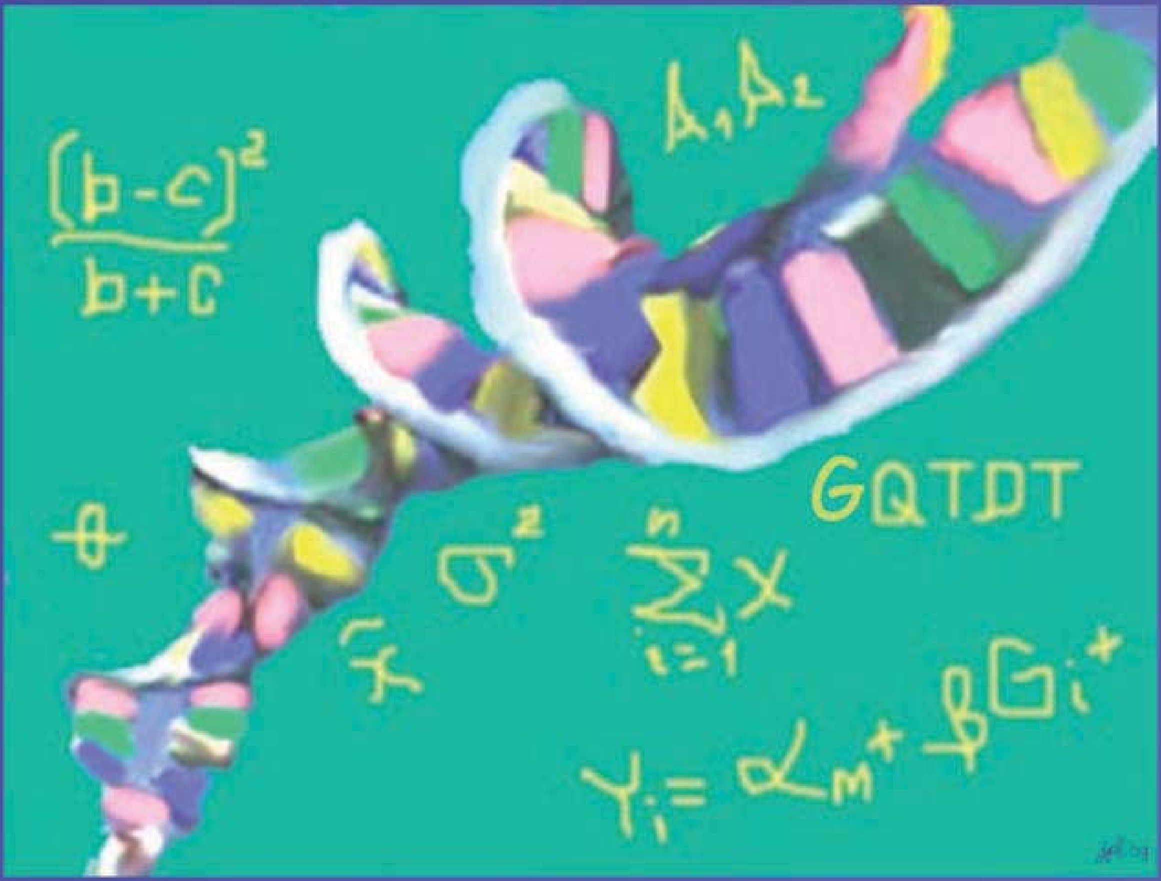
Generalized Quantitative Transmission Disequilibrium Test for Analyzing Genetic Main Effects and Epistasis



Jingky P. Lozano

Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Generalized Quantitative Transmission Disequilibrium Test for Analyzing Genetic Main Effects and Epistasis

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

vorgelegt von
Jingky Pamesa Lozano
aus
Manila, Philippinen

Göttingen, 2010

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Aufl. - Göttingen: Cuvillier, 2010
Zugl.: Göttingen, Univ., Diss., 2010

978-3-86955-582-9

D7

Referent:	Prof. Dr. Manfred Denker
Koreferentin:	Prof. Dr. Heike Bickeböller
Tag der mündlichen Prüfung:	22.11.2010

© CUVILLIER VERLAG, Göttingen 2010
Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2010
Gedruckt auf säurefreiem Papier

978-3-86955-582-9

To

my dearest God

Take no pride in knowledge or titles that label you wise;
True wisdom comes only with experience and humility.
- J. U. Noble

Abstract

Genetic studies have utilized the *Transmission Disequilibrium Test* (TDT) to investigate the effect of genes and their interactions (*epistasis*) on complex diseases such as diabetes mellitus, Alzheimer's disease, ischaemic heart disease and cancer. The TDT has been frequently used as a statistical method to analyze genetic effects in family studies due to its robustness to population stratification. The original TDT method by Spielman et al. (Am J Hum Genet, 1993) was designed for qualitative traits (e.g. disease classification: affected / unaffected) but variations have been developed for its application in quantitative traits (QTs) such as blood sugar level, radiation sensitivity and measures of coronary artery calcification. However, the occurrence of nonnormally distributed quantitative traits in candidate gene analysis poses difficulties for statistical methods that are sensitive to distributional assumptions.

This study introduces the *Generalized Quantitative Transmission Disequilibrium Test* (GQTDT) - a statistical method for quantitative traits based on generalized additive models incorporating parental mating type (parental genotype combination) indicator and different parameters of the distribution of the QT response variable in the statistical model. It aims to determine genetic effects (i.e. main effects and epistasis) affecting QTs in family-based studies. The method is based on the *Generalized Additive Model for Location, Scale and Shape* (GAMLSS, Rigby and Stasinopoulus, Appl Stat, 2005) which allows not only the mean but also other parameters of the conditional distribution of the quantitative trait to be included in the model. The power and type I error of the GQTDT to detect genetic main effects and epistasis were investigated in simulation studies. It has also been applied to real data to determine its applicability in different settings and compare its findings with existing biological data. Genotype data from family trios (parents and one offspring) as well as phenotype data of the offspring were used in the analysis. In the simulation studies, two unlinked biallelic loci and QTs influenced by one or two loci and epistasis were created. The QTs were simulated either as normally distributed or skewed to the right which are commonly encountered in genetic data. Different scenarios such as presence of population stratification and other covariates were also simulated in the data to determine its possible effects in the GQTDT analysis of genetic main effects and epistasis.

The performance of the GQTDT in determining genetic main effects is satisfactory both in the normally distributed and skewed quantitative traits. When a fitted distribution is specified in the analysis, higher power can be achieved. In terms of detecting epistasis, good power is noted when the distribution of the quantitative trait is normal. When detecting epistasis in skewed traits, the power is not as high as the power in the normally distributed traits but higher compared to the benchmark method, the *Quantitative Transmission Disequilibrium Test with parental mating type indicator* ($\mathrm{QTDT_M}$; Gauderman, Genet Epi, 2003). The power of the GQTDT is also higher with higher minor allele frequencies, correctly assumed analysis genetic model, larger "true" effect size and bigger sample size. Slightly elevated type I error may be observed in analyzing skewed quantitative traits but like other TDT-like tests, the GQTDT is also robust to the effects of population stratification which causes spurious association. Its application to real data detected genetic main effects and epistasis with known biological evidence.

Keywords: quantitative trait, family-based studies, transmission disequilibrium test, generalized quantitative transmission disequilibrium test, candidate genes, epistasis, genetic effects

Acknowledgements

I wish to express my gratitude to the many souls (some of them unnamed here) who helped me in my academic journey in Göttingen. First of all, I am thankful to my PhD supervisor Prof. Dr. Heike Bickeböller for the opportunity to do a PhD in genetic epidemiology. I appreciate her support in my academic and personal challenges in Germany. I would also like to thank Prof. Dr. Manfred Denker for staying in my PhD committee as first examiner and for his fatherly-encouragement ever since I started in the program. I couldn't thank him enough for his extra efforts to come to Göttingen to personally assist and support me in my dissertation. I would also like to specially acknowledge Prof. Dr. Walter Zucchini whose humble and approachable nature inspired me to attend economic courses and seek his help for some statistical questions. And of course, I would also like to thank Prof. Dr. Martin Schlather, Prof. Dr. Stephan Waack and Prof. Dr. Max Wardetzky for agreeing to be part of my PhD committee.

For their kindness, help and welcoming spirit, I would like to thank my colleagues in the *Department of Genetic Epidemiology* and the *Department of Medical Statistics*, most especially Andrew Entwistle who always helped me with my *Rathaus* documents, translations and computer problems. I will miss knocking on his office whenever I have questions. I am also deeply indebted to Bianca Wegener for her friendly concern and help in all aspects of my challenging beginnings in Göttingen. She has been a great help in administrative things, looking for an apartment and even searching for "special" bicycles. I am also very much thankful to Heike Born who assisted me everytime she could. I would not be able to transport my personal stuff without her help. And of course, not to forget, Albert Rosenberger, Karola Köhler, Christina Reck, Yesilda Balavarca, Melanie Sohns, Arne Schillert, Dörthe Malzahn, Monika Colmsee-Wambi, Christina Galambosi, Christoph Braun, Patricia Toegel, Debby Kronenberg, Anja Sapara, Verena Gullatz, Karin Neubert, Carola Werner, Rauf Ahmad, Karthinathan Thangavelu, Abu Hena Mahbub-ul Latif and Karola Riemenschneider whose friendly presence and readiness to help made my life in the university more enjoyable. Ms. Carmen Barann and Dr. Hartje Kriete from the Math Faculty and colleagues from the *Centre for Statistics*, especially Rico Ihle and Razmig Dichjekenian also deserve

special mention. Among others, I would also like to thank Prof. Dr. med. Jürgen Brockmöller, Sven Müller, Prof. Dr. med. Ralf Dressel and all research collaborators and colleagues from the *Graduiertenkolleg 1034*. I am also grateful to the Genetic Analysis Workshop 16, Framingham Heart Study and the Lung Cancer in the Young Study for allowing me to use their data in my analysis.

My heartfelt thanks also go to all my friends and foster families in Germany, especially to the Lacambacal family, Krull family, Lautner family, Valdez family, Come family, Baldos family, Kühne family, Jannette and Stephan Schulze and family, Imelda von Bismarck and family, Shirley and Maxi Bonn, Sohns family in Schwirzheim, Brandt family in Immenhausen, Alexis Lapiz, Julius Militante, Annabelle and Joerg Stein, Lomen and Christian Decavele, Salve Huehne, Minda Kruegel, Marites Deichsel, Tere Redaniel, Cindy Rosenmeier, Lyn Fochler, Cecilia Leon-Ramirez, Father Stephan Lipke, Gisela Wollborn, *St. Paulus Singkreis*, *Katholische Hochschulgemeinde Göttingen (KHG)*, *Blaue Sänger*, *Couples for Christ FFL* community, *Filipino Scholars and Alumni Association in Germany (FSAAG)* and all student colleagues. Their friendship and support during my study years have been a big comfort. I would also like to express my thankfulness to Lila, Neo, Susy, Brandy, Frau Rosel Lips of Studentenwohnhein in Karolinenweg, Strüver family and Nieder-Eichholz family in Glienicke, Lozano and Pamesa families, Tes Gatchalian, Cyndi Mapua, Jopee Laygo, Rina Villavicencio, Vicky Medina, The Magnolias, LHNC family, DEBs Family, UP-CPH, PCHRD, SLMC and RITM friends, Mama Cena and family, Sir Aying and Ate Lilen Sarol, Daddy Pete and Mommy Zoila Genato, Ma'am PI and Daddy Manny Natividad, Tess and Jerry Trajeco, Joycee and John Liu, the Filipino communities in Germany, SSEAYP diplomats and to all relatives and friends who visited me or in their special way supported me in my PhD endeavor.

Finally, the completion of my PhD would not be possible without the funding support from the *Graduiertenkolleg 1034* of the Deutsche Forschungsgemeinschaft and of course the prayers and support of my family - Nanay, Papa, Mama, Dante, Richard, Jun, Arlyn, Arjay and Gabriel who patiently waited while PhD work absorbed my time and energy. Unfortunately, Papa did not live long enough to see me ”kiss the goose girl” in Göttingen.

To God and all souls who have inspired and helped me in this journey ... *ang aking taos-pusong pasasalamat ... herzlichen Dank*.

J.P.Lozano
Göttingen, Germany
October 2010

Abbreviations and Acronyms

%DT	- Percent-DNA-in-Tail
BCPE	- Box-Cox Power Exponential distribution
BCT	- Box-Cox-t distribution
BMI	- Body-mass-index
CAC	- Coronary artery calcification
CPG	- Conditional on Parental Genotypes
CVD	- Cardiovascular diseases
DNA	- Deoxyribonucleic acid
DOTM	- Difference in the Olive Tail Moments
FAM71F1	- Family with sequence similarity 71, member F1; also known as FAM137A
FBAT	- Family-Based Association Tests
FHS	- Framingham Heart Study
FHSsim	- Framingham Heart Study Simulated Data
FTO	- Fat mass and obesity associated gene
GAIC	- Generalized Akaike Information Criterion
GAM	- Generalized Additive Model
GAMLSS	- Generalized Additive Model for Location, Scale and Shape
GAW	- Genetic Analysis Workshop
GD	- Global Deviance
GLM	- Generalized Linear Model
GLRT	- Generalized Likelihood Ratio Test
GQTDT	- Generalized Quantitative Transmision Disequilibrium Test
HDL	- High Density Lipoprotein
HOGG1	- Human 8-oxoguanine DNA N-glycosylase
HQTDT	- Hierarchical Quantitative Transmission Disequilibrium Test
IBD	- Identical-by-Descent
IDDM	- Insulin Dependent Diabetes Mellitus

IHGSC	- International Human Genome Sequencing Consortium
INT	- Inverse Normal Transformation
LD	- Linkage Disequilibrium
LigIV	- Ligase IV
LNPT	- Longitudinal Nonparametric Association Test
LOD score	- Log-odds score
LRT	- Likelihood Ratio Test
LUCY	- Lung Cancer in the Young
MAF	- Minor Allele Frequency
MRE11	- Meiotic Recombination 11
NHEJ	- Non Homologous End Joining
NHLBI	- National Heart, Lung, and Blood Institute
OMIM	- Online Mendelian Inheritance in Man
OR	- Odds Ratio
OTM	- Olive Tail Moment
PDT	- Pedigree Disequilibrium Test
PFKP	- Phosphofructokinase, Platelet type
PON1	- Paraoxonase 1
QCPG	- Quantitative Conditioning on Parental Genotypes
QPL	- Quantitative Polytomous Logistic
QT	- Quantitative Trait
QTDT	- Quantitative Transmission Disequilibrium Test
QTDT$_{\mathrm{M}}$	- Quantitative Transmission Disequilibrium Test with mating type indicator
RC-TDT	- Reconstruction Combined Transmission Disequilibrium Test
RFLP	- Restriction Fragment Length Polymorphism
RQTDT	- Retrospective Quantitative Transmission Disequilibrium Test
SNP	- Single Nucleotide Polymorphism
S-TDT	- Sib Transmission Disequilibrium Test
TDT	- Transmission Disequilibrium Test
TDT-AE	- Transmission Disequilibrium Test (allow for error)
XRCC1	- X-ray Repair, Complementing defective, in Chinese Hamster, 1
XRCC4	- X-ray Repair, Complementing defective, in Chinese Hamster, 4

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and rationale

Complex diseases such as diabetes mellitus, Alzheimer's disease, ischaemic heart disease and cancer are diseases that are influenced by more than one factor. Usually, the interactions between genetic and environmental factors play a role in complex disease outcome and treatment. It is also an accepted fact that individual gene effects and *epistasis* (interaction between genes) may play an important biological role in complex diseases. Genetic studies on complex diseases have utilized different statistical tests to determine genetic factors that may affect disease characteristics or traits. *Traits* can be any qualitative or quantitative (measurable) characteristic of an organism. In genetics, *trait* is often used synonymously with the term *phenotype*. One of the statistical tests in family-based studies that had become popular for the identification of *genes* affecting diseases or traits is the *Transmission Disequilibrium Test* (TDT).

The TDT is a statistical test introduced into genetic epidemiology by Spielman et al. (1993) to identify the effects of genetic factors on disease outcome using family data. The advantages of the TDT have been shown by several studies in the past (e.g. Laird and Lange, 2006). It had become a popular test due to its robustness to population stratification. In genetic association studies, population stratification or the presence of several subpopulations in the data may lead to spurious association results. Ewens and Spielman (1995) have explicitly shown that the TDT is robust against this effect of population stratification. Different forms of the TDT have been used for detecting genetic effects of *candidate genes* or genes that are thought of to affect a trait on the basis of their physiological and biological pathway functions. Originally, the TDT has been designed for the investigation of qualitative traits (e.g. disease status, that is whether a person is affected or unaffected by a disease). Analysis using the original TDT utilizes data from *family trios* consisting of the mother, father and one disease-affected child to determine the frequency of transmission of genetic characteristics from parents to offspring. Variations of the TDT or TDT-like tests have been proposed to enhance its efficiency and applicability in different settings. One example is the method called *conditioning on parental genotypes* (CPG) (Cordell et al., 2004). This method involves constructing a sample of *cases* (disease-affected individuals) and matched pseudocontrols from a sample of family trios. An appealing feature of this approach is that it allows family data to be analyzed like matched case-control design using conditional logistic regression. The TDT has also been extended to accommodate different structures of families. Some extensions apply to only one gene while others consider two or more genes and their interactions. With the importance of epistasis in mind, Wilson (2001) proposed a method to determine the effect of two interacting genes on a dichotomous disease outcome. However, since the method considered known disease genes, other genes with weak marginal effects but with a stronger epistatic effect will escape such investigation. To address this issue, Kotti, Bickeböller and Clerget-Darpoux (2007) investigated the TDT for a dichotomous disease outcome in the context of detecting disease genes with weak or no marginal effect. Other extensions of the TDT and TDT-like tests have been introduced to accommodate broader scenarios such as inclusion of maternal genetic factors (Weinberg et al., 1998), analysis using siblings (Spielman and Ewens, 1998), handling of errors in genotyping (Gordon et al., 2001) and parental-genotype reconstruction (Knapp, 1999). Another TDT-based method is the Family-Based Association Tests (FBAT). This unified approach to family-based tests of association was introduced by Rabinowitz and Laird (2000) and Laird et al. (2000). The FBAT test statistic is based on the distribution of the offspring genetic characteristics conditional on any trait information and on the parental genetic characteristics. It follows the original TDT approach by conditioning on the trait and the parental genetic characteristics. If the parental data are not available, the test statistic is conditioned on the sufficient statistics for the offspring distribution (Laird, 2007). This approach makes the FBAT applicable even if parental data are missing.

The earlier variations and extensions of the TDT focus on qualitative trait or categorical variables as an outcome. However, the TDT has also been modified to analyze quantitative traits such as blood pressure, blood glucose levels and radiation sensitivity. Quantitative traits (QTs) have continuous distribution and have quantitative or numeric values. They might be more direct and hence more informative measures than qualitative traits. This idea gave rise to the application of the TDT to quantitative traits. Earlier *Quantitative Transmission Disequilibrium Tests* (QTDTs) include the works of Allison (1997), Rabinowitz (1997), Fulker et al. (1999), Lunetta et al. (2000), and Abecasis et al. (2000) which are described in chapter 3. Gauderman (2003) looked into previous QTDT methods and proposed one which he called $\mathrm{QTDT_M}$ *(Quantitative Transmission Disequilibrium Test with mating type indicator)*. It was specifically designed for continuous quantitative traits and family trio (father, mother, child) data. This statistical method is based on linear regression incorporating *parental mating types* as fixed effects. The parental mating type is the combination of the genetic characteristics of the mother and the father. The $\mathrm{QTDT_M}$ incorporates this parental mating type information in the regression equation to test for genetic main effects and epistasis. The method can also be extended to include one or more environmental factors and gene-environment interaction. It has been shown to exhibit good power in detecting genetic effects compared to previous methods dealing with quantitative traits in family studies. Another recent approach called quantitative conditioning on parental genotypes (QCPG) by Wheeler and Cordell (2007) has also been compared to the $\mathrm{QTDT_M}$. Comparison of the $\mathrm{QTDT_M}$, QCPG and simple linear regression using simulated data showed that the $\mathrm{QTDT_M}$ was the only method suitable for estimation of effects under the alternative hypothesis with population stratification (Wheeler and Cordell, 2007). In the case of nonnormal data, the nonparametric FBAT approach will have an advantage over parametric tests like the $\mathrm{QTDT_M}$, but the issue of testing for epistasis or gene-gene interaction effect still remains a challenge with the FBAT approach. In general, analyzing epistasis is still not properly addressed in statistical genetics. However, we cannot just disregard the effect of epistasis or gene-gene interaction especially in complex diseases. Moore (2003) provided explanations supporting that interactions can be more important than the independent main effects of common disease genes. This may not be true in all diseases, but it may be observed in situations where the individual effects of several candidate genes are weak but their interaction contributes a lot to the manifestation of the disease. Knowing if epistasis is a significant factor in any disease may provide a clue in understanding the biological mechanism of the disease. It can also give us better predictions on who might develop the disease for future prevention strategies. However, determining epistasis will require both computational and biological approaches. Using a biological approach alone might prove to be very difficult considering that there is a gigantic number of gene-gene interactions possible in humans. A good statistical method hand-in-hand with biological methods is a better tandem to detect epistasis in genetic studies. Unfortunately, currently available statistical tests for family-based studies, especially those applicable for detecting epistasis using quantitative traits are not well developed (Li et al., 2007). This does not imply that there are only limited efforts done in investigating epistasis. In fact, there are many investigators who explored the topic but up to now there are still issues left unsolved especially in dealing with quantitative traits and family data. Chapter 3 gives details and issues of the TDT and TDT-like methods currently used in family-based studies. While much effort has been given to the issues of population stratification and finding efficient statistical methods to determine genetic main effects and epistasis, the problem of nonnormal distribution in the analysis of quantitative traits does not get much attention. The currently existing methods (e.g. Abecasis et al., 2000; Gauderman, 2003) for quantitative trait analysis in family-based studies which consider both genetic main effects and epistasis are often based on linear regression. Gross deviations from the normality assumption create problems for this type of analysis. Other methods may handle nonnormal data but did not consider epistasis in the analysis. Although many statistical tests have been designed for nonnormally distributed data in general, the application of these tests in genetic family-based studies is still limited.

1.2 Objectives

In lieu of the existing challenges in the analysis of candidate genes in family-based studies, this dissertation aims to provide an improved statistical method for analyzing genetic main effects and epistasis that can be applied to family data and quantitative traits. Specifically, the following are the main objectives of this work:

- To introduce the Generalized Quantitative Transmission Disequilibrium Test (GQTDT) for determining genetic effects (i.e. main effects and epistasis) of candidate genes for diseases. The new method is applicable to normally distributed and nonnormally distributed quantitative traits commonly encountered in genetics. It has been used here in few selected distributions but it can also be applied to other types of distributions.
- To investigate the power and type I error of the GQTDT in the presence of population stratification and unknown environmental covariates; and

- To apply the GQTDT to the Genetic Analysis Workshop (GAW) 16 data and to a sub-project of the Lung Cancer in the Young (LUCY) study. The GAW data have both real data and simulated data based on a heart disease study. The LUCY data contain real information on lung cancer patients diagnosed at age 50 or younger.

1.3 Organization of succeeding chapters

The next two chapters of this dissertation present a review of the literature about genetic concepts and statistical methods in genetics. Basic information about the human genome and modes of inheritance are presented in chapter 2. The Mendelian inheritance, Hardy-Weinberg equilibrium, genetic models, segregation, linkage association studies and epistasis are also discussed in the same chapter. The third chapter is about statistical methods used in genetic studies. It focuses mainly on tests for family-based studies, specifically the TDT and TDT-like tests.

Chapter 4 describes the Generalized Quantitative Transmission Disequilibrium Test, its theoretical concept, development and characteristics. Chapter 5 presents the results of simulation studies while chapter 6 contains the results of the analysis of the GAW and LUCY data.

Finally, chapter 7 gives a summary and outlook for future research directions.

2 Genetic Concepts

Basic knowledge of genetic concepts is necessary to analyze genetic factors and determine the involvement of epistasis in complex diseases. This chapter reviews the basic principles of genetics and other concepts needed for the statistical analysis of genetic factors. A separate section about epistasis or gene-gene interaction is also included. Unless otherwise specified, the details in this chapter are based on the textbooks written by Vogel and Motulsky (1986), Khoury et al. (1993) and Sham (1998).

2.1 The genetic code

The *genome* is the total hereditary information of an organism. In humans and other cellular life forms, this information is encoded in the DNA (*deoxyribonucleic acid*). The DNA can be imagined as a very long, tightly coiled string of information in every cell of the organism. Typically, the term genome refers to the complete set of DNA found in the nucleus of the cell (i.e., the "nuclear genome"). However, it can also be applied to genetic information within cell organelles that contain their own DNA, as with the mitochondrial genome or the chloroplast genome.

In humans, the genome consists of 23 pairs of DNA molecules called *chromosomes*, 22 of which are autosomal chromosomes and 1 pair are sex-specific chromosomes located in the cell nucleus. Males have one X and one Y chromosome (as shown in figure 2.1), while females have a pair of X chromosomes.

Two chromosomes forming a pair are called *homologous chromosomes*. The somatic cells of humans contain a diploid set of chromosomes (i.e. 23 pairs or 46 chromosomes) with the exception of the sex chromosomes in males. During normal cell division called *mitosis*, these chromosomes are duplicated to form new cells. On the other hand, the mature sex cells (ovum and sperm) contain only half of the set i.e. 23 chromosomes. These sex cells are also known as *gamete* or *germline cell*. All humans

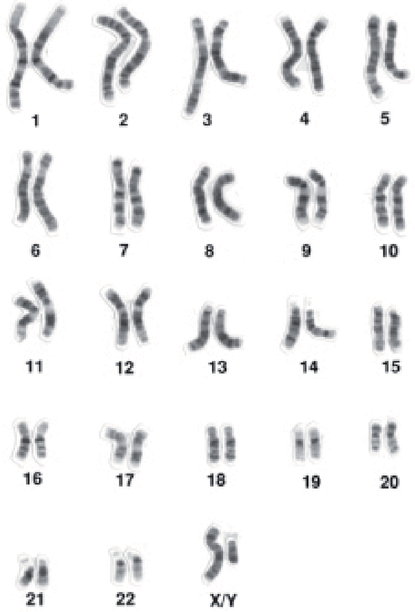


Figure 2.1: Karyogram of a human male
Source: National Library of Medicine (US), NCBI; http://www.ncbi.nlm.nih.gov/

are formed from the union of gametes (ovum and sperm) from the parents in a type of cell division called *meiosis*. During sexual reproduction, each parent contributes one gamete to form a single cell called *zygote*. Since each gamete has a haploid (single) set of 23 chromosomes, the zygote receives a diploid (double) set of 46 chromosomes. During meiosis, exchange of genetic materials can occur between the two homologous chromosomes from the parents. This results in alternating segments or blocks of inherited DNA from the mother and father.

In a chromosome, there is a special location somewhere in the middle called *centromere* which plays a role in the cell division process. Each chromosome is composed of two arms extending on either side of the centromere. The shorter arm is known as the *p-arm*, while the longer arm is known as the *q-arm*. The endpoints of the chromosome are called *telomeres*. Chromosomes can be observed during cell division under the microscope as elongated molecules that show coloured bands after staining. The chromosome was discovered in 1953 by Watson and Crick as strands of DNA consisting of two chains of nucleotides arranged in a double-helix structure. The four types of nucleotides are Adenine, Cytosine, Guanine and Thymine, which are conventionally represented by the letters A, C, G, and T, respectively. Each nucleotide only pairs with a specific nucleotide. Adenine pairs always with Thymine while Cytosine pairs always with Guanine. The nucleotide pairs are also called *base pairs*. The whole genome can be envisioned as a string of "letter codes" of these different nucleotides. According to the IHGSC (International Human Genome Sequencing Consortium, 2004) who completed sequencing the human genome in 2003, the genome has about 2.85 x 10^9 base pairs. Figure 2.2 shows a schematic representation of a double-helix DNA with the nucleotide pairs.

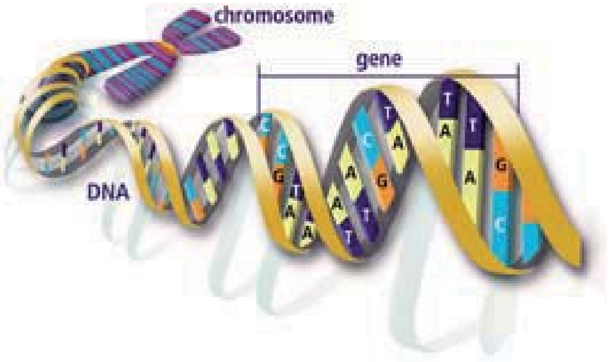


Figure 2.2: Schematic representation of the DNA
Source: U.S. Department of Energy Human Genome Programs,
http://genomics.energy.gov

Genes are parts of chromosomes that perform biological function by encoding for proteins. They are considered as the basic units of hereditary information. A chromosome may contain several thousand genes. The overall exact number of genes encoded by the human genome is still unknown but the IHGSC reported an estimate of 20,000 - 25,000 protein genes in the human genome. Genes influence human traits such as blood type, insulin level, ability to digest lactose, baldness, polydactyly (having extra fingers) and many more. Many diseases have also been identified as genetic diseases because they are caused by abnormalities in the genome. One of these is *Hemophilia* which has gained recognition because it has afflicted the descendants of Britain's Queen Victoria. Hemophilia is a bleeding disorder where blood does not clot properly due to a shortage of a clotting factor. There are three reported types of hemophilia — types A, B and C. These types of bleeding disorders are due to abnormalities in the genes encoding for the functional blood clotting factors VIII, IX and XI, respectively (Bolton-Maggs and Pasi, 2003; Zivelin et al., 2004). Other known examples of genetic disorders are color blindness, sickle cell anemia and cystic fibrosis. Color blindness is more common in males than in females because many of the genes involved in color vision are on the X chromosome. In sickle cell anemia, gene mutations are reported on chromosome 11 while mutations in the CFTR gene on chromosome 7q were reported to be associated with cystic fibrosis, an inherited disease of the mucus glands that results to progressive damage to the respiratory system and chronic digestive system problems. Complex diseases such as type 1 or insulin-dependent diabetes mellitus, autism and certain cancers (e.g. breast cancer, familial neuroblastoma) have been related also to certain genes. Type 1 diabetes mellitus is affected by many genes which are mostly located on chromosomes 6, 11 and 18 while autism has been linked to chromosome 7. BRCA1 and BRCA2 are the major genes related to hereditary breast cancer and susceptibility genes on chromosome 6p22 have been associated with the childhood cancer, neuroblastoma (OMIM, 2009).

The terms *gene* and *locus* are sometimes used interchangeably. A *locus* (plural: *loci*) is a specific position along a chromosome. It can denote the position of a gene or a genetic marker. A combination of several loci on a single chromosome strand (not necessarily adjacent) is called a *haplotype*. There are genes that usually occur at the same position in the genome so that the gene and its locus are sometimes used synonymously. Genes can exist in different forms or states. These different forms of a gene are called *alleles*. Every individual carries two copies of an autosomal gene which may be different or similar alleles. Individuals with two different alleles at a certain gene are said to be *heterozygous*, while individuals with two copies of the same allele are referred to as *homozygous*. Some alleles may be associated with certain diseases while others just contribute to the population's normal genetic variation. Genes can be *polymorphic*. "Poly" means many, and "morph" means form. A polymorphism exists when the most frequent allele occurs in less than 99% in the population.

Individuals may be characterized by their genetic make-up. The *genotype* of an individual refers to its genetic composition at a locus. It is the pair of alleles at a particular gene. For instance, if a trait is influenced by a biallelic gene (let's say that the two alleles of the gene are A_1 and A_2), then the three possible genotypes are A_1A_1, A_1A_2, and A_2A_2. The genes influence the observable characteristic or the *phenotype* of an individual and usually, the environment also plays an important role in the resulting phenotype. Examples of phenotypes, some of which have been mentioned already are stature, blood type, insulin level, polydactyly, intelligence quotient (IQ), affection by a disease such as diabetes mellitus, glucose-6-phosphate dehydrogenase

(G6PD) deficiency etc. To illustrate the distinction between genotype and phenotype, consider the gene that determines the ABO blood types which has three alleles: I^A, I^B, and i. The possible genotypes and the observed phenotypes during blood typing are shown in table 2.1.

Table 2.1: Genotypes and ABO Blood Types

Genotypes	Blood Type Phenotypes
I^AI^A or I^Ai	A
I^BI^B or I^Bi	B
I^AI^B	AB
ii	O

Note that the observed blood type A can arise from two genotypes, I^AI^A or I^Ai. Carrying at least one I^A allele without the presence of I^B allele gives rise to blood type A. The same is true for blood type B. Carrying at least one I^B allele without the I^A allele will give rise to blood type B. The presence of both I^A and I^B alleles results to blood type AB while the absence of both alleles (i.e. genotype ii) results to blood type O. Alleles I^A and I^B are said to be *codominant* to each other and both are dominant over allele i. More about genetic patterns of inheritance is discussed in the section Genetic Models.

In the study of genetic factors, another important information to note is the term *genetic marker*. Genetic markers are used to determine which inherited genes are associated or linked with certain diseases. DNA segments that lie near each other on a chromosome tend to be inherited together. A marker can therefore be used to track or map genes that have not yet been identified. A genetic marker can be a gene or a DNA segment that can be easily determined and whose location on the chromosome is known. It needs to be variable so that the alleles will be more likely different among unrelated individuals. *Restriction fragment length polymorphisms* (RFLPs) were one of the earliest molecular markers. An RFLP is a variation in the DNA sequence that is detected by cutting the DNA into segments with restriction enzymes and analyzing the resulting lengths of fragments by electrophoresis. RFLPs are also simply known as restriction polymorphisms. Another type of marker is called *microsatellite*. Microsatellites are used to detect variable numbers of DNA sequences that are repeated. A common example of a microsatellite is a CA repeat where the C and A nucleotide are repeated in the DNA sequence n number of times. At the moment, the most commonly used markers are the *single-nucleotide polymorphisms* (SNPs). These polymorphisms are characterized by a variation occurring in a single nucleotide - A, C, G, or T - in the genome. Figure 2.3 shows an illustration of a pair of homologous chromosomes which differs in a single nucleotide in the sequenced DNA fragments. One of the pair has a DNA sequence of CCTTCGAAAC while the other has CCTTTGAAAC. The difference is only in the fifth nucleotide. One has allele C and the other has allele T. Nowadays, biallelic SNP markers are commonly used because they are frequent in the genome and can be easily assayed in the laboratory. In this dissertation only biallelic SNP markers are considered.

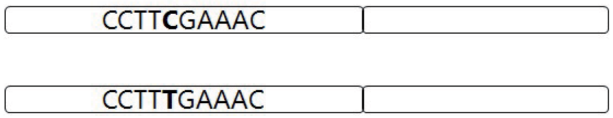


Figure 2.3: Homologous chromosomes with C/T polymorphism

2.2 Mendelian principles

How are traits inherited? In the 1860s, a monk named Gregor Johann Mendel (1822 - 1884) developed a concept about the inheritance of traits based on his experiments in breeding pea plants at the Augustinian Abbey of St. Thomas in Brno, Czech Republic. At that time when there was no evidence yet for genes, Mendel concluded that pairs of unseen ”factors” were responsible for observable traits in individuals. His work was at first not widely accepted, ignored and was rediscovered only after he died. He is now known as the ”father of modern genetics” for his *Laws of Inheritance*. His results can be summarized in two basic principles described below.

The law of segregation

Humans, being diploid organisms, normally carry a pair of alleles in a specified gene. In the formation of the sex cells or gametes, the pair of alleles *segregate* or *separate* so that each of the resulting gamete carries only half of the pair to be passed on to the next generation. At conception, one gamete (sperm) from the father randomly unites with a gamete (ovum) from the mother to form the zygote which becomes the child or offspring. The probability that each parent passes on a copy of one of the two alleles to the offspring is 0.5. Mendel's first law simply states that an individual receives with equal probability one of the segregating alleles from each parent. This law of segregation has been described by Sham (1998) using a "box model". If we imagine a parent's genotype is composed of two marbles in a box, the law of segregation says that the offspring will randomly get one of the two marbles from each of the parent's box. In statistics, this is similar to the *Bernoulli process* which is a discrete-time stochastic process consisting of a sequence of independent random variable X_i, where i = 1,2,...,n, such that for each i, the value of X_i is either 0 or 1.

Using Mendel's first law, we can make a segregation table (usually termed as Punnett Square) for alleles. This table illustrates the possible genotypes of children given the genotypes of the parents. If we take as an example a biallelic gene with alleles A_1 and A_2, the segregation for the mating of a heterozygous father (A_1A_2) and a homozygous mother (A_1A_1) is as shown in table 2.2. In this mating type, the resulting genotype *segregation ratio* is 1:1:0 which means that the probabilities of having a child with genotype A_1A_1 or A_1A_2 are both 1/2 and the probability of a child with genotype A_2A_2 is zero.

Table 2.2: Possible genotypes of offsprings for a biallelic Mendelian locus

	Father's gametes	
	A_1	A_2
Mother's gametes		
A_1	A_1A_1	A_1A_2
A_1	A_1A_1	A_1A_2

According to Mendel's Law of Segregation, when both parents are heterozygous (as illustrated in figure 2.4), the distribution of the children's genotypes given the parents' genotypes are $P(A_1A_1|A_1A_2$ x $A_1A_2) = 1/4$; $P(A_1A_2|A_1A_2$ x $A_1A_2) = 1/2$ and $P(A_2A_2|A_1A_2$ x $A_1A_2) = 1/4$.

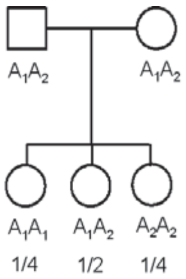


Figure 2.4: Probability distribution of genotypes of children with heterozygous parents

Some hereditary diseases follow a Mendelian mode of inheritance and are caused by just a single major gene. Known examples are cystic fibrosis and Chorea Huntington.

The law of independent assortment

The second law of Mendel says that the segregation of alleles in a gene during reproduction is independent of the segregation of alleles in other genes. Let's say we have two genes, A and B. Based on the law of independent assortment, the segregation of alleles in gene A is independent from that of gene B. This means that the segregation in gene A does not affect the probability of the segregation event in gene B and vice versa.

However, this law of independent assortment is not true for all gene pairs. Some genes do not segregate independently and are said to be linked because they tend to stay together. As a consequence, some haplotypes would be more likely observed to be passed on from parents to offsprings. The observed deviations from the law of independent assortment are the biological basis of gene mapping (Balding et al., 2001).

2.3 Hardy-Weinberg equilibrium

The theoretical prediction of the genotype distribution in a population was independently developed by the English mathematician, G.H. Hardy and the German physiologist, W. Weinberg in 1908. The Hardy-Weinberg equilibrium (HWE) simply defines the mathematical relationship between the genotype frequencies and the allele frequencies in a population. Given the case of a biallelic gene, with alleles A_1 and A_2, and with allele frequencies p_1 and p_2 respectively ($p_1+p_2=1$), the frequencies of the possible genotypes which sum up to 1 are given by:

Genotype:	A_1A_1	A_1A_2	A_2A_2
Frequency:	p_1^2	$2p_1p_2$	p_2^2

A population is said to be in Hardy-Weinberg equilibrium if the two alleles at the considered gene of a randomly chosen individual are stochastically independent and identically distributed. Therefore, given a gene with alleles A_1,...,A_k occurring with frequencies p_1,...,p_k in the population, the ordered pair of alleles (A_r,A_s) at the given gene of a randomly chosen individual has the probability p_rp_s where $r, s = 1$,...,k.

The most important implication of the Hardy-Weinberg equilibrium is the constancy of allele frequency by the mechanism of Mendelian inheritance. *Constancy* means that the allele frequency in the population after one generation remains the same in the absence of specific evolutionary forces. This also implies that the genotype frequencies are constant and thus genetic variation is preserved.

Genetically, the Hardy-Weinberg equilibrium assumes random mating in the population with respect to the gene of interest. *Random mating* describes the situation where mating is done between randomly chosen individuals. Deviations from random mating as a result of inbreeding or preferential selection of partners (assortative mating) can create correlations between uniting gametes in the population. This results in lower frequency of heterozygotes than what is expected in a population under Hardy-Weinberg equilibrium. Other assumptions of the HWE are as follows: no migration in and out of the population, no selective survival among genotypes, no genetic mutation, absence of other factors that can change the allele frequencies and large population size to avoid loss of alleles due to sampling. If all these assumptions are met, the allele frequencies can be directly calculated from the observed genotype frequencies in the population. Despite the restrictive assumptions of the HWE model, it is very relevant in practice. It has provided the foundation for experimental investigations in population genetics. It has been used as a reference model and became a baseline for comparison with realistic models in which evolutionary forces (e.g. mutation, natural selection) are considered to change allele frequencies (Hartl and Clark, 1997). It is important to check Hardy-Weinberg Equilibrium in population-based studies to avoid false positive association results. In family-based designs, HWE is not an issue in the analysis of genetic factors. However, testing for deviations from

HWE in the parents or unaffected sibling data is useful in detecting genotyping errors (Li and Leal, 2009).

2.4 Genetic models

A *genetic model* describes the mode of inheritance or the manner in which a particular genetic trait or disease is passed on from generation to generation. In statistical analysis, the genetic model specifies the parameters for the number of genes, their relationship with the trait of interest and the magnitude of their contributions. The simplest genetic model is the *Mendelian single locus model*. A single locus can be passed on from one generation to another following the principle of Mendelian segregation. The commonly used Mendelian modes of inheritance are dominant, recessive, multiplicative and additive genetic models. These genetic models are described below.

Consider for example a disease phenotype which is influenced by a single locus with k alleles. An individual can become affected or unaffected by the genetic disease depending on the genotype at the said locus. The conditional probability that an individual with a given genotype or pair of alleles will become affected by the disease is termed *penetrance*:

$$f_{A_r A_s} = P(\text{affected}|A_r A_s), \qquad r, s = 1, ..., k \tag{2.1}$$

where A_r is the allele from the father and A_s is the allele from the mother.

In the classical Mendelian model, *monogenic diseases* are caused by a single major gene and usually have a penetrance of either 0 or 1. Let us assume that our locus has only two types of alleles — A_1 and A_2, where A_1 is the ”susceptibility” allele (mutated allele type) and A_2 is the ”normal” allele (wild type). It is often assumed that the parental origin of an allele has no influence on the disease. It is also assumed in general that the probability to manifest the disease increases with the number of susceptibility alleles. Not considering the parental origin of the alleles, there are three possible genotypes for a biallelic locus. In a case-control study where the recruited subjects or participants are *cases* (individuals affected with the disease) and *controls* (not affected with the disease), the data can be presented in a contingency table as follows:

n_1 to n_6 are the genotype counts observed among cases and controls.

Table 2.3: Genotype distribution of a biallelic locus

	Genotype		
Disease status	A_2A_2	A_1A_2	A_1A_1
Case	n_1	n_2	n_3
Control	n_4	n_5	n_6

Dominant Genetic Model

In a *dominant genetic model*, an individual carrying a susceptibility allele is affected by the disease unless there is incomplete penetrance. When there is *incomplete penetrance*, it is possible that the disease does not manifest in an individual who has the susceptibility allele. In the case of *complete penetrance*, carrying at least one copy of the susceptibility allele is all that matters to manifest the disease in a dominant genetic model. Many disease susceptibility alleles affecting humans are not fully penetrant. Some have high penetrance, while others have low penetrance. An example of a highly penetrant, autosomal dominant susceptibility gene has been identified by Hall et al. (1990) among high-risk families for breask cancer and is now termed BRCA1. In the dominant mode of inheritance, the homozygous A_1A_1 and heterozygous A_1A_2 individuals have the same risks to inherit the disease but they have higher risks than the individual with the wild type A_2A_2 genotype. The disease *risk* in a case-control study can be estimated by the odds ratio (OR). In a dominant genetic model, the OR is the ratio of susceptability allele carriers to non-carriers in cases compared with that in controls. An odds ratio of 1 indicates that the odds of having the disease are the same regardless of the presence of the susceptibility allele. Rewriting the previous contingency table, a dominant genetic model where allele A_1 is the susceptibility allele is depicted in table 2.4. In this situation, the individuals with genotypes containing the A_1 allele are grouped together since they have the same risk of having the disease. In the given table, the OR $= [(n_2 + n_3)n_4] \ / \ [n_1(n_5 + n_6)]$.

Table 2.4: Genotype table for a dominant genetic model with susceptibility allele A_1

	Genotype	
Disease status	A_2A_2	$A_1A_2 + A_1A_1$
Case	n_1	$n_2 + n_3$
Control	n_4	$n_5 + n_6$

Recessive Genetic Model

Recessive genetic models are also commonly specified in genetic analysis. In this model, only the homozygotes with two copies of the susceptibility allele manifest the disease if there are no other factors involved. Table 2.5 is a revised table 2.3 showing a recessive genetic model where the A_1A_1 genotype is required for disease risk.

Table 2.5: Genotype table for a recessive genetic model with susceptibility allele A_1

	Genotype	
Disease status	$A_2A_2 + A_1A_2$	A_1A_1
Case	$n_1 + n_2$	n_3
Control	$n_4 + n_5$	n_6

Cystic fibrosis (CF) is an example of an autosomal recessive disease affecting the mucus lining of the lungs leading to breathing problems and other difficulties. It has an incidence of 1 in 2000 - 3000 newborns in Europe (WHO, 2004). Typically, only those individuals with two copies of the susceptibility allele, will manifest the disease in a recessive genetic model. However, there can be cases of *phenocopies* where individuals without the susceptibility allele can become affected.

