined to increase the statistical power in section II-3.3. Both simulation studies and real data analyses have shown that both our methods could properly remove the influence from population stratification and will not cause any increase in false positive rate.

In section II-3.3.2, particularly, we re-analyse the dispersion of the denominator of formula (II-1.22), i.e. the Armitage's trend test with trend set $\{1, 0.5, 0\}$. A previous analysis was given by Devlin & Roeder (1999), who claimed that in the presence of population stratification, the denominator of formula (II-1.22) will tend to underestimate the variance of the square root of the corresponding numerator. However, we have shown in both our theoretical and simulation analyses that in the presence of population stratification, it is an over-dispersion rather than an under-dispersion. We have calculate such an under-dispersion

rate is $\lambda = \frac{1-F}{1+F}$, where F is the Wright's coefficient of inbreeding estimated as $1 - \frac{O(N_2)}{E(N_2)}$,

where $O(N_2)$ and $E(N_2)$ are the observed number of heterozygotes and expected number of heterozygotes under the HWE. We have also noticed that in both Sasieni (1997) and Devlin & Roeder (1999), an example that case and control sample are coincidentally collected from distinguishable subpopulations have been mentioned in their discussions. However, such a sample is inappropriate because it invalidates the null hypothesis fundamentally as we have discussed in section II-2.3 and hence any modifications of statistical models, which are based on the null hypothesis, would be meaningless.

We have also re-evaluated the performances of different test statistics when no population stratification is present, since the previous discussion given by Sasieni (1997) is inappropriate as we have revealed in section II-2.3. We showed that allelic analysis, i.e. formula (II-1.19), would have higher statistical power than the Armitage's trend test with trend set $\{1, 0.5, 0\}$, i.e. formula (II-1.22), if the alternative hypothesis is true, i.e. $D \neq 0$. More importantly, we have also shown that if the HWE can be assumed, by assigning $\eta = 0$ in the absence of population stratification, our Method II with trend set $\{1, 0.5, 0\}$, i.e. formula (II-3.14), would be the most powerful of the three. Of course, the advantage of Method II over the original Armitage's trend test is generally true irrespective of the trend set implemented as long as the HWE can be assumed. However, as we have indicated in section II-3.2.3, if some genotypes are missing from the observed data, allelic analysis should be implemented as it is the least influenced by a particular missing genotype.

We would also like to note here that as we have not assumed the exact value of vector \mathbf{X} in Method I, i.e. formula (II-3.9), Method I can also be implemented for allelic analysis by denoting x_i equalling 1 or 0. We will not be bothered to show the exact proof here, but such a test statistic yields from Method I for the allelic analysis is asymptotically identical to that of Mantel-Haenszel test (Zhang and Boos, 1997).

In above, we have summarized our methods and results in Chapter II-3, and we have shown that the implementation of our contributions, i.e. the optimal choice of trend set and statistical methods to correct for population stratification in the Armitage's trend test, would efficiently decrease both false inferences. However, certain improvements are still available. For the optimal trend set, especially for the recessive model, our suggestion of using $\{1, 0.25, 0\}$ and $\{1, 0, 0\}$ is applicable but not an exactly optimal choice as much of other available information, e.g. the recombination hotspot, is not included. On the other hand, as we have proved that the optimal trend coefficient does not rely on the observed case proportion but rather the effective case proportion, i.e. the individuals contributed from the putative QTL under question, we may expect a significant increase of statistical power if we can move certain case individuals, which are contributing to the genetic heterogeneity, into the control sample, and a bootstrap strategy can be implemented to control the false positive rate. For the modified Armitage's trend test, we haven't considered the situation where each sub-sample may have their own sub-structures, which will surely cause structural LD. Such a problem can be easily solved in Method I by replacing the dummy variables with the principal component vectors (Price et al., 2006), but there might be a problem here to properly estimate the influence from a hidden population structure in Method II, especially for the numerator. A proper way to include implicit structure information, e.g. the principal component vectors, into Method II is the one of the priority topics in our following studies.

Reference

- ABECASIS, G. R., CARDON, L. R. & COOKSON, W. O. C. (2000) A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66, 279-292.
- ARMITAGE, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375-386.
- BALDING, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7, 781-791.
- CANTOR, R. M., LANGE, K. & SINSHEIMER, J. S. (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, 86, 6-22.
- COCHRAN, W. G. (1954) Some methods for strengthening the common x² tests. *Biometrics*, 10, 417-451.
- DEVLIN, B. & ROEDER, K. (1999) Genomic control for association studies. *Biometrics*, 55, 997-1004.
- EDWARDS, T. L., SCOTT, W. K., ALMONTE, C., BURT, A., POWELL, E. H., et al. (2001) Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Annals of Human Genetics*, 74, 97-109.
- FARRER, M., MARAGANORE, D. M., LOCKHART, P., SINGLETON, A., LESNICK, T. G., et al. (2001) Alpha-synuclein gene haplotypes are associated with Parkinson's disease. *Human Molecular Genetics*, 10, 1847-1851.
- FIDANI, L., KALINDERI, K., BOSTANTJOPOULOU, S., CLARIMON, J., GOULAS, A., et al. (2006) Association of the Tau haplotype with Parkinson's disease in the Greek population. *Movement Disorders*, 21, 1036-1039.
- FUNG, H. C., XIROMERISIOU, G., GIBBS, J. R., WU, Y. R., EEROLA, J., et al. (2006) Association of Tau haplotype-tagging polymorphisms with Parkinson's disease in diverse ethnic Parkinson's disease cohorts. *Neurodegenerative Diseases*, 3, 327-333.
- GORIS, A., WILLIAMS-GRAY, C. H., CLARK, G. R., FOLTYNIE, T., LEWIS, S. J. G., et al. (2007) Tau and Alpha-synuclein in susceptibility to, and dementia in, Parkinson's disease. *Annals of Neurology*, 62, 145-153.
- HEALY, D. G., ABOU-SLEIMAN, P. M., LEES, A. J., CASAS, J. P., QUINN, N., et al. (2004) Tau gene and Parkinson's disease: A case-control study and meta-analysis. *Journal of Neurology Neurosurgery and Psychiatry*, 75, 962-965.
- HILL, A. P. (1975) Quantitative linkage - statistical procedure for its detection and estimation. *Annals of Human Genetics*, 38, 439-449.
- KAHLE, P. J. & HAASS, C. (2004) How does parkin ligate ubiquitin to Parkinson's disease? First in molecular medicine review series. *EMBO Reports*, 5, 681-685.

- KENDALL, M. G. & STUART, A. (1961) *The advanced theory of statistics*, London, Griffin.
- KIM, S. J., CHUNG, Y. K., CHUNG, T. W., KIM, J. R., MOON, S. K., et al. (2009) Regulation of matrix metalloproteinase-9 expression between gingival fibroblast cells from old and young rats. *Biochemical and Biophysical Research Communications*, 378, 152-156.
- KLEJMAN, A. & KACZMAREK, L. (2006) Inducible camp early repressor (ICER) isoforms and neuronal apoptosis in cortical in vitro culture. *Acta Neurobiologiae Experimentalis*, 66, 267-272.
- KNAPP, S. J. & BRIDGES, W. C. (1990) Using molecular markers to estimate quantitative trait locus parameters - power and genetic variances for unreplicated and replicated progeny. *Genetics*, 126, 769-777.
- KNOTT, S. A. (1994) Prediction of the power of detection of marker-quantitative trait locus linkages using analysis of variance. *Theoretical and Applied Genetics*, 89, 318-322.
- KRUGER, R., VIEIRA-SAECKER, A. M. M., KUHN, W., BERG, D., MULLER, T., et al. (1999) Increased susceptibility to sporadic Parkinson's disease by a certain combined alpha-synuclein/apolipoprotein E genotype. *Annals of Neurology*, 45, 611-617.
- KU, C. S., LOY, E. Y., PAWITAN, Y. & CHIA, K. S. (2010) The pursuit of genome-wide association studies: Where are we now? *Journal of Human Genetics*, 55, 195-206.
- KWOK, J. B. J., TEBER, E. T., LOY, C., HALLUPP, M., NICHOLSON, G., et al. (2004) Tau haplotypes regulate transcription and are associated with Parkinson's disease. *Annals of Neurology*, 55, 329-334.
- LANDE, R. & THOMPSON, R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124, 743-756.
- LESAGE, S. & BRICE, A. (2009) Parkinson's disease: From monogenic forms to genetic susceptibility factors. *Human Molecular Genetics*, 18, R48-R59.
- LEVECQUE, C., ELBAZ, A., CLAVEL, J., VIDAL, J. S., AMOUYEL, P., et al. (2004) Association of polymorphisms in the Tau and Saitohin genes with Parkinson's disease. *Journal of Neurology Neurosurgery and Psychiatry*, 75, 478-480.
- LORENZL, S., ALBERS, D. S., NARR, S., CHIRICHIGNO, J. & BEAL, M. F. (2002) Expression of MMP-2, MMP-9, and MMP-1 and their endogenous counterregulators TIMP-1 and TIMP-2 in postmortem brain tissue of Parkinson's disease. *Experimental Neurology*, 178, 13-20.
- LUO, Z. W. (1993) The power of 2 experimental-designs for detecting linkage between a marker locus and a locus affecting a quantitative character in a segregating population. *Genetics Selection Evolution*, 25, 249-261.
- LUO, Z. W. (1998) Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity*, 80, 198-208.
- LUO, Z. W., TAO, S. H. & ZENG, Z. B. (2000) Inferring linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Genetics*, 156, 457-467.

- LUO, Z. W. & WU, C. I. (2001) Modeling linkage disequilibrium between a polymorphic marker locus and a locus affecting complex dichotomous traits in natural populations. *Genetics*, 158, 1785-1800.
- MAMAH, C. E., LESNICK, T. G., LINCOLN, S. J., STRAIN, K. J., DE ANDRADE, M., et al. (2005) Interaction of Alpha-synuclein and Tau genotypes in Parkinson's disease. *Annals of Neurology*, 57, 439-443.
- MANTAMADIOTIS, T., LEMBERGER, T., BLECKMANN, S. C., KERN, H., KRETZ, O., et al. (2002) Disruption of CREB function in brain leads to neurodegeneration. *Nature Genetics*, 31, 47-54.
- MARAGANORE, D. M., DE ANDRADE, M., LESNICK, T. G., STRAIN, K. J., FARRER, M. J., et al. (2005) High-resolution whole-genome association study of Parkinson disease. *American Journal of Human Genetics*, 77, 685-693.
- MARIN, I. & FERRUS, A. (2002) Comparative genomics of the RBR family, including the Parkinson's disease-related gene parkin and the genes of the ariadne subfamily. *Molecular Biology and Evolution*, 19, 2039-2050.
- MARTIN, E. R., SCOTT, W. K., NANCE, M. A., WATTS, R. L., HUBBLE, J. P., et al. (2001) Association of single-nucleotide polymorphisms of the Tau gene with late-onset Parkinson disease. *JAMA-Journal of the American Medical Association*, 286, 2245-2250.
- MCCULLOCH, C. C., KAY, D. M., FACTOR, S. A., SAMII, A., NUTT, J. G., et al. (2008) Exploring gene-environment interactions in Parkinson's disease. *Human Genetics*, 123, 257-265.
- MOORE, J. H., ASSELBERGS, F. W. & WILLIAMS, S. M. (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26, 445-455.
- MUELLER, J. C., FUCHS, J., HOFER, A., ZIMPRICH, A., LICHTNER, P., et al. (2005) Multiple regions of alpha-synuclein are associated with Parkinson's disease. *Annals of Neurology*, 57, 535-541.
- MYHRE, R., TOFT, M., KACHERGUS, J., HULIHAN, M. M., AASLY, J. O., et al. (2008) Multiple alpha-synuclein gene polymorphisms are associated with Parkinson's disease in a Norwegian population. *Acta Neurologica Scandinavica*, 118, 320-327.
- PICCENNA, L., SHEN, P. J., MA, S., BURAZIN, T. C. D., GOSSEN, J. A., et al. (2005) Localization of LGR7 gene expression in adult mouse brain using LGR7 knock-out/LacZ knock-in mice - correlation with LGR7 mRNA distribution. IN SHERWOOD, O. D., FIELDS, P. A. & STEINETZ, B. G. (Eds.) *Relaxin and related peptides: Fourth international conference*. New York, New York Acad Sciences.
- PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A., et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38, 904-909.
- SASIENI, P. D. (1997) From genotypes to genes: Doubling the sample size. *Biometrics*, 53, 1253-1261.

- SCHAID, D. J. & JACOBSEN, S. J. (1999) Biased tests of association: Comparisons of allele frequencies when departing from Hardy-Weinberg proportions. *American Journal of Epidemiology*, 149, 706-711.
- SCOTT, W. K., NANCE, M. A., WATTS, R. L., HUBBLE, J. P., KOLLER, W. C., et al. (2001) Complete genomic screen in Parkinson disease - evidence for multiple genes. *JAMA-Journal of the American Medical Association*, 286, 2239-2244.
- SEARLE, S. R. (1971) *Linear models*, New York ; Chichester, Wiley.
- SHIMURA, H., HATTORI, N., KUBO, S., MIZUNO, Y., ASAKAWA, S., et al. (2000) Familial Parkinson disease gene product, parkin, is a ubiquitin-protein ligase. *Nature Genetics*, 25, 302-305.
- SIMON-SANCHEZ, J., SCHULTE, C., BRAS, J. M., SHARMA, M., GIBBS, J. R., et al. (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics*, 41, 1308.
- SKIPPER, L., WILKES, K., TOFT, M., BAKER, M., LINCOLN, S., et al. (2004) Linkage disequilibrium and association of MAPT h1 in Parkinson disease. *American Journal of Human Genetics*, 75, 669-677.
- STEEL, R. G. D. & TORRIE, J. H. (1960) *Principles and procedures of statistics, with special reference to the biological sciences*, New York ; London, McGraw-Hill.
- SUTHERLAND, G. T., HALLIDAY, G. M., SILBURN, P. A., MASTAGLIA, F. L., ROWE, D. B., et al. (2009) Do polymorphisms in the familial Parkinsonism genes contribute to risk for sporadic Parkinson's disease? *Movement Disorders*, 24, 833-838.
- TABANGIN, M. E., WOO, J. G. & MARTIN, L. J. (2009) The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc*, 3 Suppl 7, S41.
- VALOR, L. M., JANCIC, D., LUJAN, R. & BARCO, A. (2010) Ultrastructural and transcriptional profiling of neuropathological misregulation of CREB function. *Cell Death and Differentiation*, 17, 1636-1644.
- VANDROVCOVA, J., PITTMAN, A. M., MALZER, E., WOOD, N. W., LEES, A. J., et al. (2007) Association of MAPT haplotype-tagging SNPs with Parkinson's disease. *Movement Disorders*, 22, 416.
- WEIR, B. S. (1990) *Genetic data analysis : Methods for discrete population genetic data*, Sunderland, Mass., Sinauer Associates.
- WINKLER, S., KONIG, I. R., LOHMANN-HEDRICH, K., VIEREGGE, P., KOSTIC, V., et al. (2007) Role of ethnicity on the association of MAPT h1 haplotypes and subhaplotypes in Parkinson's disease. *European Journal of Human Genetics*, 15, 1163-1168.
- WRIGHT, S. (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 19, 0506-0536.
- YANG, J. A., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42, 565-U131.

- YATES, F. (1934) The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29, 51-66.
- YATES, F. (1948) The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, 35, 176-181.
- YU, J. M., PRESSOIR, G., BRIGGS, W. H., BI, I. V., YAMASAKI, M., et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38, 203-208.
- ZABETIAN, C. P., HUTTER, C. M., FACTOR, S. A., NUTT, J. G., HIGGINS, D. S., et al. (2007) Association analysis of MAPT h1 haplotype and subhaplotypes in Parkinson's disease. *Annals of Neurology*, 62, 137-144.
- ZAPPIA, M., ANNESI, G., NICOLETTI, G., SERRA, P., ARABIA, G., et al. (2003) Association of Tau gene polymorphism with Parkinson's disease. *Neurological Sciences*, 24, 223-224.
- ZHANG, J. & BOOS, D. D. (1997) Mantel-Haenszel test statistics for correlated binary data. *Biometrics*, 53, 1185-1198.

CHAPTER III

LIKELIHOOD-BASED METHODS

FOR ASSOCIATION STUDIES

CHAPTER III-1

INFERENCE OF LINKAGE DISEQUILIBRIUM IN NON-RANDOM SAMPLE

1.1 Related Publications

Wang, M., Jia, T., Jiang, N., Wang, L. & Luo, Z. Inferring Linkage Disequilibrium from Non-random Samples. **BMC Genomics**. 11, 328(2010)

1.2 Overview

As has been introduced in section I-1.2.3, linkage disequilibrium (LD) is the core concept in association studies and hence it is among the highest demands to properly infer LD coefficient between any two genetic loci. For random samples, Hill (1974) proposed a maximum likelihood based method to estimate the LD coefficient between two genetic loci given their diploid genotypes, which is the main driven idea of establishing the likelihood-based method for association studies (Luo and Suhai, 1999, Luo et al., 2000, Luo and Wu, 2001) and inferring the haplotypes of genetic markers (Long et al., 1995, Stephens et al., 2001). However, such random samples may not be available subject to many practical limitations. For example, certain genotypes might be missing due to the highly limited sample size or, more commonly, samples are collected by case and control for a certain trait in order to improve the statistical power. In such circumstances, simply adopting Hill's method might cause severe bias due to the non-random nature of the samples, and hence a method which could properly adjust such non-randomness is required.

In this chapter, a new maximum likelihood based method will be introduced to estimate the LD coefficient for non-random samples and its performance will be evaluated through both simulation studies and real data analysis.

1.3 Inferring LD in Random Samples (Hill's Method)

Before the introduction of our new method, we would like to have a brief review of Hill's method of inferring LD in random samples in order to give a comprehensive background image about the inference of LD.

Assume there are two biallelic loci, i.e. the marker locus with alleles M and m and disease locus with alleles A and a , with the allele frequencies of M and A denoted as p and q respectively. Following the definition of LD in section I-1.2.3 that $D = f_{MA} - p \times q$, where f_{MA} denotes the frequency of haplotype MA and so on, we could calculate the frequencies of haplotypes in terms of p , q and D as shown in Table III-1, and the corresponding joint distribution of marker and disease genotypes from a random mating population are presented in Table III-2.

Table III-1. Frequencies of haplotypes between marker and disease alleles

p and q denotes the allele frequency of maker allele M and disease allele A respectively

	M	m
A	$p \times q + D$	$(1 - p) \times q - D$
a	$p \times (1 - q) - D$	$(1 - p) \times (1 - q) + D$

Table III-2. Joint distribution of marker and disease genotypes in a randomly mated population

The frequencies of haplotypes are given in Table III-1.

	<i>MM</i>	<i>Mm</i>	<i>mm</i>
<i>AA</i>	f_{MA}^2	$2f_{MA}f_{mA}$	f_{mA}^2
<i>Aa</i>	$2f_{MA}f_{Ma}$	$2(f_{MA}f_{ma} + f_{Ma}f_{mA})$	$2f_{mA}f_{ma}$
<i>aa</i>	f_{Ma}^2	$2f_{Ma}f_{ma}$	f_{ma}^2

By letting g_{ij} denote the i - j th row-column entry of Table III-2, e.g. $g_{12} = 2f_{MA}f_{mA}$, and N_{ij} denote the corresponding number of observations, we could write the log-likelihood function of a multinomial distribution as:

$$l(p, q, D) \sim \sum_{i=1}^3 \sum_{j=1}^3 N_{ij} \log g_{ij} \quad (\text{III-1.1})$$

Since the unbiased estimator of p and q could be directly given as

$$\hat{p} = \frac{2 \sum_i N_{i1} + \sum_i N_{i2}}{2N}$$

and

$$\hat{q} = \frac{2 \sum_i N_{1i} + \sum_i N_{2i}}{2N},$$

where N is the full sample size, i.e. $N = \sum_{i,j} N_{ij}$, the estimator of D could be acquired through

solving equation:

$$\frac{\partial l(\hat{p}, \hat{q}, D)}{\partial D} = 0 \quad (\text{III-1.2})$$

Instead of directly solving equation (III-1.2), Hill (1974) suggested to calculate the estimator of f_{MA} instead, and then \hat{D} could be given as $\hat{f}_{MA} - \hat{p} \times \hat{q}$. In his paper, Hill proposed an iterative process to approach \hat{f}_{MA} by updating

$$\hat{f}_{11}^{(s+1)} = \left\{ X_{11} + N_{22} \hat{f}_{11}^{(s)} (1 - \hat{p} - \hat{q} + \hat{f}_{11}^{(s)}) / \left[\hat{f}_{11}^{(s)} (1 - \hat{p} - \hat{q} + \hat{f}_{11}^{(s)}) + \left(\hat{p} - \hat{f}_{11}^{(s)} \right) \times \left(\hat{q} - \hat{f}_{11}^{(s)} \right) \right] \right\} / 2N \quad (\text{III-1.3})$$

until a preset threshold δ is reached at the t th iteration, i.e. $\delta = \hat{f}_{11}^{(t)} - \hat{f}_{11}^{(t-1)}$, where $\hat{f}_{ij}^{(s)}$ denotes the estimator of the haplotype frequency with the i th disease genotype and the j th marker genotype in the s th iteration, e.g. $\hat{f}_{11}^{(s)} = \hat{f}_{MA}^{(s)}$, and $X_{11} = 2N_{11} + N_{21} + N_{12}$. A suitable starting point, as suggested by Hill, could be estimated by replacing n_{22} with $2(\hat{f}_{11}\hat{f}_{22} + \hat{f}_{12}\hat{f}_{21})N$ in formula (III-1.3), i.e. n_{22} equals its expectation given \hat{f}_{ij} , $i, j = 1, 2$, and we could have

$$\hat{f}_{11}^{(0)} = \frac{1}{4N}(X_{11} - X_{12} - X_{21} + X_{22}) + \frac{1}{2} - (1 - \hat{p})(1 - \hat{q}),$$

where $X_{12} = 2N_{13} + N_{12} + N_{23}$, $X_{21} = 2N_{31} + N_{21} + N_{32}$ and $X_{22} = 2N_{33} + N_{32} + N_{23}$. However, Weir & Cockerham (1979) later pointed out that the estimator of f_{11} , i.e. \hat{f}_{11} could be directly given as the root of equation (III-1.4), i.e. a polynomial equation of power 3, to avoid above iterative processes:

$$b_3 f_{11}^3 + b_2 f_{11}^2 + b_1 f_{11} + b_0 = 0 \quad (\text{III-1.4})$$

where

$$b_3 = 2N, \quad b_2 = 2N(1 - 2\hat{p} - 2\hat{q}) - 2(2N_{11} + N_{12} + N_{21}) - N_{22},$$

$$b_1 = 2N\hat{p}\hat{q} - (2N_{11} + N_{12} + N_{21})(1 - 2\hat{p} - 2\hat{q}) - N_{22}(1 - \hat{p} - \hat{q}),$$

and

$$b_0 = -(2N_{11} + N_{12} + N_{21})\hat{p}\hat{q}.$$

Nevertheless, as both strategies should give theoretically identical results, in the following sections, the method introduced by Hill will be referred to as Method H.

1.4 A Likelihood-based Method for Inferring LD from Selected Samples

Since Hill's method introduced above requires the knowledge of population-based joint distribution of two loci as given in Table III-2, such a method will almost surely suffer certain biases if the information about the population is incomplete. For instance, Weir & Cockerham (1979) evaluated Hill's method in a sample with part of its genotypes being missing, and they pointed out that it should not be used to estimate LD from non-random samples. Such a conclusion could be understood by realising that if the data are collected to ensure a rare allele, i.e. the disease causal allele, to be present in the sample, the distribution of genotypes at the disease locus will be significantly distinguishable from its distribution in the population, and hence a severe bias of the estimate of LD could be expected. However, although the joint distribution of two loci in a population is not available if the data are collected non-randomly, e.g. a Case-Control sample, the conditional distribution between these two loci might still be available. For instance, the conditional distribution of disease genotypes given marker genotypes could be given in Table III-3, where $Q = q + D/p$, $R = q - D/(1-p)$ and n_{ij} denotes the number of observations with the i th marker genotype and the j th disease genotypes.

