**ABSTRACT**

LOU, XUEMEI. Methods Evaluation and Application in Complex Human Genetic Diseases. (Under the direction of Dr. Zhao-Bang Zeng and Dr. Elizabeth R. Hauser.)

One of the most important tasks in human genetics is to search for disease susceptibility genes. Linkage and association analyses are two major approaches for disease-gene mapping. Chapter 1 reviewed the development of disease-gene mapping methods in the past decades. Gene mapping of complex human diseases often results in the identification of multiple potential risk variants within a gene and/or in the identification of multiple genes within a linkage peak. Thus a question of interest is to test whether the linkage result can be explained in part or in full by the candidate SNP if it shows evidence of association, and then provide some guidance for the next time-consuming step of positional cloning of susceptibility genes. Two methods, GIST and LAMP, which access whether the SNP can partially or fully account for the linkage signal in the region identified by a linkage scan, are evaluated on Genetic Analysis Workshop 15 (GAW15) simulated rheumatoid arthritis (RA) data and discussed in Chapter 2. The simulation results showed that GIST is simple and works slightly better than the LAMP-LE test when there is little linkage evidence. The LAMP linkage test has limited power when there is not much linkage evidence, and the LAMP association test is the best not only when the linkage evidence is extremely high, but also when there is some LD between the candidate SNP and the trait locus. The fact that complex traits are often determined by multiple genetic and environmental factors with small-to-moderate effects makes it important to investigate the behavior of current association methods under multiple risk variants model. In Chapter 3, we compared APL, FBAT, LAMP, APL-Haplotype, FBAT-LC and APL-OSA conditional test in five multiple risk variants models. The simulation results showed that the power of single marker

association tests is closely correlated with the amount of LD between marker and disease loci, and these tests maintain good power to detect multiple risk variants in a small region with a moderate degree of LD for fully genotyped families. Global tests, such as FBAT-LC are sensitive to the presence of at least one susceptibility variant, but are not helpful for selecting the most promising SNPs for further study. We reported that if multiple haplotypes are associated with different disease loci, the haplotype tests results can be misleading while the APL-OSA conditional test has the greatest power to properly dissect the clustered associated markers for all cases with an acceptable type I error rate ranging from 0.033 to 0.056. We applied APL-OSA conditional test on GENECARD samples, and got reasonable results. One linkage region of particular interest on chromosome 3 was identified by two independent genome linkage scan with Coronary Artery Disease (CAD). Multiple disease susceptibility genes have been reported from this region, and there are also linkage evidence that this region may harbors a gene or genes determining HDL-C levels. Within this region, a search for HDL-C QTL and analyses of the relationship between genetic variants, HDL-C level to CAD risk are discussed in Chapter 4. We performed CAD association and HDL-C QTL analysis on two independent datasets. We identified SNP rs2979307 in the OSBPL11 gene which survives a Bonferroni correction. We observed different HDL-C trends with HDL-C associated SNPs. Even with the evident heterogeneity presented in our CAD population, we detected several association signals with SNPs in KALRN, MYLK, CDGAP and PAK2 genes in both CAD datasets for HDL-C, where all these genes belong to a Rho pathway.

Methods Evaluation and Application in Complex Human Genetic Diseases

by
Xuemei Lou

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2008

APPROVED BY:

| | |
|---|---|
| Elizabeth R. Hauser | Greg Gibson |
| Eric Stone | Zhao-Bang Zeng<br>Chair of Advisory Committee |

## DEDICATION

***To my dearest husband and son***

***Dr. Tianhao Zhang and Chris Jien Zhang***

献给我亲爱的先生章天浩和儿子章骥恩

## BIOGRAPHY

Xuemei Lou was born in Hunan Province, People's Republic of China on December 16, 1971. She received her Bachelor of Science degree and Master of Science in Automatic Control from Tsinghua University in 1994 and in 1997 respectively. After graduation from Tsinghua University, she went to Duke University in Durham, NC for her another Master of Science degree in electrical and computer engineering, and completed it in 2000. She worked as a computer engineer in EMC for a year and half, and as a web developer in Veterinary School at North Carolina State University for another year, then started her doctoral work in the Bioinformatics Research Center at North Carolina State University in 2003. While working toward her doctoral degree, she worked as a part-time research analyst at the Center for Human Genetics at Duke University under the direction of Dr. Elizabeth R. Hauser. The current research interests of Dr. Hauser include cardiovascular genetics, informatics, and methods developments for complex diseases.

## ACKNOWLEDGEMENTS

I am very grateful for the advice of my advisor, Dr. Elizabeth R. Hauser, throughout my doctoral studies. She is always patient and nice with me, and provides very helpful suggestions for the problems I encounter during my research. I am proud of working with her. I also would like to thank Dr Zhao-Bang Zeng, Dr. Greg Gibson and Dr. Eric Stone for their comments and advice on my research during our committee meetings. I would like to thank Dr. Bruce S. Weir for his encouragement and kind help during the first two years of my doctoral work, especially for his advice on my course selection.

I would like to thank the staff at the Bioinformatics Research Center, especially Juliebeth Briseno, who answered so many questions I had about graduate school regulations. I also thank Dr. Silke Schmidt at Center for Human Genetics (CHG) at Duke University for her advice and comments on my research. I would like to thank Ren-Hua Chung and Xuejun Qin for all discussions I have had with them in the past years. I also appreciate the support from members in AGENDA group at CHG for answers and suggestions for all kinds of technical problems I have met.

I would like to express my special thanks to Tianhao Zhang, my husband, who supports me with love and without reservation, and to my lovely son, Chris Jien Zhang, who is the love and joy of my life. Needless to say, none of this would be possible without the support of my family including my parents and sisters in china and my parents-in-law.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 Introduction to disease-gene mapping

Gregor Mendel was an Augustinian priest and scientist, and is often called the father of genetics. He published his paper “Experiments of Plant Hybridization” in 1865, which later became Mendel’s law of inheritance. Mendel’s work prompted the foundation of the discipline of modern genetics. Mendelian disorders refer to diseases for which alternative genotypes fall into distinct, discrete phenotypic classes, following Mendel's laws of inheritance. One of the major tasks in human genetics studies is to search for disease susceptibility genes. Positional cloning is a method of gene identification. This method identifies and maps genes solely based on a specific phenotype, and no biological knowledge regarding the phenotype is required (Botstein & Risch 2003). Largely by the process of positional cloning, disease-gene mapping has been fairly successful for Mendelian disorders (Risch 2000). Over a thousand monogenic Mendelian diseases genes have been reported in “the online version of Mendelian Inheritance in Man” database (OMIM) (Antonarakis & McKusick 2000).

The formation of complex diseases involves the interaction among multiple genes and multiple environmental risk factors. Alzheimer’s disease, type-2 diabetes and cancer and Coronary Artery Disease (CAD) are some examples of complex diseases. Consequently, disease-gene mapping for complex diseases is much more challenging than mapping genes for Mendelian disorders. The challenges include genetic heterogeneity in which mutations in different genes can cause the same disease phenotype (Lander & Schork 1994; Risch 2000), and incomplete penetrance, phenocopies, and late age of onset also limit the progress of complex disease gene mapping (Gillanders *et al.* 2006).

Despite these hindering factors, disease-gene mapping for complex diseases have obtained substantial achievements. For example, Y402H, a common variant in the complement factor H gene (*CFH*) is strongly associated with Age-related macular degeneration (AMD), reported by two totally independent research groups (Haines *et al.* 2005; Klein *et al.* 2005).

Across the human genome, about 99.9% of the genome is identical between people, but there are still millions of differences among the 3.2 billion base pairs (Kruglyak & Nickerson 2001). These genetic variations can cause phenotypic variation among people and are potentially associated with traits or diseases. Genetic markers, which are nucleotide variants with known positions, are often used for human disease analyses. There exist several types of genetic markers, for example, Restriction Fragment Length Polymorphisms (RFLP's), microsatellites, and single nucleotide polymorphisms (SNPs). Markers can be used to construct a genetic map, which can be used as a reference for disease-gene mapping (Dib *et al.* 1996). Botstein et al. (Botstein *et al.* 1980) proposed the concept using RFLP's as the markers to construct a genetic map. Later, denser microsatellites were used to construct genetic maps (Murray *et al.* 1994; Dib *et al.* 1996). SNPs, which usually contain two alleles, have become the standard polymorphisms nowadays for higher resolution genetic disease-mapping due to their abundance throughout the human genome (Kruglyak 1997; Sachidanandam *et al.* 2001). It was estimated that there are around 7.1 million SNPs with a minimal allele frequency of at least 0.05 in the human population (Kruglyak & Nickerson 2001). In the current build 129 of dbSNP, the number of SNPs in this public database is around 18 million including 6.6 million validated SNPs deposited from various studies. Falling genotype costs and the completion of the International HapMap Project have made genome-wide association (GWA) analyses of hundreds of thousands of single nucleotide polymorphism (SNP) markers possible.

Specific statistical methods for disease-gene mapping with numerous markers are required. There are two types of commonly used statistical methods, linkage and association (linkage disequilibrium) analyses (Lander & Schork 1994) and they will be introduced in the following sections.

## 1.2 Linkage analysis

Theoretical methods for linkage tests were proposed around 1930 (Penrose 1935; Fisher 1935a; Fisher 1935b). Linkage analyses are used to find chromosome regions that potential harbors a proposed disease susceptibility gene or gene(s). Linkage evidence is often measured by the logarithm (base 10) of the odds (LOD) score (Morton 1955), which is the logarithm of a likelihood ratio with the recombination rate $\theta$ estimated from the observed data with respect to random recombination ($\theta = 0.5$) in the model. A traditional LOD score posits that a single locus contributes to the disease with a specific model of inheritance (e.g., dominant or recessive). Hence, this type of method is called parametric as it requires a genetic model assumption. Parametric linkage analyses may not very powerful for complex diseases, since an obvious genetic segregation of markers cannot be observed in the polygenic disorders (Weeks & Lathrop 1995).

Another type of linkage methods are called allele sharing methods (Kruglyak *et al.* 1996;

Kong & Cox 1997; Whittemore & Tu 1998). These methods evaluated genetic markers in pairs of affected relatives in a pedigree to see how often a particular copy of a chromosome region is shared by identical-by-descent (IBD), that is, is inherited from a common ancestor (Lander & Schork 1994). A significant departure of the observed frequency of IBD sharing at a locus from the expected IBD sharing by chance implies the presence of linkage. A major advantage of the allele-sharing method is that it assumes no model for the inheritance of a trait. Therefore, it is referred as a non-parametric model. Allele sharing methods tend to be more robust than parametric linkage analysis even in the presence of incomplete penetrance, phenocopy, genetic heterogeneity, and high frequency disease alleles, but are often less powerful than a correctly specified parametric linkage analysis (Lander & Schork 1994).

Linkage analysis can also be classified as two-point or multipoint analyses (Kruglyak *et al.* 1996). For two-point linkage analysis, only one marker and the disease locus are considered when calculating the statistic. For multipoint linkage analysis, several markers are considered simultaneously with the disease locus. Thus, we can define the most likely position of the disease locus on the marker map.

## 1.3 Association analysis

Linkage analysis may lack power for common diseases caused by multiple genes and environmental factors (Cardon & Bell 2001). Since the linkage test uses LOD scores instead of a statistical significance test to measure the degree of linkage evidence in data, it is not intuitively obvious how large a LOD score should be in order for investigators to claim a significant finding (Lander & Kruglyak 1995; Risch & Botstein 1996; Curtis 1996; Morton 1998). The identified linkage regions are also often broad up to 40 megabases (Cardon & Bell 2001). Association analysis can be used as a complementary method to linkage analysis. The association test can be more powerful than the linkage test, and it requires fewer samples than linkage analysis to achieve the same power for common complex diseases (Risch & Merikangas 1996).

Association analysis tests whether the disease and marker alleles are in linkage disequilibrium (LD). Disease phenotypes are used for association analyses instead of disease loci since, in general, the disease loci are unknown (Weiss & Terwilliger 2000). LD generally spans only small distances, and the markers used for association analysis are often very tightly spaced. Therefore, association analysis provides a higher resolution for locating disease genes than linkage analysis. A very common strategy in the past for identifying complex disease genes is to conduct linkage analyses first and then follow significant results with tests for association at a denser panel of markers in an attempt to further localize the disease gene (Cardon & Bell 2001).

In the terms of samples, there are two types of statistical methods for association analysis, population-based (case-control and case-cohort studies) and family-based studies (Laird & Lange 2006).

### 1.3.1 Population-based association analysis

Population-based analysis requires samples to be independently collected. It compares the differences of distributions of allele frequencies between the affected individuals (cases) and unaffected individuals (controls) (Risch 2000). A contingency table can be created and the Pearson chi-squared statistic or Fisher's exact test can be used to test for association. Regression-based analyses such as logistic regression can also be used in the case-control test (Agresti 2002). The major concern of the case-control analysis is that the presence of confounding effects in the samples could give rise to a high false positive rate in the analysis (Risch 2000; Devlin *et al.* 2001). For example, population admixture and population substructure can produce association between unlinked loci (Ewens & Spielman 1995). Two major types of approaches were proposed to solve this problem: genomic control (GC) (Devlin & Roeder 1999; Devlin *et al.* 2001) and structured analysis (SA) (Pritchard *et al.* 2000).

### 1.3.2 Family-based association analysis

A widely used family-based method, the TDT (Spielman *et al.* 1993), compares the differences of alleles transmitted and untransmitted from parents to affected siblings in triad families (one affected offspring and both parents). A McNemar's chi-squared test is used for the paired transmitted and untransmitted statistics. The TDT was originally proposed to test for linkage in the presence of association, but it is also a valid test for association in the presence of linkage (Ewens & Spielman 2005). In terms of statistical power, the TDT has similar power compared with case-control studies for association tests when the number of triad families is equal to the number of cases and the number of cases is equal to the number of controls for case-control studies (McGinnis *et al.* 2002). Hence, it costs less to perform case-control studies for association, since collection of family samples generally requires more time and money (Laird & Lange 2006). However, the TDT test has the advantage that it is valid even when population stratification is present in the data (Ewens & Spielman 1995), because the test is conditional on parental data.

In the TDT, each pair of transmitted/untransmitted alleles from a parent to an affected sibling is treated as independent to construct the McNemar's test. However, as a test for association in a linkage region, this assumption does not hold for transmissions between affected siblings. Hence, the TDT is not a valid test for association when more than one affected sibling is used and there is linkage between marker and disease loci (Martin *et al.* 1997). One solution is to randomly select one affected sibling from each family and perform the TDT (Wang *et al.* 1996). However, affected sibling pairs can significantly increase the power and efficiency of the family-based association test (Risch 2000). Thus, it is not an optimal solution for the TDT to use only one affected sibling in the family when there are other affected siblings. Numerous extensions and generalizations of the TDT have been proposed in the literature.

Martin et al. (Martin *et al.* 1997) proposed the Pedigree Disequilibrium Test (PDT) to account for linkage in families with multiple affected siblings. PDT treats the transmissions from a parent to the affected sib pair as a unit, and the unit can be shown to be independent between parents. The PDT statistic and its variance were constructed based on the unit of transmissions and avoids the independence assumption between affected siblings used in TDT. The TDT was also generalized to extended pedigrees. In Martin et al. (Martin *et al.* 2000), the extended pedigrees are partitioned into several related nuclear families, and the transmissions in each related nuclear family sums to a statistic. The variance for the statistic was estimated based on independent transmissions between each extended pedigree. Abecasis et al. (Abecasis *et al.* 2000b) also used a similar strategy as Martin et al. (Martin *et al.* 2000) that generalized the TDT to extended pedigrees.

When parental genotypes for the affected siblings are missing, Spielman et al. (Spielman & Ewens 1998) proposed the S-TDT test to compare the difference of allele frequencies between affected and unaffected siblings without using parental data. However, S-TDT still has the requirement that only one affected sibling with one unaffected sibling should be used in each family for a valid test for association in the presence of linkage. S-TDT was generalized to multiple affected sibs in Horvath and Laird (Horvath & Laird 1998) as the SDT test.

Another approach to deal with missing parental genotypes is to infer the missing parental genotypes from siblings' genotypes and then compares the number of alleles transmitted and untransmitted from parents to affected siblings (Weinberg 1999). Knapp (Knapp 1999) proposed "reconstruction combined TDT" (RC-TDT), which reconstructs missing parental genotypes based on siblings' genotypes first and then performs the combined TDT and S-TDT test. However, because of the same property inherited from TDT and S-TDT, RC-TDT is not a valid test for association in the presence of linkage when multiple affected siblings are used in the data.

Clayton (Clayton 1999) proposed a score test derived from the likelihood of parental genotypes and offspring genotypes conditional on disease in the offspring. The score test proposed by Clayton (Clayton 1999) is implemented in the software package

TRANSMIT. The inference of missing parental genotype in TRANSMIT is based on Mendelian probabilities for the siblings, which assumes the transmissions to the affected sibs are independent. The inference is appropriate when there is no linkage between disease and marker loci, but it can inflate the type I error rate in a linkage region by ignoring linkage when inferring the missing parental genotypes with multiple affected sibs in the data in TRANSMIT (Martin *et al.* 2003).

### 1.3.3 Association in the Presence of Linkage method

Linkage between disease and marker loci should be considered when inferring the missing parental genotypes based on siblings' genotypes with multiple affected sibs present in the data (Martin *et al.* 2003). Martin et al. (Martin *et al.* 2003) proposed The Association in the Presence of Linkage (APL) method, which uses nuclear families with at least one affected offspring. The APL compares the difference between the number of copies of a specific allele in affected offspring and the expected number under the null hypothesis of no association conditional on parental genotypes. The APL can infer missing parental genotypes properly in the linkage region by taking the IBD parameters for affected siblings into consideration. Hence, APL does not have the problem of possible inflation of the type I error rate, as in TRANSMIT, if linkage is present and multiple affected siblings' data are used (Martin *et al.* 2003). Martin et al. (Martin *et al.* 2003) demonstrated that APL can have more power than PDT and FBAT (Rabinowitz & Laird 2000) for nuclear family data with missing parents. Hence, APL provides a useful family-based association tool for late-onset diseases in which parental data are usually not available.

Chung et al. (Chung *et al.* 2006) extended the original APL proposed in Martin et al. (Martin *et al.* 2003) to be flexible for mixed nuclear family structures (including the mixture of singleton and multiplex families with an arbitrary number of unaffected sibs) using a variance estimator based on the bootstrap approach (Efron & Tibshirani 1993) and a strategy of inference for missing parental genotypes based on IBD status between multiple affected siblings.

Rabinowitz and Laird (Rabinowitz & Laird 2000) proposed an unified approach to family-based association tests (FBAT), which compared the difference between the transmissions from parents to the affected siblings and the expected value conditional on the minimum sufficient statistics for the null distribution. This approach is robust to population stratification, and it is applicable to arbitrary pedigree structure and arbitrary missing allele information. FBAT implements a broad class of Family Based Association Tests, and is another powerful family based association method.

### 1.3.4 Multiple markers association methods

Gene mapping of complex human diseases often results in the identification of several potential risk variants within a gene and/or in the identification of several genes within a linkage peak. Methods to test for multiple markers simultaneously are also being developed to meet the need for additional tests including multiple markers.

There are two kinds of family-based methods to test for multiple markers. One class of tests that take advantage of the correlation of closely positioned markers in the genome is the global test, which jointly considers the effects of all markers on the disease. When there is at least one marker associated with the disease, the global test gives a significant result. FBAT-MM (Rakovski *et al.* 2007) and FBAT-LC(Xu *et al.* 2006) are examples of this kind of method for family data. Both FBAT-MM and FBAT-LC used the statistics for single-marker tests, and then linearly combine them with different weights. The other class of tests is haplotype-based tests, regression-based tests (Rakovski *et al.* 2007) and sum of single-marker statistics tests (Hao *et al.* 2004), which can be used to avoid the problem of multiple testing by combining all SNPs into a single omnibus test.

A haplotype is a set of closely linked genetic markers present on one chromosome which tends to be inherited together. Another way to think about it is that a haplotype is a "super allele". Haplotype analyses can show more power than single-marker analyses if the joint linkage disequilibrium (LD) between markers and the disease locus is stronger than the pairwise LD between a single marker and the disease locus (Morris & Kaplan 2002; Nielsen *et al.* 2004). The null hypothesis of a global haplotype test is that none of the haplotypes are associated with the disease. The global test can also have more power than individual haplotype tests since individual haplotype testing has the multiple-testing issue (Morris *et al.* 1997). A global multiple-locus haplotype test is introduced and implemented in APL package (Chung *et al.* 2006). FBAT also offer a global and individual haplotype analysis in their software package (Horvath *et al.* 2004). Several multiple markers tests are compared and described in Chapter 3.

## 1.4 Associated variants accounting for linkage methods

After fine-mapping association studies narrowed the linkage region to a disease trait to about one or a few centimorgans (cM), researchers have tried different ways to further narrow the fine-mapping results to guide the next step of positional cloning, which is time-consuming and involves intensive laboratory work (Yuan *et al.* 2004). Thus a question of interest is which associated SNP (or combination of these SNPs) in the region identified by linkage analysis influences susceptibility to the trait. Sun et al. (Sun *et al.* 2002) proposed an approach based on the observation that if a particular SNP is the only polymorphism in the region that influences the trait, then conditional on the genotypes at that SNP for the affected relatives, there should be no unexplained allele oversharing in the region among affected individuals.

Sun's approach identifies SNPs whose genotypes can fully explain the observed linkage signal, but their test does not identify SNPs that partially explain the linkage signal, which is more realistic in complex diseases. Li et al. (Li *et al.* 2004) proposed a Genotyped-IBD Sharing Test (GIST) to access whether alleles could account in part for the observed linkage signal. This approach determines if families selected on the basis of the presence of the tentatively associated allele show stronger evidence of linkage as measured by increased allele sharing identical by descent (IBD) by affected family members (Li *et al.* 2004).

The above methods are based on IBD data, which usually can be inferred only from the genotype data with some uncertainty, and these methods apply only to relative pair data.

Yuan et al., (Yuan *et al.* 2004) using genotype data, developed a statistical hypothesis-testing method to pinpoint a SNP, or SNPs, suspected of responsibility for a disease trait linkage among a set of SNPs tightly linked in a region (Yuan *et al.* 2004). This method used genotype data of affected individuals or case-control studies, which are widely available in research groups. This method can be used for singleton data, relative pair data, or general pedigree data.