Multiplicative Genetic Model

Another genetic model is the *multiplicative genetic model*. In this model, the risk of developing a disease increases by a factor r for a heterozygous carrier of the susceptibility allele (i.e. A_1A_2) and r^2 for a homozygous (i.e. A_1A_1) (Lewis, 2002). This model is frequently used for quantitative trait analysis.

Additive Genetic Model

The other frequently used genetic model is the *additive model*. In an additive model, the risk conferred by a disease allele is increased r-fold for heterozygotes A_1A_2, and increased 2r-fold for homozygotes A_1A_1. In this model the heterozygotes have half the risk of the A_1A_1 homozygotes (Lewis, 2002). An example of this genetic model is seen in the study of Talmud et al. (2002) where some polymorphisms have been found to have additive effects on plasma triglyceride levels which are major independent risk factors for coronary heart disease. This additive genetic model is also frequently used in the analysis of continuous traits.

There are other modes of inheritance such as those controlled by the sex chromosomes that are not discussed in this dissertation.

Mixed Model

For continuous or quantitative traits, one can utilize a linear regression model to relate the random variable Y_i, which represents the quantitative phenotype, to the genotype, such as:

$$Y_i = \mu + \beta_G G_i + \varepsilon_i \tag{2.2}$$

where $i = 1,...,N$ study subjects, μ is the population mean, β_G the regression coefficient for the genotypic effect and the residual ε_i is commonly assumed to be normally distributed, with mean zero and variance σ^2. The covariate G_i quantifies the genotype and is coded depending on the assumed mode of inheritance.

In statistical genetics, models that are more complex than equation 2.2 are often used to accommodate other risk factors that may affect the phenotypic trait or disease phenotype of interest. For quantitative traits, the so-called *mixed model* is commonly used. This model contains a major locus, a polygenic component (small additive effects of many genes) and an environmental effect. Usually, the random variable Y representing the quantitative trait is assumed to be a linear function of the three independent sources. Following Morton and MacLean's (1974) representation, the mixed model can be roughly written as:

$$Y = G + H + E \tag{2.3}$$

where G is the effect due to a major locus, H is the polygenic effect due to an indefinitely large number of small additive genetic factors, and E is the environmental contribution which also includes the error term.

Assume a biallelic major locus with alleles A_1 and A_2 and corresponding population frequencies p_1 and p_2, respectively, where $p_2 = 1 - p_1$. The characteristic of the major locus is shown in figure 2.5 and can be summarized as:

Genotype at the major locus:	A_2A_2	A_1A_2	A_1A_1
Frequency based on HWE:	p_2^2	$2p_1p_2$	p_1^2
Effect of the major locus:	$\mu_0 = g$	$\mu_1 = g + td$	$\mu_2 = g + t$

The notation t refers to the displacement at the major locus, d is the so-called *degree of dominance* and g is a parameter which is used to estimate the mean of the major locus. The major locus has mean $\mathrm{E}(G) = g + p_1^2t + 2p_1p_2td$ which is denoted here as μ. Since μ which is equal to $\mathrm{E}(G)$ is more easily estimated than g, we can use μ instead of g. By substituting $\mathrm{E}(G)$ with μ, then $g = \mu - p_1^2t - 2p_1p_2td$. The parameter t is taken here to be greater than zero so that the subsitution of an A_1 allele for A_2 represents a positive contribution to the trait Y. This means that $d = 0$ corresponds to a recessive contribution, $d = 1$ to a dominant one and $d = 1/2$ to an additive contribution. The variance of the major locus is $\sigma_G^2 = p_1^2(g+t)^2 + 2p_1p_2(g+td)^2 + (1-p_1)^2g^2 - \mu^2$. In figure 2.5, the mean effects of three different genotypes A_2A_2, A_1A_2 and A_1A_1 are denoted by μ_0, μ_1 and μ_2 respectively. The μ, t, p_1 and d are the parameters of the major locus that need to be estimated.

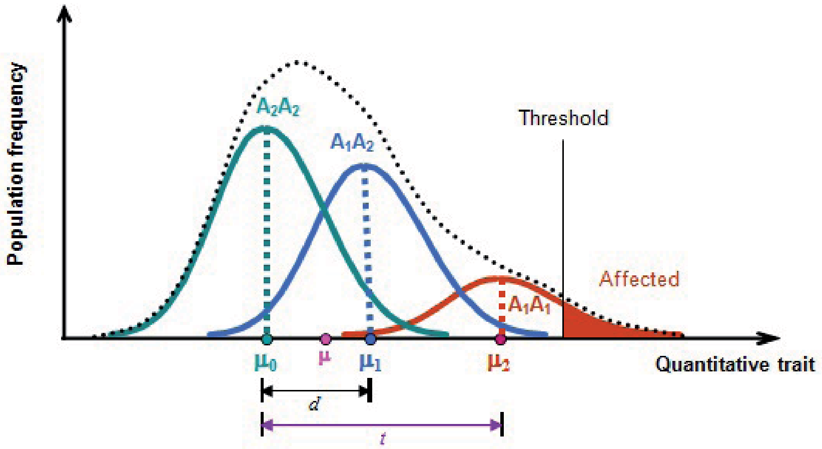


Figure 2.5: The mixed model for a quantitative trait
Based on Morton and Maclean (1974)

The polygenic effect H in Equation 2.3 is normally distributed with mean zero and variance σ_H^2, which is to be estimated. The environmental effect E can be partitioned into two parts, $E = E_c + E_r$, where E_c is the common environment and E_r is random. Each effect is assumed independently normal with mean zero and variances σ_c^2 and σ_r^2 respectively, which are also to be estimated. It follows that the total environmental effect is also normal with mean zero and $\sigma_E^2 = \sigma_c^2 + \sigma_r^2$.

In complex segregation analysis which aims to detect a major locus effect, the mixed model is usually applied. In the model, we assume that the quantitative trait is related to disease affection through a *liability* and a threshold which yields disease affection. Therefore, an underlying liability-threshold model is assumed. This underlying liability is modelled like a quantitative trait. Individuals with the disease have a liability above a certain threshold and individuals without the disease have a liability below the threshold. The single locus model is regarded as a submodel of the mixed model without the polygenic component. On the other hand, the polygenic model is also a submodel of the mixed model but without the major locus. To test if a major locus is associated with the quantitative trait, a likelihood ratio test may be applied (Sham, 1998).

Other genetic models such as the general transmission model, unified model and regressive models have been applied to quantitative traits (see Sham (1998) for details). Extensions of these models have also been proposed to accommodate two or more loci, gene-gene interaction and gene-environment interaction. In the analysis of the effect of gene-environment interaction on the phenotypic trait, several models have been used to take into account the various ways in which genetic effects can be modified by environmental exposures. Gene-environment interactions have been shown to play a role in many diseases such as skin cancer, myocardial infarction and asthma (Hunter, 2005). Detecting gene-environment interactions is oftentimes difficult. Environmental data are not always easy to measure and collect even in well-designed studies. This dissertation recognizes the importance of studying gene-environment interaction. However, the focus of the thesis is on genetic main effects and gene-gene interaction which is also a significant source of variation in many phenotypic traits. A detailed discussion of analyzing and modelling gene-gene interaction or epistasis is included in this chapter as a separate section.

In most cases, the underlying genetic model is not known. This makes the analysis of genetic factors more difficult especially in complex diseases such as depression, diabetes, cancer and ischemic heart disease. Complex diseases are usually caused by multiple genetic and environmental factors including their interactions. Knowing the underlying genetic model can help make a more efficient statistical analysis.

2.5 Analysis methods in genetic epidemiology

Genetic epidemiology resulted from the interaction of two scientific disciples: Genetics and Epidemiology. Its main aim is to investigate genetic components and risk factors that influence diseases and other phenotypic traits in families and or populations. Analysis methods in genetic epidemiology may be descriptive or analytic in nature using either family or population data. Descriptive studies characterize the distribution of genetic traits and diseases, while analytic studies usually investigate the factors affecting the distribution of genetic traits and their role in health and disease in families and populations. The following is an overview of research strategies commonly used in genetic epidemiology.

2.5.1 Segregation analysis

Studies in genetic epidemiology are usually motivated by observed clustering or aggregation of diseases in families. *Familial aggregation* of diseases may indicate biologically inherited susceptibility or common environmental exposure of family members. To further investigate diseases that cluster within families, a *segregation analysis* may be carried out. This analysis tests explicit modes of inheritance on family data. The strategy relies on fitting genetic models that could best explain the data and identify major genes that may control traits associated with the disease or contribute to the disease risk. Segregation analysis was initially designed to test whether an observed mixture of phenotypes among offsprings follows Mendelian segregation ratios.

A basic approach in performing segregation analysis can be illustrated by a simple discrete phenotype e.g. affected versus nonaffected by a disease. Let us assume that there is only one mating type or combination of the father and mother's genetic characteristic. The goal of the analysis is to estimate the probability of any given offspring being affected (denoted by P_D) and test for departure from Mendelian expectations. Assuming that the children within a sibship are independent observations, the binomial distribution that describes the probability of observing M affected offspring from a total of n sibs is:

$$P(M; n, P_D) = \binom{n}{M} (P_D)^M (1 - P_D)^{n-M} \tag{2.4}$$

Equation 2.4 serves as the likelihood function for one sibship. For a sample involving N independent sibships from different families with the same mating type, the likelihood of the total sample is the product of the above binomial function over all sibships:

$$\prod_{i=1}^{N} \binom{n_i}{M_i} (P_D)^{M_i} (1 - P_D)^{n_i - M_i} \tag{2.5}$$

where $i = 1, ..., N$ sibships of size n_i with M_i affected sibs.

Consider for example a locus of interest with two alleles A_1 and A_2, where A_1 is the susceptibility allele. For a rare autosomal dominant disease (with complete penetrance and no etiologic heterogeneity), a mating between an affected heterozygous (A_1A_2) individual and a nonaffected homozygous (A_2A_2) individual has an expected segregation ratio of 0.5 for affected and unaffected offsprings, according to Mendel's law. If we consider only families of this mating type, a test based on the binomial distribution can be applied considering the null hypothesis that the probability for a single child to be affected is 0.5. If this null hypothesis is not rejected, it may be concluded that the data are consistent with an autosomal dominant disease model. In general, the probability distribution for all possible mating types can be constructed. However, families are usually sampled or ascertained based on some recruitment criteria and not by random method. In most cases, there is an oversampling of families prone to have the disease. Therefore the test procedure should be corrected for this *ascertainment bias*. If we assume for example that families that are ascertained to participate in a study are those with "at least one affected offspring", the binomial distribution for the number of affected offspring could be corrected for ascertainment by considering a truncated binomial distribution assuming at least one affected offspring per family. However, this could mean that families with more affected children have a higher probability to be ascertained as part of the sample. (See Sham (1998) for other test procedures to correct for ascertainment bias).

In the case of extended pedigrees which encompass individuals in several generations, Elston and Stewart (1971) proposed a pedigree likelihood (denoted here as L) which is expressed as a multiple sum of products of penetrance, population and transmission parameters over all possible combinations of genotypes of the pedigree members. This *Elston-Stewart algorithm* which is widely used in segregation analysis is written as:

$$L = \sum_{g_1} \sum_{g_2} ... \sum_{g_{n_t}} \prod_{1}^{n_t} f(Y_i|G_i) \prod_{1}^{n_1} P(G_i) \prod_{1}^{n_2} \varphi(G_i|G_{if}G_{im}) \tag{2.6}$$

where g_1 represents all possible genotypes of individual 1 and g_2 all the possible genotypes of individual 2 and so on. n_t is the total number of individuals in the pedigree, n_1 the number of *founder* individuals (those without specified parents in the pedigree) and n_2 the number of *non-founder* individuals ($n_2 = n_t - n_1$). G_i, $i = 1, ..., n_t$, is the genotype of the ith individual in the pedigree, while $f(Y_i|G_i)$, $i = 1, ..., n_t$, is the penetrance which denotes the conditional probability that an individual i has an observed discrete phenotype Y_i given the genotype G_i or an analogous conditional density function for the genotype if the observed phenotype is a continuous variable. $P(G_i)$, $i = 1, ..., n_1$, is the genotype distribution for the founders which is determined by population parameters and often assuming Hardy-Weinberg equilibrium. The *transmission probability* which is the probability of an individual having a certain genotype given the parents genotypes is given by $\varphi(G_i|G_{if}G_{im})$, $i = 1, ..., n_2$, where $G_{if}G_{im}$ are the genotypes of the parents of the ith non-founding individual. The transmission probabilities are specified for all non-founder individuals in the pedigree and it is assumed that genetic transmissions to each offspring are independent of each other and that transmission of one parent to an offspring is also independent of the transmission of the other parents. The Elston-Stewart algorithm is sometimes known as "peeling" or "clipping" because it deals with large multi-generational pedigrees by considering one family at a time. It was designed to deal with large pedigrees but not with large number of loci. For an approach that considers multipoint likelihood that combines information from many loci, one can use the Lander-Green algorithm (Lander and Green, 1987).

One example of a successful application of segregation analysis for complex diseases is seen in breast cancer. The result of the segregation analysis of Newman et al. (1988) provided basis for the linkage study of Hall et al. (1990) which led to the identification of a rare autosomal dominant breast cancer gene with high penetrance. The gene for early-onset familial breast cancer has been identified on chromosome 17q21.

2.5.2 Linkage analysis

Genetic linkage or *cosegregation* is the tendency for genes or segments of DNA to be inherited together. Normally, during meiosis, homologous chromosomes pair up and partly overlap. Chromosome breakage and exchange of DNA segments can also occur during meiosis. This exchange of DNA segments between homologous chromosomes is called *chromosomal crossing overs* or *crossovers*. The process of crossover can result in *genetic recombination* between loci which is seen when the resulting gamete exhibit a different combination of DNA segments other than that of the parents. Consider the example in figure 2.6 showing two loci: A and B. Two homologous chromosomes form pairs resulting in a *tetrad* (stage II). Then crossover happens in stage III at the position between locus A and B resulting in recombination in the two middle chromosome strands. Exchange of alleles happen at locus B. In the end, each gamete will receive one of the chromosome strands. Two out of the four gametes are *recombinants*. It is possible to have multiple number of crossovers. However, if there are even number of crossovers, one sees no recombinant when genotyping the loci A and B.

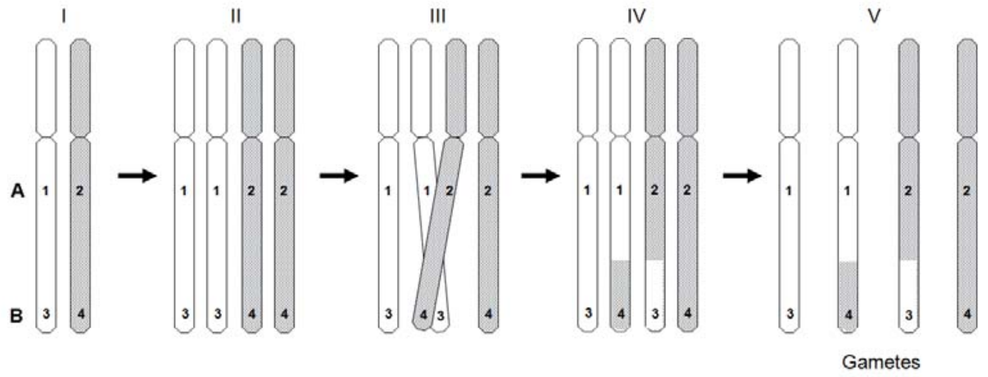


Figure 2.6: Crossover and recombination during meiosis

This information on recombination is used in linkage analysis to infer the relative positions of genes for different traits and diseases. In linkage studies, data from family members or related families (pedigree) are examined to determine the patterns of allele transmission from parents to offsprings, or the patterns of allele sharing among relatives. For linkage analysis, one parent must be a double heterozygote for both the marker and the hypothesized disease locus being studied. The probability of the marker being informative for linkage analysis is thus a function of the frequency of heterozygotes. Linkage analysis is based on measuring the cosegregation of loci or genes by determining the recombination fraction in the family data. The *recombination fraction*, θ, is the ratio of the number of recombinant gametes to the total number of gametes formed. Consider two loci: A (the putative disease locus) with possible alleles $A_1, ..., A_4$ and B (the marker locus) with possible alleles $B_1, ..., B_4$. Figure 2.7 shows an example family with their genotype data on the two loci. In this case, we know the *phases* of the individuals in the second and third generation of the pedigree. The *phase* refers to the information on the location of alleles at different loci. Two alleles at two different loci are said to be *in phase* when they are located in the same haplotype or belonging to the same chromosome, otherwise they are *not in phase*. In figure 2.7 one can see that the second offspring in the last generation of the example family has one recombination. The haplotype A_1B_4 must have been a result of a crossover in the mother (circle figure in the middle level). In the last generation of the given family, there are 11 non-recombinants and 1 recombinant out of the 12 observed informative meioses. This information is used in the computation of the likelihood of the recombination fraction.

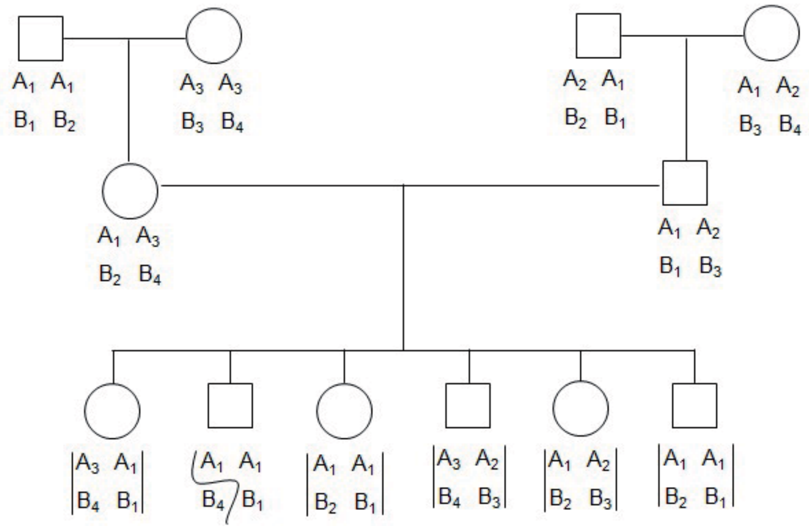


Figure 2.7: Recombination in the last generation of an example family

The aim of linkage analysis is to find evidence for linkage and to estimate recombination fractions. In the classical linkage analysis, the null hypothesis $H_0 : \theta = \frac{1}{2}$ (no linkage) is tested against the alternative hypothesis $H_1 : \theta < \frac{1}{2}$ (linkage). The usual statistical approach to linkage analysis is through the computation of the likelihood of odds (LOD) score (Morton, 1955) which tests for linkage between a susceptibility disease locus and a marker locus. The LOD score determines the likelihood of the two loci being linked given a recombination fraction versus the likelihood that they are unlinked. The likelihood of a given recombination fraction, θ is :

$$L(\theta) = \binom{l+u}{u}(1-\theta)^l\theta^u \tag{2.7}$$

where u is the number of recombinants and l the number of non-recombinants from the observed informative meioses. To get the likelihood ratio, the likelihood for linkage at a given recombination fraction θ is divided by the likelihood for no linkage (i.e. $\theta = \frac{1}{2}$). Then the maximum LOD score for linkage is computed as follows:

$$\max_{0 \leq \theta \leq \frac{1}{2}} LOD(\theta) = \max_{0 \leq \theta \leq \frac{1}{2}} log(L(\theta)/L(1/2)) \tag{2.8}$$

which is the maximum of the logarithm (base 10) of the ratio of likelihoods (Balding et al., 2001).

There are cases when the meiosis is uninformative or the *phase is unknown*. As a result, it is not possible to tell whether the offsprings' haplotypes are recombinants or non-recombinants. This happens when grandparental genotypes are not available or parental genotypes are not heterozygous for both loci concerned. In this case, the analysis becomes a little complicated because the inheritance pattern needs to be estimated based on the available insufficient information. The reader is referred to Sham (1998) or Bickeböller and Fischer (2007) for details on this. Using LOD score analysis, how can one assess if a marker is linked to a disease locus? Conventionally, a LOD score of 3 or higher is considered as evidence for linkage at any value of θ between 0 and $\frac{1}{2}$, while a LOD score of less than -2 is evidence against linkage. These values were computed by (Morton, 1955) on the basis of the sequential probability ratio test by (Wald, 1945). In large samples, a maximum LOD score of 3 is associated with the significance level of 0.0001 (Sham, 1998).

2.5.3 Association analysis

Another analysis method used in genetic studies is association analysis. Traditionally, association studies are conducted not using families but using unrelated group of subjects affected with a disease (cases) and a group of unaffected subjects (controls). In genetics, case-control studies compare marker allele frequencies between group of unrelated affected individuals and unrelated unaffected individuals to assess the contribution of genetic variants to the trait or disease of interest (Laird and Lange, 2006). An association between marker alleles and alleles of susceptibility gene will show that certain marker alleles will be present more often in cases than in controls. Association studies in genetics aim to show evidence for association or linkage disequilibrium in a population. The term *linkage disequilibrium* needs to be distinguished from the term *linkage*. While linkage is the co-inheritance at two loci observed in families, linkage disequilibrium is the relationship between alleles at two loci in a population. Linkage is independent of the allele frequencies but linkage disequilibrium is affected by the frequencies of alleles in the population.

To describe the concept of linkage disequilibrium, consider for example a putative disease locus A with k alleles $A_1, A_2, ..., A_k$ occurring in the population at frequencies $p_1, p_2, ..., p_k$ and a marker locus B with m alleles $B_1, B_2, ..., B_m$ with allele frequencies $q_1, q_2, ..., q_m$. We can test if the susceptibility allele A_i is associated with marker allele B_j by determining the probability of their joint occurrence and the product of their individual occurrence. If the occurrence of allele (A_i) is independent of the occurrence of allele (B_j), then the frequency of their joint occurrence in a haplotype (denoted by h_0) is equal to the product of their individual allele frequencies. This independence can be denoted by:

$$h_{0_{i,j}} = \quad P(A_iB_j) = P(A_i)P(B_j) = p_iq_j, \quad i = 1, ...k; j = 1, ..., m \tag{2.9}$$

If equation 2.9 does not hold, then the two alleles, (A_i) and (B_j) from the two different loci are associated. This case is a deviation from Mendel's law of independent assortment (Section 2.2.2). When there is tight linkage of loci in a large, closed, randomly mating population, allelic associations are maintained from generation to generation. Take for example the same loci A and B given above with their corresponding alleles and population frequencies. Let the recombination fraction between the two loci be θ and the frequencies of the joint occurrence of alleles (A_i) and (B_j)

(i.e. the haplotype (A_iB_j) in the current and next generation be (h_0) and (h_1), respectively. For simplicity of notation, the indeces are omitted from (h_0) and (h_1). Each haplotype in the next generation is either a recombinant (with probability θ) or a non-recombinant (with probability 1 - θ) with respect to loci A and B. A non-recombinant haplotype (A_iB_j) will have a probability of (h_0) while a recombinant haplotype will have a probability of (p_iq_j). The total probability (h_1) that a haplotype (A_iB_j) is transmitted to the next generation is :

$$h_1 = h_0(1 - \theta) + p_iq_j(\theta) \tag{2.10}$$

The difference in haplotype frequencies from the current generation to the next generation is therefore :

$$h_1 - h_0 = \theta(p_iq_j - h_0) \tag{2.11}$$

It is evident from the above equation that if the change in haplotype frequency ($h_1 - h_0$) is zero, there is no allelic association (i.e. $h_0 = p_iq_j$). It can also be deduced that the change in the haplotype frequency is proportional to the recombination fraction θ. When there is no change in haplotype frequencies from generation to generation, the considered loci are said to be in *linkage equilibrium*. Otherwise, they are in *linkage disequilibrium* (LD). Another term used to mean linkage disequilibrium is *gametic disequilibrium*.

If we denote the coefficient of linkage disequilibrium as δ_{ij}, it can be defined as:

$$\delta_{ij} = P(A_iB_j) - P(A_i)P(B_j) \tag{2.12}$$

If δ_{ij} is zero, then the probability of the joint occurrence of A_i and B_j is equal to the product of their individual probabilities. If δ_{ij} on the other hand is not equal to zero, then the marker and the disease locus are not independently occurring and we say that they are associated. When an association is found, it could mean that the associated allele is the susceptibility allele itself or it is in linkage disequilibrium with the susceptibility allele at the disease locus. In the latter case, the disease locus and the marker locus are very close to each other.

The term linkage disequilibrium may be somehow misleading since it may be taken to imply that the loci involved are linked. As previously defined, LD and linkage are two different concepts. Although LD may indicate linkage, LD may not be necessarily due to linkage. It can be affected by the presence of two or more subpopulations with different allele frequencies resulting into population stratification (Devlin et al., 2001a, 2001b). This is further explained in the next section. Among other factors, population association studies can give a false positive association result when there is population heterogeneity or stratification (Freedman, et al., 2004). Population stratification also leads to increased type-1 error and/or decreased power (Laird and Lange, 2006).

Effect of Population Stratification in Association Studies

To show the effect of population stratification in association studies, Devlin et al. (2001b) illustrated the following example. Let us assume Q is an indicator of membership in a subpopulation and $K = 1, ..., m$ subpopulations. Let G be the genotype of locus A which is simply defined as 0 or 1, where 1 indicates the presence of a susceptibility allele A_1 and 0 the absence of it. Let also D define the presence or absence of the disease (i.e. D=1 if the individual is affected by the disease and D=0 if the individual is not affected). In a case-control setting, the case-control effect can be defined by:

$$\Delta = P(G = 1|D = 1) - P(G = 1|D = 0) \tag{2.13}$$

Under the null hypothesis of no association between the disease status and the genotype, Δ will only be nonzero if there is confounding effect. *Confounding* can be due to unobserved variables such as membership in an ethnic group which creates spurious correlations between variables. A confounding variable is associated with both the probable explanatory variable and the outcome variable. To illustrate the effect of confounding due to population stratification, let us assume under the null hypothesis that the genotype G and the disease status D are independent, conditional on the membership to a subpopulation. In addition, let us define the probability of having the susceptibility allele given the subpopulation K as $\omega_K = P(G = 1|Q = K)$. Let us also define $\nu_K = P(Q = K|D = 1) - P(Q = K|D = 0)$. Summing up the product of ω_K and ν_K in all subpopulations K will give:

$$\sum_K \omega_K \nu_K = \sum_K [P(G = 1|Q = K)][P(Q = K|D = 1) - P(Q = K|D = 0)] \tag{2.14}$$

Equation 2.14 will actually lead us to the case-control effect Δ across all subpopulations:

$$
\begin{aligned}
&\sum_K \omega_K \nu_K \\
&= \sum_K [P(G=1|Q=K)P(Q=K|D=1)] - [P(G=1|Q=K)P(Q=K|D=0)] \\
&= \sum_K P(G=1|D=1) - P(G=1|D=0) \qquad (2.15)
\end{aligned}
$$

The Δ can be positive or negative for any locus of interest even under the null hypothesis. If there were only two subpopulations and the disease and the genotype 1 is more prevalent in one subpopulation, one can see from Equation 2.14 that the resulting Δ would be positive. In case-control studies, it is therefore important that cases and controls come from the same homogeneous source population. If individuals come from subpopulations with different allele frequencies, linkage disequilibrium can be detected even without linkage. Population stratification can also result from admixture of subpopulations through immigration and non-random mating or mating according to some social caste, religion or ethnic orientation. When there is population stratification or heterogeneity in a population-based study (e.g. case-control study), false positive result or *spurious association* can arise. Because of this, family-based tests of association, which are robust to population stratification have become a popular approach in detecting genes affecting diseases.

One of the most widely known tests of association in family-based studies is the *Transmission Disequilibrium Test* (TDT), introduced by Spielman et al. (1993). The TDT procedure was initially designed to test for linkage between a genetic marker and a disease locus when an association had been found between the two. The TDT is also valid as a test for association when the families considered are all simplex i.e. consisting of parents and one affected child. The effect detected by the TDT is the combined presence of linkage and association (Ewens and Spielman, 2005). The TDT and its extensions will be discussed in detail in the succeeding chapter.

2.6 Epistasis

The term *epistasis* generally means interaction between genes, a phenomenon where the effects of one gene are modified by one or several other genes. Frequently, genes interact with one another making the nature of genetic inheritance more complicated. It is an accepted fact that interaction of genetic factors plays an important biological basis in complex diseases and phenotypic variation (Barton and Keightley, 2002; Flint and Mott, 2001; Kroymann and Mitchell-Olds, 2005; Lander and Schork, 1994; Lou et al., 2008). However, in many studies that attempt to determine the genetic basis of complex traits, epistasis is often ignored (Carlborg and Haley, 2004). Eventhough the involvement of epistasis in many complex traits is not known, Carlborg and Haley (2004) argued that it should be routinely explored. Moore (2003) also had a similar opinion about the importance of epistasis in determining susceptibility to common human diseases such as essential hypertension. An example of a statistically significant interaction has been identified by Ritchie et al. (2001) among four SNPs from three estrogen metabolism genes for sporadic breast cancer. This interaction was detected in the absence of independent main effects for any of the four SNPs.

The idea that epistasis is important can be traced back to the observation of deviations from Mendel's law. However, the issues of its definition and measurement hinder many scientists to routinely consider it in genetic studies. The following presents different definitions of epistasis in the literature.

Biological definition of epistasis

Epistasis in the biological sense can be defined in many ways. Earlier, it has been used to refer to situations in which a variant of one locus masks or suppresses the effects of a variant in another locus (Cordell, 2002). This is actually based on Bateson (1909) who introduced the term *epistatic* when DNA and genes were not yet discovered. Bateson used the term *epistatic factors* to describe factors which prevent other factors from manifesting their effects in the plant and animal genetic experiments he has observed. He noted that the phenotypic differences in the coloring in animals and plants is affected by pigment factors. However, whiteness or absence of colors in them is not only due to the absence of these pigment factors but it can also be due to the suppression of the pigment factors caused by an epistatic factor.

To describe epistasis according to Bateson's definition, consider two biallelic loci, A and B, affecting hair color in rabbits. Locus A has alleles A_1 and A_2 while locus B has alleles B_1 and B_2. The possible genotypes and phenotypic outcomes (white, black or brown hair) are shown in table 2.6. It can be noted that regardless of genotype at locus A, individuals with any copies of the B_1 allele at locus B have brown hair, i.e., at locus B, allele B_1 is dominant to allele B_2, effectively masking any effect of allele B_2. If the genotype at locus B is B_2B_2 an individual with a copy of the A_1 allele is observed to have black hair. Therefore, at locus A, allele A_1 is dominant to A_2. However, if the genotype at locus B is not B_2B_2, the effect at locus A cannot be seen. Individuals with any copies of the B_1 allele have brown hair regardless of genotype at locus A. One can say that the effect at locus A is masked by that of locus B. Specifically, allele B_1 at locus B is epistatic to allele A_1 at locus A or in general, locus B is epistatic to locus A.

Table 2.6: Example of phenotype table for two epistatic loci under Bateson's definition

	Genotype at locus B		
	B_1B_1	B_1B_2	B_2B_2
Genotype at Locus A			
A_1A_1	Brown	Brown	Black
A_1A_2	Brown	Brown	Black
A_2A_2	Brown	Brown	White

Similarly, in physiological genetics which uses physiological and molecular genetic approaches to facilitate gene identification and to study gene function, epistasis is also defined based on the phenotypic differences among individuals. Physiological epistasis occurs when the phenotypic trait differences among individuals with various genotypes at one gene or locus depend on their genotypes at other genes (Cheverud and Routman, 1995).

What is meant by biological epistasis is not always exactly the same for all fields. In a biomolecular perspective, biological epistasis can be defined as the physical interactions among proteins or other molecules that affect the phenotype or trait of interest (Moore and William, 2005). Therefore, the effect of a gene on a trait is a result of the physical interaction of its biomolecules with the biomolecules of another gene(s) within the regulatory network and biochemical pathways. The interaction can occur when transcription factors physically interact with each other or when enzymes interact through a metabolic pathway. A popular example of biological epistasis is seen in sickle cell anemia. It is an inherited blood disorder that affects hemoglobin, the protein in the red blood cells which helps carry oxygen throughout the body. Sickle cell anemia is a product of pleiotropic genes (genes affecting multiple phenotypic traits by coding for a product used by various cells or has a signaling function on different targets) and epistatic genes (Nagel, 2001). Individuals with sickle cell anemia have β-globin molecules with a neutral amino acid on the outer surface of the red blood cells. This neutral amino acid increases physical interaction via intermolecular adhesion which leads to increased deoxyhemoglobin and causes deformed red blood cells (Moore and William, 2005). It has also been reported that genetic variants in the haptoglobin gene interact with the S allele of the hemoglobin β-chain gene (Giblett, 1969).

In general, biological epistasis usually refers to a situation where the mechanism of action of one factor is influenced by the presence or absence of another factor. Biological epistasis can be assessed using laboratory methods. For example, protein-protein interaction can be detected using the yeast two-hybrid system which dates as early as 1987 with the works of Stanley Fields at the State University of New York at Stony Brook (Bartel and Fields, 1997). This system uses two different proteins, one acts as a binding domain (the ”bait”) and the other as an activation domain (the ”prey”). The proteins are expressed in two different haploid yeast strains which are mated to determine if the two proteins interact. Mating of the yeast strains results in fusion of the two haploid yeasts to form a diploid yeast strain. The interaction can then be determined by measuring in the diploid strain the activation of a two-hybrid reporter gene. There are other laboratory strategies that confirm protein-protein interactions. These strategies have been very successful in model organisms but are quite complicated in humans due to the fact that there are much more interactions in humans than other model organisms (Moore and William, 2005). Bork et al. (2004) estimated roughly 10,000 - 30,000 pairwise interactions among yeast proteins and up to 200,000 or more protein interactions in humans. However, there are high hopes for detecting more biological epistasis with the rapid technological advances in laboratory assays.

Statistical definition of epistasis

Statistics has been used in many instances to complement results in the biological and medical fields. In the field of statistics, the term *epistacy* had earlier been used by Fisher (1918) to describe deviations from additivity. This implies that variation of genetic traits in the population is not predictable based only on the individual actions of the genes. Fisher considered a linear model to determine the contribution of different loci to a quantitative trait and subdivided the hereditary variance into additive effects (resulting from average effects of genes), dominance effects (allelic interaction in a gene) and epistatic effects (interaction between genes). From its earliest use by Fisher, the term *epistasis* in statistical genetics typically means that the effects of different loci are not additive. Falconer (1989) also defined epistatic deviation as the "deviation of multilocus genotypic values from the additive combination of their single-locus components" .

Other ways to describe the mathematical models for epistasis focus on the penetrance or genotype values to make it more relevant to real biological situation. To illustrate the concept, consider again two biallelic loci, A and B, affecting a disease trait. Locus A has genotypes A_1A_1, A_1A_2, and A_2A_2 while locus B has genotypes B_1B_1, B_1B_2 and B_2B_2. Alleles A_1 and B_1 are the susceptibility alleles. A two-locus disease model on two biallelic loci can be specified by 9 genotypes (see table 2.7) which may cause different genetic effects. These effects are termed *genotype values* by Hallgrímsdóttir and Yuster (2008). Table 2.7 shows the genotype values g_{ij}, where i, j = 0,1,2 refers to the number of susceptibility alleles at the first and second locus, respectively. For a dichotomous phenotype, the genotype value can be the penetrance associated with the genotype, the logarithm of the penetrance or the logarithm of the odds ratio. In the case of a quantitative trait, the genotype value g_{ij} can be the expected phenotype of an individual with genotype ij.

Suppose that the trait of interest is a dichotomous trait and that a susceptibility allele is required at both loci to exhibit the trait. This means that one or more copies of allele A_1 and B_1 leading to complete penetrance are required. Then, we obtain the example penetrance table shown in table 2.8 when the effects of the two loci are considered.

In table 2.8, the effect of allele A_1 can only be observed when allele B_1 is also present. Without B_1, the effect of A_1 is not observable. The effect at the first locus

Table 2.7: Genotype values in a two-locus disease model

	Genotype at 2nd locus		
	B_1B_1	B_1B_2	B_2B_2
Genotype at 1st Locus			
A_1A_1	g_{00}	g_{01}	g_{02}
A_1A_2	g_{10}	g_{11}	g_{12}
A_2A_2	g_{20}	g_{21}	g_{22}

Table 2.8: Example of penetrance table for two epistatic loci

	Genotype at 2nd locus		
	B_1B_1	B_1B_2	B_2B_2
Genotype at 1st Locus			
A_1A_1	1	1	0
A_1A_2	1	1	0
A_2A_2	0	0	0

(A) appeared to be 'masked' by the effect at the second locus (B). Locus B is said to be epistatic to locus A because when the genotype is B_2B_2 at locus B, the effect of the alleles at locus A cannot be seen. It can also be said that locus A is epistatic to locus B because when the genotype at locus A is A_2A_2, the effect of the alleles at locus B is not observable. This situation is not precisely similar to the original concept of Bateson that if a factor B is epistatic to factor a A, then factor A cannot be expected to also be epistatic to factor B. This is evident by the lack of symmetry in table 2.6.

The absence of epistasis has been often represented by a *heterogeneity model* wherein the two loci are independent causes of the disease and an individual becomes affected through having a predisposing genotype at either of the loci (Risch, 1990). This is illustrated in table 2.9. The biologically motivated definition, however, has some problems. If we consider a recessive disease model (i.e. two copies of allele B_1 are required to cause the disease) then having two copies of allele A_1 at locus A is enough to 'mask' the effect of B_1. With genotype A_1A_1 at locus A, one cannot observe the effect at locus B. Locus B acts differently depending on the genotype at locus A. This makes the heterogeneity model a case of interaction.

The confusing definitions of epistasis make it difficult to analyze it in practice. In statistical testing, we can only take into account mathematical models such as those

Table 2.9: Example of penetrance table for two loci in a heterogeneity model

	Genotype at 2nd locus		
	B_1B_1	B_1B_2	B_2B_2
Genotype at 1st Locus			
A_1A_1	1	0	0
A_1A_2	1	0	0
A_2A_2	1	1	1

described by Fisher and the likes. In modelling quantitative traits, the phenotypic value Y is often decomposed into additive (a_1 and a_2) and dominance (d_1 and d_2) main effects at the first and second locus respectively, and four epistatic effects: additive x additive (i_{aa}), additive x dominance (i_{ad}), dominance x additive (i_{da}), and dominance x dominance (i_{dd}). The general genetic model for a quantitative trait involving two loci is given by the following linear model notation of Cordell (2002):

$$\begin{aligned} Y = \mu + a_1 w_1 + d_1 z_1 + a_2 w_2 + d_2 z_2 \\ + i_{aa} w_1 w_2 + i_{ad} w_1 z_2 + i_{da} z_1 w_2 + i_{dd} z_1 z_2 \end{aligned} \tag{2.16}$$

where μ is the population mean and w_i and z_i are dummy variables related to the genotype at locus i = 1,2. We can set w_1 = 1 and z_1 = -0.5 for an individual with genotype A_1A_1, w_1 = 0 and z_1 = 0.5 for genotype A_1A_2 and w_1 = -1 and z_1 = -0.5 for genotype A_2A_2. The dummy variables w_2 and z_2 for the second locus are defined similarly. Without epistasis, the interaction coefficients i_{aa}, i_{ad}, i_{da}, i_{dd} become zero, reducing equation 2.16 to a simple additive model or non-epistatic model:

$$Y = \mu + (a_1 w_1 + d_1 z_1) + (a_2 w_2 + d_2 z_2) \tag{2.17}$$

What has been described so far are the classic methods used in detecting epistasis. In the past years, other statistical methods have been proposed by several authors. In the context of high dimensional data, there are combinatorial partitioning and data-mining methods (Nelson et al., 2001; Ritchie et al., 2001; Cook et al., 2004) which may be applicable in many situations (e.g. presence of covariates). However, they usually require intensive computations and in addition to the lack of clear biological interpretation, the power to detect epistasis in these methods may depend on the structure of the data (Zhao et al., 2006). With these limitations in mind, Zhao and colleagues proposed a test for epistasis that uses the easily-computed linkage disequilibrium in the cases. They examined how the interaction of two loci produces linkage disequilibrium in disease-affected individuals and used linkage disequilibrium to measure gene-gene interaction between two unlinked loci. They showed mathematically that their new definition of gene-gene interaction of two loci is similar to the linkage disequilibrium of the said loci. Their method is equivalent to testing for departure from additivity of the log penetrance values. The approach claims to have more power than the traditional logistic regression under the two-locus disease model. However, like all methods that test for departure from additivity on a particular scale, they provide no information on the type of interaction present if the additive model is rejected. The issue of scale is a long time challenge in detecting epistasis statistically. With Fisher's definition of epistasis as deviation from additivity, the choice of scale becomes important because one can actually remove or induce epistasis by simply changing the scale. It has been emphasized in the summary paper of An et al. (2009) that epistasis is a relative concept that should be carefully interpreted with respect to the particular scale of reference.

The traditional four types of epistatic effects described by Cockerham (1954) were supplemented by Hallgrímsdóttir and Yuster (2008) who described a complete classification of epistatic two-locus models. Following the works of Li and Reich (2000) who enumerated two-locus disease models for a dichotomous trait, the authors provided a classification that considers continuous quantitative traits. Their approach is geometric, showing that there are 387 distinct types of epistatic two-locus models. The comprehensive list of epistatic two-locus model provided by Hallgrímsdóttir and Yuster (2008) can aid in classifying the type of epistasis. However, its role in determining genetic variability in populations is not addressed in the study. In this thesis, only the traditional epistatic effects are considered.

Specific methods that are applicable to characterize epistasis for quantitative traits in family-based designs are not well developed (Li et al., 2007). The significance of studying epistasis, especially in complex diseases, prompted many authors to try different methods of statistical analysis. There are new methods that use interaction testing framework in two stages (Millstein et al. 2005, 2006). Other authors tried a unified model to determine both functional (also known as physiological) and statistical epistasis (Alvarez-Castro and Carlborg, 2007). In the Genetic Analysis Workshop 16, there are also several other methods proposed to detect epistasis, but the lack of consistency of results indicates that the challenge is still on in solving this analysis problem. Despite the many statistical methods available in the literature, there are still some gaps in knowledge that require further research. The next chapter presents some common statistical methods used in family-based genetic studies, focusing on the *transmission disequilibrium test* (TDT) and TDT-like tests. Most of these methods also consider testing for gene-gene interaction.

3 Transmission Disequilibrium Test

In genetic studies, family-based designs offer a unique advantage over population-based designs in terms of its robustness against population admixture and stratification (Laird and Lange, 2006; Lewinger and Bull, 2006; Thomson et al., 1989). This chapter first introduces the original version of Spielman and colleagues' statistical test for a family-based design — the *Transmission Disequilibrium Test* (TDT). Some of the commonly used extensions and modifications of the test are also presented in this chapter.

3.1 History and description of the TDT

The TDT was introduced by Spielman, McGinnis and Ewens in 1993 in their investigation of genes that contribute to the susceptibility for type 1 diabetes mellitus (also known as 'IDDM' or insulin dependent diabetes mellitus). At that time, there were already studies showing that variation in the HLA region of chromosome 6 influences susceptibility to IDDM. However, there were discrepancies in the results between association and linkage studies. This became one of the motivations for the authors to devise a method that would provide a valid test for linkage in the presence of association or vice-versa (Ewens and Spielman, 2005).

The test evaluates the frequency of transmission of alleles from heterozygous parents to the affected offspring. In its original form, the TDT uses genotype data from a sample of N random individuals from the population of *affected* individuals and their parents. The main advantage of the TDT is its robustness to population admixture as shown by several authors, e.g. McKeigue (1997) and Li et al. (2008). By using the nontransmitted marker alleles of the parents as the "control" alleles, the "within-parental" matching overcomes population stratification in the TDT. For a dichotomous trait and a biallelic locus, the test is actually equal to the conventional *McNemar test* (McNemar, 1947). For simplicity of illustration, let us consider only families with one child. Let us assume that we are interested in a disease locus A. To test for this disease locus, we use a known biallelic marker locus B, with alleles B_1 and B_2. We genotype the families with regard to the marker locus B. After laboratory testing, some individuals may be homozygous at the said marker (i.e. B_1B_1 or B_2B_2), while others may be heterozygous (i.e. B_1B_2). Using the genotype results from the laboratory, we can now determine which of the parental alleles have been transmitted to the child and which are not. Under the null hypothesis of no linkage, the transmissions from two parents of an affected child are independent. A data example of a single family is shown in figure 3.1.