Table III-3. Conditional Distribution of Disease Genotypes upon Marker Genotypes

$Q = q + D/p$ and $R = q - D/(1-p)$

<i>MM</i>			<i>Mm</i>			<i>Mm</i>		
<i>AA</i>	<i>Aa</i>	<i>aa</i>	<i>AA</i>	<i>Aa</i>	<i>aa</i>	<i>AA</i>	<i>Aa</i>	<i>aa</i>
Q^2	$2Q(1-Q)$	$(1-Q)^2$	QR	$Q + R - 2QR$	$(1-Q)(1-R)$	R^2	$2R(1-R)$	$(1-R)^2$
n_{11}	n_{12}	n_{13}	n_{21}	n_{22}	n_{23}	n_{31}	n_{32}	n_{33}

Let f_{ij} denote the expected frequency of the observation n_{ij} , it would be practical to write the conditional log-likelihood function in a multinomial distribution as

$$l(p, q, D) \sim \sum_{i=1}^3 \sum_{j=1}^3 n_{ij} \log f_{ij} . \quad (\text{III-1.5})$$

The corresponding MLE of D could hence be given through solving equation

$$\frac{\partial}{\partial D} l(p, q, D) = \sum_{i=1}^3 \sum_{j=1}^3 \frac{n_{ij}}{f_{ij}} \frac{\partial}{\partial D} (f_{ij}) = 0 , \quad (\text{III-1.6})$$

which could then be written into a polynomial equation of D with power 5 as

$$a_5 D^5 + a_4 D^4 + a_3 D^3 + a_2 D^2 + a_1 D + a_0 = 0 , \quad (\text{III-1.7})$$

where

$$a_5 = 4n$$

$$a_4 = 6n_{11} + 6n_{12} + 6n_{13} + 6n_{21} + 5n_{22} + 4n_{23} + 6n_{31} + 4n_{32} + 2n_{33} + (-12n_{11} - 12n_{12} - 12n_{13} - 10n_{21} - 10n_{22} - 10n_{23} - 8n_{31} - 8n_{32} - 8n_{33})q - p[(12n_{11} + 10n_{12} + 8n_{13} + 12n_{21} + 10n_{22} + 8n_{23} + 12n_{31} + 10n_{32} + 8n_{33} - (20n_{11} + 20n_{12} + 20n_{13} + 20n_{21} + 20n_{22} + 20n_{23} + 20n_{31} + 20n_{32} + 20n_{33})q]$$

$$a_3 = 2n_{11} + 2n_{12} + 2n_{13} + 2n_{21} + n_{22} + n_{23} + 2n_{31} + n_{32} - (12n_{11} + 12n_{12} + 12n_{13} + 9n_{21} + 8n_{22} + 7n_{23} + 6n_{31} + 4n_{32} + 2n_{33})q + (12n_{11} + 12n_{12} + 12n_{13} + 8n_{21} + 8n_{22} + 8n_{23} + 4n_{31} + 4n_{32} + 4n_{33})q^2 - p[12n_{11} + 9n_{12} + 6n_{13} + 12n_{21} + 8n_{22} + 4n_{23} + 12n_{31} + 7n_{32} + 2n_{33} - (54n_{11} + 48n_{12} + 42n_{13} + 48n_{21} + 40n_{22} + 32n_{23} + 42n_{31} + 32n_{32} + 22n_{33})q - (48n_{11} + 48n_{12} + 48n_{13} + 40n_{21} + 40n_{22} + 40n_{23} + 32n_{31} + 32n_{32} + 32n_{33})q^2] + p^2[12n_{11} + 8n_{12} + 4n_{13} + 12n_{21} + 8n_{22} + 4n_{23} + 12n_{31} + 8n_{32} + 4n_{33} - (48n_{11} + 40n_{12} + 32n_{13} + 48n_{21} + 40n_{22} + 32n_{23} + 48n_{31} + 40n_{32} + 32n_{33})q + (40n_{11} + 40n_{12} + 40n_{13} + 40n_{21} + 40n_{22} + 40n_{23} + 40n_{31} + 40n_{32} + 40n_{33})q^2]$$

$$\begin{aligned}
a_2 = & -(2n_{11} + 2n_{12} + 2n_{13} + n_{21} + n_{22} + n_{23})q + \\
& 3(2n_{11} + 2n_{12} + 2n_{13} + n_{21} + n_{22} + n_{23})q^2 - 2(2n_{11} + 2n_{12} + 2n_{13} + n_{21} + n_{22} + n_{23})q^3 - \\
& p[2n_{11} + n_{12} + 2n_{21} + n_{22} + 2n_{31} + n_{32} + \\
& (26n_{11} + 20n_{12} + 14n_{13} + 20n_{21} + 14n_{22} + 8n_{23} + 14n_{31} + 8n_{32} + 2n_{33})q - \\
& (60n_{11} + 54n_{12} + 48n_{13} + 42n_{21} + 36n_{22} + 30n_{23} + 24n_{31} + 18n_{32} + 12n_{33})q^2 + \\
& (36n_{11} + 36n_{12} + 36n_{13} + 24n_{21} + 24n_{22} + 24n_{23} + 12n_{31} + 12n_{32} + 12n_{33})q^3] \\
& + p^2[6n_{11} + 3n_{12} + 6n_{21} + 3n_{22} + 6n_{31} + 3n_{32} - \\
& (60n_{11} + 42n_{12} + 24n_{13} + 54n_{21} + 36n_{22} + 18n_{23} + 48n_{31} + 30n_{32} + 12n_{33})q + \\
& (126n_{11} + 108n_{12} + 90n_{13} + 108n_{21} + 90n_{22} + 72n_{23} + 90n_{31} + 72n_{32} + 54n_{33})q^2 - \\
& (72n_{11} + 72n_{12} + 72n_{13} + 60n_{21} + 60n_{22} + 60n_{23} + 48n_{31} + 48n_{32} + 48n_{33})q^3] - \\
& p^3[4n_{11} + 2n_{12} + 4n_{21} + 2n_{22} + 4n_{31} + 2n_{32} - \\
& (36n_{11} + 24n_{12} + 12n_{13} + 36n_{21} + 24n_{22} + 12n_{23} + 36n_{31} + 24n_{32} + 12n_{33})q + \\
& (72n_{11} + 60n_{12} + 48n_{13} + 72n_{21} + 60n_{22} + 48n_{23} + 72n_{31} + 60n_{32} + 48n_{33})q^2 - \\
& (40n_{11} + 40n_{12} + 40n_{13} + 40n_{21} + 40n_{22} + 40n_{23} + 40n_{31} + 40n_{32} + 40n_{33})q^3]
\end{aligned}$$

$$\begin{aligned}
a_1 = & p[(2n_{11} + n_{12} + n_{21} + n_{22})q + (-12n_{11} - 9n_{12} - 6n_{13} - 6n_{21} - 5n_{22} - 3n_{23})q^2 + \\
& (18n_{11} + 16n_{12} + 14n_{13} + 9n_{21} + 8n_{22} + 7n_{23})q^3 + (-8n_{11} - 8n_{12} - 8n_{13} - 4n_{21} - \\
& 4n_{22} - 4n_{23})q^4] - p^2[(12n_{11} + 6n_{12} + 9n_{21} + 5n_{22} + 6n_{31} + 3n_{32})q - \\
& (60n_{11} + 42n_{12} + 24n_{13} + 42n_{21} + 29n_{22} + 15n_{23} + 24n_{31} + 15n_{32} + 6n_{33})q^2 + \\
& (84n_{11} + 72n_{12} + 60n_{13} + 57n_{21} + 48n_{22} + 39n_{23} + 30n_{31} + 24n_{32} + 18n_{33})q^3 - \\
& (36n_{11} + 36n_{12} + 36n_{13} + 24n_{21} + 24n_{22} + 24n_{23} + 12n_{31} + 12n_{32} + 12n_{33})q^4] + \\
& p^3[(18n_{11} + 9n_{12} + 16n_{21} + 8n_{22} + 14n_{31} + 7n_{32})q - \\
& (84n_{11} + 57n_{12} + 30n_{13} + 72n_{21} + 48n_{22} + 24n_{23} + 60n_{31} + 39n_{32} + 18n_{33})q^2 + \\
& (114n_{11} + 96n_{12} + 78n_{13} + 96n_{21} + 80n_{22} + 64n_{23} + 78n_{31} + 64n_{32} + 50n_{33})q^3 - \\
& (48n_{11} + 48n_{12} + 48n_{13} + 40n_{21} + 40n_{22} + 40n_{23} + 32n_{31} + 32n_{32} + 32n_{33})q^4] - \\
& p^4[(8n_{11} + 4n_{12} + 8n_{21} + 4n_{22} + 8n_{31} + 4n_{32})q - \\
& (36n_{11} + 24n_{12} + 12n_{13} + 36n_{21} + 24n_{22} + 12n_{23} + 36n_{31} + 24n_{32} + 12n_{33})q^2 + \\
& (48n_{11} + 40n_{12} + 32n_{13} + 48n_{21} + 40n_{22} + 32n_{23} + 48n_{31} + 40n_{32} + 32n_{33})q^3 - \\
& (20n_{11} + 20n_{12} + 20n_{13} + 20n_{21} + 20n_{22} + 20n_{23} + 20n_{31} + 20n_{32} + 20n_{33})q^4]
\end{aligned}$$

$$\begin{aligned}
a_0 = & (1-p)^2 p^2 (1-q)^2 q^2 [2n_{21} + n_{22} - 4n_{21}p - 2n_{22}p - 4n_{31}p - 2n_{32}p + \\
& 4n_{11}(1-p)(1-q) - 4n_{13}q - 2n_{21}q - 2n_{22}q - 2n_{23}q + \\
& 4n_{13}pq + 4n_{21}pq + 4n_{22}pq + 4n_{23}pq + 4n_{31}pq + 4n_{32}pq + 4n_{33}pq + 2n_{12}(1-p)(1-2q)]
\end{aligned}$$

Although no analytical solution of D is available for equation (III-1.7), it could still be solved numerically. If there is more than one real root, certain criteria will be applied to distinguish the most appropriate estimator from the rest, i.e. D must fall within $[\max(-pq, -(1-p)(1-q)), \min(p(1-q), q(1-p))]$ and the corresponding LRTs could be compared among these candidate MLEs, where the most possible candidate should yield the highest test statistic

$$2[l(p, q, D) - l(p, q, D = 0)] \sim \chi^2_{df=1},$$

which asymptotically follows χ^2 distribution with 1 degree of freedom.

Note here, due to the symmetrical relationship between marker and disease loci, solving equation (III-1.6) with interchangeable p and q will yield the estimator of D given the distribution of marker genotypes conditioned on disease genotypes. It should also be noted that if the sample is collected non-randomly, the unbiased estimators of p and q , such as given in Hill's method, are not available from a sample hence collected. In such a circumstance, we could either implement biased estimations from the sample or acquire unbiased estimations from other sources. Different strategies of dealing with this issue will be evaluated in the next section. In the following sections, our newly proposed method will be referred to as Method L for convenience and comparison

1.5 Simulation Studies

In simulation studies, three schemes of sampling are conducted, where samples in Scheme I are randomly collected from an ideal Mendelian population; samples in Scheme II are randomly collected similarly as Scheme I but with one or several genotypes completely missing; samples in Scheme III are collected as the case and control individuals subject to genotypes of the disease locus. For convenience, the same notation as in Table III-1 will be

adopted in this section, and as a similar simulation procedure was introduced in section II-3.3.3, we will not be bothered to mention the details of such processes unless necessary.

Table III-4. Summary of Estimates of D for Scheme I.

All 12 simulations with 200 individuals are randomly sampled from a certain population with marker allele frequency p , disease allele frequency q and LD coefficient D . D_{min} and D_{max} represent the theoretical lower and upper bound of the LD coefficient, and the estimates of LD coefficients by Method H and L are represented as \hat{D}_H and \hat{D}_L .

Pop.	P	q	(D_{min}, D_{max})	D	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$
1	0.5	0.5	(-0.25, 0.25)	0.20	0.1999±0.0078	0.2004±0.0078
2	0.5	0.5	(-0.25, 0.25)	0.10	0.1002±0.0145	0.1003±0.0145
3	0.3	0.3	(-0.09, 0.21)	0.09	0.0898±0.0133	0.0899±0.0125
4	0.7	0.7	(-0.09, 0.21)	0.09	0.0895±0.0133	0.0896±0.0126
5	0.3	0.5	(-0.15, 0.15)	0.10	0.0997±0.0120	0.0998±0.0111
6	0.5	0.3	(-0.15, 0.15)	0.10	0.0995±0.0121	0.0993±0.0109
7	0.5	0.5	(-0.25, 0.25)	-0.20	-0.1995±0.0081	-0.1998±0.0081
8	0.5	0.5	(-0.25, 0.25)	-0.10	-0.0996±0.0146	-0.0997±0.0146
9	0.3	0.3	(-0.09, 0.21)	-0.09	-0.0896±0.0074	-0.0899±0.0068
10	0.7	0.7	(-0.09, 0.21)	-0.09	-0.0897±0.0073	-0.0899±0.0065
11	0.3	0.5	(-0.15, 0.15)	-0.10	-0.1000±0.0124	-0.1000±0.0117
12	0.5	0.3	(-0.15, 0.15)	-0.10	-0.0995±0.0120	-0.0993±0.0111

Method H: Hill's Method (Section III-1.3)

Method L: Likelihood-based Method (Section III-1.4)

In Scheme I, 12 populations are considered with their parameters p , q and D listed in Table III-4. From each population, we simulated a random sample with 200 individuals, and as the unbiased estimator of p and q could be directly estimated from the simulation sample, we could hence easily compute the MLEs of D from both Hill's method (Method H) and our newly proposed one (Method L). Such a sampling was replicated for 1000 times for each population and we summarize the corresponding results, i.e. means and standard deviations of \hat{D} , in Column $\hat{D}_H \pm s.d.$ for Method H and Column $\hat{D}_L \pm s.d.$ for Method L of Table III-4

and henceforward. From these results, we could notice that both methods have shown adequate estimates of D , where, however, Method L has slightly smaller standard deviations than Method H in several simulations. The results of Scheme I hence show that for random samples, Method L is as efficient as Method H, if not better.

In Scheme II, we implemented the first 6 populations as shown in Table III-4 with exactly the same sequences to simulate our samples. For each of these 6 populations, individuals were generated similar to the random sampling, but excluding certain genotypes, for example, $n_{i\cdot} = 0, i = 1, 2, 3$ corresponding to MM, Mm or mm and $n_{ii} = 0, i = 1, 2, 3$ corresponding to $MMAA, MmAa$ or $mmaa$ as shown in Table II-5. Similar to Scheme I, a sample with 200 individuals was generated and we replicated such a sample for 1000 times for each population. The corresponding means and standard deviations of \hat{D} of both methods are also presented in Table II-5, from which we could notice that, for $n_{i\cdot} = 0, i = 1, 2, 3$, where a marker genotype is completely missing from the sample, Method L shows better performance than Method H, especially when the heterozygote is missing, i.e. $n_{2\cdot} = 0$, where Method H suffers significant biases downwards in populations 3, 4 and 5, and for the rest part, where $n_{ii} = 0, i = 1, 2, 3$, i.e. a certain joint marker-disease genotype is missing, both methods suffer certain loss of accuracy, but Method L still retains better performance in all set up, especially for the cases $n_{33} = 0$ in population 3 and $n_{11} = 0$ in population 4. Such results show that Method L could substantially increase the accuracy of estimating the LD coefficient while certain genotypes are missing from a sample, e.g. under a strong purifying selection. Note here, both p and q used to compute D in both methods are directly estimated from the above samples despite their biases.

Table III-5. Summary of Estimates of D from Scheme II.

The simulated populations are identical with that of the same series number in Table III-4. For each population, 200 individuals are randomly collected but with certain genotype or genotypes missing and each sample such collected is replicated 1000 times. The estimates of LD coefficients by Method H and L are represented as \hat{D}_H and \hat{D}_L respectively.

Pop.	D	$n_{1..} = 0$		$n_{2..} = 0$		$n_{3.} = 0$		$n_{11} = 0$		$n_{22} = 0$		$n_{33} = 0$	
		$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$
1	0.20	0.18±0.01	0.20±0.03	0.20±0.01	0.20±0.01	0.18±0.01	0.20±0.03	0.17±0.01	0.18±0.04	0.17±0.01	0.17±0.01	0.17±0.01	0.18±0.04
2	0.10	0.09±0.02	0.10±0.02	0.10±0.01	0.10±0.01	0.09±0.02	0.10±0.02	0.05±0.02	0.06±0.02	0.07±0.01	0.07±0.01	0.05±0.02	0.06±0.02
3	0.09	0.08±0.01	0.09±0.02	0.06±0.01	0.10±0.04	0.10±0.02	0.09±0.01	0.07±0.01	0.08±0.01	0.04±0.01	0.09±0.06	0.00±0.02	0.04±0.01
4	0.09	0.10±0.02	0.09±0.01	0.06±0.01	0.09±0.01	0.08±0.01	0.09±0.01	0.00±0.02	0.04±0.02	0.04±0.01	0.06±0.01	0.07±0.01	0.08±0.01
5	0.10	0.08±0.01	0.10±0.01	0.06±0.01	0.10±0.01	0.12±0.01	0.10±0.01	0.07±0.01	0.08±0.01	0.07±0.01	0.08±0.01	0.05±0.02	0.06±0.02
6	0.10	0.09±0.01	0.10±0.01	0.10±0.01	0.10±0.01	0.09±0.01	0.10±0.01	0.07±0.01	0.08±0.01	0.07±0.01	0.08±0.01	0.05±0.02	0.06±0.02

Method H: Hill's Method (Section III-1.3)

Method L: Likelihood-based Method (Section III-1.4)

In Scheme III, samples were generated as in Case-Control studies. To mimic the reality, the disease allele A is assumed with a low frequency, i.e. less than 0.1, and the dominance model was adopted such that an individual with either AA or Aa was identified as a case and the other individuals, i.e. with genotype aa , are grouped as controls. For each simulated sample, 100 cases and 100 controls were collected, which yields a total of 200 individuals constantly. Clearly, the distribution of disease genotypes in such a Case-Control sample should severely deviate from the HWE, and as has been mentioned in the previous section III-1.4, such a non-randomness would hinder our attempt to find unbiased estimators of p and q . To tackle such an obstacle, in Scheme III, p is directly estimated from the control individuals, which is reasonable if q is small as indicated in section II-3.2.3, and q will be acquired from two sources: Firstly, the real value of q will be implemented, which is available if extra resources are available, e.g. results from previous epidemiological studies; Secondly, the q could still be estimated from the observations directly. However, due to the severe deviation from HWE of such a Case-Control sample, the estimate of q hence calculated would be expected to cause considerable loss of accuracy in estimating LD coefficient for both methods.

The corresponding parameters and results of Scheme III are all listed in Table III-6, where the mean and standard deviation of samples simulated from each population are calculated based on 1000 replicates. From Table III-6, we may notice that, even when the real value of q is adopted, the implementation of Method H results in severely biased estimates of D , which, as shown for \hat{D}_H , will always exceed the theoretical boundaries, but Method L yields almost perfect estimates. If q is directly estimated from the observations, i.e. the Case-Control sample are treated as a random sample, both methods suffer severe loss of accuracy in estimating D as shown in Table III-6. Nevertheless, the estimate from \hat{D}_L still deviates much less from its actual value than that from \hat{D}_H and hence Method L is still favoured in such a circumstance.

Note here, the reason why Method H would yield even worse estimates of D in the use of real q than the use of its biased estimate can be explained by the fact that the estimates of haplotypes, e.g. f_{MA} , with real q might exceed their theoretical boundaries in Scheme III.

Table III-6. Summary of Estimates of D from Scheme III.

Parameters are the same in Table III-4 and Table III-5.

p	q	(D_{min}, D_{max})	D	q was from population survey		q was from sample estimation	
				$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$	$\hat{D}_H \pm s.d.$	$\hat{D}_L \pm s.d.$
0.6	0.005	(-0.003, 0.002)	-0.002	-0.011±0.149	-0.002±0.000	-0.113±0.015	-0.071±0.011
0.5	0.01	(-0.005, 0.005)	0.004	0.280±0.011	0.004±0.001	0.108±0.013	0.080±0.013
0.5	0.02	(-0.010, 0.010)	0.008	0.273±0.010	0.008±0.001	0.110±0.014	0.081±0.013
0.3	0.03	(-0.009, 0.021)	0.010	0.191±0.026	0.011±0.002	0.104±0.019	0.057±0.018
0.7	0.04	(-0.028, 0.012)	0.010	0.309±0.016	0.011±0.002	0.071±0.012	0.061±0.017
0.3	0.05	(-0.015, 0.035)	0.020	0.192±0.044	0.021±0.004	0.122±0.017	0.066±0.017
0.5	0.10	(-0.050, 0.050)	0.040	0.227±0.008	0.045±0.006	0.124±0.014	0.088±0.012

 D_H for Method H: Hill's Method (Section III-1.3) D_L for Method L: Likelihood-based Method (Section III-1.4)

1.6 Real Data Analysis

1.6.1 β -Thalassemia Dataset

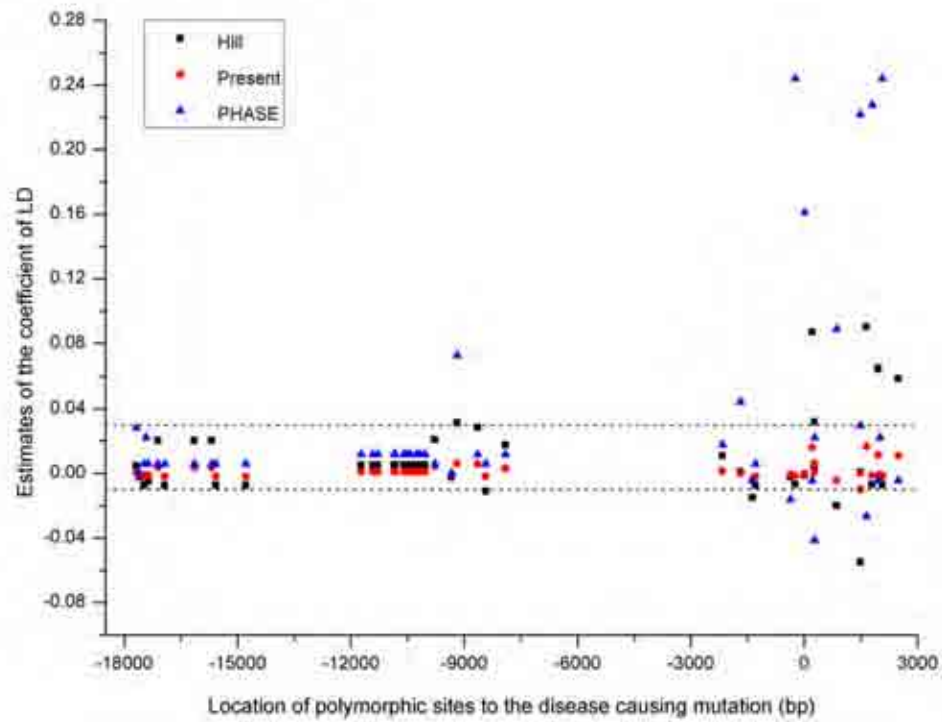
B-Thalassemia is an autosomal recessive blood disorder induced by mutations in the β -globin (HBB) gene (OMIM 141900) on the human chromosome 11, and is among the most common inherited hemoglobinopathies in the world. This disorder has been reported to affect 3% to 10% population of certain tropical and subtropical areas such as South China (Weatherall and Clegg, 2001, Xu et al., 2004, Zhang et al., 2008). In East and Southeast Asia, the most common type of β -Thalassemia, $\beta^{\text{CD41/42}}$ -Thalassemia, is caused by a frame shift mutation in codons 41 and 42, a 4-bp deletion (-CTTT), of the human β -globin gene, of which the frequency could reach 3% in South China (Zhang et al., 2008).

In Zhang et al. (2008), 16 cases with $\beta^{\text{CD41/42}}$ -Thalassemia and 24 controls were collected from China, where only the heterozygotes at the deletion locus were present due to the highly limited sampling size. From their study, 50 bi-allelic markers were located after sequencing the neighbouring area (~20kb) around the deletion locus, all of which were genotyped in all 40 individuals. Since such a dataset is collected highly non-randomly from the population, i.e. not only the data is collected as Case-Control scheme, but also the homozygote deletion is completely missing at the disease locus, it would be a perfect example for us to evaluate the capability of different methods in estimating LD coefficients from non-random samples.