Li et al. (Li *et al.* 2005) proposed a likelihood-based approach, Linkage and Association Modeling in Pedigrees (LAMP), which quantifies the degree of LD between the candidate SNP and the putative disease locus through joint modeling of linkage and association through a likelihood function of the marker data conditional on the trait data for a sample of affected sib pairs, with disease penetrances and disease-SNP haplotype frequencies as parameters. Model parameters were estimated by maximum likelihood and two likelihood-ratio tests were proposed to access whether the candidate SNP can partially or fully explain the linkage signal (Li *et al.* 2005).

This group of tests is neither a test of linkage nor a test of association. An SNP may be tightly linked or in significant linkage disequilibrium with the causal polymorphism yet still not be able to fully explain the linkage signal observed in the region (Yuan *et al.* 2004). The purpose of these tests is to pinpoint SNPs of real interests for guidance of positional cloning. Comparison of GIST and LAMP is described in Chapter 2.

## 1.5 The scope of this dissertation

One linkage peak region identified by a linkage scan could harbor multiple disease susceptibility loci, and multiple causal alleles could exist in one disease gene (Duerr *et al.* 2006; Cohen *et al.* 2006). The complexity within the linkage peak region could result in contradicting results for different association methods, thus it is crucial to investigate those methods where there are multiple risk variants which models the real complex disease situations. Two methods, GIST and LAMP, which access whether the SNP can partially or fully account for the linkage signal in the region identified by a linkage scan, are evaluated on Genetic Analysis Workshop 15 (GAW15) simulated rheumatoid arthritis (RA) data and discussed in Chapter 2. Since the GAW15 simulated RA data has an unrealistic linkage peak (LOD=60) on chromosome 6, it is hard to discriminate between methods. We simulated simplified multiple risk variants models, where several single marker association tests and multiple markers association tests are compared and described in Chapter 3. LAMP is also evaluated under this multiple risk variants setting and is described in Chapter 3 as well. From a linkage region on chromosome 3 identified by two independent genome linkage scan with Coronary Artery Disease (CAD), multiple disease susceptibility genes have been reported (Connelly *et al.* 2006; Wang *et al.* 2007; Wang *et al.* 2008). Within this region with multiple susceptibility genes, a search for HDL-C QTL and analyses of the relationship between genetic variants, HDL-C level to CAD risk are discussed in Chapter 4. Finally, some possible ideas for future work are discussed in Chapter 5.

# Chapter 2

## Comparison of GIST and LAMP on the GAW15 simulated data

## 2.1 Background

The goal of a gene mapping study is to identify genetic variants that predispose to human diseases. For a complex disease, investigators often map the locus of interest first by linkage analysis, which typically results in a large candidate genomic region of up to 40 Mb in size. To localize the susceptibility allele more precisely, disease-marker association analyses are performed, using a much denser map of genetic markers within the linkage region. One particular method of association analysis is based on comparing marker allele frequencies between unrelated cases and controls. In this design, only a subset of the samples originally collected for linkage analysis can be used. As an alternative, family-based association methods have been developed. The classic family-based transmission/disequilibrium test was proposed to test for association in the presence of linkage in family trios containing two parents and one affected offspring (Spielman *et al.* 1993). This approach has been extended to other family structures (Martin *et al.* 2000). If a SNP shows evidence of association, a hypothesis of interest is whether the linkage result can be explained in part or in full by the candidate SNP. The Genotype IBD Sharing Test (GIST) (Li *et al.* 2004) and Linkage and Association Modeling in Pedigrees (LAMP) (Li *et al.* 2005) are two methods that were specifically proposed to address different aspects of this question. The purpose of our study was to evaluate the performance of GIST and LAMP on the simulated GAW15 data. We used the answers to guide our investigations.

## 2.2 Methods

### 2.2.1 GIST

GIST considers one particular marker allele as tentatively associated with the disease variant, and calculates family-specific weight variables defined by the presence of this allele. The variable has to be defined in such a way that the weight variable and IBD sharing configuration among affected family members at that same locus are uncorrelated if there is no disease-marker association, also called the "unbiased selection scheme". If there is a significant correlation between family-specific weights and family-specific linkage evidence, this suggests that the SNP allele could account in part for the observed linkage signal. GIST calculates three kinds of family-specific weight variables, corresponding to dominant, recessive, or additive inheritance models (Li *et al.* 2004).

Once a weight variable $W$ has been defined, the sample correlation coefficient between family weights $W$ and family-specific NPL scores $Z$ is computed. Under no disease-marker association, this correlation is expected to be zero and a one-sided test may be performed. A transformation of the correlation coefficient ($X_i$, $i=dom, rec, add$) that is asymptotically standard-normally distributed is used as the test statistic. When we do not know the underlying disease model, an alternative to carrying out all three tests in GIST is to use $X_{\max} = \max(X_{dom}, X_{rec}, X_{add})$. The distribution of $X_{max}$ under no disease-marker association is estimated empirically by simulating a large number of ASPs under no linkage for various allele frequencies (Li *et al.* 2004). The test based on $X_{\max}$ should be the most appropriate test as we usually do not know the true genetic model for a complex disease.

### 2.2.2 LAMP

LAMP quantifies the degree of linkage disequilibrium (LD) between the candidate SNP and the putative disease locus through joint modeling and estimation of linkage and association parameters. LAMP constructs a likelihood of the marker data conditional on the trait data for a sample of families with disease penetrances and disease-SNP haplotype frequencies as parameters. Model parameters are estimated by maximum likelihood.

Three likelihood ratio tests are proposed to characterize the relationship between the candidate SNP and the disease locus. The first test assesses whether the candidate SNP and the disease locus are linked (LAMP Linkage test). The second test is the direct association test, which assesses whether the candidate SNP and the disease locus are in partial LD so that the SNP may account in part for the linkage signal (LAMP-LE test, $H_0$: $r^2$=0). The third test is an indirect association test, which assesses whether there are other variants which can explain the linkage signal (LAMP-LD test, $H_0$: $r^2$=1). If the null hypothesis of complete LD between the SNP and the disease allele is rejected, the SNP does not fully explain the observed linkage signal, and there may be multiple disease variants in this region (Li *et al.* 2005).

### 2.2.3 Dataset and Analysis

We used all 100 replicates of the simulated GAW 15 family dataset to evaluate the power of GIST and LAMP. Each replicate included 1500 nuclear families of size 4 (2 parents and an affected sibling pair (ASP)).

All SNP markers on chromosome 6, 7, 8, 9, 11, 16 and 18 (a total of 3069 SNPs) were analyzed by GIST and LAMP. We also analyzed genotypes at all eight trait loci with LAMP. For GIST, we used the family-specific NPL scores at the location of the maximum multipoint LOD score on each chromosome, using the microsatellite markers (Abecasis *et al.* 2002; Li *et al.* 2005). LAMP was run without any flanking markers. To compare the LAMP results to standard methods for linkage and association analysis, we also calculated the average multipoint LOD score at the SNP closest to the trait locus, using the SNP markers, and performed family-based association analysis of each SNP with the pedigree disequilibrium test (PDT) (Martin *et al.* 2000; Li *et al.* 2005). LD analysis for 50 SNP markers surrounding each trait locus was performed with GOLD, using the replicate 1 data only (Abecasis & Cookson 2000; Li *et al.* 2005).

### 2.2.4 Power Calculation

The power to reject the null hypothesis of GIST and LAMP-LE was defined in two different ways: If we consider all SNP markers in the region defined by the true trait locus ± 5 cM as the correct candidates, the power was estimated by the proportion of replicates with at least one p-value above a threshold value. If we only consider the SNP closest to the trait locus as the candidate SNP, the power was estimated by the proportion of replicates in which the p-value at this SNP was above a threshold value.

The threshold value was set differently for each test. Since chromosome 7 does not contain any trait loci, the analysis results (p-values) obtained for this chromosome across the 100 replicates define a null distribution for the hypothesis of “no linkage and no association”. Since it was previously shown that linkage and association test statistics are independent under this null hypothesis (Chung *et al.* 2007), this hypothesis implies the null hypothesis of GIST (no correlation between family-specific NPL-scores and weights derived from "non-associated" genotypes). Therefore, this is the relevant null distribution for the GIST combined test and LAMP-LE test. The 5$^{th}$ percentile value of the chromosome 7 p-value distribution for each test was set as the threshold value (false positive rate $\alpha$=0.05).

To derive a null distribution for the LAMP-LD test ($r^2$ =1), we analyzed genotypes at the actual trait loci. Due to the complexity of the GAW15 simulation model, we do not have the perfect null situation. Since a basic assumption of LAMP is the presence of only one disease variant in the region, we decided not to use the trait loci on chromosome 6 and 9. Despite the fact that none of loci A, B, E and F are "pure" disease susceptibility loci, we used these loci as the best guess approximation for the null distribution for the LAMP-LD test.

## 2.3 Results

### 2.3.1 Characteristics of trait Loci and surrounding region

In order to evaluate different methods, it is useful to gain a better understanding of the simulated data first. Table 2.1 lists the risk allele frequency at each trait locus, pairwise $r^2$ values between the closest SNP marker and each trait locus, the number of markers in the $\pm$5 cM region, the maximum $r^2$ between SNP and the trait locus in the $\pm$ 5 cM region, and the average maximum NPL-based LOD score for each chromosome. Pairwise $r^2$ values among the 50 SNP markers and the trait locus are not shown, but most of them were less than 0.05. In other words, there is not much LD among the markers in the ± 5 cM region.

From table 2.1, we can see that Locus C and F have a major risk allele. Locus D has a rare minor risk allele with frequency less than 0.05. $r^2$ values between Locus C, D, E and F and the closest SNP are over 0.1, while $r^2$ values between Locus A, B, G, H and the respective closest SNP are practically 0. The maximum $r^2$ values in the ± 5 cM region around Locus A, B, G, H are all less than 0.02. The linkage evidence for chromosome 6 is overwhelmingly high, but only modest NPL-based LOD scores were obtained for chromosomes 11 and 18. For chromosomes 8, 9 and 16, the NPL-based LOD scores are close to those for the "null" chromosome 7.

### 2.3.2 Power comparison between GIST and LAMP tests

The empirical 5% threshold value obtained from the chromosome 7 analysis was 0.0515 for the GIST combined test, and 0.04 for LAMP-LE. The empirical 5% threshold values for LAMP-LD from Locus A, B, E and F were 0.08, 0.04, 0.02 and 0.14. Since only 100 replicates were available, these values were consistent with a 0.05 nominal p-value. Using these empirical thresholds, the power to reject the null hypotheses of GIST and LAMP-LE at the closest SNP is shown in Table 2.2. The power of the GIST combined test is low for each trait locus, while LAMP-LE has 100% power to reject the null hypothesis for the closest SNP when there is at least moderate LD between the candidate SNP and the trait locus, such as Locus E. Like GIST, the LAMP-LE test has similarly low power for Locus A, B, G and H when the $r^2$ values are less than 0.01.

The power comparison of GIST and LAMP-LE across a $\pm$ 5 cM region is shown in Table 2.3. Using the same threshold values as for Table 2.2, we calculated the power with and without adjustment for multiple testing. The Bonferroni correction for multiple testing corresponds to dividing the respective thresholds by the number of markers in the 10 cM region surrounding each locus. In Table 2.3, the unadjusted power of the GIST combined test is reasonably high for all trait loci, but after adjusting for multiple testing, it is similarly low as in Table 2.2. The LAMP-LE test still has almost 100% power to reject linkage equilibrium when there is at least moderate LD (i.e., $r^2 \geq 0.145$), even with adjustment for multiple testing, but, as expected, has low power when there is no LD, e.g. Locus A, B, G and H, as in Table 2.2. The power of the LAMP-LE test increases when $r^2$ increases.

The pattern of the power estimates for the LAMP-LD test is similar in Tables 2.2 and 2.3, and consistent with expectations. The power decreases when $r^2$ increases, and also depends on the magnitude of the linkage evidence. For Locus C and D, both on chromosome 6 with very strong linkage evidence, the power to reject the null hypothesis of complete LD with a single susceptibility variant is high, even though there is almost complete LD with the closest SNP ($r^2$=0.94). Locus G and H are also located on the same chromosome, but they show little linkage evidence and hence the power to reject complete LD is low. For Locus A and B, LAMP-LD has low power to reject complete LD, presumably also because of little linkage evidence. Since Locus E has a lower $r^2$ value with the closest SNP than Locus F, and they both have moderate linkage evidence, it makes sense that the power to reject complete LD is higher for Locus E than Locus F.

### 2.3.3 Comparison of MERLIN and LAMP Linkage test, PDT and LAMP-LE test

We compared the results from the LAMP linkage test with a standard linkage analysis using MERLIN, and also compared results from the LAMP-LE association test with PDT (data not shown). These comparisons suggest that the linkage and association tests from LAMP for these family structures are very similar to the linkage test implemented in MERLIN and the family-based association test implemented in PDT, respectively.

## 2.4 Discussion

We have completed GIST and LAMP analysis on eight totally different trait loci for rheumatoid arthritis simulated in GAW15 using SNPs that mimic a 10K SNP chip set and microsatellite markers. Our results from applying linkage and association tests to the data are consistent with the genetic effects and levels of LD for each chromosome.

GIST can only be applied to affected sibship data and considers only genotypes at the tentatively associated SNP, without incorporating flanking markers. In theory, GIST can detect association in regions with little overall evidence for linkage. Our results show that the GIST combined test performs poorly even in cases which would seem to be favorable for detecting an associated SNP (both linkage and LD). A possible reason for the failure of GIST to detect associated alleles is because of the multiple loci interacting to affect the RA hazard, and the overwhelming genetic heterogeneity, perhaps even within-family heterogeneity. Since the linkage evidence is very modest on chromosomes 11, 16, and 18 (average LOD scores <1), there is also little power to detect association to the disease alleles.

LAMP can be applied to general pedigree data, including affected and unaffected individuals, and can incorporate flanking markers. Our study shows that LAMP-LE works well when there is at least moderate LD between marker and disease variant, even when $r^2$ values are as low as 0.145. The major disadvantage of LAMP is the speed. For large and sparsely genotyped pedigrees, LAMP can be painfully slow and such pedigrees must be trimmed or discarded in practice.

## 2.5 Conclusions

GIST is simple and fast once family-specific NPL scores and weight variables have been computed, but it did not perform well in the GAW15 simulated data. LAMP is more flexible and more powerful, but much slower. LAMP seems to work particularly well when there is at least moderate LD between the candidate SNP marker and the trait locus.

## 2.6 Acknowledgements

We gratefully acknowledge programming assistance from Xuejun Qin and Mike Schmidt. This research was supported by the National Institutes of Health (NEI R03 EY015216, NIMH R01 MH595228) and the Neurosciences Education and Research Foundation.

## 2.7 Tables

**Table 2.1** Characteristics of trait Loci and surrounding region

| Trait Loci | | | | Closest SNP | | | ± 5cM region | | |
|---|---|---|---|---|---|---|---|---|---|
| Chr | Locus | cM | Risk Allele Frequency | Marker | cM | $r^2$ | Number of markers | Maximum $r^2$ | Average Max NPL (SD) on Chromosome |
| A | 16 | 26.29 | 0.3658 | SNP16_31 | 26.31 | 0.008 | 15 | 0.016 | 0.60(0.60) |
| B | 8 | 170.9 | 0.3992 | SNP8_442 | 167.6 | 0.002 | 10 | 0.006 | 0.53(0.47) |
| C | 6 | 49.46 | 0.8688 | SNP6_153 | 49.46 | 0.551 | 48 | 0.953 | 62.99(7.76) |
| D | 6 | 54.57 | 0.0418 | SNP6_162 | 54.62 | 0.958 | 22 | 0.958 | 62.99(7.76) |
| E | 18 | 94.27 | 0.3797 | SNP18_269 | 94.22 | 0.145 | 26 | 0.145 | 1.21(0.91) |
| F | 11 | 115.29 | 0.6123 | SNP11_389 | 115.28 | 0.94 | 41 | 0.94 | 0.84(0.63) |
| G | 9 | 49.40 | 0.0938 | SNP9_185 | 49.32 | 0.007 | 48 | 0.014 | 0.56(0.48) |
| H | 9 | 51.41 | 0.1663 | SNP9_192 | 51.40 | 0 | 45 | 0.011 | 0.56(0.48) |
| Null | 7 | | | | | | | | 0.56(0.48) |

**Table 2.2** Power comparison of GIST and LAMP at the SNP closest to trait locus

| Locus | Chr | GIST Combined Test (%) | LAMP-LE Test (%) | LAMP- LD Test (%) |
|---|---|---|---|---|
| A | 16 | 4 | 4 | 5 |
| B | 8 | 10 | 8 | 2 |
| C | 6 | 4 | 100 | 74 |
| D | 6 | 2 | 100 | 68 |
| E | 18 | 7 | 100 | 6 |
| F | 11 | 4 | 100 | 3 |
| G | 9 | 4 | 2 | 2 |
| H | 9 | 10 | 6 | 0 |

**Table 2.3** Power comparison of GIST and LAMP within ± 5 cM of trait locus

| Locus | Chr | GIST Combined Test (%) | | LAMP-LE Test (%) | | LAMP- LD Test (%) | |
|---|---|---|---|---|---|---|---|
| | | Unadjusted | Adjusted | Unadjusted | Adjusted | Unadjusted | Adjusted |
| A | 16 | 48 | 3 | 59 | 7 | 29 | 0 |
| B | 8 | 46 | 6 | 44 | 7 | 14 | 1 |
| C | 6 | 82 | 3 | 100 | 100 | 100 | 100 |
| D | 6 | 60 | 4 | 100 | 100 | 100 | 100 |
| E | 18 | 75 | 3 | 100 | 99 | 60 | 5 |
| F | 11 | 93 | 6 | 100 | 100 | 50 | 4 |
| G | 9 | 85 | 7 | 80 | 1 | 34 | 0 |
| H | 9 | 87 | 5 | 87 | 3 | 35 | 0 |

# Chapter 3

## Genetic association tests under models including multiple susceptibility variants

## 3.1 Introduction

Complex traits are often determined by multiple genetic and environmental factors with small-to-moderate effects (Hao *et al.* 2004). Under such assumptions, association tests arguably offer enhanced statistical power compared to the classical linkage approach in searching for genetic loci (Hao *et al.* 2004). To date, evaluations of association tests have been largely limited to single disease susceptibility alleles but models for complex disease may include more than one disease susceptibility allele. The observed allelic association (AA) between disease phenotype and marker genotypes partially depends on the strength of linkage disequilibrium (LD) between the alleles of the disease-causing locus and the observed marker loci. LD decays with increasing genetic distance more rapidly than linkage, and, therefore, a set of closely-spaced markers is a prerequisite to map a disease locus using association approaches. The traditional way to uncover multiple risk variants is to test each SNP sequentially, then adjust for multiple testing to control the false-positive rate. Thus the null hypothesis of traditional association approaches is there is no marker associated with the risk variant. The alternative hypothesis is there is one marker associated with a single genetic risk variant. One of the most discussed threats to the validity of genetic association tests is population stratification, which refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease. Family-based association methods have the advantage of avoiding false positive results caused by population stratification because they have internal controls within families. Different family-based association tests adopt different approaches to utilize family data of different types.

### 3.1.1 Single-marker single genetic risk variant association methods

The introduction of the Transmission/Disequilibrium Test (TDT) was one important landmark in the popularization of association studies (Spielman *et al.* 1993). It detects apparent transmission distortion from parents to affected offspring. The TDT, which was originally restricted to parent-offspring triads, has been extended to a variety of family structures. The Pedigree Disequilibrium Test (PDT) extended the TDT idea to larger pedigrees, but requires unaffected siblings to substitute for the missing parents (Martin *et al.* 2000). The APL (Association in the Presence of Linkage) test, works only on nuclear families, but it can infer missing parental genotypes with or without unaffected siblings' genotypes (Martin *et al.* 2003). FBAT (Family-Based Association Test), also works on extended pedigrees, and it can infer missing parental genotypes as well (Rabinowitz & Laird 2000). LAMP (Linkage and Association Modeling in Pedigrees) works on extended pedigrees, and has the advantage of being a likelihood-based method (Li *et al.* 2005).

### 3.1.2 Multiple Markers and Single Genetic Risk Variant

With increasing numbers of SNPs used in association studies, power decreases quickly due to multiple-test correction penalties. To alleviate this problem, methods to combine individual markers have been proposed, such as haplotype-based methods. A haplotype is the combination of tightly linked alleles on the same chromosome, and it can be viewed as a "super allele". Under the Common Disease Common Variant (CDCV) model, a haplotype-based analysis is expected to provide increased power in the setting of a single disease allele on an ancestral haplotype (Morris & Kaplan 2002; Nielsen *et al.* 2004). Many traditional methods have been extended to offer extensions to haplotype-based approaches, such as APL-haplotype (Chung *et al.* 2006), FBAT-haplotype (Horvath *et al.* 2004), LAMP extension (Biernacka & Cordell 2007), TRIMM (Shi *et al.* 2007) on family triad data. TRIMM is also a method to test the association of a disease trait with multiple, possibly linked SNP markers and, subsequently, to nominate a set of "risk-haplotype-tagging alleles." It uses only the genotypes of affected individuals and their parents without requiring haplotype assignments or phases. It can only be applied to family trio data. Neale et al. also proposed to extend a multiple marker analysis into a Gene-Based Analysis (Neale & Sham 2004) which is basically a haplotype-based method only within gene-wide range.

### 3.1.3 Multiple Markers and Multiple Genetic Risk Variants

Gene mapping of complex human diseases often results in the identification of several potential risk variants within a gene and/or in the identification of several genes within a linkage peak. These findings are at odds with the hypothesis that a single risk allele is responsible for a linkage peak. Along with the declining costs of genotyping, the availability of detailed HapMap data and a growing number of validated human SNPs, methods to test for a disease association with multiple markers simultaneously are also being developed to meet the need for more powerful analysis approaches.

There are two kinds of family-based methods to test for multiple markers. One class of tests that takes advantage of the correlation between closely positioned markers in the genome is the global test. This test is sensitive to at least one marker being associated with the disease. FBAT-MM (Rakovski *et al.* 2007) and FBAT-LC (Xu *et al.* 2006) are examples of this kind of method for family data. Both FBAT-MM and FBAT-LC used the statistics for single-marker tests, and then linearly combine them with different weights. The other class of tests is regression-based tests (Rakovski *et al.* 2007) and sum of single-marker statistics tests (Hao *et al.* 2004), which can be used to avoid the problem of multiple testing by combining all SNPs into a single omnibus test.

We are interested in understanding the behavior of several association analysis tests under single variant and multiple variant models. The goal of our study is to compare methods for determining whether there is evidence for multiple independent risk variants in a linkage region.