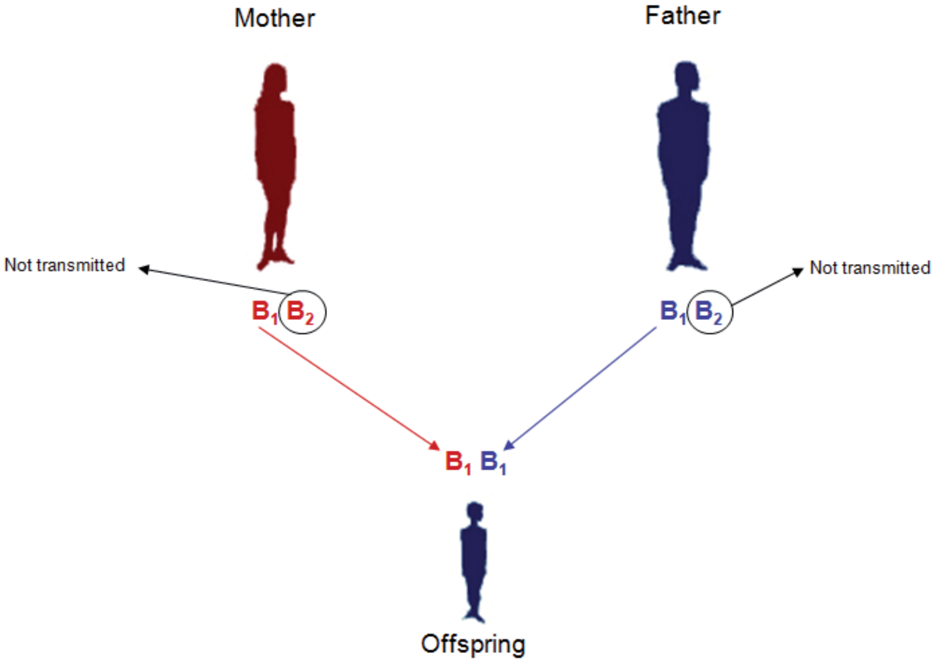


Figure 3.1: Transmitted and nontransmitted alleles in a family

In the given example, the heterozygous father transmits his B_1 allele and the heterozygous mother also transmits her B_1 allele. Therefore, the child's genotype at the marker B is B_1B_1. If we have a sample of N single-child families, there will be a total of $4N$ parental alleles. Half of these come from the fathers and the other half from the mothers. From the available $4N$ alleles, $2N$ are transmitted and $2N$ are not transmitted. Table 3.1 shows the set-up of the data on the marker alleles of the families. Each parent has an unordered genotype B_1B_1, B_1B_2 or B_2B_2 at the marker, and is counted in one of the four cells in the table. For a homozygous parent, transmitted and nontransmitted alleles are identical. These parents are included as counts in cells N_{11} or N_{22}. A heterozygous parent with genotype B_1B_2 is counted in cell N_{21} if he or she transmits allele B_2 and is counted in cell N_{12} if he or she transmits allele B_1. This gives a total count of $N_{11} + N_{12} + N_{21} + N_{22} = 2N$ parents.

Table 3.1: TDT table of marker alleles among $2N$ parents of N affected children

	Nontransmitted Allele		
	B_1	B_2	Total
Transmitted Allele			
B_1	N_{11}	N_{12}	$N_{11} + N_{12}$
B_2	N_{21}	N_{22}	$N_{21} + N_{22}$
Total	$N_{11} + N_{21}$	$N_{12} + N_{22}$	$2N$

Example data and counts of parents corresponding to the cells in table 3.1 are presented in table 3.2. For a two-allele locus, there are six possible parental mating types or genotype combinations of the parents: B_1B_1 x B_1B_1, B_1B_1 x B_1B_2, B_1B_1 x B_2B_2, B_1B_2 x B_1B_2, B_1B_2 x B_2B_2 and B_2B_2 x B_2B_2. In the first row of table 3.2, both homozygous parents transmit one of their B_1 alleles to the child and do not transmit their other B_1. Therefore, the parents contribute two counts for the cell N_{11} of the TDT table 3.1. If the father, mother and child are all heterozygous (B_1B_2), then the appropriate cell cannot be resolved for the parents individually unless parental origin of the alleles can be established. However, such a trio does contribute a count of one in each cell N_{12} and N_{21}. Hence, the father and mother can be arbitrarily assigned to either cell (see 4th row of table 3.2).

3.2 Derivation of the test statistic

The counts of transmitted and nontransmitted alleles depend both on the frequencies of the marker alleles B_1 and B_2 in the population and the relationship between the disease locus and the marker locus. If the marker locus has nothing to do with the disease, then we would expect that heterozygous B_1B_2 parents will transmit a B_1

Table 3.2: Parental mating types and scoring of parents for the TDT table

Father	Mother	Child	N_{11}	N_{12}	N_{21}	N_{22}
B_1B_1	B_1B_1	B_1B_1	2	0	0	0
B_1B_1	B_1B_2	B_1B_1	1	1	0	0
B_1B_1	B_2B_2	B_1B_2	1	0	0	1
B_1B_2	B_1B_2	B_1B_2	0	1	1	0
B_1B_2	B_2B_2	B_1B_2	0	1	0	1
B_2B_2	B_2B_2	B_2B_2	0	0	0	2

allele or B_2 allele with equal probability to an affected child. Therefore, the expected number of entries in the off-diagonal cells N_{12} and N_{21} of table 3.1 will be equal.

In constructing the test statistic, Spielman and colleagues used the probabilities derived by Ott (1989) for the cells in table 3.1 as a function of genetic model parameters. The linkage disequilibrium (δ) and the recombination fraction (θ) between the genetic marker and the disease loci were used in the calculation of probabilities. To understand the concept, let us consider a biallelic marker locus and N families with a single affected child. The total number of transmissions to the children in these families is $2N$ and under the null hypothesis of no linkage, all such transmissions are independent including those from the two parents of the same child. Suppose that our disease locus A has alleles A_1 (the susceptibility allele) and A_2 (the normal allele) with population frequencies p_1 and $(1 - p_1)$ respectively. The genetic marker B that we used for laboratory testing is as previously defined with alleles B_1 and B_2 with population frequencies q_1 and $(1 - q_1)$. The coefficient of linkage disequilibrium between A_1 and B_1 is $\delta= P(A_1B_1) - p_1q_1$. Table 3.3 shows the joint probabilites of transmitted and nontransmitted alleles of the parents from the derivation of Ott (1989).

The TDT was originally designed to test for the null hypothesis of no linkage ($\theta = \frac{1}{2}$). However, it is also a valid test when one wants to test for the association of a disease with a susceptibility locus. It means testing not only for the null hypothesis $H_0 : \theta = \frac{1}{2}$ (no linkage), but also for $H_0 : \delta = 0$ (no association). Relating tables 3.1 to 3.3, the data values that contain both θ and δ are only the cells N_{12} and N_{21}. This implies that only data from heterozygous B_1B_2 parents are useful for the test statistic. Table 3.3 also shows that under the null hypothesis of ($\theta = \frac{1}{2}$), the expectations of N_{12} and N_{21} are equal. Thus the contributions from two heterozygous parents are independent when $\theta = \frac{1}{2}$. This knowledge was used to derive the test statistic. The

Table 3.3: Probabilities of transmitted and nontransmitted marker alleles among $2N$ Parents of affected children

	Nontransmitted Allele		
	B_1	B_2	Total
Trans-mitted Allele			
B_1	$q_1^2 + q_1\delta/p_1$	$q_1(1-q_1) + (1-\theta-q_1)\delta/p_1$	$q_1 + [(1-\theta)\delta/p_1]$
B_2	$q_1(1-q_1) + (\theta-q_1)\delta/p_1$	$(1-q_1)^2 - (1-q_1)\delta/p_1$	$1-q_1-[(1-\theta)\delta/p_1]$
Total	$q_1 + (\theta\delta/p_1)$	$1-q_1-(\theta\delta/p_1)$	1

final form of the statistic which was termed "transmission/disequilibrium" or "TDT" is:

$$\chi^2_{tdt} = (N_{12} - N_{21})^2/(N_{12} + N_{21}) \tag{3.1}$$

The test statistic which is asymptotically chi-square distributed with one degree of freedom is of the same form as the *McNemar test* (McNemar, 1947).

The TDT does not assume any mode of inheritance of the disease or trait of interest. It is applicable whatever the population stratification may be. The stratification does not affect the test because the transmitted and untransmitted alleles from each parent are matched with respect to the population of origin. As a test of association, the original TDT is valid provided that all families are simplex (i.e. with only one affected child). In the case of families with multiple affected children, using more than one affected child in the test will create a problem because as a test of association it assumes independent observations for the data, and data sampled from related affected individuals are not usually independent. As a test of linkage, the advantage of the TDT is that unlike the conventional tests for linkage, it does not require data either on multiple affected family members or on unaffected sibs. However, since it depends on both linkage and linkage disequilibrium ($\delta > 0$), it is useful only as a test for linkage when there is disequilibrium between the loci. Or it means that, the TDT is a test of linkage with power only if the disease and marker loci are associated. In general, conclusions from a TDT analysis apply to disease association studies where genetic markers are closely linked to candidate genes.

3.3 Extensions and modifications of the test

One of the earlier extensions of the TDT is its generalization to more than two marker alleles (Bickeböller and Clerget-Darpoux, 1995). In the previous sections, the given examples focus on marker B having only two alleles B_1 and B_2. Often, there can be more than two alleles at a marker locus such as shown in table 3.4. Schaid (1996) illustrated a general framework for the development and computation of score statistics that can be used for testing both linkage and linkage disequilibrium when there are multiple marker alleles. Among others, he had introduced the *generalized* TDT (or GTDT) test statistic.

Table 3.4: Combinations of marker alleles B_1, B_2,...,B_m among parents of N affected children

	Nontransmitted Allele				
	B_1	B_2	...	B_m	Total
Transmitted Allele					
B_1	N_{11}	N_{11}	...	N_{1m}	$N_{1.}$
B_2	N_{21}	N_{22}	...	N_{2m}	$N_{2.}$
...					
B_m	N_{m1}	N_{m2}	...	N_{mm}	$N_{m.}$
Total	$N_{.1}$	$N_{.2}$	...	$N_{.m}$	$2N$

Another variation of the TDT is the ”*sib TDT*” (or S-TDT). The method was proposed by Spielman and Ewens (1998) to overcome the problem of missing parental data. They used marker data from affected and unaffected children, thus allowing the application of TDT to sibship data without parental data. The S-TDT determines whether the genetic marker allele frequencies among the affected sibs differ significantly from the frequencies among their unaffected sibs while conditioning on the families. In implementing the test, the authors adopted a within-family Monte Carlo permutation procedure because the chi-square test is not valid in the non-independent observations on sibs. They have further recommended a procedure to combine the TDT and the S-TDT into an overall test.

Knapp (1999) introduced an RC-TDT (Reconstruction-Combined Transmission Disequilibrium Test) which employs parental-genotype reconstruction and corrects for the biases resulting from the reconstruction of parental genotypes. The bias in reconstructing parental genotypes was shown by Curtis (1997) to inflate the type I error rate of the TDT. There are other extensions of the TDT such as the TDT-AE (allow for error in genotyping) which was proposed by Gordon et al. (2001). Other than the above-mentioned, there is also the unbiased TDT for multilocus haplotypes (Dudbridge et al., 2000), TDT with covariates (Rice et al., 1995) and many more. This dissertation focuses on TDTs applicable for quantitative traits (QTs). Commonly used methods for quantitative traits are described in the following section.

3.4 Quantitative Transmission Disequilibrium Test (QTDT)

The original TDT and its earlier modifications have been applied to dichotomous traits. However, quantitative traits (e.g. blood sugar level, body-mass-index, blood pressure, radiation sensitivity, etc.) are also common measures and often more informative than categorized characteristics. Several authors have explored the TDT in the analysis of genetic factors affecting quantitative traits. There are two types of approaches in analyzing genetic factors affecting quantitative traits, namely, *prospective* or *retrospective* approaches. Some methods have only been developed for the one-locus model. To adopt previous notations, the quantitative trait will be represented by Y. For the the genotypes of the loci, assuming that they are biallelic, A_1A_1, A_1A_2 and A_2A_2 are assigned as the possible genotypes of locus A while B_1B_1, B_1B_2 and B_2B_2 are the genotypes of locus B. The alleles A_1 and B_1 are the susceptibility alleles.

3.4.1 Prospective QTDTs

Usually, the statistical approach in QTDT is using a *prospective* model. This involves modelling the quantitative phenotype as a function of the genotypes. Selected methods of this type of approach are described below.

TDT$_{\mathrm{Q5}}$ (Allison, 1997)

Some of the early TDT methods for quantitative traits were proposed by Allison (1997). He developed several tests taking as units of observation the parent-child trios. Among the five tests he proposed, the TDT$_{\mathrm{Q5}}$ is the one that is reasonably powerful regardless of the underlying genetic model.

Let us consider a random variable Y that represents the quantitative trait. The TDT$_{\mathrm{Q5}}$ uses data of family trios with at least one heterozygous parent. In addition, only those trios in which $Y > Z_U$ (the upper cutoff on the quantitative phenotype) or $Y < Z_L$ (the lower cutoff) are selected, where $Z_U \geq Z_L$. This is because the method is based on extreme sampling from both tails of the offspring phenotypic distribution. Setting $Z_U = Z_L$ reduces this to the case of random sampling. In a single-locus model, the number of A_1 alleles (0, 1, or 2) that the offspring have is signified by X. The event of a heterozygous parent transmitting the A_1 susceptibility allele (the allele associated with higher values of the trait) is denoted by T = 1, and the absence of transmission as T = 0. The null hypothesis has two components: $\mu_T = P(T = 1) = 1/2$ and/or E($Y|T = O$)=E($Y|T = 1$). The first is a test of no linkage. It tests if the recombination fraction between the marker and disease locus is 1/2. The second component tests if $\mu_{A_1A_1} = \mu_{A_1A_2} = \mu_{A_2A_2}$. Under the alternative hypothesis, the recombination fraction is < 1/2 and the three genotypic means at the disease locus are not all equal.

The first step in the analysis is to regress Y on the dummy codes for the three parental mating types with at least one heterozygote parent: (1) A_1A_2 x A_1A_1 (or A_1A_1 x A_1A_2); (2) A_1A_2 x A_1A_2; or (3) A_1A_2 x A_2A_2 (or A_2A_2 x A_1A_2) to obtain the R^2 for the regression which is termed "R$_1^2$". By conditioning on parental mating type, confounding due to admixture is eliminated. Additionally, the number of susceptibility alleles, X and X^2, are added to the model as predictor variables, and the R^2 for this "full" regression is obtained which is termed "R$_2^2$". The R^{2}s are used to compute the F-ratios in the usual way to test for the joint additive and dominance effects of the locus.

The method TDT$_{\mathrm{Q5}}$ can accommodate families with more than one child. However, it is complicated to extend the method in this way (Zhu and Elston, 2001). In addition, it assumes that the residual distribution is normally distributed or that the sample size is large enough to be able to rely on the Central Limit Theorem. In some cases, the assumption may not be correct and nonparametric alternatives may be more useful.

A Unified Approach to Adjusting Association Test for Population Admixture (Rabinowitz and Laird, 2000)

Rabinowitz and Laird (2000) developed an approach which is not only applicable to quantitative traits but to other phenotypic traits as well. The method was probably referred to as a unified approach because it can accommodate multiple allelic markers, all pedigree structures, covariates and all patterns of missing marker allele information. The approach is also valid regardless of the type of underlying genetic model, sampling strategy and population admixture. The approach is based on conditioning on sufficient statistics for the null hypothesis which is described later. For a set of models, the conditional distribution given the sufficient statistics is the same for all models in the set. Thus, the p-values computed conditionally on the sufficient statistics for the models in the null hypothesis will lead to the same result of rejecting the null hypothesis regardless of which model is true. The authors further justified the statement by saying that if two different realizations of the marker alleles and observed traits, Y and Y' have the same value of the observed minimal sufficient statistic, then for any value of the full minimal sufficient statistic, x, either the conditional probabilities of Y and Y' given x, $P(Y|x)$ and $P(Y'|x)$, are both equal to zero, or the ratio $P(Y|x)/P(Y'|x)$ is invariant to the choice of x.

In the proposed approach, the p-values are computed by comparing the test statistic to its conditional distribution given the minimal sufficient statistic under the null hypothesis for population admixture, the sampling plan and the genetic model. Through conditioning on a sufficient statistic, the approach results in correct type I error rates regardless of the patterns of population admixture, the sampling plan, and the genetic model. In using the method, it is assumed that phenotypic traits and genotyping information are available for some members of the family pedigrees. It is also assumed that a test statistic that is sensitive to association between traits and marker alleles is defined. The test statistic can be as simple as counting a particular marker allele or a score statistic from a joint likelihood for quantitative traits. No assumption is made on the ascertainment of study subjects except for the notion that they should have been included in the study without considering their marker alleles. The application of the approach differs in two settings. The first setting is in using association methods to search for evidence of linkage. The null hypothesis here is that the marker is not linked to any trait locus. In the second setting, linkage has been established in a region and association methods are being used for more precise gene mapping. In this case, the null hypothesis is that there is independence between alleles of the marker and the alleles of any trait locus that is linked to the marker. In each setting, the full minimal sufficient statistic differs. In the case of testing for linkage, the observed traits in all pedigree members and the marker alleles in the founders (member of the pedigree without parental information) are necessary. On the other hand, in testing for association in the presence of linkage established by classical linkage analysis, the full minimal sufficient statistic is the observed traits, the marker alleles in the founders and the *identity-by-descent* (IBD) relationships. The IBD relationships are the patterns of allele sharing due to descent. When two alleles at a certain locus are said to be identical by descent, it means that the alleles are identical copies of the same allele in some earlier generation. The difficulty in this type of method is that the determination of the IBD status may not be possible in some genetic markers and candidate genes (Waldman et al., 1999).

Family-Based Test of Association and Linkage (Lunetta et al., 2000)

From other previous extensions and modifications of the TDT, Lunetta and colleagues constructed a score statistic using likelihoods for the distribution of the phenotype, given the genotype. They evaluated the distribution of the test statistic by using the appropriate permutation distributions for the offspring allele values such as those described in the previous section by Rabinowitz and Laird (2000). The score is computed based on the offspring genotypes, conditional on parental genotypes and trait values for offspring and parents. The method extends the TDT to quantitative phenotypes and to multiple genes or environmental factors allowing also for interactions. To illustrate the method, assume that there are N independent families indexed by i, each having n_i offspring indexed by j=1,...n_i. The phenotype of the jth offspring in the ith family is denoted by Y_{ij} and μ_{ij}=E(Y_{ij}). Consider a biallelic marker whose genotype is coded in the variable X_{ij}. Given the susceptibility allele A_1, for the additive model, X_{ij} counts the number of A_1 alleles in the ijth individual. In a recessive model, X_{ij}=1 if the ijth individual has genotype A_1A_1 and is 0 otherwise. Using a generalized linear model, the method assumes a link function L_{ij}, which is a transformation of μ_{ij}, such that:

$$L_{ij} = \beta_0 + \beta_X X_{ij} \tag{3.2}$$

For dichotomous phenotypes (e.g. disease vs. no disease), the natural link function is the logit:

$$L_{ij} = logit(\mu_{ij}) = log[\mu_{ij}/(1-\mu_{ij})] = \beta_0 + \beta_X X_{ij} \tag{3.3}$$

where $\mu_{ij} = \mathrm{E}(Y_{ij})$, the disease prevalence.

For quantitative traits, the natural link is the identity, i.e. $L_{ij} = \mu_{ij}$, so that the association model is the linear regression model $L_{ij} = \mu_{ij} = \beta_0 + \beta_X X_{ij}$.

To obtain the score statistic, the prospective likelihood of phenotype Y_{ij} conditioning on the genotype X_{ij} is computed. The siblings are treated as independent, given the genotype. Then the adjustment for admixture is done by computing the mean and the variance of the score statistic using the distribution of genotype in offspring, conditional on parental genotypes and on offspring phenotypes. The log likelihood for the model is written as:

$$\log \mathrm{L}(\beta_0, \beta_X) = \sum_{ij} [Y_{ij} L_{ij} - a(L_{ij})] \tag{3.4}$$

where $a(L_{ij})$ is a function of L_{ij} with $\partial a(L_{ij})/\partial L_{ij} = \mu_{ij}$ when L_{ij} is the canonical link function. To test the null hypothesis of no association ($H_0 : \beta_X = 0$), the first derivative of the log likelihood with respect to β_X is computed: $(\partial \log L)/\partial \beta_X = \sum_{ij} X_{ij}(Y_{ij} - \mu_{ij})$. Then β_X is set to 0 in the resulting equation yielding the score statistic $S = \sum_{ij} X_{ij}(Y_{ij} - \mu)$, where under the null hypothesis, μ is constant for all subjects. The score statistic depends on the nuisance parameter μ which is not a function of the genotype, so that misspecification of μ will not bias the test. However, a good choice of μ can improve the test efficiency. The distribution of the test statistic is evaluated using permutation distributions for the offspring allele values based on the algorithm of Kaplan et al. (1997) for *nuclear families* (i.e. consisting of parents and children) and on the algorithm of Rabinowitz and Laird (2000) in the case of missing parental genotype. The test is unbiased even when the associated model or phenotype distribution is misspecified. It is also unbiased even when there is population admixture because the distribution of the test statistic is computed under the correct conditional distribution of the transmitted alleles. The method can be extended to include covariates and interactions. The covariates include environmental exposures and also genotype at a gene known to affect the trait, provided that the gene is not linked to the one being tested. It is also assumed that the covariates are not affected by any gene linked to the tested locus. The incorporation of covariates in the association model is not really necessary. However, including covariates in the analysis may increase efficiency if the covariates are strongly predictive of the phenotype. The extension of the association model considering a covariate E_{ij} and interaction between the locus and the covariate can be written as:

$$L_{ij} = \beta_0 + \beta_X X_{ij} + \beta_E E_{ij} + \beta_{XE} X_{ij} E_{ij} \tag{3.5}$$

To test for gene-environment interaction, we normally set $\beta_{XE} = 0$. However, the reference distribution of the test statistic under H_0 is always computed under the assumption of no linkage and no linkage disequilibrium which means that β_X is also 0. Therefore the testable null hypothesis is $H_0 : \beta_X = \beta_{XE} = 0$ and the test statistics obtained by differentiating the log likelihood with respect to the β parameters are $S_1 = \sum_{ij} X_{ij}(Y_{ij} - \mu_{ij})$ and $S_2 = \sum_{ij} X_{ij}E_{ij}(Y_{ij} - \mu_{ij})$, where under under the null hypothesis, μ_{ij} is given as the antilink of equation 3.2. The 2-df test is insensitive to the way the covariate is coded. However, rejection of the null hypothesis of no linkage and no linkage disequilibrium may not imply interaction. If the test has enough power, the test should reject the null hypothesis even if there is no interaction, as long as a main effect is detected.

For the test of epistasis, equation 3.5 can be used if the second locus is not linked to the first locus being tested. The null hypothesis in this case is $H_0 : \beta_X = \beta_E = \beta_{XE} = 0$ since the reference distribution of the test statistic under H_0 is computed under the assumption of no linkage and no linkage disequilibrium for either locus. The statistic is $S_3 = \sum_{ij} E_{ij}(Y_{ij} - \mu_{ij})$ where $\mu_{ij} = \mu$ is a constant given by the antilink of $L_{ij} = \beta_0$.

QTDT of Abecasis et al. (2000)

Several authors worked on a revised regression model. Abecasis et al. (2000) for instance, extended the method of Fulker et al. (1999) which partitioned the association effect into two variables quantifying between- and within-family information. The extended method can accommodate any number of offspring, with or without parental genotypes. The test statistic is calculated using the likelihood ratio test assuming a normal distribution for the trait. An empirical p-value based on the permutation of patterns of allelic transmission is computed to protect against possible deviations from normality or selection on the trait.

Consider a biallelic marker locus B with alleles B_1 and B_2. Let the frequencies of the alleles be denoted by q_1 and $q_2 = 1 - q_1$, respectively. The additive genetic value is denoted by a. When the marker locus is in linkage disequilibrium with the disease locus or is the disease locus itself, $a \neq 0$. In the absence of linkage disequilibrium, $a = 0$. Given a set of $i = 1, ..., N$ nuclear families, each with n_i children, the total number of offsprings is $\sum_i n_i$. We define the marker phenotype X_{ij} as equal to the number of B_1 alleles at the marker locus and the genotype score $G_{ij} = X_{ij} - 1$ for the jth offspring ($j = 1, ..., n_i$) in the ith family. If both parental genotypes are available, the genotype scores for the father and mother can be denoted as G_{if} and G_{im}, respectively. Assuming that the expected mean of the residual resemblance and the unique environmental effects are zero, the model for the quantitative phenotypic trait Y can be written as:

$$\mathrm{E}(Y_{ij}) = \mathrm{E}(\mu + G_{ij}a) = \mu + (q_1 - q_2)a \tag{3.6}$$

where μ is the overall mean. For the offspring in each family, the n_i x n_i variance-covariance matrix, $\mathbf{\Omega}_i$ has elements:

$$\mathbf{\Omega}_{ijk} = \begin{cases} \sigma_a^2 + \sigma_s^2 + \sigma_e^2 & if \quad j = k \\ \pi_{ijk}\sigma_a^2 + \sigma_s^2 & if \quad j \neq k \end{cases} \tag{3.7}$$

where:

π_{ijk}	denotes the proportion of alleles shared identical-by-descent between siblings j and k in family i
σ_a^2	the additive genetic variance of the major gene
σ_s^2	the residual sibling resemblance
σ_e^2	the residual environmental variance component

It should be noted that the above expectations do not include dominance variance. In this variance-components approach, all the information in a set of related individuals is used to construct a test of association by simultaneous modelling of the means and the variances. For a means model like:

$$\hat{Y}_{ij} = \mu + \beta_a G_{ij} \tag{3.8}$$

and for estimates of all the variances in $\mathbf{\Omega}_i$, the likelihood of the data for the complete set of parameters, $\vartheta = [\mu, \beta_a, \sigma_a^2, \sigma_s^2, \sigma_e^2]$ is

$$L = \prod_i (2\pi)^{-n_j/2} |\hat{\mathbf{\Omega}}_{ij}|^{-1/2} e^{-1/2[(Y_i - \hat{Y}_i)' \hat{\mathbf{\Omega}}_i^{-1} (Y_i - \hat{Y}_i)]} \tag{3.9}$$

The maximum likelihood test of association can be done by maximizing the likelihood of equation 3.9 under the null and alternative hypothesis. The null-hypothesis likelihood, L_0, is computed by setting the regression coefficient of the additive genetic effect $\beta_a = 0$ while the alternative-hypothesis likelihood L_1 is computed by maximizing the same equation with no constraints on the parameters. Then the likelihood ratio test is given by $2[ln(L_1) - ln(L_0)]$ which is χ^2 distributed with *df* equal to the difference in number of parameters estimated. In the absence of population admixture, this is a valid test of linkage disequilibrium because $\mathrm{E}(\beta_a) = a$, the additive genetic effect.

To account for population admixture, Abecasis and colleagues adopted the method of Fulker et al. (1999) of decomposing the genotype score G_{ij} into orthogonal between-family (b) and within-family (w) components. The b component is sensitive to population stratification but the w component is significant only in the presence of linkage disequilibrium. Thus the means model in equation 3.8 can be rewritten as:

$$\hat{Y}_{ij} = \mu + \beta_b b_i + \beta_w w_{ij} \tag{3.10}$$

where b_i and w_{ij} are the orthogonal between- and within- family components of G_{ij}. To accommodate any number of offspring with or without parental genotypes, b_i and w_{ij} are defined as follows:

$$b_i = \begin{cases} \dfrac{\sum\limits_i G_{ij}}{n_i} & \text{if parental genotypes are unknown} \\[2ex] \dfrac{G_{if} + G_{im}}{2} & \text{if parental genotypes are available} \end{cases} \tag{3.11}$$

$$w_{ij} = G_{ij} - b_i \tag{3.12}$$

Thus b_i is the expectation of each G_{ij} conditional on family data and w_{ij} is the deviation from this expectation for offspring j. An offspring who inherits more copies of the susceptibility allele than expected would have positive values of w_{ij}, whereas excess inheritance of the other allele would have negative values of w_{ij}. The regression coefficient, β_w is a direct estimate of the additive genetic value a, while β_b accounts for all other "spurious" association between the genotype score and the quantitative phenotype.

QTDT$_{\mathrm{M}}$ of Gauderman (2003)

The earlier QTDT methods focused on testing genetic main effects and most of these methods can be extended to accommodate gene-environment and gene-gene interactions. In 2003, Gauderman introduced the *Quantitative Transmission Disequilibrium Test with Mating Type Indicator* (QTDT$_{\mathrm{M}}$) for the analysis of candidate genes using parent-offspring trios. The focus of the method was not only on tests of genetic main effects, but also on gene-environment interaction and epistasis. To illustrate the QTDT$_{\mathrm{M}}$, let us consider an observed response variable Y_i which represents the quantitative trait and the variable G_i that quantifies the genotype at a candidate locus of N individual study subjects. The study subjects together with their parents are assumed to have been randomly selected from the population. The candidate locus is assumed to be biallelic with alleles denoted as A_1 and A_2. The susceptibility allele A_1 is further assumed to have a population frequency p_1 and genotypes formed by the alleles affect some biological process resulting to the quantitative trait Y_i. To determine whether variation in G_i is associated with variation in Y_i, the QTDT$_{\mathrm{M}}$ model is constructed based on a linear regression model with multiple parental mating type specific intercepts (denoted as α_M). The parental mating type which is the combination of the genotype of the mother and the father is treated as a fixed effect. The basic QTDT$_{\mathrm{M}}$ model for a single locus can be written as:

$$Y_i = \alpha_M + \beta_G G_i + \varepsilon_i \tag{3.13}$$

where:

Y_i	the observed random response or the quantitative phenotype of the ith study subject; i = 1,...,N
α_M	the parental mating type specific intercept; M = 1,...,6
G_i	a covariate that quantifies the genotype of the study subject
β_G	regression coefficient for the covariate G_i
ε_i	residual, $\sim N(0, \sigma^2)$

Note that unlike the method of Allison (1997), there are 6 parental mating types in the QTDT_M because even those mating types without a heterozygote parent are considered. The covariate G_i is assigned a value 0.0 if the genotype of the study subject (the offspring) is A_2A_2 (the wildtype or reference genotype). If the genotype is A_1A_1, it is assigned a value of 1.0. For the heterozygous genotype (i.e., A_1A_2), G_i takes on a value of 0.0, 0.5 or 1.0 if the assumed genetic model is recessive, additive or dominant, respectively. There are scoring methods that only count the number of susceptibility alleles while others (e.g. Abecasis et al., 2000) also consider parental genotype information and assign -1, 0 and 1 as possible scores. The genotype scoring adopted by Gauderman allows for assumption about the possible genetic model or mode of inheritance of the trait being investigated.

The null hypothesis of no association between Y_i and G_i (i.e., $H_0 : \beta_G = 0$) is tested using a likelihood ratio test (LRT). The LRT of β_G is based on the model in equation 3.13 applied to family trios from all six mating types. It has the form $2(L_1 - L_0)$, where L_0 and L_1 are the values of the maximized log-likelihood under the null and alternative hypotheses, respectively. The test statistic is chi-square distributed with 1 degree of freedom under a dominant, recessive or additive genetic model.

The QTDT_M can be extended to accommodate epistasis or gene-gene interaction. Let us denote the alleles in the second locus as B_1 and B_2. Considering only main gene effects and epistasis, the model can be written as:

$$Y_i = \alpha_M + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i + \varepsilon_i \tag{3.14}$$

where:

Y_i	the observed random response or quantitative phenotype of the ith study subject; i = 1,...,N
α_M	the parental mating type specific intercept; M = 1,...,36
G_i	covariate that quantifies the genotype of the subject at locus 1
H_i	covariate that quantifies the genotype of the subject at locus 2
β_G, β_H, β_{GH}	regression coefficients
ε_i	residual, $\sim N(0, \sigma^2)$

For two loci that are both biallelic, there are 36 possible parental mating types. The covariates G_i and H_i are coded similarly as the G_i covariate in equation 3.13. A likelihood ratio test is used to test for gene-gene interaction, comparing $L_1 = max\ log[L(\alpha_M, \beta_G, \beta_H, \beta_{GH}, \sigma)]$ to $L_0 = max\ log[L(\alpha_M, \beta_G, \beta_H, \beta_{GH} = 0, \sigma)]$. If one would consider gene-environment interaction, then the H_i representing the other locus can just be replaced with a covariate for the environmental effect.

Gauderman compared the models proposed by Lunetta et al. (2000), Fulker et al. (1999), Abecasis et al. (2000), and Liu et al. (2002) with the standard linear regression and his QTDT$_\mathrm{M}$ model in a simulated population. In his paper, Gauderman referred to the method of Fulker et al. and Abecasis et al. as HQTDT (Hierarchical QTDT), the method of Liu et al. as RQTDT (Retrospective QTDT; see the next section) and Lunetta's method as the regular QTDT. Only one population was assumed in the simulation to simplify the comparison and avoid bias due to stratification which is known to confound the effect of the standard linear regression model. The result showed that in determining genetic main effects, the simple standard linear regression model performed the best. This was of course expected since no racial or ethnic stratification was included in the data. However, Gauderman (2003) was also able to show that the standard linear model can lead to substantial bias in genetic effect estimates and high type I error rates in the presence of ethnic confounding. The HQTDT and the QTDT$_\mathrm{M}$ performed similarly or slightly better than the other QTDT methods. In terms of testing for gene-environment interaction and also gene-gene interaction, the QTDT$_\mathrm{M}$ was more efficient than the previous QTDT approaches. The use of the covariate for the locus main effect which captures both between- and within-mating information contribute to the estimation of the interaction effects. However, one drawback of the QTDT$_\mathrm{M}$ is when there is only one or very few individuals belonging to a specific mating type. The author specified that at least two trios of a given mating type are needed to contribute to the test for gene-gene interaction. Having only one family trio for a mating type will not contribute any information to the test for epistasis and the corresponding intercept for the specific mating type with only one family trio will explain all the trait variation for the offspring in that mating type. This might have a big implication in the power of the test to detect epistasis, especially in cases when the sample size is not big enough or if some mating types have low frequency in the dataset being analyzed.

3.4.2 Retrospective QTDTs

Another approach used in quantitative genetics is the *retrospective* approach, where the genotypes of the study subjects are modelled as a function of their phenotypes and the parental genotypes. In this case, the genotypes are the outcome variables while the quantitative traits and other covariates are the independent variables.

FBAT for Quantitative Traits (Lange et al., 2002)

One of these retrospective methods was introduced in a unified approach to family-based tests of association by Rabinowitz and Laird (2000) and (Laird et al., 2000). The method builds on the original TDT method by Spielman et al. (1993) and was generalized to accommodate quantitative traits by Lange et al. (2002). The unified approach to Family-Based Association Tests is termed FBAT. It uses a score-based test statistic to measure the association between the phenotype and the genotype. Its distribution is computed based on the offspring genotype which is treated as a random variable, conditional on the offspring phenotypes and the parental genotypes for each offspring. The approach is applicable to many scenarios such as multi-allelic marker, dichotomous or quantitative phenotypes, multiple offspring per family and missing parental information. It is adjusted for population admixture and allows additive, dominant and recessive genetic models in the analysis.

For simplicity of equations in describing the FBAT, let us consider one biallelic marker locus with alleles A_1 and A_2. Further assume that this marker locus is also the disease locus and the allele frequency of the susceptibility allele A_1 is denoted by p_1. There are N independent families and each ith family (i=1,...,N) has n_i offspring. The variable that translates the genotype of the jth offspring (j=1,...,n_i) in the ith family to a numeric value is denoted by G_{ij}, and the quantitative trait is denoted by

Y_{ij}. The concept of the minimal sufficient statistic by (Rabinowitz and Laird, 2000) was adopted by the FBAT method. The minimal sufficient statistics is denoted here as S and the null hypothesis of no association, $H_0 : \beta_w = 0$ is tested using a score statistic. The general FBAT is a score test based on the mean model (equation 3.10) and the phenotypic variance. The phenotypic variance is given for the ith family by $\text{Var}(\mathbf{Y}_i) = \mathbf{V}_i$, where $\mathbf{V}_i$ is an n_i x n_i variance matrix. With G_{ij} as the random variable of interest, the FBAT is implemented as follows:

1. Compute the normal score S for β_w based on the mean model and the phenotypic variance.

2. Then, setting $b_i = \text{E}(G_{ij})$ and $\beta_w = 0$, we have $S = \sum_{ij} S_{ij}$ where $S_{ij} = [G_{ij} - \text{E}(G_{ij})](t_{ij})$ and $t_{ij} = (z_{ij} - \tau_{ij})$. The z_{ij} and τ_{ij} are defined by:

$$\mathbf{z}_i = (z_{i1}, ..., z_{in_i}) = \mathbf{V}_i^- \mathbf{Y}_i \tag{3.15}$$

$$\boldsymbol{\tau}_i = (\tau_{i1}, ..., \tau_{in_i}) = \mathbf{V}_i^- \boldsymbol{\mu}_i \tag{3.16}$$

where $\boldsymbol{\mu}_i$ is an n_i-dimensional offset vector of offset values $(\mu_{i1}, ..., \mu_{in_i})$ which may depend on other predictor variables for the phenotype. For illustration, a simple structure of $\mathbf{V}_i$ is considered here, where $\mathbf{V}_i$ depends only on i through its dimension n_i. The diagonal elements $\sigma^2 = Var(Y_{ij})$ are all equal, and the off-diagonal elements $\sigma^2 r = \text{Cov}\,(Y_{ij}, Y_{ij'})$ are exchangeable.

3. The general quantitative FBAT is computed as:

$$FBAT = \frac{S^2}{\text{Var}\,(S)} \tag{3.17}$$

where

$$\text{Var}\,(S) = \sum_{ijj'} t_{ij} t_{ij'} \text{Cov}\,(G_{ij}, G_{ij'}) \tag{3.18}$$

The power of the general quantitative FBAT (Lange et al., 2002) has been compared to the QTDT of Abecasis et al. (2000) and the *pedigree disequilibrium test* (PDT) by Monks and Kaplan (2000). The PDT is a method similar to FBAT which can also be used on family data containing parent and offspring genotypes, with offspring genotypes only, or a combination of these types of families, with no size restrictions. They only differ in two aspects. First, in the computation of the phenotypic residuals:

PDT assumes the offset to be the phenotypic mean, μ, while FBAT permits any value for the offset and can use phenotypes that are adjusted for within-family correlation. Second difference is in the variance computation. The PDT estimates the variance of the marker scores on the basis of the empirical variance while the quantitative FBAT computes the variances on the basis of Mendelian transmissions. It is only when linkage is present under the null hypothesis that FBAT computes the variance in the same way as the PDT. In the absence of population structures and ascertainment bias, the QTDT of Abecasis, the PDT and the general FBAT show virtually the same power. When there is extreme ascertainment (e.g. only offsprings in the upper 10% are included), the score-based statistics (PDT and FBAT) perform better than the QTDT of Abecasis. With a good offset choice under this condition, the FBAT can perform better than the other two methods. The flexibility of choosing the offset to be adapted to the ascertainment condition of the study is an advantage of the FBAT but a bad choice of offset may result in lower power. As a rule of thumb, the observed sample mean is always a powerful offset choice when analyzing total population samples. However, in the case when only ”affected” offsprings (e.g. offsprings with phenotypes in the upper the upper 10% tail of the distribution) are ascertained, the quantitative FBAT becomes sensitive to the offset choice within the phenotypic range. For example, if the offset choices are close to the phenotypic mean, the power of FBAT is virtually 0. But for offset choices outside the phenotypic range (e.g. offset smaller than the minimum value for the ascertainment condition), the power of the quantitative FBAT is identical to the power of the dichotomous FBAT. Generally, an offset outside the ascertainment condition should be specified when using quantitative FBAT. This is discussed in detail in Lange et al. (2002).

The flexibility of the offset choice, the model-free phenotype and the flexibility of modelling or not modelling the phenotypic correlation within the family are the main advantages of the FBAT. However, the method is not designed for the analysis of quantitative traits considering gene-gene interactions. A software, also called FBAT, is available for implementing the method. In addition, an integrated software package called PBAT which contains tools for power and sample size calculation is also available. It also contain tools for the data analysis of univariate, multivariate and time-to-onset statistics for nuclear families as well as for extended pedigrees. PBAT can also include covariates and gene-covariate interactions in all computed FBAT-statistics. The FBAT and PBAT software packages are the two components of the FBAT-Toolkit which can be accessed freely at the http://www.biostat.harvard.edu/ fbat/default.html (Laird, 2007; Lange et al., 2004).

Unified Framework for TDT Analysis of Discrete and Continuous Traits (Liu et al., 2002)

The method of Liu and colleagues also computes a score statistic S using an estimate of the mean of the offsprings' quantitative trait Y_i. Their test statistic is based on a conditional score test that can be applied to both discrete and continuous traits with normal or non-normal distributions. The genetic and environmental effects are modelled using a generalized linear model.

The score statistic is approximated by:

$$\hat{S} = \hat{\mathbf{U}}^T Var(\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}} \tag{3.19}$$

where:

$$\hat{\mathbf{U}}_\mathbf{i} = \sum_{i=1}^{n}[Y_i - \hat{\mathrm{E}}(Y_i)]\mathbf{J}_\mathbf{i} \tag{3.20}$$

$$\mathbf{J}_\mathbf{i} = \mathbf{z}_\mathbf{i} - \mathrm{E}(\mathbf{z}_\mathbf{i}) \tag{3.21}$$

$$\mathbf{z}_\mathbf{i} = \frac{\partial f(G_i, \mathbf{a})}{\partial \mathbf{a}}|_{\mathbf{a}=0} \tag{3.22}$$

$\mathbf{a}$ is a vector of parameters modelling the effect of the offspring's genotype G_i, f is a function that specifies the genetic model, and

$$\mathrm{Var}(\hat{\mathbf{U}}) = \sum_{i=1}^{n}[Y_i - \hat{\mathrm{E}}(Y_i)]^2 Var(\mathbf{J}_\mathbf{i}) \tag{3.23}$$

In general, the method can be extended to analyze gene-gene interaction by incorporating the product of the loci effect in the $\mathbf{z}_i$ vector. The complete details can be found in the publication. Analysis of gene-environment interaction, multiple-sib families and extreme sampling can also be handled by the method. Liu and colleagues compared their retrospective approach to the TDT_{Q5} (Allison, 1997) and the logistic regression based extension of the TDT (Waldman et al., 1999). Waldman and colleagues extended the TDT for continuous and categorical traits using logistic regression. Their approach is also retrospective because it models the probability of transmission of disease alleles versus non-disease alleles from heterozygous parents to their children as a function of a set of continuous or categorical predictor variables that are the traits of interest and other covariates. The result of the Liu et al. (2002) simulation study showed that in samples of single-child families, their approach has higher power than the TDT_{Q5} and the logistic regression based extension of the TDT when an offset is included in the model. However, Gauderman (2003) showed that the prospective $\mathrm{QTDT}_{\mathrm{M}}$ has better power than this retrospective method of Liu et al. (2002).

Quantitative Polytomous Logistic Method (Kistner and Weinberg 2004, 2005)

The retrospective approach has been further explored by Kistner and Weinberg (2004, 2005). Their *quantitative polytomous logistic* (QPL) method is an extension of the log-linear model by Weinberg et al. (1998) which is based on maximum likelihood with stratification on parental mating type. QPL also conditions on parental mating type to control for possible population stratification. It treats the offspring's value of the quantitative trait as the independent variable while conditioning on the parents' genotypes. The probabilities of the child's genotype are modelled with the generalized logistic regression method, using the parental mating type and quantitative trait as predictors. Like the QTDT of Abecasis et al. (2000), it is a family-based approach, and not really a transmission-based approach like the original TDT. Like the FBAT, the method does not require a normally distributed phenotypic trait.

Consider a biallelic marker locus B with alleles B_1 and B_2, a quantitative trait value Y and a family trio genotype data. Let X_i be equal to 0,1 or 2, quantifying the child's genotype depending on the number of copies of the susceptibility allele (say allele B_1) that the child carries. Let X_{if} and X_{im} be the number of copies of susceptibility allele for the father and mother, respectively. Assuming mating symmetry in the population (i.e. the probability of $X_{if} = x_{if}$ and $X_{im} = x_{im}$ equals the probability that $X_{if} = x_{im}$ and $X_{im} = x_{if}$), the parental genotype pairs can fall into one of the six mating types similarly described by Gauderman (2003). The QPL method is based on multinomial distribution and the authors' basic idea is to reverse the true causality. Biologically, the offspring's genotype may influence the value of the phenotype or quantitative trait but the QPL models the offspring's genotype, conditioning on the trait Y and the parents' genotypes. The model considers Y as fixed and known and C, the child's genotype, is the only random variable. Using generalized logistic, the probabilities that the child's genotype X_i is equal to 0,1 or 2 are modelled with the parental mating type and Y as predictors. For example, given a parental mating type B_1B_2 x B_1B_2, then both X_{if} and X_{im} are equal to 1. The probability that an offspring with a quantitative trait value of Y has genotype (X_i) equals 0, 1 or 2 is modelled as:

$$P[X_i = 0|X_{if} = 1, X_{im} = 1, Y] = \frac{exp(\beta_0 Y + \alpha_{110})}{1 + exp(\beta_0 Y + \alpha_{110}) + exp(\beta_2 Y + \alpha_{112})} \tag{3.24}$$

$$P[X_i = 1|X_{if} = 1, X_{im} = 1, Y] = \frac{1}{1 + exp(\beta_0 Y + \alpha_{110}) + exp(\beta_2 Y + \alpha_{112})} \tag{3.25}$$

$$P[X_i = 2|X_{if} = 1, X_{im} = 1, Y] = \frac{exp(\beta_2 Y + \alpha_{112})}{1 + exp(\beta_0 Y + \alpha_{110}) + exp(\beta_2 Y + \alpha_{112})} \tag{3.26}$$

The parameters β_0 and β_2 are assumed to be the same across different mating types. β_0 accounts for the change in the quantitative trait if the child did not inherit a copy of the susceptibility allele relative to the quantitative trait for an offspring who inherited one copy. On the other hand, β_2 accounts for the change in the quantitative trait if the child inherited two copies of the susceptibility allele, again relative to the quantitative trait for offspring with one copy. The intercept parameters α_{110} and α_{112} depend on the parental mating type and the child's genotype. The intercept α_{110} refers to a family where the father and the mother have one susceptibility allele and the offspring has none, while the intercept α_{112} refers to a family where the father and the mother have one susceptibility allele and the offspring has two. The intercept parameters allow the model to account for non-Mendelianism and possibly different distributions of the quantitative trait across the parental mating types. Thus, even if Mendelian transmission is violated, the test remain valid. The null hypothesis of no association or no linkage between the marker and the QT means that the parameters β_0 and β_2 are both equal to zero.