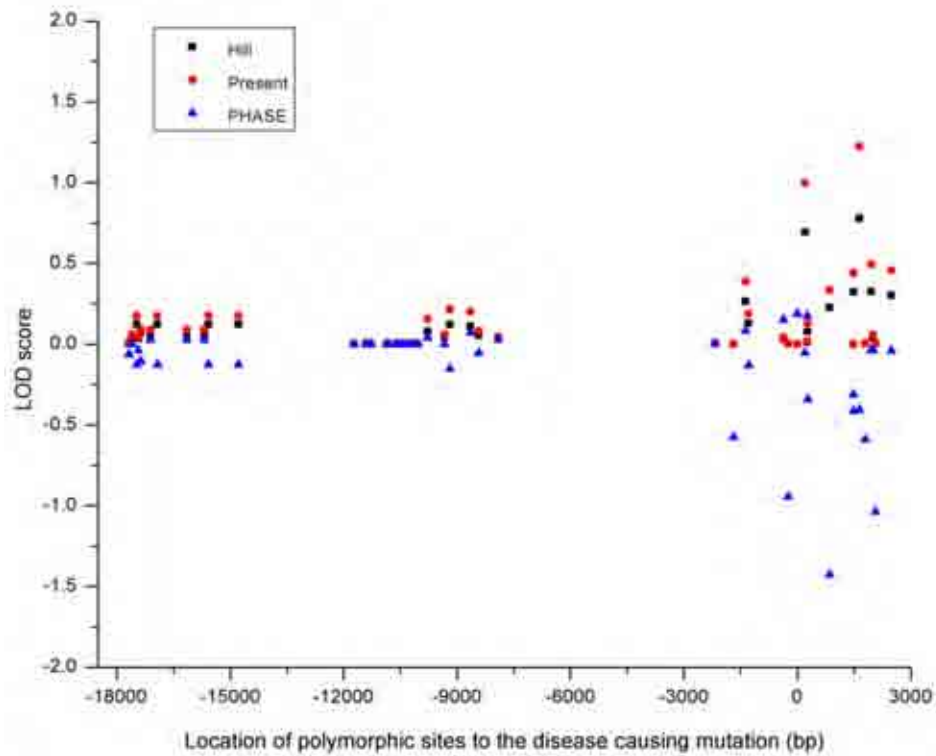
1.6.2 Comparison of Three Methods

As the 50 markers are of high density in a 20kb chromosome region, it would be practical to infer the haplotypes of each individual with the software PHASE 2.1.1 (Stephens et al., 2001), and we may hence directly calculate the LD coefficients with the frequencies of haplotypes. Along with the results acquired through predicting haplotypes by making use of PHASE 2.1.1,

both Methods H and L are also implemented to estimate the LD coefficient between the disease locus and each of 50 marker loci. By implementing the population frequency of the deletion allele as 3% as introduced in section III-1.6.2, the estimates of LD coefficients between the disease locus, located at 0 on x axis, and each marker locus are illustrated in Figure III-1 (a), where the theoretical upper and lower bounds for LD coefficients for any pairs of the disease and marker loci are also presented as the dashed lines. We may easily distinguish the newly proposed method, Method L, from the other two in Figure III-1 (a), especially for those markers closely linked to the disease locus, of which the estimates of LD coefficients from Method H and PHASE severely violate the theoretical upper and lower bounds. Figure III-1 (b) illustrates the corresponding LOD scores of \hat{D} calculated from three methods on base of likelihood function (III-1.5), and we could notice that as Method L will always yield the highest LOD score values, it is hence preferred to the other two. Clearly, the newly proposed method is sufficiently robust that the LD coefficients could be properly estimated from even such a non-random sample.



(a)



(b)

Figure III-1. Distribution of linkage disequilibrium between each of polymorphic sites and the β -thalassemia causing mutation in a 20.693 kb region surrounding the human β -globin gene. (a) Estimates of the coefficients of LD from three different methods. The dot lines represent the lowest and highest theoretical bounds of the disequilibrium parameter. (b) The LOD score values calculated for the LD estimates from the three methods.

1.7 Conclusion and Discussion

We have demonstrated that a severely biased estimate of LD coefficient could arise if methods designed for random samples, e.g. Hill's method (Hill, 1974), are implemented for a non-random sample. As shown in our simulation studies, such a biased estimate could be either positive or negative and even exceed the theoretical boundaries, which hence suggests the non-randomness might cause both false positive and negative in the estimate of LD coefficients. To tackle such a bias introduced by the non-randomness, we proposed a likelihood model based on the conditional distribution between the marker and disease loci, instead of the joint distribution of loci used in the traditional method (Hill, 1974, Weir and Cockerham, 1979). Through our comprehensive simulation studies, we have shown that our method could not only properly estimate LD coefficients in random samples as the traditional method, i.e. Hill's (1974), being capable of (Table III-4), but also provide significant improvement of estimation accuracy if the data is artificially collected and thus non-random (Table III-5, Table III-6). However, as we have shown in Table III-6, the performance of the newly proposed method relies on the proper estimates of allele frequencies for both marker and disease locus, which are not always available from non-randomly collected data. For instance, in a Case-Control sample, we could estimate marker allele frequency directly from the control sample, which could be used as a reasonable approximate for the population parameter, especially if the disease causal allele frequency is reasonably small, although an appropriate estimate for disease allele frequency is usually not available in such a circumstance. However, instead of directly computing the estimates of allele frequencies from the collected sample, we might acquire information from other sources, e.g. previous epidemiological analyses. For instance, we estimated the LD coefficients between any pairs of marker-disease loci from the β -Thalassemia data with the disease allele frequency acquired from a published report (Zhang et al., 2008), and our newly proposed method has shown an

overall out-performance over the other two, i.e. Hill's method and the calculation of LD coefficient based on the haplotype estimated from PHASE which was implemented in Zhang et al. (2008).

Of course, a direct estimate of LD coefficient between marker and putative disease loci is only available if the disease is a Mendelian trait or has a major effective gene (Cardon and Palmer, 2003, Yu and Buckler, 2006). For instance, in our study, we have assumed that genotypes at both marker and disease loci are observable, which is true for a Mendelian trait, but not so assured in the case of a quantitative trait. However, as has been revealed by several previous studies (Luo, 1998, Luo et al., 2000, Luo and Wu, 2001), it is possible to further integrate our newly proposed method and the quantitative genetic models as introduced in section II-1.2 to estimate the LD coefficient between a marker and a quantitative trait showing either continuous or dichotomous phenotype in non-random samples. Following such an idea, a likelihood-based method of inferring LD between marker and putative disease loci in a Case-Control sample will be introduced in Chapter III-2.

CHAPTER III-2

A LIKELIHOOD-BASED METHOD FOR ASSOCIATION STUDY IN CASE CONTROL SAMPLES FROM MULTIPLE COHORTS

2.1 Related Publication

This paper has been written up and is under submission. As a senior author, I have contributed substantially to the establishment of both the likelihood-based method and the modified allelic analysis, as well as their simulation studies. I have acquired the permission of all other authors to cite this paper (with some necessary rearrangement) as part of my thesis.

2.2 Overview

Although it has long been pointed out that the use of non-randomly collected samples might result in the spurious linkage disequilibrium between un-associated genetic loci, i.e. in linkage equilibrium (Avery and Hill, 1979), such an effect has not been paid a lot attention to until recently we investigated the use of non-randomly collected samples in estimating LD coefficients. We observed that such an estimate from a method established for random samples will be severely biased and the corresponding statistical power for testing for significance of LD coefficient may be substantially reduced (Wang et al., 2010). In practice, as in many association studies, sampling schemes are subject to various types of selection, a sample hence collected is no longer a random presentation of the corresponding population. A typical example of such non-random samples is the Case-Control sample as used in many association studies in human populations, in which frequencies of some disease genotypes are

artificially inflated in comparison to the population frequencies to ensure sufficient representation of the genotypes involving a rare disease allele.

In Chapter III-1, we have introduced a likelihood-based method to properly infer LD coefficients between marker and disease loci with known genotypes using non-random samples, and such a method has been comprehensively evaluated through both simulation studies and real data analyses to show its overall out-performance over the existing methods, e.g. Hill's method (Hill, 1974). By combining the idea of estimating LD coefficients in non-random samples with the genetic models as introduced in section II-1.2. In this chapter, we will propose a new likelihood-based method for association studies in Case-Control samples. Besides, as such a method could allow the population structure information to be included, for the sake of comparison, an adjusted allelic analysis will also be introduced. We will hence evaluate the newly proposed method (Method I) by comparing it to both adjusted allelic analysis (Method II) and the Armitage's trend test (Method III) in our simulation studies. Finally, we will implement all three methods to model and analyse the Parkinson's disease data, which has already been introduced and analysed in Chapter II-3.

2.3 Notations and Models

Consider a case-control sample of size n is collected from a random mating population with regard to a disease causal allele A and the wild type a . By denoting the marker alleles as M and m , we may write the conditional distributions between marker genotypes and disease genotypes as shown in Table III-7 (a) and (b), where p and q are the frequencies of allele M and A respectively, and D is the coefficient of LD. Clearly, in Table III-7 (a) and (b), $g_{ij} = \Pr(M = j | G = i)$ and $h_{ij} = \Pr(G = i | M = j)$, where M and G represent the corresponding marker and disease genotypes. By assuming that, in the presence of disease

genotypes, the marker genotypes cannot provide any further information, we could hence calculate that

$$\begin{aligned}
 \Pr(M = i | G = j, Case) &= \frac{\Pr(M = i, G = j, Case)}{\Pr(G = j, Case)} \\
 &= \frac{\Pr(Case | G = j) \Pr(G = j, M = i)}{\Pr(Case | G = j) \Pr(G = j)}, \quad (III-2.1) \\
 &= \Pr(M = i | G = j) = g_{ij}
 \end{aligned}$$

where $\Pr(Case | G = j)$ is the probability of an individual with the j th disease genotype being observed with disease exposure, which is normally denoted as the penetrance coefficient f_j . Formula (III-2.1) shows a conditional distribution of marker genotypes on disease genotypes should be identical in a case sample and in the corresponding randomly collected sample. As a similar result for a control sample could be easily acquired following the same procedure, we may hence work out the probability distributions of marker genotypes conditioned on disease genotypes in both case and control samples as presented in Table III-7 (c).

Similarly to function (III-1.5), it would be straightforward to write the conditional likelihood function into the form

$$L \sim \prod_{k=1}^2 \prod_{i=1}^3 \prod_{j=1}^3 g_{ijk}^{t_{ijk}}, \quad (III-2.2)$$

where t_{ijk} denotes the number of individuals with the i th disease genotype and the j th marker genotype from the k th sample ($k = 1$ denotes case and $k = 2$ denotes control). However, as the exact values of t_{ijk} in function (III-2.2) are not observable, estimates are hence required.

Table III-7. Conditional Distribution between Marker and QTL Genotypes.

(a) Conditional distribution of maker genotypes over disease genotypes; (b) Conditional distribution of disease genotypes over marker genotypes; (c) Conditional distribution of QTL genotypes over marker genotypes given trait statuses.

<i>AA</i>			<i>Aa</i>			<i>aa</i>		
<i>MM</i>	<i>Mm</i>	<i>mm</i>	<i>MM</i>	<i>Mm</i>	<i>mm</i>	<i>MM</i>	<i>Mm</i>	<i>mm</i>
g_{11}	g_{12}	g_{13}	g_{21}	g_{22}	g_{23}	g_{31}	g_{32}	g_{33}
Q^2	$2Q(1-Q)$	$(1-Q)^2$	QR	$Q+R-2QR$	$(1-Q)(1-R)$	R^2	$2R(1-R)$	$(1-R)^2$

where $Q = p + D/q$ and $R = p - D/(1-q)$

(a)

<i>MM</i>			<i>Mm</i>			<i>mm</i>		
<i>AA</i>	<i>Aa</i>	<i>aa</i>	<i>AA</i>	<i>Aa</i>	<i>aa</i>	<i>AA</i>	<i>Aa</i>	<i>aa</i>
h_{11}	h_{21}	h_{31}	h_{12}	h_{22}	h_{32}	h_{13}	h_{23}	h_{33}
Q^2	$2Q(1-Q)$	$(1-Q)^2$	QR	$Q+R-2QR$	$(1-Q)(1-R)$	R^2	$2R(1-R)$	$(1-R)^2$

where $Q = q + D/p$ and $R = q - D/(1-p)$

(b)

	Cases			Controls		
	<i>MM</i>	<i>Mm</i>	<i>mm</i>	<i>MM</i>	<i>Mm</i>	<i>mm</i>
<i>AA</i>	g_{11}	g_{12}	g_{13}	g_{11}	g_{12}	g_{13}
<i>Aa</i>	g_{21}	g_{22}	g_{23}	g_{21}	g_{22}	g_{23}
<i>aa</i>	g_{31}	g_{32}	g_{33}	g_{31}	g_{32}	g_{33}
# observed	n_{11}	n_{12}	n_{13}	n_{21}	n_{22}	n_{23}

(c)

As n_{ki} is the observed number of individuals with the i th marker genotype and the k th case-control status, t_{ijk} could hence be directly estimated if both $\Pr(G=i|M=j, Case)$ and $\Pr(G=i|M=j, Control)$ is known. By noticing

$$\begin{aligned} \frac{f_i h_{ij}}{\sum_i f_i h_{ij}} &= \frac{\Pr(Case, G=i|M=j)}{\Pr(Case|M=j)} \\ &= \Pr(G=i|M=j, Case) \\ &= w_{ij} \end{aligned}$$

and similarly

$$\frac{(1-f_i)h_{ij}}{\sum_i (1-f_i)h_{ij}} = \Pr(G=i \mid M=j, Control) ,$$

$$= v_{ij}$$

we may have $\tilde{t}_{ijk} = n_{kj} w_{ij}^{2-k} v_{ij}^{k-1}$. The logarithm of likelihood function (III-2.2) could hence be rewritten as

$$l(p, q, D \mid N, f_1, f_2, f_3) \sim \sum_{j=1}^3 \left\{ \sum_{i=1}^3 \left[(n_{1j} w_{ij} + n_{2j} v_{ij}) \log(g_{ij}) \right] \right\}. \quad (\text{III-2.3})$$

With a known p , which could be estimated from the control individuals as shown in Chapter III-1, the remaining two parameters, i.e. q and D , could be estimated through equating the first partial derivatives of function (III-2.3) to zero, which lead to two normal equations:

$$a_6 q^6 + a_5 q^5 + a_4 q^4 + a_3 q^3 + a_2 q^2 + a_1 q + a_0 = 0 \quad (\text{III-2.4})$$

and

$$b_5 D^5 + b_4 D^4 + b_3 D^3 + b_2 D^2 + b_1 D + b_0 = 0. \quad (\text{III-2.5})$$

The coefficients of above two equations could be calculated as

$$\begin{aligned} a_0 = & (2d^3 + 6d^4 + 4d^5 - 2d^2 p - 12d^3 p - 12d^4 p + 6d^2 p^2 + 12d^3 p^2 - 4d^2 p^3) x_{11} \\ & + (2d^3 + 6d^4 + 4d^5 - 2d^2 p - 12d^3 p - 12d^4 p + 6d^2 p^2 + 12d^3 p^2 - 4d^2 p^3) x_{12} \\ & + (2d^3 + 6d^4 + 4d^5 - 2d^2 p - 12d^3 p - 12d^4 p + 6d^2 p^2 + 12d^3 p^2 - 4d^2 p^3) x_{13} \\ & + (d^3 + 3d^4 + 2d^5 - d^2 p - 6d^3 p - 6d^4 p + 3d^2 p^2 + 6d^3 p^2 - 2d^2 p^3) x_{21} \\ & + (d^3 + 3d^4 + 2d^5 - d^2 p - 6d^3 p - 6d^4 p + 3d^2 p^2 + 6d^3 p^2 - 2d^2 p^3) x_{22} \\ & + (d^3 + 3d^4 + 2d^5 - d^2 p - 6d^3 p - 6d^4 p + 3d^2 p^2 + 6d^3 p^2 - 2d^2 p^3) x_{23} \end{aligned}$$

$$\begin{aligned}
a_1 = & (-2d^2 - 14d^3 - 18d^4 - 4d^5 + 2dp + 28d^2p + 66d^3p + 32d^4p - 12dp^2 - 66d^2p^2 \\
& - 60d^3p^2 + 18dp^3 + 40d^2p^3 - 8dp^4)x_{11} \\
& + (-d^2 - 11d^3 - 16d^4 - 4d^5 + dp + 22d^2p + 60d^3p + 32d^4p - 9dp^2 - 60d^2p^2 - 60d^3p^2 \\
& + 16dp^3 + 40d^2p^3 - 8dp^4)x_{12} \\
& + (-8d^3 - 14d^4 - 4d^5 + 16d^2p + 54d^3p + 32d^4p - 6dp^2 - 54d^2p^2 - 60d^3p^2 + 14dp^3 \\
& + 40d^2p^3 - 8dp^4)x_{13} \\
& + (-d^2 - 8d^3 - 12d^4 - 4d^5 + dp + 14d^2p + 36d^3p + 20d^4p - 6dp^2 - 33d^2p^2 - 32d^3p^2 \\
& + 9dp^3 + 20d^2p^3 - 4dp^4)x_{21} \\
& + (-d^2 - 6d^3 - 10d^4 - 4d^5 + dp + 12d^2p + 32d^3p + 20d^4p - 5dp^2 - 30d^2p^2 - 32d^3p^2 \\
& + 8dp^3 + 20d^2p^3 - 4dp^4)x_{22} \\
& + (-4d^3 - 8d^4 - 4d^5 + 8d^2p + 28d^3p + 20d^4p - 3dp^2 - 27d^2p^2 - 32d^3p^2 + 7dp^3 \\
& + 20d^2p^3 - 4dp^4)x_{23} \\
& + (-2d^3 - 6d^4 - 4d^5 + 6d^3p + 8d^4p - 4d^3p^2)x_{31} \\
& + (-d^3 - 4d^4 - 4d^5 + 4d^3p + 8d^4p - 4d^3p^2)x_{32} \\
& + (-2d^4 - 4d^5 + 2d^3p + 8d^4p - 4d^3p^2)x_{33}
\end{aligned}$$

$$\begin{aligned}
a_2 = & (8d^2 + 24d^3 + 12d^4 - 14dp - 86d^2p - 102d^3p - 20d^4p + 4p^2 + 72dp^2 \\
& + 186d^2p^2 + 88d^3p^2 - 12p^3 - 102dp^3 - 108d^2p^3 + 12p^4 + 44dp^4 - 4p^5)x_{11} \\
& + (4d^2 + 17d^3 + 10d^4 - 7dp - 62d^2p - 88d^3p - 20d^4p + 2p^2 + 51dp^2 \\
& + 162d^2p^2 + 88d^3p^2 - 8p^3 - 88dp^3 - 108d^2p^3 + 10p^4 + 44dp^4 - 4p^5)x_{12} \\
& + (10d^3 + 8d^4 - 38d^2p - 74d^3p - 20d^4p + 30dp^2 + 138d^2p^2 + 88d^3p^2 \\
& - 4p^3 - 74dp^3 - 108d^2p^3 + 8p^4 + 44dp^4 - 4p^5)x_{13} \\
& + (5d^2 + 18d^3 + 12d^4 - 7dp - 50d^2p - 72d^3p - 20d^4p + 2p^2 + 36dp^2 \\
& + 105d^2p^2 + 60d^3p^2 - 6p^3 - 51dp^3 - 60d^2p^3 + 6p^4 + 22dp^4 - 2p^5)x_{21} \\
& + (3d^2 + 12d^3 + 10d^4 - 5dp - 36d^2p - 60d^3p - 20d^4p + p^2 + 27dp^2 \\
& + 90d^2p^2 + 60d^3p^2 - 4p^3 - 44dp^3 - 60d^2p^3 + 5p^4 + 22dp^4 - 2p^5)x_{22} \\
& + (6d^3 + 8d^4 - 20d^2p - 48d^3p - 20d^4p + 15dp^2 + 75d^2p^2 + 60d^3p^2 \\
& - 2p^3 - 37dp^3 - 60d^2p^3 + 4p^4 + 22dp^4 - 2p^5)x_{23} \\
& + (2d^2 + 12d^3 + 12d^4 - 14d^2p - 42d^3p - 20d^4p + 24d^2p^2 + 32d^3p^2 - 12d^2p^3)x_{31} \\
& + (d^2 + 7d^3 + 10d^4 - 8d^2p - 32d^3p - 20d^4p + 18d^2p^2 + 32d^3p^2 - 12d^2p^3)x_{32} \\
& + (2d^3 + 8d^4 - 2d^2p - 22d^3p - 20d^4p + 12d^2p^2 + 32d^3p^2 - 12d^2p^3)x_{33}
\end{aligned}$$

$$\begin{aligned}
a_3 = & (-10d^2 - 12d^3 + 30dp + 96d^2p + 48d^3p - 16p^2 - 144dp^2 - 198d^2p^2 - 40d^3p^2 \\
& + 48p^3 + 198dp^3 + 112d^2p^3 - 48p^4 - 84dp^4 + 16p^5)x_{11} \\
& + (-5d^2 - 8d^3 + 15dp + 66d^2p + 40d^3p - 8p^2 - 99dp^2 - 168d^2p^2 - 40d^3p^2 + 32p^3 \\
& + 168dp^3 + 112d^2p^3 - 40p^4 - 84dp^4 + 16p^5)x_{12} \\
& + (-4d^3 + 36d^2p + 32d^3p - 54dp^2 - 138d^2p^2 - 40d^3p^2 + 16p^3 + 138dp^3 + 112d^2p^3 \\
& - 32p^4 - 84dp^4 + 16p^5)x_{13} \\
& + (-8d^2 - 12d^3 + 18dp + 72d^2p + 48d^3p - 8p^2 - 84dp^2 - 144d^2p^2 - 40d^3p^2 + 24p^3 \\
& + 114dp^3 + 80d^2p^3 - 24p^4 - 48dp^4 + 8p^5)x_{21} \\
& + (-4d^2 - 8d^3 + 10dp + 48d^2p + 40d^3p - 4p^2 - 58dp^2 - 120d^2p^2 - 40d^3p^2 + 16p^3 \\
& + 96dp^3 + 80d^2p^3 - 20p^4 - 48dp^4 + 8p^5)x_{22} \\
& + (-4d^3 + 24d^2p + 32d^3p - 30dp^2 - 96d^2p^2 - 40d^3p^2 + 8p^3 + 78dp^3 + 80d^2p^3 - 16p^4 \\
& - 48dp^4 + 8p^5)x_{23} \\
& + (-6d^2 - 12d^3 + 6dp + 48d^2p + 48d^3p - 24dp^2 - 90d^2p^2 - 40d^3p^2 + 30dp^3 + 48d^2p^3 \\
& - 12dp^4)x_{31} \\
& + (-3d^2 - 8d^3 + 3dp + 30d^2p + 40d^3p - 15dp^2 - 72d^2p^2 - 40d^3p^2 + 24dp^3 + 48d^2p^3 \\
& - 12dp^4)x_{32} \\
& + (-4d^3 + 12d^2p + 32d^3p - 6dp^2 - 54d^2p^2 - 40d^3p^2 + 18dp^3 + 48d^2p^3 - 12dp^4)x_{33} \\
a_4 = & (4d^2 - 26dp - 36d^2p + 24p^2 + 120dp^2 + 72d^2p^2 - 72p^3 - 162dp^3 - 40d^2p^3 \\
& + 72p^4 + 68dp^4 - 24p^5)x_{11} \\
& + (2d^2 - 13dp - 24d^2p + 12p^2 + 81dp^2 + 60d^2p^2 - 48p^3 - 136dp^3 - 40d^2p^3 + 60p^4 \\
& + 68dp^4 - 24p^5)x_{12} \\
& + (-12d^2p + 42dp^2 + 48d^2p^2 - 24p^3 - 110dp^3 - 40d^2p^3 + 48p^4 + 68dp^4 - 24p^5)x_{13} \\
& + (4d^2 - 20dp - 36d^2p + 14p^2 + 90dp^2 + 72d^2p^2 - 42p^3 - 120dp^3 - 40d^2p^3 + 42p^4 \\
& + 50dp^4 - 14p^5)x_{21} \\
& + (2d^2 - 10dp - 24d^2p + 7p^2 + 60dp^2 + 60d^2p^2 - 28p^3 - 100dp^3 - 40d^2p^3 + 35p^4 \\
& + 50dp^4 - 14p^5)x_{22} \\
& + (-12d^2p + 30dp^2 + 48d^2p^2 - 14p^3 - 80dp^3 - 40d^2p^3 + 28p^4 + 50dp^4 - 14p^5)x_{23} \\
& + (4d^2 - 14dp - 36d^2p + 4p^2 + 60dp^2 + 72d^2p^2 - 12p^3 - 78dp^3 - 40d^2p^3 + 12p^4 \\
& + 32dp^4 - 4p^5)x_{31} \\
& + (2d^2 - 7dp - 24d^2p + 2p^2 + 39dp^2 + 60d^2p^2 - 8p^3 - 64dp^3 - 40d^2p^3 + 10p^4 \\
& + 32dp^4 - 4p^5)x_{32} \\
& + (-12d^2p + 18dp^2 + 48d^2p^2 - 4p^3 - 50dp^3 - 40d^2p^3 + 8p^4 + 32dp^4 - 4p^5)x_{33}
\end{aligned}$$

