Characterization of Gene Interaction and Assessment of LD Matrix Measures for the Analysis of Biological Pathway Association

by

David R. Crosslin

Department of Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Dr. Terrence Furey, Supervisor

_____
Dr. Elizabeth Hauser

_____
Dr. Edwin Iversen

_____
Dr. Svati Shah

_____
Dr. Sayan Mukherjee

_____
Dr. Russ Wolfinger

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computational Biology & Bioinformatics
in the Graduate School of
Duke University

2009

ABSTRACT

(Computational Biology & Bioinformatics)

Characterization of Gene Interaction and Assessment of LD Matrix Measures for the Analysis of Biological Pathway Association

by

David R. Crosslin

Department of Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Dr. Terrence Furey, Supervisor

_____
Dr. Elizabeth Hauser

_____
Dr. Edwin Iversen

_____
Dr. Svati Shah

_____
Dr. Sayan Mukherjee

_____
Dr. Russ Wolfinger

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computational Biology & Bioinformatics
in the Graduate School of
Duke University

2009

Abstract

Leukotrienes are arachidonic acid derivatives long known for their inflammatory properties and their involvement with a number of human diseases, most notably asthma. Recently, leukotriene-based inflammation has also been implicated in atherosclerosis: *ALOX5AP* and *LTA4H*, two genes in the leukotriene biosynthesis pathway, have been associated with various cardiovascular disease (CVD) phenotypes. To assess the role of the leukotriene pathway in CVD pathogenesis, we performed genetic association studies of *ALOX5AP* and *LTA4H* in a non-familial sample of early onset coronary artery disease. Our results support a modest role for the leukotriene pathway in atherosclerosis pathogenesis, reveal important genomic interactions within the pathway, and suggest the importance of using pathway-based modeling for evaluating the genomics of atherosclerosis susceptibility. Motivated by this need, we investigated the statistical properties of a class of matrix-based statistics to assess epistasis. We simulated multiple two-variant disease models with haplotypes to gain an understanding of pathway interactions in terms of correlation patterns. Our goal was to detect an interaction between multiple disease-causing variants by means of their linkage disequilibrium (LD) patterns with other markers. The simulated models can be summarized into three categories: 1. No epistasis in the presence of marginal effects and LD; 2. Epistasis in the presence of LD and no marginal effects; and 3. Epistasis in the presence of marginal effects and LD. We then assessed previously introduced single-gene methods that compare whole matrices LD between two samples. These methods include comparing two sets of principal components, a

sum-of-squared-differences comparing pairwise LD, and a contrast test that controls for background LD. We also considered a partial least-square approach for modeling gene-gene interactions. Our results indicate that these measures can be used to assess epistasis as well as marginal effects under certain disease models. Understanding and quantifying whole-gene variation and association with disease using multiple SNPs remains a difficult task. Providing a single measure per gene will facilitate combining multiple types of genomic data and will serve as an alternative approach to assess epistasis in genome-wide association studies. The matrix-based measures could also be used in pathway ascertainment tools that require scores on a gene-level.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my advisor, Dr. Elizabeth Hauser. Her belief in me provided a foundation of confidence to learn and grow over the years. I would like to thank Drs. Svati Shah and William Kraus who provided unprecedented clinical training in the area of cardiovascular disease and genomics. I would also like to thank Sarah Nelson and Carol Haynes for considering me a part of the team at the Center for Human Genetics. Most importantly, I would not have been able to reach this milestone without the love and support from my wife, Rachel, and my two daughters, Meredith and Caroline.

Chapter 1

Introduction

1.1 Background

Genotyping technologies are providing unprecedented SNP coverage of the human genome, but large obstacles remain with the analyses. While SNPs have different functional implications, they are most often used to mark genetic variation within a gene or across the genome. Understanding which or if any of the many SNPs represent a contribution to disease risk can be difficult. Furthermore, the concept of defining a gene measure based on multiple SNPs is not well-developed. The probability of declaring SNPs to be significantly associated to a phenotype if they are truly associated while keeping the probability of making false declarations of association low remains a difficult task [17]. A large sample size is needed to detect the smaller gene effect sizes, which are normally overlooked when controlling for multiple testing [49][27][28]. These overlooked, hidden effects may play an important role in understanding complex pathways leading to disease phenotypes. These effects may also hinder candidate gene replication.

One extension of correlating single-marker genotypes with phenotypes (allelic association) is the characterization of multiple SNP patterns known as haplotypes in the presence of phenotypes. Haplotypes are specific combinations of alleles aligned (in phase) along a chromosomal region. Haplotypes may provide more information on the complex relationship between DNA or whole-gene variation and disease-related phenotypes compared to a single SNP [18][45]. Specific allele combinations along a chromosome for a gene may provide more information than a single SNP. Haplotype analyses localize a susceptibility gene via linkage disequilibrium (LD) with adjacent genetic markers or on the influence that the entire haplotype has on the trait [18]. LD is the non-random association of two alleles at different loci on the same gamete.

One difficulty in analyzing haplotypes in the case-control design is the ambiguity of phase in most data sets. Haplotyping methods include a parsimony algorithm (Clark 1990), a Bayesian population genetic model that uses coalescent theory (Stephens et al. 2001b; Zhang et al. 2001), and a maximum likelihood (Terwilliger and Ott 1994; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995) [18].

When selecting which SNPs to represent a haplotype, one must consider many factors. The two main approaches to haplotype construction are either choosing SNPs within (intra) or across (inter) LD bins, thus producing a higher-order haplotype. The intra-LD bin haplotypes are a set of alleles together on the same chromosome surrounded by hotbeds of recombination defined by a correlation threshold. Higher-order haplotypes use tagging SNPs (tSNP) to capture all variation at a given LD level. The crucial subset of markers to type would be those that distinguish one hap-

lotype from another [34]. Also, many methods use windowing strategies which test sets of perhaps contiguous SNPs. Obstacles with this method include non-contiguous LD patterns and and unevenly spaced SNPs throughout the genome. Understanding the benefits and shortfalls with each method are important when considering analysis strategies.

Another extension of single SNP methods is SNP tagging. For a given sample of a population, tagging SNPs are used to represent blocks of LD or haplotypes by a threshold of correlation. The idea is that a few SNPs can represent larger regions of a chromosome or gene of interest, thus reducing the number needed to genotype. SNP tagging is common practice with gene fine mapping and many GWAs panels are based on this format. Understanding what effect gene-tagging and the resulting LD pattern have on detecting gene-gene interactions will be beneficial. It is also important to note that the widely used strategies for tagging SNP (tSNP) selection are based on a single-disease-gene model, which may not apply to complex disease [39]. This approach assumes a common variant that is associated with a common disease with no interaction effects. The ability to detect interaction effects (gene-gene) may be attenuated with tagging strategies.

In addition to the one-to-many relationship of multiple SNPs per gene, there is an added layer of complexity when considering multiple genes in a metabolic pathway. For instance, complex diseases such as cardiovascular disease (CAD) involve multiple variables that could include the following: 1. Genetic variation; 2. Intermediate risk factors; 3. Family history; 5. Biomarkers; 6. Environmental conditions; and 7. The potential interaction of all of these.

An example of a complex metabolic pathway is the leukotriene biosynthesis pathway in CAD. Leukotrienes are arachidonic acid derivatives long known for their inflammatory properties and their involvement with a number of complex human diseases, most particularly asthma. Several genetic linkage and associations studies as well as gene expression studies have shown an association of the leukotriene biosynthesis pathway to CAD [36][25][24][19][37][30]. Helgadottir et al.[25] identified a four-SNP haplotype (HapA) spanning the *ALOX5AP* gene that conferred a nearly two times greater risk of MI and stroke in a case-control cohort. The *ALOX5AP* gene is mapped to a locus on chromosome 13q12.3 and encodes a protein that, with 5-lipoxygenase (*ALOX5*), is required for leukotriene synthesis. In addition, a ten-SNP haplotype (HapK) spanning the *LTA4H* gene encoding leukotriene A4 hydrolase (*LTA4H*) , a protein in the same biological pathway as *ALOX5AP*, was shown to confer a modest risk of MI in the Icelandic cohort [24]. The two haplotypes suggest two independent disease-causing mutations, but may indicate gene involvement within a pathway context. While haplotyping does characterize a single gene's association with disease using multiple SNPs, no gene-gene interaction is considered. And like single-SNP effects, smaller effects from multiple haplotypes for a given pathway may play an important role in more complex pathway analyses. Harnessing the additive effects and gaining power from considering multiple SNPs in multiple genes in a biological pathway could aid in the understanding of complex diseases like CAD.

1.2 Tests for interaction generalizing LD measures

More generalized correlation between SNPs could be due to interactions and functional relationships, or sampling on case status. Joint frequencies of alleles at different loci can be measured using a specific type of correlation called linkage disequilibrium (LD). LD is the non-random association of two alleles at different loci. Recently there have been multiple novel methods of assessing LD patterns between cases and controls as a measure of association [41][48][39]. The extent of LD can differ between the case and control groups in a region of genetic association, and the LD comparison can aid in the analysis of a candidate gene or region [39]. Genetic markers that are immediately adjacent on a chromosome might be statistically independent, whereas those that are hundreds of thousands or more base pairs apart might be highly correlated [34]. This lack of continuity is an argument against a sliding window (fixed number of contiguous SNPs) approach to test significance and an argument for pairwise comparisons across a region of interest. Nielsen et al. [3] published power results of comparing three test (single-marker case-control, haplotype-based case-control and LD contrast) with varied levels of pairwise LD in two and three SNP locus regions. While producing tables of pairwise-LD measures in genetic association studies is common practice, the observed graphic for case-control samples lack a statistical measure [41]. It is this graphic that is used to identify regions of the genome with similar and different LD patterns. The graphic may justify combining ethnic groups in candidate gene studies. Conversely, this graphic may justify splitting samples along ethnic lines.

Zaykin et al. [41] presented two methods for contrasting LD patterns. The first

method is a permutation-based statistic (Z_1) that measures the difference between two spaces (e.g. the case LD space and the control LD space) using spectral decomposition of the pairwise LD matrix. The second method is a sum-of-squared-differences statistic (Z_2) that measures the overall difference in the corresponding pairwise LD. Simulations presented by Zaykin showed that the sum-of-squared-differences statistics was the more powerful statistic.

Wang et al. [48] described a method that is similar to the sum-of-squared-differences statistic (Z_2), but the pairwise elements are the sum of subject-specific mean-corrected cross-products, by use of the best linear unbiased predictor (BLUP) of the means. In a mixed model, linear combinations of the fixed and random effects can be formed from linear combinations of the conditional means [44]. The unconditional mean is a population-wide average, whereas the conditional mean is an average specific to an observed set of random effects, which in our case is subject-specific [44]. Solving the mixed model equations yields predictors, or BLUPs, of the linear combinations of the fixed and random effects [44].

In a separate study, Zhao et al. [39] showed that interaction between two unlinked loci can be represented by LD. A similar test of measuring LD differences is presented to test the interaction between cases and controls. Zaykin et al. [41] evaluated performance of the whole-matrix LD statistic by comparing methods designed to detect either single-SNP effects or SNP interactions when the effects are associated with entire haplotypes, but these tests were based on single variant disease models. There is strong evidence that several mutations within a single gene can interact to create a "super allele" that has a large effect on the observed phenotype [18].

Complex diseases will have multiple functional sites, and it will be invaluable to understand the cross-locus interaction in terms of correlation between those sites in addition to the within-gene LD effect. Zhao et al. [39] indicated that the method proposed by Zaykin et al. [41] is similar to their method, but that while the LD contrast test was originally designed to test association between a single gene and disease, it has not been extended to testing gene-gene interaction. Wang et al. [48] suggested that their method should be useful when the LD contrast test is used to detect interaction among variants in LD, such as different variants in a candidate gene. The interaction will indicate the joint action of two genes in the development of disease [39]. Modeling a trait as an additive combination of single-locus and interaction terms is likely to limit the power to detect interaction and a combined measure may be more powerful [39]. Because the criteria for tSNP selection are based on only one pairwise LD measure between the marker and disease loci, the LD between tSNPs and loci may not be strong enough to ensure that indirect interaction between two loci will be detected. Thus, if the interacting loci are not selected as tSNPs, many loci with interactions will be missed [39].

The arguments outlined by Wang et al. [47] present a case for developing alternative methods for modeling gene-gene interaction using multiple SNPs. Wang et al. [47] presented a partial least-square (PLS) approach for modeling gene-gene as well as gene-environment interactions with multiple markers. In this study they compared the PLS approach to other methods to jointly test the interaction (termed as cross-locus gametic disequilibrium) and main effects. More importantly, they described advantages and disadvantages to other methods including Tukey's one-

degree-of-freedom model, logistic regression with factors generated from principal component analysis (PCA) and baseline logistic regression both with and without interaction terms. For instance, using logistic regression to exhaustively model all pair-wise interactions between two genes could present a large number of degrees of freedom with a large number of parameters [47]. While PCA is a commonly used tool for data reduction of predictors, no correlations of SNPs with the trait are characterized. So modeling a trait using logistic regression with the first few components may not perform well in certain LD structures [47]. Tukey's approach is based on the assumption that a SNP's interaction effect on the trait is approximately proportional to it's marginal effects on the trait. The model is not optimal in power when there exists no, or small, marginal effects [47]. The methods presented are tested on simulated data sets with the assumption that the causal variants are genotyped, and then tested with the assumption that the variants are not genotyped. Expanding on this approach will help converge on a model for complex diseases whose etiology is most likely derived from the interaction of multiple genes and environmental risk factors that ultimately affect metabolic states.

Through simulations outlined in Chapter 3, our goal was to describe the effects of the marginal disease effect size (relative risk RR), genotype frequencies and haplotype LD patterns (r^2) on interaction between two variants (or *cross-locus interaction*). We then assessed how the matrix measures performed in measuring *cross-locus interaction* in the presence of varying LD (including equilibrium), MAF and marginal (RR). We wanted to determine if we could detect an interaction between multiple unobserved disease-causing variants by means of their LD patterns with other mark-

ers. This is a subtle but important difference from measuring different LD patterns with the two disease-causing variants included in the model. This could obviously produce significant results, depending on the magnitude of the difference. In effect, our goal was to determine if the interaction in terms of RR can significantly change the LD patterns across SNP markers composing different haplotypes, such as those seen in the analysis of CAD association with *ALOX5AP* and *LTA4H* [25][24].

In summary, we evaluated these matrix-based methods in the setting of multiple risk variants in both simulated and in a genetic association study of the leukotriene pathway. We replicated several of the results observed by Helgadottir et al. in a separate case-control sample [15]: both HapA (*ALOX5AP*) and HapK (*LTA4H*) were significantly associated with CAD [25][24][19]. In addition to the observed haplotype association and multiple modest SNP effects in all three genes, we observed interactions by means of expression Quantitative Trait Locus (eQTLs) both with SNPs and haplotypes [15]. With the knowledge gained from the simulations, we explored using the matrix-based measures to determine if there was an interaction between haplotypes in multiple genes for the leukotriene biosynthesis pathway. Having the prior knowledge of haplotype and single SNP association results for the leukotriene pathway provided an excellent comparison for the matrix-based methods results. The results for all methods coupled with the LD patterns for Caucasians and African Americans will help focus further research in the area of LD-contrast methods.

Chapter 2

Genetic Effects in the Leukotriene Biosynthesis Pathway and Association with Atherosclerosis

2.1 Introduction

Cardiovascular disease is a major burden on health care in the United States and remains the leading cause of morbidity and mortality in Western society [7] [6]. Coronary artery disease (CAD), the most common manifestation of cardiovascular disease, is characterized by atherosclerotic lesions in the epicardial coronary arteries. Risk factors for atherosclerosis, including smoking, dyslipidemia, hypertension, diabetes and obesity, have been identified to be important in many large scale epidemiological and intervention studies [2]. However, despite consistent evidence of a strong heritable nature to CAD risk, the underlying genetic architecture remains largely elusive. Understanding the etiology of complex disease traits such as atherosclerosis involves modeling multiple factors or variables that include genetic variation, intermediate risk factors, family history, biomarkers, environmental conditions and the potential interaction of all of these.

It is known that the processes of atherosclerotic plaque formation and rupture are driven by inflammation. Plaque rupture correlates with increased inflammation within the plaque, implicating the genes involved in inflammatory processes as excellent candidates for study [12]. The 5-lipoxygenase (5-LO) cascade leads to formation of leukotrienes, which exhibit strong pro-inflammatory activities in cardiovascular tissue [2]. Several genetic linkage and associations studies as well as gene expression studies have shown an association of the *ALOX5* / *ALOX5AP* pathway to CAD [36] [25] [24] [19] [37] [30]

Helgadottir et al. performed a genome-wide scan in search of myocardial infarction (MI) susceptibility genes using 1,068 microsatellite markers in 296 multiplex Icelandic families including 713 individuals [25]. While no regions resulted in genome-wide significance, the most promising observation (lod score 2.86) was on chromosome 13q12-13. Within this region, a four-SNP haplotype (HapA) spanning the *ALOX5AP* gene conferred a nearly two times greater risk of MI and stroke in a separate case-control cohort. The *ALOX5AP* gene is mapped to a locus on chromosome 13q12.3 and encodes a protein that, with 5-lipoxygenase (*ALOX5*), is required for leukotriene synthesis. The same group also published an *ALOX5AP* haplotype (HapB) that conferred risk in a United Kingdom case-control cohort, although HapA was not significant in this additional sample [25]. In addition, a ten-SNP haplotype (HapK) spanning the LTA4H gene encoding leukotriene A4 hydrolase (*LTA4H*) on chromosome 12q23.1, a protein in the same biological pathway as *ALOX5AP*, was shown to confer a modest risk of MI in the Icelandic cohort [19]. In a separate study, Seo et al. analyzed human aorta samples with varying degrees of atherosclerosis

to identify gene expression patterns that predict well-defined aortic atherosclerosis [46]. The two diseased groups (minimally and severely) had significantly different pathological severity of atherosclerosis, as determined by raised lesions and Sudan IV staining. Among the genes predictive of severity of atherosclerosis was *ALOX5*, but not *ALOX5AP* or *LTA4H*.

Given these previous results, the goal of our study was to determine the association between CVD phenotypes, expression and previously reported *ALOX5AP* and *LTA4H* haplotypes. We attempted to validate the genetic association findings in both a family-based dataset of early-onset CAD (EOCAD) (GENECARD) as well as a non-familial CAD dataset of EOCAD (CATHGEN), including both MI and more general classification of CAD as outcomes. Further, to understand the role of the leukotriene pathway in atherosclerosis pathophysiology, we used expression and clinical data from human donor aortas (AORTA) to evaluate correlations among gene expression patterns in the leukotriene biosynthetic cascade and severity of histologically determined atherosclerosis. Our study takes a comprehensive approach to assess genetic association, genotype-phenotype correlation and gene interactions in three previously implicated candidate genes in the leukotriene pathway.

2.2 Methods

Three sample sets were available for the genetic association analysis: CATHGEN, GENECARD and AORTA. Descriptive statistics for the CATHGEN and GENECARD samples (Table 2.1) were generated using the *summary.formula function* in the *Hmisc library* (R Statistical Computing).

2.2.1 CATHGEN Sample

DNA samples used in this study were collected through the Cardiac Catheterization Laboratories at Duke University Hospital (Durham, NC) under a protocol approved by the Duke Institutional Review Board. Beginning in January 2001, all individuals presenting to the catheterization laboratory for cardiac catheterization were invited to participate in the CATHGEN study and signed informed consent to give a blood sample and allow abstraction of medical record information. Collected samples were later joined to the diagnostic and outcome information stored in the Duke Information System for Cardiovascular Care database maintained at the Duke Clinical Research Institute (Durham, NC). DNA samples and clinical data from over 2000 individuals were available for this study. From these, 1061 subjects were selected for the CATHGEN study on the basis of extent of coronary artery disease measured by the coronary artery disease index (CADi) and age-at-catheterization. CADi is a numerical summary of coronary angiographic data that incorporates the extent and anatomical distribution of coronary disease [20]. For the CATHGEN sample, we defined EOCAD as having an age-at-catheterization of 55 years of age or less and significant CAD (CADi $\geq$ 32) (n = 656). The unaffected control group was defined as older than 60 with insignificant CAD (CADi $\leq$ 23) and no individual epicardial coronary artery with clinically significant, (i.e. $>$ 50% stenosis) (n = 405). In addition, the unaffected controls had no history of documented cerebrovascular or peripheral vascular disease, cardiac transplant, myocardial infarction or interventional cardiac catheterization procedure. We selected a subgroup of individuals who experienced MI from the overall CATHGEN sample. Any subject regardless of age

who has ever had documentation of an MI, either prior to the index catheterization, or subsequent to the index catheterization was classified as a case ($n = 483$) and compared to the unaffected control group as defined above ($n = 405$). This includes thrombolytic therapy for an MI in the past or follow-up.

2.2.2 The GENECARD family study

The primary goal of the GENECARD study was to provide a genome-wide linkage scan to identify genetic factors for CAD by linkage analysis. The study was coordinated through Duke University and the study design has been previously described [52]. To be eligible for the linkage study, families were required to include at least two siblings, each of whom met the diagnostic criteria for early onset cardiovascular disease (EOCAD) [52]. In addition, individuals were recruited if they had been diagnosed before the age of 51 years in men and 56 year in women. A total of 420 families were included in the linkage analysis. For the purpose of association analysis we also included families with only one member meeting the diagnostic criteria but with living parents or a living older unaffected sibling. Blood samples were obtained and DNA extraction was performed by a standard protocol at the Duke Center for Human Genetics (CHG). Medical history and risk-factor information was obtained by interviewing patients and by abstracting information from medical records [52]. ACS is a serious manifestation of CAD which includes the subgroup of MI and is diagnosed by the presence of at least two of the three signs/symptoms: 1. Chest pain typical of CAD; 2. Changes on electrocardiogram; and 3. A positive serum biomarkers for MI. In addition, an ACS family must have two or more members that

qualify individually for ACS ($n = 428$). Results from the initial genome scan have been reported [51].

2.2.3 AORTA Sample

Aorta samples from heart donors with varying degrees of atherosclerosis were harvested in *University of Wisconsin* solution on ice to minimize postmortem changes [46]. RNA processing methods are referenced in Seo et al. [46]. Early atherosclerotic plaques were assessed with image processing by quantifying the area of Sudan IV staining and advanced disease was quantified as area of raised lesion using PDAY methodologies [11]. The ratio of affected area over total surface of the studied section was used as the disease burden outcome.

2.2.4 Genotyping

We selected a total of 31 SNPs in three genes: *ALOX5*, *ALOX5AP* and *LTA4H*. For the *ALOX5* gene, a minimal set of haplotype tagging SNPs (htSNPs) using minor allele frequency > 0.05 and $r2 > 0.7$, were selected using the SNPselector program [40] to cover the predicted linkage disequilibrium (LD) structure in both Caucasian and African American populations. SNPselector incorporates available information from Hapmap, The SNP Consortium, Japanese SNP database and the Affymetrix 120K SNP 20 to generate the most likely LD bins and determines the optimal tag SNP for each bin. The *ALOX5AP* and *LTA4H* SNPs were selected on the basis of the candidate haplotypes: HapA, HapB and HapK.

SNP probe and primer construction were performed and purchased from Applied

Biosystems (AB) for the TaqMan® colorimetric microtiter-plate based assay. The AB 7900 HT Sequence detection system was used for high-throughput genotyping for all three samples. The scoring of the genotypes is performed using Sequence Detection System (SDS) 7.1 software provided by AB. A total of 15 quality control samples, composed of six reference genotype controls in duplicate, two Centre d'Etude du Polymorphisme Humain pedigree individuals, and one no-template sample were included in each quadrant of the 384-well plate. Quality control consisted of evaluation of duplicate genotypes for mismatches, genotyping efficiencies and Hardy Weinberg Equilibrium (HWE). All SNPs examined were successfully genotyped for 95% or more of the individuals in the study. Error rate estimates for SNPs meeting QC benchmarks were $< 0.2\%$.

2.2.5 Statistical Analysis

Allele and Gene Frequencies

All markers had a minor allele frequency > 0.05. Haploview was used to assess linkage disequilibrium (LD) between SNPs [42]. A two marker EM method as implemented in Haploview was used to estimate the maximum-likelihood values of D' and r^2. Association between single SNPs or haplotypes with EOCAD was assessed through logistic regression models, using an additive allele model. In addition to the term for the genotype, the basic model included adjustment for race and sex and the full model included adjustment for known CAD risk factors including history of diabetes, history of smoking, body mass index, hypertension and history of dyslipidemia.

Haplotype analysis

The Haplo.stats package through R Statistical Computing was used to identify haplotypes and to provide a measure (haplo.score) of association to disease [18]. Haplo.stats expands on the likelihood approach to account for phase ambiguity in case-control studies by using a generalized linear model (GLM) to test for haplotype association which allows for adjustment of non-genetic covariables [18]. This method derives a score statistic to test the null hypothesis of no association of the trait with the genotype, $H_0 : \beta = 0$. In addition to a global statistic, haplo.stats computes score statistics for the components of the genetic vectors, such as individual haplotypes [18]. Because we wanted to assess the role of the previously identified haplotypes, we tested these individual haplotypes for association. Models were controlled for age, sex, race, and CAD risk factors as described above. We also performed race stratified analyses to control for potential confounding by race as well as to evaluate the previously reported race-specific results [19]. Both Gaussian and binomial traits were considered depending on the phenotype.

Raw gene expression data transformation and normalization

The AORTA sample microarray signal intensity values were normalized using the justRMA function in Bioconductor. We analyzed *cis* and *trans* effects of variants in the four Affymetrix tags representing the three genes of interest. The second phenotype for the AORTA study was the location of the section (distal and proximal) within the thoracic aorta as a surrogate for disease susceptibility [46]. Because some subjects had multiple aorta samples, each individual and sample was treated sepa-

rately. The expression level of each tag was modeled using multiple linear regression including age, sex, race and additive genotype. CAD risk factors were not available for the AORTA sample. To account for the multiple aorta samples per subject we fit a mixed model, which adjusts for the correlation between aorta samples from the same individual. In a mixed model, linear combinations of the fixed and random effects can be formed from linear combinations of the conditional means [44]. The random effect for an aorta with a distal and proximal section collected is subject-specific. We did not see a significant change in results when controlling for repeated measures.