The authors applied QPL in complete and incomplete trios and compared it with the methods proposed by Allison (1997), Abecasis et al. (2000), Monks and Kaplan (2000), Sinsheimer et al. (2000) and the FBAT of Laird et al. (2000). The QPL demonstrated good power and robustness under various scenarios of the genotype effect, distribution of the quantitative trait and population stratification. It has more power over the other methods when the true genotypic effect is recessive and when the variance of the quantitative trait differed across subpopulations. However, in situations where the genetic effect is dominant or additive, the QPL either had similar or slightly less power than the other methods (Kistner and Weinberg, 2004). A further strength of the method is in its application for missing parental genotype data through expectation-maximization (EM) approach which allows recovery of almost all lost power due to the missing information. Assumptions of Hardy-Weinberg equilibrium, random mating, or even Mendelian transmission is not necessary here because the marginal model for the parental mating types is an unconstrained multinomial. The method has been extended to allow for multiple offsprings, maternal effects and parent-of-origin effect (Kistner and Weinberg, 2005). However, application to gene-gene interaction was not explored by the authors.

Quantitative Conditioning on Parental Genotypes (Wheeler and Cordell, 2007)

The apparent advantage of the retrospective approach lies in the absence of explicit assumptions about the distribution of the quantitative trait of interest. Wheeler and Cordell (2007) proposed a method closely related to the retrospective approach of Kistner and Weinberg (2004, 2005). The method involves constructing a sample of cases and matched pseudocontrols from a sample of case-parent trios. The approach was derived from the case/pseudocontrol method of Cordell and Clayton (2002) and Cordell et al. (2004). The method also used ”conditioning on parental genotypes” (CPG) and generates pseudocontrols conditional on the mother's and the father's genotypes. Since it was applied to quantitative trait, the method has been called QCPG or ”quantitative conditioning on parental genotypes”.

Consider again a biallelic locus B with alleles B_1 and B_2. Let Y_i be the offspring's quantitative trait and let X_i, X_{if}, X_{im} be the genotypes quantifying the number of susceptibility allele (B_1) of the offspring, father and mother of family i, respectively. Assume also that the set of possible offspring genotypes for a mating type M is denoted by O_M and let $\sum_{X_i^*}$ be the summation over all possible offspring genotypes and $\sum_{X_i^* \in O'_M}$ the summation over all possible offspring genotypes that could have been transmitted to the offspring given the parental genotypes. The O'_M is a restricted subset of the set O_M for a given mating type. In contrast to the QPL method, the QCPG distinguishes between the two possible heterozygote offspring genotype (i.e. B_1B_2 and B_2B_1). The probability of the offspring genotype given the parental genotypes and the quantitative trait is given by:

$$
\begin{aligned}
P(X_i|X_{if}, X_{im}, Y_i) &= \frac{P(X_i, X_{if}, X_{im}, Y_i)}{P(X_{if}, X_{im}, Y_i)} \\
&= \frac{P(X_i, X_{if}, X_{im}, Y_i)}{\sum_{X_i^*} P(X_i^*, X_{if}, X_{im}, Y_i)} \\
&= \frac{P(Y_i|X_i, X_{if}, X_{im})P(X_i|X_{if}, X_{im})P(X_{if}, X_{im})}{\sum_{X_i^*} P(Y_i|X_i^*, X_{if}, X_{im})P(X_i^*|X_{if}, X_{im})P(X_{if}, X_{im})} \\
&= \frac{P(Y_i|X_i)P(X_i|X_{if}, X_{im})}{\sum_{X_i^* \in O'_M} P(Y_i|X_i^*)P(X_i^*|X_{if}, X_{im})} \qquad (3.27)
\end{aligned}
$$

The above likelihood can be calculated via conditional logistic regression. Wheeler and Cordell (2007) compared their QCPG approach with the QPL and the $\mathrm{QTDT_M}$. All the methods were extended to allow for the analysis of multilocus haplotypes, maternal genotype and parent-of-origin effects. However, gene-gene interaction was not considered in the investigation. Simulation results showed that with randomly ascertained families, with or without population stratification, the prospective $\mathrm{QTDT_M}$ approach is the most efficient, requiring smaller sample sizes to achieve convergence and asymptotic behavior. In addition, the $\mathrm{QTDT_M}$ was the only method among the three which suitably estimates the genetic effect under the alternative hypothesis with population stratification. Covariates are also easily incorporated in the $\mathrm{QTDT_M}$ and the parameter estimates can be easily interpreted as direct effect on the trait. With regard to nonnormally distributed traits, the investigation of power and Type I error showed that both $\mathrm{QTDT_M}$ and QCPG are suitable for the analysis of traits that slightly deviates from normality. However, neither of the methods was found suitable for the analysis of highly nonnormally distributed traits.

Some retrospective approaches presented in this section can also incorporate covariates and interaction in the analysis. However, it may entail more steps and subsequent analysis. Specifically, incorporating gene-gene interaction (if applicable) may not be easy. Compared to the prospective approaches, another draw back of retrospective approaches is that, they are somehow counter intuitive because the idea of true causality is reversed. In actual situations, the individual's genes may influence the value of the phenotype and not the other way around.

3.5 Synthesis of the literature

Current statistical methods in determining evidence for genetic effects are vast and varied. In the context of family-based designs, several methods have been devised extending and modifying the simple TDT introduced by Spielman et al. in 1993. Not all available TDT and TDT-like methods are detailed here. However, based on current literature, no single method can be considered most suitable in the analysis of epistasis or gene-gene interaction in nonnormally distributed quantitative phenotypes in parent-offspring trios. Table 3.5 compares selected statistical methods discussed in this thesis that are used in the analysis of quantitative traits in family-based studies.

Table 3.5: Comparison of selected methods used in the analysis of quantitative traits in family-based studies

Characteristics	$\mathrm{QTDT_M}$	FBAT	QCPG
Approach	prospective	retrospective	retrospective
Robustness to population stratification	yes	yes	no
Analysis of epistasis	yes	no	yes
Inclusion of covariates	yes	yes	yes
Applicable to nonnormal QT	*	yes	*

* The method is suitable when the quantitative trait only slightly deviates from normality

In most of the quantitative approaches, normality of the quantitative trait is assumed. In case of possible deviations from normality or selection on the trait, some programs implement permutation procedures on the genotypes to produce empirical p-values. Other methods have claimed to be robust in the deviations of the trait from the normality assumption, but unfortunately considered gene-gene interaction only in theory. In practice, the incorporation of covariates, gene-environment and gene-gene interaction may require more complicated steps and calculations.

Among the different methods enumerated, the $\mathrm{QTDT_M}$ has several merits in analyzing quantitative traits, genetic main effects and gene-gene interactions. Compared with the current available methods in genetic analysis for family-based designs, the prospective $\mathrm{QTDT_M}$ approach has been shown to have more advantages in terms of fast convergence and smaller sample size requirement. The method can also be easily extended to accommodate covariates, gene-environment and gene-gene interactions. However, the approach is most appropriate when the quantitative trait of interest is normally distributed. It is not suitable for the analysis of very nonnormally distributed traits (Wheeler and Cordell, 2007). It also requires enumeration of the mating types which could become sparse in frequency especially if multiallelic markers or multiple haplotypes will be used in the analysis. Needless to say, there is still a need for a TDT or TDT-like method that will address the issues concerning genetic main effects' analysis, gene-gene interaction and nonnormality of the data distribution in quantitative trait analysis. In addition, the effect of population stratification should not be taken for granted even if TDTs are known to be robust to it. In a recent study by Li et al. (2008), it has been shown that the presence of population admixture can influence the power of the TDT in different ways. The next chapter describes the proposed new method for analyzing quantitative traits in family-based genetic studies.

4 Generalized Quantitative Transmission Disequilibrium Test

The occurrence of nonnormally distributed traits in genetic studies poses difficulties in the statistical analysis of genetic main effects and epistasis in complex diseases. Many approaches exist in statistics to deal with nonnormally distributed data. To mention a few, one can transform the data, use nonparametric tests, permutation tests or implement bootstrap approaches. Each of these methods has its own advantages and disadvantages. The previous chapter covered several TDT and TDT-like methods to determine genetic effects in quantitative traits. However, it has been shown in the review of the literature that existing methods have not yet properly addressed the issues in genetic analysis of quantitative traits in family-based studies. This chapter introduces the *generalized quantitative transmission disequilibrium test* (GQTDT) - a statistical analysis method that combines the concept of the previously described QTDT_M regression method and the *generalized additive model for location, scale and shape* (GAMLSS) which is described later. The GQTDT draws its advantage from the flexibility of the GAMLSS method in statistical testing and modeling both categorical and quantitative variables. It can be used to detect both genetic main effects and epistasis in many types of distributions of the outcome variable.

4.1 Theoretical background

To describe the principle behind the generalized QTDT method, it is inevitable to review first some regression models in the context of genetic analysis.

Linear model

One of the conventional methods applied in the analysis of quantitative genetic traits is *linear regression.* As seen in some of the previously described methods in Chapter

3 (e.g. QTDT of Abecasis et al., QTDT$_{\mathrm{M}}$ of Gauderman), one can use a linear regression model to relate the response or outcome quantitative phenotype to the genotype to determine genetic effects.

Suppose that we have a continuous random variable Y which represents the quantitative phenotype, and a predictor variable G which represents the genotype values, and we wish to explain the variability in Y due to G. The simple regression model has the form:

$$Y_i = \beta_0 + \beta_G G_i + \varepsilon_i \tag{4.1}$$

where i=1,...n, the β_0 and β_G are the regression parameters and ε_i is the residual effect. The variable G_i is commonly defined as a score or ordinal variable based on the known genotype of the ith individual which has been randomly chosen from some population. It is also commonly assumed that the residual ε_i is normally distributed with mean zero and variance σ^2. The aim is to test the null hypothesis that there is no association between Y and G or that H$_0$:β_G=0. The hypothesis can be tested using a likelihood ratio test equal to $2(L_1 - L_0)$, where L_0 and L_1 are the values of the maximized log-likelihood under the null and alternative hypotheses, respectively.

The model 4.1 is easy to extend to accommodate more than one genetic factors and even other non-genetic or environmental factors. To account for possible confounding effects of population stratification, some regression methods (e.g. Abecasis and Gauderman's methods; see Chapter 3) incorporated parental genotypes in addition to the offspring's genotype in the regression model. The use of the parent-offspring design is to eliminate the problem of ethnic confounding effects. However, the linear regression method is not always applicable to all types of quantitative traits. Many complex traits in genetic studies are nonnormally distributed. This often implies that the residuals are also nonnormally distributed (Beasley et al., 2009). It is a common practice in statistics to transform the data to achieve an approximately normal distribution. Often, log transformation, shifted log (adding a constant before taking logs to retain zeros in the data) and inverse normal transformations (e.g. Blom transformation) are employed in the analysis. In particular, the use of rank-based inverse normal transformations (INTs) has become quite popular in recent genetic researches (Beasley et al., 2009). Rank-based INTs uses a modified rank variable and computes a new transformed value of the phenotype, Y_i^t, for the ith subject. Following the notation in Beasley et al. (2009), the transformed phenotype is computed as:

$$Y_i^t = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right) \tag{4.2}$$

where r_i is the ordinary rank of the phenotype of the ith subject among the N observations, Φ^{-1} the standard normal quantile (or probit) function and c, a constant. In Blom transformation, the recommended value of c is 3/8 (Blom, 1958). One issue concerning INTs is that the normality is assured for the wrong distribution. Most parametric tests assume that residuals from a model are normally distributed. In the case of INTs, the phenotypes and not the residuals are transformed to have a normal distribution. This is in contrast to some other transformations like the Box-Cox transformation which maximizes the normality of the sample residuals (Box and Cox, 1964). It has been shown in the study of Beasley et al. (2009) that INTs do not always maintain proper type I error and in some situations have also reduced statistical power.

In some cases, transformation of nonlinear data is not appropriate when comparison of arithmetic means is necessary (Barber and Thompson, 2000). In terms of detecting epistatic effect, several studies have shown many disadvantages of data transformation. Nonlinear transformation on the data or changing the scale can remove the interaction effect or artificially induce an interaction effect regardless of the underlying model (Thompson, 1991; An et al., 2009). Interaction and main effect relationships are usually not maintained after rank transformations (Blair et al., 1987). Poor performance has been noted in parametric tests for interaction applied to ranks due to lack of an *invariance property* which produces distorted type I and II error rates (Salter and Fawcett, 1993; Toothaker and Newman, 1994; Mansouri and Chang, 1995). In mathematics, *invariance property* is a property of mathematical objects, e.g. parameter spaces, that remains unchanged even after a given transformation. For example, maximum likelihood estimators (mle's) have invariance property. Given that $\hat{\Theta}$ is an mle of Θ and f is a certain function, the theorem on mle's invariance property states that $f(\hat{\Theta})$ is the mle of $f(\Theta)$. Interaction tests for transformed variables also performed poorly for a variety of other designs such as polynomial and response surface regression, analysis of covariance and repeated measures designs (Conover and Iman, 1981; Akritas, 1990; Thompson, 1991; Thompson, 1993; Headrick and Rotou, 2001; Headrick and Sawilowsky, 2000; Headrick and Vineyard, 2001; Beasley, 2002).

In addition, the use of rank-based transformations does not result in an adequate test for interaction (i.e. non-additivity) causing problems in evaluating epistasis and also gene-environment interactions (Hora and Conover, 1984).

Generalized Linear Model (GLM)

The linear model was extended by Nelder and Wedderburn (1972). They introduced the *generalized linear model* (GLM) to allow the response variable Y to be a member of any of the exponential family of distributions such as the binomial, Gaussian and Poisson distributions. Generalized linear models also relax the requirement of equality or constancy of variances that is required for hypothesis tests in traditional linear models. The exponential family of distributions take the general form:

$$f(Y|\Theta,\Psi) = exp\left[\frac{Y(\Theta)-b(\Theta)}{a(\Psi)} + c(Y,\Psi)\right] \tag{4.3}$$

The *canonical parameter*, Θ, represents the location, while the *dispersion parameter*, Ψ, represents the scale of the distribution. The functions a, b, and c can be specified depending on the type of exponential distribution being defined.

The GLM models the monotone link function of the response variable as a linear function of the covariates. Given a random variable, Y, and a set of genetic predictors $G_1, G_2, ..., G_p$ which may include genotypes at two unlinked loci and their interaction, the linear predictor can be expressed as:

$$\eta = g(\mathrm{E}(Y|G_1, G_2, ...G_p)) = \beta_0 + \sum_{k=1}^{p}\beta_k G_k \tag{4.4}$$

where β_0 is the overall mean and the link function, g, describes how the mean response, $\mathrm{E}(Y) = \mu$ is linked to the covariates through the linear predictor. In the Gaussian linear model, the link is the identity, i.e. $\eta = \mu$. For the Poisson GLM, the link is $\eta = log(\mu)$. Logistic regression is a binomial GLM using a logit link, $\eta = log(\mu/1-\mu)$. Many statistical models in genetics can be cast into a GLM form.

For estimation and testing, the parameters in a GLM can be estimated using maximum likelihood method. The maximum likelihood estimates of the parameters β_k can be obtained by iterative re-weighted least squares (IRLS). Analogous to the residual sum of squares in linear regression, the goodness-of-fit of a generalized linear model can be measured by the so called scaled deviance (Nelder and Wedderburn, 1972).

Generalized Additive Model (GAM)

In real life, many measured effects are generally not linear. Even the generalized linear models may be too restrictive to identify and characterize nonlinear regression effects. To transcend the limitations of traditional linear regressions, Hastie and Tibshirani (1990) introduced the methods of *generalized additive models* (GAM). Like the GLM, the GAM also assumes an exponential family distribution for the response variable. However, in the linear predictor, the linear term η is replaced by a more general functional form:

$$\beta_0 + \sum_{k=1}^{p} f_k(X_k) \tag{4.5}$$

where f_k, k=1,...,p, is a smooth function of a covariate. Examples of smooth functions can be local polynomial regression, kernel method or smoothing splines. The GAM uses a backfitting procedure in conjunction with a maximum likelihood or a maximum partial likelihood algorithm. Backfitting is an iterative procedure that estimates each f_k using previous estimates. The specification of the f_k functions makes GAMs much more flexible than linear regression models and GLMs. In genetics, the GAM has been applied in the analysis of microarray gene expression data (Tsai et al., 2004), in mapping cancer incidence rates (French, 2004), in genome-wide linkage and association studies (Rosenberger et al., 2005) and many more. However, it is also limited to the exponential family of distribution and usually allows only the modelling of the parameter mean of the distribution of the response variable as a function of the explanatory variables. The succeeding section describes a more general method that overcomes some of the limitations of both GLM and GAM.

Generalized Additive Model for Location, Scale and Shape (GAMLSS)

The GAMLSS is a general class of regression models introduced by Rigby and Stasinopoulos (2001, 2005) and Akantziliotou et al. (2002) for a univariate response variable. The method has more flexibility than the GLM or GAM. In the GAMLSS method, the exponential family assumption is replaced by a more general distribution family. The method can be very well adapted to genetic data where the response variables are highly skewed or kurtotic continuous. Real-valued response variables can be positively skewed (skewed to the right), a case where the tail on the right side of the distribution is longer than the left side. It can also be negatively skewed (skewed to the left). In this case, the tail on the left side of the distribution is longer than the right. Variables with a probability distribution exhibiting higher kurtosis may indicate that the variance is a result of extreme deviations. The systematic part of the GAMLSS model enables the mean and also other parameters of the conditional distribution of the random variable Y to be included in the statistical model as parametric and/or additive smooth nonparametric functions of explanatory variables and/or random effects terms. Adopting the authors' notation, the model is defined as follows:

Let Y_i, where i = 1,2,...,N, be independent observations with probability density function $f(Y_i|\boldsymbol{\theta}^i)$ conditional on $\boldsymbol{\theta}^i$ where $\boldsymbol{\theta}^i = (\theta_{i1}, \theta_{i2}, ..., \theta_{ip})$ is a vector of p parameters related to the explanatory variables. Let also $\mathbf{Y}^t$ be the N length vector of the response variables $Y_1, Y_2, ..., Y_N$. For k = 1,2,...,p, let $g_k(.)$ be a known monotonic link function relating the vector of distribution parameters ($\boldsymbol{\theta}_k$) to the explanatory variables and random effects through the additive model:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{\boldsymbol{J}_k} \mathbf{Z}_{jk}\gamma_{jk} \tag{4.6}$$

where
j = 1,2,...,$\boldsymbol{J}_k$ (a vector of length N)
$\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are vectors of length N, e.g. $\boldsymbol{\theta}_k^t = (\theta_{1k}, \theta_{2k}, ..., \theta_{Nk})$,
$\boldsymbol{\beta}_k^t = (\beta_{1k}, \beta_{2k}, ..., \beta_{\boldsymbol{J}'_k k})$ is a parameter vector of length $\boldsymbol{J}'_k$,
$\mathbf{X}_k$ is a known design matrix of order N x $\boldsymbol{J}'_k$,
$\mathbf{Z}_{jk}$ is a fixed known N x q_{jk} design matrix and
γ_{jk} is a q_{jk}-dimensional random variable.

Model 4.6 is referred to as the GAMLSS. The model is comprised of a parametric component $\mathbf{X}_k\boldsymbol{\beta}_k$ and additive components $\mathbf{Z}_{jk}\gamma_{jk}$. The parametric component can include linear and interaction terms for explanatory variables and factors, polynomials, fractional polynomials and piecewise polynomials for variables. Non-linear parameters can also be incorporated. The additive components can accommodate terms such as smoothing and random-effect terms as well as terms for time-series analysis. If $\boldsymbol{J}_k = 0$ for k = 1,2,...,p, then model 4.6 reduces to a fully parametric model:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k \tag{4.7}$$

If $\boldsymbol{Z}_{jk} = \boldsymbol{I}_N$, where $\boldsymbol{I}_N$ is an N x N identity matrix, and $\gamma_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for all combinations of j and k, then model 4.6 becomes:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{\mathbf{J}_k} h_{jk}(\mathbf{x}_{jk}) \tag{4.8}$$

where

$\mathbf{x}_{jk}$ for j=1,2,...,J_k and k = 1,2,...,p are explanatory vectors (assumed known) of length N,

h_{jk} is an unknown function of the explanatory variable X_{jk}, and

$\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ is the vector which evaluates the function h_{jk} at $\mathbf{x}_{jk}$.

Model 4.8 is a special case of model 4.6 and is called semi-parametric GAMLSS. Usually, the maximum number of distribution parameters is four (i.e. p = 4). In general, the vector $\boldsymbol{\theta}^i$ can have more than four distribution parameters. The current implementation of GAMLSS in the R software can accommodate up to four distribution parameters, i.e. μ, σ, ν and τ referring to the location, scale, skewness and kurtosis parameters, respectively. The skewness and kurtosis parameters are usually referred to as shape parameters. For distributions characterized by four parameters (i.e. $Y_i \sim f(Y_i|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$, the model is written:

$$\left.\begin{aligned}
g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{\boldsymbol{J}_1} \mathbf{Z}_{j1}\gamma_{j1}, \\
g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{\boldsymbol{J}_2} \mathbf{Z}_{j2}\gamma_{j2}, \\
g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{\boldsymbol{J}_3} \mathbf{Z}_{j3}\gamma_{j3}, \\
g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{\boldsymbol{J}_4} \mathbf{Z}_{j4}\gamma_{j4}.
\end{aligned}\right\} \tag{4.9}$$

One example cited by the authors on the flexibility of the GAMLSS is shown in a model used in their blood flow study (Rigby and Stasinopoulos, 2006):

$$Y_i \sim TF[(\mu = h_1(x), log(\sigma) = h_2(x), log(\nu) = 1)] \tag{4.10}$$

where the response variable Y_i has a t-distribution with parameters μ, σ and ν, each related to the explanatory variable x. They modelled the location parameter μ by a non-linear model $h_1(x) = x[1 + \beta_{11}exp(-\beta_{12}/x)]$. The log of the scale parameter σ is modelled using a quadratic polynomial model $h_2(x) = \beta_{20} + \beta_{21}x + \beta_{22}x^2$ and the log of the shape parameter ν was set to a constant 1. In the study, other distributions were also tried to determine the best fitted model for the data.

One of the main advantages of the GAMLSS method is its flexibility to define different types of distributions. Aside from the conventional families of distributions, specific distributions with more than two parameters can be defined. One example is the *Box-Cox normal* family for $Y > 0$ which was used by Cole and Green (1992). The distribution is a reparameterized distribution of Box and Cox (1964). The *Box-Cox* distribution by Cole and Green has three parameters (μ, σ, ν) and assumes that a transformed variable z obtained from Y has a standard normal distribution $N(0, 1)$

where

$$z = \begin{cases} \dfrac{1}{\sigma\nu}\left\{\left(\dfrac{Y}{\mu}\right)^{\nu} - 1\right\}, & \text{if } \nu \neq 0 \\ \\ \dfrac{1}{\sigma}\log\left(\dfrac{Y}{\mu}\right), & \text{if } \nu = 0 \end{cases} \tag{4.11}$$

Another distribution that has three parameters is the *power exponential distribution* for $-\infty < Y < \infty$ (Nelson, 1991). It is a modification of the Box and Tiao (1973) method which assumes that z has a gamma GA(1,ν) distribution where

$$z = \frac{\nu}{2}\left|\frac{Y-\mu}{\sigma\ c(\nu)}\right|^{\nu} \tag{4.12}$$

and

$$c(\nu) = \left\{2^{-2/\nu}\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}\right\}^{1/2} \tag{4.13}$$

Rigby and Stasinopoulos (2004) modified the *Box-Cox* distribution to a four-parameter (μ, σ, ν, τ) Box-Cox-t (BCT) distribution. It assumes that the z in equation 4.11 has standard t-distribution with τ degrees of freedom. In addition to the BCT, they also introduced the *Box-Cox power exponential distribution* (BCPE) for $Y > 0$. The distribution is defined by assuming that z in equation 4.11 also has a standard power exponential distribution with four parameters (μ, σ, ν, τ). The distribution can be used to model skewness combined with kurtosis in quantitative outcomes.

Investigators sometimes encounter data in which the continuous outcome variable has a lower bound. In medical data, zero is often the lower bound as in the case of coronary artery calcification (CAC) scores (Agatston et al., 1990). A sizeable fraction of sample observations with "true" zeros as values can give rise to extreme right skewness of the data and can create problems in conventional regression methods. In this respect, the data can be modelled using mixed discrete-continuous distributions: a continuous, right-skewed distribution mixed with a single probability mass at zero. The *zero-adjusted inverse Gaussian* (ZAIG) distribution by Heller et al. (2006) can be used. The model is defined as:

$$f(Y_i) = \begin{cases} 1 - \pi_i^* & \text{if } \ Y_i = 0 \\ \\ \pi_i^* \dfrac{1}{\sqrt{2\pi Y_i^3 \sigma_i}} \ exp\left[-\dfrac{1}{2Y_i} \left(\dfrac{Y_i - \mu_i}{\mu_i \sigma_i} \right)^2 \right] & \text{if } \ Y_i > 0 \end{cases} \tag{4.14}$$

where $\mathrm{E}(Y_i) = \pi_i^* \mu_i$, $\mathrm{Var}(Y_i) = \pi_i^* \mu_i^2 (1 - \pi_i^* + \mu_i \sigma_i^2)$ and π_i^* is the probability of a non-zero Y_i.

There are other types of distributions for continuous and discrete variables that can be implemented in the GAMLSS framework (see Rigby and Stasinopoulos, 2005; Stasinopoulos 2007). What have been previously enumerated are distributions that are commonly encountered in genetic studies. Table 4.1 lists some of the continuous distributions supported in R software while figure 4.1 shows graphs of selected distributions. Currently, there are more than 40 distributions supported by the GAMLSS package in R.

In the GAMLSS framework (Equation 4.6), the random-effects vectors γ_{jk} are assumed to have independent prior normal distributions with $\gamma_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-})$, where $\mathbf{G}_{jk}^{-}$ is the generalized inverse of a q_{jk} x q_{jk} symmetric matrix $\mathbf{G}_{jk}$. The matrix $\mathbf{G}_{jk}$ may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$ (e.g. degrees of freedom for smoothing terms and/or non-linear parameters). This implies that $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$ and if singular, then $\boldsymbol{\gamma}_{jk}$ is taken to have an improper prior density function proportional to exp(-$\frac{1}{2}\gamma_{jk}^T \boldsymbol{G}_{jk} \gamma_{jk}$). For fixed $\boldsymbol{\lambda}_{jk}$, the $\boldsymbol{\beta}_k$ and the λ_{jk} are estimated by maximizing the penalized likelihood function l_p:

$$l_p = l - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \gamma_{jk}^T \mathbf{G}_{jk} \gamma_{jk} \tag{4.15}$$

where $l = \sum_{i=1}^{N} \log \{f(Y_i|\boldsymbol{\theta}^i\} = \log \{f(\mathbf{Y}|\boldsymbol{\beta}, \gamma)\}$ is the log-likelihood function of the data given $\boldsymbol{\theta}^i$ for i=1,2,...,N.

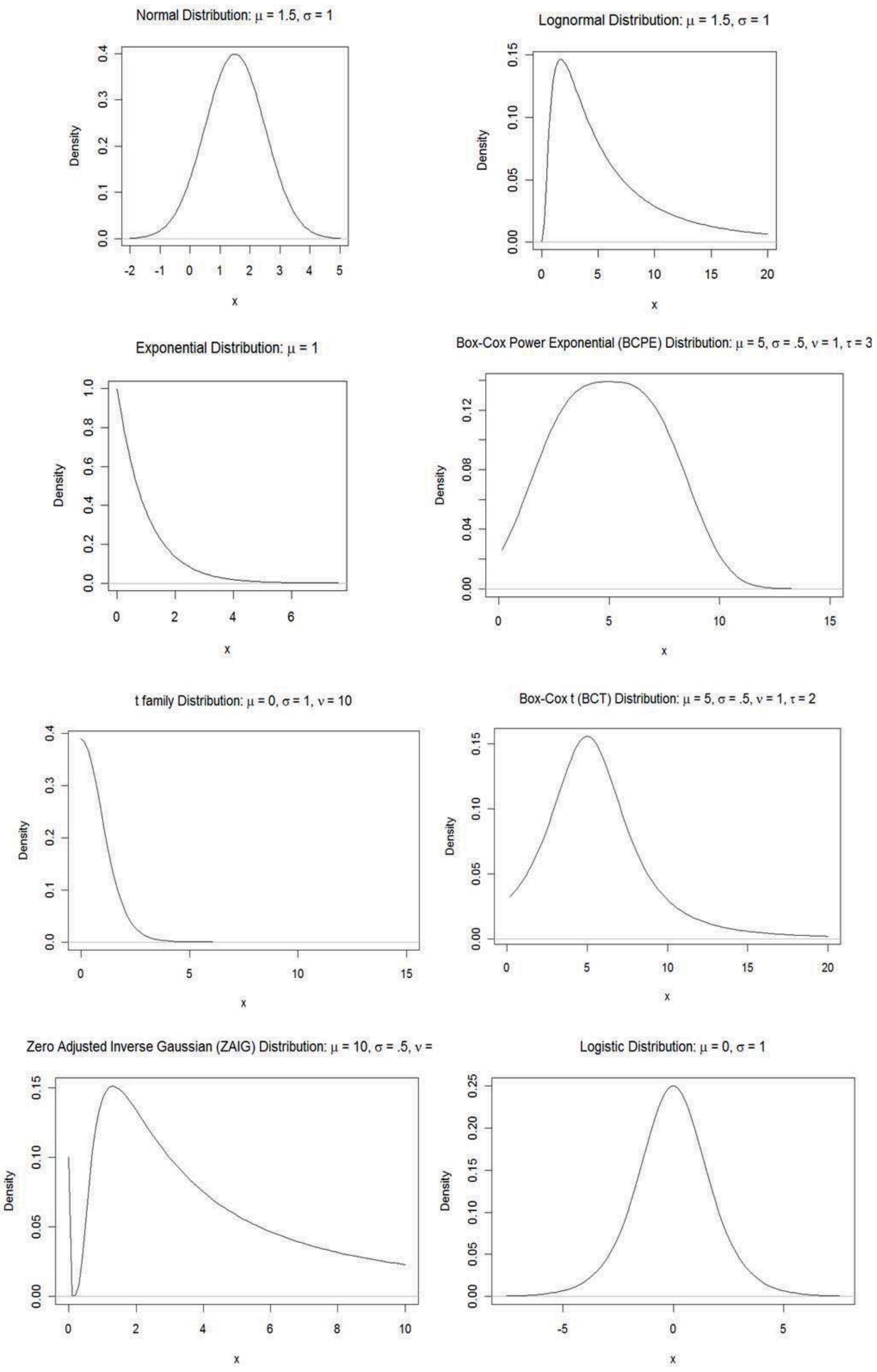
Maximizing l_p can be achieved by two basic algorithms - the CG and RS algorithms which are based on the Newton-Raphson or Fisher scoring algorithm. The CG or Cole and Green (1992) algorithm uses a backfitting algorithm and the first and (expected or approximated) second and cross derivatives of the likelihood function with respect to the distribution parameters $\boldsymbol{\theta}^i$. (A *cross derivative* of a smooth function is a mixed partial derivative obtained by differentiating at most once with respect to each

Table 4.1: Selected continuous distributions with default link functions in GAMLSS in R

Distribution	R Name	μ	σ	ν	τ
Box-Cox Cole and Green	BCCG()	identity	log	identity	-
Box-Cox power exponential	BCPE()	identity	log	identity	log
Box-Cox-t	BCT()	identity	log	identity	log
exponential	EXP()	log	-	-	-
inverse Gaussian	IG()	log	log	-	-
logistic	LO()	identity	log	-	-
log normal	LOGNO()	log	log	-	-
log normal (Box-Cox)	LNO()	log	log	fixed	-
normal	NO()	identity	log	-	-
power exponential	PE()	identity	log	log	-
skew power exponential type 1	SEP1()	identity	log	identity	log
skew power exponential type 2	SEP2()	identity	log	identity	log
skew power exponential type 3	SEP3()	identity	log	log	log
skew power exponential type 4	SEP4()	identity	log	log	log
skew t type 1	ST1()	identity	log	identity	log
skew t type 2	ST2()	identity	log	identity	log
skew t type 3	ST3()	identity	log	log	log
skew t type 4	ST4()	identity	log	log	log
skew t type 5	ST5()	identity	log	identity	log
t family	TF()	identity	log	log	-
Weibull	WEI()	log	log	-	-
zero adjusted inverse Gaussian	ZAIG()	log	log	logit	-

Note: μ, σ, ν and τ refer to the location, scale, skewness and kurtosis distribution parameters, respectively.

variable). On the other hand, the RS or Rigby and Stasinopoulos (1996) algorithm fits mean and dispersion additive models and does not use the cross derivatives. It is more appropriate for probability density functions with parameters $\boldsymbol{\theta}^i$ that are information orthogonal, i.e. the expected values of the cross derivatives of the likelihood function are zero. Examples of distributions with parameters that are information orthogonal are the normal, logistic, gamma, inverse Gaussian and negative binomial distributions. The details of the CG and RS algorithms are described in Rigby and Stasinopoulos (2005).



Normal Distribution: μ = 1.5, σ = 1

Lognormal Distribution: μ = 1.5, σ = 1

Exponential Distribution: μ = 1

Box-Cox Power Exponential (BCPE) Distribution: μ = 5, σ = .5, ν = 1, τ = 3

t family Distribution: μ = 0, σ = 1, ν = 10

Box-Cox t (BCT) Distribution: μ = 5, σ = .5, ν = 1, τ = 2

Zero Adjusted Inverse Gaussian (ZAIG) Distribution: μ = 10, σ = .5, ν =

Logistic Distribution: μ = 0, σ = 1

Figure 4.1: Example graphs of selected probability distributions

4.2 The GQTDT model

Using the framework of GAMLSS combined with the concept of the $\mathrm{QTDT_M}$ (see page 55), a *generalized quantitative transmission disequilibrium test* (GQTDT) model can be constructed to determine the main effects of genes and their epistasis in family-based genetic studies.

Let again Y_i, where $i = 1,2,...,N$, be random observations of a continuous quantitative trait. In this thesis, the investigation focuses on two candidate biallelic loci and their interaction as the main explanatory variables affecting the response variable Y_i. The control for the confounding effect of population stratification is also necessary in this case. It has been illustrated by Gauderman (2003) that using standard linear regression (see equation 4.1) can lead to bias in estimates of the genetic effect and alarmingly high type I errors (>50%) when there is stratification in the population. Therefore, the GQTDT, considering possible presence of stratification in the study population adopted the idea from previous studies of including an indicator of the parental mating types in the analysis to guard against spurious association resulting from stratification. Aside from Gauderman (2003), other investigators (Weinberg et al., 1998; Fulker et al., 1999; Li and Fan, 2000; Abecasis et al., 2000) have also applied the approach of using parental mating type indicators in analyzing family data. The GQTDT model for the analysis of two candidate loci and their interaction is expressed as:

$$Y_i = \beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i + \varepsilon_i \tag{4.16}$$

where M_i, G_i and H_i refer to the random explanatory variables corresponding to the categorical mating type of the parents and genotype scores of the subject at locus 1 and at locus 2, respectively. The betas are the unknown model parameters that must be estimated and ε_i is the error term. Compared to the $\mathrm{QTDT_M}$ which treats the mating type as fixed effect, equation 4.16 considers the mating type as a random variable and uses a general intercept rather than mating-type specific intercepts. Treating the mating type as random has the advantage of fewer parameters to estimate. Instead of estimating a parameter for each possible mating type, only one parameter for all the mating types needs to be estimated if mating type is considered as a random variable. One may be able to account for differences across mating types when it is used as a fixed variable. However, as noted in the description of the $\mathrm{QTDT_M}$ in the previous chapter, the use of mating type as fixed effect has its disadvantage. Very few subjects within a mating type contribute very few information or none at all in the test for epistasis. In practice, it is not uncommon to encounter very few subjects in certain mating types. Even if only two biallelic loci which can have up to 36 possible mating types are involved in the study, scarcity of subjects in certain mating types can be observed.

To complete the definition of the GQTDT model, it should be added that the response variable Y_i can be from a distribution other than the normal distribution. The distribution can be characterized by one or more parameters. In general, the response variable Y_i in the GQTDT model is distributed as:

$$Y_i \sim f(g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p) \tag{4.17}$$

where:
f is the distribution of Y_i,
$\theta_1, ..., \theta_p$ are the parameters of f,
$g_1, ..., g_p$ are the link functions and
$t_1, ..., t_p$ are the model formulae for the explanatory terms and/or random effects in the predictors.

For Y_i that fits a normal distribution, only two distribution parameters : μ and σ are needed. Therefore, $Y_i \sim N(\mu, \sigma)$. Considering equation 4.16 as the predictor of μ and the $\log(\sigma)$ as a constant 1, the distribution of Y_i involving two candidate loci and their interaction can be completely characterized by:

$$Y_i \sim N(\mu = (\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i), log(\sigma) = 1) \tag{4.18}$$

In some cases, a different distribution such as Box-Cox-t distribution might fit better the response variable Y_i. In this case, one can characterize the distribution using four parameters such that $Y_i \sim BCT(\mu, \sigma, \nu, \tau)$, where μ, σ, ν and τ are the BCT distribution parameters.

4.3 Model selection

A crucial detail in the analysis using GQTDT is the specification of the type of distribution (e.g. Gaussian, Box-Cox-t, t-family etc.). Plotting first the data would be very helpful to determine the most appropriate distribution. Deciding which distribution to use will require the criteria for model selection as suggested in the GAMLSS. Model selection in GAMLSS compares different competing models containing its different components. The components of the GQTDT model (see equation 4.17) include its distribution, parameters, link functions and predictor variables. Model selection in GAMLSS for nested parametric models also involves the use of the *deviance*. The usual *deviance* is defined as minus 2 times the log-likelihood of the reduced model compared to the full model. In GAMLSS, the deviance is termed *global deviance*, GD = $-2l(\hat{\boldsymbol{\theta}})$, where $l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \log f(Y_i|\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$. The global deviance is exactly minus twice the fitted log-likelihood function with all constant terms in the log-likelihood. The usual deviance in GLM is calculated as a deviation from the full model and the constant terms are not included in the fitted log-likelihood. Therefore, the usual deviance cannot be used to compare different distributions.

Two nested parametric GAMLSS models can be compared by the *generalized likelihood ratio test* statistic, GLRT = GD_0 - GD_1, where GD_0 and GD_1 are the global deviances of the models referring to the null and alternative hypothesis and with error degrees of freedom df_{e0} and df_{e1}, respectively. The GLRT has an asymptotic χ^2-distribution under the null model, with degrees of freedom $df = df_{e0}$ - df_{e1}. For each model, df_e = n - $\sum_{k=1}^{p} df_{\theta k}$, where $df_{\theta k}$ are the degrees of freedom in the predictor model for parameter θ_k for k=1,...,p.

Models that are not nested, including those with smoothing terms can be compared using the *generalized Akaike information criterion* (GAIC). To penalize overfitting, a fixed penalty "#" for each effective degree of freedom is added. The GAIC(#) = GD + #df, where df denotes the total effective degrees of freedom used in the model and GD is the global deviance. The model with the smallest GAIC(#) is recommended.

Other model selection strategies are also possible using GAMLSS. Rigby and Stasinopoulos (2008) emphasized that good model selection requires specific knowledge of the topic and that "the determination of the model adequacy should always be carried out with respect to the substantive questions of interest and not in isolation".

4.4 Data layout for analysis

In this work, the case of analyzing genetic main effects and interaction will be presented. The dependent quantitative variable is hypothesized to be affected only by two biallelic loci and their interaction. The family data needed for the GQTDT analysis are the quantitative trait of the subject and the genotypes of the subject and the subject's parents at the two loci of interest.

Using the same notations as in equation 4.16, the description and corresponding coding of the random dependent variable Y_i and the random explanatory variables, G_i, H_i and M_i are detailed below:

Y_i	the continuous quantitative trait of the ith study subject; i = 1,...,N
G_i	genotype score of the study subject at locus 1; coded as 0, 0.5 or 1.0
H_i	genotype score of the study subject at locus 2; coded like G_i
M_i	the parental mating type; categorical variable based on parents' genotype; coded as 1,2,...,36 as there are 36 possible combination of parental genotypes for two biallelic loci

The genotype data for each locus consist of pairs of alleles. The candidate locus 1 is assumed to be biallelic with possible alleles denoted as A_1 and A_2. This implies that the three possible genotypes at locus 1 are A_1A_1, A_1A_2 and A_2A_2. Candidate locus 2 is also assumed to be biallelic with alleles B_1 and B_2. Therefore, the three possible genotypes at locus 2 are B_1B_1, B_1B_2 and B_2B_2. It is further assumed that A_1 and B_1 are the susceptibility alleles. Like in the QTDT_M(Gauderman, 2003), the variable G_i is assigned a value 0.0 if the genotype of the study subject is A_2A_2. If the genotype is A_1A_1, it is assigned a value of 1.0. For the heterozygous genotype (i.e., A_1A_2), G_i takes on a value of 0.0, 0.5 or 1.0 if the assumed underlying genetic model during testing is recessive, additive or dominant, respectively. Similar coding is applied for the covariate H_i. The assumption about the underlying genetic model depends on prior biological knowledge or previous genetic data, but if data are unavailable, investigators can do the statistical testing under different genetic model assumptions.

The covariate M_i which represents the mating type of the parents is defined as a categorical variable based on the genotypes of both mother and father. If only one locus is considered, six different mating types are possible without considering the order of alleles (see table 4.2). Considering two biallelic loci, there are 36 possible mating types as shown in table 4.3.

Table 4.2: Mating types for one locus

	Genotype of the mother		
	A_1A_1	A_1A_2	A_2A_2
Genotype of the father			
A_1A_1	M_1	M_2	M_3
A_1A_2	M_2	M_4	M_5
A_2A_2	M_3	M_5	M_6

Table 4.3: Mating types considering two loci

	Locus 1 Mating Type					
	M_1	M_2	M_3	M_4	M_5	M_6
Locus 2 Mating Type						
M_1	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}
M_2	M_{21}	M_{22}	M_{23}	M_{24}	M_{25}	M_{26}
M_3	M_{31}	M_{32}	M_{33}	M_{34}	M_{35}	M_{36}
M_4	M_{41}	M_{42}	M_{43}	M_{44}	M_{45}	M_{46}
M_5	M_{51}	M_{52}	M_{53}	M_{54}	M_{55}	M_{56}
M_6	M_{61}	M_{62}	M_{63}	M_{64}	M_{65}	M_{66}

Sample data for the analysis of two loci and their epistasis are shown in table 4.4.

Table 4.4: Sample Data for GQTDT Analysis

Genotype at Locus 1			Genotypes at Locus 2			Variables for Analysis*			
Father	Mother	Child	Father	Mother	Child	Y_i	G_i	H_i	M_i
A_1A_1	A_1A_1	A_1A_1	B_1B_1	B_1B_1	B_1B_1	50.3	1.0	1.0	1
A_1A_1	A_1A_2	A_1A_1	B_1B_1	B_1B_2	B_1B_1	49.7	1.0	1.0	8
A_1A_1	A_2A_2	A_1A_2	B_1B_1	B_2B_2	B_1B_2	52.5	0.5	0.5	9
A_1A_2	A_1A_2	A_1A_2	B_1B_2	B_1B_2	B_1B_2	51.2	0.5	0.5	22
A_1A_2	A_2A_2	A_2A_2	B_1B_2	B_2B_2	B_1B_2	48.6	0.0	0.5	29
A_2A_2	A_2A_2	A_2A_2	B_2B_2	B_2B_2	B_2B_2	46.9	0.0	0.0	36

*G_i and H_i are coded assuming additive genetic model, e.g. G_i=1 if A_1A_1, G_i=0.5 if A_1A_2, G_i=0 if A_2A_2 (for more description about additive genetic model, see page 18); values for M_i are based on tables 4.2 and 4.3; $M_{11} = 1$, $M_{21} = 2$,...,$M_{12} = 7$, $M_{22} = 8$,...,$M_{66} = 36$.

4.5 Testing for genetic main effects and epistatic effects

In a two-locus setting, to test the null hypothesis of no genetic main effect, i.e. H_0: $\beta_G = \beta_H = 0$, the generalized likelihood ratio test can be applied to compare the global deviances of the model with genetic main effects ($\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i$) and the model without genetic main effects ($\beta_0 + \beta_M M_i$). For testing epistasis or gene-gene interaction, i.e. H_0: $\beta_{GH} = 0$, the model with epistasis ($\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i$) is compared to the model without epistasis ($\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i$). Usually, the model with more parameters will fit better (i.e. have a greater log-likelihood). Whether it fits significantly better can be determined by using the GLRT which means computing the probability or p-value of the obtained difference in the global deviances. This procedure of model selection using the GLRT is as previously discussed in page 82.