## 3.2 Methods

We used the simulation program, SIMLA(Schmidt *et al.* 2005), to generate family data sets for use in linkage and association studies. SIMLA can generate haplotypes associated with two distinct susceptibility variants (possibly with different disease risks) with up to 6 markers included in each haplotype. The disease model may also include an interaction between the two variants. In this simulation study, we simulated a common disease with prevalence of 0.20 with two haplotypes independently associated with two recessively acting susceptibility variants, each with risk allele frequency of 0.15. The Genotype Relative Risk (GRR) of each disease variant is 3.0, and the estimated recurrence risk ratio ($\lambda_{s)}$ is 1.13. The two disease loci are in linkage equilibrium (LE) with each other. Five hundred 3-sib nuclear families with at least 1 affected offspring and genotyped parents were simulated for all models. The general maker map is shown in Figure 3.1. Along the map, M1, M2 and D1 are in one haplotype, and M3 and D2 are in another haplotype, while M4 is not associated with either disease locus or other markers. All markers are evenly spaced at 0.25 cM to model a common complex disease of multiple susceptibility loci in a local linkage region. We simulated 1000 replicates for each model.

We simulated complex LD patterns within the first haplotype. We sought to include a range of LD patterns from complete LD ($r^2$=1) to minimal of LD under 5 different models, where M3 is in complete LD with D2 in all models but the LD between M1 and M2 in each model is different (Table 3.1). We name the models by referring to the strength of LD in the following order: between M1 and D1, between D1 and M2 and between M1 and M2, separated by a dash, High (H): $r^2 \geq 0.9$, Moderate (M): $0.4 \leq r^2 < 0.9$, Low (L): $0 < r^2 < 0.4$, Zero: $r^2$=0. In Model 1, M1 is also in complete LD with D1 but there is no LD between M2 and D1 and M1 and M2 (H-Zero-Zero). In Model2, both M1 and M2 are in complete LD with D1 and each other (H-H-H). In Model 3, M1 and M2 are in high LD, but the pairwise LD between M1, M2 and D1 drops down to around .40 (L-L-H). Both Model 4 and Model 5 have only moderate LD values. LD between M1 and M2 is higher than pairwise LD between M1, M2 and D1 in Model 4 (L-L-M) while LD between M1 and M2 is lower than pairwise LD between M1, M2 and D1 in Model 5 (M-M-L).

Tests for Hardy-Weinberg Equilibrium (HWE) showed no departure from HWE for single markers. We evaluated single-marker association methods that incorporate evidence for linkage in a family based association setting, including APL (Martin *et al.* 2003), LAMP (Li *et al.* 2005) and FBAT (Rabinowitz & Laird 2000). They all work well for single marker analysis, but here we compare their performance under models with multiple susceptibility variants. LAMP, APL and FBAT are compared in Model H-Zero-Zero, Model H-H-H and Model L-L-H for 1000 replicates.

Three likelihood ratio tests were proposed to characterize the relationship between the candidate SNP and the disease locus in LAMP. The first test assesses whether the candidate SNP and the disease locus are linked (LAMP-Linkage). The second test is the direct association test, which assesses whether the candidate SNP and the disease locus are in partial LD so that the SNP or a marker in LD with it accounts partially for the linkage signal (LAMP-LE test, $H_0$: $r^2$=0). The third test is the indirect association test, which assesses whether there are other variants which can explain the linkage signal (LAMP-LD test, $H_0$: $r^2$=1). If the null hypothesis is rejected, then the SNP completely accounts for the linkage signal(Li *et al.* 2005).

In order to distinguish among multiple risk variants associated with different disease loci, we evaluated APL-Haplotype (Chung *et al.* 2006), FBAT-LC (Xu *et al.* 2006) multiple marker test and the APL-OSA conditional association test (Chung *et al.* 2008). APL-OSA, the FBAT-LC and APL-Haplotype are compared in Model 1, 3, 4 and 5 for 1000 replicates.

Ordered-subset analysis (OSA) (Hauser *et al.* 2004b) is a powerful linkage analysis tool that considers genetic heterogeneity based on a trait-related covariate. The covariate may help to identify a more genetically homogeneous subset of families, for example, those with an earlier age at disease onset. The basic idea of the OSA method is to use a family-specific covariate, such as the average covariate value in affected family members, to identify the family subset that provides the strongest linkage signal. Families are ordered by their covariate values and the multipoint (parametric or nonparametric) LOD score for each subset of families is recalculated every time a family is added into the subset. The maximum LOD score may occur at different map positions for different subsets of families. If the covariate is related to the disease, families with high or low ranking covariate values may be genetically more homogeneous and thus provide stronger evidence for linkage to a particular disease locus. The significance of the subset and its evidence for linkage is evaluated using a permutation procedure to estimate empirical p-values. The OSA method has been applied successfully to many complex disease studies (Hall *et al.* 1990; Allingham *et al.* 2005; Woodroffe *et al.* 2006; Schmidt *et al.* 2006; Shah *et al.* 2006). Association study designs based on OSA-identified covariate cutoff points have also been proposed (Schmidt *et al.* 2008).

APL-OSA applies the linkage OSA method (Hauser *et al.* 2004b) to the APL test statistic, which is a measure of allele transmission from parents to offspring. APL-OSA (Chung *et al.* 2008) uses a trait-related covariate to identify a subset of families that has the most extreme APL statistic for a given marker. The null hypothesis of APL-OSA is that there is no correlation between the family-specific contribution to the APL statistic and the family covariate. A permutation procedure is used to approximate the distribution of the APL-OSA statistic under the null hypothesis that there is no relationship between the family-specific covariate and the family-specific evidence for allelic association. APL-OSA is a single-locus test, but in this study we use APL-OSA as a conditional test where APL statistics of other markers are used as the trait-related covariate. We use the high-to-low covariate ordering to test if there is correlation between transmission of the analysis marker and the conditioning marker. The goal is to evaluate independence of markers in the setting of multiple risk variants. The idea is similar in spirit to the conditional linkage test used by Cox et al. (Cox *et al.* 1999). They described an approach to assess the evidence for statistical interaction between unlinked regions takes the evidence for linkage at one region into account when assessing the evidence for linkage across the rest of the genome. Using this method, they showed that the interaction of genes on chromosomes 2 (*NIDDM1*) and 15 (near *CYP19*) makes a contribution to susceptibility to type 2 diabetes in Mexican Americans.

## 3.3 Results

### 3.3.1 Single-marker methods comparison

The comparison of single-marker association tests for the first 3 models, H-Zero-Zero, H-H-H and L-L-H is listed in Table 3.2. The type I error rate for APL, LAMP LE and FBAT tests was measured by M4 which is a non-associated marker in all models. The type I error rate for the LAMP-LD test was measured by M3 which is in complete LD with D2 and the null hypothesis of LAMP-LD is $r^2=1$. In Table 3.2, going from model H-H-H to model L-L-H, the power of FBAT to detect a disease association with M1 drops from 99.6% to 59.5%, the power of APL drops from 70.7% to 34.9%, the power of LAMP-LE drops from 100% to 93.8%, and the power of LAMP-LD remains at 10.0 - 10.3% for marker M4. The type I error rate of FBAT is 5.0%, 4.7% and 4.0% in the three models, the type I error rate of APL is 4.6%, 6.6% and 4.8% respectively, while the type I error rate of LAMP-LE is 16.4%, 20.3% and 15.8%, the type I error rate of LAMP-LD is 2.6%, 3.1% and 3.4%. Thus, FBAT has better power compared to APL with genotyped parents, and the LAMP-LE test has the greatest power. However, the LAMP-LE test has an unacceptably high type I error rate (≈20%) when there are multiple disease-associated marker alleles, although the type I error rate of both LAMP association tests is at the nominal level for single marker analysis, as shown in previously published work (Li *et al.* 2005). The type I error rate of the LAMP-LD test in the presence of multiple disease-associated marker alleles does not exceed the nominal level, but it only has ≈10% power to reject the null hypothesis of $r^2=1$ in all three models. Thus the results of LAMP are difficult to interpret when there are multiple disease-associated markers in the linkage region.

The power of APL and FBAT to detect one or more associated marker allele(s) decreases as the LD between the marker and disease locus decreases, but they both maintain the nominal type I error rate in all models. Both FBAT and APL maintain good power as single marker association tests in the presence of multiple disease variants as long as the LD remains relatively strong ($r^2 > 0.5$). Both APL and FBAT are robust single marker association methods to use in the presence of multiple disease variants. However, the single marker tests report only on one marker at a time and provide little information on whether there is one or more than one susceptibility allele in a linkage region.

### 3.3.2 Multiple-marker methods comparison

After single marker association tests have detected which SNPs are associated, we are interested in testing whether these SNP markers are independent. Since the LAMP-LD test is unable to provide this information, we considered a conditional test of association (APL-OSA) and a joint test (APL-Haplotype and FBAT-LC). Multiple-marker association methods are compared and results for Model H-Zero-Zero, L-L-H, L-L-M and M-M-L are listed in Table 3.3.

The power of the FBAT-LC test for the 2 marker haplotype M1-M2 is 99.4% and 71.2% in Model H-Zero-Zero and L-L-H, and is always higher than the highest result for either marker using FBAT single marker tests in both models, 99.3% and 61.6 respectively. The power of the APL-Haplotype test for the 2 marker haplotype M1-M2 in Table 3.3 is 43%, 19%, 18%

and 40% in Models H-Zero-Zero, L-L-H, L-L-M and M-M-L, while the power of FBAT-LC on (M1, M2) is 99.4%, 71.2%, 58% and 99% respectively. FBAT-LC also maintains higher power than the APL-Haplotype test on (M1, M2) in Table 3.3. The higher power for FBAT-LC is not surprising given the difference in the alternative hypothesis. The alternative hypothesis for FBAT-LC is that any marker in the set of markers tested is associated and APL-Haplotype is performing a specific joint test. Thus the FBAT-LC is a global test which is sensitive to the presence of at least one disease-associated marker but does not identify the specific one.

In Model L-L-H, the power of the APL single marker test is 34.9% and 34.2% for M1 and M2 respectively (Table 3.2), and the power of the APL-Haplotype test is 62%, 19%, 70% and 72% for M1-M2-M3, M1-M2, M1-M3 and M2-M3 respectively (Table 3.3). Thus, APL-Haplotype with 3 markers has increased power compared to the APL single marker test. However, the haplotype test can be quite misleading. M2 and M3 are associated with two distinct susceptibility variants, but in Model H-Zero-Zero, L-L-H, L-L-M and M-M-L, the power for the true M1-M2 haplotype is lower than for the M1-M3 and M2-M3 haplotypes. This demonstrates that a haplotype analysis detects joint effects of disease variants but is unable to distinguish between a single susceptibility variant model and multiple independent susceptibility variants model, and also unable to distinguish between different haplotypes.

The results of the APL-OSA conditional test are shown in Table 3.3. While the power drops to 5.0% when analyzing M1 conditioning on M3, or 3.4% when analyzing M2 conditioning on M3 in model L-L-M, the power of APL-OSA is 100% when analyzing M1 conditioning on M2 or analyzing M2 conditioning on M1 in Models L-L-H, L-L-M and M-M-L. We would like to see that markers is LD with the same risk variant would result in similar allele-sharing patterns and that the results would be similar using both markers as the conditioning marker. Thus we expect increased power for M1|M2 and M2|M1 for model L-L-H, L-L-M and M-M-L. We observed the expected pattern for all models even for Model L-L-M with modest LD between M2 and D1 ($r^2$=.18). The type I error rate ranged from 0.033 to 0.06. Thus, the APL-OSA conditional test has excellent power to detect the correct association, suggesting that the APL-OSA conditional test is a promising tool for distinguishing markers associated with different disease loci.

To further investigate the utility of the APL-OSA conditional test, we calculated the allele frequency in all pedigrees and specific family subsets identified by APL-OSA in Model L-L-H. The allele frequency for D1 and D2 is 0.19 and 0.17 respectively in 500 families for one replicate in the L-L-M model. As expected, the estimated allele frequency of each disease variant in the subset of families identified by APL-OSA is increased when conditioning on the APL statistic of markers associated with that disease variant (Table 3.4). That is, the D1 risk allele frequency increased from 0.19 to 0.31 when conditioning on M1 or M2 but remained the same as the overall frequency when conditioning on M3. The D2 risk allele frequency increased from 0.17 to 0.41 when conditioning on M3 but remained the same as the overall frequency when conditioning on M1 or M2. This result demonstrates that the

APL-OSA conditional test correctly identifies family subsets with enriched disease allele frequencies.

### 3.3.3 Application to the GENECARD data

Recently, two common sequence variants on chromosome 9p21, tagged by RS10757278-G (22114477 bp) and RS10811661-T (22124094 bp), were reported to be associated with coronary artery disease (CAD) (Helgadottir *et al.* 2007; McPherson *et al.* 2007; The Wellcome Trust Case Control Consortium 2007; Samani *et al.* 2007) and type 2 diabetes (T2D) respectively (Zeggini *et al.* 2007; Scott *et al.* 2007; Saxena *et al.* 2007). These two variants are in adjacent linkage disequilibrium (LD) blocks, and the two SNPs are not correlated ($r^2 < 0.01$) in the HapMap CEU data. The LD block with RS10757278 contains two known genes, *CDKN2A* and *CDKN2B*, which are known to play critical roles in cell proliferation, aging, senescence and apoptosis (Kim & Sharpless 2006), whereas the RS10811661 block does not contain any known genes. The two SNPs, RS10757278 and RS10811661, are separated by less than 10 kb, and *CDKN2A* and *CDKN2B* are the closest genes to both variants. Furthermore, T2D is a well-established risk factor for CAD and atherosclerosis in general, thus raising the possibility that the region on 9p21 could predispose to both CAD and T2D through a shared biological mechanism (Samani *et al.* 2007). Helgadottir *et al.* further investigated these two newly identified sequence variants with CAD, T2D and four arterial diseases. They replicated the previously described association of RS10811661-T with T2D and RS10757278-G with CAD, but did not observe association of RS10811661 with CAD or RS10757278 with T2D. These data support that the two variants on 9p21, tagged by RS10757278-G and RS10811661-T, have independent functions (Helgadottir *et al.* 2008). Thus these two variants on 9p21 appear to be a good example for us to apply conditional APL-OSA.

We analyzed data from the GENECARD project. The sample collection, study design and linkage results of the GENECARD study have been reported (Hauser *et al.* 2003a; Hauser *et al.* 2004a). The family-based GENECARD data set is composed of families with at least two affected siblings who met the criteria for early-onset CAD. The detailed classification criteria for the GENECARD samples were reported previously (Hauser *et al.* 2003a). In brief, the affected individuals had a CAD index event at age ≤ 50 for males and ≤ 55 for females. Unaffected participants were defined as siblings and relatives who had not been diagnosed with CAD and were older than 55 years of age (males) or older than 60 years of age (females). The sample set was consisted of 2,954 affected and unaffected individuals in 1102 families. The majority (~90%) of the GENECARD probands were white.

First, we obtained the CAD association results for RS18011661 (p-value = .25) and RS10757278 (p-value=.037). Our single locus results suggest the findings of a CAD association with RS10757278 but not with RS18011661 consistent with Helgadottir et al. APL statistics for each family for the two SNPs, RS18011661 and RS10757278 were calculated. Conditioning on the APL statistics of RS10757278, ordered from high to low, the association of RS18011661 with CAD was not significant (p-value = .28), while conditioning on the APL statistics of RS10811661, ordered from high to low, the association of

RS10757278 with CAD was still significant (p-value = .034). We have to conclude that conditional APL-OSA should work for multiple risk variants for the same trait.

## 3.4 Discussion

Since complex traits are often determined by multiple genetic and environmental factors with small-to-moderate effects, it is important to investigate the behavior of current association methods under multiple risk variants model. These multiple risk variants can be either from different chromosomes, or from the same chromosome under a linkage peak. The GENECARD study (Hauser *et al.* 2004a) showed linkage evidence to early onset CAD on chromosome 3q13 ranging from 100 to 160 cM (peak multipoint LOD=3.5 at D3S2460, 126.07 cM) in 2004, and Bowden et al. have shown coincident mapping of type 2 diabetes, metabolic syndrome, and measures of cardiovascular disease to a broad region on chromosome 3q (70-160 cM LOD ranging between 1.15 and 2.71) in the Diabetes Heart Study (Bowden *et al.* 2006) in 2006. In this susceptibility region, multiple genes were reported to be associated with familial early-onset coronary artery disease utilizing multiple datasets and functional studies, *GATA2* (Connelly *et al.* 2006) and *KALRN* (Wang *et al.* 2007). In addition, two other genes (*CDGAP* and *MYLK)* in this region belonging to the same Rho GTPase signaling pathway as *KALRN* were also associated with CAD (Wang *et al.* 2007)**.** Association studies suggest that *SLC22A4*, *SLC22A5* on chromosome 5q31 and *CARD15* on chromosome 16q12 act in a common pathogenic pathway to cause Crohn disease, as well as *IL23R* and the adjacent IL-12 receptor, beta-2 gene (*IL12RB2*) on chromosome 1p31 (Peltekova *et al.* 2004; Duerr *et al.* 2006). Multiple risk variants in a local region may either act independently or interact to increase the risk of complex disease. The single variant hypothesis may not apply to these situations, however multiple risk variant model have not been generally explored. Hence we modeled two independent disease loci in a close distance, and by varying LD parameters among M1 and M2, we modeled 5 different models ranging from complete LD to moderate LD. Most simulation studies assume different disease loci from different chromosomes (Hancock *et al.* 2007), while we modeled a specific situation of multiple risk variant coming from a local region.

We compared several association methods under these different multiple risk variants models. The goal of these analyses is to better identify which SNP(s) are important for follow up studies and, to define a more homogeneous family subset within a genetically heterogeneous dataset. For example, SNP(s) that are associated with different disease loci are important while SNP(s) associated with the same disease alleles and in high LD with one another are redundant. Thus, a test that can distinguish between markers associated with different disease loci and that can identify families shared with etiology is extremely valuable.

The power of single marker association tests is closely correlated with the amount of LD between marker and disease loci. These tests maintain good power to detect multiple risk variants in a small region with moderate degree of LD for fully genotyped families. Global tests, like FBAT-LC, are sensitive to the presence of at least one susceptibility variant, but are not helpful for selecting the most promising SNPs. Haplotype tests are another class of test including multiple markers which test combinations of alleles from several markers. If a particular combination of alleles is associated with the disease locus, this test is powerful. Thus a haplotype test is a joint analysis of multiple markers. If multiple haplotypes are associated with different disease loci, the results can be misleading and do not address our research question. The APL-OSA conditional test is an association test conditioning on the APL statistic of each family. If two markers are associated with the same disease locus, we would expect the transmission of their respective alleles to be much more similar compared to alleles associated with a different disease locus or alleles not associated with any disease loci. Thus the APL-OSA conditional test directly addresses our research question. We applied APL-OSA conditional test on our GENECARD samples for two independent SNP markers on 9p21 associated with CAD and T2D respectively. We see the separation of the two markers when conditioning on the CAD associated marker, but did not see the separation when conditioning on the T2D associated marker. Since T2D is a known risk factor for CAD, the joint transmission of the two markers is still possible even though T2D marker is not associated with CAD.

While APL-Haplotype produced misleading results for Model L-L-M and Model M-M-L, the APL-OSA conditional test had the greatest power to properly dissect the clustered associated markers for all models with an acceptable type I error rate ranging from 0.033 to 0.056. Our findings on APL-OSA conditional test from our simulation study suggest that this type of conditional test may be applied to identify multiple risk variants within a gene or multiple genes under a known linkage peak. Our interest is already limited to the associated markers identified by the single-marker association tests. Thus, it is recommended to run the single marker association tests first, then run the APL-OSA conditional test on those associated markers (preferably in one gene or a small region) to detect which markers may be associated with the same disease locus versus multiple independent loci. These partitioned results can help to design further searches for functional SNPs within candidate genes, additional studies of allele-specific results in these genes, and suggest additional statistical models. Standard tests may not be applying the appropriate model, and genetic/allelic heterogeneity will hamper out ability to interpret these tests. Our work has made contributions to understand the performance of association tests under complex disease models.

## 3.5 Tables

**Table 3.1** Outline of simulated data sets with varying LD

| | | Haplotype 1 | | | Haplotype 2 |
|---|---|---|---|---|---|
| **Num** | **Model** | **LD (M1-D1)** | **LD (M2-D1)** | **LD (M1-M2)** | **LD (M3-D2)** |
| **1** | **H-Zero-Zero** | **.97** | **0** | **0** | **1** |
| **2** | **H-H-H** | **.99** | **1** | **.99** | **1** |
| **3** | **L-L-H** | **.38** | **.39** | **.98** | **1** |
| **4** | **L-L-M** | **.35** | **.18** | **.61** | **1** |
| **5** | **M-M-L** | **.54** | **.73** | **.38** | **1** |

Note: Linkage disequilibrium (LD) is measured by $r^2$ for makers in Haplotype 1 associated with disease locus 1 (D1) and Haplotype 2 associated with disease locus 2 (D2). Relative risks were 3.0 for each disease locus, and disease allele frequency is 0.15 for each locus.