Gene set enrichment analysis

Using prior knowledge of the leukotriene biosynthesis pathway provided by Funk and derived from KEGG, we generated a custom gene set [6][38]. We created a leukotriene biosynthesis pathway gene set incorporating *ALOX5*, *ALO5AP* and *LTA4H* and adding the following genes: 1. Leukotriene-C4 synthase (*LTC4S*); 2. Gamma-glutamyltransferase-like activity 1 (*GGTLA1*); and 3. Leukotriene-B4 20-monooxygenase (*CYP4F2*). Figure 2.5 illustrates the correlation structure for all the genes in the pathway derived from the aorta specimens. As input, GSEA uses a sorted correlation metric between expression and phenotype. An enrichment score (ES) is calculated that reflects the degree to which a gene set is overrepresented at the extremes of the entire list [23]. In addition to our custom set, we also analyzed the C2 gene sets curated from a large number of online pathway data bases [23]. We generated correlation coefficients for all tags using raised lesion mapping and Sudan

IV staining as the phenotypes. Tags without gene assignments (including 33153_AT) were removed. Statistical significance was assessed using empirical p-values estimated by randomly sampling gene sets of the same size and correlation coefficient sign [1].

Family-based analyses

Nonparametric relative pairs linkage analysis as implemented in MERLIN (Multipoint Engine for Rapid Likelihood INference) was used to assess two-point linkage of each SNP in the GENECARD study [31]. To assess family-based association with MI and EOCAD in GENECARD, association in the presence of linkage (APL), the Pedigree Disequilibrium Test (PDT) and geno-PDT were used [43]. PDT is a family-based association test for extended pedigrees. It will perform allele-specific analysis and genotype-specific analysis for single markers. APL provides a novel test for association in the presence of linkage that also correctly infers missing parental genotypes by estimating identity-by-descent (IBD) parameters ([14]. It provides options for single locus and multiple locus haplotype analysis. Simulations show APL has more power than PDT and FBAT/HBAT in nuclear family data. However, unlike PDT, APL does not consider extended pedigrees, using only one nuclear family from each pedigree. Given the varied family structures in GENECARD, we used both PDT and APL to maximize power for detecting association in these families. For haplotype analysis, we used HBAT (Haplotype version of the Family Based Association Test). This program uses data from nuclear families, sibships, or a combination of the

[1]Stoyan Georgiev personal communication

two, to provide a general-purpose family-based testing strategy for allelic association between phenotypes and haplotypes [55].

Power calculations

The application nQuery Advisor 4.0 was used to estimate the power with our sample size, and the effect size generated from the case-control proportions for HapA, HapB and HapK published by Helgadottir et al. [24][19]. We used the continuity corrected χ^2 test with $\alpha = 0.05$ two-sided significance level to detect the difference between proportions given our sample size for the MI and CAD groups. Power estimates for the GENECARD study design have been reported [52].

2.3 Results

Table 2.1 depicts the clinical characteristics for the GENECARD probands and CATHGEN case and control subjects. The clinical characteristics of the affected individuals from GENECARD and from CATHGEN include increased prevalence of common CAD risk factors including poor lipid profiles, diabetes, hypertension, overweight, and male gender compared to the controls. These differences suggest the need for adjustment for these common clinical risk factors in understanding genetic risk.

We first addressed the previously published association results. With prior knowledge of haplotype references for both the *ALOX5AP* and *LTA4H* genes, we analyzed SNPs in the case-control (CATHGEN) and family-based samples (GENECARD) which included all SNPs in HapA, HapB and HapK [25][24]. We also analyzed SNPs

from *ALOX5* to assess further association in genes from the leukotriene biosynthesis pathway. Figure 2.3A illustrates the location of the HapA and HapB SNPs within the *ALOX5AP* gene. Figure 2.3B illustrates the location of HapK SNPs within the *LTA4H* gene. All SNPs in the haplotypes are non-coding SNPs located in introns, or 5′ and 3′ regions. Figure 2.3C illustrates the *ALOX5* haplotype tagging SNPs (htSNPs) identified using the SNPselector program (see Materials and Methods). All SNPs are intronic except for rs2228065 which is a missense mutation.

Linkage disequilibrium (LD) plots stratified by race and EOCAD affection status in the CATHGEN sample for *ALOX5AP*, *LTA4H* and *ALOX5* show minimal LD between SNPs with the strongest correlation being between rs4769874 and rs9315050 members of HapA and HapB, respectively (Figure 5.1). As often noted, LD is reduced in African Americans compared to European Americans. For the CATHGEN sample, there was significant LD between SNPs in *LTA4H*. For the Caucasians, the novel downstream SNP SG12S16 was moderately correlated to rs17677715, rs2247570, and rs2660890 with r^2 values of 0.67, 0.70 and 0.63 respectively, but there was only weak correlation in African Americans. All markers met Hardy-Weinberg equilibrium (HWE) expectations when stratified by race.

2.3.1 Single Marker and Haplotype Association with Measures of EOCAD

Our initial goal was to test the previously reported association of *ALOX5AP*, *LTA4H* and *ALOX5* SNPs with atherosclerosis. Each dataset has unique measures of atherosclerosis; however, we attempted to match the MI phenotype from the previously published associations in the datasets as well as to broaden the phenotype to include more

general EOCAD. The phenotypes for the GENECARD and CATHGEN datasets were EOCAD and MI, while the phenotypes for the aorta dataset were proportion raised lesion and Sudan IV staining as measures of atherosclerosis burden.

Single SNP analyses

Table 2.3 lists the odds ratio (OR) estimates and significance levels for each marker comprising Haplotypes A and B within *ALOX5AP*. Marker rs17216473 demonstrated evidence of association with EOCAD ($P = 0.01$) and rs17222842, demonstrated evidence of association with MI ($P = 0.02$); these SNPs are members of HapA and HapB respectively. There was suggestive association ($P = 0.06$) with rs10507391 and the Sudan IV scoring phenotype. None of the SNPs in *ALOX5AP* demonstrated significant family-based association in GENECARD (data not shown).

Table 2.4 shows results for *LTA4H*; we detected no evidence for single SNP associations with EOCAD, MI or either AORTA phenotypes. However, one SNP in the *LTA4H* gene provided significant results in the GENECARD family-based association analysis; rs6538697 was significant for both acute coronary syndrome (ACS-MI) ($P = 0.006$) and CAD ($P = 0.0098$). Table 2.5 shows single SNP association results for *ALOX5*. Three SNPs in *ALOX5* were significant in the EOCAD sample including rs10900215 ($P = 0.05$), rs3740107 ($P = 0.04$) and rs1487562 ($P = 0.03$). For the AORTA samples, *ALOX5* SNPs showed significance: rs892691 was significant for the raised lesion phenotype, while three SNPs were significant in the Sudan IV staining phenotype: rs2029253 ($P = 0.05$), rs1369214 ($P = 0.01$) and rs2115819 ($P = 0.01$). SNPs rs3780902 and rs2228065 had minor allele frequencies < 0.05 and we were

unable to estimate an accurate OR for aorta phenotypes. Again, the family-based association analysis showed no significant *ALOX5* SNPs in GENECARD.

Haplotype analyses

The haplotype-specific results for MI and EOCAD phenotypes and the haplotype frequencies for cases and controls are shown in Table 2.2. Table 2.6 contains the power estimates for our samples given the case-control haplotype frequencies and effect sizes reported by Helgadottir et al. [25].

HapA in ALOX5AP HapA comprises the SNP markers rs17222814 (G), rs10507391 (T), rs4769874 (G) and rs9551963 (A). Given the HapA case-control frequencies and effect size ($RR = 1.79$; cases - 0.158, controls - 0.095) with MI as the clinical endpoint found in the Helgadottir et al. Icelandic cohort, we estimated that we had 73% power with our CATHGEN sample of 819 subjects (414 MI subjects and 405 older controls) [25]. Assuming the same effect size for CAD, we had 82% power for the EOCAD sample ($n = 1061$). HapA showed a trend for association with MI in Caucasians ($P = 0.07$) with case-control frequencies of 0.147 and 0.119. However, we did not observe a significant association with MI in our African American sample ($P = 0.33$) with haplotype frequencies of 0.092 and 0.108 in cases and controls, respectively. We observed a significant HapA association in the EOCAD Caucasian sample ($P = 0.01$; case - 0.166, control - 0.118). HapA was not significant in the EOCAD African Americans ($P = 0.67$; cases - 0.084, controls - 0.107). The overall test (global) of all haplotype association with the HapA markers was not significant for either ethnic group.

HapB in ALOX5AP HapB comprises the SNP markers rs17216473 (A), rs10507391 (A), rs9315050 (A) and rs17222842 (G). With our CATHGEN sample of 819 subjects, we had 50% power to detect the HapB results ($RR = 1.95$; cases - 0.075, controls - 0.040) reported by Helgadottir et al. [25]. We did not detect an association of HapB with neither EOCAD nor MI in the overall sample or when stratified by race. There were no significant single haplotype results for differences between HapB cases and controls. However, we observed a significant global P-value for MI in Caucasians ($P = 0.0001$) suggesting overall differences in haplotype frequencies with that combination of SNPs. In almost every test for association, the haplotypes with the A allele in place of the G allele for marker rs17222842 were significant. This difference in the associated haplotypes supports the significant global P-value and suggests the potential for additional associated haplotypes in this sample or may indicate further divergence of the haplotypes from an ancestral haplotype containing the disease mutation detected in the Icelandic population.

HapK in LTA4H HapK comprises the SNP markers SG12S16 (deCODE) (C), rs2660880 (G), rs6538697 (T), rs1978331 (A), rs17677715 (T), rs2247570 (T), rs2660898 (T), rs2540482 (C), rs2660845 (G) and rs2540475 (G). The CATHGEN sample of 656 Caucasians subjects with MI provided 27% power to detect reported the HapK case-control frequencies ($RR = 1.37$; cases - 0.186, controls - 0.143) [24]. Given our sample of African American subjects with MI, we had 49% power to detect the HapK case-control frequencies ($RR = 6.50$; cases - 0.103, controls - 0.017) found in the Philadelphia African American sample[24]. We did not observe HapK as a

significant risk haplotype in the sample of Caucasians with MI ($P = 0.79$; cases - 0.154, controls - 0.136) nor in the African Americans with MI ($P = 0.63$; cases - 0.069, controls - 0.075). As in the case for HapA, when we expanded the phenotype to EOCAD, the Caucasians showed the largest difference in haplotypes frequencies between cases and controls ($P = 0.03$; cases - 0.158, controls - 0.137).

2.3.2 GENECARD family-based association analyses

We used HBAT to test for family-based association, but we found no significant ($P < 0.05$) global or specific candidate-haplotype results for ACS or EOCAD [55]. Given that there was no evidence for linkage in the GENECARD genome-wide linkage analysis, perhaps it is not surprising that there was no evidence for the haplotypes with EOCAD and ACS in this dataset [51]. The linkage lod scores in the regions around the genes were highest for *ALOX5AP* but were still very low (max multipoint lod score = 0.01 and 0.12 in ACS families for *ALOX5AP*; max multipoint lod score = 0.25 and 0.27 in ACS-MI families for *LTA4H*; and max multipoint lod score = 0.09 and 0.13 in ACS-MI families for *ALOX5*. The power to detect linkage in 400 affected sibpairs (ASPs) is over 80% for recurrence ratios of 1.4 or greater and is over 80% for a variety of genetic models with 250 families of 2 ASPs and one affected sib at an $\alpha = 0.001$ [14]. The estimated HapA frequencies are 0.170 for EOCAD and ACS. For HapB, the estimated frequencies are 0.070 for EOCAD and ACS. The estimated HapK frequencies are 0.140 for EOCAD and 0.170 for ACS, frequencies consistent with those observed in the CATHGEN case-control sample

In summary, the genetic associations provide some support of the published re-

sults for *ALOX5AP* and *LTA4H* with additional support for *ALOX5*. None of these results would be significant with most multiple test corrections and thus individually the studies of these genes provide weak validation in the MI and general EOCAD phenotypes; however, the multiplicity of even these weak results within the leukotriene biosynthesis pathway do support the evidence of association with this pathway as a whole. Our next step was to evaluate correlations and interactions among the three genes.

2.3.3 Expression results for the AORTA sample

In the AORTA dataset previously used in the genetic association studies, 122 aorta tissue sections (proximal1A and distal4B) from 78 subjects were assayed for gene expression using the Affymetrix HG-U95Av2 chip. Both *LTA4H* and *ALOX5AP* have one representative tag and *ALOX5* has two tags (307_at and 33153_at). Tag 307_at represents the same strand (+) and same region as the *ALOX5* gene while 33153_at probes a short region of the complement (−) at the 3′ end. Figure 2.1 illustrates the RMA normalized expression values for *LTA4H*, *ALOX5* and *ALOX5AP*. The diagonal presents a histogram of the individual RNA expression values. Pairwise XY scatter plots are illustrated below the diagonal and the calculated correlation coefficients are found above the diagonal. Tags ALOX5_307_AT and ALOX_33153_AT were negatively correlated ($r = -0.19, P = 0.04$) despite their physical proximity suggesting that these tags are identifying different transcripts. *ALOX5AP* and *LTA4H* expression levels were significantly correlated to ALOX5_307_AT ($r = 0.54, P =< 0.0001$ and $r = 0.29, P = 0.0015$ respectively) and to each other ($r = 0.42, P =< 0.0001$).

As expected, *ALOX5AP* and *LTA4H* expression levels were negatively correlated with the ALOX5_33153_AT tag ($r = -0.29, P = 0.001$ and $r = -0.35, P < 0.0001$ respectively). We then separately considered the AORTA phenotypes and the expression values as a predictor. In support of the findings of Seo et al.[46], we found that expression values for *ALOX5* had a strong effect on raised lesion mapping ($point\,estimate = 7.33, p = 0.0005$) and Sudan IV staining ($point\,estimate = 4.65, p = 0.007$). No other tags demonstrated significant association with the AORTA phenotypes.

We examined genotype and specific expression for *ALOX5AP*, *LTA4H* and *ALOX5* using the previously studied SNPs found in HapA, HapB and HapK as well as SNPs found in the *ALOX5* gene. In addition to looking at *cis*-effects (intra-gene/SNP interaction) on gene expression, we also considered *trans*-effects (inter-gene/SNP interaction), as a model of the leukotriene biosynthesis pathway using a biological systems approach. Tables 2.3, 2.4 and 2.5 lists the P-values for the SNP effects on the four tags representing *ALOX5*, *ALOX5AP* and *LTA4H*. To visualize possible interactions and to highlight cis versus trans effects we have illustrated the genotype and haplotype-specific expression association results from Table 2.2 in Figure 2.4. All three genes have at least one SNP with a *cis* and *trans*effect as shown in Supplemental Figure 2.4 and Table 2.2. Interestingly, all effects for a given gene had a single gene target in terms of expression value. For example, SNPs in *ALOX5AP* only have *trans* effects on the expression values for *ALOX5*, but not for *LTA4H*. SNP rs10507391 in *ALOX5AP* demonstrated one of the most significant associations with mRNA expression levels (Figure 2.2). We chose to illustrate this SNP because of its

trans-effect significance ($P = 0.01$) on ALOX_307 expression level. We also considered a dominant model and the association was also significant ($P = 0.009$); although the goodness of fit determined by the log likelihood was not different for the two genetic models. This SNP is the only marker shared by both HapA and HapB; however, different alleles are included in the HapA and HapB risk haplotypes. There was no effect on the ALOX_33153 espression which seems to serve as a natural control, nor was there an effect on the *ALOX5AP* (cis) or *LTA4H* (trans) expression. Given an additive haplotype effect model, HapA was associated with expression for *ALOX5AP* ($P = 0.03$) and potentially with *ALOX5* ($P = 0.06$). Neither HapK nor HapB were associated with *ALOX5*, *ALOX5AP* or *LTA4H* expression values.

For a pathway-based approach of association using the expression data, we used Gene Set Enrichment Analysis (GSEA) [23]. Using the prior-based biological knowledge of the leukotriene biosynthesis pathway, we generated a custom gene set to test concordance. The two phenotypes analyzed include raised lesion mapping and Sudan IV staining. The genes were ranked using the correlation between each individual tag and the phenotype variable as a score. For the raised lesion mapping, our custom gene set for leukotriene biosynthesis had a positive enrichment score of 0.6897 and an estimated significance level of $P = 0.004$. The Sudan IV phenotype produced a positive enrichment score of 0.6349 and an estimated significance level of $P = 0.009$. We view this strong enrichment for the leukotriene pathway as a whole as supportive of the modest genetic association results shown above for individual SNPs. These results support the hypothesis that these genes act in concert to increase risk for CAD, both at the level of affected tissue as well as at the level of clinical disease.

2.4 Discussion

We replicated several of the results observed by Helgadottir et al. in our sample: both HapA (*ALOX5AP*) and HapK (*LTA4H*) were significantly associated with CAD in the CATHGEN case-control subjects. Previously published results for association with MI, CAD and stroke combined with our results suggest that these SNPs and haplotypes are associated with vascular disease [36][25][19][24][37][30]. Overall, the case-control comparison of EOCAD in Caucasians demonstrated the strongest association for all three haplotypes with significant results for HapA and HapK. We were unable to detect an association with HapB. While we were able to determine a general association with CAD, we did not see significant specific association results for the MI phenotypes in our case-control groups.

According to the *LTA4H* results produced by Helgadottir et al., HapK is more strongly associated with MI in Philadelphia African Americans ($RR = 6.50$, $P = 0.0001$) when compared to Caucasians ($RR = 1.37$,$P = 0.010$) although both groups demonstrate significant association [19]. Similar results were found in the Cleveland and Atlanta samples, although with somewhat lower odds ratios [19]. HapK was not significant in our African American sample, but was associated in the Caucasians. Overall we observed similar frequencies found in the Helgadottir et al. study, but the case and control frequencies were not significantly different in our study, likely due to the low power to detect this difference. Another reason for differences in statistical results between Helgadottir et al. and our results for association with MI may be due to subtle differences in case and control definitions.

In addition to the evidence from prior studies about candidate genes and a

metabolic pathway, we had prior knowledge of haplotypes and SNPs to test for association with our CVD sample. In the context of our ongoing candidate gene analyses, the P-values from the logistic regression tests for these SNPs and the haplotype global associations would not have suggested pursuing these as candidates nor would they have been detected in a genome-wide association approach. The comparison of the single SNP and haplotype results shows the need for analysis of multiple SNPs per gene, even in a replication setting, and in the utility of additional genomic information such as the aorta expression study, along with knowledge of biological pathways.

Having multiple levels of genetic evidence gives us confidence in the observations that support the leukotriene-CVD association findings of Helgadottir et al. and other groups [36][25][19][24][37][30]. Given that genes with the strongest effects will be observed in more than one context, taking advantage of independent sources of genetic information is useful. As demonstrated in Figure 2.4, the one-on-one *trans* relationship between the SNP genotypes and the expression values allows for reconstructing gene networks. Our results suggest that HapA and the relationship to expression levels for ALOX5AP (*cis*) and ALOX5 (*trans*) may be an important feature that links the genetic and genomic results. Our GSEA results for our custom leukotriene biosynthesis gene set suggest another approach for assessing pathway association. Our results support previous association studies; however, the relationship between genetic variation and CVD outcomes is not driven by a single gene or SNP. By exploring the pathway in terms of risk for atherosclerosis, CAD and MI along with genotypic effects on expression, rather than looking at each component in isolation,

we observe significant complexity in the interactions that may alter risk profiles related to genetic variation. For example, it may be necessary to measure changes in allele-specific expression of ALOX5 when evaluating leukotriene inhibitors for potential primary or secondary prevention of MI. This pathway approach should be considered for evaluation of other studies of genetic variation in CAD.

Table 2.1: Clinical characteristics of the CATHGEN EOCAD cases-controls and GENECARD US probands; percentages and ($mean \pm SD$). The test statistic significance provided is for the CATHGEN EOCAD versus unaffected controls

Descriptive Statistics by EOCAD	CATHGEN EOCAD ($n = 656$)	CATHGEN Controls ($n = 405$)	Significance EOCAD vs. Controls	GENECARD ($n = 759$)
Age of Onset	(46 ± 6)		NA	(48 ± 10)
Age at Sampling	(52 ± 9)	(69 ± 7)	$P < 0.001^1$	(50 ± 7)
Self Reported Race			$P < 0.002^1$	
Black	22%(143)	19%(77)		8%(58)
Native American	5%(35)	3%(14)		3%(21)
Other	2%(13)	6%(25)		4%(31)
White	71%(465)	71%(289)		86%(649)
Male Sex	79%(519)	43%(174)	$P < 0.0001^2$	69%(522)
Positive Family History of CAD	55%(358)	27%(108)	$P < 0.001^2$	100%(759)
Body Mass Index [kg/m2]	(31 ± 6)	(29 ± 7)	$P = 0.001^1$	(30 ± 7)
Postive smoking history	69%(454)	40%(160)	$P < 0.001^2$	27%(202)
Positive History of Diabetes	33%(219)	21%(86)	$P < 0.001^2$	25%(188)
Positive History of Hypertension	70%(457)	67%(272)	$P = 0.40^2$	63%(477)
History of Myocardial Infarction	52%(341)	0%(0)	NA	63%(481)
History of CABG	40%(262)	0%(0)	NA	49%(349)
Systolic blood pressure [mm Hg]	(141 ± 24)	(150 ± 23)	$P < 0.001^1$	(142 ± 25)
Diastolic blood pressure [mm Hg]	(78 ± 14)	(77 ± 14)	$P = 0.30^1$	(78 ± 14)
Total Cholesterol [mg/dL]	(193 ± 62)	(192 ± 49)	$P = 0.30^1$	(200 ± 58)
LDL [mg/dL]	(109 ± 43)	(107 ± 36)	$P = 0.90^1$	(122 ± 59)
HDL [mg/dL]	(40 ± 12)	(52 ± 18)	$P < 0.001^1$	(44 ± 32)
Triglycerides [mg/dL]	(223 ± 261)	(157 ± 126)	$P < 0.001^1$	(212 ± 253)

$\pm$ represents $\hat{X} \pm 1SD$
Numbers after percents are frequencies.
Tests used:[1]Wilcoxon test;[2]Pearson test

Table 2.2: Haplotype A, B and K association p-values (individual and global) and case-control frequencies for the MI, EOCAD and AORTA phenotypes; results for MI and EOCAD are presented overall and stratified by race; expression tags for *ALOX5*, *ALOX5AP* and *LTA4H* were analyzed as a continuous variable.