$$\begin{aligned}
a_5 = & (8dp - 16p^2 - 36dp^2 + 48p^3 + 48dp^3 - 48p^4 - 20dp^4 + 16p^5)x_{11} \\
& + (4dp - 8p^2 - 24dp^2 + 32p^3 + 40dp^3 - 40p^4 - 20dp^4 + 16p^5)x_{12} \\
& + (-12dp^2 + 16p^3 + 32dp^3 - 32p^4 - 20dp^4 + 16p^5)x_{13} \\
& + (8dp - 12p^2 - 36dp^2 + 36p^3 + 48dp^3 - 36p^4 - 20dp^4 + 12p^5)x_{21} \\
& + (4dp - 6p^2 - 24dp^2 + 24p^3 + 40dp^3 - 30p^4 - 20dp^4 + 12p^5)x_{22} \\
& + (-12dp^2 + 12p^3 + 32dp^3 - 24p^4 - 20dp^4 + 12p^5)x_{23} \\
& + (8dp - 8p^2 - 36dp^2 + 24p^3 + 48dp^3 - 24p^4 - 20dp^4 + 8p^5)x_{31} \\
& + (4dp - 4p^2 - 24dp^2 + 16p^3 + 40dp^3 - 20p^4 - 20dp^4 + 8p^5)x_{32} \\
& + (-12dp^2 + 8p^3 + 32dp^3 - 16p^4 - 20dp^4 + 8p^5)x_{33}
\end{aligned}$$

$$\begin{aligned}
a_6 = & (4p^2 - 12p^3 + 12p^4 - 4p^5)x_{11} + (2p^2 - 8p^3 + 10p^4 - 4p^5)x_{12} + (-4p^3 + 8p^4 - 4p^5)x_{13} \\
& + (4p^2 - 12p^3 + 12p^4 - 4p^5)x_{21} + (2p^2 - 8p^3 + 10p^4 - 4p^5)x_{22} + (-4p^3 + 8p^4 - 4p^5)x_{23} \\
& + (4p^2 - 12p^3 + 12p^4 - 4p^5)x_{31} + (2p^2 - 8p^3 + 10p^4 - 4p^5)x_{32} + (-4p^3 + 8p^4 - 4p^5)x_{33}
\end{aligned}$$

and

$$\begin{aligned}
b_0 = & (4p^2q^2 - 12p^3q^2 + 12p^4q^2 - 4p^5q^2 - 12p^2q^3 + 36p^3q^3 - 36p^4q^3 + 12p^5q^3 \\
& + 12p^2q^4 - 36p^3q^4 + 36p^4q^4 - 12p^5q^4 - 4p^2q^5 + 12p^3q^5 - 12p^4q^5 + 4p^5q^5)x_{11} \\
& + (2p^2q^2 - 8p^3q^2 + 10p^4q^2 - 4p^5q^2 - 6p^2q^3 + 24p^3q^3 - 30p^4q^3 + 12p^5q^3 + 6p^2q^4 \\
& - 24p^3q^4 + 30p^4q^4 - 12p^5q^4 - 2p^2q^5 + 8p^3q^5 - 10p^4q^5 + 4p^5q^5)x_{12} \\
& + (-4p^3q^2 + 8p^4q^2 - 4p^5q^2 + 12p^3q^3 - 24p^4q^3 + 12p^5q^3 - 12p^3q^4 + 24p^4q^4 - 12p^5q^4 \\
& + 4p^3q^5 - 8p^4q^5 + 4p^5q^5)x_{13} \\
& + (2p^2q^2 - 6p^3q^2 + 6p^4q^2 - 2p^5q^2 - 8p^2q^3 + 24p^3q^3 - 24p^4q^3 + 8p^5q^3 + 10p^2q^4 - 30p^3q^4 \\
& + 30p^4q^4 - 10p^5q^4 - 4p^2q^5 + 12p^3q^5 - 12p^4q^5 + 4p^5q^5)x_{21} \\
& + (p^2q^2 - 4p^3q^2 + 5p^4q^2 - 2p^5q^2 - 4p^2q^3 + 16p^3q^3 - 20p^4q^3 + 8p^5q^3 + 5p^2q^4 - 20p^3q^4 \\
& + 25p^4q^4 - 10p^5q^4 - 2p^2q^5 + 8p^3q^5 - 10p^4q^5 + 4p^5q^5)x_{22} \\
& + (-2p^3q^2 + 4p^4q^2 - 2p^5q^2 + 8p^3q^3 - 16p^4q^3 + 8p^5q^3 - 10p^3q^4 + 20p^4q^4 - 10p^5q^4 \\
& + 4p^3q^5 - 8p^4q^5 + 4p^5q^5)x_{23} \\
& + (-4p^2q^3 + 12p^3q^3 - 12p^4q^3 + 4p^5q^3 + 8p^2q^4 - 24p^3q^4 + 24p^4q^4 - 8p^5q^4 - 4p^2q^5 \\
& + 12p^3q^5 - 12p^4q^5 + 4p^5q^5)x_{31} \\
& + (-2p^2q^3 + 8p^3q^3 - 10p^4q^3 + 4p^5q^3 + 4p^2q^4 - 16p^3q^4 + 20p^4q^4 - 8p^5q^4 - 2p^2q^5 + 8p^3q^5 \\
& - 10p^4q^5 + 4p^5q^5)x_{32} \\
& + (4p^3q^3 - 8p^4q^3 + 4p^5q^3 - 8p^3q^4 + 16p^4q^4 - 8p^5q^4 + 4p^3q^5 - 8p^4q^5 + 4p^5q^5)x_{33}
\end{aligned}$$

$$\begin{aligned}
b_1 = & (2pq - 12p^2q + 18p^3q - 8p^4q - 12pq^2 + 60p^2q^2 - 84p^3q^2 + 36p^4q^2 + 18pq^3 - 84p^2q^3 \\
& + 114p^3q^3 - 48p^4q^3 - 8pq^4 + 36p^2q^4 - 48p^3q^4 + 20p^4q^4)x_{11} \\
& + (pq - 9p^2q + 16p^3q - 8p^4q - 6pq^2 + 42p^2q^2 - 72p^3q^2 + 36p^4q^2 + 9pq^3 - 57p^2q^3 + 96p^3q^3 \\
& - 48p^4q^3 - 4pq^4 + 24p^2q^4 - 40p^3q^4 + 20p^4q^4)x_{12} \\
& + (-6p^2q + 14p^3q - 8p^4q + 24p^2q^2 - 60p^3q^2 + 36p^4q^2 - 30p^2q^3 + 78p^3q^3 - 48p^4q^3 + 12p^2q^4 \\
& - 32p^3q^4 + 20p^4q^4)x_{13} \\
& + (pq - 6p^2q + 9p^3q - 4p^4q - 9pq^2 + 42p^2q^2 - 57p^3q^2 + 24p^4q^2 + 16pq^3 - 72p^2q^3 + 96p^3q^3 \\
& - 40p^4q^3 - 8pq^4 + 36p^2q^4 - 48p^3q^4 + 20p^4q^4)x_{21} \\
& + (pq - 5p^2q + 8p^3q - 4p^4q - 5pq^2 + 29p^2q^2 - 48p^3q^2 + 24p^4q^2 + 8pq^3 - 48p^2q^3 + 80p^3q^3 \\
& - 40p^4q^3 - 4pq^4 + 24p^2q^4 - 40p^3q^4 + 20p^4q^4)x_{22} \\
& + (-3p^2q + 7p^3q - 4p^4q + 15p^2q^2 - 39p^3q^2 + 24p^4q^2 - 24p^2q^3 + 64p^3q^3 - 40p^4q^3 + 12p^2q^4 \\
& - 32p^3q^4 + 20p^4q^4)x_{23} \\
& + (-6pq^2 + 24p^2q^2 - 30p^3q^2 + 12p^4q^2 + 14pq^3 - 60p^2q^3 + 78p^3q^3 - 32p^4q^3 - 8pq^4 + 36p^2q^4 \\
& - 48p^3q^4 + 20p^4q^4)x_{31} \\
& + (-3pq^2 + 15p^2q^2 - 24p^3q^2 + 12p^4q^2 + 7pq^3 - 39p^2q^3 + 64p^3q^3 - 32p^4q^3 - 4pq^4 + 24p^2q^4 \\
& - 40p^3q^4 + 20p^4q^4)x_{32} \\
& + (6p^2q^2 - 18p^3q^2 + 12p^4q^2 - 18p^2q^3 + 50p^3q^3 - 32p^4q^3 + 12p^2q^4 - 32p^3q^4 + 20p^4q^4)x_{33}
\end{aligned}$$

$$\begin{aligned}
b_2 = & (-2p + 6p^2 - 4p^3 - 2q + 26pq - 60p^2q + 36p^3q + 6q^2 - 60pq^2 + 126p^2q^2 - 72p^3q^2 \\
& - 4q^3 + 36pq^3 - 72p^2q^3 + 40p^3q^3)x_{11} \\
& + (-2p + 6p^2 - 4p^3 - q + 20pq - 54p^2q + 36p^3q + 3q^2 - 42pq^2 + 108p^2q^2 - 72p^3q^2 - 2q^3 \\
& + 24pq^3 - 60p^2q^3 + 40p^3q^3)x_{12} \\
& + (-2p + 6p^2 - 4p^3 + 14pq - 48p^2q + 36p^3q - 24pq^2 + 90p^2q^2 - 72p^3q^2 + 12pq^3 - 48p^2q^3 \\
& + 40p^3q^3)x_{13} \\
& + (-p + 3p^2 - 2p^3 - 2q + 20pq - 42p^2q + 24p^3q + 6q^2 - 54pq^2 + 108p^2q^2 - 60p^3q^2 - 4q^3 \\
& + 36pq^3 - 72p^2q^3 + 40p^3q^3)x_{21} \\
& + (-p + 3p^2 - 2p^3 - q + 14pq - 36p^2q + 24p^3q + 3q^2 - 36pq^2 + 90p^2q^2 - 60p^3q^2 - 2q^3 \\
& + 24pq^3 - 60p^2q^3 + 40p^3q^3)x_{22} \\
& + (-p + 3p^2 - 2p^3 + 8pq - 30p^2q + 24p^3q - 18pq^2 + 72p^2q^2 - 60p^3q^2 + 12pq^3 - 48p^2q^3 \\
& + 40p^3q^3)x_{23} \\
& + (-2q + 14pq - 24p^2q + 12p^3q + 6q^2 - 48pq^2 + 90p^2q^2 - 48p^3q^2 - 4q^3 + 36pq^3 - 72p^2q^3 \\
& + 40p^3q^3)x_{31} \\
& + (-q + 8pq - 18p^2q + 12p^3q + 3q^2 - 30pq^2 + 72p^2q^2 - 48p^3q^2 - 2q^3 + 24pq^3 - 60p^2q^3 \\
& + 40p^3q^3)x_{32} \\
& + (2pq - 12p^2q + 12p^3q - 12pq^2 + 54p^2q^2 - 48p^3q^2 + 12pq^3 - 48p^2q^3 + 40p^3q^3)x_{33}
\end{aligned}$$

$$\begin{aligned}
b_3 = & (2-12p+12p^2-12q+54pq-48p^2q+12q^2-48pq^2+40p^2q^2)x_{11} \\
& + (2-12p+12p^2-9q+48pq-48p^2q+8q^2-40pq^2+40p^2q^2)x_{12} \\
& + (2-12p+12p^2-6q+42pq-48p^2q+4q^2-32pq^2+40p^2q^2)x_{13} \\
& + (2-9p+8p^2-12q+48pq-40p^2q+12q^2-48pq^2+40p^2q^2)x_{21} \\
& + (1-8p+8p^2-8q+40pq-40p^2q+8q^2-40pq^2+40p^2q^2)x_{22} \\
& + (1-7p+8p^2-4q+32pq-40p^2q+4q^2-32pq^2+40p^2q^2)x_{23} \\
& + (2-6p+4p^2-12q+42pq-32p^2q+12q^2-48pq^2+40p^2q^2)x_{31} \\
& + (1-4p+4p^2-7q+32pq-32p^2q+8q^2-40pq^2+40p^2q^2)x_{32} \\
& + (-2p+4p^2-2q+22pq-32p^2q+4q^2-32pq^2+40p^2q^2)x_{33} \\
b_4 = & (6-12p-12q+20pq)x_{11} + (6-12p-10q+20pq)x_{12} + (6-12p-8q+20pq)x_{13} \\
& + (6-10p-12q+20pq)x_{21} + (5-10p-10q+20pq)x_{22} + (4-10p-8q+20pq)x_{23} \\
& + (6-8p-12q+20pq)x_{31} + (4-8p-10q+20pq)x_{32} + (2-8p-8q+20pq)x_{33}
\end{aligned}$$

$$b_5 = 4n$$

$$\text{where } x_{ij} = n_{1j}w_{ij} + n_{2j}v_{ij}.$$

Although function (III-2.3) is not derived through an ECM procedure, from section II-1.3.3, an ECM process could be considered as an iterative process for solving equations (III-2.4) and (III-2.5), and hence we may still estimate q and D following the scheme of an ECM procedure:

E-step:

Calculate w_{ij}^t and v_{ij}^t in t th iterative process from p , D^t and q^t .

CM-step:

Update q^{t+1} and D^{t+1} by solving equations (III-2.4) and (III-2.5) one after the other.

Iteratively updating both the E and CM steps until the increase of likelihood at the t -th iteration reaches a presumed convergence criterion, and then we could claim both D^t and q^t have converged to their MLEs.

Clearly, there are no analytical solutions for both equations (III-2.4) and (III-2.5). However, they could still be solved numerically. It is also possible that multiple roots may be acquired while solving equations (III-2.4) and (III-2.5), and those fall into the theoretical boundary, i.e.

$$0 < q < 1$$

and

$$\max\{-pq, -(1-p)(1-q)\} \leq D \leq \min\{p(1-q), (1-p)q\},$$

with the highest likelihood will be selected as the MLEs.

The statistical test against the null hypothesis $D = 0$ could be established through the LRT as

$$LR = -2 \left[l(\hat{p}, \hat{q}, D=0 | N, f_1, f_2, f_3) - l(\hat{p}, \hat{q}, \hat{D} | N, f_1, f_2, f_3) \right].$$

It should be noted here that as the likelihood function (III-2.3) under the null hypothesis could be expressed as

$$l(p, q, D=0 | N, f_1, f_2, f_3) \sim (n_{11} + n_{21}) \log[p^2] + (n_{12} + n_{22}) \log[2p(1-p)] + (n_{13} + n_{23}) \log[(1-p)^2],$$

which is free from both parameters q and D , we could hence claim $LR \sim \chi^2$ with 2 degrees of freedom as discussed in sections I-1.4.2 and II-1.3.3. Note here, if the case and control data are collected independently from multiple cohorts, say number equals k , one could calculate LR for each cohort, and the sum of all LR should follow χ^2 with $2k$ degrees of freedom.

Clearly, the likelihood-based method proposed above retains the flexibility of adopting any a penetrance model, *e.g.* the co-dominance model, as the penetrance coefficients are explicitly modelled as f_j . Such a flexibility is important as shown in Chapter II-3 that the penetrance coefficients are hard to be explicitly modelled in the existing methods, *e.g.* the Armitage's trend test and hence approximation has to be made, where, on the contrary, the new method could have any penetrance models fitted in easily.

2.4 Allelic Analysis for Multiple Cohorts

The relationship between the allelic analysis and the Armitage's trend test has been discussed in section II-2.3 and II-3.3.3, where we have argued that the allelic analysis loses the information about both penetrance models but may have a higher statistical power if the alternative hypothesis, i.e. the testing marker locus does associate with the phenotype, is true under the co-dominance situation, where the HWE is also assumed. However, as the general validity of allelic analysis is out of question as discussed in section II-2.3, and also due to the simpler form of allelic analysis when compared with the Armitage's trend test, it would be easier for us to manipulate the structure of allelic analysis to deal with the population stratification in this section.

Before initiating the main derivations, an important result will be given here at the very beginning. Recall that $\Pr(MM | AA, D) = \Pr(MM | AA, C) = \Pr(MM | AA)$ as given at formula (III-2.1), we may notice

$$\begin{aligned}\Pr(MM | AA, D) &= \Pr(MMAA | D) / \Pr(AA | D) \\ &= \Pr^2(MA | D) / \Pr^2(A | D) \\ &= \Pr^2(M | A, D)\end{aligned}$$

by assuming random union of gametes, and similarly $\Pr(MM | AA, C) = \Pr^2(M | A, C)$ and

$\Pr(MM | AA) = \Pr^2(M | A)$. Clearly, the above indicates that

$$\Pr(M | A, D) = \Pr(M | A, C) = \Pr(M | A).$$

The idea of allelic analysis is to test the difference between the marker allele frequencies in case and control samples as indicated by formula (II-1.20). In the absence of population stratification, we could calculate theoretically

$$\begin{aligned}
\Pr(M | D) &= \Pr(MA | D) + \Pr(Ma | D) \\
&= \Pr(M | A) \Pr(A | D) + \Pr(M | a) \Pr(a | D) \quad , \\
&= \Pr(M | a) + (\Pr(M | A) - \Pr(M | a)) \Pr(A | D)
\end{aligned}$$

and similarly $\Pr(M | C) = \Pr(M | a) + (\Pr(M | A) - \Pr(M | a)) \Pr(A | C)$. By noticing that

$$\Pr(M | A) = \frac{D}{p_A} + p_M \text{ and } \Pr(M | a) = -\frac{D}{p_a} + p_M \text{ from the definition of } D, \text{ where } p_\theta \text{ denotes}$$

the frequency of allele θ and θ equals one of A, a, M, m , we may calculate the difference between $\Pr(M | D)$ and $\Pr(M | C)$ as

$$\Pr(M | D) - \Pr(M | C) = \frac{D}{p_A p_a} (\Pr(A | D) - \Pr(A | C)), \quad (\text{III-2.6})$$

and the statistical test for the difference between marker allele frequencies in case and control samples is hence proportional to that between disease allele frequencies in case and control samples to a factor $\frac{D}{p_A p_a}$. Clearly, under the null hypothesis where $D = 0$, formula (III-2.6)

yields $\Pr(M | D) = \Pr(M | C)$.

However, such a result does not hold in the presence of population stratifications. Let's consider a Case-Control sample collected from k subpopulations, each of which with frequency $p_M^{(i)}$ of marker allele M , where $i = 1, 2, \dots, k$. By letting r_i and s_i denote the proportion of case and control samples collected from i th subpopulation in the whole case and control samples respectively, we may calculate

$$\begin{aligned}
p_{M|D} - p_{M|C} &= \sum_{i=1}^k r_i p_{M|D}^{(i)} - \sum_{i=1}^k s_i p_{M|C}^{(i)} \\
&= \sum_{i=1}^k (r_i - s_i) p_M^{(i)} + \frac{D_i}{p_A p_a} [r_i p_{A|D}^{(i)} - s_i p_{A|C}^{(i)} - (r_i - s_i) p_A] \quad , \quad (\text{III-2.7})
\end{aligned}$$

where $p_{\theta|D}^{(i)}$ and $p_{\theta|C}^{(i)}$ represent the frequency of allele θ for the i th subpopulation in case and control samples respectively. It is clear that under the null hypothesis, i.e. $D_i = 0 \forall i$, we could

have $p_{M|D} - p_{M|C} = \sum_{i=1}^k (r_i - s_i) p_M^{(i)}$. An obviously sufficient condition to equate formula (III-2.7) to 0 under the null hypothesis is either $p_M^{(i)} = p_M^{(j)} \forall i, j$ or $r_i = s_i \forall i$, as already suggested in section II-3.3.2 in the modified Armitage's trend test. We may hence expect a severe bias introduced by population stratification if none of such two conditions are satisfied. Given so, in the presence of population stratification, the numerator of allelic analysis has to be modified as $\Delta \hat{p}_M - \sum_{i=1}^k (r_i - s_i) \hat{p}_M^{(i)}$ in order to keep the numerator equalling to 0 under the null hypothesis, where $\Delta \hat{p}_M$ denotes $\hat{p}_{M|D} - \hat{p}_{M|C}$, i.e. the observed difference between marker allele frequencies in case and control, and $\hat{p}_M^{(i)}$ is calculated from each subpopulation with $D_i = 0$.

On the other hand, as has been mentioned in Chapter II-3, the denominator of formula (II-1.20) will tend to overestimate the variance of $\Delta \hat{p}_M$ in the presence of population stratification. A corrected estimate could be calculated as

$$\begin{aligned}
 Var(\Delta \hat{p}_M) &= Var\left(\sum_{i=1}^k r_i \hat{p}_{M|D}^{(i)} - \sum_{i=1}^k s_i \hat{p}_{M|C}^{(i)}\right) \\
 &= \sum_{i=1}^k r_i^2 Var(\hat{p}_{M|D}^{(i)}) + \sum_{i=1}^k s_i^2 Var(\hat{p}_{M|C}^{(i)}) \\
 &= \sum_{i=1}^k r_i^2 \frac{\hat{p}_{M|D}^{(i)}(1 - \hat{p}_{M|D}^{(i)})}{2t_i} + \sum_{i=1}^k s_i^2 \frac{\hat{p}_{M|C}^{(i)}(1 - \hat{p}_{M|C}^{(i)})}{2(T_i - t_i)}, \\
 &= \sum_{i=1}^k \frac{1}{2} \hat{p}_M^{(i)}(1 - \hat{p}_M^{(i)}) \left(\frac{r_i^2}{t_i} + \frac{s_i^2}{T_i - t_i} \right)
 \end{aligned} \tag{III-2.8}$$

where $p_{M|D}^{(i)} = p_{M|C}^{(i)} = p_M^{(i)}$ under the null hypothesis; t_i and T_i represent the sizes of the case and whole samples collected the i th subpopulation respectively. It could be noticed that, if $p_M^{(i)} = p_M^{(j)} \forall i, j$, i.e. in the absence of population stratification, formula (III-2.8) will be reduced to

$$Var(\Delta\hat{p}_M) = \hat{p}_M(1 - \hat{p}_M) \frac{T}{2t(T-t)},$$

which is identical to the denominator of formula (II-1.20), and we may hence conclude that formula (III-2.8) is appropriate even if there is no real difference at the testing marker locus among those subpopulations.

With formulae (III-2.7) and (III-2.8), we may hence establish the allelic analysis in the presence of population stratification as

$$\chi_A^2 = \frac{[\Delta\hat{p}_M - \sum_{i=1}^k (r_i - s_i) \hat{p}_M^{(i)}]^2}{\sum_{i=1}^k \hat{p}_M^{(i)} (1 - \hat{p}_M^{(i)}) [r_i^2 / 2t_i + s_i^2 / 2(T_i - t_i)]}, \quad (\text{III-2.9})$$

which follows χ^2 distribution with 1 degree of freedom.

2.5 Simulation Study

To investigate statistical properties and limitations of the method developed in the present study, two sampling schemes will be conducted to collect case and control individuals from computer simulated random mating populations. Sampling Scheme A collects cases and controls from a single population, and Scheme B samples cases and controls from two genetically divergent populations with regard to a tested marker and a putative disease locus. The simulated populations are characterized by population genetic parameters p , q and D (allele frequencies at a genetic marker locus and a disease trait locus, and the coefficient of linkage disequilibrium between the former two loci) and quantitative genetic parameters f_1, f_2 and f_3 (*i.e.* the penetrance coefficients for genotypes at the disease locus). In the present simulation, the penetrance coefficients f_{1-3} are fixed as $\{1, 0.5, 0\}$, representing a co-dominance model of disease allele. For any given set of the simulation parameters, genotype data are generated for case and control samples using the modified programme as described in Chapter II-3. For each simulation, three statistical methods are implemented, *i.e.* the present

likelihood-based method (Method 1), the modified allelic analysis (Method 2) and the original Armitage's trend test with the default trend set (Method 3).

Table III-8 presents the parameters defining 10 simulated random mating populations as well as means and standard deviations of estimates of the model parameters from 1,000 repeated samples of 200 cases and 200 controls. Note that only Method 1 and 3 are implemented here, which is simply because Method 2 and 3 are approximately identical if no subpopulation is present. When the testing marker locus and disease locus are in linkage equilibrium, i.e. $D = 0$ in simulated populations 1-3, the marker genotype will provide no information about the unobservable genotype at the disease locus and thus no estimate of disease allele frequency q is attempted under such a circumstance. Means of the test statistics in these populations approximately equal to 2.0 or 1.0 for Method 1 or 3 respectively, corresponding to the means of the chi-square variable with 2 or 1 degree of freedom as expected in section I-1.4 and III-2.3. Above results demonstrate the adequacy of proposed distribution of the test statistics constructed in these methods under the null hypothesis and, in turn, the appropriate control of the type I error of the statistical tests. While the linkage disequilibrium is actually present (populations 4-10), Method 1 estimates the modelling parameters, q and D , adequately, providing a consistently higher statistical power (ρ) to test for significance of the association than Method 3.

Table III-8. Results for Simulation under Scheme A with Simulation Parameters.