**Table 3.2** Comparison of APL, LAMP and FBAT single marker tests in Models 1, 2, 3, with 2 disease susceptibility alleles each with relative risk of 3.0. (1000 replicates)

| Model 1 H-Zero-Zero | Marker | Percentage (%) of significant results among 1000 replicates Significance level set at 5% | | | | |
|---|---|---|---|---|---|---|
| | | APL | LAMP-Linkage | LAMP-LE $H_0$: $r^2$=0 | LAMP-LD $H_0$: $r^2$=1 | FBAT |
| | M1 | 72.5 | 12.8 | 100 | 3.2 | 99.3 |
| | M2 | 4.2 | 4.1 | 19.6 | 10.4 | 4.7 |
| | M3 | 72.4 | 12.4 | 100 | 2.6 | 99.5 |
| | M4 | 4.6 | 4.5 | 16.4 | 10.2 | 5.1 |
| **Model 2 H-H-H** | **Marker** | **Percentage (%) of significant results among 1000 replicates Significance level set at 5%** | | | | |
| | | APL | LAMP-Linkage | LAMP-LE $H_0$: $r^2$=0 | LAMP-LD $H_0$: $r^2$=1 | FBAT |
| | M1 | 70.7 | 12.7 | 100 | 3.8 | 99.6 |
| | M2 | 71.2 | 13.3 | 100 | 3.2 | 99.7 |
| | M3 | 74.0 | 12.5 | 100 | 3.1 | 99.6 |
| | M4 | 6.6 | 4.1 | 20.3 | 10.0 | 4.7 |
| **Model 3 L-L-H** | **Marker** | **Percentage (%) of significant results among 1000 replicates Significance level set at 5%** | | | | |
| | | APL | LAMP-Linkage | LAMP-LE $H_0$: $r^2$=0 | LAMP-LD $H_0$: $r^2$=1 | FBAT |
| | M1 | 34.9 | 6.8 | 93.8 | 9.6 | 59.5 |
| | M2 | 34.2 | 8.2 | 94.5 | 11.0 | 61.6 |
| | M3 | 74.7 | 17.1 | 100 | 3.4 | 99.3 |
| | M4 | 4.8 | 4.1 | 15.8 | 10.3 | 4.0 |

**Table 3.3** Comparison of APL-OSA conditional test, APL-Haplotype and FBAT-LC in Model 1, 3, 4 and 5, RR (D1, D2) = 3.0, 1000 replicates

| Association Tests | Markers in the tests | Model1 H-Zero-Zero | Model 3 L-L-H | Model 4 L-L-M | Model 5 M-M-L |
|---|---|---|---|---|---|
| APL-Haplotype | M1-M2-M3 | 70.2 | 60.4 | 52.8 | 67.8 |
| | M1-M2 | 49.2 | 18.6 | 23.3 | 44.6 |
| | M1-M3 | 83.6 | 66.7 | 67.4 | 73.1 |
| | M2-M3 | 53.8 | 67.5 | 62.3 | 77.9 |
| FBAT-LC | M1, M2, M3 | 100 | 99.8 | 99.8 | 100 |
| | M1, M2 | 99.4 | 71.2 | 60.4 | 98.3 |
| | M1, M3 | 100 | 100 | 100 | 99.9 |
| | M2, M3 | 99.6 | 100 | 99.8 | 100 |
| APL-OSA conditional test* | M1\|M2 | 5.6 | 100 | 100 | 100 |
| | M1\|M3 | 4.6 | 3.3 | 4.6 | 3.8 |
| | M2\|M1 | 5.6 | 100 | 100 | 100 |
| | M2\|M3 | 4.9 | 3.4 | 3.8 | 4.2 |
| | M3\|M1 | 4.8 | 4.5 | 4.7 | 4.6 |
| | M3\|M2 | 4.9 | 4.1 | 4.4 | 4.4 |

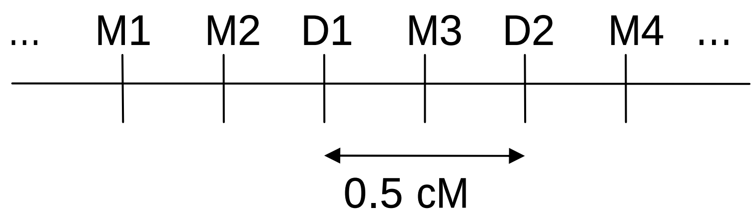
Note: values in the table are percentage (%) of significant results among 1000 replicates with significance level set at 5%. * Conditional test of the analyzing marker given the conditioning marker, i.e., M1|M2 is the conditional test of M1 conditioning on M2.

**Table 3.4** Allele frequency of D1 and D2 in family subset identified by APL-OSA in one replicate of Model L-L-H

| | p(D1) | p(D2) | | p(D1) | p(D2) | | p(D1) | p(D2) |
|---|---|---|---|---|---|---|---|---|
| M1\|M2 | 0.30 | 0.18 | M2\|M1 | 0.31 | 0.18 | M3\|M1 | 0.19 | 0.17 |
| M1\|M3 | 0.19 | 0.17 | M2\|M3 | 0.19 | 0.17 | M3\|M2 | 0.19 | 0.17 |

Note: Analysis marker | Conditioning maker, i.e., M1|M2.

## 3.6 Figures




D1: Disease Locus 1;
D2: Disease Locus 2;

M1, M2: Marker 1 and Marker 2 are in Haplotype 1, associated with D1;
M3: Marker 3, associated with D2, in Haplotype 2;
M4: Marker 4;

See Table 1 for LD parameters.

**Figure 3.1** Map of simulated markers

# Chapter 4

## Modeling joint contributions of genes and HDL-C level to CAD risk

## 4.1 Introduction

Cardiovascular disease, in particular coronary artery disease (CAD) is one of the leading causes of death in the United States (Thom *et al.* 2006). CAD is a complex disease, and the formation of CAD involves both genetic and environmental factors. Many risk factors have been revealed by classical epidemiologic studies, including age, gender, hypertension, dyslipidemia, diabetes mellitus, smoking, and physical inactivity (Rissanen 1979; Ten Kate *et al.* 1982; Shea *et al.* 1984; Marenberg *et al.* 1994; Zdravkovic *et al.* 2002; Voss *et al.* 2002; Lloyd-Jones *et al.* 2004; Yusuf *et al.* 2004; Murabito *et al.* 2005). In addition, numerous studies have repeatedly shown that there is a strong genetic component underlying this complex disease (Rissanen 1979; Shea *et al.* 1984; Marenberg *et al.* 1994).

The genetic contribution to development of CAD has been extensively studied by many investigators. Ten genome-wide linkage scans for CAD and myocardial infarction (MI) have been conducted (Pajukanta *et al.* 2000; Francke *et al.* 2001; Harrap *et al.* 2002; Broeckel *et al.* 2002; Helgadottir *et al.* 2004; Wang *et al.* 2004; Hauser *et al.* 2004a; Samani *et al.* 2005; Farrall *et al.* 2006; Bowden *et al.* 2006), and these scans have identified linkage regions on chromosome 1, 2, 3, 4, 5, 7, 12, 13, 14, 16, 17, 19, and X. Within these linkage regions the TNFSF4 gene on chromosome 1 (Wang *et al.* 2005), the VAMP 8 gene on chromosome 2 (Shiffman *et al.* 2006), and the ALOX5AP gene on chromosome 13 (Helgadottir *et al.* 2004), have been reported as putative CAD susceptibility genes. A highly significant association between myocardial infarction (MI) and a common sequence variant on chromosome 9p21 has been identified, and this variant is adjacent to the tumor suppressor genes CDKN2A and CDKN2B (Helgadottir *et al.* 2007; McPherson *et al.* 2007).

Of particular interest is a region on chromosome 3q13-21, our GENECARD study (Hauser *et al.* 2004a) showed linkage evidence to early onset CAD on 3q13 ranging from 100 to 160 cM (peak multipoint LOD=3.5 at D3S2460, 126.07 cM). Bowden et al. showed coincident mapping of type 2 diabetes, metabolic syndrome, and measures of cardiovascular diseases to a broad region on chromosome 3q (70-160 cM LOD raging between 1.15 and 2.71) in the Diabetes Heart Study (Bowden *et al.* 2006). The susceptibility region mapped to Chromosome 3q13-21 in these two independent studies is shown in Figure 4.1. We have furthermore performed association mapping of genetic variants within this region using an independent case-control cohort (the CATHGEN study). Connelly et al. reported GATA2 is associated with familial early-onset coronary artery disease utilizing gene expression patterns within human donor aortas and an association study of CATHGEN samples and GENECARD probands (Connelly *et al.* 2006), and the LSAMP gene was identified in late onset and severe CAD (Wang *et al.* 2008). Wang et al. identified KALRN as a novel candidate gene for coronary artery disease with multiple datasets and a functional study (Wang *et al.* 2007). In addition, two other associated genes (CDGAP and MYLK) participating in the same Rho GTPase signaling pathway as KALRN and located under the 3q13 linkage peak are also reported. The KALRN gene was also identified in a genome wide scan association of early onset CAD (Stewart *et al.* 2007).

Meanwhile, there is linkage evidence that this region harbors a gene or genes determining high-density lipoprotein (HDL) cholesterol levels. Imperatore et al. showed the strongest evidence for linkage with HDL cholesterol is also on chromosome 3q (LOD score 2.64) near D3S3053 (174.4 cM) (Imperatore *et al.* 2000). QTL linkage analysis done by Shah et al. showed the strongest QTL for HDL-C on chromosome 3p14 (95cM, LOD 1.65, p=0.005) (Shah *et al.* 2006). However an additional modest QTL linkage peak for HDL/TC ratio was observed from 150 – 200 cM on chromosome 3. An ordered subset analysis (OSA) to identify covariates related to strength of linkage evidence showed an increased LOD score at 140.4 cM in GENECARD families specifically with favorable HDL-C (Shah *et al.* 2006). The increased LOD score from OSA analysis was also observed in the 3q region at 124 -148 cM subsetting on low HDL-C in the Diabetes Heart Study (Bowden *et al.* 2006).

The genetic contributions to HDL-C levels have been well studied. A review in 2005 noted 45 HDL-C QTLs in genome-wide scans in 15 different human populations by QTL analysis, and 30 unique HDL-C QTLs in humans (Wang & Paigen 2005). In addition, 54 candidate genes that are known to regulate HDL metabolism and functions were reported, including genes on 3p25.2 and 3q29 (Wang & Paigen 2005). To date, no other genes known to regulate HDL-C level have been found on chromosome 3 except for APOD (3q29) and PPARG (3p25.2). More recently genome wide association (GWA) studies of HDL-C have identified additional genomic region as candidate loci. In early 2008, several GWAs published genes for biomarkers of cardiovascular disease and dyslipidemia. Wallace et al. identified a SNP rs905648 (COLQ, 15540326 bp, p-value=4.58E-6) associated with HDL-C on chromosome

3, and several other associated SNPs with HDL on chromosome 1, 8, 9, 15, 16 and 19 (Wallace *et al.* 2008). Willer et al. used genome-wide association data from three studies (n=8816 individuals), and they identified strongly associated variants in eleven loci previously reported in lipid metabolism and two newly identified loci (near MVK-MMAB on chromosome 12 and GALN2 on chromosome 1) with variants primarily associated with HDL-C (Willer *et al.* 2008). Kathiresan et al. used the same data as Willer et al. and identified the rs4846914 in GALNT2 on 1q42 with HDL-C as well (Kathiresan *et al.* 2008). The Framingham Heart Study reported a genome wide association with significant SNPs on chromosome 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 14, 17, 18, 19, 20 and 21, including three SNPs in our chromosome 3q13 linkage peak region, rs6804032 (122.54 cM, ZBTB20), rs4459845 (133.15 cM, ALDH1L1) and rs1394041 (153.36 cM, in a non-gene region) (Kathiresan *et al.* 2008).

The ubiquitous linkage evidence for CAD and multiple quantitative trait loci for HDL-C demonstrate the complexity of the pathogenesis of CAD and HDL-C. The hypothesis that complex traits like CAD and HDL-C are often determined by multiple genetic and environmental factors with small to moderate effects is well accepted. The models for the relationship between a quantitative trait like HDL-C and a disease process like CAD can be very complicated as shown in Figure 4.2 (Schmidt *et al.* 2007). HDL-C level is a known risk factor for CAD (Yang *et al.* 2005). Thus the genetic models to be considered include both HDL-independent as well as HDL-dependent risks of CAD. These models could include a model where genetic variants affect the levels of HDL-C (QTL model), or genetic factors affect HDL-C which in turn modify the risk of CAD (Modifier model, dashed line). Alternatively these models may represent the interaction effect of genetic variants and the level of HDL-C affecting the risk of CAD (Interaction model). The genetic variants for either CAD risk or HDL-C level could be single or multiple, and they could act independently or collectively (Heterogeneity model) (Schmidt *et al.* 2007). The purpose of this paper is to identify genetic variants for quantitative trait HDL-C levels in this region and apply these models in the statistical analysis of SNPs association. The results of these models may begin to dissect the complex relationship among HDL-C levels, CAD and genetic risk variants in the chromosome 3q13-21 region.

## 4.2 Materials and Methods

### 4.2.1 Early-onset CAD family-based sample (GENECARD).

GENECARD is a collaborative study involving investigators affiliated with the Duke Center for Human Genetics, the Duke University Center for Living, the Duke Clinical Research Institute, the Duke University Consortium for Cardiovascular Studies, and five additional international investigative sites of the GENECARD Study Network. The study is coordinated at Duke University, and the study design has been previously reported (Hauser *et al.* 2004a). In brief, collection of families began in March 1998 and was completed on 31 March, 2002. All study participants signed a consent form approved by the responsible institutional review board or local ethics committee.

The sample set used for the initial genome-wide linkage scan in the GENECARD project consisted of 493 affected sibling pairs in 420 families, where at least two siblings met the criteria for early-onset CAD. The detailed classification criteria for the GENECARD samples have been published (Hauser *et al.* 2003a). In brief, the affected individuals had a CAD index event at age ≤ 50 for males and ≤ 55 for females. The characteristics of this study group are summarized in Table 4.1 and elsewhere (Connelly *et al.* 2006). The follow-up collection includes an additional 681 families together with a large number of unaffected participants from all families. Unaffected participants were defined as siblings and relatives who have not been diagnosed with CAD with a stringent definition of unaffected status requiring age older than 55 years (males) or older than 60 years (females). This additional collection has increased our sample size to 2,954 affected and unaffected individuals (Table 4.1).

**4.2.2 Case-control sample (CATHGEN).**

CATHGEN participants were recruited sequentially through the cardiac catheterization laboratories at Duke University Hospital (Durham, North Carolina, United States) with approval from the Duke Institutional Review Board. All participants undergoing catheterization were offered participation in the study and signed informed consent. Medical history and clinical data were collected and stored in the Duke Information System for Cardiovascular Care database maintained at the Duke Clinical Research Institute.

Controls and cases were chosen on the basis of extent of CAD as measured by the CAD index (CADi). CADi is a numerical summary of coronary angiographic data that incorporates the extent and anatomical distribution of coronary disease (Smith *et al.* 1991). The CADi may be considered a quantitative phenotypic measure of coronary atherosclerosis that increases as coronary atherosclerosis increases. An affected case was defined by the presence of significant CAD with a CADi ≥ 32 (Felker *et al.* 2002). For patients older than 55 years of age, a higher CADi threshold (CADi ≥ 74) was used to adjust for the higher baseline extent of CAD in this group. Medical records were reviewed to determine the age of onset (AOO) of CAD, i.e., the age at first documented surgical or percutaneous coronary revascularization procedure, myocardial infarction, or cardiac catheterization meeting the above-defined CADi thresholds (Connelly *et al.* 2006). In summary, the CATHGEN cases with an age at onset ≤ 55 years for females and ≤ 50 years for males were classified as a young affected group (YA), which provides a consistent comparison for the GENECARD family study, whereas cases with an age at onset ≥ 56 years for females and ≥ 51 years for males and CADi >= 74 were stratified in the old affected group (OA). The CATHGEN controls (ON) were defined as ≥ 60 years of age, with no CAD as demonstrated by coronary angiography and no documented history of cerebrovascular or peripheral vascular disease, myocardial infarction, or interventional coronary revascularization procedures. A comparison of clinical characteristics between GENECARD probands and affected and unaffected CATHGEN samples is presented in Table 4.2, showing the expected risk profiles for sets of unaffected and affected individuals.

### 4.2.3 Measurement of lipids

The total HDL cholesterol levels for CATHGEN subjects were calculated lipids derived from nuclear magnetic resonance (NMR) spectroscopic analysis (Jeyarajah *et al.* 2006). The correlation of calculated lipids to those chemically measured is strong for HDL. HDL levels for GENECARD samples were either measured by standard enzymatic methods at time of ascertainment or reported in the medical history. The measured lipids data are all fasting cholesterols. Since the distribution of HDL-C values does not conform to a normal distribution, a $\log_{10}$ transformation was used for HDL-C in both data sets to approximate a normal distribution (Shah *et al.* 2006).

### 4.2.4 SNP Selection and Genotyping

SNPs were assembled for the HDL-C analysis from 3 genetic studies of this region. The first set of SNPs is selected across the one-LOD unit down linkage peak region of 3q13-21 at an average of 100 kb intervals (Wang *et al.* 2007). The second set of SNPs comes from candidate genes derived from an expression study in aorta samples (Seo *et al.* 2004). The detailed SNP selection mechanism using the software program SNPSelector (Xu *et al.* 2005) for these two sets of SNPs was previously described (Wang *et al.* 2007). The third set of SNPs is the 3 SNPs in the 3q13 region identified by the Framingham Heart study which demonstrated association with mean HDL-C (Kathiresan *et al.* 2008).

Genomic DNA for the GENECARD and CATHGEN samples was extracted from whole blood using the PureGene system (Gentra Systems, Minneapolis, Minnesota, United States).

Genotyping in GENECARD was performed using the ABI 7900HT Taqman SNP genotyping system (Applied Biosystems, Foster City, California, United States). Genotyping in CATHGEN was performed using the Illumina BeadStation 500G SNP genotyping system (Illumina, San Diego, California, United States). For the purpose of quality control, 2 Centre d'Etude du Polymorphisme Human (CEPH) pedigree individuals (Dausset *et al.* 1990), and a minimum of 4 duplicated quality-control samples were included in every genotyping run. Results of the CEPH and quality-control samples were compared to identify possible sample plating errors and genotype calling inconsistencies. SNPs that showed mismatches on quality-control samples were reviewed by an independent genotyping supervisor for potential genotyping errors. All SNPs examined were successfully genotyped for 95% or more of the individuals in the study. Error rate estimates for SNPs meeting the quality control benchmarks were determined to be less than 0.2%. Hardy-Weinberg equilibrium (HWE) testing was performed for all markers in the affected and unaffected groups. SNPs with departures from HWE (cut-off p-value $\leq .0001$) or with minor allele frequency less than 0.05 were excluded from this study. 672 SNPs genotyped in CATHGEN and 215 SNPs genotyped in GENECARD met the quality control standards. The SNPs genotyped in GENECARD are a subset of CATHGEN SNPs. For each pair of SNPs located less than 500 kb apart, a measure of the strength of linkage disequilibrium (LD), $r^2$, was calculated with the Graphical Overview of Linkage Disequilibrium (GOLD) program (Abecasis & Cookson 2000) and displayed using Haploview (Barrett *et al.* 2005).

### 4.2.5 Statistical Analysis

To evaluate the relationship between CAD, HDL-C levels and genetic variants we used different approaches depending upon the sample set and the outcome. For the outcome of CAD, we used multivariable logistic regression and for the outcome of HDL-C we used multivariable ANOVA analysis in the case-control data. When analyzing these outcomes in the family data, we used APL for the outcome of CAD, and used QTDT for the out come of HDL-C. For the interaction between HDL-C and genetic markers on CAD status, multivariate logistic regression was used for case-control data, and multivariate logistic regression with the generalized estimating equations (GEE) method was used for family data. Generalized Estimating Equations (GEE) was introduced by Liang and Zeger (1986) (Liang & Zeger 1986) as a method of estimation of regression model parameters when dealing with correlated data. Multivariate logistic regression with GEE method is a valid and robust method for family data for genotype correlations within pedigrees (Hancock *et al.* 2007). Models included adjustment for potential confounding variables that depend on the outcome. For analysis of CAD, all procedures were adjusted for race and sex. For analysis of HDL-C, all procedures were adjusted for sex, race, age at exam, body mass index (BMI) and smoking history (more than half a pack per day). SAS version 9.1 (SAS Institute, Cary, NC) was used for all statistical analyses.

To evaluate the relative strength of each SNP under the various models (Figure 4.2), we took a stepwise approach to perform the statistical analysis as shown in Figure 4.4. The regional genomic coverage of the CATHGEN study is more complete across the 3q13-21 region than the GENECARD study. Thus we started by screening for HDL-C associated SNPs in CATHGEN, and then investigated the relationship between genetic markers with HDL-C and CAD by testing the CAD association directly with logistic regression and then by testing the interaction between HDL-C level and genetic markers on CAD status on the resulting associated HDL-C SNPs. SNPs that did not show a significant interaction effect were assigned to the QTL model. Under the assumption of a QTL model, we would expect to see an association between HDL-C level and genotype in both affected and unaffected individuals. To test this assumption, we divided the QTL SNPs into the ALL, YA, OA and ON group by their association results with HDL-C. SNPs which showed a significant interaction effect were separated into two groups based on if they are associated in the ON group. SNPs with an interaction effect and no association in ON belong to Modifier I group, and this group of SNPs increased the risk of CAD via HDL-C levels. SNPs with interaction effect and association only in ON group belong to Modifier II group, and this group of SNPs reduced risk of CAD via HDL-C levels. SNPs with interaction effect and association in either affected or control group were assigned to Interaction group. The standard for model assignment is illustrated in Table 4.8. Finally, for all these groups of SNPs, we studied their joint association with HDL-C with backward stepwise linear regression. Details of each of these models are described below.

To validate the significance of the multiple HDL-C associated SNPs tested in the CATHGEN data set, SNPs that had a nominal p-value <.05 were examined in the GENECARD family data set. The CAD association was examined with APL on the validated SNPs, and QTDT was used for HDL-C association. The QTL and modifier models were tested with multivariable logistic regression analysis. Backward stepwise linear regression of the HDL-C outcome was preformed on SNPs with or without interaction effect.

**Relationship between CAD and genetic variants (CAD association)**

Allelic and genotypic association with CAD was analyzed using logistic regression analysis, with adjustment for race and sex. To test for linkage and family-based association, the Association in the Presence of Linkage (APL) was performed for GENECARD data. In nuclear families, APL provides a test for association that correctly infers missing parental genotypes by estimating identity-by-descent parameters (Martin *et al.* 2003). A p-value < .05 is considered significant for inclusion of this SNP in subsequent tests.

**Relationship between HDL and genetic variants (QTL model)**

$$HDL_i = \alpha + \beta_1 * sex_i + \beta_2 * race_i + \beta_3 * age_i + \beta_4 * BMI_i + \beta_5 * smoking_i + \beta_G * genotype_i + e_i;\ e \sim N(\mu, \sigma^2)$$

Regression-based analysis is a very flexible tool and allows inclusion of risk factors and to add interaction terms into analysis. A multivariable ANOVA analysis was used to exam the relationship between SNPs and HDL-C levels for case-control data. In this analysis, generalized linear models were used to compare differences in mean levels of quantitative traits by genotypes adjusting for sex, race, age-at-exam, BMI and smoking history. The QTL model tests the null hypothesis of $\beta_G = 0$ versus the alternative that $\beta_G \neq 0$, assuming an additive model.

The Quantitative Transmission Disequilibrium Tests (QTDT) (Abecasis *et al.* 2000a) was used to test for association with quantitative trait HDL-C for family data with adjustment for sex, race, age at exam, BMI and smoking history. QTDT can use all the information in a pedigree to construct powerful tests of association that are robust in the presence of population stratification.