Haplotype	Phenotype	Race	Haplo. Freq. case	Haplo. Freq. cont.	p-value	Global p-value
ALOX5AP HapA	MI	All	0.12	0.14	0.11	0.75
		AA	0.11	0.09	0.33	0.48
		CA	0.12	0.15	0.07	0.48
	EOCAD	All	0.12	0.15	0.02	0.45
		AA	0.11	0.08	0.67	0.78
		CA	0.12	0.17	0.01	0.25
	Raised Lesion	All	0.18	0.20	0.06	0.40
	Sudan IV		0.21	0.17	0.11	0.08
	307_AT (*ALOX5*)	All	NA	NA	0.06	0.97
	33153_AT (*ALOX5*)				0.31	0.55
	37099_at (*ALOX5AP*)				0.03	0.23
	38081_at (*LTA4H*)				0.16	0.77
ALOX5AP HapB	MI	All	0.08	0.09	0.64	1.00
		AA	0.14	0.14	0.87	0.80
		CA	0.06	0.08	0.61	0.0001
	EOCAD	All	0.08	0.10	0.39	0.72
		AA	0.14	0.16	0.37	0.89
		CA	0.06	0.08	0.44	0.65
	Raised Lesion	All	0.08	0.07	0.76	0.00001
	Sudan IV		0.09	0.07	0.58	0.04
	307_AT (*ALOX5*)	All	NA	NA	0.38	0.11
	33153_AT (*ALOX5*)				0.41	0.08
	37099_at (*ALOX5AP*)				0.24	0.21
	38081_at (*LTA4H*)				0.18	0.37
LTA4H HapK	MI	All	0.12	0.14	0.57	0.07
		AA	0.08	0.07	0.63	0.16
		CA	0.14	0.15	0.79	0.83
	EOCAD	All	0.12	0.13	0.04	0.33
		AA	0.07	0.05	0.27	0.62
		CA	0.14	0.16	0.03	0.54
	Raised Lesion	All	0.19	0.10	0.32	0.06
	Sudan IV		0.18	0.17	0.71	0.02
	307_AT (*ALOX5*)	All	NA	NA	0.41	0.18
	33153_AT (*ALOX5*)				0.16	0.07
	37099_at (*ALOX5AP*)				0.94	0.13
	38081_at (*LTA4H*)				0.27	0.0009

Table 2.3: *ALOX5AP*: Single SNP odds ratio estimates and P-values using logistic regression for the CATHGEN subjects with EOCAD ($n = 656$) versus unaffected controls ($n = 405$), all affected with myocardial infarction (MI) ($n = 483$) versus unaffected controls ($n = 405$), and the AORTA case-control samples (raised lesion mapping and Sudan IV staining); The relationship between expression level of each tag and SNP was modeled using multiple linear regression

ALOX5AP Hap A,B	EOCAD ($n = 1061$)		MI ($n = 888$)		Raised Lesion ($n = 201$)		Sudan IV ($n = 150$)		307_AT (*ALOX5*)	33153_AT (*ALOX5*)	37099_AT (*ALOX5AP*)	38081_AT (*LTA4H*)
SNP	P	OR	P	OR	P	OR	P	OR	P	P	P	P
RS17222814*	0.08	0.72	0.21	0.75	0.71	0.8	0.48	0.72	0.49	0.52	0.41	0.60
RS17216473*	0.01	1.47	0.13	1.31	0.95	1.03	0.35	0.70	0.12	0.98	0.45	0.86
RS10507391*$^{\psi}$	0.26	1.12	0.60	0.94	0.25	1.58	0.06	1.78	0.01	0.75	0.54	0.2
RS4769874*	0.30	1.29	0.51	1.22	0.56	0.63	0.68	0.78	0.01	0.39	0.09	0.6
RS9551963$^{\psi}$	0.80	0.97	0.63	1.06	0.40	1.36	0.81	0.94	0.26	0.13	0.03	0.66
RS9315050$^{\psi}$	0.59	1.09	0.85	1.04	0.56	0.71	0.24	0.55	0.06	0.64	0.25	0.37
RS17222842$^{\psi}$	0.10	1.41	0.02	1.77	0.20	2.49	0.52	1.41	0.32	0.93	0.15	0.27
SNPs represented are in candidate haplotypes HapA(*) and HapB($^{\psi}$)												

Table 2.4: *LTA4H*: Single SNP odds ratio estimates and P-values using logistic regression for the CATHGEN subjects with EOCAD ($n = 656$) versus unaffected controls ($n = 405$), all affected with myocardial infarction (MI) ($n = 483$) versus unaffected controls ($n = 405$), and the AORTA case-control samples (raised lesion mapping and Sudan IV staining); The relationship between expression level of each tag and SNP was modeled using multiple linear regression

LTA4H Hap A,B	EOCAD ($n = 1061$)		MI ($n = 888$)		Raised Lesion ($n = 201$)		Sudan IV ($n = 150$)		307_AT (*ALOX5*)	33153_AT (*ALOX5*)	37099_AT (*ALOX5AP*)	38081_AT (*LTA4H*)
SNP	*P*	*OR*	*P*	*OR*	*P*	*OR*	*P*	*OR*	*P*	*P*	*P*	*P*
12P0557	0.87	1.02	0.52	1.10	0.74	1.14	0.19	1.52	0.73	0.61	0.94	0.22
RS2660880	0.96	1.01	0.55	0.86	0.36	1.91	0.39	1.55	0.25	0.67	0.29	0.49
RS6538697	0.40	0.86	0.10	0.7	0.27	0.50	0.85	1.10	0.36	0.89	0.05	0.05
RS1978331	0.08	1.19	0.65	0.95	0.60	1.19	0.08	1.61	0.17	0.34	0.9	0.40
RS17677715	0.65	1.06	0.23	1.21	0.48	0.73	0.97	0.99	0.34	0.94	0.29	0.12
RS2247570	0.14	1.18	0.33	0.88	0.88	1.06	0.51	1.20	0.36	0.06	0.82	0.11
RS2660898	0.82	0.98	0.81	0.97	0.58	0.82	0.35	1.32	0.43	0.41	0.24	0.85
RS2540482	0.95	0.99	1.00	1.00	0.23	0.63	0.89	1.04	0.98	0.79	0.40	0.48
RS2660845	0.83	0.98	0.64	0.94	0.34	0.72	0.88	0.96	0.45	0.39	0.63	0.79
RS2540475	0.98	1.00	0.63	0.93	0.54	1.30	0.64	1.18	0.77	0.62	0.27	0.42

Table 2.5: *ALOX5*: Single SNP odds ratio estimates and P-values using logistic regression for the CATHGEN subjects with EOCAD ($n = 656$) versus unaffected controls ($n = 405$), all affected with myocardial infarction (MI) ($n = 483$) versus unaffected controls ($n = 405$), and the AORTA case-control samples (raised lesion mapping and Sudan IV staining); The relationship between expression level of each tag and SNP was modeled using multiple linear regression

ALOX5 Hap A,B	EOCAD ($n = 856$)		MI ($n = 698$)		Raised Lesion ($n = 75$)		Sudan IV ($n = 50$)		307_AT (*ALOX5*)	33153_AT (*ALOX5*)	37099_AT (*ALOX5AP*)	38081_AT (*LTA4H*)
SNP	*P*	*OR*	*P*	*OR*	*P*	*OR*	*P*	*OR*	*P*	*P*	*P*	*P*
RS1864414	0.32	1.19	0.21	1.26	0.96	1.04	0.15	0.47	0.60	0.66	0.56	0.01
RS3824613	0.36	1.18	0.3	1.21	0.94	1.06	0.19	0.50	0.70	0.80	0.50	0.03
RS2029253	0.74	1.04	0.97	1	0.51	1.63	0.05	3.17	0.02	0.11	0.96	0.31
RS1369214	0.53	0.92	0.43	0.9	0.23	2.64	0.01	8.05	0.25	0.25	0.91	0.64
RS2115819	0.58	0.93	0.47	0.91	0.25	2.58	0.01	10.4	0.04	0.31	0.8	0.99
RS10900215	0.05	1.39	0.17	1.27	0.53	0.63	0.23	0.52	0.92	0.24	0.08	0.64
RS892691	0.42	0.89	0.96	0.99	0.03	4.39	0.93	0.96	0.24	0.59	0.72	0.26
RS3780906	0.08	0.78	0.21	0.83	0.07	3.10	0.95	0.98	0.44	0.90	0.93	0.26
RS3740107	0.04	0.74	0.3	0.85	0.08	3.50	0.89	1.06	0.33	0.73	0.88	0.45
RS1487562	0.03	1.46	0.07	1.39	0.93	0.92	0.12	0.34	0.44	0.20	0.54	0.85
RS2242332	0.33	0.87	0.56	0.92	0.12	2.81	0.81	1.13	0.78	0.78	0.89	0.06
RS2242334	0.26	0.85	0.44	0.89	0.33	1.89	0.96	0.98	0.75	0.34	0.20	0.03

Table 2.6: Table of power for the CATHGEN sample given the HapA, HapB and HapK case-control frequencies and effect sizes found in the Helgadottir et al. Icelandic cohort with MI and EOCAD as the clinical endpoints. The last row contains power estimates for the GENECARD sample at a recurrence risk of ≥ 1.4

Haplotype	Effect size Relative Risks	Case/Control frequencies reported by Helgadottir et al.	Clinical endpoint in CATHGEN sample	Power
HapA	1.79	0.158 / 0.095	MI	73%
HapA	1.79	0.158 / 0.095	EOCAD	82%
HapB	1.95	0.075 / 0.040	MI	50%
HapB	1.95	0.075 / 0.040	EOCAD	58%
HapK	1.37	0.186 / 0.143	MI in Caucasians	27%
HapK	1.37	0.186 / 0.143	EOCAD in Caucasians	29%
HapK	6.50	0.103 / 0.017	MI in African Americans	49%
HapK	6.50	0.103 / 0.017	EOCAD in African Americans	55%
	Recurrence Risk ≥ 1.4		GENECARD 400 ASPs	>80%

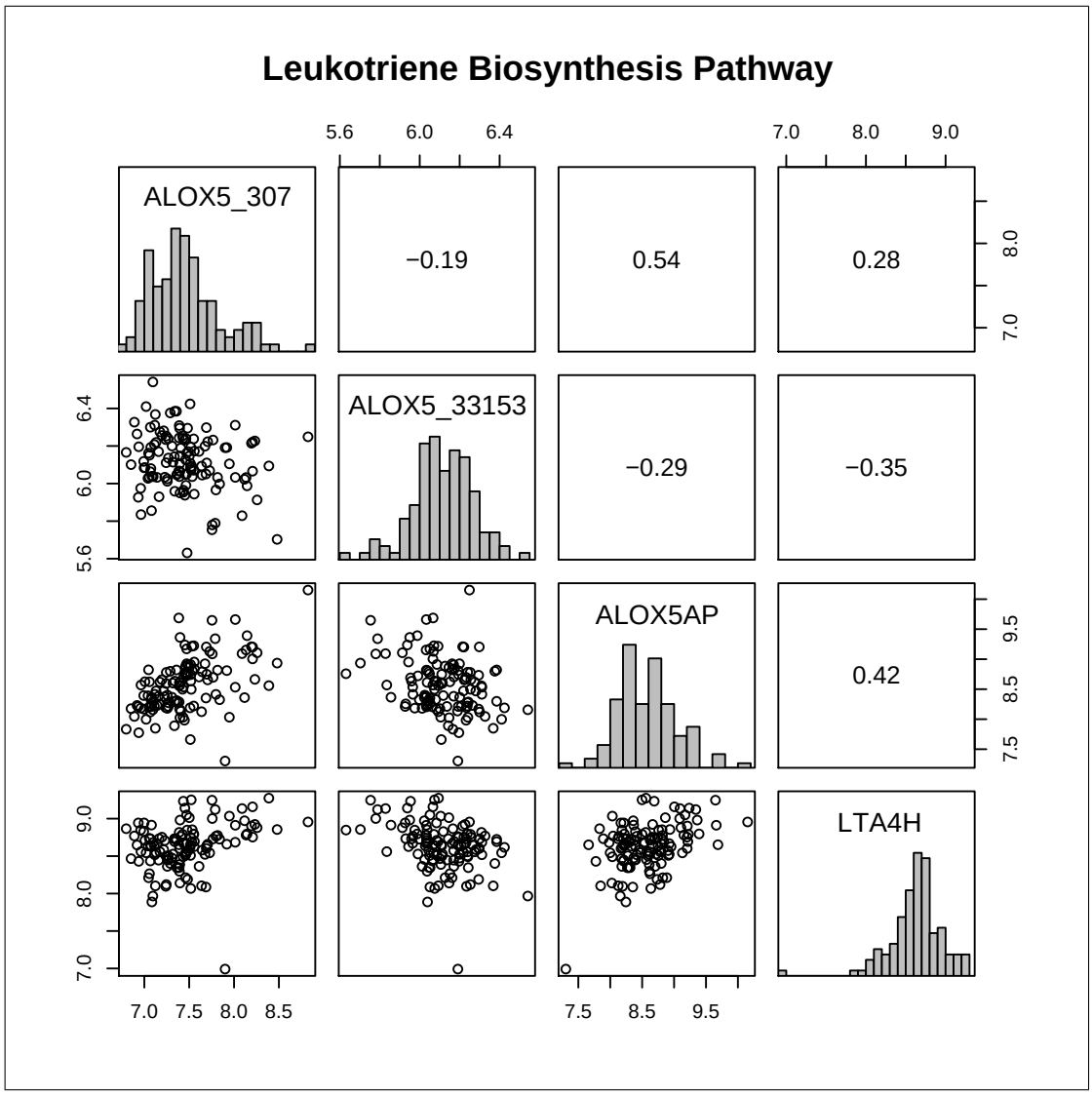

Figure 2.1: Pairwise scatterplots (below the diagonal), histograms (across the diagonal) and pairwise correlation coefficients (above the diagonal) using RMA normalized gene expression data for *ALOX5*, *ALOX5AP* and *LTA4H*; all correlation coefficients were statistically significant ($P \leq 0.05$)

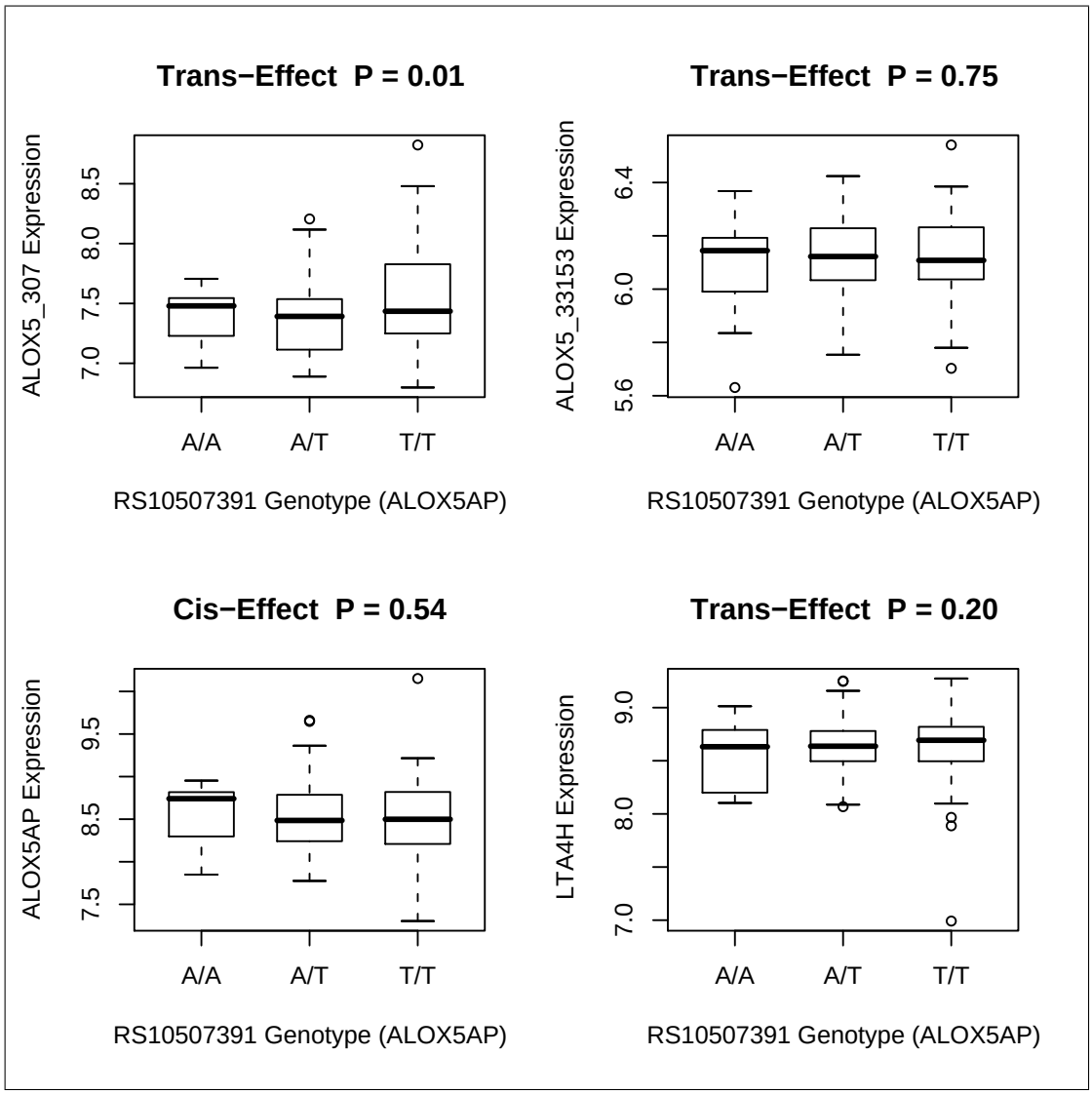

Figure 2.2: Boxplots of rs10507391 genotype-specific expression levels of RMA normalized transcripts in 122 aorta-sections. The x-axis annotates the genotype for rs10507391 and the y-axis annotates the expression levels for *ALOX5* (*trans*-effect), *ALOX5AP* (*cis*-effect) and *LTA4H* (*trans*-effect).

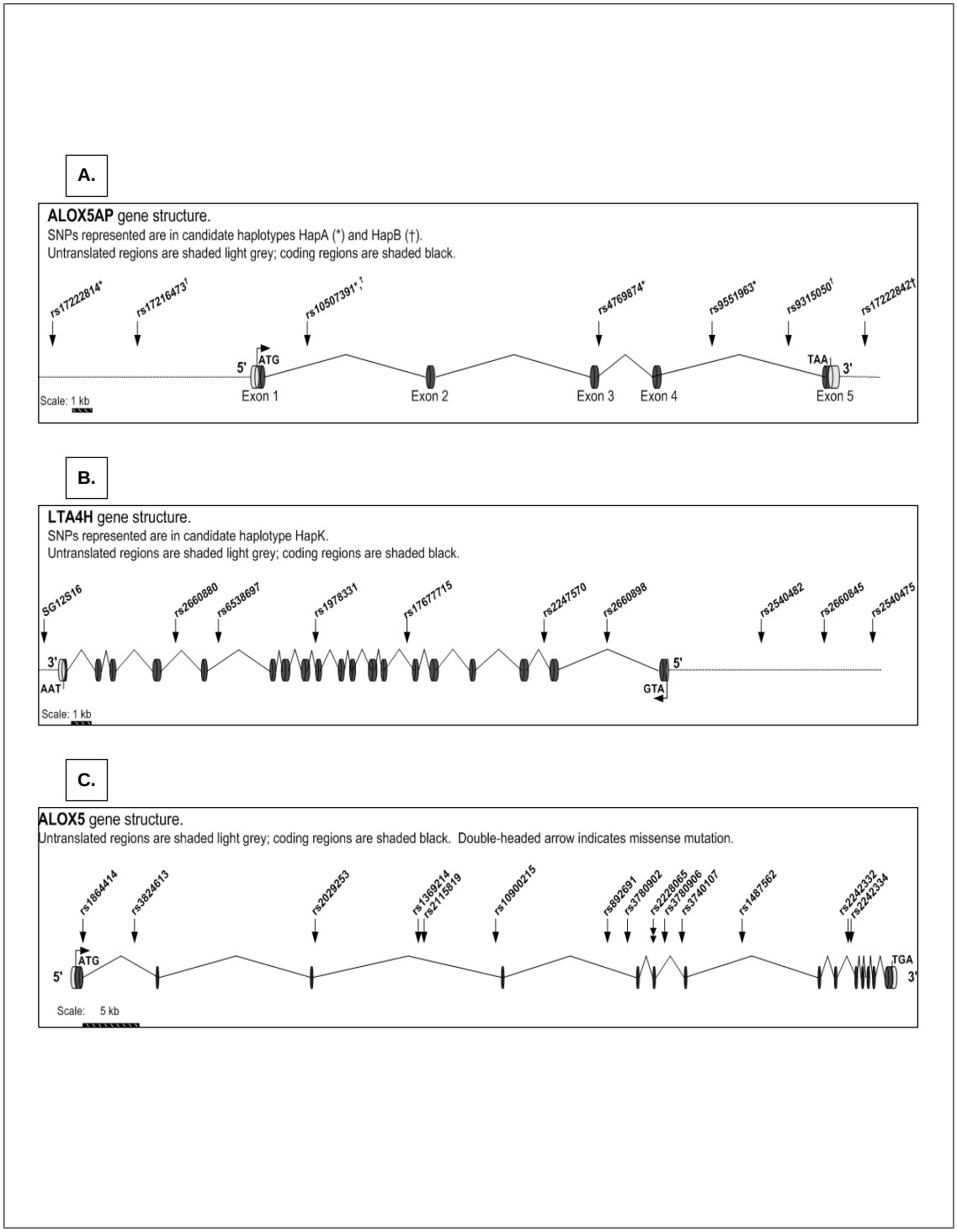

Figure 2.3: Supplemental Figure 1: Illustration of the *ALOX5AP* (A) *LTA4H* (B.) and *ALOX5* (C.) gene structures with the SNPs composing haplotypes A, B and K. Untranslated regions are shaded light grey; coded regions are shaded black; double-headed arrow indicates missense mutation

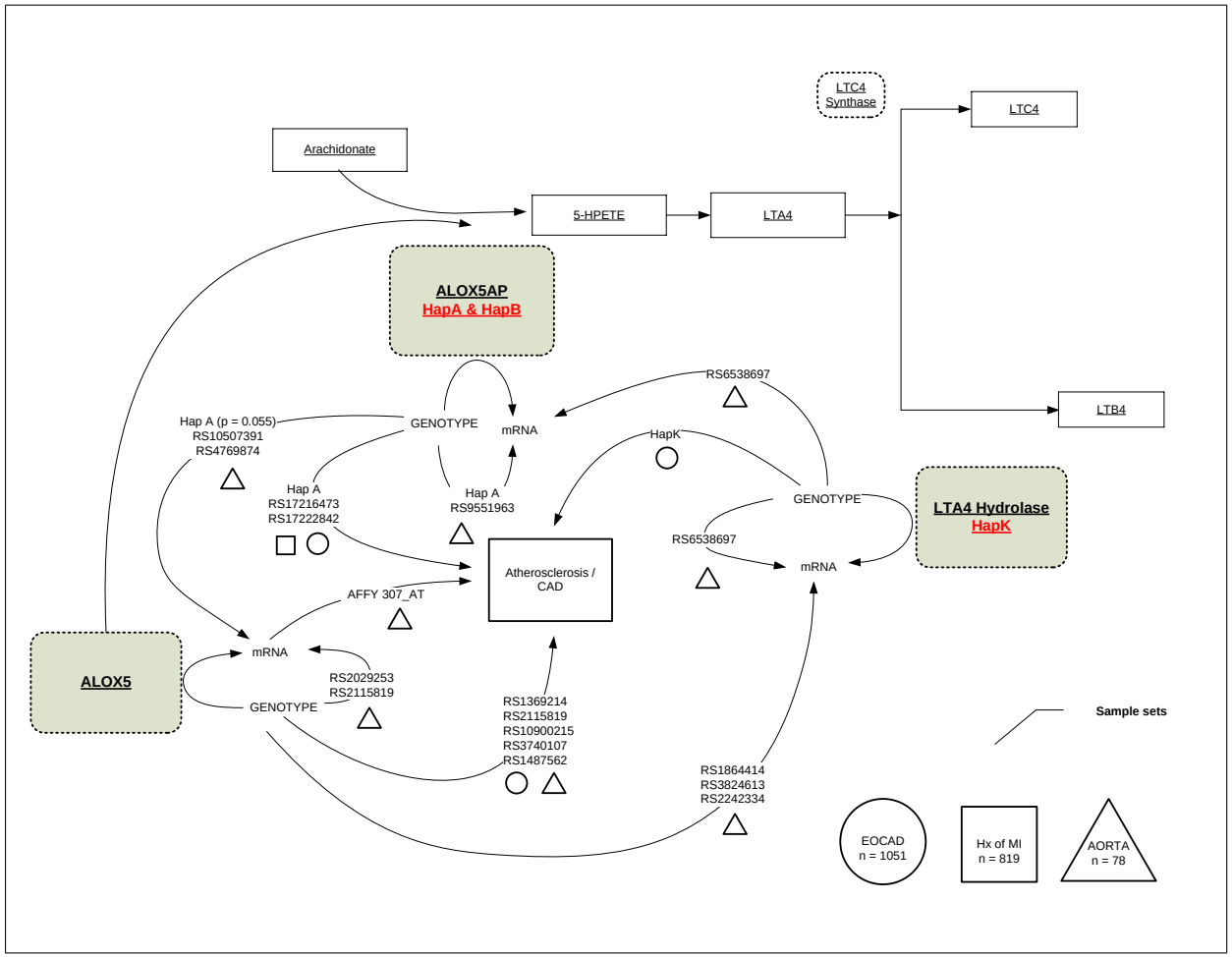

Figure 2.4: Summary of *cis* and *trans* association findings in the leukotriene biosynthesis pathway for the CATHGEN subjects with EOCAD versus unaffected controls (◯), subjects with myocardial infarction (MI) versus unaffected controls (□), and the AORTA casecontrol samples (raised lesion mapping and Sudan IV staining) (△); SNPs and haplotypes identified are significant ($P \leq 0.05$) in statistical models for CAD or transcript expression outcomes.