Population genetic parameters for 10 simulated populations and statistical inference of model parameters from 200 cases and 200 controls repeatedly sampled from the simulation populations. p and q are allelic frequencies at the marker and disease loci, D is the coefficient of linkage disequilibrium (LD) between the two loci. Means and standard deviations (s.d.) of the model parameters, q and D , and χ^2 test statistic were calculated from 1000 repeated samples. ρ (%) is the proportion in 1000 repeats in which the association test surpassed the Bonferroni threshold of P -value at 5×10^{-5} .

Pop.	p	q	D	Method 1				Method 3	
				$\hat{q} \pm s.d.$	$\hat{D} \pm s.d.$	$\chi^2_{[2]} \pm s.d.$	ρ (%)	$\chi^2_{[1]} \pm s.d.$	ρ (%)
1	0.5	0.5	0	-	0.001±0.012	2.0±2.7	0	0.9±1.3	0
2	0.3	0.7	0	-	0.002±0.011	2.0±2.7	0	1.0±1.3	0
3	0.7	0.3	0	-	0.001±0.011	2.2±2.7	0	1.0±1.3	0
4	0.5	0.5	0.15	0.50±0.05	0.148±0.015	184.4±42.8	100	73.3±14.0	100
5	0.5	0.5	0.10	0.50±0.09	0.097±0.018	73.9±26.5	99.7	33.3±10.6	96.6
6	0.5	0.5	0.05	0.50±0.20	0.043±0.020	18.1±12.0	36.8	8.8±5.6	10.8
7	0.3	0.7	0.07	0.72±0.12	0.064±0.026	68.4±25.4	99.6	29.6±10.2	91.5
8	0.3	0.7	0.05	0.70±0.15	0.047±0.023	33.2±17.6	77.3	15.1±7.5	38.2
9	0.7	0.3	-0.07	0.28±0.14	-0.062±0.028	54.8±23.4	96.8	26.3±9.6	85.2
10	0.7	0.3	-0.05	0.31±0.20	-0.042±0.024	27.8±15.6	66.1	13.7±6.9	31.0

Method 1: Likelihood-based Method (Section III-2.3)

Method 3: The Original Armitage's Trend Test (Section II-1.4.3)

On the other hand, all three methods are implemented for simulations under Scheme B to explore the influence of multiple subpopulations, or cohorts, on their performance. Table III-9 illustrates 14 sets of simulation parameters, which define the genetic structures of two random mating populations and the corresponding empirical powers of these three methods. There are implemented to perform the association tests with case-control samples from these populations separately and jointly, where, in the admixed samples, 57% cases and 76% controls are collected from population 1 and the rest are collected from population 2. The table shows that while LD is null in both of the populations (Pop. 1-6), all three methods share a low probability of claiming false positive inference using case-control samples from these populations separately. The false positive rate remains at the same lower level for Methods 1 and 2 but is increased remarkably for the Armitage's trend test, or Method 3, when the cases and controls are contributed by the two populations. Moreover, the increase in the false positive rate for Method 3 is in proportion to the difference in marker allele frequencies between the two contributing populations. The larger the difference is, the higher the false

positive rate would be, reflecting the fact that the test statistic of this method is proportionate to the size of difference between the allele frequencies. When the LD does truly exist in either or both simulated populations (Pop. 7-14), Method 1 is able to detect it with remarkably higher statistical power than the other two methods. In particular, when the LD has an opposite sign in the two contributing populations (Pop. 14-15), i.e. the scenario where the disease causing gene is in association with different marker alleles in different populations, the highest detecting power is observed for Method 1 no matter whether the case and control samples are collected from the contributing populations separately or as an admixture of the populations. In contrast, both Method 2 and 3 fail to detect the LD under this circumstance. These findings strongly support the improved statistical efficiency of the likelihood-based method and its robustness to inherent genetic structure in the case and control samples.

Table III-9. Results for Simulations under Scheme B with Simulation Parameters.

Population genetic parameters defining two genetically divergent populations, i.e. columns 2-7, and empirical statistical powers of Methods 1-3 (M 1-3) for detecting significance of linkage disequilibrium between a polymorphic marker and a putative disease locus, i.e. columns 8-16. The empirical power was calculated from 1,000 repeated samples of 1,000 cases and 1,000 controls as the proportion of the test statistic surpassing the Bonferroni threshold 5×10^{-5} . The admixed samples were made up of 57% cases and 76% controls from Population 1 and the rest from Population 2.

Pop.	$p^{(1)}$	$q^{(1)}$	$D^{(1)}$	$p^{(2)}$	$q^{(2)}$	$D^{(2)}$	Population 1			Population 2			Admixed Samples		
							M 1	M 2	M 3	M 1	M 2	M 3	M 1	M 2	M 3
1	0.40	0.10	0.00	0.70	0.10	0.00	0.1	0.0	0.0	1.6	0.0	0.0	1.2	0.0	25.3
2	0.45	0.10	0.00	0.70	0.10	0.00	0.0	0.0	0.0	1.0	0.0	0.0	0.6	0.0	12.6
3	0.50	0.10	0.00	0.70	0.10	0.00	0.3	0.0	0.0	1.4	0.0	0.0	1.2	0.0	3.7
4	0.55	0.10	0.00	0.70	0.10	0.00	0.2	0.0	0.0	2.1	0.0	0.0	1.1	0.0	0.9
5	0.60	0.10	0.00	0.70	0.10	0.00	0.0	0.0	0.0	1.1	0.0	0.0	1.0	0.0	0.3
6	0.65	0.10	0.00	0.70	0.10	0.00	0.1	0.1	0.1	0.9	0.0	0.0	0.5	0.0	0.0
7	0.40	0.10	0.00	0.50	0.10	0.02	0.1	0.0	0.0	94.3	44.8	45.6	91.1	2.9	50.8
8	0.45	0.10	0.00	0.50	0.10	0.02	0.0	0.0	0.0	93.4	45.7	47.2	90.8	1.4	28.4
9	0.40	0.10	0.02	0.50	0.10	0.00	99.5	93.9	94.7	1.1	0.0	0.0	99.4	70.1	90.0
10	0.45	0.10	0.02	0.50	0.10	0.00	99.7	95.4	95.5	1.1	0.0	0.0	99.3	69.3	77.4
11	0.40	0.10	0.02	0.50	0.10	0.02	99.6	95.0	95.1	93.2	43.7	45.7	100.0	99.7	100.0
12	0.45	0.10	0.02	0.50	0.10	0.02	99.6	95.2	95.6	93.1	47.5	49.0	100.0	99.7	100.0
13	0.40	0.10	0.02	0.50	0.10	-0.02	99.4	95.1	95.3	92.2	45.6	47.0	100.0	4.2	6.1
14	0.45	0.10	0.02	0.50	0.10	-0.02	99.1	93.9	94.0	94.2	45.8	47.8	100.0	3.0	1.4

Method 1: Likelihood-based Method (Section III-2.3)

Method 2: Adjusted Allelic Analysis (Section III-2.4)

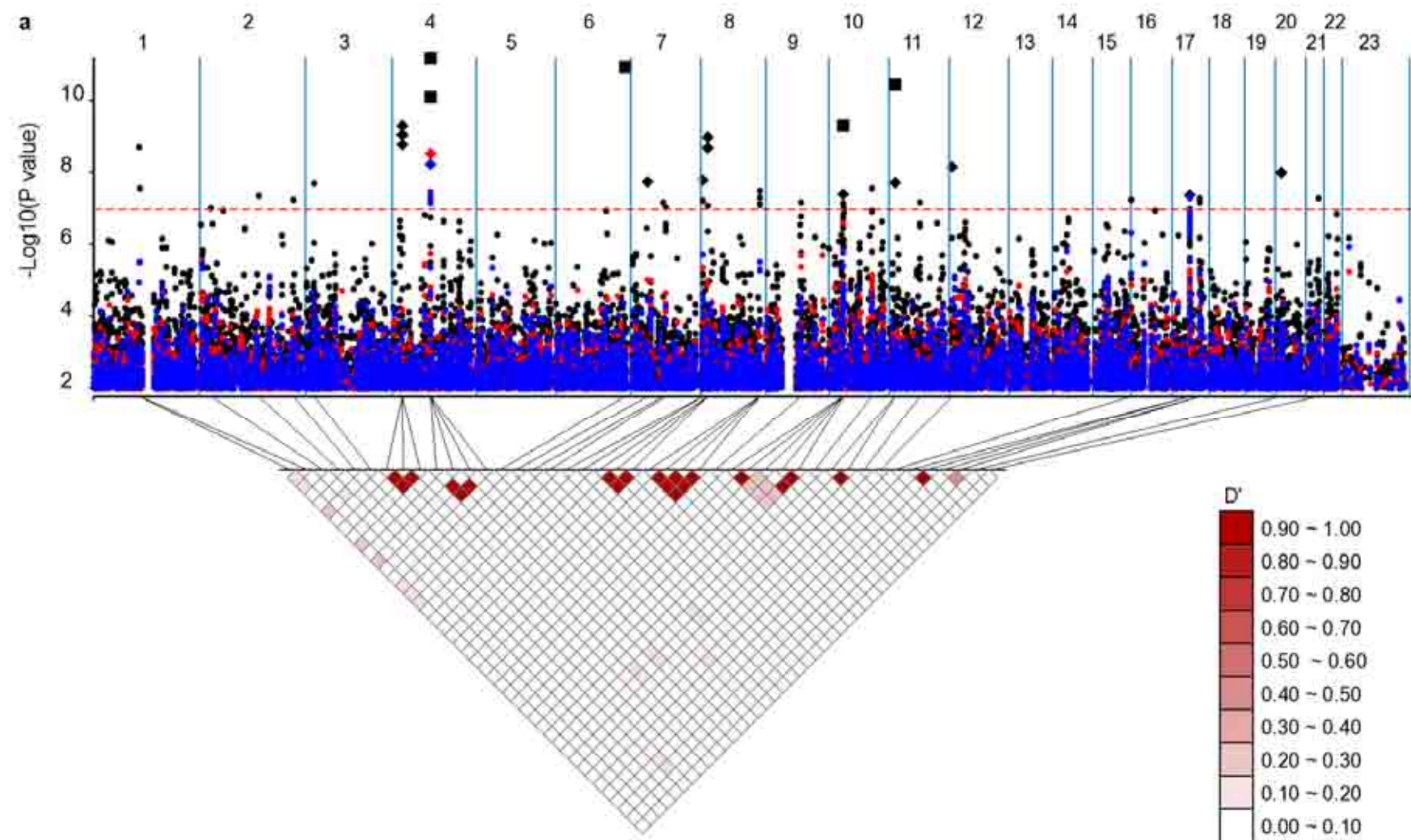
Method 3: The Original Armitage's Trend Test (Section II-1.4.3)

2.6 Real Data Analysis

All three methods are implemented to re-analyse the Parkinson's disease (PD) dataset which was recently published by Simon-Sanchez et al. (Simon-Sanchez et al., 2009). The study carried out a genome-wide screening for subtle genetic variants predisposing susceptibility to the Parkinson's diseases through a two-stage case-control design. In stage I, 4,005 individuals from the United States and 1,686 individuals recruited from the Germany were genotyped at 463,185 quality checked SNPs by using the Infinium BeadChips. Of the 5,691 objects, 1,713 were PD cases and remaining 3,978 controls. Because the estimate of allele frequency from a small sample may vary greatly, those markers, at which there were less than five individuals for any genotype, were hence excluded from further analysis. After this quality control, a total of 447,270 SNPs were used in the present study. In the stage II which was designed as a confirmation stage, 3,392, 3,223 and 1,319 individuals were recruited from three different cohorts: the United States, Germany and UK respectively. They, of which 3,341 were cases, were all genotyped at the 345 SNPs which showed significant associations in analysis with stage I dataset. After applying the same quality check on the stage II data, two SNPs were excluded in the present study. The genetic association with each of the SNP markers was evaluated by the original Armitage's trend test (Method 3 here) and the genome-wide significance level was determined by the Bonferroni correction for the probability of an overall type I error at 5%.

Figure III-2 (a)-(c) illustrate the distributions of the logarithmic significance levels ($\log-P$) of genetic association tests across the 23 human chromosomes using the three sets of case and control SNP data from the stage I, stage II and the combination of stage I and II, respectively. In analysis of each of three datasets, all three methods are implemented. It could be seen from the analysis of stage I data (Figure III-2 (a)) that 44 SNPs, which are distributed in 25 chromosomal regions with a size of less than 1 Mb (Table III-10), are detected by Method 1

developed in this chapter (black labels) to be significantly associated with the disease phenotype. Methods 2 and 3 are able to detect only two (4q21 and 17q21) of the 25 regions at the same Bonferroni threshold ($P \leq 1.1 \times 10^{-7}$). In order to explore the genetic dependences among the 44 significant SNPs, the coefficients of LD between any pair of the SNPs are calculated through the approach that was introduced in Chapter III-1 for the Case-Control samples. The disequilibrium structure illustrated at the bottom of Figure III-2 (a) shows that the significant SNPs are not associated with each other across the different regions, which hence excludes the concern that the detected SNP- disease associations might be due to the random association in genotypic distribution among the SNPs between these regions. In particular, Method 1 uniquely detects three candidate SNPs on chromosome region 8p22 (the most significant $P = 9.9 \times 10^{-10}$, rs2736050) which are only 1.2 Mb apart from a previously reported PD susceptible gene *FGF20* that was reported to be associated with Parkinson's disease synergistically with *SNCA* (Mizuta et al., 2008). To assess the variation of the predicted genetic associations, the bootstrap sampling with replacement is performed for the stage I dataset, where the empirical posterior probability is calculated at each of the 44 significant SNPs from 1,000 bootstrap samples. Table III-10 summarizes the significance level (P value) and the bootstrap posterior probability (BPP) calculated from the three methods and shows that Method 1 confers a powerful test for the genetic association than the other two methods. The BPP values predicted for analysis with Method 1 are consistently higher, suggesting the method is more robust to the variation caused by sampling than the other two methods.



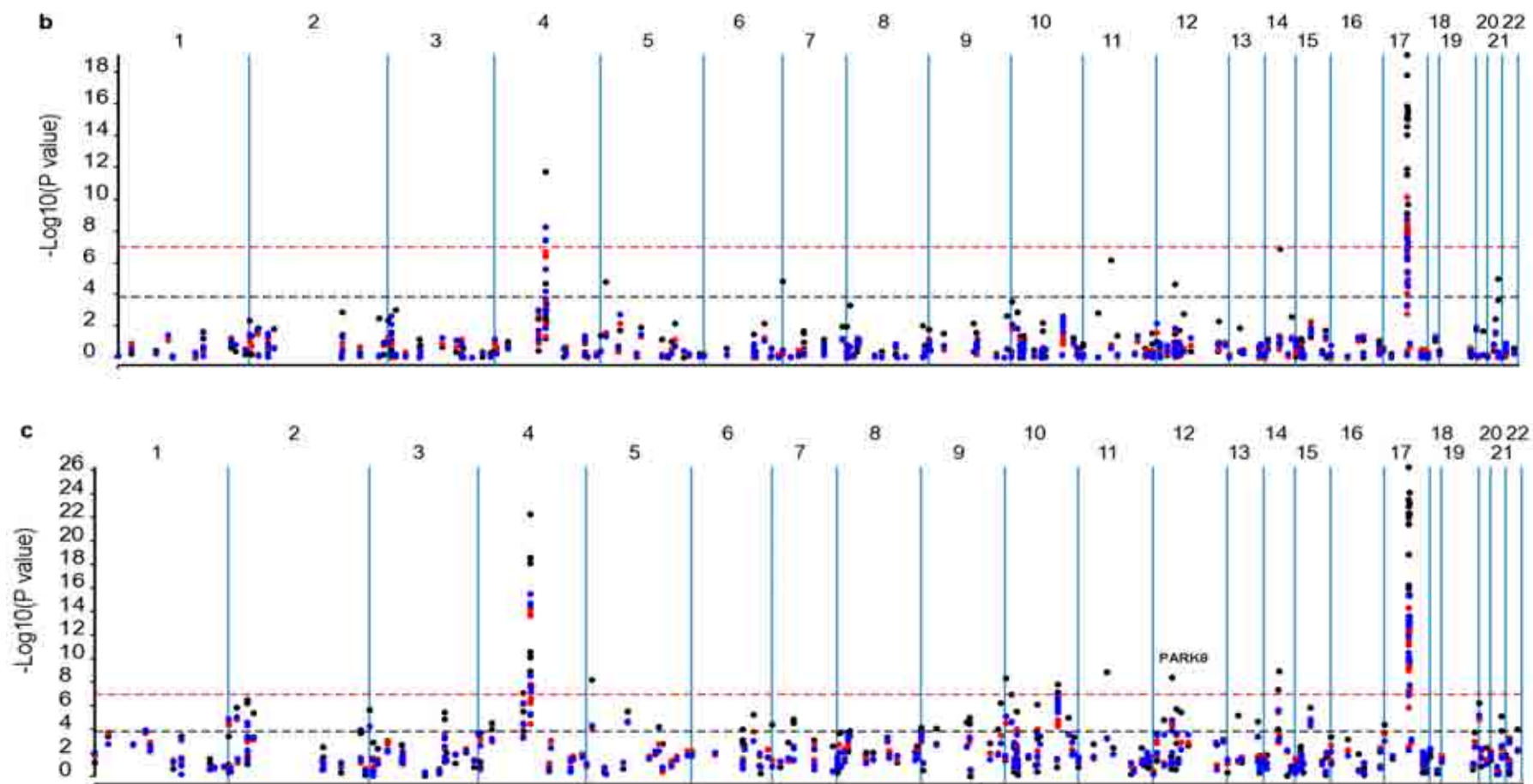


Figure III-2. Genome-wide association results from (a) stage I, (b) stage II and (c) two-stage combined case and control samples. The analysis with each of the three datasets was done using Method 1 (black circles), 2 (red circles) and 3 (blue circles) accordingly. The red and black horizontal dashed lines indicate the Bonferroni significance threshold of P value equal to 1.1×10^{-7} and 1.5×10^{-4} respectively. The triangle at the bottom of (a) is the estimated linkage disequilibrium structure for the 44 most significant SNPs listed in Table 1. The diamonds and squares in (a) illustrate the SNPs at which the bootstrap posterior probability for genetic association are either $> 80\%$ or within 60–80%.

Table III-10. Significant Markers Detected by Method 1 from Stage I Data.

Significance and Bootstrap posterior probabilities (BPP) for the 44 SNPs exceeding the Bonferroni genome-wide significance threshold (1.1×10^{-7}) detected by Method 1 (M 1) from stage I dataset. Shadowed are the regions at which the genetic association was tested by Method 2 (M2) and Method 3 (M 3) at the same significance level.

Locus	SNP name	Dist(kb) [*]	<i>P</i> value			BPP (%)		
			M 1	M 2	M 3	M 1	M 2	M 3
1p13.2-	rs17654531	-	1.9×10^{-9}	3.2×10^{-6}	1.2×10^{-5}	37	22	14
13.3	rs10857899	328	2.7×10^{-8}	3.1×10^{-6}	3.1×10^{-6}	57	25	27
2p23.3	rs7564397	-	9.7×10^{-8}	0.013	0.033	55	0	0
2q21.2	rs1474406	-	4.3×10^{-8}	2.3×10^{-3}	0.001	57	1	3
2q36.1	rs1447108	-	5.5×10^{-8}	2.5×10^{-4}	4.4×10^{-4}	59	4	3
3p24.3	rs1605527	-	2.0×10^{-8}	1.0×10^{-4}	9.4×10^{-5}	53	9	10
4p15.2	rs6820719	-	1.6×10^{-9}	0.23	0.30	74	0	0
	rs7676830	23	8.6×10^{-10}	0.12	0.15	77	0	0
	rs12649499	11	4.8×10^{-10}	0.20	0.26	77	0	0
4q21	rs11931074	-	3.9×10^{-8}	5.1×10^{-8}	4.8×10^{-8}	56	54	54
	rs356220	2	7.7×10^{-11}	3.4×10^{-8}	7.0×10^{-8}	81	56	52
	rs3857059	34	5.3×10^{-8}	4.0×10^{-8}	3.6×10^{-8}	56	55	56
	rs2736990	3	6.3×10^{-12}	2.9×10^{-9}	5.7×10^{-9}	88	71	67
6q27	rs2072638	-	1.1×10^{-11}	0.014	0.012	86	0	0
7p14-p13	rs859522	-	1.8×10^{-8}	9.7×10^{-6}	3.4×10^{-5}	62	21	14
7q21	rs3779331	-	6.6×10^{-8}	0.028	0.01	56	0	0
7q21.11	rs10246477	-	9.3×10^{-8}	2.3×10^{-5}	5.3×10^{-5}	56	13	10
8p23.2	rs7013027	-	5.8×10^{-8}	4.3×10^{-6}	1.9×10^{-6}	56	23	29
	rs4875773	63	1.6×10^{-8}	0.02	0.044	63	0	0
8p22	rs7828611	-	8.4×10^{-8}	1.2×10^{-4}	6.2×10^{-4}	55	6	3
	rs2736050	1	9.9×10^{-10}	1.0×10^{-5}	2.0×10^{-4}	74	18	5
	rs2009817	3	2.0×10^{-9}	1.3×10^{-5}	2.1×10^{-4}	72	16	5
8q24.23-	rs4556079	-	4.8×10^{-8}	5.0×10^{-6}	4.8×10^{-6}	60	20	22
24.3	rs11781101	14	7.3×10^{-8}	5.4×10^{-6}	5.3×10^{-6}	56	21	22
	rs7004938	12	3.1×10^{-8}	3.0×10^{-6}	3.0×10^{-6}	59	24	25
	rs11783351	1	7.7×10^{-8}	5.0×10^{-6}	5.5×10^{-6}	53	21	21
9q21.31	rs2378554	-	6.6×10^{-8}	2.0×10^{-6}	2.9×10^{-5}	54	29	13
10p11.21	rs2492448	-	3.8×10^{-8}	1.6×10^{-6}	3.8×10^{-6}	61	29	24
	rs11591754	12	4.8×10^{-10}	2.5×10^{-7}	1.7×10^{-6}	80	43	30
	rs7923172	102	7.0×10^{-8}	1.1×10^{-5}	1.4×10^{-5}	54	17	16
	rs4934704	23	7.3×10^{-8}	1.2×10^{-5}	1.5×10^{-5}	54	17	16
	rs10827492	97	9.7×10^{-8}	1.3×10^{-5}	1.7×10^{-5}	52	16	16
10q24.3	rs17115100	-	2.7×10^{-8}	6.9×10^{-6}	2.5×10^{-5}	37	19	13
11p15.2	rs11605276	-	3.4×10^{-11}	0.079	0.19	86	0	0
	rs10500796	45	1.9×10^{-8}	0.18	0.30	61	0	0
11q13	rs1726764	-	6.6×10^{-8}	0.088	0.20	53	0	0
12p13	rs10849446	-	6.7×10^{-9}	1.1×10^{-4}	3.7×10^{-5}	68	6	12
16p13.3	rs11648673	-	5.5×10^{-8}	1.3×10^{-5}	4.8×10^{-7}	56	15	38
17q21	rs169201	-	1.0×10^{-7}	6.5×10^{-6}	1.2×10^{-7}	57	19	49
	rs199533	39	4.1×10^{-8}	2.8×10^{-6}	5.0×10^{-8}	60	24	55
17q24.3	rs558076	-	6.6×10^{-8}	1.0×10^{-4}	2.5×10^{-5}	57	7	14
	rs817097	42	5.0×10^{-8}	8.1×10^{-6}	6.2×10^{-6}	56	18	18
20p12.1	rs6041636	-	9.9×10^{-9}	0.16	0.24	66	0	0
21q22.3	rs2070535	-	5.0×10^{-8}	0.060	0.096	54	0	0

*Distance (kb) from previous significant SNP in the same chromosome region.

Method 1: Likelihood-based Method (Section III-2.3)

Method 2: Adjusted Allelic Analysis (Section III-2.4)


Method 3: The Original Armitage's Trend Test (Section II-1.4.3)

It is worth stressing that the 345 SNPs for the confirmation study (stage II dataset) are selected only from the previous analysis using Method 3 (Simon-Sanchez et al., 2009). The dataset contains only twenty seven of the significant forty four SNPs declared by Method 1 in the stage I data analysis. In addition, the SNPs within 4q21 and 17q21 are repeatedly detected in significant association by all the three methods in the stage II dataset, the other six SNPs are detected by Method 1 to be in significant association with the disease trait at the Bonferroni genome-wide threshold (1.5×10^{-4}) (Figure III-2 (b)). In particular, a new significant SNP, rs11564162, is detected within chromosome 12q12, ($P = 2.2 \times 10^{-5}$), which is only 176 Kb from the PD candidate gene *PARK8* (Khan, N.L. et al. 2005). Neither Method 2 nor 3 could detect the significance of the SNP marker rs11564162. A full list of significant SNPs detected by the three methods in stage II dataset analysis were shown in Table III-12.