**Relationship between CAD, HDL and genetic variants (Interaction model)**

$$\ln\left(\frac{p_i}{1-p_i}\right) == \alpha + \beta_1 * sex_i + \beta_2 * race_i + \beta_3 * age_i + \beta_4 * BMI_i + \beta_5 * smoking_i$$

$$+ \beta_E * HDL_i + \beta_G * genotype_i + \beta_I * HDL_i * genotype_i + e_i;$$

$$e \sim N(\mu, \sigma^2)$$

A multivariable logistic regression analysis was performed for interaction between HDL and genetic variants on CAD disease status with adjustment for race, sex, smoking history, age at exam and BMI on every sample for both the case-control dataset and the family dataset. The family dataset applied the GEE method to adjust for the correlation within families.

**Relationship between HDL and multiple genetic variants**

To evaluate the independent effect of each associated marker, backward stepwise linear regression of multiple markers was performed for case-control data, and backward stepwise linear regression with GEE modeling the correlation among family and recruiting centers was used for family data. SNPs in high LD (>.7) are not included in the model selection to avoid collinearity problems.

## 4.3 Results

### 4.3.1 Clinical Characteristics of Study Subjects

The CATHGEN dataset consists of 920 cases (656 young affected samples, 264 old affected samples), 405 old controls, and 1265 of 1325 individuals in CATHGENE have total HDL-C values. The GENECARD dataset consists of 1101 families and 1338 of 2954 individuals with HDL-C values. Baseline clinical characteristics are reported in Table 4.1. Comparison between affected and unaffected individuals was done using a univariate t-test or chi-square test as appropriate. Samples are mostly Caucasians in both sets. Consistent with other studies, there was a high prevalence of CAD risk factors among affected individuals, especially young affected individuals, including hypertension, diabetes mellitus, history of smoking, dyslipidemia and family history of CAD. However, the mean values of total cholesterol, LDL cholesterol and triglycerides were not highest in the affected groups, consistent with the higher use of lipid medications in the affected group. HDL Cholesterol level is robust to the lipid medications, and as expected the affected groups have lower mean HDL-C levels than the ON group.

### 4.3.2 Evaluation of HDL-C Association with single markers

**Results of the QTL model**

Preliminary evidence of association with HDL-C (p-value≤.05) was observed for 37 SNPs in the combined group (ALL), and the CAD association p-values, HDL-C association p-values and interaction analysis p-values of the top 20 HDL-C associated SNPs in the ALL group in case-control dataset are listed in Table 4.3. The original studies were focused an age-specific effects in CAD and then we divided the analysis into separate groups of young affected, old affected subjects and controls. In the analysis of the YA subset, evidence of association was found for 29 SNPs, and 34 SNPs in the OA group and 27 SNPs in the ON group, and the CAD association p-values, HDL-C association p-values and interaction analysis p-values of the top 20 HDL-C associated SNPs in the YA, OA and ON group in case-control dataset are listed in Table 4.4, 4.5 and 4.6 respectively. Jointly we identified 108 HDL-C associated markers with p-value<.05. The majority of the SNPs are separated in different sub-groups as shown in Figure 4.5, supporting our further analysis done by sub-groups. There is very little overlap in the groups, with no SNPs significant in the ALL group and also in all three subsets.

The most significant association is at SNP rs2979307 in ON group (Table 4.6), which resides in the OSBPL11 gene (p-value=5.13E-5). OSBPL11 is a member of the OSBP family of intracellular lipid receptors. The mean level of HDL-C increases from 49.81 (14.1) to 51.0 (43.7) as the number of minor allele G increases at SNP rs2979307, and this SNP is not directly associated with CAD, nor does it have interaction effect on CAD. This SNP is a candidate marker for HDL-C level, however there is no evidence for association with CAD.

SNP rs1132200 in TMEM39A appeared most of times for it is significant in ALL (p-value=0.007), YA (p-value=0.0329) and OA (p-value=0.0441) three groups (Table 4.3, 4.4 and 4.5). However we did not detect an association with CAD in the case-control and family based association analysis. The genotype specific HDL-C level of rs1132200 varies in different subgroups. The mean HDL-C level for rs1132200 major allele homozygote is higher than that of minor allele homozygote in ALL, YA and ON group, but it is lower in OA group. It is another candidate marker for HDL-C QTL.

To evaluate the significance of the HDL-C associated SNPs tested in the CATHGEN data set, SNPs that had a nominal p-value ≤.05 were examined in the GENECARD family data set with QTDT analysis. Forty-four HDL-C associated SNPs were genotyped in the GENECARD dataset, and QTDT analysis was performed on these SNPs in GENECARD. The CAD association p-values, HDL-C association p-values and interaction analysis p-values of the top 6 HDL-C associated SNPs are listed in Table 4.7, and 4 SNPs validated their significance at the level of .05 in the GENECARD data set. SNP rs869851 (p-value=.0494) in the KALRN gene, rs9831023 (p-value=.0301) in the CDGAP gene, and rs1132200 (p-value=.0456) in the TMEM39A gene were also significant in the analysis of the young affected cases in the CATHGEN data set. TMEM39A stands for transmembrane protein 39A, and it has not been annotated.

**Results of the Interaction model**

For the total 108 HDL-C associated SNPs (p-value≤.05), only one SNP rs9289235 in the KALRN gene is associated with CAD in the case-control association analysis (p-value=.0408). The number of markers assigned to different model group depending on the results from the 3 statistical analyses for CATHGEN dataset is listed in Table 4.8. The interaction analysis between the 108 SNP markers and HDL-C levels on CAD disease status resulted in 22 markers with significant interaction effect (p-value≤.05). Ten SNPs out of the 22 SNPs do not have HDL-C associations in the ON group, and we consider the 10 SNPs are candidate SNPs for modifier I model which increase the risk of CAD affected individuals by affecting the level of HDL-C. Eight SNPs out of the 22 SNPs have HDL-C associations just in ON group, and we consider the 8 SNPs are candidate markers for modifiers of CAD which reduce the risk of CAD. The other 4 SNPs have HDL-C associations in the ON group and the ALL group, and we consider them as candidate SNPs for interaction model. As for the 86 SNPs without significant interaction effect (p-value≤.05), including rs2979307 and rs1132200, we consider them candidate SNPs for HDL-C QTLs, except for SNP rs9289235 because it also has CAD association. The association evidence for both CAD (p-value=.0408) and HDL-C (p-value=.0107 in ALL group) at rs9289235 in KALRN gene makes it a candidate marker for pleiotropic effect.

In the GENECARD dataset, SNP rs869851 (p-value=.0494) in the KALRN gene has significant interaction with HDL-C (p-value=0.0281) on CAD, but no direct association with CAD. It could be a modifier marker. SNP rs2860228 in GATA2 gene (p-value=.0428) showed HDL-C association also in CATHGEN ON group (p-value=.0313) and interaction effect (p-value=.0481). It could also be a modifier marker.

### 4.3.3 Evaluation of multiple markers and HDL-C Association

Pairwise LD between SNP genotyped in GENECARD and CATHGEN datasets are shown in Figure 4.3. The Haploview plot of these data in the unaffected population (Figure 4.3) shows several blocks of LD in SNPs in CATHGEN dataset, but most SNPs (>85%) were not in high LD ($r^2$<0.7) with each other, and >90% SNPs in GENECARD dataset were not in high LD ($r^2$<0.7) with each other.

Within the statistical models, we evaluated the independence of the associated markers in different groups, and backward stepwise linear regression was performed for both data sets with adjustment for race, sex, smoking, BMI and age at exam on transformed HDL-C by groups. In CATHGEN data set, for the Modifier I group, 2 SNPs were left in the final model, rs17681395 in ZBTB20 and rs1850719 in LSAMP gene. For Modifier II group, all 9 SNPs were kept in the final model, including rs2873688 and rs12639070 in the CDGAP gene. For the interaction group, two SNPs rs626364 in CD80 and rs3772759 in the KALRN gene were left in the final model. For the QTL model, model selection was done by groups. 10 SNPs were left in final model for the ALL group, including rs6784285 in PHLDB2 gene, rs4687991 in LSAMP gene, rs7650193, rs2044621 and rs10049044 in CDGAP gene, rs1132200 in TMEM39A, rs4678047 in MYLK gene and rs3772587 in AGTR1 gene. 3 SNPs were left for YA group, including rs935616 in CDGAP and rs1132200 in TMEM39A. 16 SNPs were left in final model for the OA group, including rs2682215 and rs16834826 in MYLK gene, rs4678100 and rs332410 in KALRN gene, rs2659689 in GATA2 gene, rs275645 in AGTR1, and rs2084382 in PAK2 gene. 5 SNPs were left in final model for the

ON group, including rs27647 in GHRL gene, rs11922594 in CDGAP gene. No markers were left in the final model for GENECARD data. The SNPs in the final models for each analysis are listed in Table 4.9. The SNPs in different groups are totally exclusive, and thus there is little evidence for a single SNP or a few SNPs controlling both HDL-C levels and CAD risk. We do observe several familiar genes appearing more than once, such as KALRN, CDGAP, TMEM39A, MYLK and AGTR1.

The 3 associated SNPs reported from the Framingham study are not confirmed in CATHGEN and GENECARD dataset for HDL-C association, but we did observe some association with the ZBTB20 gene.

## 4.4 Discussion

We performed the ANOVA analysis on the transformed HDL-C levels, adjusting for race, sex, smoking, BMI and age at exam for case-control data, and we investigated the relationship between genetic markers, CAD and HDL-C by direct association analysis with CAD and indirect association with CAD by interaction with HDL-C. We identified 37 SNPs associated with HDL-C at nominal p-value .05 in YA group. We validated the association with HDL-C of three of these SNPs, rs869851 (KALRN), rs9831023 (CDGAP) and rs1132200 (TMEM39A), in an independent family dataset. We separated the HDL-C associated SNPs into different groups, suggesting possible different functions for them. We identified several different markers in the KALRN gene that could act in different settings.

Combining previous strong association evidence at rs9289231 in KALRN gene (Wang *et al.* 2007) and several associated markers in MYLK, CDGAP and PAK2 genes, we suggest KALRN gene and the Rho GTPase signal-transduction pathway play an important role in the pathogenesis of CAD and HDL-C, although the relationship of any individual SNP with either CAD risk or HDL-C level remains obscure.

One of the most difficult aspects in genetic association studies is how to properly evaluate the significance of many tested SNPs. Different methods have been proposed, including Bonferroni correction, correction by false-discovery rate (Benjamini & Hochberg 1995), and replication in a second data set. A simulation study from the Genetic Analysis Workshop 14 found that a multiple-testing correction is usually too conservative, whereas replication of a nominal p-value in a second data set is less stringent (Shephard *et al.* 2005). In our case we report uncorrected p-values and we sought to validate the associations in another independent data set to filter out the false-positive association. However, it is worth noting that the strongest association at rs2979307 in the OSBPL11 gene survives a Bonferroni correction. Another complication in our study is that we have multiple phenotypes, CAD status and HDL-C levels, and there is a question if the association tests for the two phenotype are independent. We simulated (Schmidt *et al.* 2005) 500 cases and 500 controls under null hypothesis that no marker is associated with either trait for 1000 replicates. The simulation results for CAD association test and the HDL-C association test demonstrated that the two tests are independent (data not shown). Thus we can set the significance level for each test at 0.05.

We observed different HDL-C trends with HDL-C associated SNPs. Some SNPs have the highest HDL-C level for heterozygous genotype, and some have increasing or decreasing HDL-C values when the number of minor allele increases. This phenomenon suggests that the genetic basis for a complex trait such as HDL-C is highly heterogeneous. Additionally, the lack of a consistent clinical definition of CAD across studies further confounds genetic analysis. The detection of multiple, significant but non-overlapping chromosomal regions in ten genome screens indicate the mixed genetic and phenotypic characteristics of CAD. We also see evidence of this heterogeneity in our own dataset, both phenotypic (Shah *et al.* 2006) and in the linkage scores. We failed to detect evidence for linkage on Chromosome 3 in our follow-up dataset (maximum multipoint LOD score = 0.0 at 126 cM). Bowden et al. (Bowden *et al.* 2006) has suggested a replication of the linkage region on Chromosome 3q13 we initially reported. Even with the evident heterogeneity in our CAD population, we detected several association signals with SNPs in KALRN, MYLK, CDGAP and PAK2 genes in two separate CAD datasets for HDL-C. We suggest that our validation of these association warrants further study in our datasets, including an expansion to other racial groups such as Asians and African Americans, as well as in other studies to understand what roles these genes might play in CAD and HDL development.

The results of our study are important for two main reasons. First, our study results are important because of the strength of identifying significant associations with complex disease arising from a hypothesis-driven experimental design. We identified SNPs in a genomic region selected on the basis of multiple lines of evidence (Hauser *et al.* 2003b; Hauser *et al.* 2004a; Bowden *et al.* 2006). We used a stratified case-control data set and a phenotypically similar family dataset to validate the HDL-C associations we observed in CATHGEN YA group, as well as a test of association (QTDT) that allows us to make full use of the GENECARD general pedigrees, which allowed us to identify three separate associations with HDL-C. We identified associated SNPs by different models.

Second, complex genetic diseases, such as cardiovascular disease and HDL-C QTL, are most likely the result of an accumulation of multiple small genetic effects that influence an individual's capability to cope with biological and environmental effects (Cohen *et al.* 2006). The interaction or sequential regulation among multiple genes in the same pathway can have an ultimate effect on the multiple genetic outcomes. Our work suggests that common genetic variants within Rho GTPase signal transduction pathway play a role in two important complex genetic traits, CAD and HDL-C. The genes in this pathway, which remain to be explored in this context, represent a rich set of candidate genes from which to dissect genetic contributions to coronary disease susceptibility along with the contributions from the covariate HDL-C. Hence, KALRN, CDGAP and PAK2 genes will most likely be important in the dissection of complex traits, CAD and HDL-C.

## 4.5 Tables

**Table 4.1** GENECARD Sample Size

| Sample | Initial Genome Screen Families | Follow-up Collection | All Families |
|---|---|---|---|
| Number of families | **420** | **681** | **1,101** |
| Individuals sampled | **1,129** | **1,825** | **2,954** |
| Affected individuals | **1,005** | **1,255** | **2,260** |
| Unaffected individuals | **122** | **481** | **603** |
| Unaffected > age 60 y (M); 65 y (F) | **84** | **192** | **276** |
| Affected sibling pairs | **539** | **427** | **966** |
| Discordant sibling pairs | **169** | **467** | **825** |
| Discordant sibling pairs > 60 y (M); 65 y (F) | **45** | **56** | **137** |

**Table 4.2** Clinical Characteristics of GENECARD and CATHGEN subjects

| Variable | GENECARD Probands | CATHGEN Young Affected | CATHGEN Old Affected | Unaffected Controls |
|---|---|---|---|---|
| **Number of samples** | (N=1101) | (N=656) | (N=264) | (N=405) |
| **Mean age at onset (SD)** [1,2] | 44 (6) | 46 (6) | 66 (8) | N/A |
| **Mean age at exam (SD)** [1,2,3,4] | 50 (7) | 52 (9) | 72 (8) | 69 (7) |
| **Self Reported Race: Caucasian**[1,2,3,4] | 86% | 71% | 87% | 71% |
| **Sex: Male**[1,2,3,4] | 69% | 79% | 72% | 43% |
| **Family History of CAD**[1,2,3,4] | 100% | 55% | 43% | 27% |
| **BMI, kg/m$^2$ (SD)** [2,3,4] | 32 (8) | 31 (6) | 28 (6) | 29 (7) |
| **Ever-smoked**[1,2,3,4] | 78% | 69% | 52% | 40% |
| **Diabetes**[1,3,4] | 25% | 33% | 31% | 21% |
| **Hypertension**[1,4] | 63% | 70% | 75% | 67% |
| **Myocardial Infarction**[1,2,3,4] | 63% | 52% | 38% | N/A |
| **Coronary Artery Bypass Graft** [1,2,3,4] | 49% | 40% | 52% | N/A |
| **Mean systolic BP, mmHg (SD)** [1,2,3] | 127 (19) | 141 (24) | 151 (26) | 150 (23) |
| **Mean diastolic BP, mmHg (SD)** [1] | 76 (12) | 78 (14) | 78 (13) | 77 (14) |
| **Total Cholesterol, mg/dL (SD)** [1,2,4] | 219 (56) | 193 (62) | 173 (41) | 192 (49) |
| **LDL, mg/dL (SD)** [1,2,4] | 142 (74) | 109 (43) | 98 (34) | 107 (36) |
| **HDL, mg/dL (SD)** [2,3,4] | 35 (37) | 40 (12) | 45 (14) | 52 (18) |
| **NHC, mg/dL (SD)** | | 37 (10) | 38 (12) | 47 (14) |
| **Triglycerides mg/dL (SD)** [1,2,3,4] | 240 (310) | 223 (261) | 163 (96) | 157 (126) |

[1] significant difference ($p<0.05$) between GENECARD probands and CATHGEN young affected cases.
[2] significant difference between CATHGEN young affected and CATHGEN old affected cases.
[3] significant difference between CATHGEN young affected cases and controls.
[4] significant difference between CATHGEN old affected cases and controls.
Differences between the two groups were determined using a Chi-square test (categorical variables) or a t-test (numerical variables).
SD, standard deviation; N/A, not applicable

**Table 4.3** Top 20 SNPs for Mean of HDL-C by Genotypes in the CATHGEN ALL group

| | | | | | Number of Individuals with HDL-C values and Unadjusted Mean (SD) of HDL-C for Each Genotype | | | | | | ANOVA Analysis p-value | CAD*** Association p-value | Interaction Analysis p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Locus** | **bp** | **MA** | **MAF** | N11 | 11 | N12 | 12 | N22 | 22 | **Adjusted*** | **Adjusted**** | **Adjusted*** |
| RS6784285 | PHLDB2 | 113116525 | C | 0.3501 | 369 | 40.69(13.09) | 401 | 39.05(12.21) | 120 | 38.69(13.32) | **0.0128** | 0.9052 | 0.0777 |
| RS6786655 | PHLDB2 | 113146618 | C | 0.336 | 449 | 40.70(13.09) | 446 | 39.22(12.46) | 124 | 39.60(13.50) | **0.013** | 0.5846 | 0.1066 |
| RS1316642 | FLJ39873 | 115506753 | A | 0.3754 | 381 | 40.08(12.64) | 490 | 39.27(12.89) | 131 | 42.21(13.53) | **0.024** | *0.0633* | 0.3761 |
| RS1106851 | LSAMP | 117943999 | A | 0.1877 | 682 | 40.13(12.73) | 314 | 40.03(13.41) | 39 | 35.77(11.25) | **0.0012** | *0.1036* | 0.0816 |
| RS6788787 | LSAMP | 118353538 | A | 0.1518 | 731 | 40.29(12.99) | 263 | 39.83(12.94) | 29 | 32.07(8.55) | **0.0159** | 0.4829 | 0.3883 |
| RS4687991 | LSAMP | 118947921 | G | 0.4177 | 381 | 39.79(12.94) | 443 | 39.06(12.52) | 200 | 42.28(13.30) | **0.0297** | *0.0982* | 0.3213 |
| RS1401951 | LOC389142 | 119708716 | G | 0.4183 | 302 | 38.51(11.88) | 417 | 40.80(13.01) | 160 | 38.74(13.03) | **0.0141** | 0.6526 | 0.9818 |
| RS7650193 | CDGAP | 120517446 | T | 0.292 | 590 | 39.83(12.39) | 438 | 40.49(13.59) | 113 | 39.88(11.52) | **0.0244** | *0.3648* | 0.2223 |
| RS2044621 | CDGAP | 120533422 | T | 0.3115 | 554 | 40.51(12.98) | 469 | 39.44(12.44) | 115 | 40.80(13.13) | **0.0287** | *0.6396* | 0.3869 |
| RS1132200 | TMEM39A | 120633526 | T | 0.1347 | 666 | 39.39(12.89) | 200 | 40.68(12.22) | 18 | 38.72(10.03) | **0.007** | *0.1577* | 0.4118 |
| RS2049502 | CD80 | 120737075 | G | 0.1646 | 563 | 39.17(13.04) | 206 | 41.65(12.21) | 25 | 41.04(12.27) | **0.0011** | *0.3883* | 0.8222 |
| RS626364 | CD80 | 120755573 | G | 0.477 | 226 | 41.00(12.80) | 367 | 40.01(12.93) | 196 | 37.97(12.42) | **0.0152** | 0.5556 | **0.0066** |
| RS10511415 | DIRC2 | 124011725 | G | 0.0939 | 636 | 40.18(13.02) | 128 | 38.98(12.31) | 5 | 27.80(9.39) | **0.0208** | *0.4175* | **0.0004** |
| RS7641248 | MYLK | 124909674 | G | 0.1239 | 885 | 40.57(12.82) | 227 | 38.37(12.34) | 28 | 38.54(14.14) | **0.0277** | *0.0676* | **4.49E-05** |
| RS9289235 | KALRN | 125636850 | C | 0.3401 | 381 | 40.41(12.71) | 400 | 39.83(13.28) | 108 | 36.47(10.32) | **0.0107** | **0.0408** | 0.3319 |
| RS3772759 | KALRN | 125653172 | T | 0.1172 | 680 | 40.25(13.04) | 192 | 37.78(11.47) | 11 | 39.91(14.73) | **0.0279** | 0.4671 | **0.0187** |
| RS885076 | NEK11 | 132406532 | G | 0.133 | 605 | 39.36(12.83) | 184 | 41.49(13.33) | 11 | 36.91(4.83) | **0.0231** | *0.4975* | 0.737 |
| RS6802186 | CPNE4 | 132916957 | C | 0.3424 | 338 | 38.66(12.98) | 348 | 40.19(12.54) | 98 | 42.27(13.56) | **0.0081** | *0.8489* | 0.4926 |
| RS6786236 | RAB6B | 135024399 | G | 0.3121 | 556 | 39.28(12.46) | 475 | 41.36(13.11) | 120 | 38.40(12.14) | **0.0126** | 0.0691 | 0.9583 |
| RS12634982 | PAK2 | 197994749 | C | 0.0766 | 964 | 40.46(12.68) | 156 | 37.76(12.39) | 7 | 50.00(21.06) | **0.025** | 0.176 | 0.7112 |

1. In table, 2 is the minor allele; N11 is the number of people with genotype 11 who have HDL-C value.
2. Italic font means OR < 1, otherwise OR >= 1.0. Red font color means p-value ≤ .05.
3. * race, sex, smoking history, BMI, age at exam adjusted
4. ** race and sex adjusted
5. *** versus ON individuals