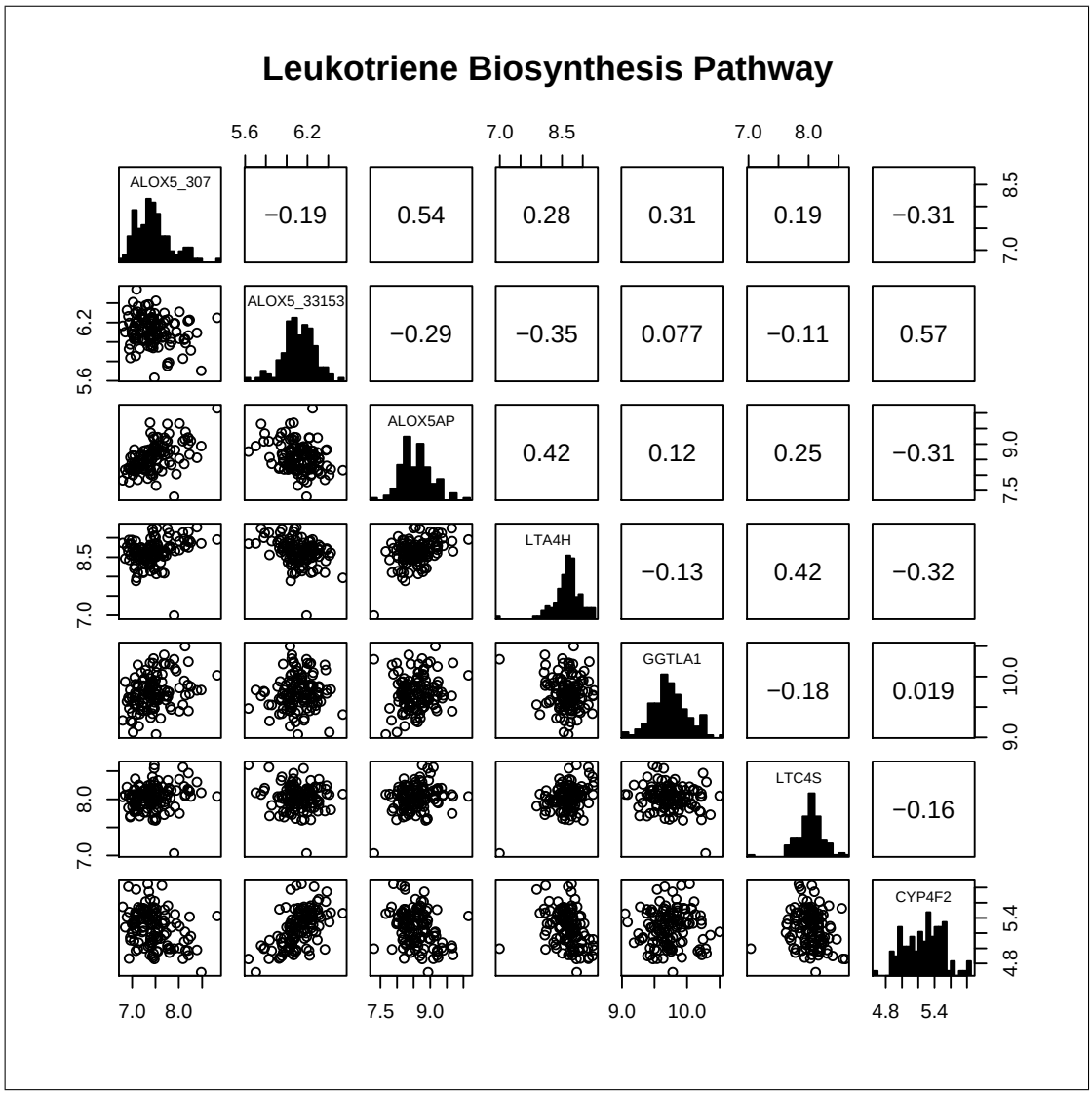

Figure 2.5: Pairwise scatterplots (below the diagonal), histograms (across the diagonal) and pairwise correlation coefficients (above the diagonal) using RMA normalized gene expression data for *ALOX5*, *ALOX5AP*, *LTA4H*, *GGTLA1*, *LTC4S* and *CYP4F2*

Chapter 3

Simulations of case-control data and coding of matrix measures

3.1 Introduction

Complex diseases will have multiple functional sites, and it will be invaluable to understand this LD interaction between those sites in addition to the haplotype-LD effects. An example of this is the leukotriene biosynthesis pathway and association with CVD-phenotypes (Results in Chapter 2). While our individual haplotype results with *ALOX5AP*(HapA) and *LTA4H*(HapK) support a modest role of single-gene effects, the fact that both genes are in the same pathway reveal important genomic interactions for the leukotriene pathway in atherosclerosis pathogenesis. With this prior knowledge, we simulated multiple two-variant disease models with haplotypes to gain an understanding of pathway interactions in terms of correlation patterns. Our first goal was to model a complex disease, with the potential for multiple risk variants which may or may not interact to cause risk, all in the presence of LD. For a complex trait system independent of epistasis, it is useful to consider four framework

parameters [34]:

1. The effect size of a disease locus

   - Relative Risk (RR) = $\frac{p_{exposed}}{p_{non-exposed}}$

   - Odds Ratio (OR) = $\frac{p_{exposed}/(1-p_{exposed})}{p_{non-exposed}/(1-p_{non-exposed})}$

2. The frequency of a disease allele(s)

3. The frequency of marker allele(s)

4. The extent of linkage disequilibrium (LD) between the marker and disease locus

   - Non-random association of alleles at two loci

   - Lewontin's D'

   - Correlation coefficient (r)

After review of the literature where many of the LD-contrast methods were derived (Section 1.2), we discovered there were many approaches to haplotype simulations in the presence of risk effects. Zaykin et al. [41] drew haplotype frequencies from a Dirichlet (1,...,1) distribution, but no specific LD patterns were assigned. Effect sizes were drawn from the Gamma (1) distribution and were inspected to ensure that two large effect sizes were allocated to the most-distinct six-SNP haplotypes - 111111 and 222222 - corresponding to two independent mutations in high LD with two very distinct haplotypes. Nielsen et al. [8] used a different approach based on HapMap data. Groups of three SNPs in near proximity were selected at random from phased Hapmap Japanese and Chinese samples from chromosome 22. For each of

the SNP trios, the locus with the smallest MAF was determined to be the functional polymorphism with higher susceptibility. Others use prior knowledge of haplotype frequencies from real data [3][16]. A variation of penetrance was assigned for each genotype combination.

For the PLS method, Wang et al. [47] simulated interactions in the presence of various levels of correlation between the two gene regions. The disease status of each subject was simulated based on a dominant inheritance model:

$$P(D=1) = \frac{exp[\theta_0 + \theta_1 I_{u1} + \theta_2 I_{u2} + \theta_{12} I_{u1} I_{u2}]}{1 + exp[\theta_0 + \theta_1 I_{u1} + \theta_2 I_{u2} + \theta_{12} I_{u1} I_{u2}]} \tag{3.1}$$

I_{u1} and I_{u2} are indicator variables for the disease alleles at the two causal loci. The intercept term θ_0 is the probability of disease for a group of subjects who do not carry any disease alleles. For Type I error rates, Wang et al. considered two situations: 1. neither u_1 nor u_2 is associated with disease ($\theta_1 = \theta_2 = \theta_{12} = 0$); and 2. only u_2 is associated to disease (($\theta_2 = 2, \theta_1 = \theta_{12} = 0$). To compare power for the different models, three classes of genetic models were considered: 1. a multiplicative model where the joint effect of two factors is the product of the main effects ($\theta_1, \theta_2, \theta_{12} > 0$); 2. an epistatic model where the main effects are simulated to be 0 ($\theta_1 = \theta_2 = 0$) with joint effects $\theta_{12} > 0$; and 3. a cross-over model where u_1 has opposite effects depending on the genotype u_2 ($\theta_1 = -0.5, \theta_{12} > 0$). Finally, they considered two scenarios for input data for the model analyses: 1. the causal variants u_1 and u_2 were not genotyped; and 2. the causal variants were genotyped.

For our research, the software package SIMLA was used to generate case-control genotype and phenotype data. SIMLA is the only publicly available program that

can simulate variable levels of both linkage (recombination) and LD between marker and disease loci in general pedigrees [50]. Equally important, SIMLA allows for the joint action of up to two genes in the simulated data, with all possible multiplicative interaction effects between them. The LD and interaction features of SIMLA provide functionality we did not observe in other simulations. These two features will also allow us to assess interaction between multiple disease-causing variants by means of their linkage disequilibrium (LD) patterns which is our hypothesis for the leukotriene biosynthesis pathway. We are unaware of any off-the-shelf software that provides such functionality.

Next, we investigated the statistical properties of a class of matrix-based statistics (Section 1.2) to assess epistasis generated by these simulations. Our goal was to detect an interaction between multiple disease-causing variants by means of their linkage disequilibrium (LD) patterns with other haplotype markers. We considered three statistics based on contrasting pairwise matrices of LD patterns between cases and controls. The statistics include Zaykin's LD permutation-based test (Z_1), the LD contrast test (Z_2) and Wang's LD contrast test that controls for background LD. In addition, we considered Wang's interaction test using partial least squares (PLS). The two primary analysis packages used to implement these tests were SAS 9.1.3 (IML) and R for Statistical Computing Software version 2.7.0.

3.2 Sample simulations

SIMLA can simulate two bi-allelic disease loci spread over three chromosomes, with the third chromosome providing the option to analyze markers completely unlinked

to disease loci. Blocks of LD can be simulated by selecting a subset of markers (SNPs) to be in LD with a disease locus, while other markers are in linkage equilibrium with the disease locus [50]. The disease risk model we simulated included two disease loci, each of which was included in distinct three-SNP haplotypes. Figure 3.1 illustrates a two-variant (V) disease model in one disease gene. Each variant is in LD with two markers forming a haplotype. There is also interaction between the variants in terms of RR. This can also be illustrated in the context of pathway interaction (Figure 3.2). Each variant is hypothetically located in separate disease genes with equal marginal effects (RR) and an interaction effect. We varied the values of these parameters for the simulation as described in Table 3.1.

3.2.1 Disease model

Each disease locus was generated with varying LD with genotypes assigned according to the specific LD pattern (see Table 3.1). Because in real data analysis the haplotype phase cannot be directly observed, the composite disequilibrium $\widehat{D}_{AB}$ and correlation $\widehat{r}$ were calculated using dilocus counts and sample allele frequencies [5][4][13][41]. Zaykin et al. [41] emphasizes that with the EM algorithm, the likelihood is constructed assuming HWE on the level of haplotypes that cannot be inferred from single SNP HWE assessment. The composite disequilibrium approach provides results similar to those of the EM-based method under HWE, is computationally simpler, and avoids the assumption of haplotypic HWE [41]:

$$\widehat{D}_{AB} = \frac{1}{n}(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}) - 2\tilde{p}_A\tilde{p}_B \tag{3.2}$$

In this equation $\tilde{p}_A$ and $\tilde{p}_B$ are sample allele frequencies and n_{AABB} is the dilocus

sample genotype count. The composite correlation is given as

$$\widehat{r} = \frac{D_{AB}}{\sqrt{(Pr_A \cdot Pr_a + D_A)(Pr_B \cdot Pr_b + D_B)}} \tag{3.3}$$

where $D_A = P_{AA} - p_A^2$ is the Hardy-Weinberg disequilibrium coefficient.

Pairwise composite correlation $r_{composite}$ matrices were generated for input data. In the statistical analyses, we assumed that the disease variants V_1 and V_2 were not genotyped; however, the genotypes were generated to measure the interaction in terms of $r_{composite}$ and to validate the parameter settings for SIMLA.

Null models

Null models were used to investigate the statistical properties of the tests under the null hypothesis of no interaction. The central null condition is the absence of any interaction ($Interaction_{RR} = 1$) between the two variants with and without marginal effects. There is also a null model in the presence of LD, but no marginal effects. The null models can be summarized into three categories (Table 3.1): 1. Null models with no marginal effects $V_{1RR} = 1; V_{2RR} = 1$ and no LD (model 1); 2. Null models with marginal effects $V_{1RR} = V_{2RR} > 1$ and no LD (models 2 & 3); and 3. Null models in the presence of LD and no marginal effects (model 4). As shown in Table 3.1, all null models have interaction RR set to 1.0. We also considered the effect of varying allele frequencies. For the several null models there are genetic effects to detect, but there is no interaction between the SNPs.

Alternative models

Alternative models were used to evaluate power of the matrix measures and to identify interaction effects in complex disease models. The alternative models can be summarized into three categories (Table 3.1): 1. Alternative models in the presence of LD, marginal effects and no interaction (models 5 & 6); 2. Alternative models in the presence of LD, interaction effects and no marginal effects (model 7); and 3. Alternative models in the presence of LD, marginal effects and interaction (models 8 & 9). For models 7, 8 and 9, we generated a RR interaction of 3.0 and 10.0 (the limit for SIMLA).

LD and effect simulations

The *haplotypeselect* feature in SIMLA was used to simulate a three locus disequilibrium system with two markers and one disease locus. Given an already assigned V_i (susceptibility) or v_i (non-susceptibility) allele at the disease variant locus, a marker haplotype (set of alleles) for all individual founder chromosomes were randomly generated based on the conditional haplotype probabilities. Marker M_1 is in LD with disease variant V, and M_1 is in LD with M_2 (see figure 3.1). Given SIMLA's algorithm for generating haplotypes, LD cannot be controlled between M_2 and the disease variant V variant (see figure 3.1). We set $p(M_{1i}, M_{2i}, V_i)$ for alleles i and j to maximize the three locus r^2.

$$r^2(M_{1i}, M_{2j}, Vi) = \frac{D^2(p(M_{1i}, M_{2i}, V_i)}{var(D(p(M_{1i}, M_{2j}, V_i)))} \tag{3.4}$$

where

$$D(M_{1i}, M_{2i}, V_i) = p(M_{1i}, M_{2i}, V_i) - p(M_{2i})D(M_{1i}, V_i) - \\ p(V_i)D(M_{1i}, M_{2i}) - p(M_{1i})D(M_{2i}, V_i) - p(M_{1i})p(M_{2i})p(V_i) \quad (3.5)$$

then

$$p(M_{1j}M_{2i}V_j) = p(M_{2i}V_j) - p(M_{1i}M_{2i}) + p(M_{1i}M_{2i}V_i), \quad (3.6)$$

$$p(M_{1i}M_{2i}V_j) = p(M_{1i}M_{2i}) - p(M_{1i}M_{2i}V_i), \quad (3.7)$$

$$p(M_{1j}M_{2i}V_i) = p(M_{2i}V_i) - p(M_{1i}M_{2i}V_i); \quad (3.8)$$

Interaction was simulated in terms of the RR in addition to marginal effect RR. The actual susceptibility variants (V_1 and V_2) were generated, but were not used in the analysis; LD was measured between the disease variants to validate the simulation parameters. Chromosome-specific genetic maps with inter-marker distances were specified in Morgans (M). While maintaining an interaction RR of 1.0 between variant 1 (V_1) and variant 2 (V_2), we simulated minor allele frequencies (MAF) of $0.05, 0.15, 0.35$ and 0.50 for the variants and the haplotype markers associated with each (see Table 3.1). Allele frequencies for V_1,V_2 and all haplotype markers were kept equivalent. LD between the haplotype-specific markers was generated at $r^2 = 0.4, 0.7$ and 0.9 with equal values for both haplotypes. We then considered the effect of increasing RR for the two variants. We generated RR of $1.0, 1.5$ and 3.0 effect sizes for our models. We generated a prevalence frequency of 0.10 for all models. Finally, we generated a RR interaction between both variants of 3.0 and 10.0 (the limit for SIMLA).

For each parameter set, we generated samples of 500 cases and 500 controls, and statistics were calculated and stored. We simulated 1000 replicates for each model

listed in Table 3.1. Statistic values and achieved significance levels (P-values) were collected. We compared mean values and standard deviations to assess the stability of the measures. Type I error rates and empirical power were estimated using the proportion of achieved significance levels <= 0.05 based on 1000 simulations of the null and alternative models respectively.

SIMLA can simulate four types of disease models for both variants (V_1 & V_2) independently. We used the multiplicative model where risk of developing disease is increased by a factor for each allele carried. Dominant, recessive and additive disease models can be generated and will be explored in the future.

3.3 LD contrast test statistics

3.3.1 LD-contrast and featurevector permutation-based tests

As applicable to a square matrix, composite-LD matrices have spectral decompositions which can be reduced to canonical form, represented by eigenvalues and eigenvectors. The Z_1 statistic measures the difference between two spaces defined by the first k eigenvectors with the sum of squared cosines of the angles (θ) between the eigenvectors. The matrix formed by the first k eigenvectors is a featurevector. This statistic and the subsequent permutation-based test was originally described by Krzanowski [41][53][54]:

$$\sum_{i=1}^{k} \lambda_i = traceS = traceTT' = \sum_{i=1}^{k} \sum_{j=1}^{k} cos^2\theta_{ij} \tag{3.9}$$

In this equation the value $T = E_{case}E'_{control}$. The values i and j represent the principal components of two separate groups. Krzanowski described this sum as a measure of

similarity between two spaces. The value can lie between k (coincident spaces) and 0 (orthogonal spaces)[53]. This can be interpreted as a geometric correlation. The first column k eigenvectors are defined by E_{case} and $E_{control}$. Because our matrix consists of four SNPs, we generated featurevectors consisting of one, two and three vectors. The tests were implemented in SAS/IML. The sum-of-squared-differences statistic (Z_2) measures the overall difference in the corresponding pairwise LD where r is the correlation matrix:

$$Z_2 = trace[(\hat{r}_{case} - \hat{r}_{control})^T(\hat{r}_{case} - \hat{r}_{control})] \tag{3.10}$$

3.3.2 LD contrast controlling for subject-specific background LD

We next considered Wang's LD contrast test that controls for background LD. We used the linear mixed-effects model (lme) function in R to estimate the best linear unbiased predictor (BLUP) (Dr. Tao Wang personal communication). Instead of using the overall sample mean to center the genotypes which is equivalent to the composite correlation-based LD contrast test statistic, Wang et al.[48] proposed centering the genotype by use of the individual (i) specific means which absorb background LD (equation 3.11). When calculating the score statistic U in equation 3.11, the individual genotype (X_{Ai}) for marker A is corrected for background LD by subtracting the individual specific mean for that marker $\hat{E}(X_{Ai})$. This individual specific mean takes both types of information into account, information across subjects and information across markers[48] . To compare cases and controls, Wang et al. [48] defined the matrices Λ_{case} and $\Lambda_{control}$ in equation 3.12 as the corresponding sum of pairwise mean-corrected cross-products by using the BLUPs of the means fitted using the

lme function in R statistical computing software. We will refer to this statistic as $T_{BACKGROUND-CORRECTED}$. The tests were implemented in R.

$$U = \sum_i [X_{Ai} - \hat{E}(X_{Ai})][X_{Bi} - \hat{E}(X_{Bi})] \tag{3.11}$$

$$T_{BACKGROUND-CORRECTED} = trace[(\Lambda_{case} - \Lambda_{control})^T(\Lambda_{case} - \Lambda_{control})] \tag{3.12}$$

3.3.3 Partial Least Squares

Finally we considered the PLS approach presented by Wang et al. for modeling gene-gene interactions with multiple markers [47]. Partial least squares accounts for spectral decomposition of response values (case-control status) in addition to the independent correlation variables. Wang et al. suggest detecting an association by a likelihood ratio test based on a logistic regression model:

$$logit(Pr(D)) = \beta_0 + \sum_{i=1}^{k_1} \beta_{1i} s_{1i} + \beta_{g1g2} T_1^{1PLS} T_2^{1PLS} \tag{3.13}$$

or

$$logit(Pr(D)) = \beta_0 + \sum_{j=1}^{k_2} \beta_{2j} s_{2j} + \beta_{g1g2} T_1^{1PLS} T_2^{1PLS} \tag{3.14}$$

In this equation k = the number of SNPs for a given gene 1 or 2. $\beta_{g1g2} T_1^{1PLS} T_2^{1PLS}$ = the interaction term between the first factor for each gene 1 or 2. We reported both the likelihood ratio-based results testing the global null hypothesis ($\beta = 0$) as well as the single effect significance for the haplotype interaction term (probability of observing a Wald $\chi^2 \geq$ to observed).

3.3.4 Coding and data management

It quickly became apparent that computational tools to managing the data output and summarizing the statistical measures were needed, especially with the number of null and alternative models simulated. Perl was the logical choice to not only manage the software, but store and summarize the data. Sample code is referenced in Appendix A.1. In this sample, we demonstrate how Perl manages the system following this general set of progression: 1. Update the SIMLA control file for a given set of parameters; 2. Execute SIMLA; 3. Format the SIMLA genotype output to be centered (-1(AA),0(Aa),1(aa)) for LD measurements and matrix measures; 5. Execute SAS or R package to generate matrix statistic; 6. Parse results and store measures for Type I and power estimates; 7. Loop for 1000 replicates; and 8. Loop to next model and repeat.

Sample SAS/IML code is found in Appendix A.2. In this example we demonstrate how the composite D' and r^2 are calculated for creating the LD matrix. Also demonstrated is the Z_1 permutation test that uses SAS/IML matrix functions for spectral decomposition. The Z_1 weighted *eigenvector* by the *eigenvalue* method is shown. The Z_2 LD contrast test was also calculated using SAS/IML. In appendix A.3, we demonstrate R coding for the linear mixed-effects model (*lme*) function to estimate the best linear unbiased predictor (BLUP). Because R is extremely slow with loops, we used the reshape function to stack the SNPs (long format) for linear mixed effects (*lme*) input. All plots in Chapter 4 were created using R Statistical Computing.

Table 3.1: Descriptive summary of the nine study designs for simulation using the package SIMLA. Each model is simulated with MAF = 0.05, 0.15, 0.35 and 0.50 at a prevalence of 0.10.

Model	$LD(r^2)$ V_1M_1	$LD(r^2)$ V_1M_2	$LD(r^2)$ V_2M_1	$LD(r^2)$ V_2M_2	RR V_1	RR V_2	RR $V_1 \times V_2$	MAF
Null models with no LD, marginal or interaction effects								
1	0	0	0	0	1.0	1.0	1.0	0.05, 0.15 0.35, 0.50
Null models with marginal effects, but no LD or interaction effects								
2	0	0	0	0	1.5	1.5	1.0	0.05, 0.15 0.35, 0.50
3	0	0	0	0	3.0	3.0	1.0	0.05, 0.15 0.35, 0.50
Null models in the presence of LD, but no marginal or interaction effects								
4	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	1.0	1.0	1.0	0.05, 0.15 0.35, 0.50
Alternative models in the presence of LD, marginal effects and no interaction								
5	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	1.5	1.5	1.0	0.05, 0.15 0.35, 0.50
6	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	3.0	3.0	1.0	0.05, 0, 15 0.35, 0.50
Alternative models in the presence of LD, interaction and no marginal effects								
7	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	1.0	1.0	3.0 10.0	0.05, 0.15 0.35, 0.50
Alternative models in the presence of LD, marginal effects and interaction								
8	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	1.5	1.5	3.0 10.0	0.05, 0.15 0.35, 0.50
9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	0.4 0.7 0.9	3.0	3.0	3.0 10.0	0.05, 0.15 0.35, 0.50

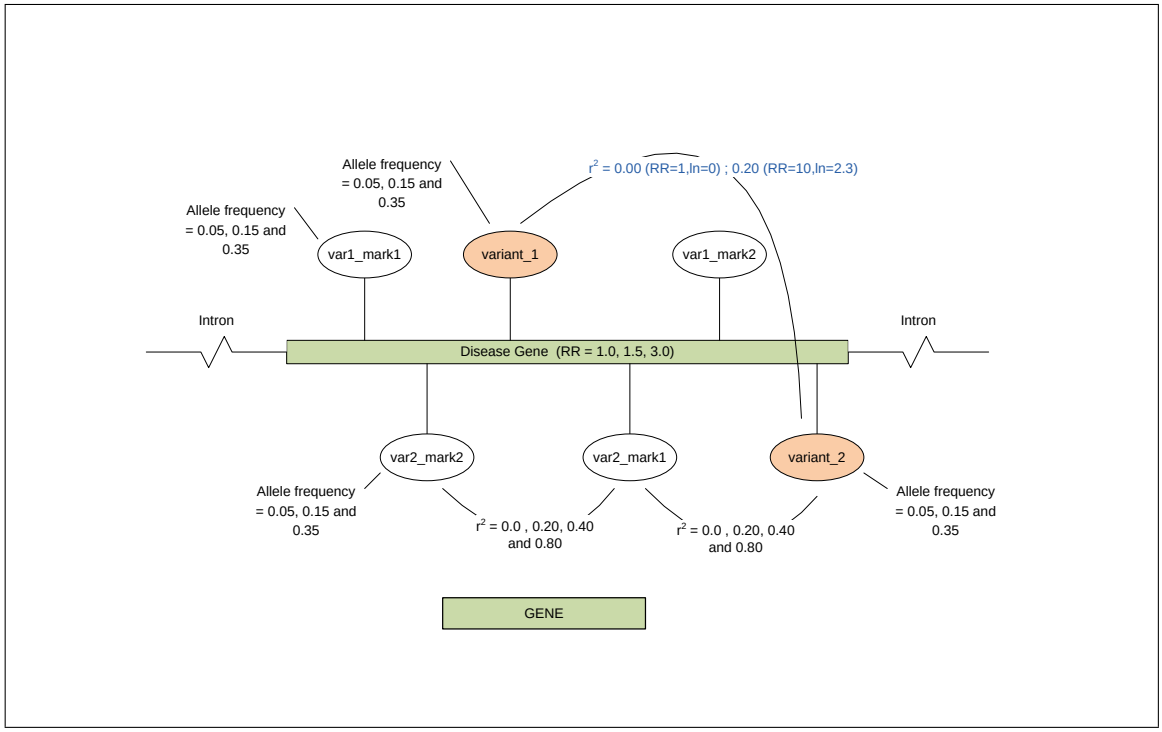

Figure 3.1: Gene annotation for interaction simulations using the SIMLA package.

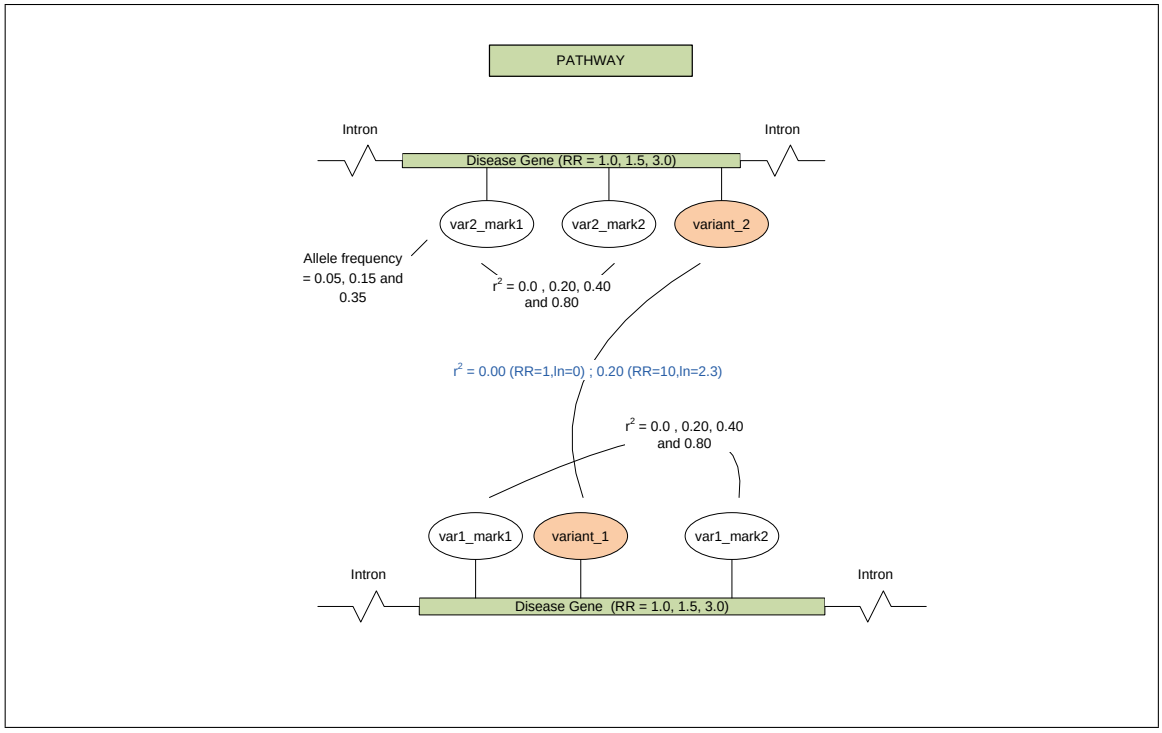

Figure 3.2: Pathway annotation for interaction simulations using the SIMLA package.