When the two datasets (stage I and stage II) are combined, ninety SNPs were detected significant at the Bonferroni threshold (1.5×10^{-4}) by Method 1, including all the twenty nine significant SNPs detected by the same method in stage II analysis and eight significant SNPs detected by the same method in stage I data analysis (Figure III-2 (c) and Table III-13. The SNP marking the PD candidate gene, *PARK8*, detected in stage II dataset analysis, is also repeated in analysis with the combined dataset. As expected, the associated SNPs are detected with remarkably more stringent significant levels, reflecting increased size of the combined datasets.

Table III-11. P-values of Known Candidate Genes from Three Methods.

The most significant SNP within ± 2.5 Mb chromosome regions surrounding each of 25 Parkinson's disease (PD) candidate genes. In parentheses is the physical distance (Mb) of the SNP to the corresponding PD candidate gene. P values are calculated from analysis of stage I dataset with Method 1 (square) and 2 (up triangle) and 3 (down triangle), and presented in the color bar depicting varying levels of significance probability. Note some data points are overlapped. n_{FD} refers as to estimate of the number of false discoveries for a given P value.

Gene	Chr	SNP (Distance)			P value 	n_{FD}		
		Method I	Method II	Method III		Method 1	Method 2	Method 3
PARK6	1	rs17393888 (1.42)	rs6687825 (1.61)	rs6687825 (1.61)	★ ■	2.8	202.3	162.3
PARK7	1	rs756096 (1.91)	rs6662747 (0.3)	rs6662747 (0.3)	★ ■	3	245.5	151.6
PARK10	1	rs17109582 (0.79)	rs11582713 (1.81)	rs17109582 (0.79)	★ ■	3.8	657.3	1044.6
GBA	1	rs4584384 (0.71)	rs4584384 (0.71)	rs4584384 (0.71)	▲ ▼	46.2	123.6	30.2
PARK16	1	rs11240359 (0.72)	rs10900544 (1.78)	rs823114 (0.08)	★	203.6	998.8	640
PARK9	1	rs16861613 (1.12)	rs4661747 (0.57)	rs4661747 (0.57)	▲ ▼ ■	2.4	85.1	30
PARK13	2	rs951409 (0.57)	rs10208443 (2.43)	rs13416937 (0.53)	★ ■	15.3	65.9	482.1
NR4A2	2	rs17307226 (0.55)	rs1568531 (1.62)	rs7608383 (0.52)	★	234.1	2269.5	1636.9
PARK3	2	rs3755388 (0)	rs10208443 (0.33)	rs13416937 (2.23)	★ ■	5.5	85.9	482.1
PARK11	2	rs11685523 (1.74)	rs10169231 (2.02)	rs11685523 (1.74)	★	46.2	739.7	881.2
ADH1C	4	rs3822069 (2.39)	rs10516486 (0.13)	rs1469019 (0.13)	★ ■	24.6	3392.6	1396.6
PARK5	4	rs6824001 (1.06)	rs2256007 (2.48)	rs2256007 (2.48)	★ ■	32.2	186.4	499
SNCA	4	rs2736990 (0)	rs2736990 (0)	rs2736990 (0)	★ ■	0	0	0
SNCAIP	5	rs17147449 (1.13)	rs37391 (1.73)	rs37391 (1.73)	★ ■	2.9	224.3	53.5
TBP	6	rs3012434 (0.32)	rs9294955 (1.36)	rs9294955 (1.36)	★	93.5	335.3	686.2
PARK2	6	rs3798923 (0.17)	rs9458499 (0)	rs9458499 (0)	▲ ▼	7.1	133.1	17.9
FGF20	8	rs2736050 (1.18)	rs2736050 (1.18)	rs2736050 (1.18)	▼ ▲ ■	0	3.8	63.3
DBH	9	rs11185726 (0.56)	rs11244079 (0.32)	rs11185726 (0.56)	▼ ▲	0.5	0.6	50.3
DRD4	11	rs12790950 (0.04)	rs6578273 (1.95)	rs6578273 (1.95)	★	69	435.6	259.4
PDDC1	11	rs12790950 (0.06)	rs6578273 (1.81)	rs6578273 (1.81)	★	89	435.8	259.4
PARK8	12	rs11564162 (0.18)	rs2896905 (0.13)	rs2896905 (0.13)	★ ■	0.1	3.3	2.1
MAPT	17	rs199533 (0.72)	rs199533 (0.72)	rs415430 (0.75)	▲ ▼	0	1.1	0
NDUFV2	18	rs12957179 (1.29)	rs9953827 (2.34)	rs1249489 (0.39)	★ ■	5.9	175.1	152
PARK14	22	rs1157557 (1.13)	rs2269529 (1.82)	rs5750616 (0.34)	★ ■	31.1	170.5	224.9
PARK15	22	rs241730 (0.48)	rs997120 (0.21)	rs997120 (0.21)	★ ■	13.7	154.2	78.9

Method I: Likelihood-based Method (Section III-2.3)

Method II: Adjusted Allelic Analysis (Section III-2.4)

Method III: The Original Armitage's Trend Test (Section II-1.4.3)

Table III-12. Significant Markers Detected by Method 1 from Stage II Data.

Locus	SNP	Position (bp)	<i>P</i> value		
			Method 1	Method 2	Method 3
4q21	rs356229	90825620	3.5×10^{-7}	6.3×10^{-4}	2.6×10^{-6}
	rs11931074	90858538	2.1×10^{-4}	4.1×10^{-7}	3.7×10^{-8}
	rs3857059	90894261	1.9×10^{-4}	3.4×10^{-7}	3.4×10^{-8}
	rs2736990	90897564	1.9×10^{-12}	2.1×10^{-7}	5.5×10^{-9}
	rs3775439	90928764	2.1×10^{-5}	7.0×10^{-3}	1.5×10^{-3}
	rs894278	90953558	2.1×10^{-3}	2.6×10^{-4}	5.9×10^{-5}
5p15.2	rs26286	14219402	1.7×10^{-5}	0.025	0.044
7p21.3	rs2681051	11615690	1.5×10^{-5}	0.60	0.31
11q12-q13.1	rs1005511	57123232	6.8×10^{-7}	0.14	0.25
12q12	rs11564162	38729159	2.2×10^{-5}	0.019	0.095
14q22.2	rs2878172	54443420	1.3×10^{-7}	0.06	0.12
17q21	rs11012	40869224	3.0×10^{-12}	4.0×10^{-6}	2.1×10^{-5}
	rs393152	41074926	7.9×10^{-20}	7.2×10^{-11}	1.7×10^{-9}
	rs417968	41084159	1.3×10^{-12}	2.2×10^{-7}	4.5×10^{-7}
	rs7215239	41123556	8.8×10^{-15}	9.4×10^{-8}	4.8×10^{-7}
	rs12373139	41279910	2.8×10^{-15}	1.9×10^{-8}	1.2×10^{-7}
	rs17690703	41281077	1.5×10^{-18}	2.9×10^{-9}	2.3×10^{-8}
	rs17563986	41347100	1.3×10^{-16}	4.8×10^{-9}	3.2×10^{-8}
	rs1981997	41412603	6.4×10^{-16}	1.1×10^{-8}	9.1×10^{-8}
	rs8070723	41436901	9.0×10^{-16}	1.1×10^{-8}	9.2×10^{-8}
	rs7225002	41544850	7.5×10^{-9}	8.5×10^{-5}	5.8×10^{-4}
	rs2532274	41602941	7.4×10^{-10}	3.6×10^{-6}	1.6×10^{-5}
	rs2532269	41605885	8.8×10^{-15}	8.9×10^{-8}	6.2×10^{-7}
	rs2668692	41648797	4.2×10^{-16}	7.7×10^{-9}	5.3×10^{-8}
	rs183211	42143493	2.0×10^{-10}	6.3×10^{-7}	3.9×10^{-6}
	rs169201	42145386	9.0×10^{-16}	1.0×10^{-8}	8.8×10^{-8}
	rs7224296	42155230	9.4×10^{-9}	1.1×10^{-5}	3.3×10^{-5}
	rs199533	42184098	2.3×10^{-16}	7.4×10^{-9}	8.0×10^{-8}
21q22.3	rs681210	43603771	1.1×10^{-5}	0.60	0.43

Method 1: Likelihood-based Method (Section III-2.3)

Method 2: Adjusted Allelic Analysis (Section III-2.4)

Method 3: The Original Armitage's Trend Test (Section II-1.4.3)

There have been a total of twenty five candidate genes discovered so far to predispose the Parkinson disease (the OMIM database with entry 168600). Listed in Table III-13 are the most significant SNP within a 2.5 Mb chromosome region surrounding each of the 25 Parkinson' disease (PD) candidate genes and the estimate of the number of false discoveries evaluated at the probability at which the SNP was claimed significant (Storey and Tibshirani, 2003). It can be seen that all the three methods detected the SNP, rs2736990, within the PD candidate gene *SNCA* on human chromosome 4q21 and, the SNP, rs199533, only 0.72Mb distant from the PD candidate gene *MAPT* on chromosome 17q21 with negligible risk of false

positive. In addition, Method 1 discovered additional two SNPs which were only 1.18 Mb and 0.18 Mb distant to the PD candidates, *FGF20* and *PARK8* respectively without invoking the risk of false positive. This positive proof analysis again supports the improved efficiency of the newly developed method for genetic association study.

Table III-13. Significant Markers Detected by Method 1 from Combined Stage I & II Data.

Locus	SNP	Position (bp)	P value		
			Method I	Method II	Method III
1p21.1	rs12172730	104913058	1.6×10^{-4}	2.2×10^{-4}	1.1×10^{-4}
2p25.3	rs6542651	3737705	4.0×10^{-4}	3.6×10^{-5}	1.4×10^{-5}
2p24.2	rs2042079	16969333	1.3×10^{-6}	1.4×10^{-5}	8.2×10^{-6}
2p22.3	rs935920	35985642	3.2×10^{-7}	2.3×10^{-5}	2.5×10^{-5}
2p22.3	rs2949065	35986447	5.7×10^{-7}	3.8×10^{-5}	4.5×10^{-5}
2p21	rs935378	46962655	3.8×10^{-6}	6.5×10^{-4}	4.7×10^{-4}
2q36.3	rs1035833	230417442	1.2×10^{-4}	4.2×10^{-2}	2.7×10^{-2}
3p25.1	rs7651825	14787122	2.2×10^{-6}	5.6×10^{-5}	5.2×10^{-5}
3q23	rs6440096	143814050	3.3×10^{-6}	2.9×10^{-4}	4.3×10^{-4}
	rs6800573	143828507	1.4×10^{-5}	3.2×10^{-4}	5.7×10^{-4}
	rs1453815	24566751	9.0×10^{-5}	1.0×10^{-3}	7.1×10^{-4}
4p15.2	rs4697508	24576450	2.9×10^{-5}	7.4×10^{-4}	5.8×10^{-4}
	rs7666265	77395305	1.4×10^{-4}	3.1×10^{-4}	5.4×10^{-4}
4q21.1	rs6851219	77398854	2.8×10^{-6}	2.9×10^{-5}	3.6×10^{-5}
	rs6812193	77418010	8.5×10^{-8}	6.7×10^{-7}	5.9×10^{-7}
4q21	rs1430961	90771943	5.6×10^{-6}	3.6×10^{-5}	4.4×10^{-6}
	rs12644119	90822442	1.2×10^{-9}	3.8×10^{-7}	2.3×10^{-8}
	rs356229	90825620	6.4×10^{-11}	5.5×10^{-7}	2.3×10^{-9}
	rs11931074	90858538	7.0×10^{-19}	2.3×10^{-14}	2.6×10^{-15}
	rs3857059	90894261	2.4×10^{-19}	1.6×10^{-14}	1.9×10^{-15}
	rs2736990	90897564	5.7×10^{-23}	6.1×10^{-15}	3.3×10^{-16}
	rs3775439	90928764	2.8×10^{-11}	2.2×10^{-8}	2.2×10^{-8}
	rs894278	90953558	7.4×10^{-11}	1.4×10^{-8}	2.9×10^{-9}
4q22	rs6532197	91016324	2.1×10^{-8}	2.1×10^{-7}	5.7×10^{-8}
5p15.2	rs26286	14219402	5.9×10^{-9}	7.2×10^{-5}	4.7×10^{-5}
5q11.2-q13.3	rs3792738	76283540	2.8×10^{-6}	1.7×10^{-5}	2.5×10^{-5}
5q23.3	rs264122	129675680	5.8×10^{-5}	1.6×10^{-3}	6.6×10^{-3}
6q22	rs6903627	120169406	1.0×10^{-4}	3.0×10^{-2}	4.9×10^{-2}
6q23.3	rs996243	137895153	5.4×10^{-6}	1.5×10^{-4}	9.0×10^{-4}
7p21.3	rs2681051	11615690	4.1×10^{-5}	4.0×10^{-2}	1.3×10^{-1}
7p12.3	rs2708909	48018204	1.4×10^{-5}	3.6×10^{-3}	2.6×10^{-3}
	rs2708851	48052327	2.8×10^{-5}	4.4×10^{-3}	3.8×10^{-3}
8p21.2	rs925030	25298518	1.3×10^{-4}	3.7×10^{-4}	3.4×10^{-4}
9p24.3	rs4742236	676753	7.6×10^{-5}	3.8×10^{-4}	3.3×10^{-4}
	rs9299039	680460	7.9×10^{-5}	1.2×10^{-3}	6.6×10^{-4}
9p21.2	rs4534200	25642508	8.6×10^{-5}	1.5×10^{-3}	2.0×10^{-3}
	rs700802	78424532	2.5×10^{-5}	2.6×10^{-3}	2.9×10^{-3}
9q21.31	rs7024926	82766092	1.0×10^{-5}	1.1×10^{-3}	4.6×10^{-4}
	rs2378554	82805138	2.3×10^{-5}	4.2×10^{-4}	5.0×10^{-4}

	rs9918939	82836491	3.2×10^{-5}	4.2×10^{-4}	4.7×10^{-4}
9q34.11	rs2240914	131938127	9.1×10^{-5}	1.1×10^{-2}	1.5×10^{-2}
9q34.2	rs11185726	136240925	6.0×10^{-7}	3.3×10^{-4}	1.4×10^{-3}
10p13	rs7077361	15601549	4.7×10^{-9}	3.0×10^{-5}	7.3×10^{-6}
10p12.1	rs11595185	25231376	1.0×10^{-7}	2.3×10^{-5}	2.5×10^{-5}
10p11.21	rs11591754	35247159	3.2×10^{-6}	9.4×10^{-5}	2.7×10^{-4}
10q22.1	rs2491015	70436819	8.2×10^{-7}	1.0×10^{-4}	2.1×10^{-4}
10q24.32	rs999867	104494554	6.6×10^{-8}	2.6×10^{-5}	1.8×10^{-6}
	rs17115100	104581383	1.4×10^{-8}	3.6×10^{-6}	1.4×10^{-7}
	rs3824754	104604340	4.6×10^{-7}	5.1×10^{-5}	3.6×10^{-6}
	rs4409766	104606653	2.6×10^{-7}	2.2×10^{-5}	1.3×10^{-6}
	rs11191425	104615960	1.2×10^{-7}	1.5×10^{-5}	5.4×10^{-7}
	rs12411886	104675289	9.0×10^{-7}	3.7×10^{-5}	1.8×10^{-6}
	rs12413409	104709086	8.6×10^{-7}	3.6×10^{-5}	1.6×10^{-6}
10q26	rs12777747	123989646	1.0×10^{-5}	2.7×10^{-4}	2.5×10^{-4}
11q12-q13.1	rs1005511	57123232	1.3×10^{-9}	5.3×10^{-4}	5.9×10^{-4}
12p12.1	rs699038	25050907	1.5×10^{-5}	2.8×10^{-4}	2.8×10^{-4}
12q12	rs11564162	38729159	3.9×10^{-9}	2.1×10^{-5}	1.3×10^{-4}
	rs1491923	38877384	2.5×10^{-5}	5.3×10^{-5}	1.5×10^{-5}
12q13.11	rs1793949	46657862	1.8×10^{-6}	3.0×10^{-4}	2.1×10^{-4}
12q13.2	rs2710697	53684257	3.6×10^{-6}	1.3×10^{-3}	4.8×10^{-3}
13q14.11	rs9525776	42928966	6.7×10^{-6}	5.4×10^{-2}	4.3×10^{-2}
13q22.2	rs9530494	75434275	2.1×10^{-5}	4.6×10^{-4}	3.5×10^{-4}
14q22-q23	rs2150279	53343338	9.0×10^{-5}	1.2×10^{-3}	6.7×10^{-4}
	rs12431733	53360580	4.5×10^{-8}	3.5×10^{-6}	2.3×10^{-6}
14q22.2	rs2878172	54443420	1.2×10^{-9}	2.8×10^{-4}	6.4×10^{-4}
15q22	rs1481088	71629314	1.4×10^{-5}	5.4×10^{-5}	4.7×10^{-5}
	rs922687	71635861	1.3×10^{-6}	2.1×10^{-5}	1.4×10^{-5}
17p13.3	rs4247113	228978	4.2×10^{-5}	2.3×10^{-4}	9.8×10^{-4}
17q21	rs11012	40869224	1.1×10^{-16}	9.7×10^{-10}	1.7×10^{-10}
	rs393152	41074926	5.5×10^{-27}	5.0×10^{-15}	4.4×10^{-16}
	rs417968	41084159	6.2×10^{-17}	4.7×10^{-10}	7.0×10^{-11}
	rs7215239	41123556	1.5×10^{-19}	1.4×10^{-10}	3.0×10^{-11}
	rs1526123	41139123	5.9×10^{-8}	1.4×10^{-6}	1.3×10^{-7}
	rs12373139	41279910	9.7×10^{-23}	4.7×10^{-13}	1.5×10^{-13}
	rs17690703	41281077	3.4×10^{-22}	6.9×10^{-12}	1.4×10^{-12}
	rs17563986	41347100	3.3×10^{-24}	1.0×10^{-13}	2.4×10^{-14}
	rs1981997	41412603	1.2×10^{-23}	2.0×10^{-13}	5.3×10^{-14}
	rs8070723	41436901	4.5×10^{-23}	3.6×10^{-13}	1.0×10^{-13}
	rs7225002	41544850	3.1×10^{-13}	7.4×10^{-8}	1.0×10^{-7}
	rs2532274	41602941	3.8×10^{-16}	1.9×10^{-10}	1.2×10^{-10}
	rs2532269	41605885	3.6×10^{-22}	3.1×10^{-12}	9.5×10^{-13}
	rs2668692	41648797	4.7×10^{-23}	3.7×10^{-13}	1.2×10^{-13}
	rs183211	42143493	4.2×10^{-16}	3.4×10^{-10}	1.4×10^{-10}
	rs169201	42145386	6.2×10^{-24}	1.1×10^{-13}	5.9×10^{-14}
	rs7224296	42155230	2.8×10^{-13}	4.5×10^{-8}	1.8×10^{-8}
	rs199533	42184098	7.7×10^{-25}	3.9×10^{-14}	2.5×10^{-14}
20p12.1	rs1223271	13244912	6.0×10^{-7}	1.3×10^{-5}	6.8×10^{-6}
21q22.3	rs681210	43603771	7.4×10^{-6}	3.9×10^{-3}	2.2×10^{-3}
	rs595046	43626445	1.3×10^{-4}	1.5×10^{-2}	2.4×10^{-2}
22q13.32	rs4823506	46643740	1.0×10^{-4}	4.9×10^{-3}	6.5×10^{-3}

Method I: Likelihood-based Method (Section III-2.3)

Method II: Adjusted Allelic Analysis (Section III-2.4)

2.7 Conclusion and Discussion

In summary, we have shown that the existing widely used strategy in the current literature of genome-wide genetic association studies with a case-control setting is highly vulnerable to sampling schemes and genetic structure embedded in the samples. These can result in severe loss of statistical power or false positive inference of association. We have developed a novel method that is robust to these influential factors and confers a more powerful test. Although the method is developed for complex quantitative traits with discrete phenotype, it will not involve major technical problems to extend the ideas and principles behind the newly developed method to cope with continuous phenotype.

The robustness and improved statistical power of the newly developed method have been demonstrated through re-analysing the large SNP genotype dataset of the Parkinson's disease, of which cases and controls were collected from multiple geographical cohorts (Simon-Sanchez et al., 2009), as well as through intensive simulation studies. The method is built upon the population genetic model of linkage disequilibrium between any tested polymorphic genetic marker and a putative QTL. The simulation study indicates that the key model parameters of q , allele frequency of a disease locus, of which genotypes are not observable, and D , the coefficient of linkage disequilibrium, can be estimated adequately under quite different settings. As the accurate estimate of the LD coefficient is crucial for the reliability of any LD analysis including LD-based mapping of complex genetic disease traits (Hill and Weir, 1994), this may explain the outperformance of the parametric approach over the existing non-parametric rival, i.e. the allelic analysis and the Armitage's trend test.

The new method is built on a model-based likelihood framework. This confers several primary and practically useful statistical properties over the existing non-parametric approaches. Firstly, it is feasible to be adopted to incorporate different fixed or random effects in the model. For example, onset of many common diseases is sex, age, diet or life habit dependent. Incorporation of these as fixed effects into the model shall improve the statistical efficiency of genetic association analysis. Secondly, the approach enables data from different sources to be combined, as demonstrated in the case and control data analysis of the Parkinson's disease. This provides flexibility for performing association studies using common control datasets. Thirdly, the parameter estimates from the method make it feasible to evaluate statistical and genetic properties of detected associations.

However, besides all the advantages above, the background gene effect or genetic heterogeneity has not been implemented into the likelihood model, which will introduce certain degrees of bias as has been demonstrated in Chapter II-3. As our newly proposed method shows dominant performance over the existing ones, if the effect from genetic heterogeneity could be properly introduced into a likelihood-based method, we may expect a further increase of statistical power. On the other hand, the population stratification has already assumed to be known in our new method, which might not always be true in practice. In fact, we have been actively exploring appropriate statistical strategies that tackle this issue properly and appropriately.

CHAPTER III-3

A COMPOSITE LIKELIHOOD-BASED MODEL FOR ASSOCIATION-BASED MAPPING OF QUANTITATIVE TRAIT LOCI

3.1 Overview

Interval mapping was initially introduced by Lander & Botstein (1989) for linkage analysis so that people could screen for QTLs on the whole genome continuously instead of discretely at each of marker locus through the use of two flanking markers, out of which no further information of inferring LD within the interval could be acquired. Such a method, as well as its successors (Zeng, 1993, Zeng, 1994), are based on linkage analysis and hence are highly limited in their mapping resolution and applicable to only certain type of data as discussed in section I-1.2. In association studies, Terwilliger (1995) and Devlin et al. (1996) introduced their composite likelihood-based methods of inferring QTL within marker intervals using data collected from natural populations. However, both of these methods are restricted to binary data. In this chapter, we will introduce a composite likelihood-based method for mapping loci affecting a quantitative trait with continuous phenotypic distributions using a population genetics model. A series of simulations are conducted to demonstrate the reliability of our newly proposed method.

3.2 Notations and Models

Recall formula (I-1.2) as introduced in section I-1.2.3 that

$$D_T = (1-r)^T D_0, \quad (\text{III-3.1})$$

where r is the recombination fraction between the testing marker locus and a putative QTL; T is the number of generations from the introduction of the mutant allele at the QTL, i.e. the recombination generation; D_T and D_0 denotes the LD coefficients at generation T and 0 respectively. Following the same notation in Table I-1, at generation 0, the LD coefficient should theoretically reach its maximum or minimum and hence one may write

$$D_0 = \min(p, q) - p \times q, \quad (\text{III-3.2})$$

where p , q denote the allele frequency of marker allele M and QTL allele A respectively. Without loss of generality, we have assumed alleles M and A to be positively associated in formula (III-3.2), and we will also keep this assumption in the following discussion for convenience.

From formulae (III-3.1) and (III-3.2), D_T is now a function of allele frequencies p and q instead of an independent parameter, and hence if both p and q could be properly estimated, the LD coefficient between a testing marker locus and a putative QTL could be directly calculated with known r and T . For instance, we might assume a putative QTL located at any a locus in the genome without restricting on a particular marker locus, and by estimating the most possible allele frequency at the putative QTL, we might directly calculate the LD coefficient between the putative QTL and its surrounding marker loci. With a LD coefficient hence estimated, we may expect a significant increase in mapping resolutions. However, as the validation of such an estimate of LD coefficient relies on known values of r and T , we thus need to acquire the information of both those parameters prior to the establishment of our new approaches.

