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 mea-

sures 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 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 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 QTDT_M. Comparison of the QTDT_M, QCPG and simple linear regression using simulated data showed that the 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 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 con-

sidering 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.