**Table 4.4** Top 20 SNPs for Mean of HDL-C by Genotypes in the CATHGEN YA group

| | | | | | Number of Individuals with HDL-C values and Unadjusted Mean (SD) of HDL-C for Each Genotype | | | | | | ANOVA Analysis p-value | CAD*** Association p-value | Interaction Analysis p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Locus** | **bp** | **MA** | **MAF** | N11 | 11 | N12 | 12 | N22 | 22 | **Adjusted*** | **Adjusted**** | **Adjusted*** |
| RS1875111 | BOC | 114461454 | C | 0.3006 | 231 | 37.35(9.91) | 176 | 36.28(10.81) | 44 | 33.91(8.66) | **0.0178** | *0.672* | 0.061 |
| RS7633227 | LSAMP | 118313302 | G | 0.3918 | 189 | 36.01(10.50) | 242 | 37.77(10.47) | 79 | 33.80(8.80) | **0.0022** | 0.509 | 0.6253 |
| RS6788787 | LSAMP | 118353538 | A | 0.1508 | 367 | 37.20(10.55) | 127 | 35.98(9.62) | 17 | 28.94(7.85) | **0.0181** | 0.4738 | 0.3883 |
| RS11716267 | LSAMP | 118586537 | T | 0.2346 | 300 | 37.53(10.23) | 183 | 35.68(10.41) | 29 | 32.66(9.61) | **0.0317** | 0.4564 | 0.3503 |
| RS869851 | LSAMP | 118804008 | C | 0.3569 | 206 | 37.76(10.12) | 257 | 35.36(9.79) | 55 | 38.00(12.88) | **0.0073** | *0.8228* | **0.0281** |
| RS9831023 | CDGAP | 120594452 | C | 0.3447 | 251 | 36.45(11.05) | 232 | 35.72(8.84) | 73 | 40.70(10.65) | **0.0204** | 0.2968 | 0.6693 |
| RS1132200 | TMEM39A | 120633526 | T | 0.1277 | 386 | 36.54(10.15) | 104 | 37.91(10.36) | 10 | 32.30(5.54) | **0.0329** | *0.292* | 0.4118 |
| RS2049502 | CD80 | 120737075 | G | 0.1575 | 333 | 35.87(9.93) | 111 | 38.77(10.35) | 13 | 38.62(14.26) | **0.0125** | *0.4871* | 0.8222 |
| RS8935 | PDIA5 | 124362881 | C | 0.2695 | 267 | 35.65(9.94) | 190 | 38.23(10.44) | 42 | 36.71(9.62) | **0.0119** | 0.9606 | 0.1946 |
| RS4678030 | ADCY5 | 124659437 | G | 0.4085 | 173 | 34.92(9.60) | 243 | 37.16(9.86) | 83 | 39.10(11.60) | **0.0259** | 0.4593 | **0.0058** |
| RS11927656 | PTPLB | 124734483 | G | 0.1903 | 286 | 35.84(9.71) | 155 | 38.07(10.82) | 14 | 40.00(10.65) | **0.0339** | 0.8545 | **0.0033** |
| RS2605417 | MYLK | 125032085 | C | 0.1676 | 315 | 36.88(10.41) | 126 | 35.29(9.19) | 12 | 43.00(11.54) | **0.0249** | *0.6323* | 0.5022 |
| RS12496156 | KALRN | 125478341 | A | 0.3466 | 194 | 34.94(9.65) | 186 | 37.81(10.16) | 60 | 39.20(11.64) | **0.0044** | *0.4063* | 0.1636 |
| RS920232 | PODLX2 | 128861968 | T | 0.1679 | 317 | 37.23(10.43) | 131 | 36.11(9.59) | 13 | 29.08(9.79) | **0.026** | *0.7953* | 0.2375 |
| RS2811469 | PLXND1 | 130764156 | A | 0.2664 | 299 | 36.52(10.04) | 198 | 37.53(10.18) | 54 | 33.50(11.05) | **0.0019** | 0.3116 | 0.0944 |
| RS1872119 | PLXND1 | 130802834 | G | 0.0789 | 472 | 36.43(9.93) | 81 | 38.16(11.76) | 5 | 29.00(4.90) | **0.0162** | *0.3823* | 0.4044 |
| RS1486230 | MRPL3 | 132678451 | G | 0.2363 | 262 | 36.59(10.13) | 165 | 37.81(10.56) | 32 | 32.94(9.38) | **0.0271** | *0.7285* | 0.0732 |
| RS1870713 | CPNE4 | 133157225 | A | 0.4545 | 142 | 37.63(9.95) | 219 | 35.22(9.96) | 98 | 39.08(10.92) | **0.0046** | 0.6511 | 0.121 |
| RS309986 | ACPP | 133516338 | G | 0.3469 | 195 | 38.36(10.75) | 202 | 35.41(9.84) | 54 | 36.00(9.64) | **0.0069** | 0.4206 | 0.9737 |
| RS7646247 | PAK2 | 198007519 | T | 0.2062 | 351 | 36.54(9.74) | 174 | 36.32(10.10) | 25 | 42.16(15.33) | **0.004** | 0.2327 | 0.6214 |

1. In table, 2 is the minor allele; N11 is the number of people with genotype 11 who have HDL-C value.
2. Italic font means OR < 1, otherwise OR >= 1.0. Red font color means p-value ≤ .05.
3. * race, sex, smoking history, BMI, age at exam adjusted
4. ** race and sex adjusted
5. *** versus ON individuals

**Table 4.5** Top 20 SNPs for Mean of HDL-C by Genotypes in the CATHGEN OA group

| | | | | | **Number of Individuals with HDL-C values and Unadjusted Mean (SD) of HDL-C for Each Genotype** | | | | | | **ANOVA Analysis p-value** | **CAD*** Association p-value** | **Interaction Analysis p-value** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Locus** | **bp** | **MA** | **MAF** | N11 | 11 | N12 | 12 | N22 | 22 | **Adjusted*** | **Adjusted**** | **Adjusted*** |
| RS6786655 | PHLDB2 | 113146618 | C | 0.344 | 104 | 38.69(11.44) | 95 | 38.28(11.88) | 33 | 35.03(12.94) | **0.0219** | 0.2434 | 0.1066 |
| RS10511302 | PHLDB2 | 113170446 | G | 0.4307 | 66 | 39.59(11.55) | 94 | 39.48(13.18) | 40 | 34.40(10.82) | **0.0023** | 0.0939 | 0.1518 |
| RS1994731 | PHLDB2 | 113173458 | G | 0.4485 | 64 | 39.25(11.49) | 93 | 39.38(13.01) | 45 | 35.16(11.66) | **0.0055** | 0.0701 | 0.1844 |
| RS2272022 | CD200 | 113546540 | A | 0.3364 | 72 | 39.76(11.97) | 73 | 35.21(11.22) | 18 | 41.72(12.56) | **0.0266** | 0.3239 | 0.0544 |
| RS4687991 | LSAMP | 118947921 | G | 0.3882 | 95 | 38.84(11.47) | 100 | 35.77(12.06) | 41 | 41.37(11.88) | **0.0097** | *0.2707* | 0.3213 |
| RS553070 | LOC389142 | 119637627 | A | 0.2164 | 123 | 38.73(12.01) | 67 | 36.85(12.38) | 9 | 41.44(11.31) | **0.0141** | 0.76 | 0.6384 |
| RS935616 | CDGAP | 120595249 | T | 0.4286 | 82 | 41.01(11.95) | 130 | 38.24(11.59) | 45 | 34.49(12.12) | **0.02** | *0.9079* | 0.8691 |
| RS1132200 | TMEM39A | 120633526 | T | 0.1443 | 143 | 37.43(11.99) | 54 | 40.26(12.46) | 2 | 39.50(7.78) | **0.0441** | *0.1721* | 0.4118 |
| RS1501899 | CASR | 123390018 | A | 0.3604 | 101 | 40.50(13.09) | 103 | 36.15(10.68) | 34 | 36.82(11.22) | **0.0217** | 0.8139 | 0.2453 |
| RS11718105 | MYLK | 124946398 | T | 0.3138 | 114 | 36.98(10.17) | 96 | 41.06(13.69) | 27 | 32.70(9.97) | **0.012** | 0.6213 | 0.2385 |
| RS2682215 | MYLK | 125027266 | G | 0.1779 | 107 | 37.42(10.34) | 50 | 39.48(13.08) | 4 | 28.00(15.60) | **0.0044** | *0.1669* | 0.3884 |
| RS2605417 | MYLK | 125032085 | C | 0.1747 | 110 | 37.48(11.42) | 50 | 39.78(12.82) | 4 | 28.00(15.60) | **0.0023** | *0.4615* | 0.5022 |
| RS2682239 | MYLK | 125042419 | A | 0.1807 | 109 | 37.73(11.34) | 50 | 39.54(13.06) | 5 | 29.00(13.69) | **0.0085** | *0.1815* | 0.3618 |
| RS4461370 | MYLK | 125048862 | G | 0.1752 | 156 | 37.81(11.03) | 70 | 39.73(13.66) | 6 | 34.00(16.06) | **0.0134** | *0.1974* | 0.1515 |
| RS2682218 | MYLK | 125066569 | A | 0.1654 | 180 | 38.22(11.22) | 75 | 39.52(13.25) | 6 | 34.00(16.06) | **0.0197** | *0.4351* | 0.2107 |
| RS2682229 | MYLK | 125084440 | G | 0.1841 | 155 | 37.55(11.12) | 76 | 39.47(13.32) | 6 | 34.00(16.06) | **0.0149** | *0.47* | 0.2262 |
| RS332410 | KALRN | 125742463 | G | 0.1479 | 173 | 37.12(11.68) | 60 | 41.03(12.18) | 5 | 31.00(5.00) | **0.0266** | 0.791 | 0.0939 |
| RS2358766 | . | 128448992 | G | 0.3792 | 56 | 39.20(11.01) | 72 | 35.10(12.57) | 20 | 38.95(8.33) | **0.0179** | 0.8226 | 0.4393 |
| RS2659689 | GATA2 | 129685704 | C | 0.3975 | 84 | 39.85(12.42) | 117 | 36.49(11.82) | 36 | 38.67(10.39) | **0.0256** | 0.8606 | 0.5091 |
| RS2811469 | PLXND1 | 130764156 | A | 0.2954 | 131 | 40.61(11.36) | 100 | 36.58(11.96) | 26 | 36.46(13.67) | **0.0058** | 0.9343 | 0.0944 |
| RS4682664 | RYK | 135402607 | G | 0.1345 | 196 | 37.39(11.68) | 63 | 41.95(12.22) | 3 | 35.00(11.14) | **0.0154** | 0.2379 | 0.6285 |

1. In table, 2 is the minor allele; N11 is the number of people with genotype 11 who have HDL-C value.
2. Italic font means OR < 1, otherwise OR >= 1.0. Red font color means p-value ≤ .05.
3. * race, sex, smoking history, BMI, age at exam adjusted
4. ** race and sex adjusted
5. *** versus ON individuals

**Table 4.6** Top 20 SNPs for Mean of HDL-C by Genotypes in the CATHGEN ON group

| | | | | | Number of Individuals with HDL-C values and Unadjusted Mean (SD) of HDL-C for Each Genotype | | | | | | ANOVA Analysis p-value | Interaction Analysis p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Locus** | **bp** | **MA** | **MAF** | N11 | 11 | N12 | 12 | N22 | 22 | **Adjusted*** | **Adjusted*** |
| RS11708769 | CD200 | 113569632 | T | 0.1694 | 116 | 52.07(15.09) | 54 | 45.85(13.66) | 4 | 43.75(9.32) | **0.0121** | 0.1955 |
| RS2078008 | CD200R2 | 114021682 | T | 0.2345 | 162 | 46.96(14.26) | 98 | 48.51(15.11) | 17 | 53.65(15.89) | **0.0121** | 0.2156 |
| RS1106851 | LSAMP | 117943999 | A | 0.2149 | 177 | 48.58(14.62) | 88 | 47.27(14.83) | 15 | 39.80(12.66) | **0.0112** | 0.0816 |
| RS1915585 | LSAMP | 118391522 | T | 0.1519 | 196 | 48.53(14.14) | 73 | 46.64(15.24) | 4 | 38.50(28.24) | **0.0375** | **0.0013** |
| RS1698041 | LSAMP | 118682441 | C | 0.4744 | 81 | 45.52(13.47) | 127 | 50.33(15.41) | 66 | 45.47(13.58) | **0.0118** | 0.4002 |
| RS2937673 | LSAMP | 118715077 | C | 0.4753 | 80 | 45.98(12.95) | 122 | 50.07(15.57) | 66 | 45.30(13.33) | **0.03** | 0.428 |
| RS6438468 | . | 119873265 | C | 0.396 | 70 | 49.77(13.91) | 74 | 49.85(16.02) | 31 | 49.65(14.77) | **0.0165** | 0.6692 |
| RS705233 | LOC389142 | 119952613 | C | 0.3225 | 87 | 46.86(14.30) | 79 | 52.75(15.53) | 19 | 44.58(11.06) | **0.0012** | **0.0446** |
| RS12639070 | CDGAP | 120614010 | C | 0.1269 | 242 | 46.18(14.51) | 75 | 48.87(13.18) | 2 | 65.50(0.71) | **0.0113** | **0.0407** |
| RS626364 | CD80 | 120755573 | G | 0.4679 | 54 | 49.76(14.25) | 73 | 51.73(13.34) | 45 | 46.07(16.92) | **0.0298** | **0.0066** |
| RS2229308 | GTF2E1 | 121983136 | A | 0.3132 | 120 | 47.14(14.68) | 129 | 49.26(14.93) | 24 | 40.33(7.81) | **0.0142** | 0.2882 |
| RS865066 | . | 122034873 | A | 0.2682 | 150 | 46.81(14.05) | 108 | 47.21(14.78) | 21 | 56.48(15.48) | **0.013** | 0.07 |
| RS7641248 | MYLK | 124909674 | G | 0.1352 | 247 | 48.68(14.00) | 67 | 41.91(14.66) | 10 | 50.00(16.48) | **0.0055** | **4.49E-05** |
| RS4118366 | MYLK | 125066921 | A | 0.2113 | 171 | 46.81(13.51) | 94 | 49.95(16.08) | 11 | 42.45(17.91) | **0.0247** | 0.1296 |
| RS3772759 | KALRN | 125653172 | T | 0.1065 | 145 | 50.55(14.58) | 37 | 44.38(15.17) | 3 | 49.00(20.95) | **0.0167** | **0.0187** |
| RS2713655 | KALRN | 125756597 | A | 0.319 | 127 | 49.47(13.79) | 108 | 48.05(15.17) | 38 | 40.71(12.39) | **0.0243** | **0.0248** |
| RS2979307 | OSBPL11 | 126792264 | G | 0.1674 | 115 | 49.81(14.07) | 51 | 50.27(14.17) | 3 | 51.00(43.21) | **5.128E-05** | 0.8696 |
| RS2860228 | GATA2 | 129692365 | T | 0.4677 | 83 | 46.28(14.65) | 129 | 47.19(15.20) | 66 | 50.89(13.61) | **0.0313** | **0.0481** |
| RS11711812 | KIAA1257 | 130191252 | G | 0.3618 | 73 | 53.99(15.11) | 71 | 47.35(13.55) | 25 | 43.28(15.47) | **0.0337** | **0.0002** |
| RS12634982 | PAK2 | 197994749 | C | 0.0634 | 286 | 47.52(14.24) | 33 | 41.67(13.64) | 3 | 65.33(24.19) | **0.0092** | 0.7112 |

1. In table, 2 is the minor allele; N11 is the number of people with genotype 11 who have HDL-C value.
2. Red font color means p-value ≤ .05.
3. * race, sex, smoking history, BMI, age at exam adjusted

**Table 4.7** Top 6 SNPs with HDL-C association of the QTDT test in GENECARD data

| | | | | | Number of Individuals with HDL-C values and Unadjusted Mean (SD) of HDL-C for Each Genotype | | | | | | QTDT p-value | CAD Association p-value | | Interaction GEE p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Locus** | **bp** | **MA** | **MAF** | N11 | 11 | N12 | 12 | N22 | 22 | **Adjusted*** | **APL** | **PDT** | **Adjusted*** |
| RS1513156 | LSAMP | 118549311 | G | 0.1553 | 935 | 39.55(11.93) | 315 | 39.13(13.04) | 39 | 40.18(10.81) | **0.0393** | 0.084 | 0.203 | 0.9125 |
| RS869851 | LSAMP | 118804008 | C | 0.3579 | 512 | 39.20(12.02) | 654 | 39.48(12.99) | 152 | 39.80(9.13) | **0.0494** | 0.28 | 0.536 | 0.7861 |
| RS9831023 | CDGAP | 120594452 | C | 0.2778 | 718 | 39.03(11.38) | 504 | 39.86(13.62) | 97 | 39.77(10.04) | **0.0301** | 0.394 | 0.447 | 0.2866 |
| RS1132200 | TMEM39A | 120633526 | T | 0.1526 | 953 | 39.52(12.11) | 308 | 39.25(12.76) | 35 | 37.86(9.58) | **0.0456** | 0.356 | 0.269 | 0.2431 |
| RS2682248 | . | 125099763 | A | 0.0759 | 1117 | 39.41(12.24) | 147 | 39.37(13.02) | 23 | 40.87(8.36) | **0.006** | 0.482 | 0.107 | 0.3825 |
| RS2860228 | GATA2 | 129692365 | T | 0.4463 | 389 | 38.72(10.60) | 676 | 39.23(11.79) | 253 | 40.97(15.21) | **0.0428** | 0.703 | 0.903 | 0.1483 |

1. In table, 2 is the minor allele; N11 is the number of people with genotype 11 who have HDL-C value.
2. Red font color means p-value ≤ .05.
3. * race, sex, smoking history, BMI, age at exam adjusted

**Table 4.8** Model assignment of Markers in CATHGEN/GENECARD dataset

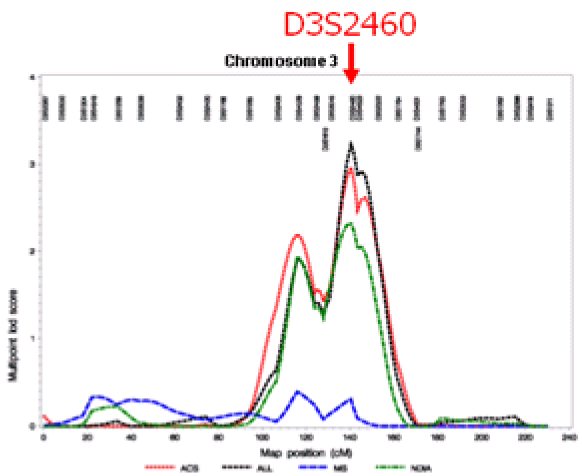
| Models | Tests | | | | Results |
|---|---|---|---|---|---|
| | QTL analysis | | CAD association | Interaction analysis | Number of Markers |
| | Stratified affected /All | Stratified control | | | |
| QTL model | Yes | Yes | No | No | 86*/3 |
| Modifier I model | Yes | No | No | Yes | 10/0 |
| Modifier II model | No | Yes | No | Yes | 8/1 |
| Interaction model | Yes | Yes | Yes/No | Yes | 4/0 |
| Pleiotropy model | Yes | Yes | Yes | Yes or No | (1)/0 |

Note: * The 86 markers include the one belonging to pleiotropy model;

**Table 4.9** SNPs in final model by groups with backward stepwise analysis

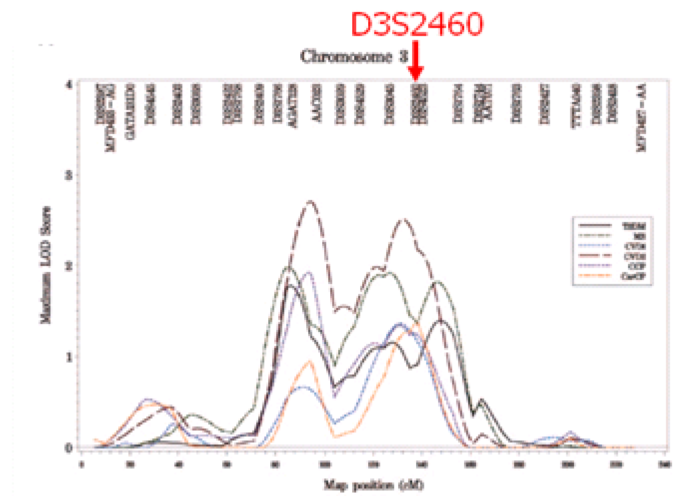
| SNP | Locus | SNP | Locus | SNP | Locus | SNP | Locus | SNP | Locus |
|---|---|---|---|---|---|---|---|---|---|
| **Modifier I group** | | **QTL-ALL group** | | **QTL-YA group** | | **QTL-OA group** | | **QTL-ON group** | |
| RS17681395 | ZBTB20 | RS6784285 | PHLDB2 | RS935616 | CDGAP | RS6786655 | PHLDB2 | RS11708769 | CD200 |
| RS1850719 | LSAMP | RS4687991 | LSAMP | RS1132200 | TMEM39A | RS2272022 | CD200 | RS11922594 | CDGAP |
| **Modifier II group** | | RS7650193 | CDGAP | RS2811469 | PLXND1 | RS4687991 | LSAMP | RS865066 | . |
| RS1915585 | LSAMP | RS2044621 | CDGAP | | | RS553070 | LOC389142 | RS2712418 | RPN1 |
| RS705233 | LOC389142 | RS10049044 | CDGAP | | | RS1501899 | CASR | | |
| RS2873688 | CDGAP | RS1132200 | TMEM39A | | | RS2682215 | MYLK | | |
| RS12639070 | CDGAP | RS9844193 | SEC22L2 | | | RS16834826 | MYLK | | |
| RS9968051 | GOLGB1 | RS4678047 | MYLK | | | RS4678100 | KALRN | | |
| RS2860228 | GATA2 | RS885076 | NEK11 | | | RS332410 | KALRN | | |
| RS11711812 | KIAA1257 | RS3772587 | AGTR1 | | | RS2358766 | . | | |
| RS712984 | ATP2C1 | | | | | RS2659689 | GATA2 | | |
| | | | | | | RS1126828 | RPN1 | | |
| **Interaction group** | | | | | | RS2811469 | PLXND1 | | |
| RS626364 | CD80 | | | | | RS4682664 | RYK | | |
| RS3772759 | KALRN | | | | | RS275645 | AGTR1 | | |
| | | | | | | RS2084382 | PAK2 | | |