Since the development of genetic technologies (Kong et al., 2002, Frazer et al., 2007, Altshuler et al., 2010), either physical or genetic mapping positions of a particular marker is available nowadays. Since the genetic distance between any two marker loci may be properly

defined as the recombination fraction between them, the genetic mapping could directly provide the information of r . Alternatively, if only physical mapping data are available, we may transfer the physical mapping to genetic mapping through an appropriate mapping function, e.g. $r = 0.5(1 - e^{-2d})$ proposed by Haldane (1919), where r and d are the genetic and physical distances between two genetic loci respectively. With above discussions, if the genetic distance between any a pair of genetic markers is either known or could be estimated, the genetic distances between a putative QTL, within the marker interval, and both the flanking markers are hence known. On the other hand, the recombination generation, i.e. T , could be either known, e.g. in a well established experimental population, or directly estimated from samples collected from a random mating population. For instance, if a sample is randomly collected from an ideal population, we may apply Hill's method (Hill, 1974) as introduced in section III-1.3 to estimate the LD coefficient between any a pair of marker loci. In case non-randomness is present, our newly proposed method as introduced in section III-1.4 could be implemented to correct the bias introduced by the non-randomness of samples. As D_0 could be directly calculated from formula (III-3.2), with the estimate of D , we could compute the corresponding recombination generation as $T = (\log D - \log D_0) / \log(1 - r)$, where r is assumed to be known as we have discussed above. Practically, we would prefer to use the average T throughout the whole genome instead of an estimate between a particular marker pair. Such a consideration is quite reasonable if we are not dealing with a newly introduced mutation. From above discussions, as the information of r and T could either be directly acquired or estimated from highly accessible resources, we could hence establish a likelihood function between a putative quantitative QTL and a marker locus in a randomly collected sample as following.

Consider a randomly collected sample with quantitative phenotypes y_{ij} of the j th individual with the i th marker genotype, and the full sample size $n = \sum_{i=1}^3 n_i$, where n_i denotes the number of individuals with the i th marker genotype. Assuming the normality of phenotypes that $y_{ij} \sim N(\mu_k, \nu)$ conditioned on the k th QTL genotype, we could hence write the probability of a single individual with phenotype y_{ij} and the k th QTL genotype.

$$\Pr(y_{ij} | G = k, \Omega) = \frac{1}{\sqrt{2\pi\nu}} e^{[-(y_{ij} - \mu_k)^2 / 2\nu]}, \quad (\text{III-3.3})$$

where $\mu_k = \mu + (2-k)a + \frac{1-(-1)^k}{2}d$, with $k = 1, 2, 3$, denotes the genotypic value of the k th QTL genotype as shown in Table II-1; G represents the QTL genotype and $\Omega = (p, q, \mu, a, d, \nu, r, T)$ represent the set of all modelling parameters.

As from function (II-1.10), by assuming the independence among individuals, we could write the complete likelihood function of the observations as

$$L(\Omega | y, M) = \prod_{i=1}^3 \prod_{j=1}^{n_i} \Pr(y_{ij}, M = i | \Omega),$$

where M denotes the marker genotype. Following the ECM scheme of a randomly collected sample as introduced in section II-1.3.3, we could write the expectation log-likelihood function at the s th iterative ECM process, i.e. function (II-1.11), as

$$l_c(\Omega^s | y, M) = \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 \left[\omega_{ijk}^s \log h_{ik}^s \Pr(y_{ij} | G = k, \Omega^s) \right], \quad (\text{III-3.4})$$

where $h_{ik}^s = \Pr(M = i, G = k | \Omega^s)$ is the joint distribution of marker-QTL genotypes at s th iterative process, and

$$\begin{aligned}\omega_{ijk}^s &= \Pr(G = k \mid y_{ij}, M = i, \Omega^s) \\ &= \frac{h_{ik}^s \Pr(y_{ij} \mid G = k, \Omega^s)}{\sum_{k=1}^3 h_{ik}^s \Pr(y_{ij} \mid G = k, \Omega^s)} .\end{aligned}\quad (\text{III-3.5})$$

By noticing that p could be directly estimated as $(n_1 + n_2 / 2) / n$, and the estimates of r and T are also available as introduced above, we could hence establish the ECM algorithm of function (III-3.4) as:

E-step: Calculate the posterior probability of one individual having the k th QTL genotype given the i th marker genotype and its phenotype y_{ij} at the s th iteration, says ω_{ijk}^s . If the values of parameters at the s th iteration are denoted as $\Omega^s = (p, q^s, \mu^s, a^s, d^s, v^s, r, T)$, the posterior probabilities can be written as:

$$\omega_{ijk}^s = \frac{h_{ik}^s e_{ijk}^s}{\sum_{l=1}^3 h_{il}^s e_{ijl}^s} ,$$

where $e_{ijk}^s = \exp[-(y_{ij} - \mu_k^s)^2 / 2v^s]$ and $\mu_k^s = \mu^s + (2 - k)a^s + \frac{1 - (-1)^k}{2} d^s$.

CM-step: With ω_{ijk}^s calculated from an E-step, the updated estimate of q^{s+1} could be achieved through solving the following equation:

$$\frac{\partial l_c(\Omega^s \mid Y, M)}{\partial q} = \sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 \left\{ \omega_{ijk}^s \times \frac{\partial [\log(h_{ik}^s)]}{\partial q} \right\} = 0 , \quad (\text{III-3.6})$$

which is equivalent to a three degree polynomial equation

$$a_3 q^3 + a_2 q^2 + a_1 q + a_0 = 0 , \quad (\text{III-3.7})$$

where

$$a_3 = 4n(1 - rt)^2 (p + (1 - p)rt)^2,$$

$$a_2 = -(p(1 - rt) + rt)(1 - rt) \left(rt(6\tilde{\omega}_{11} + 6\tilde{\omega}_{12} + 6\tilde{\omega}_{13} + 6\tilde{\omega}_{21} + 5\tilde{\omega}_{22} + 4\tilde{\omega}_{23} + 6\tilde{\omega}_{31} + 4\tilde{\omega}_{32} + 2\tilde{\omega}_{33}) \right. \\ \left. + 2p(1 - rt)(6\tilde{\omega}_{11} + 5\tilde{\omega}_{12} + 4\tilde{\omega}_{13} + 6\tilde{\omega}_{21} + 5\tilde{\omega}_{22} + 4\tilde{\omega}_{23} + 6\tilde{\omega}_{31} + 5\tilde{\omega}_{32} + 4\tilde{\omega}_{33}) \right),$$

$$a_1 = rt^2(2\tilde{\omega}_{11} + 2\tilde{\omega}_{12} + 2\tilde{\omega}_{13} + 2\tilde{\omega}_{21} + \tilde{\omega}_{22} + \tilde{\omega}_{23} + 2\tilde{\omega}_{31} + \tilde{\omega}_{32}) \\ + 4p^2(1 - rt)^2(3\tilde{\omega}_{11} + 2\tilde{\omega}_{12} + \tilde{\omega}_{13} + 3\tilde{\omega}_{21} + 2\tilde{\omega}_{22} + \tilde{\omega}_{23} + 3\tilde{\omega}_{31} + 2\tilde{\omega}_{32} + \tilde{\omega}_{33}) \\ + p(1 - rt)rt(12\tilde{\omega}_{11} + 9\tilde{\omega}_{12} + 6\tilde{\omega}_{13} + 12\tilde{\omega}_{21} + 8\tilde{\omega}_{22} + 4\tilde{\omega}_{23} + 12\tilde{\omega}_{31} + 7\tilde{\omega}_{32} + 2\tilde{\omega}_{33})$$

and

$$a_0 = -p(2p(1 - rt) + rt)(2\tilde{\omega}_{11} + \tilde{\omega}_{12} + 2\tilde{\omega}_{21} + \tilde{\omega}_{22} + 2\tilde{\omega}_{31} + \tilde{\omega}_{32}),$$

where $rt = (1 - r)^T$ and $\tilde{\omega}_{ik} = \sum_j^{n_i} \omega_{ijk}$. Because the highest degree of formula (III-3.7) is 3, it

could hence be analytically solved. Note here, as equation (III-3.7) is derived under the assumption that $q \leq p$, i.e. equation (III-3.2) takes the form $D_0 = q(1 - p)$, in each iteration roots acquired through solving equation (III-3.7) have to fall within the range $[0, p]$. If none of the roots could meet such a requirement, we simply carried on the iterative process by letting $q^{s+1} = p$.

The updated estimates of the remaining parameters, i.e. μ , a , d and ν , keep similar forms as given in Luo et al. (2000) as

$$\mu^{s+1} = \frac{1}{n} \left[\sum_{i=1}^3 \sum_{j=1}^{n_i} y_{ij} + \sum_{i=1}^3 \sum_{j=1}^{n_i} (\omega_{ij3}^s - \omega_{ij1}^s) a^s + \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^{n_i} \omega_{ij2}^s d^s \right],$$

$$a^{s+1} = \frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} (\omega_{ij1}^s - \omega_{ij3}^s)(y_{ij} - \mu^{s+1})}{\sum_{i=1}^3 \sum_{j=1}^{n_i} (\omega_{ij1}^s + \omega_{ij3}^s)},$$

$$d^{s+1} = \frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} \omega_{ij2}^s (y_{ij} - \mu^{s+1})}{\sum_{i=1}^3 \sum_{j=1}^{n_i} \omega_{ij2}^s},$$

and

$$v^{s+1} = \frac{\sum_{i=1}^3 \sum_{j=1}^{n_i} \sum_{k=1}^3 \omega_{ijk}^s (y_{ij} - \mu_k^{s+1})^2}{n}.$$

Repeating these E and CM steps iteratively until a given convergence criterion is satisfied, the converged values of these parameters are hence their MLEs according to the theory of ECM algorithm (Meng and Rubin, 1993).

Note here, if the assumption $q \leq p$ is not true, we may expect that when the convergent criterion is satisfied, the MLE of q would be equal to the estimate of p given the restriction we have set in the E-steps, and hence such an MLE is in-conclusive. However, in the case where positively associated marker allele M and QTL allele A have the relationship in their frequency as $f(A) > f(M)$, there must be coincidentally $f(a) < f(m)$ with alleles a and m positively associated, where a and m are the complementary alleles of A and M at their loci respectively, and hence if the MLE of q equals the estimates of p , we could simply adopt $p' = 1 - p$ to repeat the ECM algorithm to avoid the problem.

The corresponding LRT for the null hypothesis $D = 0$ could hence be given as

$$LR = 2[l(\hat{\Omega} | y, M) - l(\hat{\Omega}_{D=0} | y, M)] \sim \chi^2$$

with 1 degree of freedom.

Above, we have established the likelihood function between a putative QTL and a testing marker locus based on the known genetic distance between them as well as the known recombination generation. A more comprehensive idea would be the establishment of a likelihood function which could integrate the information of all genetic markers surrounding the putative QTL. For example, we may seek to model the joint probability of phenotypes

with both flanking markers, i.e. $f(M_1, Y, M_2 | \Omega)$, where M_1 and M_2 denote the two flanking markers of a putative QTL respectively. However, as the directly modelling of such a joint probability is hard to implement, we may turn to alternative strategies, e.g. the composite likelihood, which could provide efficient combination of information from multiple marker loci together to infer the most possible location of a putative QTL. From Devlin et al. (1996), it is possible to use a composite likelihood model instead of a complete likelihood model. Suppose there are two observations, says y_1 and y_2 , the complete likelihood function should be $L(\Omega | y_1, y_2) = f(y_1, y_2 | \Omega)$, the logarithm form of which could be written into $l(\Omega | y_1, y_2) = \log f(y_2 | y_1, \Omega) + \log f(y_1 | \Omega)$. Alternatively, we could choose to follow either of two different ‘natural choices’ of composite likelihood models instead of the complete one that

$$l'(\Omega | y_1, y_2) = \log f(y_1 | \Omega) + \log f(y_2 | \Omega)$$

or

$$l'(\Omega | y_1, y_2) = \log f(y_1 | y_2, \Omega) + \log f(y_2 | y_1, \Omega).$$

The choice may be mainly based on the convenience of computing (Devlin et al., 1996). In our situation, where the putative QTL falls within the interval of two flanking markers, i.e. M_1 and M_2 , we may write the composite log-likelihood function as

$$l'(\Omega | y, M_1, M_2) = l(\Omega | y, M_1) + l(\Omega | y, M_2),$$

and the corresponding composite LRT could be given as

$$\begin{aligned} LR_{M_1 Q M_2} &= 2(l_{M_1 Q} + l_{M_2 Q} - l_{M_1 Q} |_{D_{M_1 Q}=0} - l_{M_2 Q} |_{D_{M_2 Q}=0}) \\ &= LR_{M_1 Q} + LR_{M_2 Q}, \end{aligned}$$

where $l_{M_1 Q} = l(\Omega | y, M_1)$ and etc.. Note here, if a putative QTL is assumed to be located at one of the marker loci exactly, say M_m , we may alternatively write

$$\begin{aligned}
LR_{M_1QM_mM_2} &= \frac{LR_{M_1QM_m} + LR_{M_mQM_2}}{2} \\
&= \frac{LR_{M_1Q} + 2LR_{M_mQ} + LR_{M_2Q}}{2} .
\end{aligned}$$

Unfortunately, due to the dependence between marker loci M_1 and M_2 , both composite LRTs presented above, i.e. $LR_{M_1QM_2}$ and $LR_{M_1QM_mM_2}$, do not follow the exact χ^2 distribution with 2 degrees of freedom, as they were supposed to be only if M_1 and M_2 are independent, and hence alternative evaluations have to be adopted to determine the significant thresholds (Devlin et al., 1996). However, in the presence of a true QTL, a locally highest LR statistic does provide an estimate of QTL location, and as our main interest would be the accuracy of our presented method to infer the location of a QTL within the marker interval, only the locus with the highest LR statistic will be mentioned but the distribution of test statistic will not be focused in this context.

3.3 Simulation Studies

3.3.1 Simulation Models

In this simulation study, we will mimic a species with a single pair of chromosomes, and then a finite population of such a species under certain generations of randomly mating will be simulated to evaluate our newly presented method. For simplicity, we will assume all genetic loci, i.e. both marker and QTL, to be bi-allelic with initial allele frequencies 0.5 at generation 0 and the genetic distances between any two adjacent marker loci are identical. Each individual in such a population is assumed to have the same chance to give its offspring but cross-generation mating is prohibited.

To generate such a population as described above, we start from an initial population, i.e. generation 0, with size n , and then a pool of gametes are simulated in the presence of

recombination events to randomly generate the zygotes for the next generation with the same population size n . Such a process is repeated until the population at generation T has been generated. Then, parameters of the phenotypic effects, i.e. population mean μ , additive effect a , dominant effect d , phenotype variance v , are implemented to simulate a value of phenotype for each individual as the observations based on a normal distribution as introduced at function (III-3.3). Throughout simulated schemes, μ is equal to 10.0, a is equal to 0.63, d is equal to 0.0 and v is 1.0, which indicates that the co-dominance model is adopted and the genotypic variance can explain 20% of the phenotypic variance. Irrespective of the various genetic distances between pairs of adjacent markers in different simulation schemes, there are always 41 markers equally spaced on the testing chromosome, and the true QTL is located between the 28th and 29th markers, except for the linked QTLs analyses, where the QTLs will be specified otherwise. Simulations for each set of parameters are replicated 100 times, and the test results of each replicate as well as the value of parameters used for the simulations are summarized in Table III-14 to Table III-19.

Table III-14. Comparison of Simulation Results with Real and Estimated *T*.

Fixed means during the process of estimating parameters, the recombination generation *T* is not from estimation but considered as known, and it was kept constant for each of 100 replicates. Simulation parameters are summarized in columns 'Simulation Model' and 'Simulation Parameters' and the means and standard deviations of the corresponding estimates are summarized in column 'Means of Estimates'. If the estimated QTL location for each replicate is within 0.10 Mpd to its true location, i.e. listed in column QTL of 'Simulation Model', it will be claimed as 'accurate', and the overall percentages of claimed 'accurate' throughout 100 replicates are listed in column 'Percentage of Accuracy'. The mean of estimated *T* for each population is listed in column 'Average *T*'.

Simulation Model					Simulation Parameters				Means of Estimates (with Standard Deviations)					Percentage of Accuracy (Biased less or equal to 0.1)	Average <i>T</i>
Pop.	N	mpd(c M)	T	QTL	U	A	d	v	QTL	U	A	D	v		
1	300	5	35	28.40	10.000	0.630	0.000	1.000	28.42 (0.19)	9.976 (0.063)	0.295 (0.046)	0.002 (0.058)	0.957 (0.039)	71%	35
2	300	5	35	28.40	10.000	0.630	0.000	1.000	28.39 (0.19)	9.972 (0.070)	0.295 (0.049)	0.001 (0.049)	0.949 (0.046)	73%	35(fixed)
3	1000	5	40	28.40	10.000	0.630	0.000	1.000	28.41 (3.159)	9.996 (0.148)	0.233 (0.156)	-0.005 (0.125)	0.937 (0.073)	32%	37
4	1000	5	40	28.40	10.000	0.630	0.000	1.000	28.97 (2.656)	9.994 (0.139)	0.239 (0.144)	0.021 (0.125)	0.928 (0.078)	28%	40(fixed)

Table III-15. Simulation Results for QTL Locates at 28.2 with 5cM Mpd.

Simulation parameters are summarized in columns 'Simulation Model' and 'Simulation Parameters' and the means and standard deviations of the corresponding estimates are summarized in column 'Means of Estimates'. If the estimated QTL location for each replicate is within 0.10 Mpd to its true location, i.e. listed in column QTL of 'Simulation Model', it will be claimed as 'accurate', and the overall percentages of claimed 'accurate' throughout 100 replicates are listed in column 'Percentage of Accuracy'. The mean of estimated T for each population is listed in column 'Average T '.

Simulation Model					Simulation Parameters				Estimates (with Standard Deviation)					Percentage of Accuracy (Biased less or equal to 0.1)	Average T
Pop.	N	mpd(c M)	T	QTL	u	a	d	v	QTL	U	A	D	v		
1	300	5	30	28.20	10.000	0.630	0.000	1.000	28.15 (0.62)	9.971 (0.103)	0.359 (0.113)	0.001 (0.100)	0.906 (0.088)	30%	29
2	300	5	35	28.20	10.000	0.630	0.000	1.000	28.07 (0.34)	9.991 (0.106)	0.349 (0.090)	-0.012 (0.105)	0.907 (0.068)	30%	34
3	300	5	40	28.20	10.000	0.630	0.000	1.000	28.08 (0.82)	9.958 (0.150)	0.320 (0.109)	0.017 (0.104)	0.913 (0.074)	48%	37
4	1000	5	35	28.20	10.000	0.630	0.000	1.000	28.22 (0.23)	9.979 (0.048)	0.333 (0.053)	-0.003 (0.066)	0.928 (0.041)	50%	35
5	1000	5	40	28.20	10.000	0.630	0.000	1.000	28.20 (0.19)	9.986 (0.071)	0.316 (0.061)	0.003 (0.059)	0.939 (0.042)	69%	41
6	1000	5	45	28.20	10.000	0.630	0.000	1.000	28.10 (0.25)	9.969 (0.070)	0.293 (0.058)	0.001 (0.055)	0.948 (0.043)	68%	46
7	3000	5	40	28.20	10.000	0.630	0.000	1.000	28.24 (0.15)	9.979 (0.042)	0.304 (0.034)	-0.003 (0.003)	0.945 (0.028)	82%	40
8	3000	5	45	28.20	10.000	0.630	0.000	1.000	28.23 (0.14)	9.975 (0.038)	0.287 (0.040)	0.004 (0.034)	0.947 (0.023)	88%	46

Table III-16. Simulation Results for QTL Locates at 28.4 with 5cM Mpd.

Simulation parameters are summarized in columns 'Simulation Model' and 'Simulation Parameters' and the means and standard deviations of the corresponding estimates are summarized in column 'Means of Estimates'. If the estimated QTL location for each replicate is within 0.10 Mpd to its true location, i.e. listed in column QTL of 'Simulation Model', it will be claimed as 'accurate', and the overall percentages of claimed 'accurate' throughout 100 replicates are listed in column 'Percentage of Accuracy'. The mean of estimated *T* for each population is listed in column 'Average T'.

Simulation Model					Simulation Parameters				Estimates (with Standard Deviation)					Percentage of Accuracy (Biased less or equal to 0.1)	Average T
Pop.	N	mpd(c M)	T	QTL	u	a	d	v	QTL	U	a	d	v		
1	300	5	30	28.40	10.000	0.630	0.000	1.000	28.19 (0.86)	9.957 (0.095)	0.334 (0.120)	0.009 (0.108)	0.938 (0.079)	39%	28
2	300	5	35	28.40	10.000	0.630	0.000	1.000	28.31 (1.42)	9.985 (0.123)	0.284 (0.132)	-0.024 (0.101)	0.930 (0.087)	35%	34
3	300	5	40	28.40	10.000	0.630	0.000	1.000	28.41 (3.16)	9.996 (0.148)	0.233 (0.156)	0.004 (0.125)	0.937 (0.073)	32%	37
4	1000	5	35	28.40	10.000	0.630	0.000	1.000	28.42 (0.19)	9.976 (0.063)	0.295 (0.046)	0.002 (0.058)	0.957 (0.039)	71%	35
5	1000	5	40	28.40	10.000	0.630	0.000	1.000	28.36 (0.23)	9.976 (0.077)	0.271 (0.059)	-0.007 (0.054)	0.964 (0.037)	75%	41
6	1000	5	45	28.40	10.000	0.630	0.000	1.000	28.36 (0.24)	9.967 (0.081)	0.228 (0.056)	-0.003 (0.051)	0.969 (0.039)	63%	45
7	3000	5	40	28.40	10.000	0.630	0.000	1.000	28.41 (0.08)	9.977 (0.044)	0.265 (0.028)	0.001 (0.028)	0.961 (0.027)	94%	40
8	3000	5	45	28.40	10.000	0.630	0.000	1.000	28.40 (0.10)	9.967 (0.051)	0.237 (0.032)	-0.006 (0.034)	0.970 (0.025)	90%	46

Table III-17. Simulation Results for 10cM Mpd.

Simulation parameters are summarized in columns ‘Simulation Model’ and ‘Simulation Parameters’ and the means and standard deviations of the corresponding estimates are summarized in column ‘Means of Estimates’. If the estimated QTL location for each replicate is within 0.10 Mpd to its true location, i.e. listed in column QTL of ‘Simulation Model’, it will be claimed as ‘accurate’, and the overall percentages of claimed ‘accurate’ throughout 100 replicates are listed in column ‘Percentage of Accuracy’. The mean of estimated T for each population is listed in column ‘Average T ’.

Simulation Model					Simulation Parameters				Estimates (with Standard Deviation)					Percentage of Accuracy (Biased less or equal to 0.1)	Average T
Pop	N	mpd(c M)	T	QTL	u	a	d	v	QTL	U	A	d	v		
1	300	10	10	28.20	10.000	0.630	0.000	1.000	28.28 (0.40)	9.996 (0.070)	0.419 (0.068)	0.011 (0.095)	0.885 (0.071)	29%	9
2	300	10	15	28.20	10.000	0.630	0.000	1.000	28.23 (0.33)	9.990 (0.088)	0.367 (0.077)	-0.023 (0.089)	0.914 (0.066)	30%	14
3	300	10	20	28.20	10.000	0.630	0.000	1.000	27.98 (0.93)	9.981 (0.093)	0.316 (0.010)	-0.017 (0.092)	0.928 (0.076)	41%	20
4	300	10	25	28.20	10.000	0.630	0.000	1.000	28.40 (2.04)	10.001 (0.115)	0.255 (0.109)	-0.018 (0.104)	0.948 (0.078)	51%	23
5	300	10	10	28.40	10.000	0.630	0.000	1.000	28.43 (0.32)	9.991 (0.070)	0.410 (0.075)	-0.016 (0.091)	0.909 (0.070)	45%	9
6	300	10	15	28.40	10.000	0.630	0.000	1.000	28.27 (0.93)	9.992 (0.088)	0.314 (0.086)	0.001 (0.089)	0.936 (0.070)	44%	15
7	300	10	20	28.40	10.000	0.630	0.000	1.000	28.52 (1.99)	9.990 (0.087)	0.251 (0.104)	0.003 (0.089)	0.940 (0.080)	42%	20
8	300	10	25	28.40	10.000	0.630	0.000	1.000	28.64 (2.71)	9.995 (0.116)	0.193 (0.117)	-0.013 (0.122)	0.962 (0.086)	22%	23

Table III-18. Simulation Results for 50cM Mpd.

Simulation parameters are summarized in columns ‘Simulation Model’ and ‘Simulation Parameters’ and the means and standard deviations of the corresponding estimates are summarized in column ‘Means of Estimates’. If the estimated QTL location for each replicate is within 0.10 Mpd to its true location, i.e. listed in column QTL of ‘Simulation Model’, it will be claimed as ‘accurate’, and the overall percentages of claimed ‘accurate’ throughout 100 replicates are listed in column ‘Percentage of Accuracy’. The mean of estimated *T* for each population is listed in column ‘Average *T*’.