In the test for genetic main effects, a p-value less than or equal to a set alpha error (usually 0.05) rejects the null hypothesis of no genetic main effects and concludes the alternative. The same is true for the test for epistasis. Rejection of the null model without epistatic effect favors the conclusion of having statistical epistasis between the two loci tested. As in any other statistical test in genetic analysis, the results may not necessarily mean presence of biological effects. Although the use of candidade genes in the analysis give higher chance of getting results with biological meaning, the need for biological experiments to confirm the statistical results cannot be overemphasized.

The described generalized quantitative transmission disequilibrium test or GQTDT can of course be extended to determine the effects of other non-genetic and environmental factors.

5 Simulation Studies

The objective of these simulation studies is to investigate the power and type I error of the GQTDT in determining genetic main effects and epistasis associated to quantitative traits (QTs) in a family-based study design. Datasets of families based on realistic situations were simulated. Different simulation scenarios were created to investigate the performance of the statistical test in normally distributed and nonnormally distributed QTs. Other scenarios such as presence of population stratification, different allele frequencies, different genetic models and presence of other covariate effects were also simulated. Detection rates, i.e. the percentage of datasets where genetic main effects and epistasis were significant were noted in the different simulation schemes. The R software, Version 2.10.0 (R Development Core Team, 2009) was used for all simulations and numerical work.

5.1 The simulation scheme

5.1.1 The simulation model

The simulated datasets contain phenotype and genotype data from simulated family trios. A family trio includes one offspring and its parents. The phenotype data are the quantitative traits (observed random responses) of the offsprings (or study subjects), while the genotype data are genotypes of all offsprings and parents on two candidate loci. In reality, complex traits studied in genetics are not just affected by genes and their epistasis but other environmental covariates as well. However, the emphasis of the simulations is on determining the genetic main effects and epistasis and not on the effect of the environmental covariates on the quantitative trait. The environmental covariates were only included in some of the simulation schemes to create a complex quantitative trait and to determine in which way it influences the result of the test for genetic main effects and epistasis. The general model used to simulate the quantitative traits is:

$$Y_i = \beta_0 + \beta_G G_i + \beta_H H_i + \beta_I I_i + (\sum_j \beta_{E_j} E_{ji}) + \beta_\varepsilon \varepsilon_i \tag{5.1}$$

where:

Y_i	the observed random response or quantitative trait of the ith offspring; i = 1,...,N independent observations
β_0	the intercept
G_i	explanatory variable representing the effect of locus A
H_i	explanatory variable representing the effect of locus B
I_i	explanatory variable representing the epistatic effect of locus A and B
E_{ji}	explanatory variable representing the effect of the jth environmental covariate of offspring i
β_G, β_H, β_I, β_{E_j}, β_ε	regression coefficients
ε_i	residual effects, $\sim N(0, \sigma^2)$ or $\sim ln(0, \sigma^2)$

Each random explanatory variable in Equation 5.1 is a value that was independently simulated. The details of which are described in the succeeding sections. The random component ε_i which represents unmeasured environmental and genetic effects was drawn from either a normal or lognormal distribution. QTs that are approximately normally distributed or skewed to the right were simulated. Different types of QTs were created using modified models based on Equation 5.1. Figure 5.1 illustrates one type of QT that is influenced by two loci, age, gender, smoking habit and their interactions. Other datasets have less complex QT, i.e. are only influenced by a single locus and no interaction and other covariate effects. Still, other datasets were created without the influence of any locus at all.

The simulation of datasets was started by creating three different types of populations - PopA, PopB and PopMix. The population type is defined by the minor allele frequencies of the loci. The details of this are explained later. PopA and PopB are both homogenous population while PopMix is a 50:50 mixture of subjects from PopA and PopB. In each type of population, different datasets were created considering the distributional type of the QT and the covariates affecting it. In general, there are 30 different main simulation schemes corresponding to 30 different types of datasets.

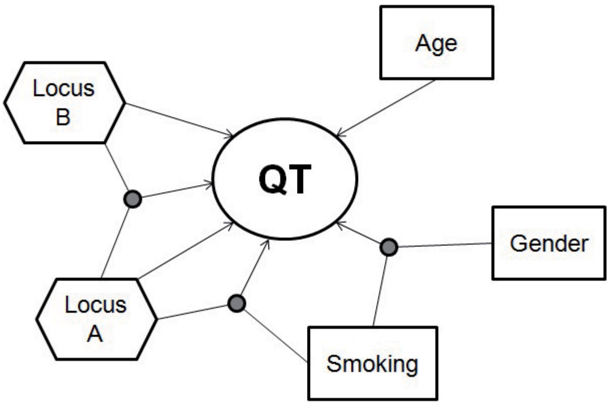


Figure 5.1: A quantitative trait affected by genetic and non-genetic variables and their interactions

Each type of dataset was created in 1000 replicates. An overview of the major criteria for creating the datasets is shown in table 5.1.

Table 5.1: Major criteria for creating the simulated datasets

Type of Population	Distribution of the QT response variable	Type of the QT according to the explanatory variables affecting it
PopA	Normal	QT$_N$, QT$_L$, QT$_{LL}$, QT$_{LLI}$ or QT$_{All}$
	Skewed	QT$_N$, QT$_L$, QT$_{LL}$, QT$_{LLI}$ or QT$_{All}$
PopB	Normal	QT$_N$, QT$_L$, QT$_{LL}$, QT$_{LLI}$ or QT$_{All}$
	Skewed	QT$_N$, QT$_L$, QT$_{LL}$, QT$_{LLI}$ or QT$_{All}$
PopMix	Normal	QT$_N$, QT$_L$, QT$_{LL}$, QT$_{LLI}$ or QT$_{All}$
	Skewed	QT$_N$, QT$_L$, QT$_{LL}$, QT$_{LLI}$ or QT$_{All}$

QT$_N$ refers to a quantitative trait without any locus or genetic effect. The modified equation 5.1 that was used to create the quantitative trait contains only non-genetic covariates. QT$_L$ is influenced by a single locus while QT$_{LL}$ is influenced by 2 loci. QT$_{LLI}$ is generated from two-locus main effects and their epistasis. QT$_{All}$ is similar to QT$_{LLI}$ with the addition of other covariate effects. The details of the simulation of each type of QT are also explained later. Other than the population type and the QT, additional covariates such as the type of genetic model and also the number of families in a dataset were also considered. Thus, increasing the types of datasets to more than 30 in some simulation schemes.

As mentioned before, the first step in the simulation is to create the different populations. Initially, genotypes of individuals are needed to create each population type. The parents' genotype data were first simulated. From these, the offspring's genotypes were generated. Afterwards, genetic main effects, epistatic effect and covariate effects were created. Then finally, from the different genetic and/or non-genetic variables, the quantitative trait Y_i was generated. The details of the simulations are as follows:

5.1.2 Creating the genotypes

The genotypes or pairs of alleles for the parents and the offsprings are the first explanatory variables created. Let us define two independent (unlinked) biallelic loci, A and B. Locus A has alleles A_1 and A_2 with population frequencies p_1 and p_2, respectively. Locus B has alleles B_1 and B_2 with population frequencies denoted by q_1 and q_2, respectively. Alleles A_1 and B_1 are marked as the susceptibility alleles and in this case also the minor allelles. The three possible genotypes (pairs of alleles) for Locus A are: A_1A_1, A_1A_2 and A_2A_2. On the other hand, the possible genotypes for Locus B are: B_1B_1, B_1B_2 and B_2B_2. Assuming Hardy-Weinberg equilibrium (see Chapter 2, page 14), the specific allele frequencies were used to calculate the frequencies of the three possible genotypes for each locus. Using this calculated distribution of genotypes, the genotypes of male and female individuals were randomly generated for each locus. Assuming that the population is randomly mating and that any parental pairs or mating types are possible, the male and female individuals were randomly paired to form the parents. Datasets with 1000 pairs and 2000 pairs of parents were created. Three types of populations based on the minor allele frequencies (MAF) were simulated. PopA was created using MAFs of p_1=0.1 and q_1=0.2 which refer to the minor allele frequencies of alleles A_1 and B_1, respectively. For PopB, the minor allele frequencies specified are p_1=0.3 and q_1=0.4. The third type of population (PopMix) contains a 50:50 mixture of the first two types to create a population with admixture. Table 5.2 shows a summary of the types of populations created.

After creating the pairs of parents, the mating type M_i was determined for each set of parents. Considering two biallelic loci, there are 36 possible mating types as shown in table 4.3 in page 85.

The genotype of the offspring is then created by randomly sampling from the possible genotype distribution given the parents' genotypes. The probability of the possible

Table 5.2: Types of simulated populations

Population	Locus A MAF (p_1)	Locus B MAF (q_1)	Number of parental pairs
PopA	0.10	0.20	1000
			2000
PopB	0.30	0.40	1000
			2000
PopMix	0.10 & 0.30	0.20 & 0.40	1000
			2000

MAF (Minor Allele Frequency); p_1 - frequency of allele A_1; q_1 - frequency of allele B_1

genotype of the child is defined by the usual Mendelian law of inheritance (see page 13). In locus A for example, if both parents are heterozygous (A_1A_2), the probability of a heterozygous child is 50%, while the probabilities of a homozygous (A_1A_1 or A_2A_2) child are both 25%.

5.1.3 Creating the genetic main effects (G_i and H_i)

The genetic main effect of each locus is a certain value contributed to the total value of the QT of interest. The contributed effect was created by drawing a value from a given distribution conditional on the assumed genetic model (i.e. additive, dominant and recessive; see also page 16) and the genotype of the offspring. For creating a dataset with normally distributed QT, the distribution where the genetic main effect was drawn was also specified as normal. For creating a dataset with skewed to the right QT distribution, the genetic main effect was drawn from a lognormal distribution. Table 5.3 specifies the parameters used in creating the distributions where the locus main effects were drawn. The values of the parameter means are assumed according to the magnitude of effect differences desired to be seen conditional on the genetic model. The standard deviations are fixed to 1. The parameter means used for the lognormal distribution are also based on the values used for the normal distribution which are either zero or the log scale equivalent of the non-zero mean used in the normal distribution. This is to achieve a relatively similar magnitude of effect. Under the dominant genetic model assumption, the parameter mean used to simulate the genetic effect is higher with the presence of at least one susceptibility allele (i.e. A_1 or B_1) in the genotype. In the additive genetic model, the parameter mean increases with the number of the susceptibility allele while in the recessive genetic model, both alleles in the genotype should be susceptibility alleles to exhibit a genetic effect. Locus A's main effect, G_i, was simulated either under a dominant or recessive genetic model while locus B's main effect, H_i, was simulated only under an additive genetic model assumption. For example (see table 5.3), under an assumed dominant genetic model, the genetic main effect contribution of the genotype A_1A_1 to a normally distributed QT is randomly drawn from a distribution $\sim N(\mu = 1.5\ ,\ \sigma = 1)$. Ideally, one can create both locus A and locus B under different genetic models. However, just one example of each type of genetic model would suffice for simulation purposes. Also for simplicity, the genetic main effect of locus A under the recessive genetic model was generated only from a normal distribution. So in the case of the recessive genetic model, both normal and skewed QTs have locus A genetic main effect taken from a normal distribution. It is not only the genetic main effects that controls the distribution of the resulting QT. Other variables described later also contribute to the QT value. The generation of the final QT was based on testing several values of the different parameters in the generating model to achieve a QT distribution similar to what is encountered in practice. The distribution of the QT in all simulation schemes was checked so that it satisfies the desired normal or skewed to the right characteristic.

Table 5.3: Simulation scheme used in creating the genetic main effects

Assumed Genetic Model	Genotype of offspring	Parameters* used in simulating the genetic effects
Dominant (D)	A_1A_1	$\mu = 1.5\ ,\quad \sigma = 1$
	A_1A_2	$\mu = 1.5\ ,\quad \sigma = 1$
	A_2A_2	$\mu = -2.5\ ,\quad \sigma = 1$
Additive (A)	B_1B_1	$\mu = 2.5\ ,\quad \sigma = 1$
	B_1B_2	$\mu = 0\ ,\quad \sigma = 1$
	B_2B_2	$\mu = -2.5\ ,\quad \sigma = 1$
Recessive (R)	A_1A_1	$\mu = 2.5\ ,\quad \sigma = 1$
	A_1A_2	$\mu = 0\ ,\quad \sigma = 1$
	A_2A_2	$\mu = 0\ ,\quad \sigma = 1$

*For creating the genetic main effects contributing to a normally distributed QT. Example graph for the distribution with $\mu = 1.5$ and $\sigma = 1$ is shown on page 80.

5.1.4 Creating the epistatic effects (I_i)

Epistasis or gene-gene interaction effect (I_i) was simulated considering genetic models and not simply multiplying the two genetic main effects. The locus A was defined to have either dominant or recessive main effect while locus B was defined to have additive main effect. The epistatic effect was simulated either from a normal of lognormal distribution conditional on the genotype at both loci. Table 5.4 shows the parameter means used in simulating epistatic effects for a normally distributed QT. For skewed QT, a lognormal distribution was used. The parameter means for the lognormal distribution are either zero or the log scale equivalent of the non-zero mean shown in table 5.4). The standard deviation used in the simulations is 1. Based on the table below, an individual subject with A_1A_1 genotype at locus A and B_1B_2 genotype at locus B will have an epistatic effect drawn from a distribution that is $\sim N(\mu = 0, \sigma = 1)$.

Table 5.4: Parameter means used in simulating normal distributions for the epistatic effects of two loci with individual genetic main effects

	Genotype at Locus B		
	B_1B_1	B_1B_2	B_2B_2
Genotype at Locus A			
A_1A_1	-2.5	0	2.5
A_1A_2	1.5	0	-1.5
A_2A_2	-2.5	0	2.5

From the values shown in table 5.4, one can say that the genotype in locus B can invert the effect of the genotype in locus A. On the other hand, locus A also influences the effect of locus B. The interaction pattern shown in the table is based on a simulation study of Kraja et al. (2009). This is just one type of ”biological” epistasis or interaction between two biallelic loci. In the study of Hallgrímsdóttir and Yuster (2008), 387 distinct types of epistatic patterns have been described to characterize a two-locus model.

5.1.5 Creating the covariates (E_{ji})

Complex traits or diseases are usually affected by other environmental covariates in addition to genetic effects and gene-gene interactions. Several non-genetic variables of the offspring (i.e. age, gender and smoking status) were also created. These variables were used in these simulation studies to create quantitative traits without genetic effects and also quantitative traits influenced by both genetic and non-genetic variables.

The age of the offspring was randomly sampled from a normal distribution with mean age 46 years ± 5.0 years. The age distribution was truncated so that the youngest age will not be below 18 years old and the oldest is 51 years old. Age 18 was used as the lower age limit because most studies among adults are conducted from age 18 and above. The maximum of 51 years old was chosen after a genetic study on lung cancer. Many genetic traits or diseases are usually expressed at a younger age. It is therefore common for many genetic studies to put a younger maximum age limit. In addition, recruiting study participants belonging to older age groups can increase the probability of the trait or disease being investigated to be influenced by environmental covariates rather than genetic factors. For the gender, equal proportions of male and female offsprings were created in the dataset. This was done by randomly sampling from a binomial distribution using a probability of 0.50. The smoking status of the offspring was also created using a binomial distribution with a probability of 0.3 for smokers. This is to create datasets with roughly 30% smokers and 70% non-smokers. The choice of the smoking proportion was based on the key figures of adult daily smoking prevalence in Europe which are between 20% to 44%. In Germany, current tobacco smoking prevalence among males is 32% and 22% among females (WHO, 2009b).

Thus, three covariates affecting the quantitative trait were created, i.e. age effect, gender-smoking effect and locus-smoking effect, the latter constituting a gene-environment interaction. To create the *age effect*, a value was randomly sampled from a normal distribution with standard deviation of 1 and a mean which is dependent on the age group of the offspring. There are six normal distributions of age effects corresponding to six age groups. Older offsprings have higher mean age effect contributing to the quantitative trait. The specific parameter means used in the simulation of the age effects are shown in table 5.5.

Gender-smoking effect was created by sampling from a normal distribution conditional on both gender and smoking status of the offspring. The variable was simulated

Table 5.5: Parameter means for simulating the age effect

Age Group in years	Mean age effect
18-30	1
31-35	2
36-40	3
41-45	4
46-50	5
= 51	6

in such a way that smoking males have higher mean contributing value to the response variable while in females the value is lower and does not vary between smokers and non-smokers. See table 5.6 for the parameter means used in the simulation of the gender-smoking effect. The standard deviation used in the simulation is also 1.

Table 5.6: Parameter means for simulating the gender-smoking effect

Gender	Smoker	Mean gender-smoking effect
Male	No	2
	Yes	4
Female	No	1
	Yes	1

The last covariate, the *locus-smoking effect*, was also drawn from a normal distribution conditional on the genotype at locus A and the smoking status. Offsprings carrying more susceptibility alleles at locus A and at the same time smokers have the highest mean contribution to the QT value compared to the other subgroups (see table 5.7). Similar to the first two covariates, the standard deviation used in the simulation of the locus-smoking effect is also 1. The first two covariates are considered non-genetic covariates.

5.1.6 Creating the other environmental effects or residuals (ε_i)

The residuals were simulated from a normal distribution or a lognormal distribution with mean zero and standard deviation of 1.

Table 5.7: Parameter means for simulating the locus-smoking effect

Genotype at Locus A	Smoker	Mean locus-smoking effect
A_1A_1	No	1
	Yes	3
A_1A_2	No	1
	Yes	2
A_2A_2	No	1
	Yes	1

5.1.7 Creating the quantitative traits (Y_i)

The final quantitative traits, Y_i, are either approximately normally-distributed or skewed to the right. These distributions are commonly encountered in genetics and medical data. Different QTs were created according to the following scenarios:

1. QTs without locus effect (QT_N)

The first quantitative traits simulated are those without genetic effects. In this simulation scheme, only non-genetic covariates (i.e. age and gender-smoking effects) and the residual were used to create the QT of the offsprings belonging to different types of population as shown in table 5.8. Population PopA has lower minor allele frequencies than PopB. PopMix consists of two populations (i.e. mixture of PopA and PopB). Normally distributed and skewed traits were created in all simulation schemes. In addition, all simulations were done in sample sizes of 1000 or 2000 family trios.

Equation 5.1 was used to generate Y_i using previously calculated non-genetic covariate effects and the following regression parameters: $\beta_0 = 50$, $\beta_{E_{age}} = 0.1$, $\beta_{E_{gender-smoke}}$=0.01, and $\beta_\varepsilon = 0.89$. The other regression parameters were set to 0. Therefore,

$$Y_i = 50 + 0.10E_{age} + 0.01E_{gender-smoke} + 0.89\varepsilon_i \tag{5.2}$$

As previously explained, the residual or random error ε_i is $\sim N(0,1)$ for QTs intended to be approximately normal in distributional shape. For skewed QTs, the ε_i is $\sim ln(0,1)$. Graphs of the distributions of the created QTs are shown in Figure 5.2. The mean of the QT for normally distributed data in the figure is 50.5 ± 0.9. For the

Table 5.8: No-locus effect simulation scenarios

Population	Locus A MAF (p_1)	Locus B MAF (q_1)	Distribution of the QT	Sample size (n)
PopA	0.10	0.20	normal	1000
				2000
			skewed	1000
				2000
PopB	0.30	0.40	normal	1000
				2000
			skewed	1000
				2000
PopMix	0.10 & 0.30	0.20 & 0.40	normal	1000
				2000
			skewed	1000
				2000

MAF (Minor Allele Frequency); p_1 - frequency of allele A_1; q_1 - frequency of allele A_1

skewed data, the mean of the QT is 51.9 ± 1.8. The graphs of other QTs created for the other scenarios are similar in shape as the ones illustrated in Figure 5.2. For this simulation scheme of no genetic effects, a total of 12 datasets were created. This is considering the 3 types of populations, the 2 types of QTs (normal and skewed) and the sample sizes (N=1000 and N=2000).

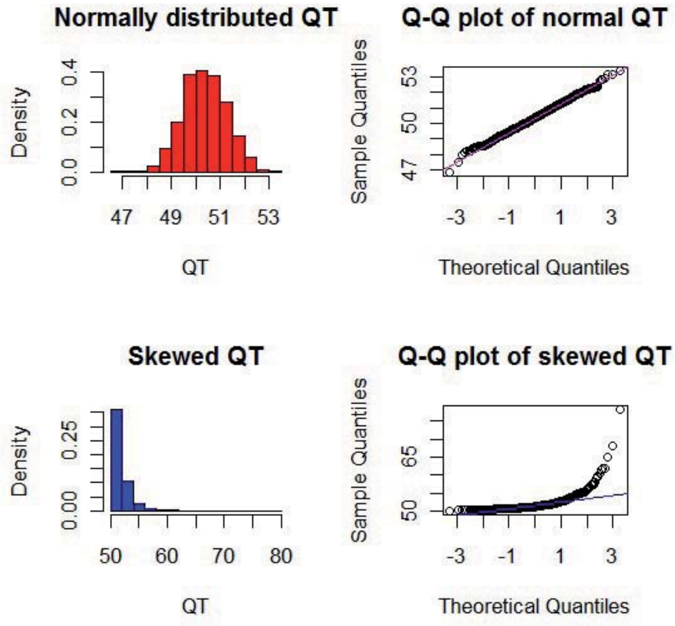


Figure 5.2: Distribution of simulated quantitative traits with no genetic effects (N=1000)

2. QTs with single-locus effect (QT_L)

Some datasets were created with QTs affected only by a single locus and a random error. Table 5.9 shows the different scenarios for this simulation. Again, there are three types of populations simulated: two homogeneous (PopA and PopB) and one heterogenous (PopMix). Either locus A or locus B was used in different simulation set-ups. The generating model was created so that when locus A is used, the genetic main effect size is 0.20. The model to generate the response variable Y_i is:

$$Y_i = 50 + 0.20G_i + 0.80\varepsilon_i \tag{5.3}$$

where G_i and ε_i are as previously defined. The normal QTs are produced by specifying a normal distribution in the independent simulation of the genetic main effect and the residual, while the skewed QTs are produced by specifying a lognormal distribution for both genetic main effect and residual. For generating the response variable when locus B is used, the model is as follows:

$$Y_i = 50 + 0.05H_i + 0.95\varepsilon_i \tag{5.4}$$

where H_i in this case is the genetic main effect of locus B which was previously defined.

The main difference of locus A and locus B are their minor allele frequencies. In addition, locus A has been used in simulating genetic main effects under a dominant genetic model and in some cases under a recessive genetic model. Locus B on the other hand was only used in simulating genetic main effects under the additive genetic model. While there are two possible genetic models for locus A, locus B only has one. It is possible to create exactly similar characteristics of locus A and B. However, it would be more worthwhile to create different scenarios for testing the statistical method. Table 5.9 shows only 18 types of dataset but considering the 2 sample sizes (i.e. N = 1000 and N = 2000), a total of 36 types of datasets were created for this simulation scheme.

Table 5.9: Single-locus effect simulation scenarios

	Affecting	Genetic Model of	Distribution
Population	Locus	Affecting Locus	of the QT
PopA	A	Dominant	normal or skewed
	B	Additive	normal or skewed
	A	Recessive	normal or skewed
PopB	A	Dominant	normal or skewed
	B	Additive	normal or skewed
	A	Recessive	normal or skewed
PopMix	A	Dominant	normal or skewed
	B	Additive	normal or skewed
	A	Recessive	normal or skewed

3. QTs with two-locus effect (QT_{LL})

Datasets with QTs affected independently by two loci were created. No other covariate effects were included other than the individual genetic main effects of the two loci. For locus A, the genetic model was set to either dominant or recessive, while for locus B, the genetic model was fixed to additive genetic model (see table 5.10). This was done so as not to complicate too much the simulation scenarios. The β_G for locus A

was set to 0.20 and for locus B, β_H was set to 0.05. The other effect on the trait is attributed to random error ($\beta_\varepsilon = 0.75$). The equation used for simulating the Y_i is:

$$Y_i = 50 + 0.20G_i + 0.05H_i + 0.75\varepsilon_i \tag{5.5}$$

Similar to the creation of QTs with single locus effect, QTs which are normally distributed were simulated by using genetic main effects and residuals sampled from normal distributions. For skewed QTs, affecting variables were randomly drawn from lognormal distributions as detailed in previous sections. In this simulation scenario, there are 24 types of datasets including datasets with different sample sizes.

Table 5.10: Two-locus effect simulation scenarios

Population	Genetic Model of Locus A	Genetic Model of Locus B	Distribution of the QT
PopA	Dominant	Additive	normal or skewed
	Recessive	Additive	normal or skewed
PopB	Dominant	Additive	normal or skewed
	Recessive	Additive	normal or skewed
PopMix	Dominant	Additive	normal or skewed
	Recessive	Additive	normal or skewed

4. QTs with two-locus effect and epistasis (QT_{LLI})

This simulation scenario is similar to that of table 5.10 where in the individual genetic main effects of the two loci contribute to the QT variability. In this case, epistatic effect additionally contributes to the variability of the QTs. The QTs were created using modified Equation 5.1. The generating equation for Y_i is shown below:

$$Y_i = 50 + 0.20G_i + 0.05H_i + 0.05I_i + 0.70\varepsilon_i \tag{5.6}$$

The I_i accounts for the "biological" epistasis or interaction of locus A and B. In contrast to the usual definition of interaction in statistics, the simulation of I_i (see page 93) is somewhat "biologically" inspired and therefore not just derived from simple multiplication of individual effects. The value of I_i depends on the genotype of the offspring on both loci. In addition to the individual locus effect, the epistasis between the two loci also contributes to the total QT value in this simulation scenario. In this case, there are also 24 types of datasets simulated.

There may be cases when the genetic main effects are too small to detect and only the epistatic effect is evident in the data. Therefore, in addition to the above simulation, another simulation scheme was done where the epistatic effect is much bigger than the individual genetic main effects. The model used for simulating the quantitative trait is as follows:

$$Y_i = 50 + 0.01G_i + 0.05H_i + 0.30I_i + 0.64\varepsilon_i \tag{5.7}$$

where as previous, G_i and H_i are the genetic main effects of locus A and B, respectively, I_i the variable accounting for epistatic effect and ε_i is the residual.

5. QTs with two-locus effect, epistasis and covariates (QT_{All})

Datasets with QTs that include other covariate effects were also created. The covariates age, gender-smoking and locus-smoking effects are included having regression coefficients 0.1, 0.01 and 0.01 respectively. The generating equation for Y_i can be written as:

$$\begin{aligned} Y_i = {} & 50 + 0.20G_i + 0.05H_i + 0.05I_i + 0.10E_{age} + 0.01E_{gender-smoke} + \\ & 0.01E_{locus-smoking} + 0.58\varepsilon_i \end{aligned} \tag{5.8}$$

Like in the previous simulation scenarios, normally distributed QTs have G_i, H_i, I_i and ε_i drawn from normal distributions. For skewed QTs, they were drawn from a lognormal distribution. There are also 24 types of dataset created for this simulation scenario.

5.2 Statistical tests

The statistical tests were done considering a situation where there are two unlinked, biallelic loci of interest. As mentioned in the introduction, the objective of the simulation studies is to determine the power and type I error of the GQTDT method in determining if two candidate genes show genetic main effects and an epistatic effect. Therefore, the statistical method was used to test both the hypotheses of ”no genetic main effects” and ”no epistatic effect”. The power and type I error of the GQTDT were compared to a modified QTDT$_\mathrm{M}$. The modified QTDT$_\mathrm{M}$ is referred here as QTDT$_\mathrm{M}$* and includes a random variable for the mating type indicator instead of a fixed variable as in the original QTDT$_\mathrm{M}$ described in chapter 3. The QTDT$_\mathrm{M}$* is theoretically equivalent to the GQTDT when the specified distribution of the response variable or quantitative trait Y_i in the GQTDT analysis has a normal distribution. The same statistical analysis methods are applied in all types of simulated datasets. The full model used in the analysis is as follows:

$$Y_i = \beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i + \varepsilon_i \tag{5.9}$$

where Y_i, i = 1,2,...,N, are the random observations of the response variable which is the continuous quantitative trait and M_i, G_i and H_i refer to the random explanatory variables corresponding to the categorical mating type of the parents and genotype scores of the subject at locus A and at locus B, respectively. No other covariates were considered in the analyses. The βs are the unknown model parameters that must be estimated and ε_i is the error term.

In the GQTDT analysis, the distribution of the response variable Y_i is further defined as follows:

$$Y_i \sim f(g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p) \tag{5.10}$$

where:
f is the distribution of Y_i,
$(\theta_1, ..., \theta_p)$ are the parameters of f,
$g_1, ..., g_p$ are the link functions and
$t_1, ..., t_p$ are the model formulae for the explanatory terms and/or random effects in the predictors.

In these simulation studies, the $\mathrm{QTDT_M}$* is used in the analysis of normally simulated quantitative traits. The result is the same when a normal distribution is specified for the response variable in the GQTDT analysis. Specifying a normal distribution in the GQTDT analysis denotes using an identity and log link for the mean and dispersion parameters of the distribution, respectively. For the skewed quantitative traits, the $\mathrm{QTDT_M}$* was applied as is while the GQTDT was applied using the best fitted distribution. The generalized AIC was used in selecting the best fitted distribution. For example, in one simulation scenario with skewed QT, the AIC for the fitted t-family, lognormal and normal distributions are 2809, 3531 and 3615, respectively. The t-family with the smallest AIC was chosen. The t-family distribution uses the identity link for the mean, and the log link for both the dispersion and skewness parameters.

In practice, investigators more commonly look first at the individual genetic main effects of genes before considering testing for interaction of genes. However, some genes may have weak or no marginal effects but have significant joint or interaction effect. In these simulation studies, we consider two candidate genes from the same biological pathway and assume prior information that the genes may probably be associated to the quantitative trait of interest. To test the null hypothesis of no genetic main effects, i.e. $\mathrm{H_0}$: $\beta_G = \beta_H = 0$, the generalized likelihood ratio test was used to compare the global deviances of the model with genetic main effects $(\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i)$ and the model without genetic main effects $(\beta_0 + \beta_M M_i)$. For testing epistasis, i.e. $\mathrm{H_0}$: $\beta_{GH} = 0$, the model with epistasis $(\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i)$ is compared to the model without epistasis $(\beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i)$. Statistical testing for detecting genetic main effects and epistatic effect was performed under three commonly used genetic model assumptions, i.e. dominant, additive or recessive genetic model was specified as the "analysis genetic model". In reality, one does not know what is the true (in this case simulated) genetic model behind the effect of each locus of interest, so this needs to be assumed. For simplicity, the same analysis genetic model is assumed for both loci of interest in the analysis.

All statistical tests were evaluated at a level of significance $\alpha = 0.05$. For each simulation scenario, the detection rates of genetic main effects and epistasis were noted. The number of the types of datasets varies in each simulation scenario but the total number of replications for each type of dataset used in the analysis was fixed to 1000. Type I error rates were calculated for the probability of rejecting the null hypothesis of no genetic main effects and also for the probability of rejecting the null hypothesis of no epistasis given that the hypothesis is true. In datasets where the alternative hypothesis speaks the truth about the data, the power to detect the simulated effect was computed as the proportion of datasets that gave statistically significant results ($p \leq 0.05$) out of the total number of replications successfully tested. Comparisons of the power or type I error of the GQTDT in different scenarios are shown in the result tables.

5.3 Results

Generally, this section is organized according to the type of quantitative traits simulated. Depending on the simulation scenario, power and/or type I error of the statistical methods ($\mathrm{QTDT_M}$* and GQTDT) are shown. The results of the tests for detecting genetic main effects and epistasis are both presented. The results are tabulated according to population type, affecting locus, QT distribution and number of subjects (N) in the dataset. Results are also arranged according to the analysis genetic model used in the test. Since there are three genetic model assumptions (i.e. dominant, additive and recessive) being considered here, there are also three test results for each type of dataset. They are indicated in the tables under columns Dom, Add and Rec for dominant, additive and recessive analysis genetic models, respectively. For normally distributed QTs, only the result columns for $\mathrm{QTDT_M}$* have been filled-up since the results are the same with the GQTDT analysis.

5.3.1 QTs without locus effect (QT_N)

Tables 5.11 and 5.12 show the type I errors in detecting genetic main effects and epistasis in datasets with quantitative traits not affected by any locus. In these datasets, the proportion of falsely detecting genetic main effects and epistasis in the analysis of normally distributed QTs using both methods is roughly around 5% to 7%. Slightly elevated false detection rates were noted in the analysis of datasets with skewed quantitative traits. The error rates for some of these datasets were improved when the sample sizes were increased from 1000 to 2000 family trios. The use of the better fitted t-family distribution in the GQTDT analysis of skewed traits also shows slightly elevated type I errors in detecting genetic main effects and epistasis in some datasets. The result also shows that the slightly elevated type I error occur more often in PopB which has the highest minor allele frequencies among the three populations. In few cases, the type I error cannot be accurately determined because of the limited number of test that successfully converged. Convergence is not a problem in the GQTDT when analyzing a normally distributed dataset. But in other types of distributions, the problem of convergence may be encountered. In general, in this simulation, 1% - 6% non-covergence has been observed in the analysis of datasets with skewed QT. However, extreme number of non-convergence (> 50%) may be observed in cases when certain genotype frequencies are too few or there is no observed variation in the genotype. This can usually happen when the assumed genetic model in the analysis is recessive and the minor allele frequency of the locus investigated is low, as in the case of PopA (see table 5.12).

Table 5.11: Type I error in detecting genetic main effects in QTs with no-locus effect

Population Type	QT Type	Sample Size	Type I error of QTDT_M* with analysis genetic model			Type I error of GQTDT with f = TF and analysis genetic model		
			Dom	Add	Rec	Dom	Add	Rec
PopA	*Normal*	1000	0.046	0.050	0.059			
		2000	0.057	0.057	0.052			
	Skewed	1000	0.062	0.059	0.069	0.050	0.049	0.072
		2000	0.057	0.058	0.066	0.071	0.048	0.048
PopB	*Normal*	1000	0.064	0.056	0.056			
		2000	0.061	0.053	0.057			
	Skewed	1000	0.057	0.056	0.047	0.081	0.079	0.073
		2000	0.060	0.052	0.056	0.081	0.045	0.068
PopMix	*Normal*	1000	0.049	0.052	0.065			
		2000	0.055	0.065	0.063			
	Skewed	1000	0.071	0.066	0.068	0.053	0.065	0.046
		2000	0.063	0.046	0.056	0.053	0.058	0.071

Population types differ in the minor allele frequencies at the locus of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 96); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively.

Table 5.12: Type I error in detecting epistasis in QTs with no-locus effect

Population Type	QT Type	Sample Size	Type I error of QTDT_M* with analysis genetic model			Type I error of GQTDT with f = TF and analysis genetic model		
			Dom	Add	Rec	Dom	Add	Rec
PopA	*Normal*	1000	0.044	0.046	0.072			
		2000	0.043	0.047	0.050			
	Skewed	1000	0.061	0.053	0.054	0.064	0.064	-
		2000	0.066	0.058	0.046	0.060	0.083	-
PopB	*Normal*	1000	0.050	0.056	0.059			
		2000	0.055	0.049	0.038			
	Skewed	1000	0.049	0.061	0.050	0.073	0.079	0.079
		2000	0.051	0.057	0.061	0.081	0.083	0.062
PopMix	*Normal*	1000	0.055	0.052	0.039			
		2000	0.054	0.050	0.037			
	Skewed	1000	0.045	0.044	0.051	0.023	0.059	0.086
		2000	0.054	0.046	0.047	0.061	0.070	0.079

Population types differ in the minor allele frequencies at the locus of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 96); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ”-” Type I error cannot be accurately determined due to limited number of tests that successfully converged.

5.3.2 QTs with single-locus effect (QT_L)

Detection of genetic main effects were compared across different datasets with QTs affected only by a single locus. The performance of the methods in detecting the single locus effect (see table 5.13) differ across populations, QT distributional types and the genetic model used to simulate the locus. The methods showed high power (> 80%) in detecting genetic main effects when the QT is normally distributed and the correct genetic model is specified in the analysis. It can be noted in some cases that regardless of the underlying population, if the wrong genetic model is specified, the power to detect the locus effect is minimal. Take for example the datasets with locus A simulated under dominant genetic model and with normally distributed QTs (first row of table 5.13) . The power to detect the genetic main effects when a recessive genetic model is assumed during statistical testing is only 10.5%.

In the normally distributed QTs, the power under the dominant and additive genetic model assumptions does not differ when the locus involved was originally simulated with the dominant genetic model. However, the difference in power between using dominant and additive genetic model assumptions in the test can be seen when the locus involved was originally simulated with additive or recessive genetic model. Detection rates in datasets with locus effect simulated with the additive genetic model is not as satisfactory as datasets simulated with the dominant model. This is partly because the locus with the additive model was simulated with low contribution to the QT variability. The effect of the allele frequency in the power is also seen especially if one would look at the results of the analysis of datasets with recessive genetic model in PopA. Even if the QT is normally distributed and the correct genetic model was specified, the power could go as low as 26.5% (marked $^\diamond$ in table 5.13). The minor allele frequency is lower in the datasets from PopA than those from PopB and PopMix. The power was improved when the sample size was increased, as seen in the result in the same table (marked $^\blacklozenge$). The most evident result is the very low power of detecting genetic main effects when the distribution of the QT is skewed and the QTDT_M* or GQTDT with specified normal distribution was used in the analysis. This is obvious regardless of the type of population and type of genetic model used to generate the data. Although increased sample size helped a bit to improve the power of the tests in most datasets, the same trend of low power for skewed traits is observed in all the other simulation scenarios. The use of the t-family distribution in the GQTDT analysis helped improved the power of detecting genetic main effects in skewed traits. For example, the dataset from PopB whose skewed QT is affected by locus *A* which was simulated with recessive genetic model showed improved power when the appropriate distribution was used in the GQTDT analysis (See marked † in table 5.13). The improvement in power is noticeable especially when the correct genetic model is specified in the analysis (i.e. from 44.6% to 98.7% power under the recessive genetic model assumption). However, this big improvement is only obvious when the locus involved contributes a sizeable effect to the variability of the quantitative trait as in the case of locus *A*. In the case of locus *B* which has a smaller contribution to the QT variability, the power of the GQTDT method in detecting genetic main effects is quite low. For example, in dataset marked "‡", the power are only 10.1%, 10% and 8.5% under the dominant, additive and recessive genetic model assumptions, respectively.

Tests for epistatic effects were also made to determine false positive detection of epistasis in QTs affected only by a single locus. The type I errors of both methods

Table 5.13: Power to detect genetic main effects in QTs with single-locus effect

Population Type	Affecting Locus	Simulated Genetic Model	QT Type	Sample Size	Power of QTDT$_{\mathrm{M}}$* with analysis genetic model			Power of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	A	Dom	*Normal*	1000	1.000	1.000	0.105			
				2000	1.000	1.000	0.154			
			Skewed	1000	0.081	0.083	0.072	0.216	0.191	0.105
				2000	0.123	0.110	0.068	0.306	0.321	0.071
	B	Add	*Normal*	1000	0.272	0.304	0.100			
				2000	0.487	0.560	0.167			
			Skewed	1000	0.058	0.050	0.073	0.071	0.079	0.121
				2000	0.063	0.061	0.068	0.071	0.084	0.072
	A	Rec	*Normal*	1000	0.058	0.070	0.265◇			
				2000	0.062	0.098	0.468♦			
			Skewed	1000	0.052	0.058	0.112	0.092	0.087	0.259
				2000	0.062	0.070	0.160	0.080	0.111	0.518
PopB	A	Dom	*Normal*	1000	1.000	1.000	0.222			
				2000	1.000	1.000	0.407			
			Skewed	1000	0.084	0.088	0.056	0.364	0.282	0.080
				2000	0.141	0.127	0.057	0.682	0.469	0.116
	B	Add	*Normal*	1000	0.314	0.469	0.242			
				2000	0.562	0.735	0.418			
			Skewed	1000	0.054	0.054	0.056	0.101	0.100	0.085 ‡
				2000	0.056	0.055	0.056	0.090	0.081	0.099
	A	Rec	*Normal*	1000	0.086	0.648	0.986			
				2000	0.102	0.921	1.000			
			Skewed	1000	0.065	0.200	0.446	0.085	0.632	0.987 †
				2000	0.071	0.347	0.747	0.129	0.957	1.000
PopMix	A	Dom	*Normal*	1000	1.000	1.000	0.164			
				2000	1.000	1.000	0.277			
			Skewed	1000	0.091	0.092	0.066	0.255	0.244	0.052
				2000	0.117	0.104	0.060	0.548	0.400	0.052
	B	Add	*Normal*	1000	0.279	0.370	0.182			
				2000	0.521	0.668	0.304			
			Skewed	1000	0.075	0.072	0.064	0.070	0.083	0.064
				2000	0.057	0.050	0.062	0.087	0.070	0.071
	A	Rec	*Normal*	1000	0.052	0.324	0.816			
				2000	0.056	0.562	0.988			
			Skewed	1000	0.059	0.128	0.268	0.090	0.423	0.899
				2000	0.042	0.180	0.467	0.062	0.663	0.996

Population types differ in the minor allele frequencies at the locus of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 98); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ◇, ♦, ‡ and † are referred to in the text.

Table 5.14: Type I error in detecting epistasis in QTs with single-locus effect

Population Type	Affecting Locus	Simulated Genetic Model	QT Type	Sample Size	Type I error of QTDT$_{\mathrm{M}}$* with analysis genetic model			Type I error of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	A	Dom	*Normal*	1000	0.047	0.052	0.048			
				2000	0.048	0.050	0.044			
			Skewed	1000	0.065	0.053	0.054	0.043	0.056	-
				2000	0.069	0.060	0.044	0.060	0.072	-
	B	Add	*Normal*	1000	0.048	0.048	0.064			
				2000	0.042	0.047	0.057			
			Skewed	1000	0.064	0.052	0.054	0.049	0.057	-
				2000	0.070	0.060	0.044	0.060	0.060	-
	A	Rec	*Normal*	1000	0.067	0.064	0.063			
				2000	0.058	0.062	0.057			
			Skewed	1000	0.050	0.054	0.045	0.056	0.072	-
				2000	0.053	0.050	0.052	0.049	0.071	-
PopB	A	Dom	*Normal*	1000	0.052	0.057	0.051			
				2000	0.051	0.047	0.033			
			Skewed	1000	0.046	0.057	0.052	0.071	0.064	0.069
				2000	0.053	0.052	0.057	0.073	0.069	0.037
	B	Add	*Normal*	1000	0.052	0.058	0.055			
				2000	0.053	0.045	0.037			
			Skewed	1000	0.047	0.059	0.051	0.069	0.063	0.079
				2000	0.054	0.054	0.057	0.063	0.068	0.036
	A	Rec	*Normal*	1000	0.059	0.062	0.057			
				2000	0.060	0.066	0.057			
			Skewed	1000	0.041	0.053	0.060	0.058	0.054	0.076
				2000	0.047	0.058	0.051	0.066	0.057	0.077
PopMix	A	Dom	*Normal*	1000	0.057	0.060	0.031			
				2000	0.051	0.063	0.031			
			Skewed	1000	0.049	0.047	0.055	0.052	0.076	0.079
				2000	0.054	0.053	0.047	0.070	0.078	0.078
	B	Add	*Normal*	1000	0.052	0.050	0.045			
				2000	0.054	0.047	0.039			
			Skewed	1000	0.049	0.047	0.057	0.063	0.074	0.076
				2000	0.054	0.049	0.048	0.070	0.069	0.079
	A	Rec	*Normal*	1000	0.066	0.068	0.057			
				2000	0.048	0.065	0.050			
			Skewed	1000	0.053	0.062	0.060	0.070	0.054	0.067
				2000	0.057	0.053	0.050	0.074	0.057	0.054

Population types differ in the minor allele frequencies at the locus of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 98); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ”-” Type I error cannot be accurately determined due to limited number of tests that successfully converged.

in detecting epistatic effect are shown in table 5.14. The GQTDT method showed acceptable type I error rates in testing the hypothesis of no epistatic effect. However, like in the results seen in the datasets with no locus effect, some type I errors are elevated to 6%-7%.

5.3.3 QTs with two-locus effect (QT_{LL})

In datasets with two-locus effect, similar results as the single locus were noted (see tables 5.15 and 5.16). The results show high power in detecting genetic main effects when the QT is normally distributed and at least one of the modes of inheritance or genetic model is correctly specified in the analysis. In cases when one of the affecting loci was simulated from a recessive genetic model and the other from an additive genetic model, the power is lower in determining the genetic main effects. The power of the GQTDT is observed to be higher in datasets with higher minor allele frequencies. Analysis in PopB and PopMix will usually have higher power compared to the analysis of similar datasets from PopA which has the lowest minor allele frequencies among the three types of populations.

In skewed traits, the power of the GQTDT is markedly improved when using the t-family distribution and when the correct genetic model is assumed. In other simulations, up to 4x or more increase in power is usually observed when a better fitted distribution and correct genetic model is specified in the analysis. Examples of this increased power can be seen in results marked '‡' in table 5.15.

As to errors in detecting the non-existing epistasis in this simulation scenario, the performance of the GQTDT is the same as with previous simulation schemes. The observed type I error in the analysis of normally distributed traits is around 5%-6% with slight elevation in some cases of up to 7%. Type I error in detecting epistasis in skewed traits are observed to be slightly higher than type I error in normally distributed traits. In the mixed population, one dataset with skewed QT was noted to have the highest type I error (9%) in the simulation. In general, the higher type I errors (6%-8%) are also observed in other datasets from homogenous populations.