## 4.6 Figures




**A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD study**

*Hauser ER. et al., AJHG, 2004 Sep;75(3):436-47*




**A genomewide scan for type 2 diabetes, metabolic syndrome, and measures of cardiovascular disease: the Diabetes Heart study**

*Bowden DW. et al., Diabetes, 2006 July; 55 (7): 1985-94*

**Figure 4.1** A Major Susceptibility Region has been Mapped to Chromosome 3q13-21 in Two Independent Studies

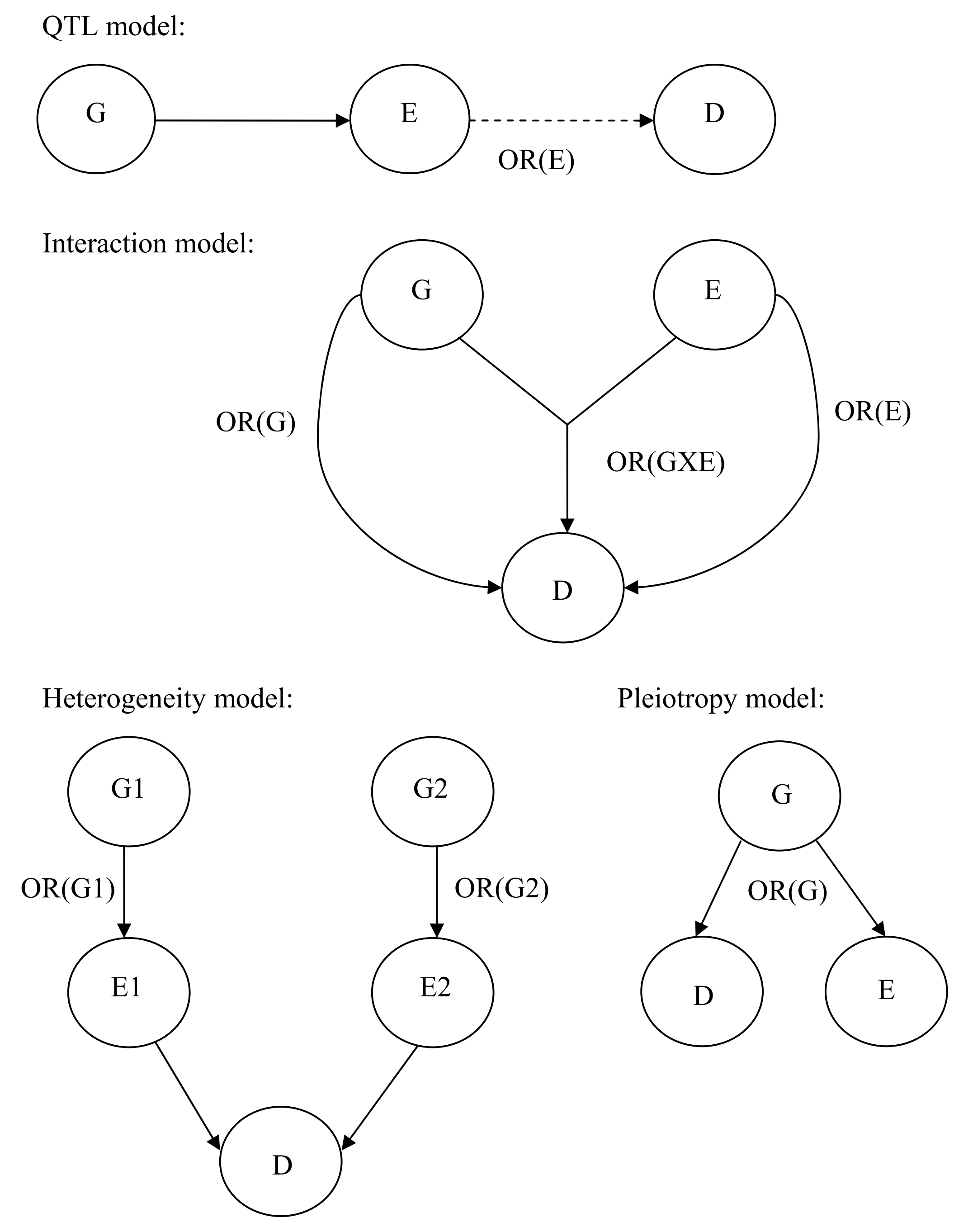
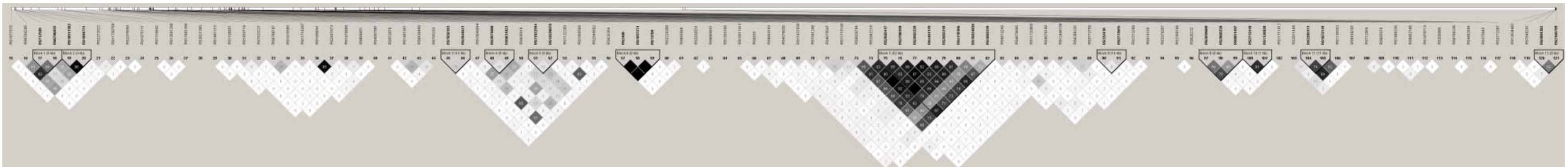
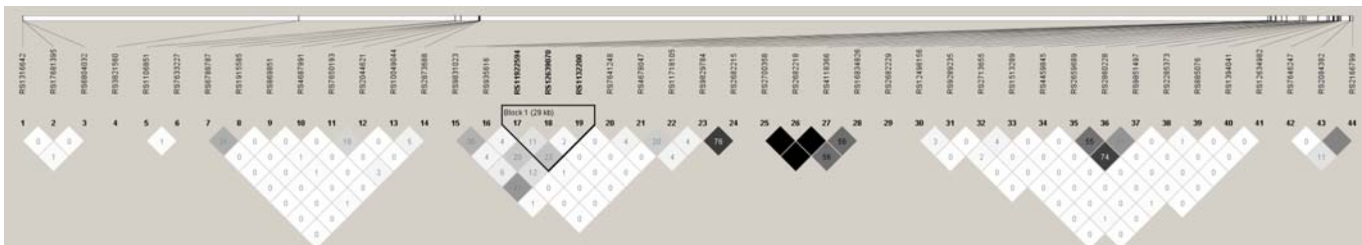



**Figure 4.2** Graphical illustration of covariate models (Schmidt *et al.* 2007).
G, G1, G2 = chromosome 3 variants (QTL or susceptibility locus).
E, E1, E2 = HDL or other CAD risk factors.
D = CAD.



(a) Pair-wise LD between 108 SNPs genotyped in CATHGEN dataset

LD was estimated in unaffected individuals from CATHGEN (n = 264). A similar pattern of LD was observed using the affected individuals.



(b) Pair-wise LD between 44 SNPs genotyped in GENECARD dataset

LD was estimated in unaffected individuals from each non-redundant GENECARD discordant sibling pair (n = 307). A similar pattern of LD was observed using the affected individuals.

**Figure 4.3** Pair-wise LD between SNPs genotyped in the CATHGEN and GENECARD dataset

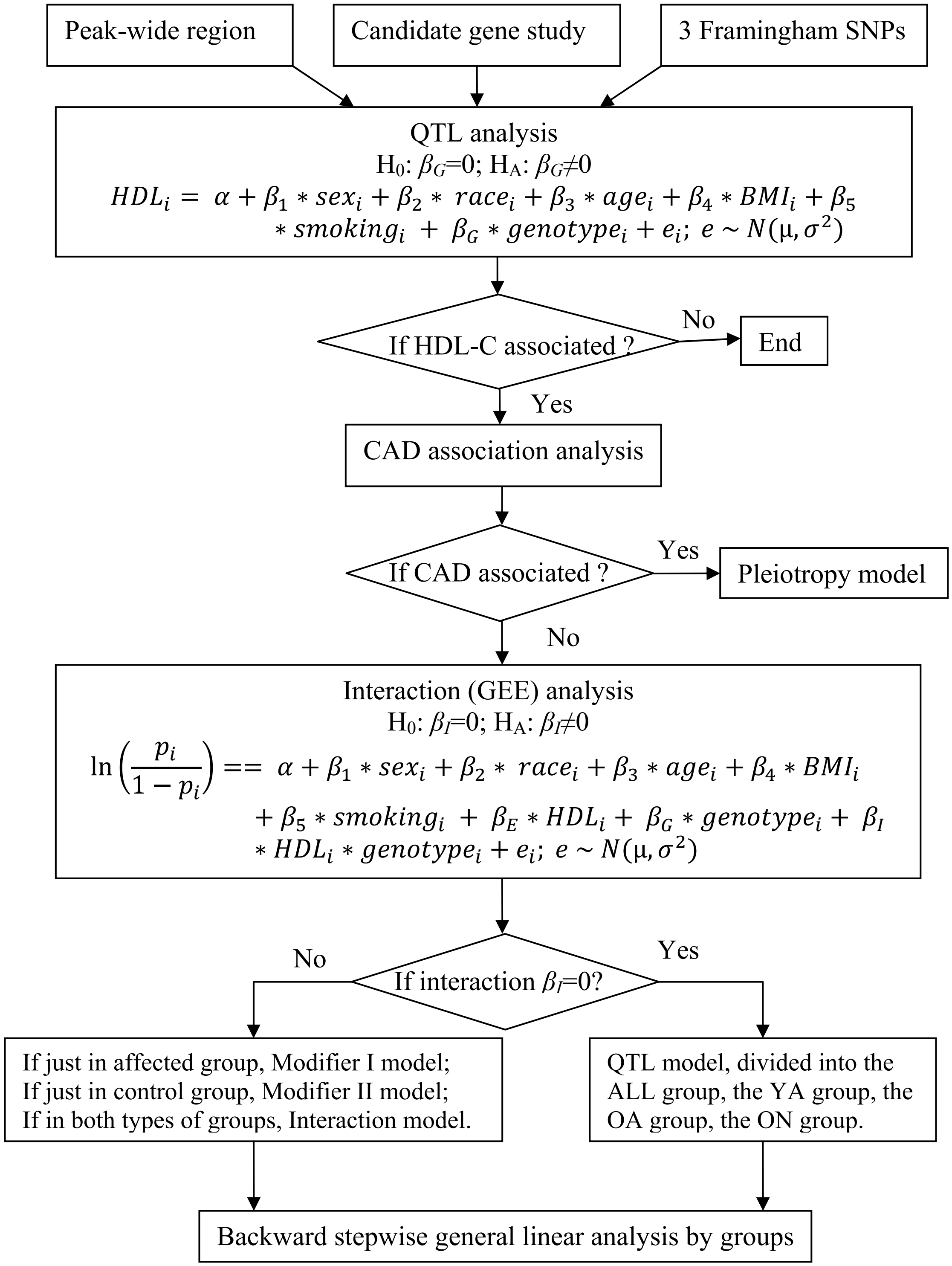



**Figure 4.4** Statistical Analysis Flowchart

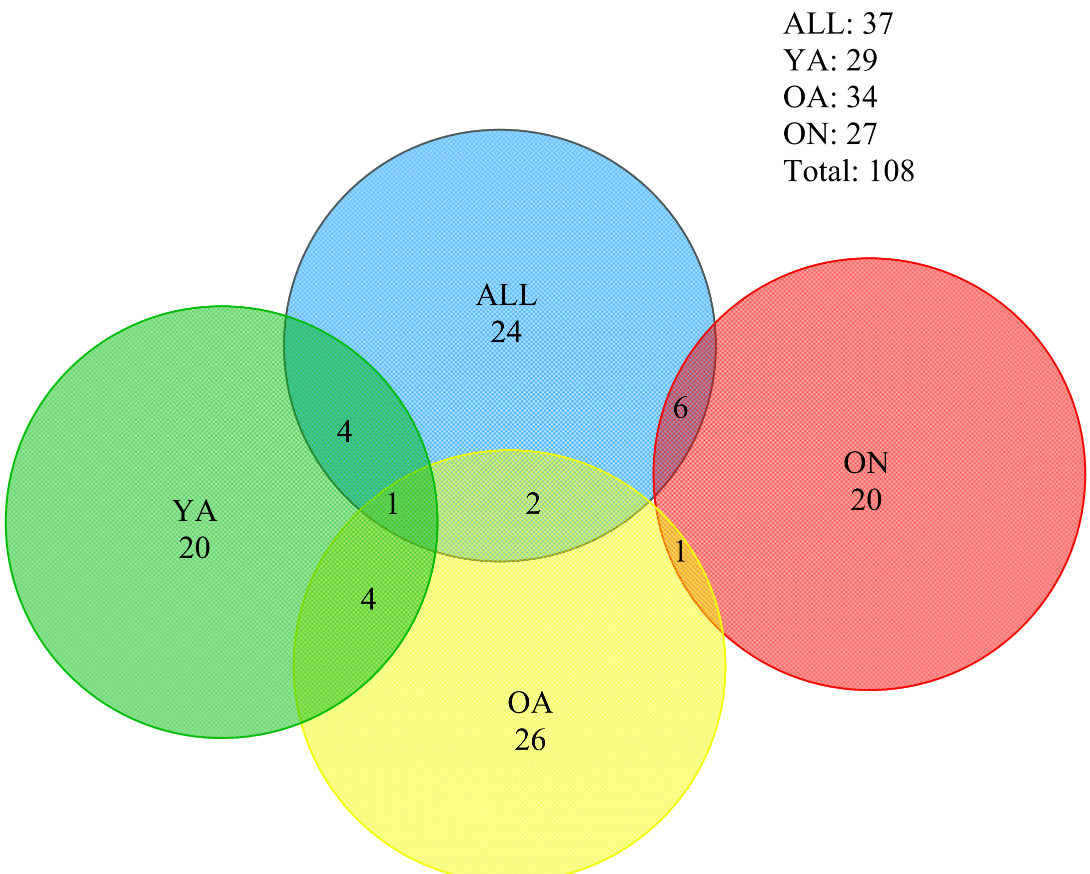



**Figure 4.5** Venn Diagram of HDL-C associated SNPs in different groups

# Chapter 5

## Conclusion

## 5.1 Summary

In summary, we evaluated different association methods under various settings, and applied statistical analyses on two independent datasets for HDL-C QTL gene mapping. Comparison of GIST and LAMP was done on GAW15 on the simulated rheumatoid arthritis (RA) data. Single marker association methods LAMP, APL and FBAT are compared under a multiple risk variants model within a local linkage peak region, and multiple marker association tests APL-Haplotype, FBAT-LC and conditional APL-OSA are compared under the same setting. Within the 3q13-21 CAD linkage region, stepwise statistical analyses were carried out in search for HDL-C associated variants, and attempts to dissect the complex relationship among HDL-C levels, CAD and genetic risk variants in this region.

GIST and LAMP were methods to test whether the linkage result can be explained in part or in full by the candidate SNP if it shows evidence of association. We compared the two approaches and evaluated their ability to identify genetic effects in the GAW 15 simulated RA data. We used 100 replicates of the simulated family data, and the simulation results showed that GIST is simple and works slightly better than LAMP-LE test when there is little linkage evidence. The LAMP linkage test has limited power when there is not much linkage evidence. The LAMP association test is the best not only when the linkage evidence is extremely high, but also when there is some LD between the candidate SNP and the trait locus. The type I error rate we obtained from a null chromosome with no trait locus is at the nominal level. We concluded that LAMP is more flexible and reliable to use in practice.

We modeled a specific situation of two independent disease loci with relative risk (RR=3.0) in a local region, and by varying LD parameters among M1 and M2 we modeled 5 different cases ranging from complete LD to moderate LD. We compared several association methods under these different multiple risk variants models. The simulation results showed that the power of single marker association tests is closely correlated with the amount of LD between marker and disease loci, and these tests maintain good power to detect multiple risk variants in a small region with moderate degree of LD for fully genotyped families. Global tests, like FBAT-LC, are sensitive to the presence of at least one susceptibility variant, but are not helpful for selecting the most promising SNPs for further study. We reported that if multiple haplotypes are associated with different disease loci, the haplotype tests results can be misleading and do not address the research question either, while APL-OSA conditional test has the greatest power to properly dissect the clustered associated markers for all cases with an acceptable type I error rate ranging from 0.033 to 0.056. We concluded that these findings on APL-OSA conditional test from our simulation study suggest that this type of conditional test may be applied to identify multiple risk variants within a gene or multiple genes under a known linkage peak, and our work makes contributions to understanding the performance of association tests under complex disease models.

Multiple CAD candidate genes have been reported from chromosome 3q13-21 (Connelly *et al.* 2006; Wang *et al.* 2007), and there are also linkage evidence that this region may harbors a gene or genes determining high-density lipoprotein (HDL) cholesterol levels (Imperatore *et al.* 2000; Shah *et al.* 2006). These results suggest that genetic variants associated with CAD risk in the chromosome 3q13 region, may be associated with, and mediated through, HDLC phenotypes. We performed the ANOVA analysis on the transformed HDL-C levels, adjusting for race, sex, smoking, BMI and age at exam for case-control data, and we investigated the relationship between genetic markers, CAD and HDL-C by direct association analysis with CAD and indirect association with CAD by interaction with HDL-C. We used a stratified case-control data set and a phenotypically similar family dataset to validate the HDL-C associations we observed in CATHGEN YA group. We identified SNPs in a genomic region selected on the basis of multiple lines of evidence (Hauser *et al.* 2003b; Hauser *et al.* 2004a; Bowden *et al.* 2006), including the strongest association at rs2979307 in the OSBPL11 gene which survives a Bonferroni correction. We identified associated SNPs by different models. We observed different HDL-C trends with HDL-C associated SNPs. Some SNPs have the highest HDL-C level for heterozygous genotype, and some have increasing or decreasing HDL-C values when the number of minor allele increases. This phenomenon suggests that the genetic basis for a complex trait such as HDL-C is highly heterogeneous. Even with the evident heterogeneity in our CAD population, we detected several association signals with SNPs in KALRN, MYLK, CDGAP and PAK2 genes in two separate CAD datasets for HDL-C, where all these genes belong to a Rho pathway. Thus we suggested that common genetic variants within the Rho pathway play an important role in two important complex genetic traits, CAD and HDL-C.

## 5.2 Future directions

Complex traits are often determined by multiple genetic and environmental factors with small-to-moderate effects. The way we try to model the complexity within a multiple modal linkage distribution of a complex trait to compare different association methods is that we simulated 2 independent disease loci within a local region, but the space for disease modeling is quite extensive. The number of genetic risk variants could be more than two, and they could interact with each other additively or multiplicatively. They could also interact with a QTL. All these risk variants could have different degrees of effect on the disease status. Thus, a polygenic model could be constructed to model after the real world of the complex disease. The problem is that all these genetic models could be lack of biological meaning. Genetic variations are not only nucleotide variations, but also include copy number variation and genetic variations at transcription level, or post-translational regulation, or protein level, et al. The molecular mechanisms behind the interaction between these different levels of genetic factors are not all clear to us (McCarthy *et al.* 2008), so all those modeling possibilities are difficult to construct with biological meaning and to implement.

Genome-wide association (GWA) studies are studies in which a dense set of SNPs across the genome is genotyped to survey the most common genetic variation for a role in disease or to identify the heritable quantitative traits that are risk factors for a disease (Hirschhorn & Daly 2005). With the availability of abundant genetic variations information from the International HapMap resource and dbSNP public database, the dense genotyping chips containing sets of hundreds of thousands of single nucleotide polymorphisms (SNPs) that provide good coverage of much of the human genome, and the large and well-characterized clinical samples already assembled for many common diseases, GWA studies for thousands of cases and controls are technically and financially feasible (The Wellcome Trust Case Control Consortium 2007). GWA studies can be performed without information from linkage analyses, and no assumptions are made about the genomic location of the causal variants (Hirschhorn & Daly 2005). Much research efforts for disease-gene mapping have been shifted from disease modeling to the unbiased yet fairly comprehensive option. Due to the multiple comparison corrections required for the huge number of hypotheses tested, the definition of a significant result in a GWA study is not straightforward. A threshold of $\sim 5 \times 10^{-7}$ has been chosen for several GWA studies to use in practice after considering confirmation of successfully validated variants at this nominal level (The Wellcome Trust Case Control Consortium 2007; Willer *et al.* 2008; Wallace *et al.* 2008). The first GWA study appeared in year 2005. Klein et al. (Klein *et al.* 2005) reported a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with Age-related macular degeneration (AMD), which is a major cause of blindness in the elderly, and Y402H, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal p-value $<10^{-7}$). The same variant was validated by another group published in the same issue (Haines *et al.* 2005). GWA studies have become the major force for disease-gene mapping, especially in the past 3 years, and made substantial advances in understanding the genetic basis of many common phenotypes of biomedical importance (McCarthy *et al.* 2008). 145 GWA studies are listed in the National Cancer Institute (NCI)-

National Human Genome Research Institute (NHGRI)'s catalog of published genome-wide association studies, over 50 disease-susceptibility loci have been revealed by these studies and provided valuable insights to the allelic architecture of complex traits (McCarthy *et al.* 2008). Despite all the achievements GWA studies have been made during the past years, GWA studies are facing many challenges in subject ascertainment and design, marker selection, analysis and interpretation, and validation and replication (McCarthy *et al.* 2008). GWA study is not an answer to every problem we meet.

The results from out study regarding genes and HDL-C level to CAD risk revealed that both CAD and HDL-C are truly complex traits. Even though they both have strong genetic components, the environmental effect is overwhelming. Inspired by the spirit of the GWA study, we can collect more samples, including collect/combine samples from other races, and add more SNPs to cover overall 3q13-21 region to overcome the selection bias of markers selected in Rho-GTPase pathway. But, facing the perplexing results from our study that only 1 SNP has a p-value at $10^{-5}$ level, GWA study may not be a solution either. The way for future successful disease-gene mapping I believe is to utilize all kinds of information available to us for disease-gene mapping.

Microarray expression data contains simultaneous information of many thousands of genes. Gene set enrichment analysis (GSEA) (Mootha *et al.* 2003) is a powerful analytical method to interpret gene expression data . It is a computational method that determines whether an a priori defined set of genes (for example, genes in one pathway) shows statistically significant and concordant differences between two phenotypes (Subramanian *et al.* 2005). The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation (Subramanian *et al.* 2005). In GSEA, expression information on all the genes under study is retained. GSEA is one kind of pathway analysis method, and there are other pathway analysis methods (Curtis *et al.* 2005). The idea of GSEA that the statistic for each risk variant in one set can be computed to measure the significance of two phenotypes can be extended to apply either linkage scores or association results from existing studies to utilize information from multiple genes simultaneously.