Chapter 4

Results of simulations

4.1 Introduction

In this chapter we evaluate the realized correlation for complex models as described in Chapter 3. While the models we simulated seem quite extreme (i.e. interaction $RR = 10.0$), the purpose was to observe a trend in correlation patterns based on different combinations of model parameters including joint RR. The rationale behind this approach is that the joint effect of two variants would generate different LD patterns in cases and controls [47]. The effect of MAF and RR are strong influences on correlation, but we also observed the influence of selection bias based on marginal effects in the presence of interaction. Besides altering MAF and penetrance parameters, we observed static simulation systems proposed by others [3][41][8][47][39].

Zaykin et al. [41] simulated haplotype frequencies with risk effect sizes sampled from a Gamma distribution, but LD between markers was not described in detail. Wang et al. [47] simulate two-SNP haplotype in LD with an untyped causal variant. First, four haplotype frequencies were determined by the allele frequencies and LD,

and a pair of haplotypes were randomly sampled from the corresponding multinomial distribution for each subject. The disease status was defined by a similar model used by Zaykin et al. [41]. To avoid the influence of allele frequency, both frequencies were simulated at 0.5, unlike our models ranging from 0.05 to 0.50. Our simulations better capture the range of allele frequencies that were observed in the leukotriene pathway example. Zhao et al. [39] introduce interaction between two unlinked loci. They propose to define interaction between two genes as the deviance of the penetrance for a haplotype at two loci from the product of the marginal penetrance of the individual alleles that span the haplotype. Interaction between two unlinked loci will result in deviation of the penetrance of the two-locus haplotype from independence of the marginal penetrance of the alleles at an individual locus, which in turn will create LD even if two loci are not on the same chromosome.

SIMLA allows the user to assign interaction RR independent of marginal effects, thus allowing a pure epistatic pure model. Finally, Wang et al. [47] simulated similar two-variant disease models as we proposed, but LD was not specifically assigned. By no means did we exhaust the model space, but we did observe trends in correlation patterns between the two disease variants V_1 and V_2. The results for realized correlation patterns are presented in Section 4.2.

We then summarize the matrix measures results for testing interaction in these complex models. Null models were simulated to evaluate Type I error rates. Based on a nominal significance level of $\alpha = 0.05$, we calculated the proportion of rejections to evaluate whether the size/level of a test achieves the advertised α when the null hypothesis is true [35]. To evaluate power, we generated alternative models. We

calculated the proportion of rejections based on the nominal significance level of $\alpha = 0.05$ when in fact the alternative hypothesis is true. The results for each matrix-contrast method are presented in Section 4.3.

4.2 Composite correlation disease model

Tables 4.1, 4.2 and 4.3 contain the composite correlation between variants V_1 and V_2 for both cases and controls that we simulated for the two disease locus model. While the variants were not included in the correlation matrices assessed, we wanted to understand what effects the disease effect size (relative risk RR), haplotype LD, disease prevalence and MAF have on cross-locus interaction (epistasis) that we specified though a RR of $1.0, 3.0$ and 10.0. The table is divided into three sections for marginal disease RR of $1.0, 1.5$ and 3.0 for each variant. Within each section there are four columns representing MAF of $0.5, 0.15, 0.35$ and 0.50 for both markers and variants. And finally each row represents LD in terms of correlation for the haplotypes with values $0.0, 0.4, 0.7$ and 0.9. Table 4.1 contains the correlations for the null models. These models can be summarized into three categories (Table 3.1): 1. Null models with no marginal effects and no LD (model 1); 2. Null models with marginal effects and no LD (model 2 & 3); and 3. Null models in the presence of LD and no marginal effects (model 4). Table 4.1 also contain the correlations for alternative models 5 and 6 (Table 3.1). These alternative models are in the presence of LD, marginal effects and no interaction.

Tables 4.2 and 4.3 contain the composite correlations for the interactive RR of 3.0 and 10.0 respectively. Because these are alternative models with interactive $RR > 1$,

we did not consider $r^2 = 0.0$ for the haplotypes. These alternative models can be summarized into two categories (Table 3.1): 1. Alternative models in the presence of LD, marginal effects and no interaction; and 2. Alternative models in the presence of LD, marginal effects and interaction.

Figure 4.1 illustrates the MAF ($x-axis$) by case composite correlations ($y-axis$) for the interaction RR of $1.0, 3.0$ and 10.0 represented as a solid, dashed and dotted line respectively. Each trellis window represents a different combination of marginal disease RR of $1.0, 1.5$ and 3.0 (moving up the trellis), and LD in terms of r^2 for each haplotype (moving across the trellis) of $0.4, 0.7$ and 0.9. At an interaction RR of 1.0 (solid line), we detected a composite correlation close to 0.0 for almost every combination of marginal relative risk and haplotype LD as evident throughout the trellis plot in Figure 4.1. However, there is a slight dip below 0.0 in correlation that is illustrated in the top row as a marginal RR of 3.0 as MAF increases from 0.05 to 0.5. This is most likely due to the case selection bias with a marginal RR of 3.0. At a marginal $RR \geq 1.0$, a single disease variant is sufficient to cause disease and thus to be a case.

For an interaction RR of 3.0 (dashed line) in 4.1, we observed different interaction effects (r^2) based on different combinations of parameters (MAF, LD and RR). As we increase the marginal RR from 1 to 1.5 (bottom and middle row of trellis), there are similar effects on the interaction correlation with the maximum of approximately 0.10 at MAF 0.35 and 0.5. At RR of 3.0, there is less correlation (approximately 0.05) because of the case selection bias as described above.

For an interaction RR of 10.0 (dotted line) in Figure 4.1, we also observed sim-

ilar effects as we increased the marginal RR from 1.0 to 1.5 (bottom and middle row of trellis). This is also where we observed the maximum composite correlation of approximately 0.20 at MAF = 0.35. In all of the trellis windows (every combination of LD and marginal RR) we observed a decrease in interaction correlation as we increased MAF from 0.35 to 0.5. This is due to the fact that we have equal representation of each allele in the population and the chance that a case subject is selected based on the presence of variants at either V_1 or V_2 is greater in addition to the joint effects (interaction) disease. Cases could have disease due to variant 1, variant 2 or a combination.

4.3 Matrix measures results

4.3.1 Z_1 featurevector permutation-based test

Table 4.4 contains the Type I error rates for the Z_1 permutation-based test using the null models with no LD (models 1, 2 & 3 in Table 3.1). With no apparent differences between cases and controls, the Type I error rate remained close to the nominal significance level of 0.05. Table 4.5 contains the Type I error rates for the Z_1 permutation-based test using the null models in the presence of LD and no marginal or interaction effects (models 4 in Table 3.1). At a marginal RR of 1.0, the Type I error remained close to the nominal significance level of 0.05.

Figure 4.2 illustrates empirical power for the Z_1 permutation-based test where LD is present and no interaction RR, but the marginal $RR > 1.0$ (models 5 and 6 in Table 3.1). This model is testing the correlation differences between cases and controls due to the correlation (LD) between haplotype markers and the disease variant not

included in the measure. For the four-snp matrix, we considered featurevector of $k = 1, 2$ and 3 eigenvalues, but we only present $k = 1$ and 2. Figure 4.3 illustrates $k = 1$ or one eigenvector to represent the featurevector. The respective correlation measured can be referenced in Figure 4.1. Like the interaction LD plot (Figure 4.1), each window in the trellis plot is represented by a marginal RR (moving up the trellis) and haplotype LD (moving across the trellis) in terms of r^2. At higher LD and MAF, we observed a slight increase in power (0.11), but not statistically significant.

Figure 4.3 illustrates empirical power for the Z_1 permutation-based test where interaction $RR > 1.0$ (models 7, 8 and 9 in Table 3.1). The interaction RR of 3.0 and 10.0 are represented by a solid and dashed line respectively. For the interaction RR of 3.0 (solid line) in Figure 4.3, we observed an increase in power to detect interaction between the variants. Zaykin et al. [41] indicated that both the featurevector-permutation (Z_1) and LD-contrast (Z_2) have limitations when allele frequencies are low. Lower MAF can give rise to spurious correlation structures. Our results illustrated in Figure 4.3 at MAF = 0.05, and for the most part 0.15, support those findings. At low haplotype LD ($r^2 = 0.4$; left column of trellis), the power to detect an interaction is low (< 0.20). As haplotype LD is increased (LD = 0.9, $RR = 3.0$) in Figure 4.3, the power to detect an interaction at MAF = 0.5 is 0.70, which is the maximum power across all simulation conditions for $RR = 3.0$.

At an interaction RR of 10.0 (dashed line) in Figure 4.3, we observed an increase in power throughout all combinations of parameter settings. As MAF increases, we observed an increase in power. As illustrated in Figure 4.3, there is an increase in

power as LD is increased. The marginal RR has less of an effect on power. However, the power for the Z_1 permutation-based test to detect an interaction between V_1 and V_2 increases as both parameters are increased. At higher LD = 0.7 and 0.9, we observe power > 0.80 for MAF 0.35 and 0.5 when marginal $RR > 1.0$. The empirical power plots for the Z_1 permutation-based test mimic the interaction LD patterns illustrated in Figure 4.1 suggesting that this measure is sensitive to the extent of correlation.

Figure 4.4 illustrates $k = 2$ to represent the featurevector. There is a total loss of power to detect interaction between the variants V_1 and V_2 when $k = 2$. As Zaykin et al. indicated, Krzanowski suggested using the value of k that is the largest integer smaller than $L/2$ when L is the total number of markers. Values $k \geq L/2$ will cause the subspaces defined by the two sets of subspaces to intersect in at least one dimension[41][53]. Our results in Figure 4.4 support this fact. We also weighted the eigenvectors by their corresponding eigenvalues (Figures 4.5 and 4.6) to assess a benefit of creating a featurevector $k = 2$ by the level of significance. This indicates that the second eigenvector contributes minimal information with respect to the first, which is likely for a matrix containing four variables. The empirical power where LD is present and no interaction RR, but the marginal $RR > 1.0$ (models 5 and 6 3.1) is different than the unweighted $k = 1$. The weighted $k = 2$ Z_1 power approaches 0.10 for $RR = 1.5$. For $RR = 3.0$, the power increases as MAF increases, but there is a decrease of power as LD increases (moving across the trellis). Further examination of the correlation structure simulated by the haplotype selection feature in SIMLA using these parameters may explain this. For the the interaction $RR > 1.0$ (models

7, 8 & 9 in Table 3.1), there is improved power compared to the $k = 1$ approach. Adding an additional eigenvector and weighting by significance (eigenvalue) increased power under this disease model.

4.3.2 Z_2 LD contrast test

Table 4.4 contains the Type I error rates for the Z_2 permutation-based test using the null models with no LD (models 1, 2 & 3 in Table 3.1). The Type I error remained close to the nominal significance level of 0.05. Table 4.5 contains the Type I error rates for the Z_1 permutation-based test using the null models in the presence of LD and no marginal or interaction effects (model 4 in Table 3.1). Like the Z_1, the Type I error remains close to the nominal significance level of 0.05.

Figure 4.7 illustrates empirical power for the Z_2 permutation-based test where LD is present and no interaction RR, but the marginal $RR > 1.0$ (models 5 and 6 in Table 3.1). At LD 0.4 and MAF of 0.5., we observed an increase in power approaching 0.23. All other combination of parameters had very low power (0.05).

Figure 4.8 illustrates empirical power for the Z_2 LD contrast test where interaction $RR > 1.0$ (models 7,8 and 9 in Table 3.1). For an interaction RR of 3.0 (solid line), we observed mixed results. The empirical power does not exceed 0.40 for any combination of marginal RR (moving up the trellis) and haplotype LD (moving across the trellis) except for the top left panel (LD = 0.4, $RR = 3.0$). At MAF = 0.35 and 0.5, the observed power is approximately 0.5 and 0.5. In the presence of low LD (r^2), high marginal $RR(3.0)$ and high MAF (> 0.35) this test seems to perform well. While less encouraged by these results compared to Z_1, we must note that we

were measuring the case-control correlation (r) difference of approximately $0.1 - 0.2$.

For an interaction RR of 10.0 (dotted line), we observed higher power. As Zaykin et al. [41] suggest, the LD contrast test is optimal at higher MAF which we observed. At higher LD (middle and right column) of $r^2 = 0.7$ and 0.9, the power results strongly mimic the interaction correlation pattern we observed in Figure 4.1. Interestingly, at this level of LD, we observed diminishing effectiveness as the marginal RR of 1.0 (power is approximately 0.8) is increased to 3.0 (power is approximately 0.6). At a lower level of LD ($r^2 = 0.4$), we observe the opposite effect (power increases from ~ 0.6 to ~ 1.0). The higher marginal effects in the presence of higher haplotype LD masked the interaction correlation. The probability that a case is selected because of the marginal effects explains this.

4.3.3 Partial Least Squares approach

Table 4.4 contains the Type I error rates for the PLS approach using the null models with no LD (models 1, 2 & 3 in Table 3.1). We reported the P-value for the likelihood ratio test based on a logistic regression model including marginal and interaction effects as suggested by Wang et al. [47]. The likelihood ratio test is based on a regression model to test $H_0 : \beta_{g1g2}$ and $\beta_{1i} = 0$ where $i = 1$ to the total number of SNPs for a given gene/haplotype. We also reported the single effects of the interaction term ($\beta_{g1g2} T_1^{1_{PLS}} T_2^{1_{PLS}}$) based on the logistic regression model in equation 3.13 (Figure 4.12). Table 4.5 contains the Type I error rates for the PLS likelihood test and interaction term using the null models in the presence of LD but no RR (models 4 Table 3.1). The Type I error rates remained close to the nominal significance level

of 0.05 for both measures.

Figure 4.9 and 4.11 illustrate empirical power for the PLS likelihood ratio test and interaction term where LD is present and no interaction RR, but the marginal $RR >$ 1.0 (models 5 and 6 in Table 3.1). The PLS likelihood ratio test had greater power for both marginal $RR = 1.5$ (approaching 0.40) and 3.0 (approaching > 0.90). There is a slight increase in power as LD increases (moving across trellis). As expected, the PLS interaction term had low power (0.05).

Figures 4.10 and 4.12 illustrate the empirical power for the PLS likelihood ratio test and interaction term where interaction $RR > 1.0$ (models 7, 8 and 9 in Table 3.1). Overall, there was an increase in power as the MAF and marginal RR (moving up the trellis) was increased. We only observed a slight increase in power as the LD was increased (moving across the trellis).

While powerful, this test may not be appropriate to separate marginal ($RR = 3.0$) from interaction effects which is illustrated in the top row of Figure 4.10. For the likelihood ratio test, at least one β term must be significant. A significant result could be reporting on the marginal effect β's. Reporting on the interaction term β helped control the strong influence of the marginal RR which is illustrated in Figure 4.12. For an interaction RR of 3.0 (solid line) and 10.0 (dashed line), we observed an increase in power as LD is increased (moving across the trellis). Because PLS places an emphasis on predicting the responses may explain this observation. More SNPs in LD may provide a more significant interaction via PLS. Our results in Figure 4.12 (bottom and middle row of trellis) with power approaching 0.80 at LD = 0.7 and 0.9 support this.

4.3.4 Background corrected LD contrast test

Figure 4.13 illustrates the empirical power for the LD-contrast test while taking the background LD into account. Unfortunately, we were unable to calculate an achieved significance level (P) using case-control label permutation because of computational resources required to run the permutations in R Statistical Computing. Instead of estimating power with the proportion of ($P \leq 0.05$) based on 1000 simulations, we compared distributions of the background-corrected LD contrast statistic for the null (models 4 Table 3.1) versus the respective alternative models. When estimating power (proportion of the alternative distribution $> 95^{th}$ *percentile* of the null distribution), the null model distribution with marginal $RR = 1.0$ and interaction $RR = 1.0$ was matched to the equivalent LD and MAF alternative model distributions.

As evident in the trellis plot (Figure 4.13), the results are inconclusive. This method breaks down at higher RR (top row of trellis; $RR = 3.0$) and higher LD (middle and right column) of 0.7 and 0.9. This is perhaps due to the correlation (LD) between haplotype markers and the disease variant for these null models. The actual values of the test statistics are extremely variable and get larger as parameters are increased. This causes great instability in the variances. The test statistic needs to be reformulated to be stable. We do note that Wang et al. presented this method in the context of analyzing two SNPs and only suggested expanding to the LD-contrast test presented by Zaykin et al.[41] when comparing multiple SNPs.

4.4 Conclusion

Zaykin et al. [41] suggested that the LD contrast test (Z_2) had the highest power based on the models they simulated. The permutation-based method (Z_1) had similar or better power that the Z_2 based on the alternative models we simulated. Neither method without modification was able to detect an association in terms of correlation with the disease variants when the marginal $RR = 1.5$ and 3.0. Weighting the eigenvectors by their level of significance (eigenvalue) produced intriguing results and will need to be studied further. This may allow the differentiation of effects (marginal vs. joint) if needed. This PLS method makes mathematical sense (i.e. isolate variance that correlates with outcome) and does demonstrate power in assessing interactions.

In future work, we want to determine whether the background corrected method is viable to detect differences in correlation between cases and controls. In theory, the more SNPs you add to the correlation matrix the more chance that additional "smaller effects" can be captured. As the number of SNP increases, there is the increase of differences between cases and controls due to factors unrelated to disease. Controlling for these differences (noise) including random effects will be necessary and the background corrected method is a candidate to achieve this goal.

Allele frequency differences between cases and controls due to systematic ancestry differences, can cause spurious associations in disease studies [9]. Because the effects of stratification vary in proportion to the number of samples, stratification is an increasing problem in large scale association studies [9]; however, as more markers are added to a model subtle background population differences can be observed.

SIMLA allows for additional SNPs to be output, but we felt it was important to understand the basic "complex" model first.

Table 4.1: Composite correlation between V_1 and V_2 for cases (top) and controls (bottom) for the null models 1, 2, 3 and 4, and for the alternative models 5 and 6 (Table 3.1). Numbers in the table are the mean for 1000 simulations. Prevelance = 0.10 and interaction relative risk = 1.0.

Interactive Relative Risk (RR) = 1.0												
RR for V_1, V_2	1.0 (null 1,2, 3 & 4)				1.5 (alternative 5)				3.0 (alternative 6)			
MAF	0.05	0.15	0.35	0.50	0.05	0.15	0.35	0.50	0.05	0.15	0.35	0.50
$r^2_{Hap_{1,2}} = 0.0$	-0.003 0.000	0.002 -0.002	-0.002 0.000	-0.001 -0.001	0.000 0.000	0.002 0.000	0.000 -0.003	-0.002 -0.002	-0.006 -0.003	-0.014 -0.008	-0.016 -0.013	-0.014 -0.014
$r^2_{Hap_{1,2}} = 0.4$	0.001 0.000	0.001 0.000	-0.001 0.001	0.001 0.000	-0.001 0.001	-0.001 0.002	-0.002 -0.002	-0.002 -0.004	-0.009 -0.003	-0.012 -0.010	-0.018 -0.011	-0.012 -0.010
$r^2_{Hap_{1,2}} = 0.7$	0.000 0.003	-0.000 0.002	0.001 0.002	0.001 0.002	0.001 -0.001	0.001 -0.003	-0.003 -0.002	-0.002 0.000	-0.008 -0.003	-0.010 -0.010	-0.017 -0.013	-0.014 -0.012
$r^2_{Hap_{1,2}} = 0.9$	0.001 -0.003	0.001 -0.001	0.001 0.000	-0.001 0.001	-0.001 -0.001	0.001 -0.001	0.000 0.001	-0.002 -0.005	-0.007 -0.004	-0.011 -0.007	-0.018 -0.013	-0.014 -0.013

Table 4.2: Composite correlation between V_1 and V_2 for cases (top) and controls (bottom) for the alternative models 7, 8 and 9 (Table 3.1). Numbers in the table are the mean for 1000 simulations. Prevelance = 0.10 and interaction relative risk = 3.0.

Interactive Relative Risk (RR) = 3.0												
RR for V_1, V_2	1.0				1.5				3.0			
MAF	0.05	0.15	0.35	0.50	0.05	0.15	0.35	0.50	0.05	0.15	0.35	0.50
$r^2_{Hap_{1,2}} = 0.4$	0.027 -0.001	0.069 -0.009	0.112 -0.018	0.116 -0.017	0.024 -0.006	0.072 -0.013	0.107 -0.021	0.103 -0.026	0.023 -0.010	0.052 -0.026	0.065 -0.038	0.065 -0.040
$r^2_{Hap_{1,2}} = 0.7$	0.025 -0.004	0.067 -0.007	0.114 -0.017	0.117 -0.014	0.027 -0.003	0.069 -0.012	0.105 -0.025	0.101 -0.026	0.025 -0.011	0.053 -0.024	0.070 -0.041	0.065 -0.039
$r^2_{Hap_{1,2}} = 0.9$	0.026 -0.002	0.069 -0.008	0.114 -0.015	0.117 -0.018	0.026 -0.005	0.068 -0.010	0.104 -0.024	0.103 -0.026	0.024 -0.010	0.049 -0.027	0.071 -0.042	0.067 -0.040

Table 4.3: Composite correlation between V_1 and V_2 for cases (top) and controls (bottom) for the alternative models 7, 8 and 9 (Table 3.1. Numbers in the table are the mean for 1000 simulations. Prevelance = 0.10 and interaction relative risk = 10.0.

Interactive Relative Risk (RR) = 10.0												
RR for V_1, V_2	1.0				1.5				3.0			
MAF	0.05	0.15	0.35	0.50	0.05	0.15	0.35	0.50	0.05	0.15	0.35	0.50
$r^2_{Hap_{1,2}} = 0.4$	0.061 -0.005	0.151 -0.020	0.220 -0.041	0.207 -0.044	0.063 -0.009	0.146 -0.028	0.192 -0.049	0.171 -0.054	0.058 -0.019	0.114 -0.046	0.119 -0.071	0.107 -0.072
$r^2_{Hap_{1,2}} = 0.7$	0.062 -0.008	0.151 -0.022	0.219 -0.040	0.206 -0.043	0.059 -0.010	0.146 -0.030	0.192 -0.052	0.170 -0.056	0.056 -0.017	0.117 -0.045	0.117 -0.072	0.108 -0.073
$r^2_{Hap_{1,2}} = 0.9$	0.065 -0.010	0.152 -0.022	0.218 -0.039	0.205 -0.043	0.064 -0.010	0.146 -0.028	0.191 -0.052	0.173 -0.052	0.059 -0.017	0.116 -0.044	0.118 -0.069	0.108 -0.073

Table 4.4: Type 1 error rates for null models 1, 2 and 3 (Table 3.1) with no LD for the haplotypes. Interaction relative risk = 1.0.

$r^2_{Hap_1} = r^2_{Hap_2} = 0.0$												
RR for V_1, V_2	1.0				1.5				3.0			
MAF	**0.05**	**0.15**	**0.35**	**0.50**	**0.05**	**0.15**	**0.35**	**0.50**	**0.05**	**0.15**	**0.35**	**0.50**
Permutation Z_1	0.04	0.04	0.03	0.06	0.05	0.05	0.06	0.05	0.05	0.04	0.05	0.04
Contrast Z_2	0.04	0.04	0.04	0.06	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.06
PLS χ^2	0.04	0.05	0.05	0.04	0.03	0.05	0.05	0.05	0.04	0.04	0.04	0.04
PLS Interaction	0.03	0.04	0.05	0.04	0.04	0.04	0.06	0.03	0.02	0.04	0.04	0.05

Table 4.5: Type 1 error rates for null model 4 (Table 3.1). Interactive relative risk = 1.0.