Table 5.15: Power to detect genetic main effects in QTs with 2-locus effect

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of QTDT_{M}* with analysis genetic model			Power of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	1.000	1.000	0.185			
				2000	1.000	1.000	0.327			
			Skewed	1000	0.086	0.085	0.073	0.238	0.246	0.112
				2000	0.138	0.124	0.071	0.386	0.358	0.084
	Rec	Add	*Normal*	1000	0.406	0.453	0.370			
				2000	0.658	0.740	0.661			
			Skewed	1000	0.055	0.066	0.122	0.130	0.138	0.328
				2000	0.067	0.074	0.174	0.084	0.146	0.585
PopB	Dom	Add	*Normal*	1000	1.000	1.000	0.498			
				2000	1.000	1.000	0.810			
			Skewed	1000	0.086	0.098	0.057	0.405 ‡	0.346 ‡	0.095
				2000	0.159	0.140	0.066	0.727 ‡	0.545 ‡	0.098
	Rec	Add	*Normal*	1000	0.468	0.923	0.994			
				2000	0.777	0.998	1.000			
			Skewed	1000	0.072	0.236	0.490	0.108	0.712 ‡	0.996 ‡
				2000	0.072	0.405	0.792	0.146	0.976 ‡	1.000 ‡
PopMix	Dom	Add	*Normal*	1000	1.000	1.000	0.359			
				2000	1.000	1.000	0.611			
			Skewed	1000	0.096	0.094	0.070	0.329	0.284	0.052
				2000	0.137	0.120	0.067	0.586	0.526	0.096
	Rec	Add	*Normal*	1000	0.416	0.734	0.909			
				2000	0.713	0.958	0.998			
			Skewed	1000	0.058	0.137	0.294	0.101	0.475	0.924
				2000	0.046	0.206	0.515	0.095	0.744	1.000

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 100); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ‡ is referred to in the text.

Table 5.16: Type I error in detecting epistasis in QTs with 2-locus effect

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Type I error of QTDT$_M$* with analysis genetic model			Type I error of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	0.047	0.053	0.048			
				2000	0.048	0.050	0.044			
			Skewed	1000	0.065	0.053	0.054	0.056	0.049	-
				2000	0.069	0.060	0.044	0.060	0.073	-
	Rec	Add	*Normal*	1000	0.065	0.068	0.054			
				2000	0.058	0.058	0.060			
			Skewed	1000	0.049	0.053	0.045	0.057	0.073	-
				2000	0.053	0.051	0.052	0.049	0.067	-
PopB	Dom	Add	*Normal*	1000	0.047	0.057	0.054			
				2000	0.052	0.050	0.035			
			Skewed	1000	0.046	0.058	0.052	0.066	0.069	0.077
				2000	0.052	0.053	0.057	0.073	0.071	0.037
	Rec	Add	*Normal*	1000	0.058	0.061	0.054			
				2000	0.059	0.071	0.053			
			Skewed	1000	0.040	0.054	0.061	0.059	0.045	0.076
				2000	0.047	0.059	0.051	0.061	0.062	0.070
PopMix	Dom	Add	*Normal*	1000	0.057	0.061	0.035			
				2000	0.055	0.069	0.031			
			Skewed	1000	0.049	0.047	0.055	0.059	0.076	0.069
				2000	0.054	0.052	0.047	0.069	0.068	0.070
	Rec	Add	*Normal*	1000	0.066	0.072	0.061			
				2000	0.048	0.063	0.047			
			Skewed	1000	0.052	0.061	0.060	0.070	0.051	0.068
				2000	0.056	0.053	0.050	0.074	0.094	0.053

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 100); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ”-” Type I error cannot be accurately determined due to limited number of tests that successfully converged.

5.3.4 QTs with two-locus effect and epistasis (QT_{LLI})

The performance of the $\mathrm{QTDT_M}$* and the GQTDT when applied to datasets with both genetic main effects and epistatic effect is tabulated in tables 5.17 and 5.18. Given a normally distributed trait, the power to detect the genetic main effects in the GQTDT analysis is fairly high when there is epistasis in the data. This holds true even if the genetic model is wrongly specified in the analysis. For example, in PopA with a dominant-additive interacting loci and normal QT (see ‡ in table 5.17), the power to detect the genetic main effects under a wrongly specified recessive genetic model is almost double (34%) than in similar dataset (see table 5.15) where the QTs have no epistatic effect. In the case of skewed traits, the GQTDT performs better than $\mathrm{QTDT_M}$* when the appropriate distribution is used. The improvement in power is double in some cases when the correct genetic model was specified. Take for example the case of PopB with simulated recessive-additive interacting loci. Even at a smaller sample size (see $^\diamond$ in table 5.17), the big improvement in power is clearly seen when the TF distribution is used and the assumed analysis genetic model is the same as any of the ”true” genetic model of the simulated loci, that is, either recessive or additive genetic model.

The detection of epistasis has much lower power than the detection of genetic main effects in both normally distributed and skewed traits (see table 5.18). Considering the small magnitude of the simulated epistatic effect, the power of both statistical tests cannot be expected to be high. At a larger sample size, the power to detect epistasis in the normally distributed traits increased considerably in all types of populations assuming either the dominant or additive genetic model. Under the recessive model assumption, the increase in power in both methods to detect epistasis when the sample size was increased to 2000 is minimal. In the case of the skewed traits, the power to detect epistasis effect is not satisfactory even at increased sample size. Specifying a TF distribution for better fit only minimally increased the power of the GQTDT to detect the epistatic effect.

Table 5.17: Power to detect genetic main effects in QTs with 2-locus effect and epistasis

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of QTDT$_{\mathrm{M}}$* with analysis genetic model			Power of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	1.000	1.000	0.340‡			
				2000	1.000	1.000	0.628			
			Skewed	1000	0.114	0.108	0.076	0.397	0.352	0.127
				2000	0.213	0.199	0.073	0.624	0.659	0.118
	Rec	Add	*Normal*	1000	0.925	0.969	0.523			
				2000	0.999	1.000	0.843			
			Skewed	1000	0.059	0.078	0.123	0.158	0.188	0.358
				2000	0.081	0.110	0.192	0.179	0.309	0.578
PopB	Dom	Add	*Normal*	1000	1.000	1.000	0.636			
				2000	1.000	1.000	0.922			
			Skewed	1000	0.114	0.110	0.059	0.501	0.423	0.131
				2000	0.191	0.175	0.075	0.830	0.679	0.132
	Rec	Add	*Normal*	1000	0.814	0.993	0.997			
				2000	0.982	1.000	1.000			
			Skewed	1000 ◇	0.078	0.278	0.535	0.131	0.740	0.996
				2000	0.078	0.465	0.836	0.187	0.986	1.000
PopMix	Dom	Add	*Normal*	1000	1.000	1.000	0.494			
				2000	1.000	1.000	0.816			
			Skewed	1000	0.121	0.121	0.069	0.412	0.448	0.084
				2000	0.182	0.172	0.074	0.759	0.664	0.122
	Rec	Add	*Normal*	1000	0.860	0.981	0.939			
				2000	0.992	1.000	0.998			
			Skewed	1000	0.063	0.165	0.323	0.128	0.529	0.922
				2000	0.073	0.259	0.567	0.203	0.807	1.000

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 100); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ◇ and ‡ are referred to in the text.

Table 5.18: Power to detect epistasis in QTs with 2-locus effect and epistasis

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of QTDT_M* with analysis genetic model			Power of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	0.314	0.287	0.036			
				2000	0.553	0.516	0.042			
			Skewed	1000	0.063	0.055	0.054	0.103	0.092	-
				2000	0.061	0.062	0.044	0.106	0.107	-
	Rec	Add	*Normal*	1000	0.330	0.317	0.054			
				2000	0.558	0.546	0.064			
			Skewed	1000	0.057	0.055	0.045	0.100	0.120	-
				2000	0.062	0.061	0.054	0.108	0.107	-
PopB	Dom	Add	*Normal*	1000	0.455	0.276	0.090			
				2000	0.759	0.446	0.116			
			Skewed	1000	0.062	0.066	0.052	0.113	0.087	0.109
				2000	0.061	0.053	0.062	0.169	0.135	0.062
	Rec	Add	*Normal*	1000	0.468	0.300	0.104			
				2000	0.760	0.490	0.148			
			Skewed	1000	0.050	0.054	0.063	0.086	0.049	0.081
				2000	0.059	0.065	0.056	0.157	0.106	0.087
PopMix	Dom	Add	*Normal*	1000	0.449	0.408	0.055			
				2000	0.726	0.665	0.067			
			Skewed	1000	0.056	0.050	0.051	0.078	0.116	0.071
				2000	0.064	0.069	0.052	0.121	0.103	0.078
	Rec	Add	*Normal*	1000	0.437	0.301	0.083			
				2000	0.706	0.501	0.108			
			Skewed	1000	0.058	0.057	0.061	0.095	0.067	0.065
				2000	0.060	0.064	0.048	0.132	0.100	0.092

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 100); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; "-" Type I error cannot be accurately determined due to limited number of tests that successfully converged.

Additional simulations were done to further check the performance of the GQTDT in analyzing QTs with simulated genetic main effects and epistasis. A different statistical model was used to generate the datasets. In this case, the model contains higher epistatic effect than the individual genetic main effects (See equation 5.7 on page 101). Tables 5.19 and 5.20 show the results of applying the GQTDT in these simulated datasets. In determining genetic main effects in the normally distributed traits, the GQTDT showed consistently high power in all simulation schemes but in the analysis of the skewed traits with the appropriately fitted distribution, the highest power observed is only 51%. Compared to the results in determining genetic main

Table 5.19: Power of the GQTDT to detect genetic main effects in QTs with 2-locus effect and strong epistasis

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of GQTDT with f = NO or TF and analysis genetic model		
					Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	1.000	1.000	0.997
			Skewed	1000	0.495	0.509	0.217
	Rec	Add	*Normal*	1000	1.000	1.000	1.000
			Skewed	1000	0.458	0.499	0.221
PopB	Dom	Add	*Normal*	1000	1.000	1.000	0.999
			Skewed	1000	0.186	0.280	0.206
	Rec	Add	*Normal*	1000	1.000	1.000	0.995
			Skewed	1000	0.200	0.293	0.175
PopMix	Dom	Add	*Normal*	1000	1.000	1.000	0.997
			Skewed	1000	0.334	0.370	0.207
	Rec	Add	*Normal*	1000	1.000	1.000	0.992
			Skewed	1000	0.302	0.392	0.220

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 101); NO = normal distribution; TF = t-family distribution used for the analysis of skewed traits; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively.

effects in skewed QTs in table 5.17, the power for the skewed QTs in this case is lower. A major factor in this difference is the magnitude of ”true” effect size being measured. In this situation, locus A was simulated to have only 0.01 effect size. This decreases the total genetic main effects contributed by the two loci. In the case of the normally distributed trait, the condition did not diminish the power in determining genetic main effects. In the analysis under the recessive genetic model, the power was even improved. The test in the normally distributed QTs possibly gained power from the higher interaction effect in the simulated dataset.

In determining epistasis, noticeable increase in power was seen in the GQTDT analysis of both normal and skewed QTs (see table 5.20). This is expected because of the higher epistatic effect simulated in this scenario. In the normally distributed QTs, the power to detect epistasis is 100% under both dominant and additive genetic model regardless whether the population is mixed or not. Under the recessive model, the power is not as high as the power under the dominant and additive genetic models, but the power is observed to be much higher than in the previous simulation scenario where the "true" epistatic effect is smaller. The observed power in testing under the recessive genetic model is not dependent on whether the population is homogenous or heterogenous. PopB which has the highest minor allele frequencies among the three populations shows the highest power followed by PopMix. PopA in this case has uncomputable power due to the high proportion of tests that failed to converge. In the analysis of the skewed QTs, the power is not as high compared to the result when the QT is normally distributed. However, compared to the previous scenario, the power of the GQTDT to determine epistasis is three times higher in general. Again, this is expected because of the higher epistatic effect simulated in this case.

Table 5.20: Power of the GQTDT to detect epistasis in QTs with 2-locus effect and strong epistasis

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of GQTDT with f = NO or TF and analysis genetic model		
					Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	1.000	1.000	-
			Skewed	1000	0.289	0.288	-
	Rec	Add	*Normal*	1000	1.000	1.000	-
			Skewed	1000	0.273	0.275	-
PopB	Dom	Add	*Normal*	1000	1.000	1.000	0.928
			Skewed	1000	0.427	0.287	0.137
	Rec	Add	*Normal*	1000	1.000	1.000	0.934
			Skewed	1000	0.383	0.237	0.153
PopMix	Dom	Add	*Normal*	1000	1.000	1.000	0.555
			Skewed	1000	0.377	0.291	0.155
	Rec	Add	*Normal*	1000	1.000	1.000	0.613
			Skewed	1000	0.359	0.275	0.146

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 100); NO = normal distribution; TF = t-family distribution used for the analysis of skewed traits; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; "-" Type I error cannot be accurately determined due to limited number of tests that successfully converged.

5.3.5 QTs with two-locus effect, epistasis and covariates (QT_{All})

Datasets with genetic main effects, epistasis and additional covariates were simulated to determine the behavior of the statistical methods in this situation. The same null hypotheses of no genetic main effects and no epistasis were tested. Pretending that there are no additional covariates affecting the quantitative trait, the same statistical analysis model as in all previous analysis was used. The results show that as long as the trait is normally distributed and at least one of the genetic model is correctly specified in the analysis, both $\mathrm{QTDT_M}$* and GQTDT show excellent power in detecting genetic main effects. In cases were one of the loci was simulated with a recessive genetic model, lower power to detect the genetic main effects was noted when the minor allele frequency in the population is low. This happens even if the correct genetic model is specified during statistical testing. However, the power to detect the genetic main effects in this case was satisfactorily improved when the sample size was increased. For example, in table 5.21, the result for PopA with recessive-additive genetic model and normal QT showed only 65.8% power (marked $^\diamond$ in the table) when the analysis was done using recessive genetic model assumption for both loci. At larger sample size, the power became 93.6%. Comparing the results of populations PopA and PopB, the power of the tests to determine the genetic main effects is higher (100%; marked $^\blacklozenge$ in table 5.21) in an example dataset from PopB where the minor allele frequency is high.

In this simulation scenario, there are observed improvements in the power of the GQTDT to detect the genetic main effects when the QT is not normally distributed. When using a better fitted distribution and at least one correct genetic model assumption, the power can range from 85% to 100%. This is especially observed when the minor allele frequencies of the loci of interest is high in the population, as in the case of PopB (see † in table 5.21). The results in the PopMix datasets are not much different from the PopB dataset results. The slightly higher power noted in PopB can be due to the fact that the minor allele frequencies in PopB are higher compared to PopMix datasets.

Regarding the detection of epistasis, higher powers are observed in the result of the analysis of QTs affected additionally by other covariates (table 5.22) than QTs without covariates (table 5.18). The increase in power is appreciable in QTs that are normally distributed but not in QTs that are skewed.

Table 5.21: Power to detect genetic main effects in QTs with 2-locus, epistasis & covariates

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of QTDT_M* with analysis genetic model			Power of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	1.000	1.000	0.422			
				2000	1.000	1.000	0.777			
			Skewed	1000	0.149	0.141	0.077	0.493	0.451	0.131
				2000	0.273	0.266	0.075	0.843	0.786	0.167
	Rec	Add	*Normal*	1000	0.974	0.986	0.658$^{\diamond}$			
				2000	1.000	1.000	0.936			
			Skewed	1000	0.066	0.090	0.152	0.173	0.205	0.426
				2000	0.092	0.136	0.241	0.214	0.362	0.696
PopB	Dom	Add	*Normal*	1000	1.000	1.000	0.753			
				2000	1.000	1.000	0.963			
			Skewed	1000	0.162	0.149	0.073	0.594	0.500	0.134
				2000	0.273	0.239	0.081	0.910	0.848	0.153
	Rec	Add	*Normal*	1000	0.907	0.997	1.000$^{\blacklozenge}$			
				2000	0.999	1.000	1.000			
			Skewed	1000	0.086	0.378	0.700	0.149	0.848†	1.000†
				2000	0.092	0.630	0.938	0.227	0.990†	1.000†
PopMix	Dom	Add	*Normal*	1000	1.000	1.000	0.609			
				2000	1.000	1.000	0.905			
			Skewed	1000	0.148	0.144	0.073	0.526	0.497	0.096
				2000	0.268	0.242	0.082	0.904	0.774	0.139
	Rec	Add	*Normal*	1000	0.945	0.998	0.983			
				2000	1.000	1.000	1.000			
			Skewed	1000	0.075	0.206	0.449	0.173	0.618	0.955
				2000	0.089	0.348	0.775	0.219	0.914	1.000

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables (See page 101); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; $^{\diamond}$, $^{\blacklozenge}$ and † are referred to in the text.

Table 5.22: Power to detect epistasis in QTs with 2-locus effect, epistasis & covariates

Population Type	Locus A Genetic Model	Locus B Genetic Model	QT Type	Sample Size	Power of QTDT$_{\mathrm{M}}$* with analysis genetic model			Power of GQTDT with f = TF and analysis genetic model		
					Dom	Add	Rec	Dom	Add	Rec
PopA	Dom	Add	*Normal*	1000	0.401	0.350	0.040			
				2000	0.676	0.623	0.039			
			Skewed	1000	0.064	0.061	0.051	0.090	0.098	-
				2000	0.064	0.059	0.048	0.120	0.155	-
	Rec	Add	*Normal*	1000	0.415	0.410	0.054			
				2000	0.686	0.684	0.062			
			Skewed	1000	0.061	0.054	0.045	0.115	0.099	-
				2000	0.061	0.067	0.056	0.138	0.121	-
PopB	Dom	Add	*Normal*	1000	0.584	0.342	0.088			
				2000	0.869	0.539	0.138			
			Skewed	1000	0.066	0.064	0.050	0.110	0.100	0.100
				2000	0.078	0.059	0.061	0.143	0.142	0.054
	Rec	Add	*Normal*	1000	0.587	0.365	0.112			
				2000	0.857	0.594	0.182			
			Skewed	1000	0.046	0.059	0.063	0.068	0.081	0.072
				2000	0.061	0.067	0.058	0.124	0.120	0.058
PopMix	Dom	Add	*Normal*	1000	0.561	0.512	0.048			
				2000	0.829	0.791	0.077			
			Skewed	1000	0.061	0.051	0.047	0.077	0.109	0.089
				2000	0.071	0.072	0.054	0.096	0.121	0.078
	Rec	Add	*Normal*	1000	0.553	0.389	0.093			
				2000	0.830	0.603	0.121			
			Skewed	1000	0.058	0.067	0.065	0.089	0.075	0.068
				2000	0.061	0.061	0.049	0.115	0.112	0.083

Population types differ in the minor allele frequencies at the two loci of interest (See page 91); The normal and skewed QTs were simulated using the same model parameters and explanatory variables. They only differ in the distribution of some explanatory variables and residuals (See page 101); TF = t-family distribution; Dom, Add, and Rec refer to dominant, additive and recessive genetic model, respectively; ”-” Type I error cannot be accurately determined due to limited number of tests that successfully converged.

5.4 Discussion

The simulation studies showed the powers and type I errors of the GQTDT in different scenarios. In phenotypes or QTs that are normally distributed, the natural choice in the GQTDT analysis is to specify a normal distribution which means using an identitiy link for the mean parameter of the distribution. In this case, the result of the GQTDT is the same as that of the $\mathrm{QTDT_M}$ by Gauderman (2003) when the mating type indicator is considered as a random variable. The observed power is very good (80% to 100%) in detecting genetic main effects except in cases when the loci were simulated from a recessive genetic model or the minor allele frequency is very low. This issue of low power in the case of a recessive genetic model is not a new or unexplained phenomena in genetic analysis. The condition in a recessive genetic model that both alleles should be susceptibility alleles before an effect on the response variable is observed makes it less frequent for the effect to be seen. This in turn affects the power of any test used to detect genetic effects. In general, the power the GQTDT is consistently high in detecting genetic main effects in the case of dominant and additive genetic model. However, low power may also be seen when analyzing data where the minor allele frequencies are low. This observation is not unique to the GQTDT and in general also true for other tests that are based on counting alleles and genotypes.

In detecting epistasis in normally distributed quantitative traits, the power of the GQTDT can also be considered satisfactory given the small epistatic effect simulated in the data. It was noted to range from 30% to 86% depending on the simulated loci and the sample size. The power to detect epistasis improved when the sample size was increased. In the additional simulation (page 117) where the simulated epistatic effect of the two loci was increased, the GQTDT as expected showed higher power in detecting epistasis in both normally distributed and skewed traits. Except in the case of the recessive genetic model, the power to detect strong epistasis in normally distributed QTs is 100%. In the skewed QTs, although the power of the GQTDT is not high, it is considerably higher than the $\mathrm{QTDT_M}$*. In simulations where the ”true” epistatic effect is strong, the power of the GQTDT to detect epistasis is higher, as expected, compared to simulations where the the ”true” epistasis in the data is weak.

With regard to the results of the analysis of the skewed quantitative traits, several important points can be noted. The results have clearly shown the increase in power to detect genetic main effects when one specifies a distribution and a model that fits the data well. Logically, it can be expected that a test that uses a better fitted model and distribution will perform much better than a test using an inappropriate model and distribution. The simulation studies presented here have emphasized the danger of using regression-based methods without care about the distributional characteristics of the data. In genetic analysis literature, one will encounter methods for determining genetic effects that seem to be applicable in general situations. However, these methods must be used with care because many complex traits in genetics have markedly non-normal distributions. The simulation scenarios here were designed to investigate only quantitative response variables with normal or skewed to the right distributions. The response variables were created not from a simple known distribution but from a mixture of different variables which are normal or non-normally distributed to simulate naturally occurring distributions of complex disease traits. The shape of the skewed distribution used here is actually based on a measure of radiation sensitivity among lung cancer patients. The advantage of the GQTDT over existing TDT methods is that it is flexible and can be applied to many different types of distributions. The method showed good power in determining genetic main effects in cases when the quantitative response variable fits a normal distribution or a t-family distribution. Even if the ”true” genetic main effect is small, the observed power to detect genetic main effect using GQTDT in general is still satisfactory. However, in determining epistasis, the power of the method can be minimal when the ”true” epistatic effect being detected is quite small and the quantitative response variable is not normally distributed. The power of the GQTDT to detect epistasis can be as high as 100% when the ”true” epistatic effect is considerably large and the response variable is normally distributed. In skewed traits, the power of the GQTDT to detect epistasis may be quite low but compared to the $\mathrm{QTDT_M}$*, the power is up to three times or more higher.

The need to assume a genetic model in coding genotypes in the GQTDT and also in the $\mathrm{QTDT_M}$* can be both an advantage and disadvantage. The power of the statistical tests is definitely higher when the test assumes the correct genetic model. But in reality, one does not know beforehand the underlying genetic model of the genes (if a gene is indeed involved) affecting the quantitative trait. Assuming one type of genetic model can miss the existing genetic effect if what was assumed was the wrong genetic model. However, one can test for several genetic models and get several results that can lead to the idea of what type of underlying genetic model plays a role in the genotype-phenotype relationship. Although this testing scheme can be tedious if one would consider all possible genetic models and more number of candidate genes to test. This will also have an impact on the p-values as one would need to adjust for multiple testing.

In terms of the type I error, the GQTDT gave fairly acceptable type I errors. The false positive detection rates of 5%-7% for both genetic main effects and epistasis are considerably acceptable. For some datasets with skewed QTs, the type I error can go slightly higher up to 9% when the fitted distribution in the GQTDT analysis is the t-family distribution. There are cases when a statistic cannot be computed because there is no variability in the offspring's genotype score. Certain mating types give only one possible type of offspring genotype and some mating types exist in very low frequency in the population. This can happen most of the times in recessive genetic models and rare susceptibility alleles. In recessive genetic model, the genotype scoring is made in such a way that subjects with two susceptibility alleles will have a genotype score of 1, otherwise the score is zero. If the susceptibility allele of a specific locus is rare in the population, the chance is high that there will be no subject in the dataset with a genotype score of 1 and all genotype score for the said locus is zero. This causes a convergence problem in the analysis of some datasets.

The presence of stratification did not pose any problem in the GQTDT analysis as seen in the comparable results in datasets from the homogeneous populations (PopA and PopB) and the heterogenous population (PopMix). This is one main advantage of family-based analysis methods like the GQTDT over population-based methods such as case-control analysis. TDT and TDT-like tests are robust to the influence of population stratification which can cause spurious association in the analysis. There may be an instance when an extreme result may be observed in the simulation studies but this is more attributed to random variation rather than population stratification. Concerning power of the test when there is stratification, no peculiar result was noted. The power in detecting genetic main effects and epistasis was noticeably affected by other factors such as the magnitude of the minor allele frequencies in the population and the type of genetic model simulated in the data rather than stratification. Other covariates affecting the quantitative response also have an overall effect on the power of the GQTDT. The power to detect the genetic main effect is better in datasets where the QTs were simulated with other covariate effects. This is probably due to the fact that one covariate was simulated to interact with one of the loci affecting the response variable. Higher power to detect epistasis is also noted in QTs affected by additional covariates than QTs without covariate effects. The higher power to detect epistasis is appreciable in normally distributed QTs but not much in skewed QTs especially when the "true" epistatic effect being detected is minimal.

In summary, the simulation studies demonstrated that the GQTDT performs well in detecting genetic main effects when an appropriately fitted distribution is used and the correct genetic model assumption is specified. It performs much better than the $\mathrm{QTDT_M}$* in detecting genetic main effects when the quantitative trait is skewed to the right. The flexibility of the GQTDT to accommodate other types of distribution is its major advantage over the $\mathrm{QTDT_M}$* which is based on linear regression. The GQTDT also performed satisfactorily in detecting epistasis in normally distributed traits especially when there is strong epistatic effect in the data. However, when applied to skewed QTs, the GQTDT did not perform as well as it did in normally distributed QTs when detecting epistasis. While the power is satisfactory in detecting genetic main effects even if the trait is skewed, the same cannot be said when detecting epistasis in skewed QTs. Capturing the "biological" epistasis is not that easy especially when the analysis is complicated by a nonnormal distribution of the response variable. It could be that the "best" fitted model used in the GQTDT analysis is still not the best for the particular data.

Although, careful planning was done to simulate naturally occurring complex quantitative traits, some of the simulation scenarios created here are likely to be simplified and idealized compared to real life situations. The allele frequencies in real populations may be more extreme than what is assumed in the simulations. Out of the many possible types of gene-gene interactions, the simulated epistatic effects are also simplified ones and not all types of genetic models are included. In addition, the relationship of the variables affecting the quantitative trait may be more complex in real life than what is specified in the simulations. Hence, it would be good to apply the statistical method to realistic situations. In order to address this issue, the GQTDT method is applied in real data and in another simulated data based on real situations in the next chapter.

6 Applications

In this chapter, the *Generalized Quantitative Transmission Disequilibrium Test* (GQTDT) was applied in the analysis of three datasets. The first two datasets are from the Genetic Analysis Workshop (GAW) 16 which was held in St. Louis, Missouri, U.S.A. The GAW is a joint effort of genetic epidemiologists worldwide to evaluate and compare statistical genetic methods. In the recently concluded GAW16, the datasets used in the workshop which are also used in this thesis are the real data from the Framingham Heart Study (FHS) and the simulated data by Kraja et al. (2009). The use of the GAW data has been approved by the local and international ethics committees after evaluation of the data protection, management and analysis plan of GAW participants. The other dataset used in this thesis is from the Lung Cancer in the Young (LUCY) Study, a multicenter study on lung cancer in Germany.

6.1 The FHS real dataset

6.1.1 Description and objectives

The FHS real dataset contains selected data from the the Framingham Heart Study (FHS) which began in 1948 among adults from the town of Framingham, Massachusetts. The research is under the direction of the National Heart, Lung, and Blood Institute (NHLBI) and is now conducted in collaboration with Boston University. The objective of the original study was to identify the common factors or characteristics that contribute to cardiovascular diseases (CVD) by following-up over a long period of time group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke. From the list of addresses recorded for the town of Framingham, two out of every three households were approached for participation in the study. All household members in the ages 30-60 years old within each house that was selected for study were recruited as participants. A total of 5,209 subjects (2,336 men and 2,873 women) between the ages of 29 and 62 have been recruited between the year 1948 and 1953 and comprised the *Original Cohort* of the study. Detailed medical history, physical examination, and laboratory tests of the participants were done at the beginning of the study and every two years thereafter. After 60 years of follow-up, there remain about 500 participants from the original group. During the period of 1971 and 1975, the study enrolled a second-generation group. An additional 5,124 individuals who are children (including their spouses) of the original cohort participated in the study. This second generation, called *Offspring Cohort* has 2,616 participants who are offsprings of spouse pairs from the original cohort, 34 are stepchildren, 898 are children of the Original Cohort members where only one parent was a study participant and 1,576 are spouses of the offsprings of the Original Cohort. The Offspring Cohort has been followed less frequently, that is, every four years until 2001. Similar protocols as the Original Cohort have been used for the Offspring Cohort. Between 2002 and 2005, FHS enrolled the *Third Generation* (Gen3) of study participants. This time, 4,095 offsprings of the second generation were recruited but their spouses were not recruited. An additional 103 parents of this third generation, who did not participate in the second phase of the study were also recruited at this time. However, this group of parents is not included in the GAW16 data. The Gen3 group had only one round of examination on file during GAW16. Genotyping data are not available for all participants since genetic studies began in the FHS only in the 1990s. The FHS entered a new phase in 2007 with the conduct of genotyping by Affymetrix for the FHS SHARe (SNP Health Association Resource) project, using approximately 550,000 SNPs (GeneChip® Human Mapping 500K Array Set and the 50K Human Gene Focused Panel). The SHARe data are stored at the National Center for Biotechnology Information database of genotypes and phenotypes (NCBI dbGaP) (http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?id=phs000007). Further information on the FHS study can be found at http://www.nhlbi.nih.gov/about/framingham/index.html.

The objective of the analysis of the FHS real dataset in this thesis is to determine genetic main effects and epistasis associated with body-mass-index (BMI). The data used in GAW include those FHS subjects who have consented to anyone's use of their information. The data are however anonymized or stripped of any personal identifier. Files for analysis contain $\sim$ 550,000 SNP genotype data, a pedigree file that provides the family structure and files with phenotypic data for each cohort

(Original, Offspring, Gen3). There are a total of 7130 subjects with phenotype data: 373 Original Cohort, 2760 Offspring Cohort and 3997 Gen3 participants. However, only 332 pedigrees (17 from the Offspring Cohort and 315 from the Gen3 Cohort) contain family trios which satisfy the conditions that all trio members (father, mother and child) are genotyped and that all children have baseline body-mass-index data in the phenotype file. The BMI which is the quantitative phenotype of interest in the FHS real dataset has been computed by dividing the weight (in kilograms) of the ”child” study subject with the square of its height (in meters). One family trio per pedigree was randomly chosen for analysis. Among the children of the sample trios, there are 139 (42%) males and 193 (58%) females. The average age of these children is 35.7 ± 8.2 years old. For the analysis of body-mass-index, five candidate SNPs from SNPedia (http://www.snpedia.com) which have been previously associated with modification of BMI were selected. The SNPs belong to genes that carry susceptibility variants previously identified to modify BMI in single-locus studies (Malzahn et al., 2009). Table 6.1 below shows some information about the SNPs used in the analysis (Source: http://www.snpedia.com).

Table 6.1: Selected SNPs for BMI analysis

SNP*	Gene	Located in	MAF*
rs Number		Chromosome	
rs6602024	PFKP	10	0.42
rs1121980	FTO	16	0.48
rs9930506	FTO	16	0.42
rs854560	PON1	7	0.24
rs6971091	FAM71F1	7	0.22

*SNP - Single Nucleotide Polymorphism; MAF - Minor Allele Frequency

The PFKP gene or Phosphofructokinase is involved in the regulation of glycolysis. The FTO or ’fat mass and obesity associated’ gene has been shown in genomewide association studies as a type 2 diabetes susceptibility gene. The other gene, PON1 or Paraoxonase 1 encodes the enzyme arylesterase that hydrolyzes paroxon to produce p-nitrophenol. Paroxon is a compound that is produced in vivo by oxidation of the insecticide parathion. Polymorphisms in this gene were found to be risk factors in coronary artery disease. The FAM71F1 (family with sequence similarity 71, member F1) gene, also known as FAM137A gene has been associated with obesity (NCBI, 2010). The SNPs used in the data analysis which belong to the above mentioned genes were all found associated with BMI in previous single-locus studies. In addition, significant statistical interaction effects on BMI involving FTO, PON1, and PFKP genes have been reported by Malzahn et al. (2009).

6.1.2 Analysis method

The GQTDT method was applied in the analysis of the FHS real datasets. Extension of the statistical model was done to accommodate *cohort effect* in the data. The term *cohort effect* refers to the effect or influence of shared characteristics among individuals. Cohorts or groups of individuals in a study are often defined by their entry date in the study, year of birth, or year of exposure to certain disease-causing agent. In the previous study by Malzahn et al. (2009), a strong cohort effect was found in the FHS real data (p-value = 0.015). A significant cohort effect has also been found in this analysis. The cohort effect signals differences in the quantitative trait (i.e. BMI) across generations. Therefore, the GQTDT model was extended to adjust for cohort effect and also other known covariates affecting BMI such as sex and age. The extended two-locus statistical model used is:

$$\begin{aligned} Y_i = \beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i + \gamma_1(sex) + \gamma_2(age) + \gamma_3(sex * age) \\ + \delta_1(cohort) + \delta_2(cohort * sex) + \delta_3(cohort * age) + \varepsilon_i \qquad (6.1) \end{aligned}$$

where:

Y_i	the random observation of a continuous quantitative response or phenotype (i.e. baseline BMI) of the ith study subject; i = 1,...,N
M_i	explanatory variable representing the parental mating type
G_i	the genotype score of the study subject at SNP or locus 1
H_i	the genotype score of the study subject at SNP or locus 2
β_0	the intercept
β_M	regression coefficient for the parental mating type
$\beta_G, \beta_H, \beta_{GH}$	regression coefficients for the effects of locus 1, locus 2 and their interaction
$\gamma_1, \gamma_2, \gamma_3$	regression coefficients for the effects of sex, age and their interaction
$\delta_1, \delta_2, \delta_3$	regression coefficients for the effects of cohort, cohort*sex and cohort*age interactions
ε_i	residual, $\sim N(0, \sigma^2)$

The variable age is used as a continuous quantitative covariable. Sex has its usual two categories: male and female. The variable cohort has also two categories: Offspring Cohort and Gen3 Cohort. The genotype scores G_i and H_i is 0 for genotype code *'aa'*, 1 for *'AA'* and 0, 0.5 or 1 for *'Aa'* depending on the assumed model being recessive, additive or dominant, respectively.

The response variable Y_i in the statistical model is distributed as:

$$Y_i \sim f(g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p) \tag{6.2}$$

where:
f is the distribution of Y_i,
$(\theta_1, ..., \theta_p)$ are the parameters of f,
$g_1, ..., g_p$ are the link functions and
$t_1, ..., t_p$ are the model formulae for the explanatory terms in the predictors.

The distribution and QQ plot of the body-mass-index in the FHS real dataset are shown in figure 6.1. As one can see, the distribution of BMI in the data does not nicely fit a normal distribution. Therefore, other distributions were tried. The choice of the distribution in the GQTDT analysis was based on the graphical characteristic of the data and the result of the generalized AIC criterion after fitting several candidate distributions. The computed AIC for the model using normal distribution, t-family, lognormal, and Box-Cox Power Exponential (BCPE) are 1487, 1457, 1453 and 1433, respectively. Based on the smallest AIC, the distribution fitted to the data in the GQTDT analysis is the BCPE. The BCPE is usually used for modeling skewness combined with kurtosis in continuous data (Rigby and Stasinopoulos, 2005). The QQ plot after fitting with the BCPE distribution is shown in figure 6.2. Compared with the fit of the normal distribution in figure 6.1, improvement in the fit is seen when the BCPE was used.

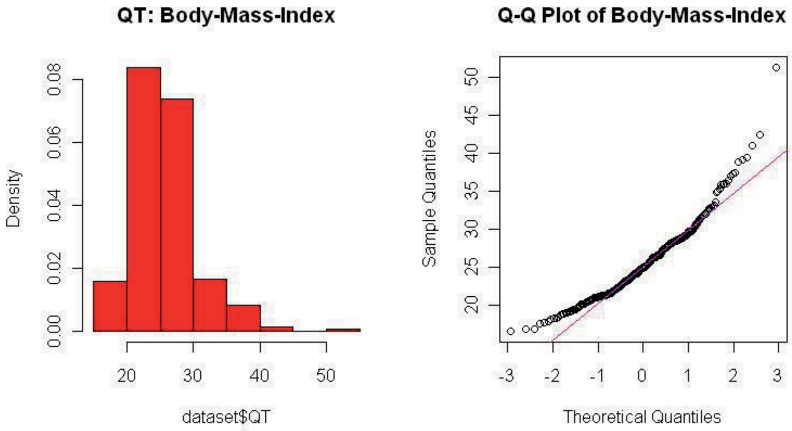


Figure 6.1: Distribution of baseline BMI in the FHS real dataset, N=332

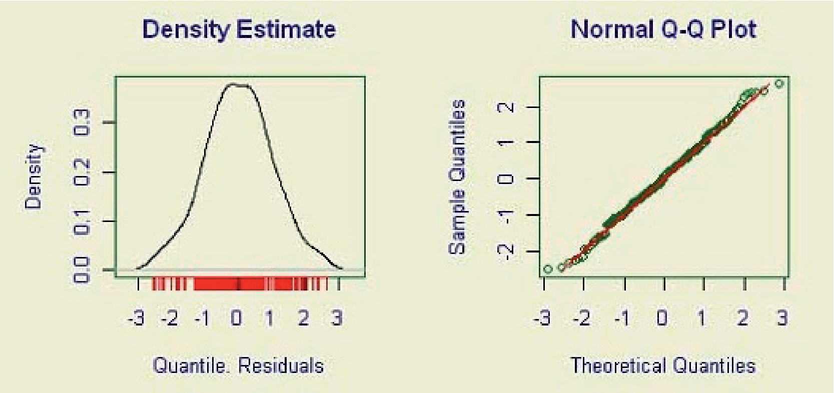


Figure 6.2: Density estimate and QQ Plot of baseline BMI after fitting BCPE, N=332

Statistical testing was done pairwise or with two loci at a time to determine genetic main effects and epistasis. Tests to determine the effect of each individual SNP were also done in the dataset. In the test for the effect of individual SNPs, there are 6 possible mating types for each biallelic locus in contrast to the 36 possible mating types when SNPs are analyzed pairwise. The generalized likelihood ratio test was used to test for the hypothesis of no genetic main effects and the hypothesis of no epistasis. The genetic main effect of individual SNPs was tested using equation 6.1 without the explanatory variable H_i and the interaction term as a full model. The likelihood of this model with only one locus and the mating type was compared with the model with only the mating type to get the genetic main effect of a specific SNP. A pairwise analysis using two SNPs at a time was also done by comparing the model with the two SNP main effects (G_i and H_i) and mating type with the model without the two SNP main effects. For testing epistasis, the likelihood of the full model with the interaction of the two loci (equation 6.1) was compared against the likelihood of the null hypothesis model of no locus interaction. The GQTDT was applied to the data under three different genetic model assumptions, i.e. dominant, additive or recessive genetic model. The computed p-values were adjusted for multiple testing using Holm's (1979) procedure.

6.1.3 Results

On the average, the baseline BMI of children in the analyzed family trios is 25.6 ± 4.9. Figure 6.3 shows the distribution of parents' genotype combination or mating types across different SNP pairs. The mating types are represented by the different segments of the bars in the figure. Though the graph cannot clearly depict the details of the mating types in each SNP pair, one can notice that there is no uniform distribution of mating types. Depending on the SNP pair, some mating types may be more frequent while other mating types do not occur at all.

In the individual SNP analysis, two out of the five SNPs analyzed showed significant genetic main effects ($p<0.05$). The SNP rs1121980 and rs9930506 which showed significant genetic main effects both belong to the gene FTO. For the pairwise analysis of the SNPs, the results are summarized in table 6.2. The GQTDT detected genetic main effects in six out of nine SNP pairs tested but did not detect any significant epistatic effect under any of the genetic model assumption. Differences in the results are noted depending on the genetic model specified in the analysis. For instance,

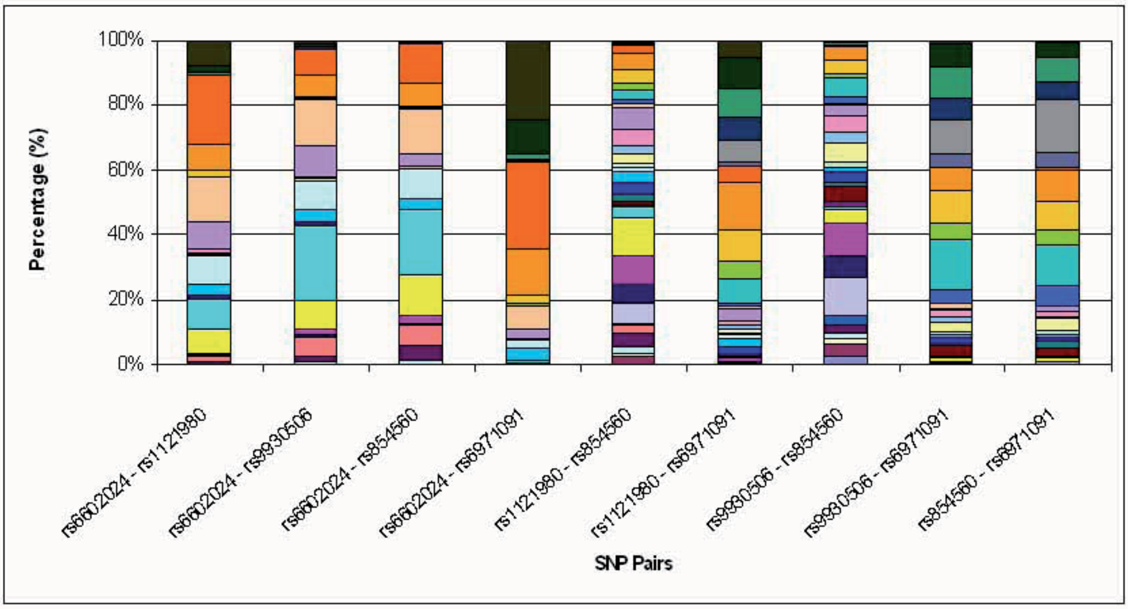


Figure 6.3: Distribution of mating types in the FHS real dataset
*The shades within the bars represent the different mating types; N = 332

Table 6.2: Results of GQTDT analysis of BMI in the Framingham real dataset

SNP Pairs tested		GQTDT p-values* under assumed analysis genetic model					
		Dominant		Additive		Recessive	
		main	inter.	main	inter.	main	inter.
rs6602024	rs1121980	<0.001	-	<0.001	n.s.	0.027	n.s.
	rs9930506	0.008	-	<0.001	n.s.	<0.001	n.s.
	rs854560	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs6971091	n.s.	-	n.s.	n.s.	n.s.	n.s.
rs1121980	rs854560	0.023	n.s.	n.s.	n.s.	n.s.	0.065
	rs6971091	0.003	n.s.	0.014	n.s.	n.s.	n.s.
rs9930506	rs854560	n.s.	n.s.	n.s.	n.s.	0.043	n.s.
	rs6971091	n.s.	n.s.	0.017	n.s.	0.022	n.s.
rs854560	rs6971091	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

*Listing only p-values ≤ 0.1, otherwise marking the test as 'n.s.' (not significant) or '-' where analysis was not possible; main = p-value for genetic main effect; inter. = p-value for interaction; Fitted distribution is the Box-Cox power exponential

the genetic main effect in testing SNPs rs1121980 and rs854560 was only picked-up by the test when the assumed genetic model specified is dominant. In the case of rs9930506 and rs854560, the significant genetic main effect was noted when the assumed genetic model is recessive. Some tests for epistatic effect are not computable under the dominant genetic model assumption and are therefore marked '-' in the table.

The type I errors of the GQTDT were calculated in this dataset using the SNPs rs1121980 and rs854560. One thousand datasets were created from the original dataset. Each generated dataset contains permuted BMI data. The permutation of the BMI trait is to make sure that any relationship (if any) between the trait and the genotypes are removed. The statistical tests were then applied to the permuted datasets to determine the type I error. Slightly elevated type I errors (8% - 9%) are observed with the GQTDT using Box-Cox Exponential Distribution.

6.1.4 Discussion

In the FHS real dataset, the GQTDT detected genetic main effects in the single SNP and two-SNP analyses. The SNPs rs1121980 and rs9930506 (both belonging to FTO gene) which showed significant genetic main effects in single SNP analysis are also involved in the two-SNP analyses with signicant genetic main effects. All SNP pairs which showed significant genetic main effects involve either of the two SNPs which showed significant effects in the single SNP analysis. The strongest genetic main effect was seen when the two mentioned SNPs were tested with rs6602024. The result of the two-SNP test with rs6602024 showed significant effects in all the analysis genetic model. In other SNP pairs (e.g. rs9930506 - rs854860), significant genetic main effects are seen only under one genetic model assumption.