# References


Abecasis, G. R., L. R. Cardon, and W. O. C. Cookson, 2000a A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* **66**: 279-292.

Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon, 2002 Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**: 97-101.

Abecasis, G. R., and W. O. C. Cookson, 2000 GOLD - Graphical Overview of Linkage Disequilibrium. *BioInformatics* **16**: 182-183.

Abecasis, G. R., W. O. C. Cookson, and L. R. Cardon, 2000b Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics* **8**: 545-551.

Agresti A., 2002 *Categorical data analysis*. Wiley, New York.

Allingham, R. R., J. L. Wiggs, E. R. Hauser, K. R. Larocque-Abramson, C. Santiago-Turla *et al.* 2005 Early Adult-Onset POAG Linked to 15q11-13 Using Ordered Subset Analysis. *Invest Ophthalmol.Vis.Sci.* **46**: 2002-2005.

Antonarakis, S. E., and V. A. McKusick, 2000 OMIM passes the 1,000-disease-gene mark. *Nature Genetics* **25**: 11.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, 2005 Haploview: analysis and visualization of LD and haplotype maps. *BioInformatics* **21**: 263-265.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**: 289-300.

Biernacka, J. M., and H. J. Cordell, 2007 Exploring causality via identification of SNPs or Haplotypes responsible for a linkage signal. *Genetic Epidemiology* **31**: 727-740.

Botstein, D., and N. Risch, 2003 Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* **33 Suppl**: 228-237.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314-331.

Bowden, D. W., M. Rudock, J. Ziegler, A. B. Lehtinen, J. Xu *et al.* 2006 Coincident linkage of type 2 diabetes, metabolic syndrome, and measures of cardiovascular disease in a genome scan of the diabetes heart study. *Diabetes* **55**: 1985-1994.

Broeckel, U., C. Hengstenberg, B. Mayer, S. Holmer, L. J. Martin *et al.* 2002 A comprehensive linkage analysis for myocardial infarction and its related risk factors. *Nat.Genet* **30**: 210-214.

Cardon, L. R., and J. I. Bell, 2001 Association study designs for complex diseases. *Nat.Rev.Genet.* **2**: 91-99.

Chung, R. H., E. R. Hauser, and E. R. Martin, 2006 The APL test: Extension to general nuclear families and haplotypes and examination of its robustness. *Human Heredity* **61**: 189-199.

Chung, R. H., E. R. Hauser, and E. R. Martin, 2007 Interpretation of simultaneous linkage and family-based association tests in genome screens. *Genetic Epidemiology* **31**: 134-142.

Chung, R. H., S. Schmidt, E. R. Martin, and E. R. Hauser, 2008 Ordered-subset analysis (OSA) for family-based association mapping of complex traits. *Genet.Epidemiol.*

Clayton, D., 1999 A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics* **65**: 1170-1177.

Cohen, J. C., A. Pertsemlidis, S. Fahmi, S. Esmail, G. L. Vega *et al.* 2006 Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc.Natl.Acad.Sci.U.S.A* **103**: 1810-1815.

Connelly, J. J., T. Wang, J. E. Cox, C. Haynes, L. Y. Wang *et al.* 2006 GATA2 is associated with familial early-onset coronary artery disease. *Plos Genetics* **2**: 1265-1273.

Cox, N. J., M. Frigge, D. L. Nicolae, P. Concannon, C. L. Hanis *et al.* 1999 Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics* **21**: 213-215.

Curtis, D., 1996 Genetic dissection of complex traits. *Nature Genetics* **12**: 356-357.

Curtis, R. K., M. Oresic, and A. Vidal-Puig, 2005 Pathways to the analysis of microarray data. *Trends in Biotechnology* **23**: 429-435.

Dausset, J., H. Cann, D. Cohen, M. Lathrop, J. M. Lalouel *et al.* 1990 Program Description - Center-Detude-Du-Polymorphisme-Humain (Ceph) - Collaborative Genetic-Mapping of the Human Genome. *Genomics* **6**: 575-577.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. *American Journal of Human Genetics* **65**: A83.

Devlin, B., K. Roeder, and L. Wasserman, 2001 Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology* **60**: 155-166.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.* 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152-154.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg *et al.* 2006 A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**: 1461-1463.

Efron B., and R. J. Tibshirani, 1993 *An introduction to the Bootstrap*. Chapman & Hall/CRC.

Ewens, W. J., and R. S. Spielman, 1995 The Transmission Disequilibrium Test - History, Subdivision and Admixture. *American Journal of Human Genetics* **57**: 455-464.

Ewens, W. J., and R. S. Spielman, 2005 What is the significance of a significant TDT? *Human Heredity* **60**: 206-210.

Farrall, M., F. R. Green, J. F. Peden, P. G. Olsson, R. Clarke *et al.* 2006 Genome-Wide Mapping of Susceptibility to Coronary Artery Disease Identifies a Novel Replicated Locus on Chromosome 17. *PLoS.Genet.* **2**: e72.

Felker, G. M., L. K. Shaw, and C. M. O'Connor, 2002 A standardized definition of ischemic cardiomyopathy for use in clinical research. *J.Am.Coll.Cardiol.* **39**: 210-218.

Fisher, R. A., 1935a The detection of linkage with "dominant" abnormalities. *Annals of Eugenics* **4**: 187-201.

Fisher, R. A., 1935b The detection of linkage with recessive abnormalities. *Annals of Eugenics* **6**: 339-351.

Francke, S., M. Manraj, C. Lacquemant, C. Lecoeur, F. Lepretre *et al.* 2001 A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. *Human Molecular Genetics* **10**: 2751-2765.

Gillanders, E. M., J. V. Pearson, A. J. Sorant, J. M. Trent, J. R. O'Connell *et al.* 2006 The value of molecular haplotypes in a family-based linkage study. *American Journal of Human Genetics* **79**: 458-468.

Haines, J. L., M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson *et al.* 2005 Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**: 419-421.

Hall, J. M., M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson *et al.* 1990 Linkage of Early-Onset Familial Breast-Cancer to Chromosome-17Q21. *Science* **250**: 1684-1689.

Hancock, D. B., E. R. Martin, Y. J. Li, and W. K. Scott, 2007 Methods for interaction analyses using family-based case-control data: Conditional logistic regression versus generalized estimating equations. *Genetic Epidemiology* **31**: 883-893.

Hao, K., X. Xu, N. Laird, X. B. Wang, and X. P. Xu, 2004 Power estimation of multiple SNP association test of case-control study and application. *Genetic Epidemiology* **26**: 22-30.

Harrap, S. B., K. S. Zammit, Z. Y. H. Wong, F. M. Williams, M. Bahlo *et al.* 2002 Genome-wide linkage analysis of the acute coronary syndrome suggests a locus on chromosome 2. *Arteriosclerosis Thrombosis and Vascular Biology* **22**: 874-878.

Hauser, E. R., D. C. Crossman, C. B. Granger, J. L. Haines, C. J. Jones *et al.* 2004a A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. *American Journal of Human Genetics* **75**: 436-447.

Hauser, E. R., V. Mooser, D. C. Crossman, J. L. Haines, C. H. Jones *et al.* 2003a Design of the genetics of early onset cardiovascular disease (GENECARD) study. *Am.Heart J.* **145**: 602-613.

Hauser, E. R., R. M. Watanabe, W. L. Duren, M. P. Bass, C. D. Langefeld *et al.* 2004b Ordered subset analysis in genetic linkage mapping of complex traits. *Genetic Epidemiology* **27**: 53-63.

Hauser, M. A., Y. J. Li, S. Takeuchi, R. Walters, M. Noureddine *et al.* 2003b Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Human Molecular Genetics* **12**: 671-677.

Helgadottir, A., A. Manolescu, G. Thorleifsson, S. Gretarsdottir, H. Jonsdottir *et al.* 2004 The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nature Genetics* **36**: 233-239.

Helgadottir, A., G. Thorleifsson, K. P. Magnusson, S. Gretarsdottir, V. Steinthorsdottir *et al.* 2008 The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genetics* **40**: 217-224.

Helgadottir, A., G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal *et al.* 2007 A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**: 1491-1493.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**: 95-108.

Horvath, S., and N. M. Laird, 1998 A discordant-sibship test for disequilibrium and linkage: No need for parental data. *American Journal of Human Genetics* **63**: 1886-1897.

Horvath, S., X. Xu, S. L. Lake, E. K. Silverman, S. T. Weiss *et al.* 2004 Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics. *Genetic Epidemiology* **26**: 61-69.

Imperatore, G., W. C. Knowler, D. J. Pettitt, S. Kobes, J. H. Fuller *et al.* 2000 A locus influencing total serum cholesterol on chromosome 19p - Results from an autosomal genomic scan of serum lipid concentrations in Pima Indians. *Arteriosclerosis Thrombosis and Vascular Biology* **20**: 2651-2656.

Jeyarajah, E. J., W. C. Cromwell, and J. D. Otvos, 2006 Lipoprotein particle analysis by nuclear magnetic resonance spectroscopy. *Clinics in Laboratory Medicine* **26**: 847-870.

Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N. P. Burtt *et al.* 2008 Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics* **40**: 189-197.

Kim, W. Y., and N. E. Sharpless, 2006 The regulation of INK4/ARF in cancer and aging. *Cell* **127**: 265-275.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler *et al.* 2005 Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385-389.

Knapp, M., 1999 The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test. *American Journal of Human Genetics* **64**: 861-870.

Kong, A., and N. J. Cox, 1997 Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics* **61**: 1179-1188.

Kruglyak, L., 1997 The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics* **17**: 21-24.

Kruglyak, L., M. J. Daly, M. P. ReeveDaly, and E. S. Lander, 1996 Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics* **58**: 1347-1363.

Kruglyak, L., and D. A. Nickerson, 2001 Variation is the spice of life. *Nature Genetics* **27**: 234-236.

Laird, N. M., and C. Lange, 2006 Family-based designs in the age of large-scale gene-association studies. *Nat.Rev.Genet.* **7**: 385-394.

Lander, E., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**: 241-247.

Lander, E. S., and N. J. Schork, 1994 Genetic dissection of complex traits. *Science* **265**: 2037-2048.

Li, C., L. J. Scott, and M. Boehnke, 2004 Assessing whether an allele can account in part for a linkage signal: The Genotype-IBD Sharing Test (GIST). *American Journal of Human Genetics* **74**: 418-431.

Li, M. Y., M. Boehnke, and G. R. Abecasis, 2005 Joint modeling of linkage and association: Identifying SNPs responsible for a linkage signal. *American Journal of Human Genetics* **76**: 934-949.

Liang, K. Y., and S. L. Zeger, 1986 Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* **73**: 13-22.

Lloyd-Jones, D. M., B. H. Nam, R. B. D'Agostino, Sr., D. Levy, J. M. Murabito *et al.* 2004 Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *Journal of the American Medical Association* **291**: 2204-2211.

Marenberg, M. E., N. Risch, L. F. Berkman, B. Floderus, and U. De Faire, 1994 Genetic susceptibility to death from coronary heart disease in a study of twins. *N.Engl.J.Med.* **330**: 1041-1046.

Martin, E. R., M. P. Bass, E. R. Hauser, and N. L. Kaplan, 2003 Accounting for linkage in family-based tests of association with missing parental genotypes. *American Journal of Human Genetics* **73**: 1016-1026.

Martin, E. R., N. L. Kaplan, and B. S. Weir, 1997 Tests for linkage and association in nuclear families. *American Journal of Human Genetics* **61**: 439-448.

Martin, E. R., S. A. Monks, L. L. Warren, and N. L. Kaplan, 2000 A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* **67**: 146-154.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.* 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat.Rev.Genet.* **9**: 356-369.

McGinnis, R., S. Shifman, and A. Darvasi, 2002 Power and efficiency of the TDT and case-control design for association scans. *Behavior Genetics* **32**: 135-144.

McPherson, R., A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts *et al.* 2007 A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**: 1488-1491.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag *et al.* 2003 PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**: 267-273.

Morris, A. P., R. N. Curnow, and J. C. Whittaker, 1997 Randomization tests of disease-marker associations. *Annals of Human Genetics* **61**: 49-60.

Morris, R. W., and N. L. Kaplan, 2002 On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology* **23**: 221-233.

Morton, N. E., 1955 Sequential Tests for the Detection of Linkage. *American Journal of Human Genetics* **7**: 277-318.

Morton, N. E., 1998 Significance levels in complex inheritance. *American Journal of Human Genetics* **62**: 690-697.

Murabito, J. M., M. J. Pencina, B. H. Nam, R. B. D'Agostino, Sr., T. J. Wang *et al.* 2005 Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. *Journal of the American Medical Association* **294**: 3117-3123.

Murray, J. C., K. H. Buetow, J. L. Weber, S. Ludwigsen, T. Scherpbierheddema *et al.* 1994 A Comprehensive Human Linkage with Centimorgan Density. *Science* **265**: 2049-2054.

Neale, B. M., and P. C. Sham, 2004 The future of association studies: Gene-based analysis and replication. *American Journal of Human Genetics* **75**: 353-362.

Nielsen, D. M., M. G. Ehm, D. V. Zaykin, and B. S. Weir, 2004 Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics* **168**: 1029-1040.

Pajukanta, P., M. Cargill, L. Viitanen, I. Nuotio, A. Kareinen *et al.* 2000 Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland. *Am.J.Hum.Genet* **67**: 1481-1493.

Peltekova, V. D., R. F. Wintle, L. A. Rubin, C. I. Amos, Q. Q. Huang *et al.* 2004 Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nature Genetics* **36**: 471-475.

Penrose, L. S., 1935 The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics* **4**: 133-138.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.

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

Rakovski, C. S., X. Xu, R. Lazarus, D. Blacker, and N. M. Laird, 2007 A new multimarker test for family-based association studies. *Genetic Epidemiology* **31**: 9-17.

Risch, N., and D. Botstein, 1996 A manic depressive history. *Nature Genetics* **12**: 351-353.

Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. *Science* **273**: 1516-1517.

Risch, N. J., 2000 Searching for genetic determinants in the new millennium. *Nature* **405**: 847-856.

Rissanen, A. M., 1979 Familial Occurrence of Coronary Heart-Disease - Effect of Age at Diagnosis. *American Journal of Cardiology* **44**: 60-66.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.* 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Samani, N. J., J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino *et al.* 2007 Genomewide association analysis of coronary artery disease. *New England Journal of Medicine* **357**: 443-453.

Samani, N. J., A. Hall, A. J. Balmforth, M. Mangino, J. Barrett *et al.* 2005 A genome-wide linkage study of 1933 families affected by premature coronary artery disease: The British Heart Foundation Family Heart Study. *Circulation* **112**: U413.

Saxena, R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. W. de Bakker *et al.* 2007 Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**: 1331-1336.

Schmidt, M., E. R. Hauser, E. R. Martin, and S. Schmidt, 2005 Extension of the SIMLA package for generating pedigrees with complex inheritance patterns: Environmental covariates, gene-gene and gene- environment interaction. *Statistical Applications in Genetics and Molecular Biology* **4**.

Schmidt, S., M. A. Hauser, W. K. Scott, E. A. Postel, A. Agarwal *et al.* 2006 Cigarette Smoking Strongly Modifies the Association of LOC387715 and Age-Related Macular Degeneration. *Am.J.Hum.Genet.* **78**: 852-864.

Schmidt, S., X. Qin, M. A. Schmidt, E. R. Martin, and E. R. Hauser, 2007 Interpreting analyses of continuous covariates in affected sibling pair linkage studies. *Genet.Epidemiol.* **31**: 541-552.

Schmidt, S., M. A. Schmidt, X. J. Qin, E. R. Martin, and E. R. Hauser, 2008 Increased efficiency of case-control association analysis by using allele-sharing and covariate information. *Human Heredity* **65**: 154-165.

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li *et al.* 2007 A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**: 1341-1345.

Seo, D., T. Wang, H. Dressman, E. E. Herderick, E. S. Iversen *et al.* 2004 Gene Expression Phenotypes of Atherosclerosis. *Arterioscler.Thromb.Vasc.Biol.* **24**: 1922-1927.

Shah, S. H., W. E. Kraus, D. C. Crossman, C. B. Granger, J. L. Haines *et al.* 2006 Serum lipids in the GENECARD study of coronary artery disease identify quantitative trait loci and phenotypic subsets on chromosomes 3q and 5q. *Annals of Human Genetics* **70**: 738-748.

Shea, S., R. Ottman, C. Gabrieli, Z. Stein, and A. Nichols, 1984 Family history as an independent risk factor for coronary artery disease. *J.Am.Coll.Cardiol.* **4**: 793-801.

Shephard, N., S. John, L. Cardon, M. I. McCarthy, and E. Zeggini, 2005 Will the real disease gene please stand up? *BMC.Genet.* **6 Suppl 1**: S66.

Shi, M., D. M. Umbach, and C. R. Weinberg, 2007 Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. *American Journal of Human Genetics* **81**: 53-66.

Shiffman, D., C. M. Rowland, J. Z. Louie, M. M. Luke, L. A. Bare *et al.* 2006 Gene Variants of VAMP8 and HNRPUL1 Are Associated With Early-Onset Myocardial Infarction. *Arterioscler.Thromb.Vasc.Biol.*

Smith, L. R., F. E. Harrell, J. S. Rankin, R. M. Califf, D. B. Pryor *et al.* 1991 Determinants of Early Versus Late Cardiac Death in Patients Undergoing Coronary-Artery Bypass Graft-Surgery. *Circulation* **84**: 245-253.

Spielman, R. S., and W. J. Ewens, 1998 A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**: 450-458.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens, 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506-516.

Stewart A.F.R., McPherson R., Chen L., Williams K., Kavaslar N., Rutberg J., Doelle H., Ewart G., Wells G.A. & Roberts R. Genome-wide scan for coronay artery disease genes using 500,668 markers. American Society of Human Genetics . 2007.
Ref Type: Conference Proceeding

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.* 2005 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 15545-15550.

Sun, L., N. J. Cox, and M. S. Mcpeek, 2002 A statistical method for identification of Polymorphisms that explain a linkage result. *American Journal of Human Genetics* **70**: 399-411.

Ten Kate, L. P., H. Boman, S. P. Daiger, and A. G. Motulsky, 1982 Familial aggregation of coronary heart disease and its relation to known genetic risk factors. *Am.J.Cardiol.* **50**: 945-953.

The Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-6U7.

Thom, T., N. Haase, W. Rosamond, V. J. Howard, J. Rumsfeld *et al.* 2006 Heart disease and stroke statistics--2006 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* **113**: 85-151.

Voss, R., P. Cullen, H. Schulte, and G. Assmann, 2002 Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Munster Study (PROCAM) using neural networks. *International Journal of Epidemiology* **31**: 1253-1262.

Wallace, C., S. J. Newhouse, P. Braund, F. Zhang, M. Tobin *et al.* 2008 Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *American Journal of Human Genetics* **82**: 139-149.

Wang, L., E. R. Hauser, S. H. Shah, D. Seo, P. Sivashanmugam *et al.* 2008 Polymorphisms of the Tumor Suppressor Gene LSAMP are Associated with Left Main Coronary Artery Disease. *Ann.Hum.Genet.*

Wang, L. Y., E. R. Hauser, S. H. Shah, M. A. Pericak-Vance, C. Haynes *et al.* 2007 Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. *American Journal of Human Genetics* **80**: 650-663.

Wang, Q., S. Rao, G. Q. Shen, L. Li, D. J. Moliterno *et al.* 2004 Premature Myocardial Infarction Novel Susceptibility Locus on Chromosome 1P34-36 Identified by Genomewide Linkage Analysis. *American Journal of Human Genetics* **74**: 262-271.

Wang, S. B., S. D. DeteraWadleigh, H. Coon, C. E. Sun, L. R. Goldin *et al.* 1996 Evidence of linkage disequilibrium between schizophrenia and the SCA1 CAG repeat on chromosome 6p23. *American Journal of Human Genetics* **59**: 731-736.

Wang, X., M. Ria, P. M. Kelmenson, P. Eriksson, D. C. Higgins *et al.* 2005 Positional identification of TNFSF4, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nature Genetics* **37**: 365-372.

Wang, X. S., and B. Paigen, 2005 Genetics of variation in HDL cholesterol in humans and mice. *Circulation Research* **96**: 27-42.

Weeks, D. E., and G. M. Lathrop, 1995 Polygenic Disease - Methods for Mapping Complex Disease Traits. *Trends in Genetics* **11**: 513-519.

Weinberg, C. R., 1999 Allowing for missing parents in genetic studies of case-parent triads. *American Journal of Human Genetics* **64**: 1186-1193.

Weiss, K. M., and J. D. Terwilliger, 2000 How many diseases does it take to map a gene with SNPs? *Nature Genetics* **26**: 151-157.

Whittemore, A. S., and I. P. Tu, 1998 Simple, robust linkage tests for affected sibs. *American Journal of Human Genetics* **62**: 1228-1242.

Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle *et al.* 2008 Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics* **40**: 161-169.

Woodroffe, A., C. M. Krafchak, N. Fuse, P. R. Lichter, S. E. Moroi *et al.* 2006 Ordered subset analysis supports a glaucoma locus at GLC1I on chromosome 15 in families with earlier adult age at diagnosis. *Exp.Eye Res.* **82**: 1068-1074.

Xu, H., S. G. Gregory, E. R. Hauser, J. E. Stenger, M. A. Pericak-Vance *et al.* 2005 SNPselector: a web tool for selecting SNPs for genetic association studies. *BioInformatics* **21**: 4181-4186.

Xu, X., C. Rakovski, X. P. Xu, and N. Laird, 2006 An efficient family-based association test using multiple markers. *Genetic Epidemiology* **30**: 620-626.

Yang, Q. N., C. Q. Lai, L. Parnell, L. A. Cupples, X. Adiconis *et al.* 2005 Genome-wide linkage analyses and candidate gene fine mapping for HDL3 cholesterol: the Framingham Study. *Journal of Lipid Research* **46**: 1416-1425.

Yuan, A., G. J. Chen, Y. X. Chen, C. Rotimi, and G. E. Bonney, 2004 Identifying the susceptibility gene(s) in a set of trait-linked genes using genotype data. *Genetics* **167**: 1445-1459.

Yusuf, S., S. Hawken, S. Ounpuu, T. Dans, A. Avezum *et al.* 2004 Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* **364**: 937-952.

Zdravkovic, S., A. Wienke, N. L. Pedersen, M. E. Marenberg, A. I. Yashin *et al.* 2002 Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *Journal of Internal Medicine* **252**: 247-254.

Zeggini, E., M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott *et al.* 2007 Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**: 1336-1341.