RR for V_1, V_2	1.0			
MAF	0.05	0.15	0.35	0.50
LD Permutation Z_1 $k = 1$				
$r^2_{Hap_1}, r^2_{Hap_2} = 0.4$	0.05	0.04	0.05	0.06
$r^2_{Hap_1}, r^2_{Hap_2} = 0.7$	0.06	0.05	0.06	0.05
$r^2_{Hap_1}, r^2_{Hap_2} = 0.9$	0.05	0.05	0.05	0.06
LD Permutation Z_1 $k = 2$ **weighted**				
$r^2_{Hap_1}, r^2_{Hap_2} = 0.4$	0.05	0.06	0.06	0.06
$r^2_{Hap_1}, r^2_{Hap_2} = 0.7$	0.06	0.06	0.06	0.05
$r^2_{Hap_1}, r^2_{Hap_2} = 0.9$	0.05	0.05	0.05	0.06
LD Contrast Z_2				
$r^2_{Hap_1}, r^2_{Hap_2} = 0.4$	0.05	0.05	0.05	0.04
$r^2_{Hap_1}, r^2_{Hap_2} = 0.7$	0.05	0.05	0.05	0.06
$r^2_{Hap_1}, r^2_{Hap_2} = 0.9$	0.06	0.04	0.05	0.04
PLS χ^2				
$r^2_{Hap_1}, r^2_{Hap_2} = 0.4$	0.06	0.04	0.05	0.05
$r^2_{Hap_1}, r^2_{Hap_2} = 0.7$	0.07	0.05	0.07	0.06
$r^2_{Hap_1}, r^2_{Hap_2} = 0.9$	0.06	0.04	0.06	0.06
PLS Interaction				
$r^2_{Hap_1}, r^2_{Hap_2} = 0.4$	0.03	0.04	0.05	0.04
$r^2_{Hap_1}, r^2_{Hap_2} = 0.7$	0.04	0.04	0.06	0.04
$r^2_{Hap_1}, r^2_{Hap_2} = 0.9$	0.03	0.05	0.06	0.05

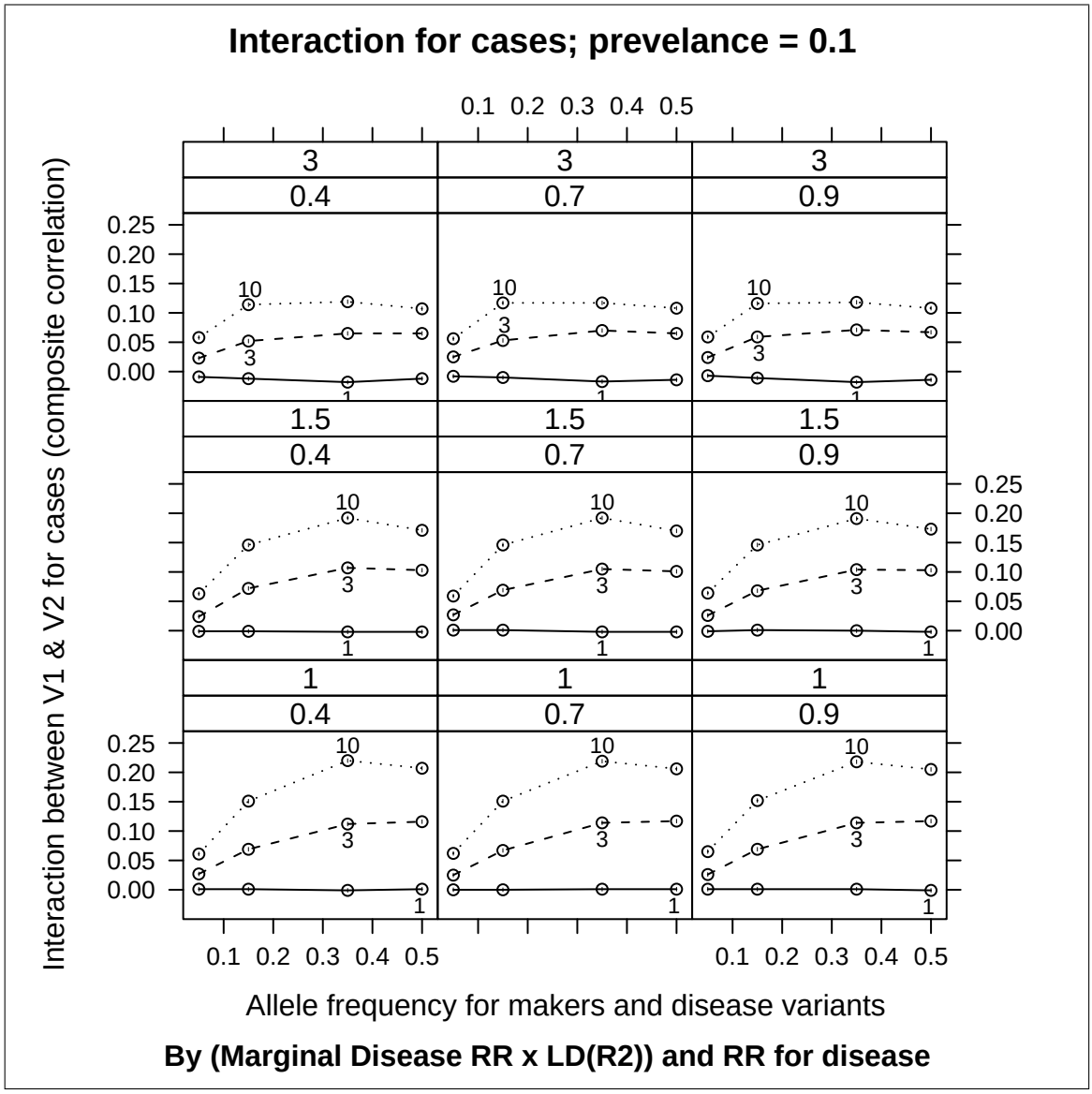

Figure 4.1: Composite correlation coefficients between V_1 and V_2 for disease model simulations using the SIMLA package.

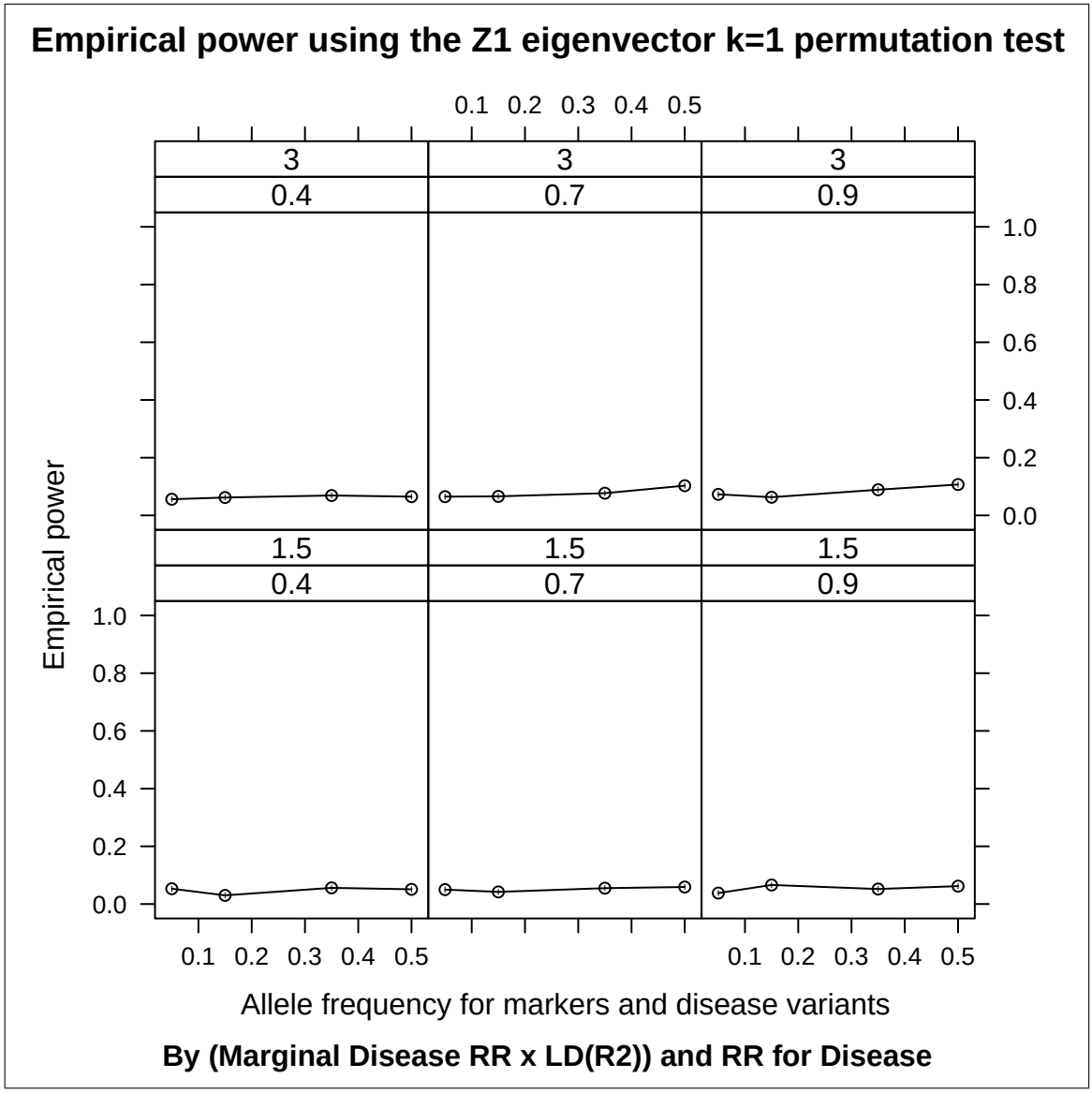

Figure 4.2: Empirical power using the Z_1 LD eigenvector ($k = 1$) permutation test for alternative models 5 and 6 (Table 3.1). Interaction $RR = 1.0$

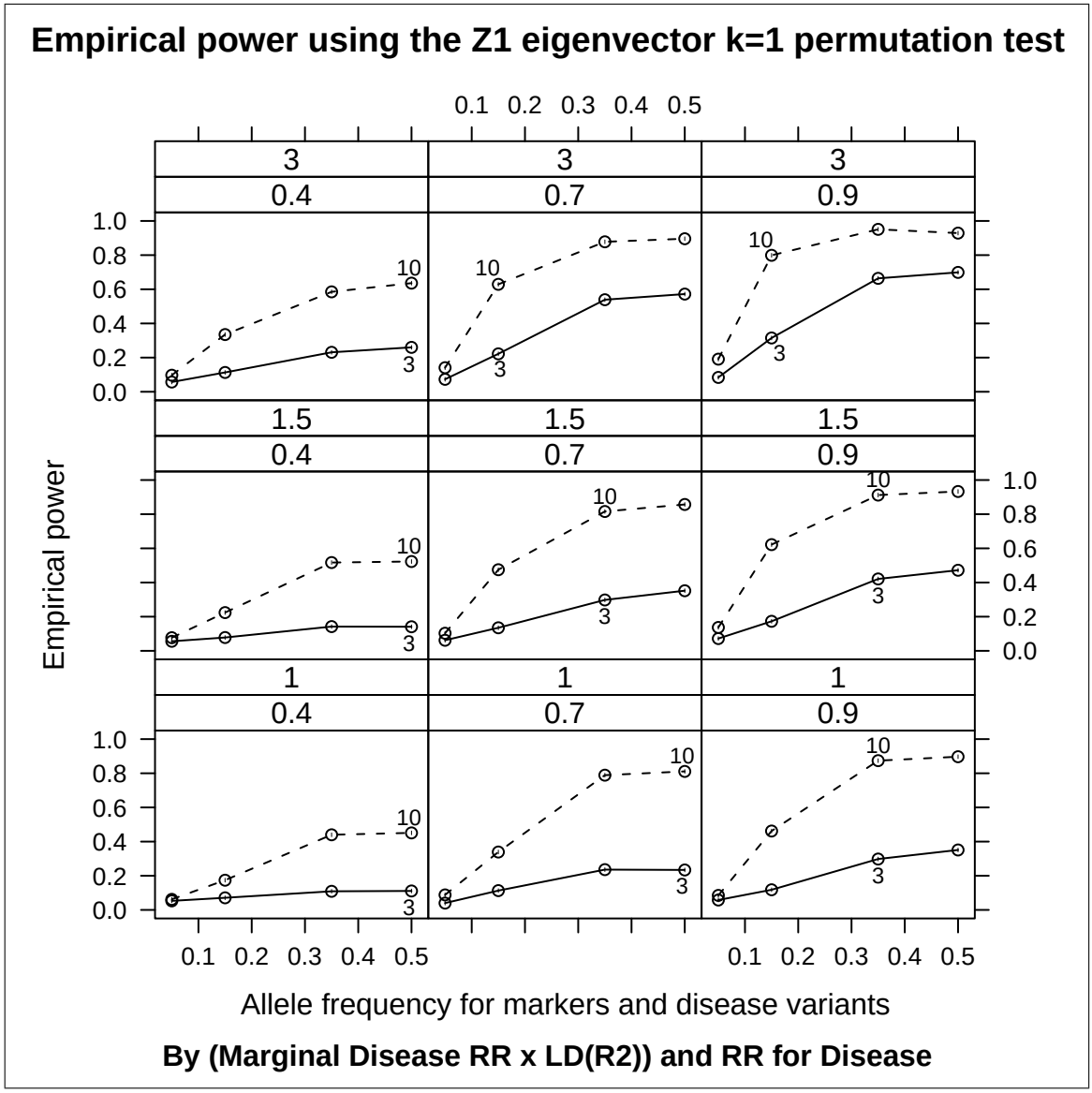

Figure 4.3: Empirical power using the Z_1 LD eigenvector ($k = 1$) permutation test for alternative models 7, 8 and 9 (Table 3.1). Interaction $RR = 3.0$ and 10.0

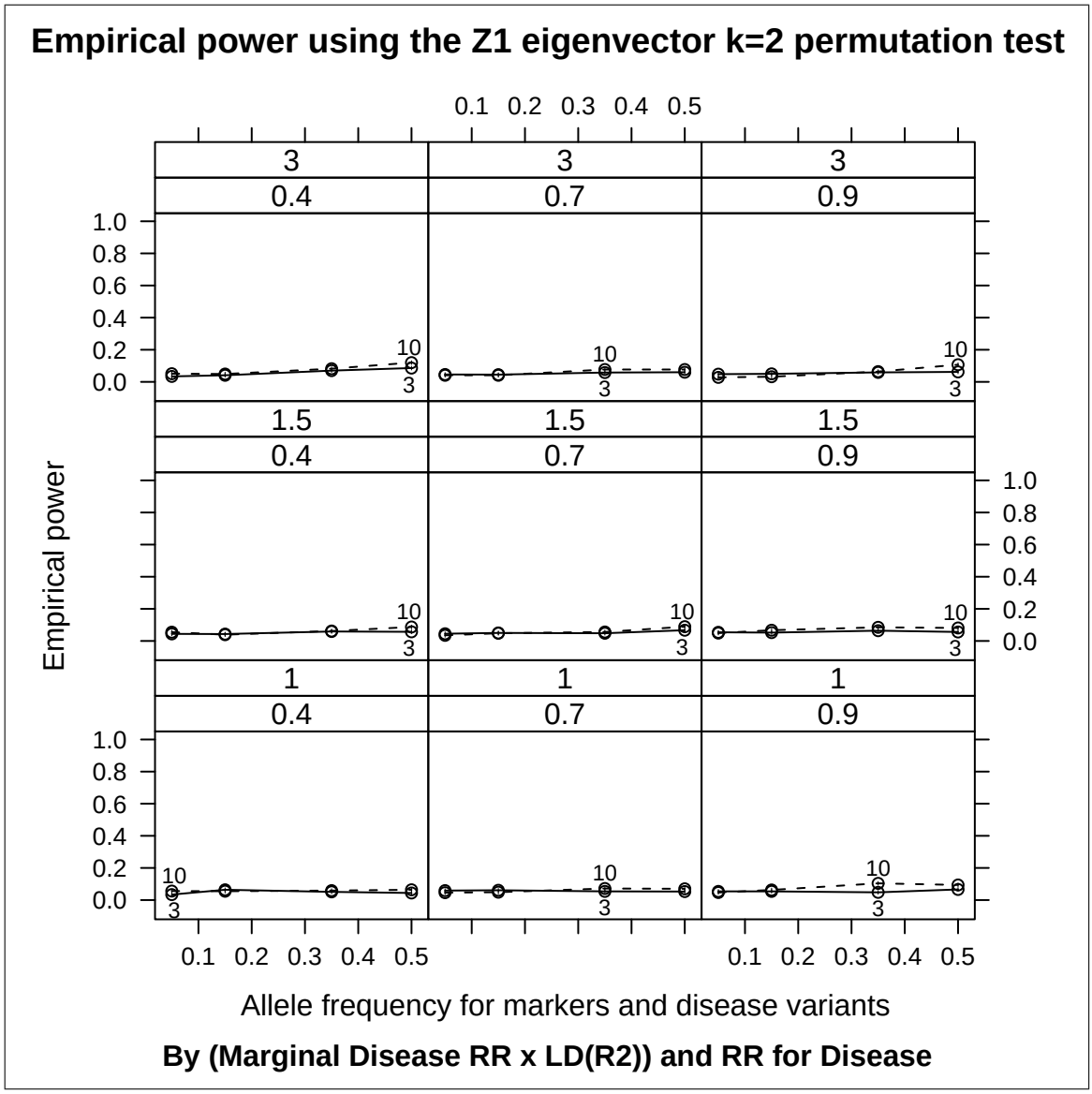

Figure 4.4: Empirical power using the Z_1 LD eigenvector ($k = 2$) permutation test. Interaction $RR = 3.0$ and 10.0.

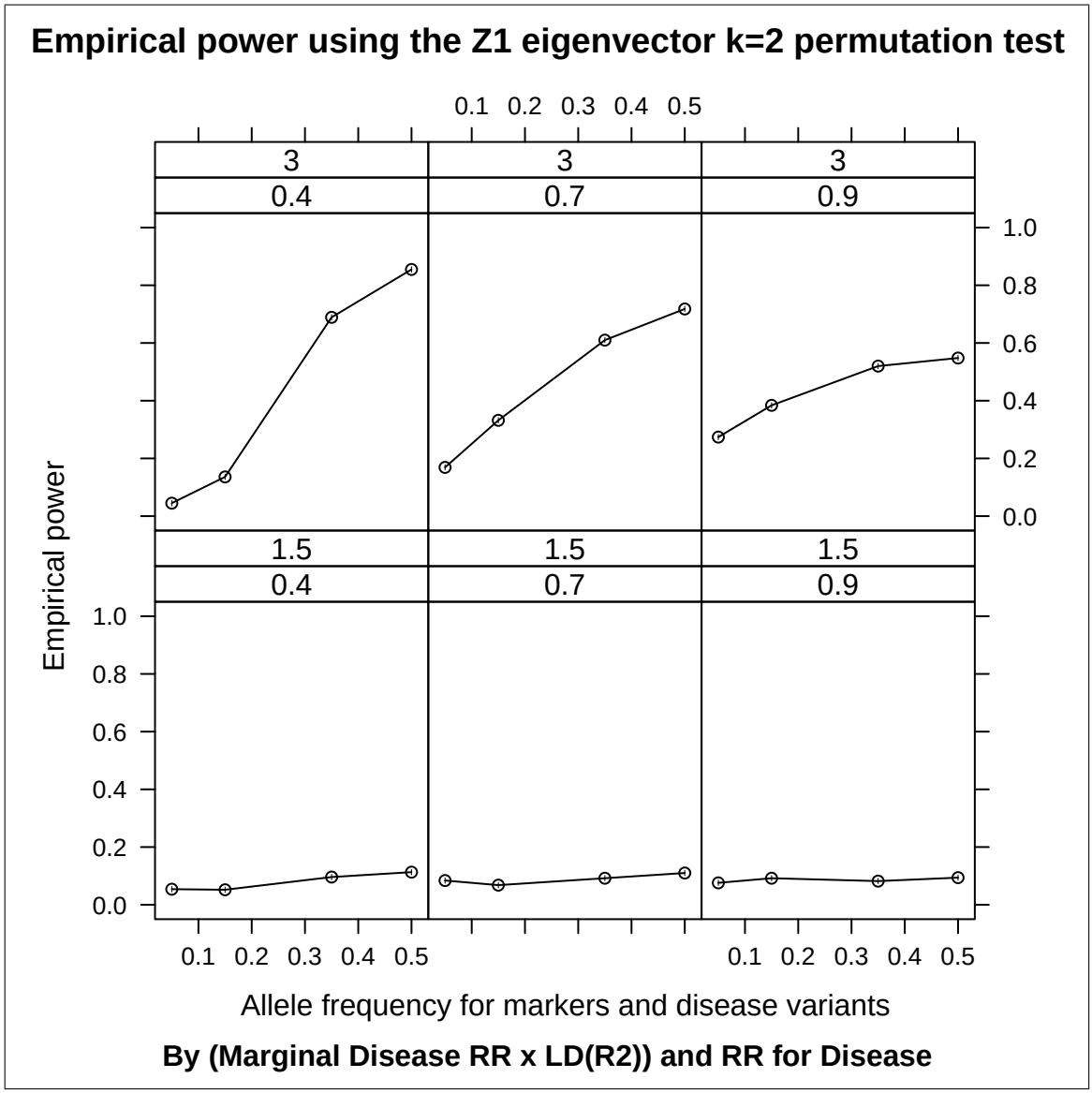

Figure 4.5: Empirical power using the Z_1 LD eigenvector ($k = 2$) weighted by the eigenvalue permutation test for alternative models 5 and 6 (Table 3.1). Interaction $RR = 1.0$

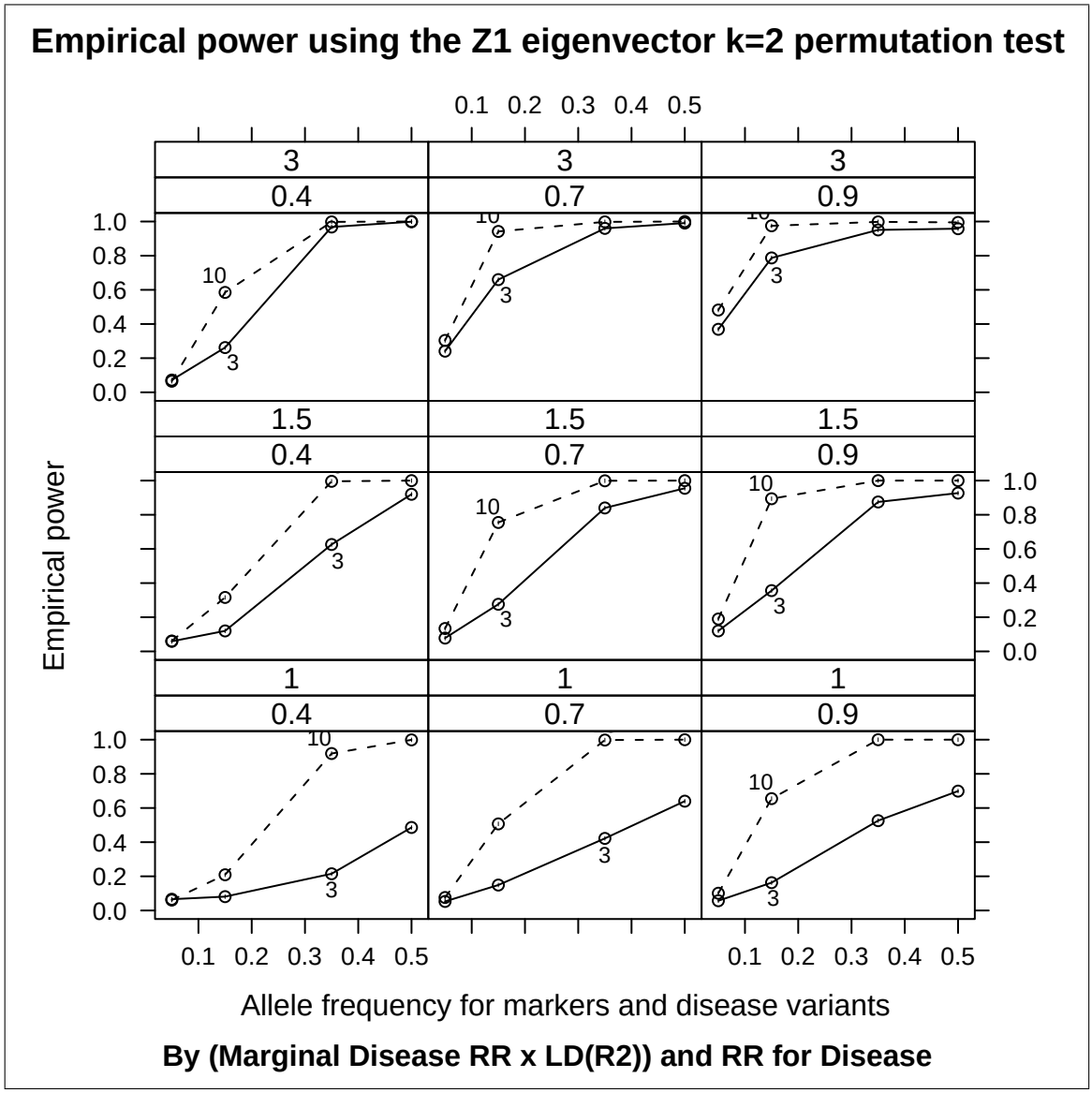

Figure 4.6: Empirical power using the Z_1 LD eigenvector ($k = 2$) weighted by the eigenvalue permutation test for alternative models 7, 8 and 9 (Table 3.1). Interaction $RR = 3.0$ and 10.0.