The analysis detected genetic main effects but not epistatic effect. However, one SNP pair (rs1121980 - rs854560) gave a p-value slightly indicative of epistasis. The same SNP pair was reported to have significant interaction effect on BMI by Malzahn et al. (2009) who used the same FHS datasets utilizing longitudinal nonparametric association test (LNPT) and semiparametric survival analysis methods. They tested 8 gene pairs for epistasis which were also tested in this application. Their two approaches showed evidence for pairwise interaction between three other genes FTO (rs1121980 and rs9930506), PON1 (rs854560) and PFKP (rs6602024). These interactions are not seen in the GQTDT results. If indeed epistasis is present in the mentioned SNPs from the real dataset, one reason that could have contributed to the GQTDT's non-detection of epistasis and also genetic main effects in some of the SNP pairs is the small sample size. There are only 332 family trios in the GQTDT analyzed dataset. The given number of family trios may not be enough to detect epistasis if one is considering loci whose epistatic effect contributes only a very small proportion in the quantitative trait variability. Based on Gauderman's (2003) sample size recommendation for a linear regression method, two loci with 20% minor allele frequencies that are both defined by a dominant genetic model and contributes 2% main effects to the quantitative trait will need roughly 484 families to reach an 80% power to detect epistasis. The sample size requirement could get really high if the effects of the loci are smaller. Considering that Gauderman's recommended sample size was computed for a linear regression method, a different sample size requirement may be expected for other methods like the GQTDT. The study of Malzahn et al. (2009) has the advantage of having more study subjects since it is population-based. The two methods in the study used different statistical models which incorporates longitudinal data. Their LNPT method profited from using phenotypes from several time points and the survival analysis method used additional information from other time-varying covariates which are smoking status and cholesterol treatment. The LNPT method is also a robust method without distributional assumptions. It is applicable to data on individual subjects but not for families. Methods like the LNPT and survival analysis which is based on longitudinal data might be more difficult to collect in the long run than collecting observations from family trios at one point in time. Longitudinal studies face a big risk of loss-to-follow up study subjects and problem in harmonizing the data especially when observations need to be collected at the same time intervals. In addition, population-based longitudinal methods will also need to adjust for population stratification. Including the cohort effect as a covariate will be able to control for differences between generations but not for effects of possible population stratification. In this case, family-based methods like the GQTDT will have more advantage. The GQTDT also has the advantage of using the quantitative response variable as originally defined without worries about categorization. In other methods, like in the Cox-proportional hazards model (Malzahn et al., 2009), one has to first adopt a way to logically categorize the quantitative response variable. In some cases, like in the case of BMI, it is easy because there are already existing standards about meaningful BMI categories. But in other cases, like in gene expression or radiation sensitivity, one would have to think of a way how to categorize the response variable.

6.2 The FHS simulated dataset

6.2.1 Description and objectives

The FHS simulated data (FHSsim) by Kraja et al. (2009) are based on real data. The basic demographic profile (e.g. sex, age) and family structures are data of the Framingham Heart Study participants. Only the phenotypes of the participants were replaced by simulated values. The simulated phenotype in the FHSsim datasets is the Coronary Artery Calcification (CAC) score which in practice is detected by electron beam computed tomography that quantifies the coronary artery calcium levels. The CAC score is computed by assigning a weighted value to the highest density (measured in Hounsfield units) of calcification in a given coronary artery. The total CAC score is the sum of the calcium scores of every calcification in each coronary artery for all of the tomographic slices. The CAC score which can range from zero to several thousands is a predictor of clinical coronary artery disease (Agatston et al., 1990). The higher the CAC score, the higher the amount of calcium causing increased coronary atherosclerotic burden. In the simulated data of Kraja et al. (2009), the CAC was simulated as a function of total cholesterol, high density lipoprotein and five other SNPs ($\tau1$ - $\tau5$) having direct effects on its development. Below is the model used by Kraja et al. (2009) to simulate the CAC score.

$$CAC = 500 + 20(Chol200)25(HDL53) + ME + PE + Het + 300(\varepsilon) \tag{6.3}$$

The explanatory variables $Chol, HDL, ME, PE$ and Het are measures of the total cholesterol, high density lipoprotein, epistasis with main effects, pure epistatic effect and heterosis (over dominance effect), respectively. The explanatory variables of interest in this analysis are the ME and PE. The ME is a joint genetic effect from an epistatic interaction between SNPs $\tau1$ and $\tau2$. The SNP $\tau1$ was simulated to display only a minimal main effect and SNP $\tau2$ has an additional measurable additive main effect. PE is the joint effect of SNPs $\tau3$ and $\tau4$ which were simulated as a pair of purely epistatic SNPs. All four SNPs ($\tau1$ - $\tau4$) have minor allele frequencies of around 0.50. The ε in the model is the residual variation not explained by the explanatory variables in the model. It is 300 times a random draw from a normal distribution

∼N(0,1). It represents the sum of normal deviations from the mean of each genetic effects and noise from other unmeasured effects. The mean effects on CAC due to the ME and PE variables used by Kraja et al. (2009) are shown in tables 6.3 and 6.4.

Table 6.3: Mean effect of ME ($\tau1$ and $\tau2$) on CAC

	Genotype at SNP $\tau2$			marginal
	2/2	2/4	4/4	effects
Genotype at SNP $\tau1$				
2/2	-250	0	250	0
2/4	150	0	-150	0
4/4	-250	0	250	0
marginal effects	-100	0	100	

Table 6.4: Mean effect of PE ($\tau3$ and $\tau4$) on CAC

	Genotype at SNP $\tau4$		
	1/1	1/2	2/2
Genotype at SNP $\tau3$			
2/2	200	-200	200
2/4	-200	200	-200
4/4	200	-200	200

The Chol and HDL explanatory variables were simulated based on complex relationship of other variables such as age, sex and polygenes. This thesis only focuses on the analysis of effects of the four SNPs ($\tau1$ - $\tau4$) directly affecting the CAC score. Therefore, the simulation of the other variables is not anymore detailed here. The complete information about the simulation of the dataset can be found in the publication of Kraja et al. (2009). A negative CAC score is not possible in practice, so to avoid getting a negative CAC, Kraja et al. (2009) conditioned the CAC to be zero if the generated value was negative. The objective of this analysis is to determine the power of the GQTDT in detecting the genetic main effects and epistasis in the SNP pairs $\tau1$ - $\tau2$ and $\tau3$ - $\tau4$. All the 200 replicates of the FHSsim dataset were used in the analysis.

6.2.2 Analysis method

The analysis method for the FHSsim dataset is exactly the same as the FHS real dataset. All the explanatory variables are the same as the FHS real dataset except for the specific SNP investigated and the response variable which is the CAC score. The two-locus statistical model used is:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i + \gamma_1(sex) + \gamma_2(age) + \gamma_3(sex * age) \\ & + \delta_1(cohort) + \delta_2(cohort * sex) + \delta_3(cohort * age) + \varepsilon_i \end{aligned} \tag{6.4}$$

where:

Y_i	the random observation of a continuous quantitative response or phenotype (i.e. CAC score) of the ith study subject; i = 1,...,N
M_i	explanatory variable representing the parental mating type
G_i	the genotype score of the study subject at SNP or locus 1
H_i	the genotype score of the study subject at SNP or locus 2
β_0	the intercept
β_M	regression coefficient for the parental mating type
$\beta_G, \beta_H, \beta_{GH}$	regression coefficients for the effects of locus 1, locus 2 and their interaction
$\gamma_1, \gamma_2, \gamma_3$	regression coefficients for the effects of sex, age and their interaction
$\delta_1, \delta_2, \delta_3$	regression coefficients for the effects of cohort, cohort*sex and cohort*age interactions
ε_i	residual, $\sim N(0, \sigma^2)$

Like in the FHS real data, the variable age is used as a continuous quantitative covariable. Sex has its usual two categories: male and female. The variable cohort has also two categories: Offspring Cohort and Gen3 Cohort. Likely, the genotype scores G_i and H_i are given the value 0, 0.5 or 1 depending on the assumed genetic model in the analysis.

The response variable Y_i in the statistical model is distributed as:

$$Y_i \sim f(g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p) \tag{6.5}$$

where:
f is the distribution of Y_i,
$(\theta_1, ..., \theta_p)$ are the parameters of f,
$g_1, ..., g_p$ are the link functions and
$t_1, ..., t_p$ are the model formulae for the explanatory terms in the predictors.

In contrast to the BMI, the distribution of CAC score is skewed to the right with inflated number of zeros at the left of the distribution (see figure 6.4). On the average, there are about 57% zero CAC scores in the simulated datasets. The characteristic of the CAC score entails the use of GQTDT with a distribution that can handle plenty of zeros in the quantitative response variable. Among the available models for continuous variables, the zero-adjusted Gaussian (ZAIG; see page 77) distribution is the one most appropriate for the characteristic of the data. The AIC was not anymore used as a criteria for model selection in this case because the other types of distributions that may be used for skewed data (e.g. t-family, lognormal, BCPE) failed when specified in the GQTDT analysis. The QQ plot of the data after fitting a ZAIG distribution is presented in figure 6.5. The fit is not perfect but better compared to the fit in figure 6.4 using a normal distribution.

We have here a situation of a SNP pair with weak main effects and another SNP pair with only epistatic effect. Separate analysis is done for each SNP pair which are hypothesized to be interacting. The analysis did not looked into individual SNP genetic effect but on the combined SNP genetic main effects. Again, the generalized likelihood ratio test was used to test for the hypothesis of no genetic main effect by comparing the model with the two SNP main effects (G_i and H_i) and other covariates with the model without the two SNP main effects. The hypothesis of no epistatic effect was tested by comparing the likelihood of the full model with the loci interaction (equation 6.4) against the null hypothesis model of no loci interaction. The GQTDT

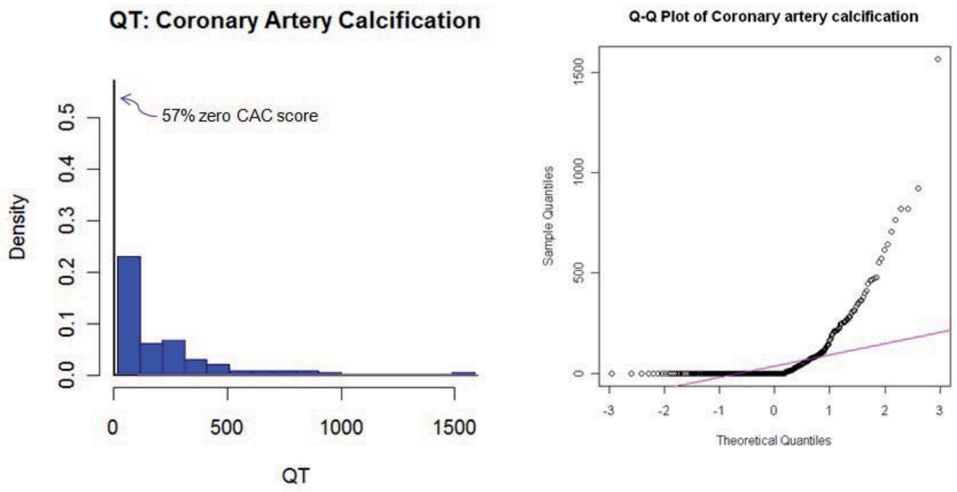


Figure 6.4: Distribution of CAC in the FHS simulated dataset, N=323

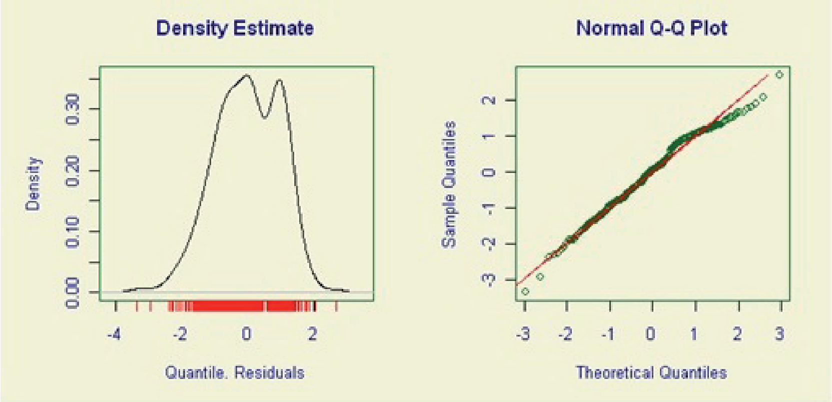


Figure 6.5: Density estimate and QQ Plot of CAC after fitting ZAIG, N=323

was applied to the data under three different genetic model assumptions, i.e. dominant, additive or recessive genetic model. The power was estimated as percentage of significant results ($p \leq 0.05$) of the test on the 200 FHSsim replicates. The average number of family trios among the 200 replicates is 323.

Although quite inappropriate for the distribution of the CAC, the QTDT_M* was also applied to the dataset for the purpose of comparing the power of the test with GQTDT. The same statistical model with the specified explanatory variables was used.

6.2.3 Results

The distribution of the FHSsim study subjects is also sparse in some mating types considering the two pairs of SNPs (figure 6.6). The CAC scores in the datasets ranges from 0 to more than 1500 with an average of 77. The analysis of the CAC scores using GQTDT detected the $\tau 1$ and $\tau 2$ genetic main effects in 54% - 60% of the 200 replicates. This is slightly higher than the observed power (47% - 53%) for detecting genetic main effects for the $\tau 3$ and $\tau 4$ SNP set. For the power to detect epistasis, the GQTDT showed 32% to 45% power for the first SNP set and 26% to 39% power for the second SNP set (table 6.5). Much lower power was observed when $\mathrm{QTDT_M}$* was used (see table 6.6).

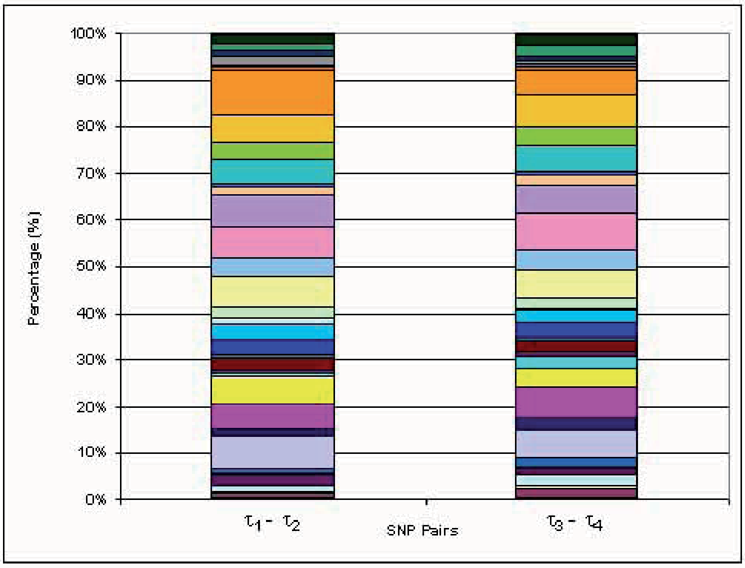


Figure 6.6: Distribution of mating types in the FHS simulated dataset
*The shades within the bars represent the different mating types; N = 323

Table 6.5: Power of the GQTDT in the simulated Framingham (FHSsim) data

SNPs	Power under analysis genetic model					
tested	Dominant		Additive		Recessive	
	main	inter.	main	inter.	main	inter.
$\tau1$ x $\tau2$	0.597	0.316	0.573	0.450	0.542	0.333
$\tau3$ x $\tau4$	0.469	0.259	0.533	0.391	0.494	0.326

main = genetic main effect; inter. = interaction effect or epistatis

Table 6.6: Power of the QTDT_M* in the simulated Framingham (FHSsim) data

SNPs	Power under analysis genetic model					
tested	Dominant		Additive		Recessive	
	main	inter.	main	inter.	main	inter.
$\tau1$ x $\tau2$	0.095	0.075	0.095	0.055	0.080	0.045
$\tau3$ x $\tau4$	0.095	0.080	0.090	0.090	0.070	0.090

main = genetic main effect; inter. = interaction effect or epistatis

6.2.4 Discussion

The high proportion of zero CAC scores in the data created difficulty in finding a fitted distribution. In the analysis of the two SNP pairs using GQTDT, the highest observed power to detect the genetic main effect is 60% and the highest observed power to detect epistasis is 45%. Considering the performance of a traditional linear regression method such as the QTDT_M, the observed result in the GQTDT is much better. Applying a typical linear regression model using the same explanatory variables to explain the variability in the CAC score gives only 5% to 10% power in detecting genetic main effects or epistasis. In detecting genetic main effects, the observed power in the FHSsim data analysis is not as high as the power in the simulation studies in chapter 5. This can be explained by the different simulation models used in generating the data. In chapter 5, the simulated datasets which showed 100% power in detecting genetic main effects have higher "true" genetic main effects in the data compared to the FHSsim data. In addition, the response variable, CAC, in the FHSsim dataset is characterized by inflated number of zeroes which makes it difficult to find a best fitting distribution. Malzahn et al. (2009) also analyzed the FHSsim data for gene-gene interaction. They used longitudinal methods to determine the presence of epistatic effects in the two SNP pairs. They compared longitudinal nonparametric association test (LNPT) and survival analysis methods. The LNPT detected both interactions in SNP pairs τ_1 - τ_2 and τ_3 - τ_4 with 100% power using 856 study subjects. Their survival analysis using Cox model had a power of 94% for τ_1 - τ_2 interaction and 100% for τ_3 - τ_4 interaction with 808 sample size. The survival analysis method benefited from the cohort effect adjustment which led to an increase of power from 69% to 94% in testing for interaction between τ_1 and τ_2. In detecting epistasis, the power of the GQTDT is lower compared to the power in the methods used in the study of Malzahn et al. (2009) using the same simulated dataset. The reasons for lower power observed in the analysis of the FHS simulated datasets are similar to the reasons mentioned for the FHS real dataset. The limited number of family trios available for analysis is one factor contributing to the lower power in the FHSsim analysis compared to the power of the methods in Malzahn et al. (2009). In addition, the GQTDT model did not consider longitudinal covariates which also contributes to the variation of the CAC score. The longitudinal approach using case-control data benefited from additional information, but as previously mentioned it has disadvantages in terms of controlling for population stratification and applicability when there is incomplete and uneven follow-up of repeated measurements from the study participants.

Comparing the result of this section with the simulation studies involving two loci and epistasis in chapter 5, the power for detecting epistasis is slightly higher in this application. Given a different scenario in this case, one cannot directly compare the results with the simulation in chapter 5. The testing in FHSsim dataset involve an extended model with other covariates and a different distribution as well. In addition, the minor allele frequencies of the SNPs in FHSsim are higher compared to the SNPs in the simulation studies in chapter 5. Higher minor allele frequencies improves the chances of detecting genetic effects in the data. However, for the test of genetic main effects, the results of the simulations in chapter 5 has higher power than the FHSsim because the simulated genetic main effect in chapter 5 contributes more to the response variable. In the FHSsim, there is no separate genetic main effect of the SNPs in addition to their joint effect. In the test for genetic main effects in the "purely epistatic" SNPs τ_3 and τ_4, the test seems to get power from the interaction effects of the SNPs. The "purely epistatic" effect created by Kraja et al. (2009) also signals genetic main effects in the GQTDT test. The differences in means across genotypes considering one locus has been detected as a genetic main effect. The definition of "purely epistatic" is quite artificial in this scenario. If this "purely epistatic"" effect exists in nature, the method in detecting epistatic effects using GQTDT and related methods will not be specific enough to detect this type of epistasis without falsely detecting a genetic main effect. Compared to the $\mathrm{QTDT_M}$*, the GQTDT using the zero-adjusted Gaussian (ZAIG) distribution showed better power to detect both genetic main effects and epistasis. In the case of the ”purely epistatic” SNP pairs (τ_3 and τ_4), both $\mathrm{QTDT_M}$* and GQTDT detected genetic main effects. This special case of epistasis can probably be investigated in future studies.

6.3 The LUCY dataset

6.3.1 Description and objectives

Lung cancer is a major cause of cancer death worldwide resulting to 1.3 million deaths per year. Based on the number of global deaths, it ranks first in cancer mortality among men and second to breast cancer among women (WHO, 2009a). Given the global burden of the disease, numerous studies investigate the risk factors related to the prognosis and treatment of lung cancer. The Lung Cancer in the Young (LUCY) study is a multi-center study involving 31 major lung clinics all over Germany. The LUCY study forms part of the lung cancer investigation at the *Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie der Ludwig-Maximilians-Universität München* whose main objectives are to identify and replicate susceptibility genes for lung cancer considering potential effect modifications by the amount of smoke exposure or radiation susceptibility. Patients diagnosed histologically or cytologically with primary lung cancer at age 50 or younger were recruited to the study as well as their families. This was done in cooperation with the *Department of Genetic Epidemiology, University of Göttingen Medical Center.*

Almost 800 patients with primary lung cancer and their relatives were recruited in the LUCY study (Sauter, 2008). Consent forms were signed by all participants before inclusion in the study. A questionnaire in which detailed information on personal history, history of lung diseases, family history of cancer and smoking habits was administered to the patients and participating family members. Blood samples were taken from all participants for DNA extraction, genotyping and other laboratory testing. Results of all testing were stored in a DNA data bank. As a sub study to LUCY, in 309 participants (patients and their relatives) of the LUCY study, radiation sensitivity was measured with the single cell gel electrophoresis (also known as *comet assay*) which quantifies the DNA fragments that have migrated out of the cell nucleus. In the LUCY comet assay, an experiment consists of a random sample of up to 200 cells from one blood sample. The cells are fixed in agarose gel and subjected to electrophoresis under alkaline conditions. In alkaline conditions, the DNA double helix is denatured and becomes single stranded. The application of electric current during electrophoresis enables the negatively charged DNA fragments to migrate away from the cell nucleus toward the anode and spread on the gel. Smaller fragments of damaged DNA migrate farther. When viewed under a fluorescent microscope, the DNA looks like a ”comet” with a bright head and a tail (figure 6.7). The length and intensity of the tail is directly proportional to the amount of DNA damage in the cell. The image is stained using a DNA-specific fluorescent stain and analysed using a microscope connected to a computer with image analysis software. In this project, the Comet Imager Metasystems (Altlussheim, Germany) was used. There are different measures of radiation sensitivity that can be generated from the comet assay. For this analysis, the Percent-DNA-in-Tail (%DT), Olive Tail Moment (OTM), and the difference in the Olive Tail Moments (DOTM) before and after exposure to 4 Gray radiation are used as quantitative measures of radiation sensitivity. The %DT is the relative amount of DNA in the tail or body of the comet which is measured by the image analyzer based on the optical intensity of the tail. The OTM is the product of the distance of DNA migration in the tail (see label in figure 6.7) and the fraction of total DNA in the tail. The formula for calculating the Olive Tail Moment is:

$$OTM = (|CG_T - CG_H|)(DNA_T/100) \tag{6.6}$$

where:
OTM - Olive Tail Moment
CG_T - center of gravity of the comet tail weighted by Gray values
CG_H - center of gravity of the comet head weighted by Gray values
DNA_T - amount of DNA in the tail

Using the OTM, one can also calculate the DOTM which is the difference between the Olive Tail Moments of cells exposed and unexposed to radiation. Olive Tail Moments and Percent-DNA-in-Tail are commonly recommended measures of DNA damage (Lovell and Omori, 2008).

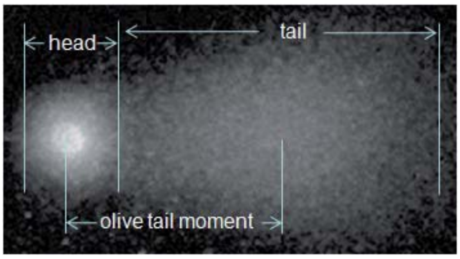


Figure 6.7: Comet assay
Source: Comet assay image is from Enciso et al. (2009)

This analysis focused on genes involved in DNA damage repair. Based on known or expected functional effects, 7 genes and SNPs were selected by experts in the LUCY study (see table 6.7 for list of genes and SNPs). XRCC1 and HOGG1 are *Base Excision Repair* (BER) pathway genes, while the others are *Non Homologous End Joining* (NHEJ) pathway genes. The chosen SNPs are all in Hardy-Weinberg equilibrium and are not in linkage disequilibrium with each other.

Table 6.7: Selected genes and SNPs for LUCY analysis

Gene	Structure	Location in the Chromosome	SNP* rs Number	Alleles	MAF*
XRCC1	17 Exons; 32.25 kb	19q13.2	rs1001581	C>T	0.42
XRCC4	8 Exons; 276.3 kb	5q13-q14	rs10040363	A>G	0.48
LigIV	2 Exons; 7.34 kb	13q33-q34	rs1151403	T>C	0.42
HOGG1	7 Exons; 8.13 kb	3p26.2	rs2072668	C>G	0.24
RAD50	25 Exons; 86.96 kb	5q31	rs2706348	G>A	0.22
MRE11	20 Exons; 76.57 kb	11q21	rs3017077	C>T	0.34
NBS1	16 Exons; 51.34 kb	8q21.3	rs709816	A>G	0.35

*SNP - Single nucleotide polymorphism; MAF - Minor Allele Frequency

Out of 795 lung cancer patients, 156 patients have radiation sensitivity data. Including the 153 relatives of the patients who have radiation sensitivity data, there were a total of 309 participants in the sub-study. However, from the 309 participants, only 123 individuals form complete family trios (patients with parents). This gives a total of only 41 family trios with complete phenotyping and genotyping data to determine if there are genetic main effects and epistatic effects of the SNPs of interest that contribute to the variation of the radiation sensitivity measures, i.e. %DT, OTM and DOTM. The type of lung cancer and other factors that may be related to radiation sensitivity were not included in the analysis.

6.3.2 Analysis method

The GQTDT method as described in the previous chapters was used to analyze the LUCY data. The statistical model used is:

$$Y_i = \beta_0 + \beta_M M_i + \beta_G G_i + \beta_H H_i + \beta_{GH} G_i H_i + \varepsilon_i \tag{6.7}$$

where:

Y_i	the random observation of a continuous quantitative response or phenotype (i.e. %DT, OTM or DOTM) of the ith patient; i = 1,...,n
M_i	explanatory variable representing the parental mating type
G_i	the genotype score of the study subject at SNP or locus 1
H_i	the genotype score of the study subject at SNP or locus 2
β_0	the intercept
β_M	regression coefficient for the parental mating type
$\beta_G, \beta_H, \beta_{GH}$	regression coefficients for the effects of locus 1, locus 2 and their interaction
ε_i	residual, $\sim N(0, \sigma^2)$

The genotype scores G_i and H_i is 0 for genotype code *’aa’*, 1 for *’AA’* and 0, 0.5 or 1 for *’Aa’* depending on the assumed model being recessive, additive or dominant, respectively.

The response variable Y_i in the statistical model is distributed as:

$$Y_i \sim f(g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p) \tag{6.8}$$

where:
f is the distribution of Y_i,
$(\theta_1, ..., \theta_p)$ are the parameters of f,
$g_1, ..., g_p$ are the link functions and
$t_1, ..., t_p$ are the model formulae for the explanatory terms in the predictors.

In the GQTDT analysis of the %DT and the OTM traits, the distribution fitted to the data is the inverse Gaussian (IG) distribution. The choice of the distribution was based on the generalized AIC criterion after fitting candidate distributions. For both %DT and OTM traits, the lowest AIC was noted with the use of inverse Gaussian distribution. The AIC for %DT using IG distribution is 212 and using lognormal the AIC is 223. For the OTM trait, the computed AIC using IG and lognormal distributions are 38 and 57, respectively. For the DOTM trait, the AICs are 152, 154 and 157 for the normal, lognormal and inverse Gaussian distributions, respectively. Therefore the normal distribution was used in the GQTDT analysis of the differences in Olive Tail Moments.

The generalized likelihood ratio test was used to test for the hypothesis of no genetic main effects and the hypothesis of no epistasis. The genetic main effect was tested using individual SNPs. This uses equation 6.7 without the explanatory variable H_i and the interaction term as a full model. The likelihood of this model with only one locus and the mating type was compared with the model with only the mating type to get the genetic main effect of a specific SNP. A pairwise analysis using two SNPs at a time was also done by comparing the model with the two SNP main effects (G_i and H_i) and mating type with the model without the two SNP main effects. For testing epistasis, the likelihood of the full model with the interaction of the two loci (equation 6.7) was compared against the null hypothesis model of no locus interaction. The GQTDT was applied to the data under three different genetic model assumptions, i.e. dominant, additive or recessive genetic model. The computed p-values were adjusted for multiple testing using Holm's (1979) procedure.

6.3.3 Results

The distributions of the three radiation sensitivity measures are shown in figure 6.8. It can be seen that the baseline %DT and OTM are severely skewed while the DOTM is approximately normally distributed.

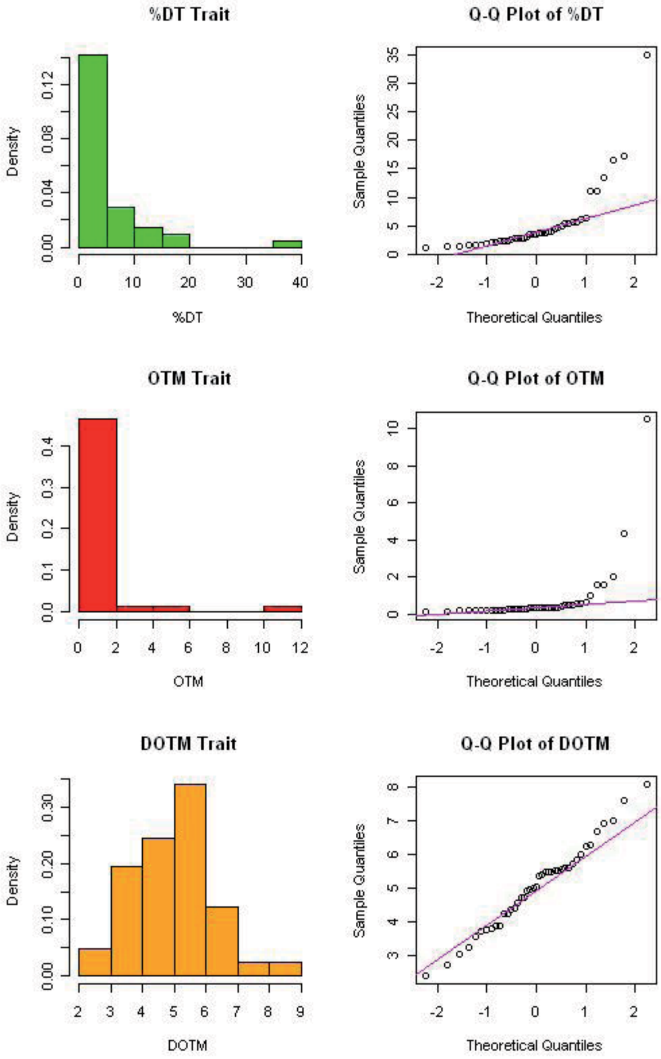


Figure 6.8: Distributions of radiation sensitivity measures, N=41

Individual SNP analysis showed different significant results in the three radiation sensitivity measures analyzed. In the analysis of the %DT, only the SNP rs1001581 showed significant ($p<0.05$) result under the dominant genetic model assumption. In the OTM trait, four SNPs showed significant genetic main effects (see table 6.8). The SNP rs1001581 which was found to affect %DT also affects the OTM trait. In the DOTM trait, none of the SNPs showed significant genetic main effects.

The results of the analysis for the %DT (table 6.9) show that some SNPs and their interactions have significant ($p<0.05$) effects on radiation sensitivity among the families investigated. Six interacting SNP pairs were detected out of the 21 pairs tested using GQTDT. Four of these six pairs show significant epistatic ef-

Table 6.8: Results of the analysis of genetic effect of individual SNP in the LUCY data

Trait	SNP	GQTDT p-values* under analysis genetic model		
		Dominant	Additive	Recessive
%DT	rs1001581 (XRCC1)	.006	n.s.	n.s.
	rs10040363 (XRCC4)	n.s.	n.s.	n.s.
	rs1151403 (LigIV)	n.s.	0.078	0.078
	rs2072668 (HOGG1)	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.
	rs3017077 (MRE11)	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	n.s.	n.s.
OTM	rs1001581 (XRCC1)	.003	.049	0.018
	rs10040363 (XRCC4)	n.s.	n.s.	n.s.
	rs1151403 (LigIV)	n.s.	0.036	0.036
	rs2072668 (HOGG1)	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	0.017	0.009	0.095
	rs3017077 (MRE11)	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	0.007	0.003
DOTM	rs1001581 (XRCC1)	n.s.	n.s.	n.s.
	rs10040363 (XRCC4)	n.s.	n.s.	n.s.
	rs1151403 (LigIV)	n.s.	n.s.	n.s.
	rs2072668 (HOGG1)	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	0.091	0.073	n.s.
	rs3017077 (MRE11)	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	n.s.	n.s.

*Listing only p-values $\leq$0.1, otherwise marking the test as 'n.s.' (not significant)

fects under the additive and recessive genetic model assumption and one SNP pair showed epistatic effect under the dominant genetic model. The interacting SNP pairs are from the following gene pairs: XRCC4-LigIV (rs10040363-rs1151403), XRCC4-MRE11 (rs10040363-rs3017077), XRCC4-NBS1 (rs10040363-rs709816), HOGG1-MRE11 (rs2072668-rs3017077), HOGG1-NBS1 (rs2072668-rs709816) and MRE11-NBS1 (rs3017077-rs709816). The XRCC4-NBS1 and HOGG1-MRE11 pairs showed epistasis under both additive and recessive genetic model. The other pairs showed epistasis only under one of the genetic model assumptions.

Genetic main effects are also noted in the pairwise analysis of SNPs. There are also SNP pairs which did not show interaction but showed genetic main effects such as those involving the XRCC1 gene. The results for testing genetic main effects using GQTDT also differs depending on the assumed genetic model. For example, in testing SNP pairs involving the gene XRCC1, most of the significant genetic main effects were noted when the assumed genetic model is dominant.

Table 6.9: Results of the analysis of SNP pairs and the Percent-DNA-in-Tail (%DT)

SNP (Gene) Pair tested		GQTDT p-values* under genetic model					
		Dominant		Additive		Recessive	
		main	inter.	main	inter.	main	inter.
rs1001581 (XRCC1)	rs10040363 (XRCC4)	0.007	n.s.	n.s.	n.s.	n.s.	n.s.
	rs1151403 (LigIV)	<0.001	n.s.	0.009	n.s.	n.s.	n.s.
	rs2072668 (HOGG1)	<0.001	-	0.013	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	<0.001	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	0.088	-	0.088	n.s.	n.s.	n.s.
	rs709816 (NBS1)	0.047	-	n.s.	0.089	n.s.	n.s.
rs10040363 (XRCC4)	rs1151403 (LigIV)	n.s.	0.006	n.s.	n.s.	n.s.	0.052
	rs2072668 (HOGG1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	<0.001	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	0.015	n.s.	0.019
rs1151403 (LigIV)	rs2072668 (HOGG1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
rs2072668 (HOGG1)	rs2706348 (RAD50)	n.s.	-	0.048	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	0.025	n.s.	0.026
	rs709816 (NBS1)	n.s.	-	n.s.	0.009	n.s.	n.s.
rs2706348 (RAD50)	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	n.s.	-
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	-
rs3017077 (MRE11)	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	0.043

*Listing only p-values ≤0.1, otherwise marking the test as 'n.s.' (not significant) or '-' where analysis was not possible; main = p-value for genetic main effect; inter. = p-value for interaction

For the OTM quantitative trait, four interacting SNP pairs were noted (see table 6.10). Two pairs showed epistasis in both additive and recessive genetic models while the other two pairs showed epistasis only under the recessive genetic model. The interacting SNPs involve the following genes: XRCC1-NBS1 (rs1001581-rs709816), LigIV-NBS1 (rs1151403-rs709816), HOGG1-MRE11 (rs2072668-rs3017077) and MRE11-NBS1 (rs3017077-rs709816). The last two interacting pairs are also included in the interacting pairs detected using the %DT quantitative trait.

Under different genetic model assumptions, different results also come up in testing for genetic main effects in the OTM trait. In the dominant model, three SNP pairs showed significant genetic main effects ($p<0.05$). Under the additive genetic model, two SNP pairs showed significant genetic main effects. A similar observation with the %DT analysis is that no genetic main effects showed up when testing under the recessive genetic model.

Table 6.10: Results of the analysis of SNP pairs and the Olive Tail Moment (OTM)

SNP (Gene) Pair tested		GQTDT p-values* under genetic model					
		Dominant		Additive		Recessive	
		main	inter.	main	inter.	main	inter.
rs1001581 (XRCC1)	rs10040363 (XRCC4)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	rs1151403 (LigIV)	0.003	n.s.	n.s.	n.s.	n.s.	n.s.
	rs2072668 (HOGG1)	<0.001	-	n.s.	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	<0.001	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	0.055	-	n.s.	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	0.002	n.s.	0.004
rs10040363 (XRCC4)	rs1151403 (LigIV)	0.072	n.s.	n.s.	n.s.	n.s.	n.s.
	rs2072668 (HOGG1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
rs1151403 (LigIV)	rs2072668 (HOGG1)	0.077	-	0.038	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	0.003
rs2072668 (HOGG1)	rs2706348 (RAD50)	0.056	-	0.011	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	0.029	n.s.	0.009
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
rs2706348 (RAD50)	rs3017077 (MRE11)	n.s.	-	0.094	n.s.	n.s.	-
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	-
rs3017077 (MRE11)	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	0.013

*Listing only p-values ≤0.1, otherwise marking the test as 'n.s.' (not significant) or '-' where analysis was not possible; main = p-value for genetic main effect; inter. = p-value for interaction

Among the three quantitative traits, the DOTM is the only one which is approximately normal in distribution. Using this quantitative trait, only one SNP pair showed epistasis in the analysis. The significant interaction was noted between SNPs rs2706348 (RAD50) and rs3017077 (MRE11). This finding is unique in the DOTM trait for this is not seen in the other two traits. Like in the %DT and OTM analysis, most of the tests for interaction in the DOTM trait under the dominant model were not analyzable. In contrast with the individual SNP analysis, the SNP pair analysis detected significant genetic main effects on the DOTM. There are more similarities in the results assuming dominant and additive genetic models. Under the recessive genetic model, there are more significant genetic main effects found. A total of eight SNP pairs have significant genetic main effects detected using the DOTM quantitative trait. Three of these are the same with the result in %DT and one is the same with the result in OTM. The pair rs2072668 (HOGG1) and rs2706348 (RAD50) was commonly identified to have genetic main effect in the three quantitative traits analyzed. With regard to epistasis, no SNP pair was commonly identified in the analysis of the three traits.

Table 6.11: Results of the analysis of SNP pairs and the Difference in Olive Tail Moments (DOTM)

SNP (Gene) Pair tested		GQTDT p-values* under genetic model					
		Dominant		Additive		Recessive	
		main	inter.	main	inter.	main	inter.
rs1001581 (XRCC1)	rs10040363 (XRCC4)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	rs1151403 (LigIV)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	rs2072668 (HOGG1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
rs10040363 (XRCC4)	rs1151403 (LigIV)	0.021	n.s.	<0.001	n.s.	0.012	n.s.
	rs2072668 (HOGG1)	n.s.	-	n.s.	n.s.	0.042	n.s.
	rs2706348 (RAD50)	0.008	n.s.	<0.001	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	0.072	0.018	n.s.
	rs709816 (NBS1)	n.s.	-	0.032	n.s.	0.015	n.s.
rs1151403 (LigIV)	rs2072668 (HOGG1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs2706348 (RAD50)	n.s.	n.s.	n.s.	n.s.	n.s.	-
	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	n.s.	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	0.060
rs2072668 (HOGG1)	rs2706348 (RAD50)	n.s.	-	n.s.	n.s.	0.021	-
	rs3017077 (MRE11)	n.s.	-	n.s.	n.s.	0.042	n.s.
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	n.s.
rs2706348 (RAD50)	rs3017077 (MRE11)	0.070	-	0.063	0.002	0.070	-
	rs709816 (NBS1)	n.s.	-	n.s.	n.s.	n.s.	-
rs3017077 (MRE11)	rs709816 (NBS1)	n.s.	-	n.s.	0.084	0.018	n.s.

*Listing only p-values ≤0.1, otherwise marking the test as 'n.s.' (not significant) or '-' where analysis was not possible; inter. = p-value for interaction; main = p-value for genetic main effect

6.3.4 Discussion

It is evident in this analysis that the use of different quantitative measures can lead to different conclusions about possible genetic factors that can affect radiation sensitivity. In detecting genetic main effects, the SNPs from the genes HOGG1 and RAD50 were commonly identified in the analysis of the three radiation sensitivity measures to have significant genetic main effects. It is not surprising to see a lot of significant genetic main effects in this analysis because the chosen candidate genes are the ones with proven biological functions related to DNA damage repair. The GQTDT method has proven its applicability in real data even if the sample size in the study is very limited.

With regard to epistasis, the results of the analysis differ depending on the radiation sensitivity measure or trait used. In total, there are nine SNP pairs which showed epistasis in this statistical analysis. Different SNP pairs showed epistasis in the %DT, OTM and DOTM traits. Table 6.12 summarizes the result of analyzing epistasis in the three radiation sensitivity measures.

Table 6.12: Summary of epistatic effects in the LUCY data

SNP (Gene) Pair		Radiation Sensitivity Measure			
		%DT	OTM	DOTM	
rs10040363 (XRCC4)	rs1151403 (LigIV)	E	N	N	*
rs10040363 (XRCC4)	rs3017077 (MRE11)	E	N	N	*
rs10040363 (XRCC4)	rs709816 (NBS1)	E	N	N	
rs2072668 (HOGG1)	rs3017077 (MRE11)	E	E	N	
rs2072668 (HOGG1)	rs709816 (NBS1)	E	N	N	
rs3017077 (MRE11)	rs709816 (NBS1)	E	E	N	*
rs1001581 (XRCC1)	rs709816 (NBS1)	N	E	N	
rs1151403 (LigIV)	rs709816 (NBS1)	N	E	N	
rs2706348 (RAD50)	rs3017077 (MRE11)	N	N	E	*

E - with epistasis; N - no epistasis detected in the GQTDT analysis;

*Reported in the literature with biological or functional interaction; In addition RAD50 and NBS and also XRCC1 and HOGG1 have been reported to have biological interactions but were not detected in this analysis.

Some of the SNP pairs with significant epistatic effect in this analysis were also found to be epistatic in biological studies. A recent article about a biological experiment on MRE11 by Xie et al. (2009) showed that MRE11 (Meiotic Recombination 11) has a crucial role in both XRCC4-dependent and XRCC4-independent non-homologous end joining (NHEJ) in mammalian cells. The interaction effects noted in the other genes are also supported by experimental data. MRE11, RAD50 and NBS1 were also reported to form a complex in response to double-strand chromosomal breaks (Lee and Paull, 2005). In the statistical analysis of the LUCY data, there was also significant interaction between rs3017077 (MRE11) and rs709816 (NBS1) and between rs3017077 (MRE11) and rs2706348 (RAD50) but not between rs2706348 (RAD50) and rs709816 (NBS1). Other known biological interactions that were detected in the statistical analysis of the LUCY data is the interaction between the DNA double strand break repair protein XRCC4 and DNA ligase IV which was demonstrated by several authors (Critchlow et al., 1997; Tseng et al., 2009). The epistasis between XRCC4 and DNA ligase IV was detected by the GQTDT in the %DT analysis. Some interactions (e.g. LigIV-NBS1 and HOGG1-MRE11) were detected in the analysis of the LUCY data but there is no publication yet of their biological interaction in the literature. On the other hand, the genes XRCC1 and HOGG1 have already been shown to interact both physically and functionally in biological experiments on DNA damage repair (Marsin et al., 2003) but their epistasis was not detected in the statistical analysis of the LUCY data.

The agreement of the results of the GQTDT with several studies and biological findings is a good measure of its applicability in real data. Although some results in this analysis have no existing publication yet to compare with, it does not necessarily mean a disadvantage of the statistical method. The two-SNP pairs that are known to have biological interaction but were not detected by the GQTDT method could be attributed to factors such as effect size and sample size. In the simulation studies done in chapter 5, one saw the improvement in the power of the test at larger sample sizes and in situations when the genetic factor being measured contributes a bigger effect on the response variable.

The use of a statistical test such as the GQTDT in the analysis of genetic main effects and epistasis can give important clues on biological mechanisms of genetic factors affecting diseases. The results of statistical tests can also generate hypothesis that scientists might wish to pursue in the laboratory. It can also become a basis for objective decision-making with respect to disease management and treatment. In the LUCY study for example, one can develop a criteria which includes significant genes to guide the doctors whether to recommend radiation therapy or not to a patient. In the absence of biological evidence, a statistical evidence about radiation sensitivity can be used as a rough basis.

In the LUCY study, there is still a question which radiation sensitivity measure is best to use to determine genetic factors related to DNA damage repair. Although both OTM and %DT are commonly recommended end points in comet assay, some prefer the %DT more because of its advantage of being 'standardized' over different studies. Still others suggest that comparisons of results with different comet assay end points is useful (Lovell and Omori, 2008). The choice of the response variable is important as it can lead to different conclusions in the end. As seen in the result of the LUCY analysis, the three radiation sensitivity measures detected different SNP pairs with genetic main effects and epistatic effect. As to which response variable should be labelled as most appropriate in measuring radiation sensitivity is beyond the scope of this analysis.