Simulation Model					Simulation Parameters				Estimates (with Standard Deviation)					Percentage of Accuracy (Biased less or equal to 0.1)	Average <i>T</i>
Pop.	N	mpd(c M)	T	QTL	u	a	d	v	QTL	U	A	d	v		
1	300	50	3	28.20	10.000	0.630	0.000	1.000	28.15 (0.32)	9.993 (0.053)	0.342 (0.065)	0.011 (0.074)	0.919 (0.076)	45%	3
2	300	50	5	28.20	10.000	0.630	0.000	1.000	28.25 (1.74)	9.988 (0.061)	0.275 (0.103)	0.000 (0.092)	0.938 (0.068)	58%	5
3	300	50	3	28.40	10.000	0.630	0.000	1.000	28.46 (0.88)	9.994 (0.062)	0.321 (0.082)	0.007 (0.096)	0.955 (0.073)	53%	3
4	300	50	5	28.40	10.000	0.630	0.000	1.000	28.66 (2.90)	9.999 (0.068)	0.202 (0.086)	0.005 (0.091)	0.956 (0.083)	29%	5
5	1000	50	5	28.40	10.000	0.630	0.000	1.000	28.32 (0.16)	9.994 (0.038)	0.203 (0.035)	0.005 (0.050)	0.974 (0.043)	66%	6

Table III-19. Simulation Results for Twin QTLs.

Simulation parameters are summarized in columns ‘Simulation Model’ and ‘Simulation Parameters’ and the means and standard deviations of the corresponding estimates are summarized in column ‘Estimates’. The mean of estimated T for each population is listed in column ‘Average T ’.

Simulation Model					Simulation Parameters (Two QTL contribute equally)					Estimates (with Standard Deviation)										Average T
Pop.	N	mpd(c M)	T	QTL1	QTL2	U	a	d	v	QTL1	QTL2	u1	u2	a1	a2	d1	d2	v1	v2	
1	500	1	10	17.400	22.400	10.000	0.450	0.000	1.000	18.366 (0.995)	21.169 (1.002)	9.985 (0.050)	9.988 (0.051)	0.740 (0.062)	0.738 (0.065)	0.009 (0.087)	0.009 (0.081)	0.845 (0.059)	0.847 (0.061)	7
2	500	1	30	17.400	22.400	10.000	0.450	0.000	1.000	18.000 (0.837)	21.932 (0.841)	9.982 (0.077)	9.981 (0.074)	0.526 (0.071)	0.531 (0.069)	0.004 (0.101)	-0.016 (0.083)	0.885 (0.056)	0.885 (0.055)	23
3	500	1	50	17.400	22.400	10.000	0.450	0.000	1.000	17.400 (0.621)	22.281 (0.660)	9.960 (0.103)	9.958 (0.099)	0.438 (0.085)	0.448 (0.083)	0.004 (0.093)	0.003 (0.104)	0.913 (0.065)	0.909 (0.060)	41
4	500	1	10	17.400	22.400	10.000	0.320	0.000	1.000	18.371 (1.106)	21.352 (1.005)	10.004 (0.046)	10.003 (0.050)	0.524 (0.061)	0.524 (0.062)	0.005 (0.092)	0.005 (0.091)	0.925 (0.060)	0.924 (0.059)	7
5	500	1	30	17.400	22.400	10.000	0.320	0.000	1.000	17.905 (0.898)	21.827 (0.957)	9.987 (0.070)	9.985 (0.077)	0.395 (0.077)	0.388 (0.075)	0.006 (0.093)	0.016 (0.096)	0.945 (0.060)	0.950 (0.065)	23
6	500	1	50	17.400	22.400	10.000	0.320	0.000	1.000	17.600 (0.749)	22.205 (0.856)	9.966 (0.085)	9.964 (0.091)	0.322 (0.080)	0.327 (0.071)	0.003 (0.116)	0.016 (0.096)	0.947 (0.061)	0.943 (0.061)	41

3.3.2 The Relationship between T , r and n

For convenience, each putative QTL location is measured in a map distance (Mpd). For example, saying the location of QTL is 20.4 means the real QTL locates between the 20th and the 21st markers, and the distances to them are in proportion 4:6 in Mpd, or equivalently 0.4 Mpd to the 20th marker and 0.6 Mpd to the 21st marker. The step or gap to scan for the QTL is set to 0.1 Mpd if not specified otherwise, and the phrase ‘accurate’ is claimed once the estimated QTL is no more than 0.1 Mpd away from its real location.

In Table III-14, two populations with different parameters are simulated with 100 replicates. For each replicate of each population, we estimate the QTL and parameters with either an estimated T , which is estimated through the implementation of Hill’s method as introduced in III-3.2, or a real T , i.e. the one used to simulate the population. The averages of both T through 100 replicates are shown in the last column of Table III-14, and as the real T is constant throughout all replicates, we hence indicate its average as ‘fixed’ in the table. It can be noticed from Table III-14 that there is no remarkable difference between estimations of parameters from the using of estimated or real T . We may hence claim that the estimated T could efficiently represent the number of generations the population has experienced through, and hence we will simply adopt the estimated T in the following simulation analyses.

According to our simulation models as introduced above, all the genetic markers are in complete linkage at generation 0 and the initial LD coefficient between any a pair of genetic loci is of its maximum 0.25, where we have assumed the allele frequency at each genetic locus is 0.5. In order to acquire a validate estimate of QTL location, certain recombination generations are required to break down the complete linkage between marker loci. Generally, the closer two adjacent markers are with each other, a larger T is required to break down their linkage. However, when T is too large, the information of markers will be quickly

shrunk, and hence the trade off between T and r has to be balanced. Behind such an effect, there are two main factors: the population size and the marker distance. The two main factors affect the results jointly, and hence it would not be very easy to distinguish their effects separately. As generally believed, the population size will affect the possibility of inbreeding, and in a population with a small population size, the random drift of allele frequency might have remarkable impacts at each generation (Kimura, 1983) and results in a significant spurious LD after certain generations. What is more, a smaller population surely has a lower statistical power to detect a true LD comparing to a larger one, and the corresponding results might lose accuracy under such a circumstance. For instance, in Table III-15 and Table III-16, the accuracy increases dramatically as the increase of population size among populations with the same T , e.g. populations 3, 5 and 7. On the other hand, the marker distance also affects the reduction of LD over generations. With low marker coverage, LD between a marker and QTL will diminish quickly. Thus, a spurious LD randomly raised due to the genetic drift will be more easily detected in a population with both a small population size and low marker coverage than otherwise, and the estimate of QTL location in such a population is hence more sensitive to the increase of T . For instance, we could notice that populations 2 and 4 in Table III-18, i.e. $n=300$, $Mpd=50cM$ and $T=5$, have shown worse results than populations 1 and 3, i.e. $n=300$, $Mpd=50cM$ and $T=3$, where the standard deviations of the estimated QTL locations in the former populations are much larger than those in the later ones. Similar reductions of accuracy in the presence of too high a T could also be observed in populations with $n = 300$ in Table III-15, Table III-16 and Table III-17. Note here, when the true QTL locates at 28.2, although we could observe a significant increase of standard deviations for the estimated QTLs in populations with $n = 300$, the claimed accuracy might still increase, e.g. population 2 in Table III-18 and population 4 in Table III-17, which is mainly because one of the flanking markers in such a circumstance could still provide adequate information compared to those populations with true QTLs located at 28.4, e.g. population 4 in Table

III-18 and population 8 in Table III-17. Comparing populations with varying number of individuals scored in Table III-15, Table III-16 and Table III-18, the increase in standard deviations of estimated QTL caused by the increase in T could be efficiently compensated by an increased population size.

With above results and discussions, we could approximately outline an adequate T , which could efficiently break down the linkage between adjacent markers but not increase the chance of detecting spurious LDs. For instance, for a population with size $n=300$, the adequate T given a known marker distance could be listed as following

$T=3$ for Mpd=50cM

$T=15-20$ for Mpd=10cM,

$T=35-40$ for Mpd=5cM

As indicated in previous discussions, the adequate T listed above is ideal for a small population size. In a larger population, the adequate T will be larger as we could easily observe from Table III-15 and Table III-16, for example, with $n=3000$ and Mpd=5cM, $T=45$ is approximately the best choice comparing to $T=35-40$ for $n=300$ and Mpd=5cM as we have indicated above. From above results, given a constant sample size, we might approximately assume that a given Mpd and its correspondingly adequate T are of negative-proportionality, e.g. $T \times \text{Mpd} = 175$ approximately when $n=300$. Following above derivations, we might expect that in case a smaller Mpd, says 1cM, is presented, an adequate T might be over 150. Also, a sufficiently large population size is required to control the rate of inbreeding. As both above requirements are hardly to be achieved in experimental conditions, this method is hence highly limited by the resolution it could achieve in such a circumstance, although those requirements might be satisfied in nature samples.

3.3.3 Detecting Linked QTLs

The purpose of this session of analysis is to demonstrate how the present method could distinguish two closely linked QTLs. In the simulated chromosome, we adopt $Mpd=1cM$ between any two adjacent genetic markers. As these two QTLs are located at 17.4 and 22.4 in the chromosome, they are hence 5cM apart from each other. Except for the locations of QTLs, the remaining simulation processes are exactly the same as introduced in section III-3.3.1. For the sake of clear comparison to be made from the simulation study, we set the highest recombination generations T to be 50.

Although 50 generations of recombination might not be large enough to yield the most prominently accurate estimates of QTL locations as has been indicated in section III-3.3.2, it is sufficient to separate these two closely linked QTLs as shown in Table III-19 and Figure III-3 (e),(f). Figure III-3 illustrates the lod score values of composite likelihood ratio test statistics for each simulated population in Table III-19, where (a) (c) (e) are illustrated for population 1, 2, 3, respectively and (b) (d) (f) are illustrated for population 4, 5, 6, respectively. We could directly observe from Figure III-3 that unless the recombination generation is sufficiently high, i.e. $T = 50$ as shown in (e) and (f), the presented method failed to separate these two QTLs effectively at $T=10$ and $T=30$ as shown in (a) – (d). Alternatively, the estimates of a for both QTLs, denoted as a_1 and a_2 in Table III-19, could provide a statistical criterion of whether these two QTLs have been detected separately through the presented method. It could be easily noticed that in Table III-19, the estimates of a_1 and a_2 are significantly larger than their true values given $T = 10$ or 30 , i.e. population 1, 2, 4, 5. Especially, in population 1 and 4, either estimates of a_1 or a_2 solely could represent the whole phenotypic contributions from both QTLs, which implies that these two QTLs are not sufficiently separated.

Above results, again, show that a sufficient recombination generation is crucial in the ability of the presented method to properly estimate a QTL. However, these results also show that even the recombination generation is far from ideal as proposed in section III-3.3.2, where the accuracy of estimate might be reduced when compared to an ideal T , the presented method could still integrate sufficient information to distinguish two closely linked QTLs.

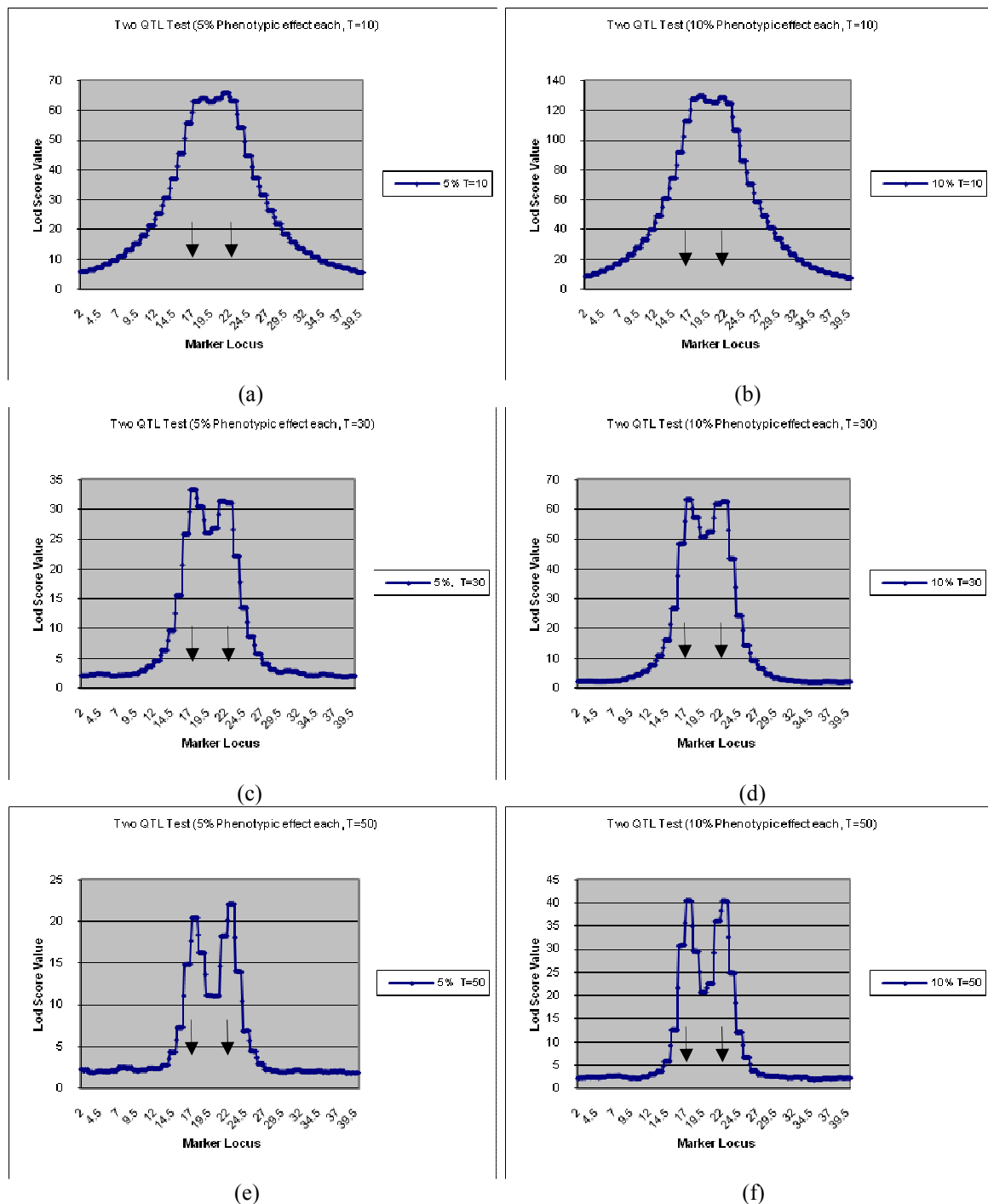


Figure III-3. Figure III-3 illustrates the means of lod score values of composite likelihood ratio statistics for each simulated population in Table III-19, where (a) (c) (e) for population 1, 2, 3, respectively and (b) (d) (f) for population 4, 5, 6, respectively. The locations of true QTLs are pointed by downward arrows in each figure.

3.4 Conclusion and Discussion

In Chapter III-3, we present a likelihood-based method for association studies, which could properly estimate the location of a putative QTL within a marker interval. The validation of our presented method relies on the marker density and recombination generations, i.e. T , the population has experienced ever since the introduction of mutate alleles. As both information is either available or could be directly estimated from other resources, the presented method could hence be generally applied for any random samples with a quantitative trait.

The validation and performance of our presented method have been intensively evaluated through simulation studies in III-3.3, where the method has shown a remarkable accuracy in estimating QTL locations within marker intervals, and the mapping resolution is hence improved significantly compared to traditional methods, e.g. a linear regression, which are based on detecting the association between a genetic marker and a putative QTL. However, we have also shown that the performance of this presented method heavily relies on the recombination generation T . That is, with a known genetic distance between any a pair of genetic markers, a sufficiently large T is required to allow the presence of enough recombination events to break down the linkage between these two markers. For instance, we have shown in section III-3.3.2 that $T > 100$ is required for $Mpd=1cM$ in order to maximize the performance of our presented method. Although a less sufficient T will reduce the accuracy of estimates of QTL parameters, it does not influence the validation of the presented method. The present method is fairly robust to these population parameters and should have taken advantages of improved mapping resolution if the population of interest has evolved sufficient generations after the introduction of mutated alleles at QTLs. Nowadays, with the development of sequencing technology and the accomplishment of HapMap Phase III (Altshuler et al., 2010), human data with over half a million SNPs are quite common. As

SNPs in these human data are so close to each other, there might not be sufficient recombination for us to infer a putative QTL between two SNPs. However, for less sophisticated species, where the available markers are of less density than in human genome, the present method could definitely provide extra information than traditional methods.

Reference

- ALTSHULER, D. M., GIBBS, R. A., PELTONEN, L., DERMITZAKIS, E., SCHAFFNER, S. F., et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 52-58.
- AVERY, P. J. & HILL, W. G. (1979) Distribution of linkage disequilibrium with selection and finite population-size. *Genetical Research*, 33, 29-48.
- CARDON, L. R. & PALMER, L. J. (2003) Population stratification and spurious allelic association. *Lancet*, 361, 598-604.
- DEVLIN, B., RISCH, N. & ROEDER, K. (1996) Disequilibrium mapping: Composite likelihood for pairwise disequilibrium. *Genomics*, 36, 1-16.
- FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449, 851-U3.
- HALDANE, J. B. S. (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8, 299-309.
- HILL, W. G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 33, 229-239.
- HILL, W. G. & WEIR, B. S. (1994) Maximum-likelihood-estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics*, 54, 705-714.
- KIMURA, M. (1983) *The neutral theory of molecular evolution*, Cambridge, Cambridge University Press.
- KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., et al. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*, 31, 241-247.
- LANDER, E. S. & BOTSTEIN, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185-199.
- LONG, J. C., WILLIAMS, R. C. & URBANEK, M. (1995) An e-m algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56, 799-810.
- LUO, Z. W. (1998) Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity*, 80, 198-208.
- LUO, Z. W. & SUHAI, S. (1999) Estimating linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Genetics*, 151, 359-371.
- LUO, Z. W., TAO, S. H. & ZENG, Z. B. (2000) Inferring linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Genetics*, 156, 457-467.

- LUO, Z. W. & WU, C. I. (2001) Modeling linkage disequilibrium between a polymorphic marker locus and a locus affecting complex dichotomous traits in natural populations. *Genetics*, 158, 1785-1800.
- MENG, X. L. & RUBIN, D. B. (1993) Maximum-likelihood-estimation via the ECM algorithm - a general framework. *Biometrika*, 80, 267-278.
- MIZUTA, I., TSUNODA, T., SATAKE, W., NAKABAYASHI, Y., WATANABE, M., et al. (2008) Calbindin 1, fibroblast growth factor 20, and alpha-synuclein in sporadic Parkinson's disease. *Human Genetics*, 124, 89-94.
- SIMON-SANCHEZ, J., SCHULTE, C., BRAS, J. M., SHARMA, M., GIBBS, J. R., et al. (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics*, 41, 1308.
- STEPHENS, M., SMITH, N. J. & DONNELLY, P. (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68, 978-989.
- STOREY, J. D. & TIBSHIRANI, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 9440-9445.
- TERWILLIGER, J. D. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, 56, 777-787.
- WANG, M. H., JIA, T. Y., JIANG, N., WANG, L., HU, X. H., et al. (2010) Inferring linkage disequilibrium from non-random samples. *BMC Genomics*, 11, 12.
- WEATHERALL, D. J. & CLEGG, J. B. (2001) Inherited haemoglobin disorders: An increasing global health problem. *Bulletin of the World Health Organization*, 79, 704-712.
- WEIR, B. S. & COCKERHAM, C. C. (1979) Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 42, 105-111.
- XU, X. M., ZHOU, Y. Q., LUO, G. X., LIAO, C., ZHOU, M., et al. (2004) The prevalence and spectrum of alpha and beta thalassaemia in Guangdong province: Implications for the future health burden and population screening. *Journal of Clinical Pathology*, 57, 517-522.
- YU, J. M. & BUCKLER, E. S. (2006) Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology*, 17, 155-160.
- ZENG, Z. B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, 90, 10972-10976.
- ZENG, Z. B. (1994) Precision mapping of quantitative trait loci. *Genetics*, 136, 1457-1468.
- ZHANG, W., CAI, W. W., ZHOU, W. P., LI, H. P., LI, L., et al. (2008) Evidence of gene conversion in the evolutionary process of the codon 41/42 (-CTTT) mutation causing beta-thalassemia in southern china. *Journal of Molecular Evolution*, 66, 436-445.

FINAL CONCLUSION

Throughout the whole thesis, I mainly focused on the strategies and statistical methods for association analysis of complex traits. Besides the general introduction of QTL mappings, I have divided this thesis into two chapters, i.e. Chapter II: The Comparison of Linear Models in Association Study and the Optimal Trend Coefficients in Armitage's Trend Test, and Chapter III: Likelihood-based Methods for Association Studies.

The thesis started to introduce the commonly used genetic and statistical models in association analysis in Chapter II-1, and was then followed by thorough comparison and evaluation of the performance of these commonly used methods in Chapter II-2. From these analyses, I concluded that SLR was generally better than a fixed effect model as the former method normally showed an increased statistical power especially if a proper set of x_i , which represented the genetic effects of a genetic marker, was used. The two different genetic models, i.e. explicit and implicit, were asymptotically equivalent among various statistical models, e.g. SLR, ANOVA and ML, under the scheme of the Rao's score test. These results hence suggest use of SLR with the explicit model for association analysis due to their mathematical convenience. With the results above, I further explore the optimal choice of x_i under SLR for Case-Control studies in Chapter III-3, where in such a situation, SLR was demonstrated to be equivalent to the so-called Armitage's trend test with x_i being the corresponding trend coefficient. I have managed to show that the presence of heterogeneity will largely influence the choice of optimal trend set, and hence I suggest the use of trend set $\{1, 0.5, 0\}$ for both dominance and co-dominance models, and trend set $\{1, 0.25, 0\}$ for the recessive model. Such suggestions differ from those given by Sasieni (1997), and the results from simulation analyses favour our suggestions over Sasieni's. In Chapter III-3, I also compared two strategies to correct for population stratification when multiple samples are

combined to increase the statistical power, i.e. incorporating dummy variables as introduced in section II-3.3.1 and non-central χ^2 test as introduced in section II-3.3.2 above. Through both simulation study and real data analysis, we have shown that both methods enable one to properly adjust the bias introduced by population stratification. However, the former method, i.e. dummy variable, is favoured if the differences of case-control ratios among multiple samples are large, and the other, i.e. non-central χ^2 test, is favoured otherwise. During the establishment of non-central χ^2 test, I have shown that population stratification will have distinguishable impacts on one or both of the numerator and denominator of the Armitage's trend test given as formula (II-1.21). For the numerator, influence from population stratification could be either positive or negative, and it is the main cause of false positive. On the contrary, the influence on the denominator from population stratification could only be positive, and because the denominator was proportional to the variance of the numerator and hence was positive as well, such influence would rather reduce the statistical power than increase it. The influence on the denominator was initially proposed to increase the statistical power with certain flaws by Devlin & Roeder (1999). When the newly developed method was implemented to association study with Parkinson's disease data (Simon-Sanchez et al. 2009), it revealed several new candidate genes which showed significant association with the disease. All these genes have long been proposed as candidates by neuro-biologist but never detected in the current literature of the GWAS.

In Chapter III-1, I proposed a likelihood method using the conditional probability distribution to estimate the coefficient of LD between the marker and disease loci for non-random samples. Through the comprehensive simulation studies and real data analyses, it has been shown that the new method could not only properly estimate LD coefficients in random samples as the traditional method does, i.e. Hill's (1974), but also provide significant improvement of estimation accuracy if the data is artificially collected and thus non-random. Following the

same idea, a further conditional likelihood of association study for Case-Control samples is established in Chapter III-2. This method delivers the ability to easily adapt any penetrance models and to count for the non-random nature of samples from a case and control design. Its capacity to significantly improve statistical power to detect the genetic association was illustrated through simulation studies and real data analyses. In the real data analysis, I implemented the newly proposed method to analyse the same Parkinson's data as introduced in Chapter II-3 and compared the method with two other popularly cited methods in the literature of genetic association. Finally, in Chapter III-3, I present a likelihood-based method for association studies, which provided a genome-wide scan for QTL based on linkage disequilibrium analysis. The validation and performance of this method have been assessed through simulation studies, and a significant increase in mapping resolution was shown by the LD based QTL analysis.

Along with acquiring more comprehensive genetic knowledge through above researches, I have gained plenty of sophisticated trainings and practices in both statistics and computational programming. Not only the using of common statistical tools, e.g. SPSS, STATA, I am also professional in programming statistical methods and simulating statistical ideas in FORTRAN and R languages. These experiences hence strengthen my ability to develop new statistical methods and handle various kinds of data in real data analyses.