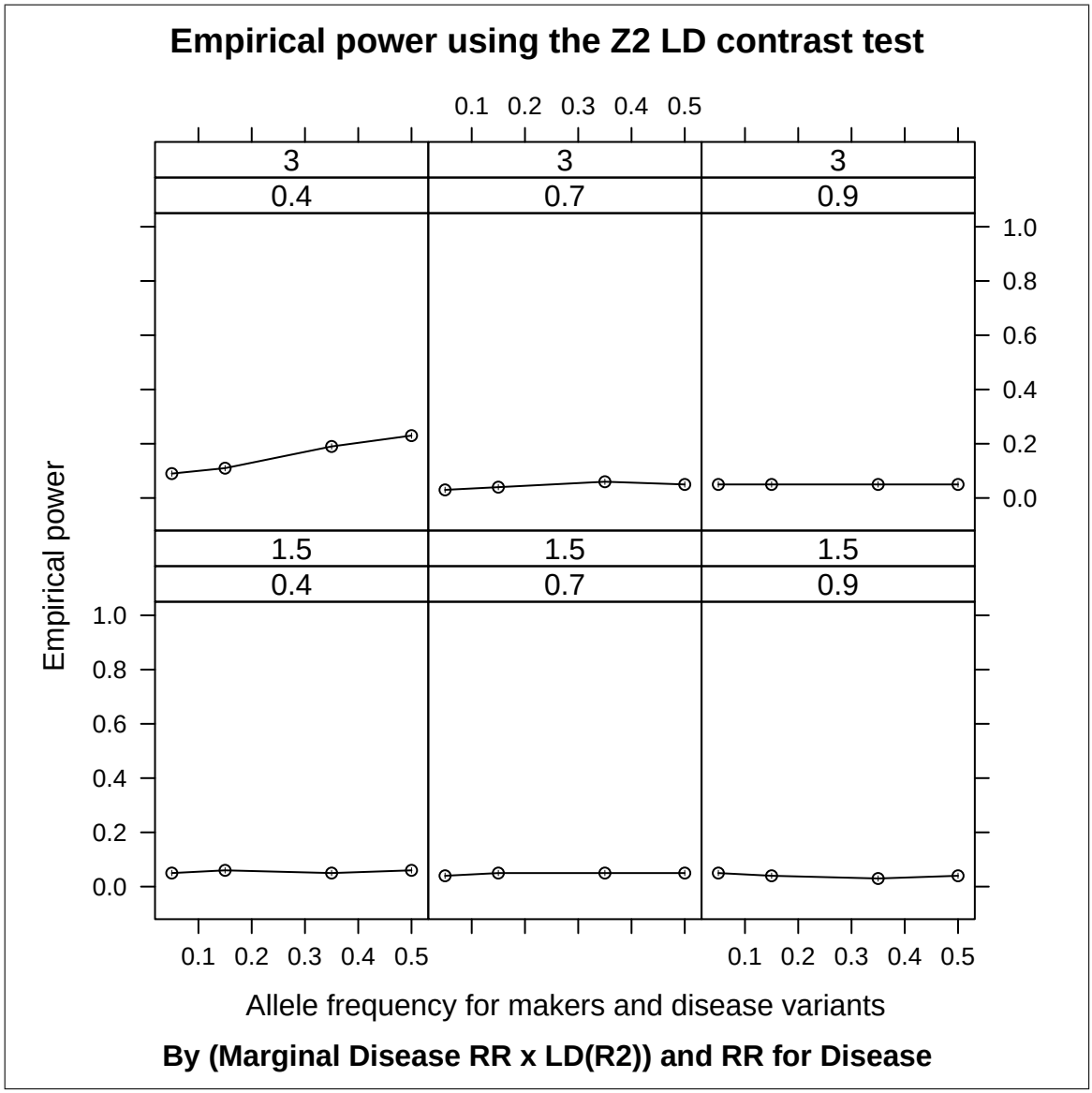

Figure 4.7: Empirical power using the Z_2 LD contrast test for alternative models 5 and 6 (Table 3.1). Interaction $RR = 1.0$

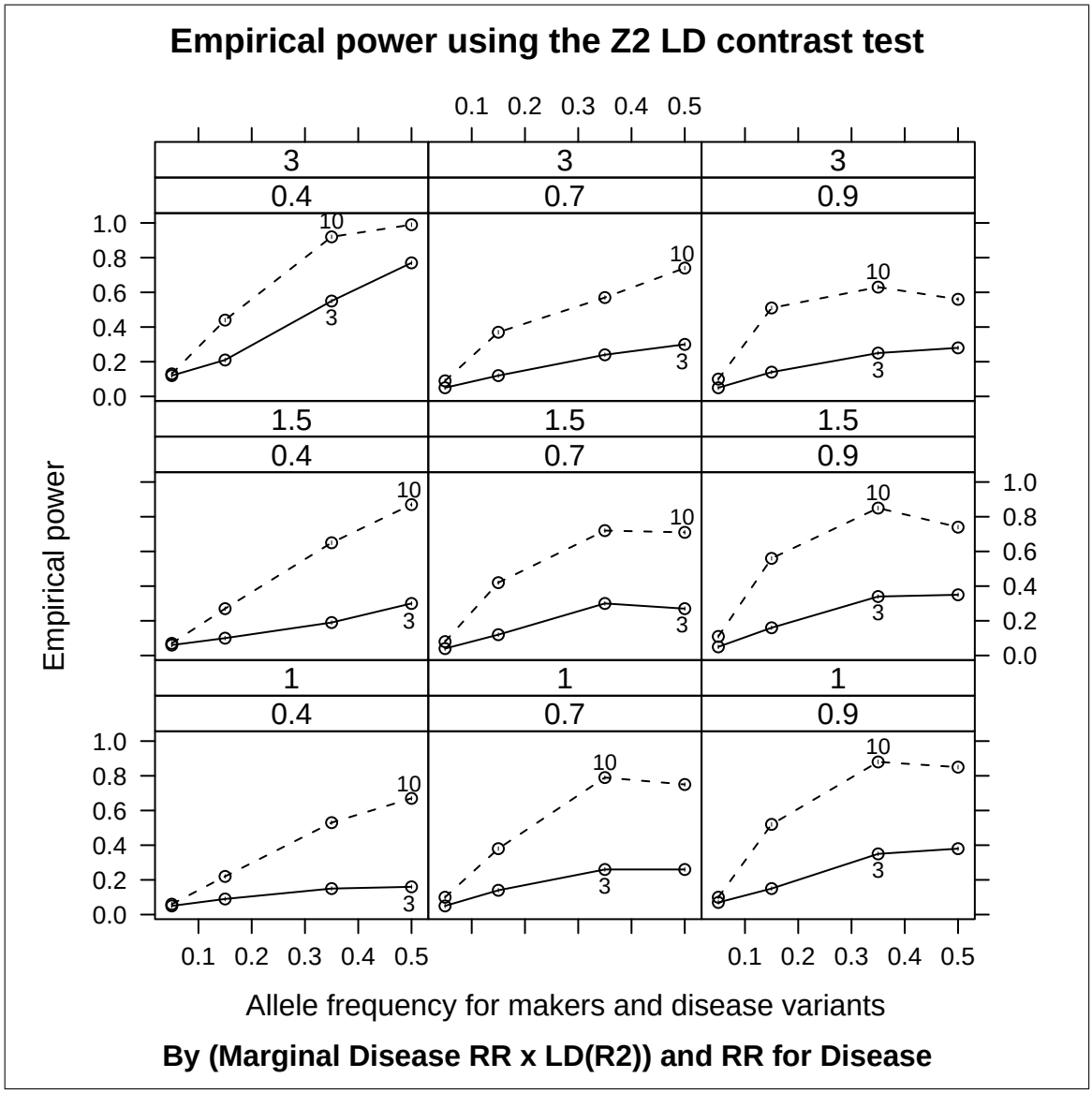

Figure 4.8: Empirical power using the Z_2 LD contrast test for alternative models 7, 8 and 9 (Table 3.1). Interaction $RR = 3.0$ and 10.0.

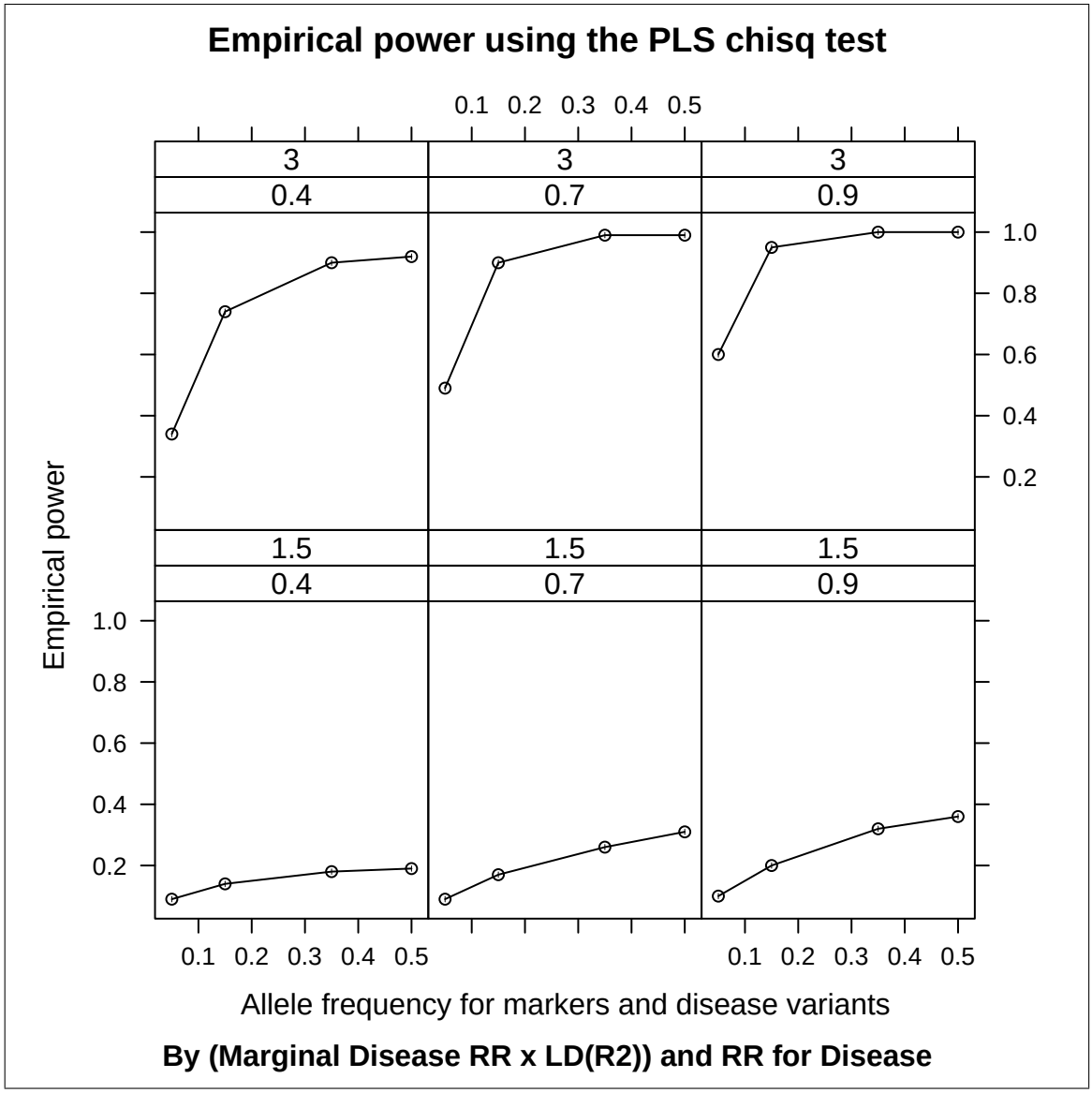

Figure 4.9: Empirical power using the partial least square χ^2 test for alternative models 5 and 6 (Table 3.1). Interaction $RR = 1.0$

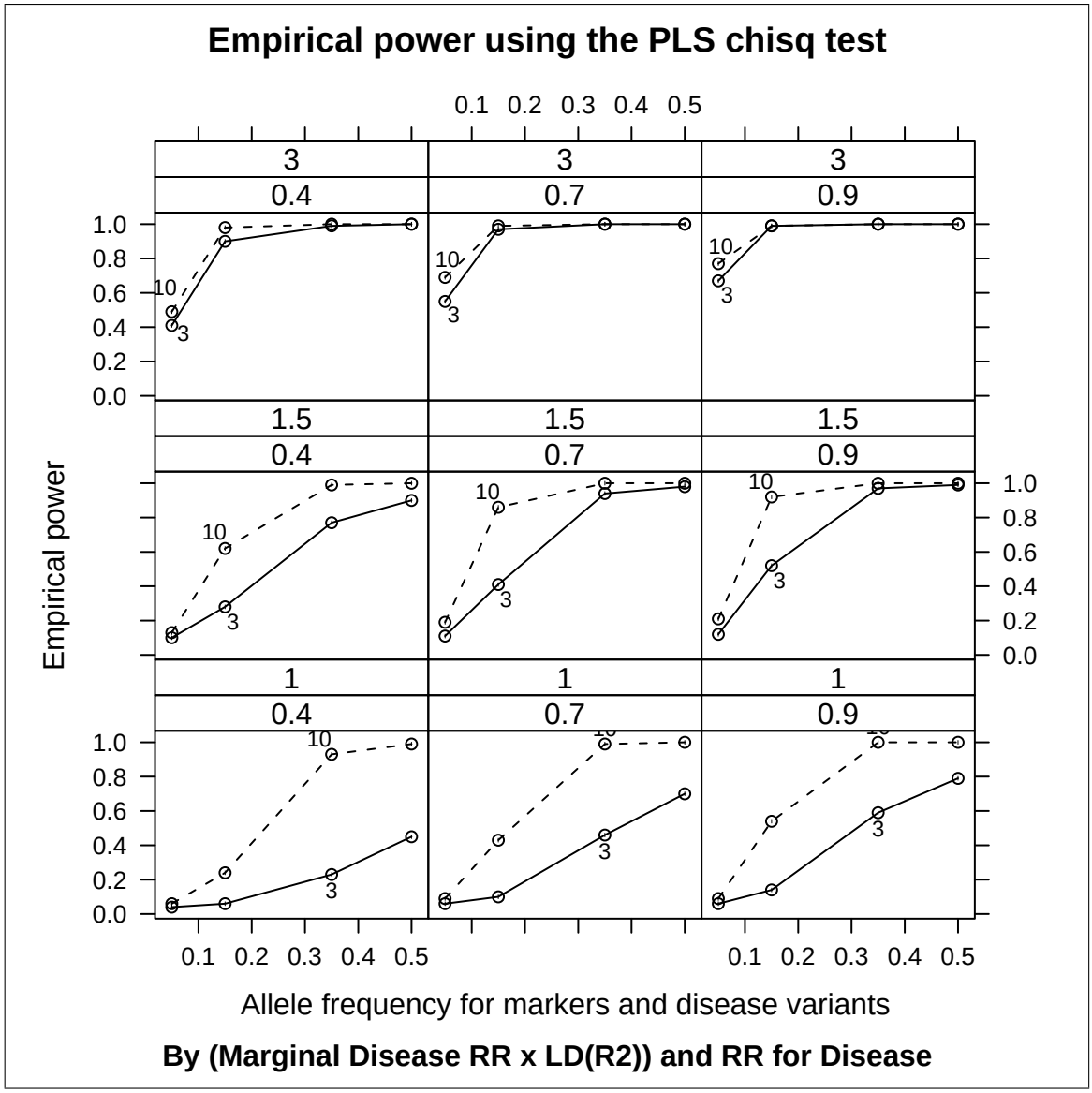

Figure 4.10: Empirical power using the partial least square χ^2 test for alternative models 7, 8 and 9 (Table 3.1). Interaction $RR = 3.0$ and 10.0.

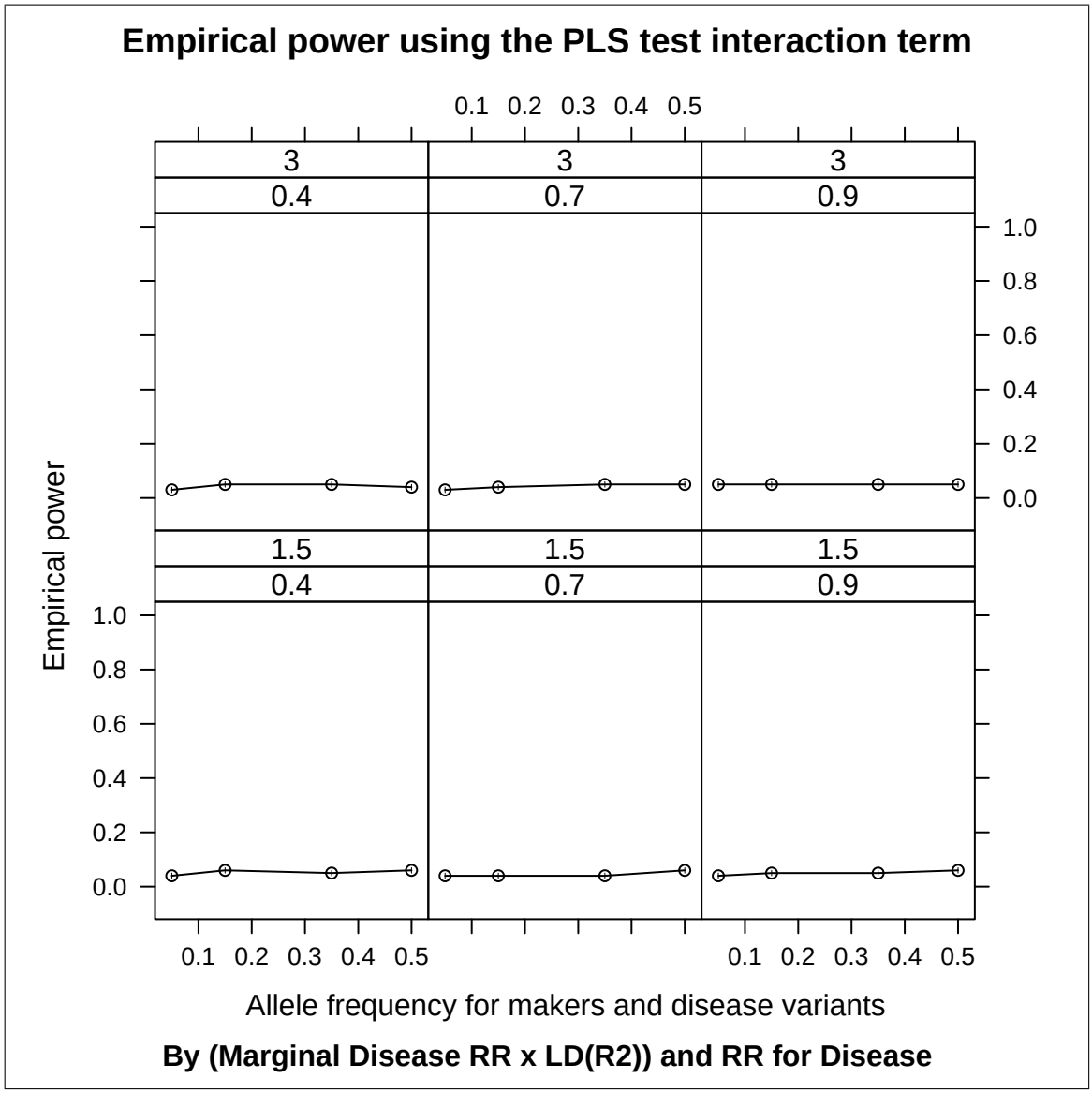

Figure 4.11: Empirical power using the partial least square test interaction term for alternative models 5 and 6 (Table 3.1). Interaction $RR = 1.0$

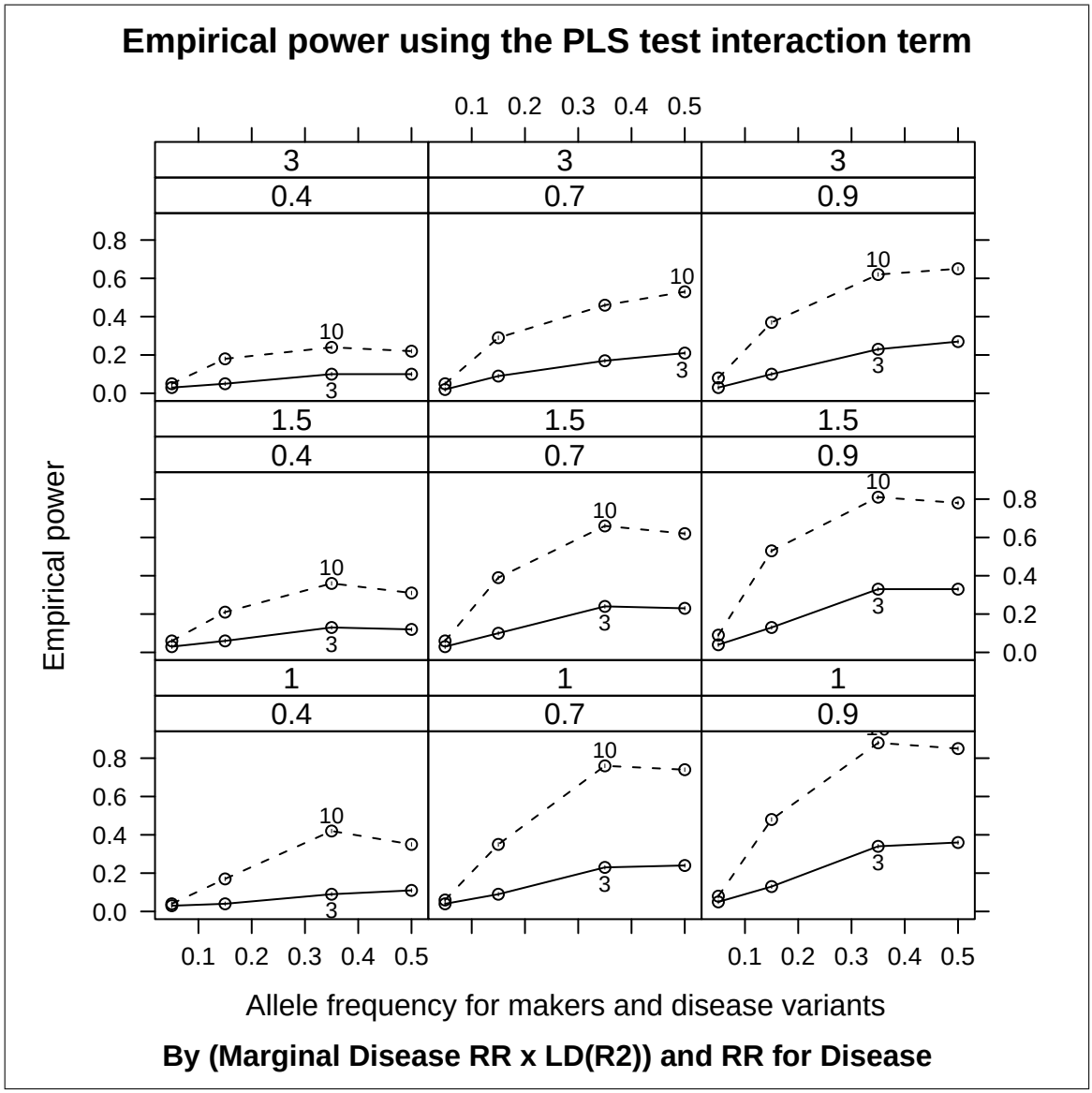

Figure 4.12: Empirical power using the partial least square test interaction term for alternative models 7, 8 and 9 (Table 3.1). Interaction $RR = 3.0$ and 10.0.

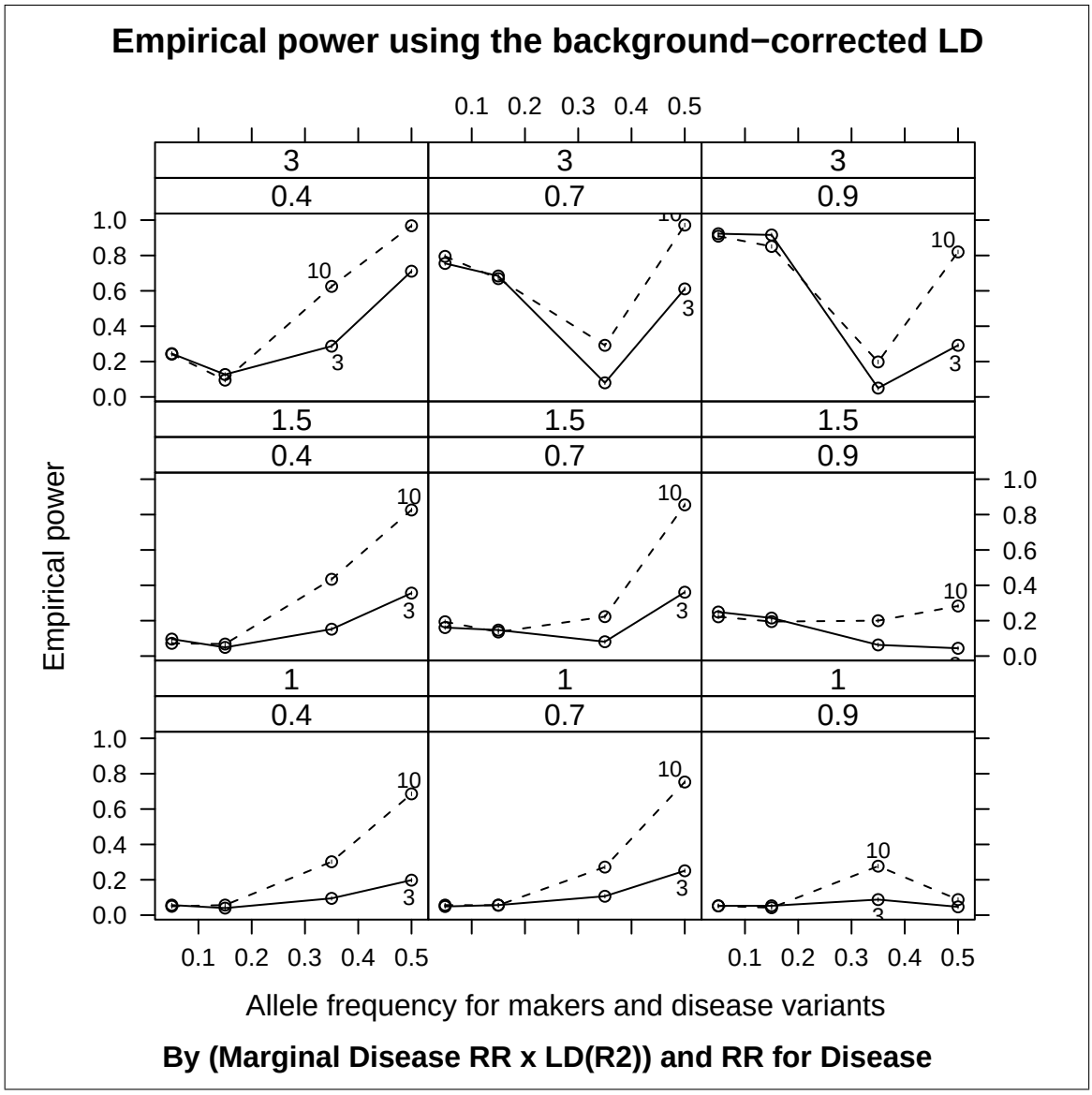

Figure 4.13: Empirical power using the background-corrected LD contrast method to test interaction

Chapter 5

Results of leukotriene pathway using matrix measures

With the knowledge gained from our simulations, we applied these methods to SNPs in the leukotriene biosynthesis pathway. Crosslin et al. assessed the role of the leukotriene pathway in CVD pathogenesis with association studies of *ALOX5AP* and *LTA4H* in a non-familial dataset of EOCAD [15]. The previously reported four-SNP haplotype (HapA) in *ALOX5AP* showed association with EOCAD in CATHGEN ($P = 0.02$), while controlling for age, race and CVD risk factors. HapK, the previously reported ten-SNP haplotype in *LTA4H* was associated with EOCAD in CATHGEN ($P = 0.04$). Another previously reported four-SNP haplotype in *ALOX5AP* (HapB) was not significant in our sample ($P = 0.39$). The overall lack of (or weak) association of single SNPs as compared with the haplotype results demonstrates the need for analyzing multiple SNPs within each gene in such studies. As noted, our results support a modest role for the leukotriene pathway in atherosclerosis pathogenesis, reveal important genomic interactions within the pathway, and suggest the importance of using pathway-based modeling for evaluating the genomics

of atherosclerosis susceptibility. As evident in the LD plots for *ALOX5AP* (Figure 5.1), there is minimal correlation for both Caucasians and African Americans. We did observe a difference in the LD block consisting of SNPs rs4769874, rs9551963 and rs9315050 for the African Americans. For *LTA4H* (Figure 5.2), we observe increased correlation in the presence of ten SNPs. Unlike the simulations, we did not know how many variants might exist and what the LD pattern with the observed variants and the susceptibility variants are for our sample. Based on our simulations, we would expect the LD permutation-based method Z_1 to have the highest power to detect differences between cases and controls as a whole-gene measure. While not as powerful, the LD contrast method (Z_2) does demonstrate the ability to distinguish between cases and controls. For the detection of gene-gene interactions, we would expect the PLS method to provide the highest power.