(*Data Acknowledgement: The author wishes to acknowledge the GAW, FHS and LUCY study. The GAW was supported by grant R01 GM031575. The generation of the GAW16 simulated data was carried out in the Division of Statistical Genomics at Washington University School of Medicine and supported by NIH grants HL08770003, HL08768803, 1RR02499203, DK06833603, and HL08821502. The Framingham Heart Study research was supported by NHLBI Contract: 2 N01-HC-25195-06 and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). The author is grateful to the many investigators within the Framingham Heart Study who have collected and managed the data and especially to the participants for their invaluable time, patience, and dedication to the Study. The LUCY study is funded by the Deutsche Forschungsgemeinschaft and the Helmholtzgemeinschaft. In particular, the radiation sensitivity data for LUCY was funded by the Bundesamt für Strahlenschutz (BfS-Auftrag Nr. StSch 4454). Genotyping for LUCY was done at the Genome Analysis Center of the GSF Research Center for Environment and Health. The author is also grateful to the investigators and all participants of the LUCY study.*)

7 Summary and Outlook

Previous studies in the literature and simulation studies in this thesis have shown the difficulties and issues involved in the statistical analysis of genetic main effects and epistasis in family-based studies. The existing Transmission Disequilibrium Test and other related tests have made good contributions to the body of knowledge available for analyzing genetic effects in family data. However, there is still a need for improved methods that can be applied specifically for quantitative traits. The proposed Generalized Quantitative Transmission Disequilibrium Test (GQTDT) offers several advantages in analyzing genetic effects given different distributions of the quantitative response variable. Being a generalized method, it encompasses other existing methods because of its flexibility. For example, the QTDT_M by Gauderman (2003) would give the same result as the GQTDT if normal distribution parameter links are specified in the GQTDT analysis. The power of the GQTDT can be influenced by several factors such as the minor allele frequencies of the loci or genes being investigated, the analysis genetic model, the fit of the distribution used in the analysis and the sample size.

Higher power of the GQTDT was observed in the simulation scenarios with higher minor allele frequencies (MAF). This effect of the MAF on power is not unique to the GQTDT. It can also be seen in population-based studies. Data with minor allele frequencies close to 0.50 have the advantage of having enough individuals in the genotype groups in a reasonably-sized study. This makes it easier for the test to make comparisons. In data where the minor allele frequencies are small or rare, some genotypes also appear in small frequencies or may not appear at all in the dataset being analyzed. Thus, causing decreased power and sometimes problems in convergence of the test statistic. Slightly higher type I error in detecting genetic main effects and epistasis were noted more often in some simulation scenarios involving PopB datasets which has the highest minor allele frequencies among the three simulated populations in chapter 5. This is however cannot be generalized for all the simulation scenarios.

The assumed genetic model in the analysis is also another factor that affects the power of the GQTDT and similar tests like the QTDT_M where assigning a genotype score depends on the assumed genetic model. Higher power is observed when the analysis genetic model is the same as the "true" genetic model of the gene in the data. In testing two genes at a time, good power is achieved in determining genetic main effects when at least one of the genes has a correctly assumed genetic model. In the case of a recessive genetic model, lower power in detecting genetic effects is expected, especially when the minor allele frequencies of the genes are also low. In a study by Lettre et al. (2007) on genetic model testing and power of population-based association studies in quantitative traits, a similar observation was reported. Even if the study uses unrelated individuals and different statistical methodology, the result showed a similar result with the GQTDT that maximal power is achieved when one uses a genetic model that matches the actual underlying mode of inheritance of the gene.

The use of a fitted distribution gives the GQTDT higher power in detecting genetic main effects and epistasis compared to linear regression based method like the $\mathrm{QTDT}_{\mathrm{M}}$. The higher power in statistical testing is more noticeable in detecting genetic main effects than in detecting epistatic effects. A major advantage of the GQTDT is its flexibility to model additional environmental covariates and to adopt different types of quantitative continuous distributions. The GQTDT model can also be extended and modified to analyze response variables with mixed distributions and also for handling non-continuous or discrete variables (e.g. Poisson distributed variables).

In general, the power of the GQTDT increases with increased sample size. What is not investigated in the simulation studies are scenarios of very small sample sizes. Based on the sample size recommendation in the study of Gauderman (2003), a sample size of less than 100 family trios may be used to investigate genetic main effects or epistasis in certain situations. However, there are no published studies yet investigating the effect of genes on quantitative traits in line with this type of statistical method using only 20 or less family trios.

A common advantage of the GQTDT with other family-based analysis is its robustness to population stratification. In terms of power and type I error, the simulation results of testing in mixed populations show the same trend as that of the homogenous populations. In using a regular linear regression without mating type indicator, the type I error can go beyond 50% when there is population stratification. This has been clearly shown in the study of Gauderman (2003). In using the GQTDT, slightly elevated type I error (up to 9%) was seen in cases of analyzing simulated skewed quantitative traits. The result is similarly noted in the PopMix population and in the homogenous PopA and PopB populations, indicating that this is not peculiar to datasets with population stratification.

The performance of the GQTDT in determining genetic main effects is satisfactory both in the normally distributed and skewed quantitative traits. When a fitted distribution is specified in the analysis, higher power can be achieved. In terms of detecting epistasis, good power is noted when the distribution of the quantitative trait is normal, but in case of skewed traits the observed power is not as high as that observed in the normal trait. The measurement of epistasis both biologically and statistically is not an easy task. The different definitions of the term alone create confusions on how it can be detected in practice. For statistical testing, we are limited on available mathematical models in detecting epistasis in genetic data. The way of defining the model and choosing the correct distribution in using the GQTDT can affect its power and accuracy. Fitting a distribution that is very inappropriate (e.g. using a normal distribution for skewed data) can elevate type I errors and reduce the power of the test.

In this thesis, the application of the GQTDT in real data shows the performance of the method in realistic scenarios. In the Framingham Heart Study, the method was able to detect genetic main effects associated with body-mass-index which were also detected by other methods. However, no significant interaction was detected in the data. In the simulated data which were also based on the real Framingham data, epistatic effects were detected by the method in the two SNP pairs analyzed. The power of the GQTDT in this case is again much higher compared to using $\mathrm{QTDT_M}$. In the lung cancer study, the GQTDT was also able to determine known genetic main effects and epistasis as proven by the agreement of the results with biological studies and experiments. Not all known genetic effects were captured by the test in the LUCY data, but it has performed quite well given a data with only 41 family trios.

The GQTDT has been compared in this thesis with the $\mathrm{QTDT_M}$ which has been a benchmark method in determining genetic effects in quantitative traits. The GQTDT has shown advantages over the $\mathrm{QTDT_M}$ especially in dealing with skewed quantitative traits. Being a generalized method, the GQTDT transcends the limitations of the $\mathrm{QTDT_M}$ when dealing with nonnormally distributed quantitative traits. The method can also be extended to accommodate more complicated statistical models.

In comparison with population-based methods, the GQTDT has an advantage when population stratification is present in the data. It also doesn't need more complicated models and longitudinal data as used in the methods of Malzahn et al. (2009). The method of Malzahn and colleagues showed higher power in detecting epistasis but also needed more information (e.g. follow-up measurements) which may not be always easy to collect and requires longer study period. In addition, the methods used by Malzahn et al. (2009) have not been tested for effects of population stratification and it cannot be directly compared with the GQTDT method because it is not a family-based method.

In the simulation studies done here, ideal situations have been assumed where the complete family data is available and the loci involved are not linked or in linkage disequilibrium with each other. The proposed GQTDT may be extended to accommodate more than two candidate loci and other environmental factors. In some instances, several marker loci being in linkage disequilibrium with each other are genotyped on the same gene. It would be a challenge to extend the GQTDT by considering possible linkage disequilibrium between genetic markers. It is also an open task to evaluate epistasis between two or more genes when several SNPs in each gene are involved. It may be straightforward to test which SNPs are interacting, but if one wants to determine which genes consisting of several SNPs are interacting, further evaluation method is needed. Moreover, this thesis focused on statistical hypothesis testing but did not consider effect estimation. The GQTDT method has been applied to few nonnormal distributions in this study but it cannot generalize about the performance of the method for all types of parametric and also semi-parametric distributions. The problem of missing family data would also be interesting to pursue using the method.

APPENDIX

Notations

The summary of notations used in this thesis is listed below in the order that they appeared in the text. The symbols introduced in one chapter is used throughout the whole thesis unless otherwise specified.

Chapter 2

X	independent random variable
(A_r,A_s)	ordered pair of alleles at a given gene or locus where $r,s = 1,\ldots,k$
(B_r,B_s)	ordered pair of alleles at another locus where $r,s = 1,\ldots,m$
p_r,p_s	frequencies of alleles for A_r and A_s, respectively, where $r,s = 1,\ldots,k$
q_r,q_s	frequencies of alleles for B_r and B_s, respectively, where $r,s = 1,\ldots,m$
$f_{A_rA_s}$	penetrance of genotype A_rA_s where $r,s = 1,\ldots,k$
Y	the outcome variable (e.g. affection status or quantitative trait)
μ	population mean
β_G	regression coefficient for the genotypic effects or covariate G
G	genotype or covariate that quantifies the genotype at a locus
G_i	genotype or covariate that quantifies the genotype of the ith individual
σ^2	residual variance
H	polygenic effect due to a large number of small additive genetic factors
E	environmental effect which also includes the error term
d	degree of dominance
t	measure of displacement at the major locus
σ_G^2	variance of the effect of the major locus G
μ_0, μ_1, μ_2	mean effects of genotypes A_2A_2, A_1A_2 and A_1A_1
σ_H^2	variance of the polygenic effect H
σ_E^2	variance of the environmental effect E
σ_c^2	common environmental component of the variance
σ_r^2	random component of the environmental variance
M	number of affected offspring
n	total number of offsprings

P_D	probability of being affected by a disease
N	total number of sibship or families
g_i	represents all possible genotypes of the ith individual
G_{if}, G_{im}	genotypes of parents of the ith individual
n_t	total number of individuals in a pedigree
n_1	total number of founders in a pedigree
n_2	total number of non-founders in a pedigree
φ	transmission parameter
θ	recombination fraction
u	number of recombinants
v	number of non-recombinants
h_0, h_1	haplotype frequencies of the current and next generation, respectively
δ_{ij}	coefficient of linkage disequilibrium of alleles A_i and B_j
Q	unobservable indicator of subpopulation
D	the presence or absence of the disease
Δ	the case-control effect
g_{ij}	genotype values where *i,j* = 0,1 or 2 refers to the number of susceptibility alleles at the first and second locus respectively
a_i	additive effects at locus i
d_i	dominance effects at locus i
i_{aa}	additive x additive epistatic effects
i_{ad}	additive x dominance epistatic effects
i_{da}	dominance x additive epistatic effects
i_{dd}	dominance x dominance epistatic effects
w_i, z_i	dummy variables for the genotype at locus i
β_{1i}	regression coefficient for the effect associated with having i susceptibility alleles at the first locus, where i = 0,1,2
β_{2j}	regression coefficient for the effect associated with having j susceptibility alleles at the second locus, where j = 0,1,2

Chapter 3

N_{ij}	number of parents who transmitted i allele and did not transmit j allele
δ	coefficient of linkage disequilibrium

χ^2_{tdt}	TDT Chi-square statistic
Z_U	upper cut-off of on the quantitative phenotype
Z_L	lower cut-off of on the quantitative phenotype
T	absence or presence of transmission of susceptibility allele
X	number of susceptibility alleles
μ_T	mean transmission of susceptibility allele
$\mu_{A_rA_s}$	mean quantitative phenotype of genotype A_rA_s, where r,s = 1,...,k
R^2	regression coefficient
Y'	observed trait
x	minimal sufficient statistic
β_X	regression coefficient for the number of susceptibility alleles
L_{ij}	link function
S, $\mathbf{S}$	score statistic and vector of score statistics
m_s	number of score statistics
$\mathbf{V}$	vector of variances of score statistics
E	covariates affecting the phenotypic trait
β_E	regression coefficient for the covariate E
β_{XE}	regression coefficient for the interaction of the genotype X and covariate E
$\mathbf{\Omega}_i$	variance-covariance matrix for family i
π_{ijk}	proportion of alleles shared IBD between siblings j and k in family i
σ^2_a	additive genetic variance of the major gene
σ^2_s	residual sibling resemblance
σ^2_e	residual environmental variance component
β_a	regression coefficient of the additive genetic effect
b	between-family component
w	within-family component
β_b	regression coefficient of the between-family component
β_w	regression coefficient of the within-family component
ε_i	residual, $\sim N(0, \sigma^2)$
H_i	genotype of the ith individual at another locus
β_H	regression coefficient of covariate H

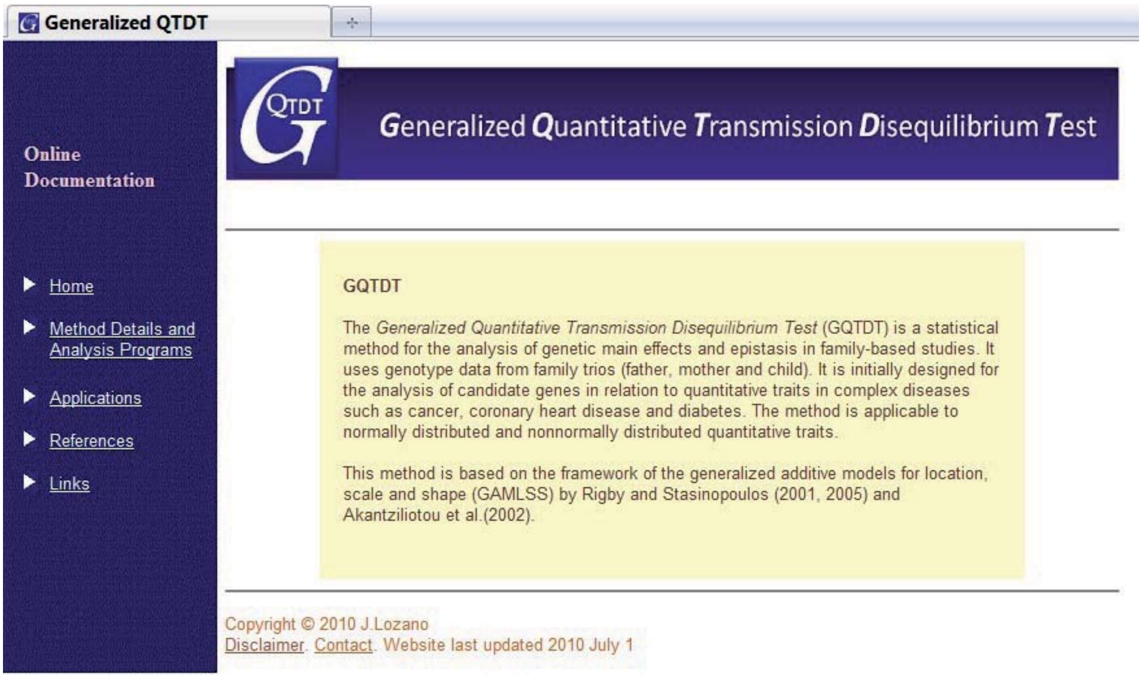
β_{GH}	regression coefficient of the interaction of covariates G and H
α_M	parental mating type specific intercept
S_i	minimal sufficient score statistic for the the ith family
$\mathbf{V}_i$	n_i x n_i variance matrix; n_i is the number of offsprings in the ith family
$\hat{U}$	a score statistic
O_M	set of possible offspring genotypes for a mating type M
$\sum_{X_i^*}$	summation over all possible offspring genotypes
$\sum_{X_i^* \in O'_M}$	all possible offspring genotypes that could have been transmitted given the parental genotypes

Chapter 4

Y_i^t	transformed value of the phenotype for the ith subject
Φ^{-1}	standard normal quantile (or probit) function
η	linear predictor
Θ	canonical parameter representing the location in the exponential distribution
Ψ	dispersion parameter representing the scale in the exponential distribution
μ	the location parameter
σ	the scale parameter
ν	the skewness parameter
τ	the kurtosis parameter
l	log-likelihood function of the data
z	transformed variable
π_i^*	probability for a non-zero Y_i
$\mathbf{X}$, $\mathbf{Z}$	design matrices
β	linear parameter
γ	random effect
λ	hyperparameter

GQTDT online documentation and programs

The description of the GQTDT method, R programs and applications are available online in the website http://gqtdt-statistics.int-org.co.cc


Generalized QTDT
QTDT
Generalized Quantitative Transmission Disequilibrium Test
Online Documentation
Home
Method Details and Analysis Programs
Applications
References
Links

GQTDT

The *Generalized Quantitative Transmission Disequilibrium Test* (GQTDT) is a statistical method for the analysis of genetic main effects and epistasis in family-based studies. It uses genotype data from family trios (father, mother and child). It is initially designed for the analysis of candidate genes in relation to quantitative traits in complex diseases such as cancer, coronary heart disease and diabetes. The method is applicable to normally distributed and nonnormally distributed quantitative traits.

This method is based on the framework of the generalized additive models for location, scale and shape (GAMLSS) by Rigby and Stasinopoulos (2001, 2005) and Akantziliotou et al.(2002).

Copyright © 2010 J.Lozano
Disclaimer. Contact. Website last updated 2010 July 1

References

Abecasis G, Cardon L, Cookson W. 2000. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279-292.

Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte Jr M, Detrano R. 1990. Quantification of coronary artery calcium using ultrafast computed tomography. *J Am Coll Cardiol* 15:827-832.

Akantziliotou K, Rigby RA and Stasinopoulos DM. 2002. The R implementation of Generalized Additive Models for Location, Scale and Shape. In: Stasinopoulos M and Touloumi G (eds.), *Statistical modelling in Society: Proceedings of the 17th International Workshop on statistical modelling* pp. 75-83. Chania, Greece.

Akritas MG. 1990. The rank transform method on some two factor designs. *J Am Stat Assoc* 85:73-78.

Allison DB. 1997. Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 60:676-690.

Almasy L, Blangero J. 1998. Multipoint quantitatitve-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198-1211.

Alvarez-Castro JM, Carlborg O. 2007. A unified model for functional and statistical epistasis and its application in quantitative trait Loci analysis. *Genetics* 176(2):1151-67. Epub 2007 Apr 3.

An P, Mukherjee O, Chanda P, Yao L, Engelman C, Huang C, Zheng T, Kovac IP, Dubé M, Liang X, Li J,8 de Andrade M, Culverhouse R, Malzahn D, Manning AK, Clarke GM, Jung J, Province MA. 2009. The challenge of detecting epistasis (GxG interactions): Genetic Analysis Workshop 16. *Genet Epidemiol* 33 Suppl 1:S58-67.

Andreasen CH, Mogensen MS, Borch-Johnsen K, Sandbaek A, Lauritzen T, Sϕrensen TI, Hansen L, Almind K, Jϕrgensen T, Pedersen O, Hansen T. 2008. Non-replication of genome-wide based associations between common variants in INSIG2 and PFKP and obesity in studies of 18,014 Danes. *PLoS One* 6;3(8):e2872.

Balding DJ, Bishop M, Cannings C (eds). 2001. *Handbook of Statistical Genetics.* 2nd Ed. Wiley, Chichester, UK.

Barber JA, Thompson SG. 2000. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Stat Med* 19:3219-3236.

Bartel PL, Fields S (eds). 1997. *The Yeast Two-Hybrid System.* Oxford University Press, Inc., UK.

Barton NH, Keightley PD. 2002. Understanding quantitative genetic variation. *Nat Rev Genet* 3:11-21.

Bateson W. 1909. *Mendel's Principles of Heredity.* Cambridge University Press, Cambridge.

Beasley TM. 2002. Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multiv Behav Res* 37:197-226.

Beasley TM, Erickson S, Allison DB. 2009. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet* 39(5):580-95. Epub 2009 Jun 14.

Bickeböller H, Clerget-Darpoux F. 1995. Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epi* 12(6):865-870.

Bickeböller H, Fischer C. 2007. *Einführung in die Genetische Epidemiologie.* Springer, Berlin.

Blair RC, Sawilowsky SS, Higgins JJ. 1987. Limitations of the rank transform statistic in test for interactions. *Comm Stat-Simul Comp* 16(4):1133-1145.

Blom G. 1958. *Statistical estimates and transformed beta-variables.* Wiley, New York.

Bolton-Maggs PH, Pasi KJ. 2003. Haemophilias A and B. *Lancet* 24;361(9371):1801-9.

Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM. 2004. Protein interaction networks from yeast to human. *Curr Opin Struct Biol* 14:292-299.

Box GEP, Cox DR. 1964. An analysis of transformations. *J R Stat Soc B* 26:211-252.

Box GEP, Tiao GC. 1973. *Bayesian Inference in Statistical Analysis.* New York: Wiley.

Bradley JV. 1958. *Distribution-free Statistical Tests.* Prentice-Hall, New Jersey.

Carlborg Ö, Haley C. 2004. Epistasis: too often neglected in complex traits studies? *Nature Reviews Genetics* 5:618-625.

Cheverud JM, Routman EJ. 1995. Epistasis and its contribution to genetic variance components. *Genetics* 139:1455-1461.

Cockerham CC. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39:859-882.

Cole TJ, Green PJ. 1992. Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat Med* 11:1305-1319.

Conover WJ, Iman RL. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *Am Stat* 35: 124-133.

Cook NR, Zee RY, Ridker PM. 2004. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med* 23:1439-1453.

Cordell HJ. 2002. Epistasis: what it means, what it doesn't mean, and statistical models to detect it in humans. *Hum Mol Genet* 11:2463-2468.

Cordell HJ, Barratt BJ, Clayton DG. 2004. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epi* 26:167-185.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* 70:124-141.

Critchlow SE, Bowater RP, Jackson SP. 1997. Mammalian DNA double-strand break repair protein XRCC4 interacts with DNA ligase IV. *Curr Biol* 7(8):588-98.

Curtis D. 1997. Use of siblings as controls in case-control association studies. *Ann Hum Genet* 61:319-333.

Devlin B, Roeder K, Bacanu SA. 2001. Unbiased methods for population-based association studies. *Genet Epi* 21(4):273-284.

Devlin B, Roeder K, Wasserman L. 2001. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 60(3):155-166.

Draganov DI, Teiber JF, Speelman A, Osawa Y, Sunahara R, La Du BN. 2005. Human paraoxonases (PON1, PON2, and PON3) are lactonases with overlapping and distinct substrate specificities. *J Lipid Res* 46: 1239-1247.

Dudbridge F, Koeleman BPC, Todd JA, Clayton DG. 2000. Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet* 66:2009-2012.

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Stat* 7:1-26.

Elston RC, Stewart J. 1971. A general model for genetic analysis of pedigree data. *Hum Hered* 21:523-542.

Enciso M, Sarasa J, Agarwal A, Fernandez JL, Gosalvez J. 2009. A two-tailed Comet assay for assessing DNA damage in spermatozoa. *Reprod Biomed Online* 18(5):609-16.

Ewens WJ, Spielman RS. 1995. The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet* 57:455-464.

Ewens WJ, Spielman RS. 2004. The TDT is a statistically valid test: Comments on Wittkowski and Liu. *Human Hered* 58:59-60.

Ewens WJ, Spielman, RS. 2005. What is the significance of a significant TDT? *Hum Hered* 60:206-210.

Falconer DS. 1989. *Introduction to Quantitative Genetics*, Ed. 3. Longman, New York.

Field LL. 1989. Genes predisposing to IDDM in multiplex families. *Genet Epi* 6:101-106.

Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Bruning JC, Ruther U. 2009. Inactivation of the Fto gene protects from obesity. *Nature* 458: 894-898.

Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburg* 52:399-433.

Flint J, Mott R. 2001. Finding the molecular basis of quantitative traits: successes and pitfalls. *Nat Rev Genet* 2:437-445.

Frankel WN, Schork NJ. 1996. Who's afraid of epistasis? *Nat Genet* 14:371-373.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 36:388-393.

French JL. 2004. Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics* 5(2):177-191(15).

Fulker D, Cherny S, Sham P, Hewitt J. 1999. Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64:259-67.

Gauderman WJ. 2003. Candidate gene association analysis for quantitative trait using parent-offspring trios. *Genet Epi* 25:327-338.

Giblett ER. 1969. Genetic Markers in Human Blood. Philadelphia: FA Davis Co.

Gordon D, Heath SC, Liu X, Ott J. 2001. A transmission/ disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 69(2):371-380.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. 1990. Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21. *Science* New Series, 250(4988):1684-1689.

Hallgrímsdóttir IB, Yuster DS. 2008. A complete classification of epistatic two-locus models. *BMC Genetics* 9:17.

Hartl DL, Clark AG. 1997. *Principles of Population Genetics.* 3rd Ed. Sinauer Associates, Inc., Sunderland, Massachusetts.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3-19.

Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models.* Chapman and Hall, London.

Headrick TC, Rotou O. 2001. An investigation of the rank transformation in multiple regression. *Comput Stat Data Anal* 38:203-215.

Headrick TC, Sawilowsky SS. 2000. Properties of the rank transformation in factorial analysis of covariance. *Comm Stat-Simul Comp* 29:1059-1087.

Headrick TC, Vineyard G. 2001. An empirical investigation of four tests of interaction in the context of factorial analysis of covaraince. *Mult Linear Regress View* 27:3-15.

Heller G, Stasinopoulos M, Rigby RA. 2006. The zero-adjusted inverse Gaussian distribution as a model for insurance claims. *Proceedings of the 21th International Workshop on Statistial Modelling*, eds J. Hinde, J. Einbeck and J. Newell, pp 226-233, Galway, Ireland.

Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand J Statist* 6: 65-70.

Hora SC, Conover WJ. 1984. The F-statistic in the two-way layout with rank-score transformed data. *J Am Stat Assoc* 79:668-673.

Horvath S, Xu X, Laird NM. 2001. The family based association test method: strategies for studying general genotype-phenotype associations. *Eur J Hum Genet* 9:301-306.

Hunter DJ. 2005. Gene-environment interactions in human diseases. *Nat Rev Genet* 6(4):287-98.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome, *Nature* 431:931-945.

Jin K, Speed TP, Klitz W, Thomson G. 1994. Testing for segregation distortion in the HLA complex. *Biometrics* 50:1189-1198.

Kaplan NL, Martin ER, Weir BS. 1997. Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet* 60:691-702.

Khoury MJ, Beaty TH, Cohen, BH. 1993. *Fundamentals of Genetic Epidemiology.* Oxford University Press, New York.

Kistner EO, Weinberg CR. 2004. Method for using complete and incomplete trios to identify genes related to a quantitative trait. *Genet Epi* 27:33-42.

Kistner EO, Weinberg CR. 2005. A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. *Genet Epid* 29:155-165.

Knapp M. 1999. The transmission/disequilibrium test and parental genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test. *Am J Hum Genet* 64:861-870.

Kotti S, Bickeböller H, Clerget-Darpoux F. 2007. Strategy for detecting susceptibility genes with weak or no marginal effect. *Hum Hered* 63:85-92.

Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA, Borecki IB. 2009. The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. *BMC Proceedings* 3(Suppl 7):S4.

Kroymann J, Mitchell-Olds T. 2005. Epistasis and balanced polymorphism influencing complex trait variation. *Nature* 435:95-98.

Laird NM. 2007. Family-Based Association Tests and the FBAT-toolkit : Users' Manual. Available from: `http://www.biostat.harvard.edu/~fbat/fbat.htm`.

Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family-based tests of association. *Genet Epi[Suppl]* 19:36-42.

Laird NM, Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7(5):385-94.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 84:2363-2367.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. *Science* 265:2037-2048.

Lange C, DeMeo D, Laird N. 2002. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet* 71:1330-41.

Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. 2004. PBAT: Tools for Family-Based Association Studies. *Am J Hum Genet* 74(2):367-369.

Lee JH, Paull TT. 2005. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science* 308:551-553.

Lettre G, Lange C, Hirschhorn JN. 2007. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epi* 31(4):358-362.

Lewinger JP, Bull SB. 2006. Validity, efficiency, and robustness of a family-based test of association. *Genet Epi* 30:62-76.

Lewis CM. 2002. Genetic association studies: design, analysis and interpretation. *Brief Bioinform* 3(2):146-53.

Li B, Leal SM. 2009. Deviations from Hardy-Weinberg Equilibrium in Parental and Unaffected Sibling Genotype Data. *Hum Hered* 67:104-115.

Li H, Fan J. 2000. A general test of association for complex diseases with variable age at onset. *Genet Epi [Suppl]* 19:43-49.

Li H, Gao G, Li J, Page GP, Zhang K. 2007. Detecting epistatic interactions contributing to human gene expression using the CEPH family data. *BMC Proceedings* 1(Suppl 1):S67.

Li M, Jiang L, Song Y, Sham PC. 2008. Power of Transmission / Disequilibrium Tests in Admixed Populations. *Genet Epi* 32: 434-444.

Li W, Reich J. 2000. A complete enumeration and classification of two-locus disease models. *Hum Hered* 50:334-349.

Liang X, Gao Y, Lam TK, Li Q, Falk C, Yang XR, Goldstein AM, Goldin LR. 2009. Identifying Rheumatoid Arthritis susceptibility genes using high-dimensional method. *BMC Proceedings* 3(Suppl 7):S79.

Liu Y, Tritchler D, Bull SB. 2002. A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. *Genet Epi [Suppl]* 22:26-40.

Lou X, Chen G, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD. 2008. A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am J Hum Genet* 83:457-467.

Lovell DP, Omori T. 2008. Statistical issues in the use of the comet assay. *Mutagenesis* 23(3):171-182.

Lu AT, Cantor RM. 2007. Weighted variance FBAT: a powerful method for including covariates in FBAT analyses. *Genet Epi* 31:327-337.

Lunetta K, Faraone S, Biederman J, Laird N. 2000. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet* 66:605-14.

Lurie G, Wilkens LR, Thompson PJ, McDuffie KE, Carney ME, Terada KY, Goodman MT. 2008. Genetic polymorphisms in the Paraoxonase 1 gene and risk of ovarian epithelial carcinoma. *Cancer Epidemiol Biomarkers Prev* 7(8):2070-7.

Malzahn D, Balavarca Y, Lozano JP, Bickeböller H. 2009. Tests for candidate gene interaction for longitudinal quantitative traits measured in a large cohort. *BMC Proceedings 2009*, 3(Suppl 7): S80.

Mansouri H, Chang G-H. 1995. A comparative study of some rank tests for interaction. *Comput Stat Data Anal* 19:85-96.

Marsin S, Vidal AE, Sossou M, Menissier-de Murcia J, Le Page F, Boiteux S, de Murcia G, Racidella JP. 2003. Role of XRCC1 in the coordination and stimulation of oxidative DNA damage repair initiated by the DNA glycosylase hOGG1. *J Biol Chem* 278(45):44068-74. Epub 2003 Aug 21.

McKeigue PM. 1997. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am J Hum Genet* 60:188-196.

McNemar, Q. 1947.”Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12 (2): 153157.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. 2006. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* Jan;78(1):15-27. Erratum in: *Am J Hum Genet* 2009 Feb;84(2):301.

Millstein J, Siegmund KD, Conti DV, Gauderman WJ. 2005. Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis. *BMC Genet* 30;6 Suppl 1:S147.

Mitchell BD, Ghosh S, Schneider JL, Birznieks G, Blangero J. 1997. Power of variance component linkage analysis to detect epistasis. *Genet Epi* 14:1017-1022.

Monks SA, Kaplan NL. 2000. Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *Am J Hum Genet* 66:576-592.

Moore JH. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56:73-82.

Moore JH, William SM. 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 27:637-646.

Morton NE. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* 7(3):277-318.

Morton NE, MacLean CJ. 1974. Analysis of family resemblance. 3. Complex segregation of quantitative traits. *Am J Hum Genet* 26(4):489-503.

Nagel RL. 2001. Pleiotropic and epistatic effects in sickle cell anemia. *Curr Opin Hematol* 8:105-110.

NCBI (National Center for Biotechnology Information). U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda MD, 20894 USA. 2010. Available from: `http://www.ncbi.nlm.nih.gov/`.

Nelder JA, Wedderburn RWM. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society A*, 135, 370-384.

Nelson DB. 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59: 347-370.

Nelson MR, Kardia SL, Ferrell RE, Sing CF. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 11:458-470.

Neuman RJ, Rice JP. 1992. Two-locus models of disease. *Genet Epi* 9:347-365.

Newman B, Austin MA, Lee M, King MC. 1988. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc Natl Acad Sci USA* 85(9):3044-8.

NLM (National Library of Medicine, U.S.), NCBI; 18 Oct 2008 [cited 24 Apr 2009]. Genes and disease [internet]. Bethesda (MD). Available from: `http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/gnd/tocstatic.html`.

Oetiker T, Partl H, Hyna I, Schlegl E. 2005. The not so short introduction to $\LaTeX2_\varepsilon$. Version 4.17. `http://www.ctan.org/tex-archive/info/lshort/english/`.

Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). 23 Jun 2009. `http://www.ncbi.nlm.nih.gov/omim/`.

Ott J. 1989. Statistical properties of the haplotype relative risk. *Genet Epi* 6:127-130.

Phillips PC. 1998. The language of gene interaction. *Genetics* 149:1167-1171.

Purcell S, Sham PC. 2004. Epistasis in quantitative trait locus linkage analysis: interaction or main effect. *Behav Genet* 34:143-152.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.

Rabinowitz D. 1997. A transmission disequilibrium test for quantitative trait loci. *Hum Hered* 47:342-350.

Rabinowitz D, Laird NM. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 50:211-223.

Rice JP, Neuman RJ, Hoshaw SL, Daw EW, Gu G. 1995. TDT with covariates and genome screens with MOD scores: Their behavior on simulated data. *Genet Epi* 12:659-664.

Rigby RA, Stasinopoulos DM. 1996. A semi-parametric additive model for variance heterogeneity. *Statistal Computing* 6:57-65.

Rigby RA, Stasinopoulos DM. 2001. The GAMLSS project: a flexible approach to statistical modelling. In: Klein B and Korsholm L (eds.), *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling* pp. 249-256. Odense, Denmark.

Rigby RA, Stasinopoulos DM. 2004. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Stat Med* 23:3053-3076.

Rigby RA, Stasinopoulos DM. 2005. Generalized additive models for location, scale and shape. *Appl Stat* 54:507-554.

Rigby RA, Stasinopoulos DM. 2006. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Stat Modelling* 6;209.

Rigby RA, Stasinopoulos DM. 2008. *A flexible regression approach using GAMLSS in R.* International Workshop of Statistical Modelling 2008.

Risch N. 1990. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 46:222-228.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69:138-147.

Rosenberger A, Janicke N, Köhler K, Korb K, Kulle B, Bickeböller H. 2005. Surrogate phenotype definition for alcohol use disorders: a genome-wide search for linkage and association. *BMC Genetics* 6(Suppl 1):S55.

Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk CT, Ginsburg F. 1981. Genetics of HLA disease associations: The use of the Haplotype Relative Risk (HRR) and the 'Haplo-Delta' (Dh) estimates in juvenile diabetes from three racial groups. *Hum Immunol* 3:384.

Ryan TP. 1997. Modern Regression Models. John Wiley & Sons, Inc., New York.

Salter KC, Fawcett RF. 1993. The ART test of interaction: a robust and powerful test of interaction in factorial models. *Comm Stat-Simul Comp* 22:137-153.

Sauter W. 2008. Case-control study of genetic susceptibility in early onset lung cancer: investigation of Matrix Metalloproteinase-1 (MMP1). Presentation in the Department of Genetic Epidemiology, University of Göttingen on 09 June 2008.

Schaid DJ. 1996. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epi* 13:423-449.

Scuteri A, Sanna S, Chen WM, Uda M, Albai G, et al. 2007. Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLoS Genet* 3:e115.

Sham P. 1998. *Statistics in Human Genetics*. Arnold, London.

Sham PC, Curtis D. 1995. An extended transmission/disequilibrium test (TDT) for multi-allelic marker loci. *Ann Hum Genet* 59:323-336.

Sinsheimer JS, Blangero J, Lange K. 2000. Gamete-competition models. *Am J Hum Genet* 66:1168-1172.

Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 62:450-498.

Spielman RS, McGinnis RE, Ewens WJ. 1993. The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506-512.

Stasinopoulos DM, Rigby RA. 2007. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* 23:7.

Stasinopoulos DM, Rigby RA, Akantziliotou C. 2008. Instructions on how to use the GAMLSS package in R. Technical Report 01/06. STORM Research Centre, London Metropolitan University, London.

Talmud PJ, Hawe E, Martin S, Olivier M, Miller GJ, Rubin EM, Pennacchio LA, Humphries SE. 2002. Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. *Human Molecular Genetics* 11(24):3039-3046.

Thompson GL. 1991. A note on the rank transform for interactions. *Biometrika* 78:697-701.

Thompson GL. 1993. Amendments and corrections: a note on the rank transform for interactions. *Biometrika* 80:711-713.

Thomson G, Robinson WP, Kuhner MK, Joe S, Klitz W. 1989. HLA and insulin gene associations with IDDM. *Genet Epi* 6:155-160.

Toothaker LE, Newman D. 1994. Nonparametric competitors to the two way ANOVA. *J Educ Behav Stat* 19:237-273.

Tsai C, Hsuehb H, Chenac JJ. 2004. A generalized additive model for microarray gene expression data analysis. *J Biopharmaceutical Stat* 14(3):553 - 573.

Tseng RC, Hsieh FJ, Shih CM, Hsu HS, Chen CY, Wang YC. 2009. Lung cancer susceptibility and prognosis associated with polymorphisms in the nonhomologous end-joining pathway genes: a multiple genotype-phenotype study. *Cancer* 115(13):2939-48.

Vogel F, Motulsky AG. 1986. *Human Genetics, Problems and Approaches.* 2nd Ed. Springer-Verlag, Berlin.

Wald A. 1945. Sequential Tests of Statistical Hypotheses. *Ann Math Statist* 16(2): 117-186.

Waldman ID, Robinson BF, Rowe DC. 1999. A logistic regression based extension of the TDT for continuous and categorical traits. *Ann Hum Genet* 63:329-340.

Weinberg C. 1999. Methods for detection of parent-of-origin effects in genetic studies of case-parent triads. *Am J Hum Genet* 65:229-235.

Weinberg C, Wilcox A, Lie R. 1998. A log-linear approach to caseparent- triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 62:969-978.

Wheeler E, Cordell HJ. 2007. Quantitative trait association in parent offspring trios: extension of case/pseudocontrol method and comparison of prospective and retrospective approaches. *Genet Epi* 31:813-833.

Whittock NV, Izatt L, Mann A, Homfray T, Bennett C, Mansour S, Hurst J, Fryer A, Saggar AK, Barwell JG, Ellard S, Clayton PT. 2003. Novel mutations in X-linked dominant chondrodysplasia punctata (CDPX2). *J Invest Dermatol* 121(4):939-42.

WHO (World Health Organization). 2004. The molecular genetic epidemiology of cystic fibrosis. Report of a joint meeting of WHO/ECFTN/ICF(M)A/ECFS. Genoa, Italy, 19 June 2002.

WHO (World Health Organization). 2009. Fact Sheet on Cancer: Fact Sheet 297. Available from: `http://www.who.int/mediacentre/factsheets/fs297/en/index.html`.

WHO (World Health Organization). 2009. Report on the Global Tobacco Epidemic: Implementing smoke-free environments. Geneva. Available from: `http://whqlibdoc.who.int/publications/2009/9789241563918_eng_full.pdf`.

Wilk JB, Laramie JM, Latourelle JC, Williamson S, Nagle MW, Tobin JE, Foster CL, Eckfeldt JH, Province MA, Borecki IB, Myers RH. 2008. NYD-SP18 is associated with obesity in the NHLBI Family Heart Study. *Int J Obes (Lond)* 32(6):930-5. Epub 2008 Mar 4.

Wilson SR. 2001. Epistasis and its possible effects on transmission disequilibrium tests. *Ann Hum Genet* 62:565-575.

Wittkowski KM, Liu X. 2002. A statistically valid alternative to the TDT. *Human Hered* 54(3):157-164.

Wittkowski KM, Liu X. 2004. Beyond the TDT: Rejoinder to Ewens and Spielman. *Human Hered* 58:60-61.

Xie A, Kwok A, Scully R. 2009. Role of mammalian Mre11 in classical and alternative nonhomologous end joining. *Nat Struct Mol Biol* 16(8):814-8. Epub 2009 Jul 26.

Yang Q, Rabinowitz D, Isasi C, Shea S. 2000. Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. *Hum Hered* 50:227-33.

Yoon S, Suh YJ, Mendell NR, Ye KQ. 2005. A Bayesian approach for applying Haseman-Elston methods. *BMC Genet* 30(Suppl 1):S39.

Zhao J, Jin L, Xiong M. 2006. Test for interaction between two unlinked loci. *Am J of Hum Genet* 79:831-845.

Zhu X, Elston RC. 2001. Transmission/Disequilibrium Tests for Quantitative Traits. *Genet Epi* 20:57-74.

Zivelin A, Ogawa T, Bulvik S, Landau M, Toomey JR, Lane J. 2004. Severe factor XI deficiency caused by a Gly555 to Glu mutation (factor XI-Glu555): a cross-reactive material positive variant defective in factor IX activation. *J Thromb Haemost* 2(10):1782-9.

Index

Author's Profile

Name:	Jingky Pamesa Lozano
Email:	jingky.lozano@stud.uni-goettingen.de

Education

1984 - 1988	High School Education, Manila Science High School, Philippines
1988 - 1992	Bachelor of Science in Public Health, University of the Philippines Manila
1998 - 2004	Master of Science in Public Health (Biostatistics) University of the Philippines Manila
2005 - 2010	Doctoral Studies (Faculty of Mathematics) concurrent with the Interdisciplinary Ph.D. Program in Applied Statistics and Empirical Methods (Genetic Epidemiology) Center for Statistics, Georg-August-Universität Göttingen, Germany

Professional Affiliations

1992 - 1993	Programmer, Software Ventures International Corporation, Makati City, Philippines
1993 - 1995	Science Research Specialist, HIV/Virology Section, Dept. of Microbiology Research Institute for Tropical Medicine, Department of Health, Philippines
1995 - 1997	Research Supervisor, Research Ventures Incorporated, Philippines
1997 - 2000	Research Manager & Proprietor, Development Research & Analysis Systems, Philippines
1997 - 2000	Professional Staff, Research and Operations Audit Section Medical and Underwriting Division, Philamlife, Manila, Philippines
2000 - 2005	Research Consultant/Biostatistician, Research and Biotechnology Division, St. Luke's Medical Center, Philippines
2001 - 2005	Assistant Professor, College of Public Health, University of the Philippines Manila
2002 - 2005	Affiliate Faculty Member, University of the Philippines Open University (UPOU)
2005 - 2007	Wissenschaftliche Mitarbeiterin (Scientist) Department of Genetic Epidemiology, Division of Medicine Georg-August-Universität Göttingen, Germany
2007 - 2010	Research Fellow, Graduiertenkolleg 1034 (The impact of inherited polymorphisms in oncology), Georg-August-Universität Göttingen, Germany

Other Offices Held

1991 - 1992	President, University of the Philippines Hygiene Society / Public Health Society
1996	Goodwill Ambassador to the 23rd Ship for Southeast Asian Youth Program
2001 - 2005	Board of Trustees & Directors (2001-2005) University of the Philippines BS Public Health Alumni Association
2003	Philippine delegate, 21st Century Renaissance Youth Leaders Invitation Program Tokyo, Japan

Scholarships, Grants and Selected Awards

State Scholarship Program, Department of Education,Culture & Sports, Philippines (1988 - 1991)

Francisco J. Nicholas Scholarship, University of the Philippines Manila (1991 - 1992)

Diamond Jubilee Faculty Grant Recipient, University of the Philippines Manila (2004 - 2005)

Graduiertenkolleg 1034 Scholarship, Deutsche Forschungsgemeinschaft, Germany (2007 - 2010)

Best Paper Presentor Award, Young Statisticians' Session, 18th Annual Meeting of the Belgian Statistical Society, Belgium (2010)

Selected Publications

Guazon JLB, **Lozano JP**, Posas FEB, Calleja, HB. (2004). Chest radiograph aortic arch calcification is a surrogate marker of coronary artery disease. *The Leading Edge in Cardiology 4* edited by Saavedra RD and Luque MPM. Heart Institute, St. Luke's Medical Center, Quezon City, Philippines : 117-127.

Bickeböller H, Goddard KAB, Igo RP, Kraft P, **Lozano JP**, Prankratz N. (2007). Issues in association mapping with high-density SNP data and diverse family structures. *Genetic Epidemiology* 31 (Supplement 1): S22-S33.

Elsner L, Muppala V, Gehrmann M, **Lozano J**, Malzahn D, Bickeböller H, Brunner E, Zientkowska M, Herrmann T, Walter L, Alves F, Multhoff G, Dressel R. (2007). The heat shock protein HSP70 promotes mouse NK cell activity against tumors which express inducible NKG2D ligands. *Journal of Immunology* 179: 5523-5533.

Malzahn D, Balavarca Y, **Lozano JP**, Bickeböller H. (2009). Tests for candidate gene interaction for longitudinal quantitative traits measured in a large cohort. *BMC Proceedings 2009*, 3(Suppl 7): S80.

Elsner L, Flügge PF, **Lozano J**, Muppala V, Eiz-Vesper B, Demiroglu SY, Malzahn D, Herrmann T, Brunner E, Bickeböller H, Multhoff G, Walter L, Dressel R. (2010). The endogenous danger signals HSP70 and MICA cooperate in the activation of cytotoxic effector functions of NK cells. *Journal of Cellular and Molecular Medicine.* Vol 14, No 4, pp. 992-1002.