5.1 Results

Table 5.1 depicts the matrix measure results for the leukotriene biosynthesis pathway for both race groups. For the LD permutation Z_1, LD contrast Z_2 and background-corrected contrast test, we analyzed the SNPs composing the following: 1. HapA (*ALOX5AP*); 2. HapB (*ALOX5AP*); 3. HapK (*LTA4H*); 4. HapA $\times$ HapB; and HapA $\times$ HapK. For the PLS interaction method, we considered both HapA $\times$ HapB and HapA $\times$ HapK. We reported both the likelihood ratio-based results testing the global null hypothesis (all $\beta = 0$) as well as the single effect significance for the haplotype interaction term (probability of observing a Wald $\chi^2 \geq$ to observed). The P-values presented for Z_1 and Z_2 are the significance achieved from permuting

the phenotype labels of EOCAD. For reference, we added the EOCAD haplotype association results (last column in Table 5.1) from Crosslin et al.[15].

Tables 5.2 and 5.3 present the race stratified results for African Americans and Caucasians respectively. The race stratified LD plots for *ALOX5AP* and *LTA4H* comparing cases and controls can be found in Figures 5.1 and 5.2 respectively. From a visual perspective, the greatest correlation difference between cases and controls is in *LTA4H* for African Americans (Figure 5.2), although it is difficult to draw conclusions on the patterns. In a pairwise LD plot, the darker the box the greater the correlation (r^2) between two SNPs. SNPs SG12S16 and rs2660880 have different correlation patterns (case $r^2 = 0.25$, control $r^2 = 0.79$). SNP SG12S16 also has a different correlation pattern with rs17677715 (case $r^2 = 0.66$, control $r^2 = 0.19$). SNP rs2660890 has a number of different cases and control correlations with upstream SNPs. We did not observe the same qualitative difference for the Caucasians' case and control LD plots (Figure 5.1). Again, we added the EOCAD haplotype association results from Crosslin et al. [15].

For both race groups combined and stratified, we observed mixed results for Z_1 and Z_2 to support what Crosslin et al. observed in the haplotype analyses (Table 5.1). For instance, the Z_1 results ($P = 0.284$) are less supportive of LD differences between cases and controls than the Z_2 results ($P = 0.066$) compared to the haplotype result for HapA ($P = 0.02$) in both race groups (Table 5.1). However, the Z_1 results ($P = 0.064$) is more supportive than the Z_2 ($P = 0.115$) compared to the haplotype result for HapK ($P = 0.04$). For HapA in Caucasians, we could not reproduce the haplotype results ($P = 0.01$) using the Haplo.stats package through R Statistical

Computing. Both the permutation Z_1 ($P = 0.178$) and LD contrast Z_2 ($P = 0.216$) tests were not significant. We observed similar results with HapK (Table 5.3).

We were most intrigued by the HapK for both tests, most notably in the African Americans (Table 5.2). Helgadottir et al. reported a significant effect ($RR = 6.50$) in the Philadelphia African American sample[24]. This haplotype was not significant in the CATHGEN African American sample ($P = 0.27$), but generally we observe a qualitative difference in the LD patterns as illustrated in Figure 5.2. We believe power is lost beyond a seven-SNP haplotype because the correct estimation of phase becomes an issue, especially with a smaller n. The risk allele-pattern in *LTA4H* may not be driven from a single disease-associated haplotype, but rather a global haplotype pattern driven by LD for a subset of SNPs.

We did not observe a significant interaction between HapA and HapB for any method. This could be explained by the fact that the marker shared by both HapA and HapB (rs10507391) has different alleles included in the HapA and HapB risk haplotypes. This suggest they are independent and are supported by the results presented by Helgadottir et al. [25]. We did observe interesting results when assessing the HapA and HapK interaction, using both the permutation Z_1 ($P = 0.087$) and LD contrast Z_2 tests ($P = 0.139$), although this may be due to the marginal effects of HapK (Z_1 $P = 0.121, 0.018$ ($k = 2$); Z_2 $P = 0.023$).

The PLS method did not show evidence with either the global test or the interaction test. We did not expect to see significant results with PLS using SNPs from *ALOX5AP* and *LTA4H* in a multivariate logistic regression model. Our lack of multiple significant SNP association results for EOCAD summarized in Tables 2.3

(*ALOX5AP*) and 2.4 (*LTA4H*) do not support using this method.

5.1.1 *LTA4H*, HapK and deCODE genetics

deCODE genetics currently has two compounds (DG031 and DG051) in the pipeline for the prevention of CAD and ultimately a heart attack. Both compounds' ultimate goal is to attenuate the production of the pro-inflammatory compound LTB4, thus reducing atherosclerosis. DG501 is an *ALOX5AP* inhibitor and had reached Phase III clinical trials before a voluntary termination due to formulation issues. Those issues have been resolved according to their web site (http://www.decode.com). Based on the HapK effect size observed in African American subjects with MI from Philadelphia ($RR = 6.50$), deCODE genetics used this haplotype to identify CAD-susceptible study subjects [24]. Our results suggest greater complexity in the relationship between the HapK haplotype in *LTA4H* with genetic variation at other loci in the leukotriene pathway. Our ability to take these complex relationships into account in study design may cause variability in the trial results.

Table 5.1: Results of significance tests using matrix measures to assess genetic effects in the leukotriene biosynthesis pathway in the CATHGEN sample. For the PLS test, the first P-value is for the likelihood test and the second is for the interaction term

	Matrix Measure				Haplotype
	LD Permutation Z_1 $(k = 1)$	LD Contrast Z_2	PLS χ^2	PLS Inter.	Results
ALOX5AP HapA	$P = 0.284$	$P = 0.066$	N/A	N/A	$P = 0.02$
ALOX5AP HapB	$P = 0.063$	$P = 0.396$	N/A	N/A	$P = 0.39$
LTA4H HapK	$P = 0.064, 0.040$ $(k = 2)$	$P = 0.115$	N/A	N/A	$P = 0.04$
HapA × HapB	$P = 0.122$	$P = 0.464$	$P = 0.528$	$P = 0.231$	N/A
HapA × HapK	$P = 0.089$	$P = 0.276$	$P = 0.699$	$P = 0.566$	N/A

Table 5.2: Results of significance tests using matrix measures to assess genetic effects in the leukotriene biosynthesis pathway for African Americans in the CATHGEN sample

	Matrix measure				Haplotype
	LD Permutation Z_1 $(k = 1)$	LD Contrast Z_2	PLS χ^2	PLS Inter.	Results
ALOX5AP HapA	$P = 0.784$	$P = 0.697$	N/A	N/A	$P = 0.67$
ALOX5AP HapB	$P = 0.016$	$P = 0.090$	N/A	N/A	$P = 0.37$
LTA4H HapK	$P = 0.121, 0.018$ $(k = 2)$	$P = 0.023$	N/A	N/A	$P = 0.27$
HapA × HapB	$P = 0.515$	$P = 0.230$	$P = 0.715$	$P = 0.629$	N/A
HapA × HapK	$P = 0.087$	$P = 0.139$	$P = 0.750$	$P = 0.952$	N/A

Table 5.3: Results of significance tests using matrix measures to assess genetic effects in the leukotriene biosynthesis pathway for Caucasians in the CATHGEN sample

	Matrix measure				Haplotype
	LD Permutation Z_1 $k = 1$	LD Contrast Z_2	PLS χ^2	PLS Inter.	Results
ALOX5AP HapA	$P = 0.178$	$P = 0.216$	N/A	N/A	$P = 0.01$
ALOX5AP HapB	$P = 0.375$	$P = 0.203$	N/A	N/A	$P = 0.44$
LTA4H HapK	$P = 0.178, 0.105$ $(k = 2)$	$P = 0.309$	N/A	N/A	$P = 0.03$
HapA $\times$ HapB	$P = 0.233$	$P = 0.602$	$P = 0.119$	0.667	N/A
HapA $\times$ HapK	$P = 0.247$	$P = 0.090$	$P = 0.127$	0.916	N/A

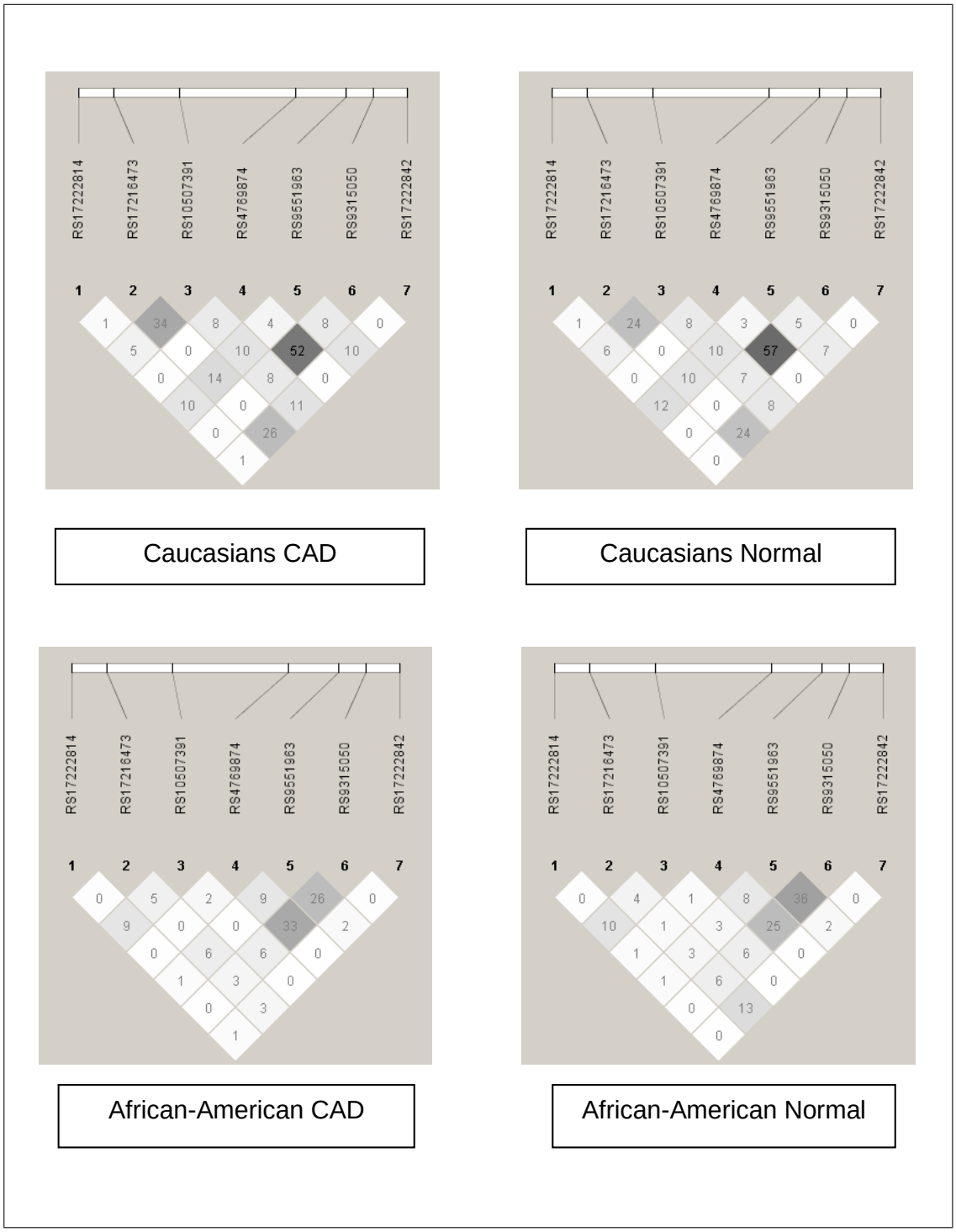

Figure 5.1: *ALOX5AP* (including SNPs for HapA and HapB) linkage disequilibrium figures for CAD and controls by ethnicity in the CATHGEN sample

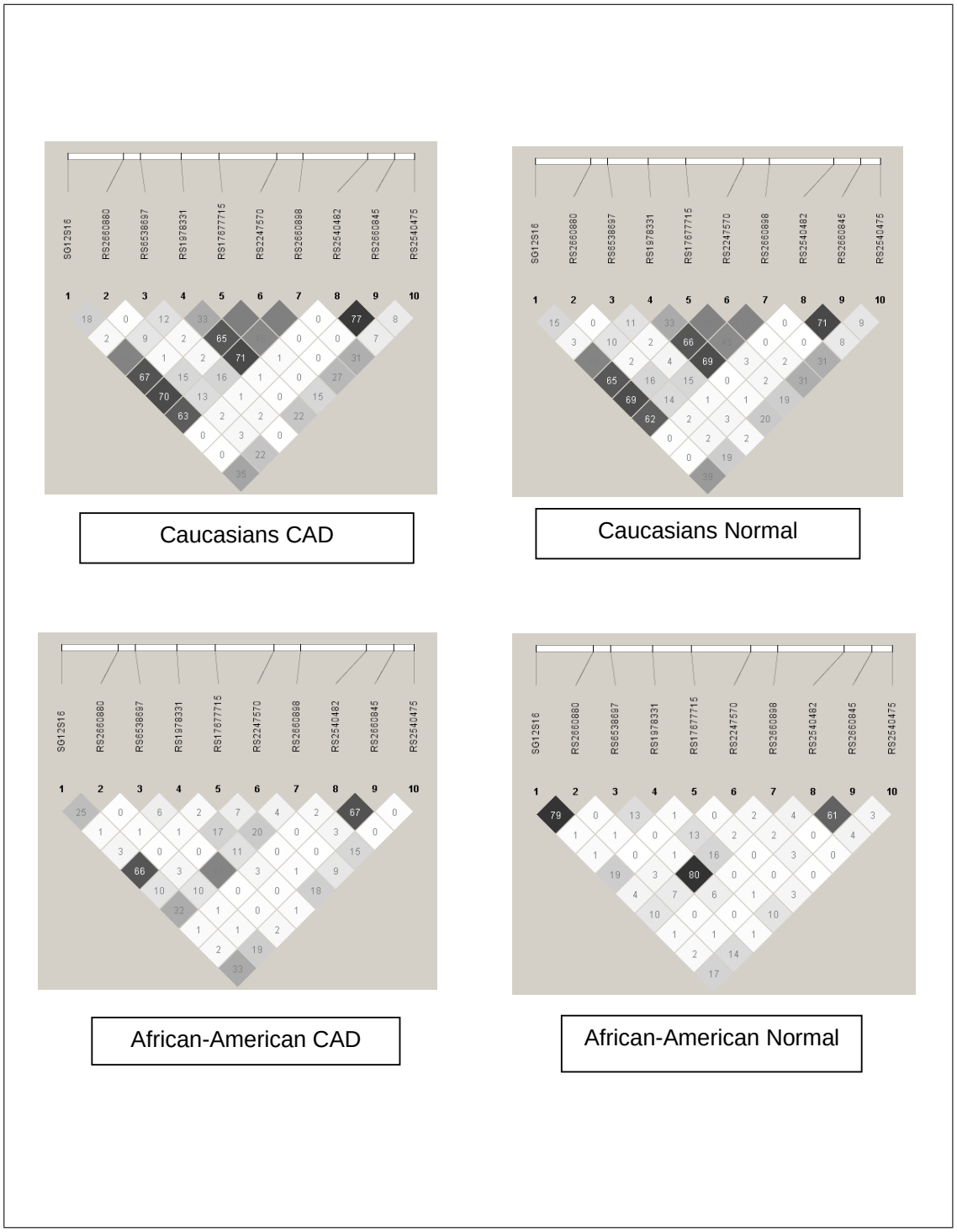

Figure 5.2: *LTA4H* (including SNPs for HapK) linkage disequilibrium figures for CAD and controls by ethnicity in the CATHGEN sample

Chapter 6

Conclusion

Our initial motivation for studying these methods was to find alternative robust measures for whole genes using multiple SNPs, in addition to haplotypes. In the context of whole-genome scans, alternative measures incorporating multiple SNPs at the gene-level will allow for gene set (pathway) analysis. Assessing alternative methods to detect interaction using multiple SNPs was equally important. To do this and to expand on Zaykin et al. [41] LD contrast methods, we took a comprehensive approach by simulating two separate disease variants with marginal effects (RR) in the presence of LD, with and without joint effects (RR). Null and alternative models were simulated to assess Type I error rates and power respectively. The null models can be summarized into three categories (Table 3.1): 1. Null models with no marginal effects $V_{1RR} = 1; V_{2RR} = 1$ and no LD (model 1); 2. Null models with marginal effects $V_{1RR} = V_{2RR} > 1$ and no LD (models 2 & 3); and 3. Null models in the presence of LD and no marginal effects (model 4). The alternative models can be summarized into three categories (Table 3.1): 1. Alternative models in the

presence of LD, marginal effects and no interaction (models 5 & 6); 2. Alternative models in the presence of LD, interaction effects and no marginal effects (model 7); and 3. Alternative models in the presence of LD, marginal effects and interaction (models 8 & 9).

We also expanded on previous versions of the test methods. For instance, we introduced the concept of weighting the eigenvector by the corresponding eigenvalue for the permutation method (Z_1). This approach may allow the detection of marginal effects in the presence of LD (see Figure 4.5). While incomplete, we expanded the LD background-corrected contrast method to include multiple SNPs from two disease causing variants. Zhao et al. [39] presented the method and results in the context of two SNPs. Finally, we compared the PLS approach introduced by Wang et al. [47] to the LD contrast methods.

We also used these measures to contrast CVD-phenotype case and control subjects using SNPs found in a candidate pathway (leukotriene biosynthesis). Based on correlation patterns observed, we were able to detect contrasts between cases and controls when stratifying by race groups. To truly interpret these results and for completeness, we will have to assess haplotype association (haplo.stats) using the simulated models. Having a biological pathway example to challenge these methods will validate their effectiveness.

6.1 The simulated disease model

Disease models incorporating two-loci interaction have been simulated and tested by others [33][32][39]. Zhao et al. described the LD pattern generated by two un-

linked loci in the context of penetrance, correctly concluding that formal statistics for testing gene interactions by use of LD are not yet developed [39]. Our purpose in simulating the two-variant disease models with haplotypes was to gain an understanding of pathway interactions in terms of correlation patterns.

One of the biggest hurdles with this project was developing a reasonable two-locus genetic model and specifying the parameters for input into SIMLA. Our goal was to understand what effect haplotype LD, MAF and marginal RR have on the interaction between disease variant 1 (V_1) and variant 2 (V_2) in terms of composite correlation. We simulated a RR interaction between both variants of 1.0, 3.0 and 10.0 (the limit for SIMLA). By no means did we exhaust the entire parameter space of possible models [33][32]. We simulated the data using a multiplicative disease model based on RR.

Our null models fell into three categories with an underlying theme of no interaction between the disease variants V_1 and V_2 (Table 3.1). We simulated a class of null models with no haplotype LD and no marginal effects. These models have no genetic effects and are considered the base null model. Next we simulated null models with no haplotype LD, but in the presence of marginal effects (RR). There are genetic effects with risk associated to the disease variants V_1 and V_2, but are independent from the markers used in the correlation matrix. Finally, we simulated null models in the presence of haplotype LD, but no marginal effects. The markers included in the correlation matrix were in LD with the disease variants, but there was no associated risk.

Our alternative models can be summarized into three categories (Table 3.1). We

simulated alternative models in the presence of haplotype LD and marginal effects, but no interaction. In this model, the markers in LD with the risk-associated disease variants will have an indirect association to disease. This model is the paradigm for common-disease/common-variant genetic association studies using SNPs. The idea is that the genotyped SNP marker is near (or in LD with) a causal variant. We then simulated alternative models in the presence of LD and interaction effects, but no marginal effects. This essentially is a synergistic epistatic model where each variant independently confers no risk. Finally, we considered alternative models in the presence of LD, marginal effects and interaction effects. The purpose of this class of models was to determine if these measures can capture both marginal and interaction effects.

Simulating a range of minor allele frequencies is an obvious parameter to describe with these measures (see Table 3.1). As noted, both Zaykin et al. [41] and Wang et al. [48] indicated that their respective tests have limitations when allele frequencies are low because of spurious correlation structures. Wang et al. suggested that the primary association information exists in the difference between the marginal allele frequencies [48]. Equally important, LD between the haplotype-specific markers was simulated at $r^2 = 0.4, 0.7$ and 0.9. For instance, criteria for tSNP selection may affect LD patterns when analyzing candidate pathways such as the leukotriene biosynthesis pathway. Finally, under the two-disease variant model, understanding the effect of increasing RR for the two variants was crucial. We generated RR of $1.0, 1.5$ and 3.0 marginal effect sizes to cover a range of power for our models. We generated a prevalence frequency of 0.10 for all models. We do note that certain age-dependent

diseases like CVD can approach frequencies of 0.25 at older ages. This age-dependent prevalence merits further investigation. Understanding how to present the data for interpretation was also a major hurdle. Given the number of parameters, we used trellis plots to illustrate a trend in interaction correlation in four (model parameter) dimensions. In the future we will consider three-dimensional surface plots by group to illustrate the results.

The models were difficult to assess statistically. In a complex system such as this with multiple correlations (both marginal and interaction), it was difficult to determine the source of a significant result. There are very few data driven models that produced high correlation. The underlying correlation is low even in strong interaction models. In the future we will consider alternating levels between settings. For instance, we will assess what effect a low MAF for one variant and a high MAF for another has on these tests, instead of simulating equal frequencies. The same could be considered for marginal RR for each respective variant. We need to do further testing with alternating LD patterns in the presence of disease risk. Future research will include data simulations under the recessive, dominant and additive disease models.

6.2 Matrix measures

We investigated the statistical properties of the matrix-based statistics to measure epistasis under various model conditions. Zaykin et al.[41] presented the LD matrix measures to provide statistical support in quantifying difference in a graphical display of LD of cases and controls. This was our original motivation for exploring these

measures. One major drawback with these matrix methods is that they do not produce effect sizes.

6.2.1 LD permutation

We did not observe Type I error rate inflation for any of the null models using the Z_1 method. For alternative models 5 and 6 (Table 3.1), we observed low power. These models are the paradigm for common-disease/common-variant genetic association studies using SNPs. The LD permutation test gained power with the synergistic epistatic model (model 7 in Table 3.1) where each variant independently has no risk, although interaction $RR = 3.0$ only reaches 0.40 under high LD (0.9). The $RR = 10.0$ performed better, but probably is unrealistic. There was additional gain in power for the full models of LD, interaction and marginal effects (Table 4.6). One drawback with this measure is deciding *a priori* how many eigenvectors to use in the featurevector (matrix of eigenvectors). Krazanowski's suggestion of using the first integer below 50% of the original number of variables is a valid argument and supported by our Z_1 power results when $k = 2$ (Figure 4.4). We introduced weighting the eigenvector by its corresponding eigenvalue as an alternative approach. We observed an increase in power for alternative models 5 and 6 (table 3.1) at marginal $RR = 3.0$. There was a gain in adding the additional weighted eigenvector to detect the indirect risk via LD. There was additional gain in power for the full models of LD, interaction and marginal effects (Figure 4.5). Our results suggest that a four-SNP matrix will most likely have a single important component, but variance explained by additional eigenvalues/eigenvectors is important. This approach on higher order

multivariate data will need to be explored.

6.2.2 LD contrast

We again did not observe Type I error rate inflation for any of the null models using the Z_2 method. We observed low power for alternative models 5 and 6 (table 3.1), but there was a slight increase (0.20) at low LD (0.4) and high marginal risk (3.0) (Figure 4.7). We observed an increase in power for the synergistic epistatic model, but again only reaching 0.40 under high LD (0.9) for $RR = 3.0$. There was additional gain in power for the full models of LD, interaction and marginal effects illustrated in Figure 4.8. We will need to explore the increase in power at low LD in the alternative models in the presence of LD and marginal effects. This method in combination with the permutation test would be a viable approach to assess correlation differences.

6.2.3 Partial least squares

Understanding what question you want to ask is important to which model you wish to use. As evident in the power plots for the PLS χ^2 method (Figures 4.9 and 4.10), this is a powerful test to detect the marginal effects as well as effects due to interaction. If the purpose of the analysis is to specifically test for interaction in the presence of overall genetic effects, this test may not be powerful. Reporting the interaction term P-value will assist in identifying the interaction significance while controlling for the main effects. Mathematically it is logical to choose factors that best predict outcome, but adding SNPs from one gene in a multivariate model may not be the best approach.

6.2.4 LD background-corrected

The results for the LD background-corrected method are inconclusive and warrant further investigation. Our inability to derive nominal Type 1 significance prevented us from calculating empirical power. In addition, under certain conditions the model failed to converge. We would like to explore what combinations of parameters led to this failure. Wang et al. presented this method in the context of analyzing two SNPs and only suggested expanding to the LD-contrast test presented by Zaykin et al. [41] when comparing multiple SNPs. For the mixed model, we analyzed all four SNPs from two different haplotypes when computing the BLUP for an individual. Perhaps background LD should be calculated on a two-SNP or single variant basis instead of across variants. Understanding the difference of using unlinked variants with this model will be crucial.

6.3 Other areas of research to consider in the future

6.3.1 Adjustment of covariates and stratification

In most complex diseases like CAD it is important to adjust for covariates. In the context of the LD-contrast methods, there are multiple methods to explore in the future. Correlations matrices could include continuous variables such as gene expression or copy number variation (CNV) data, as well as binary data. Weighting these correlations based on *cis* or *trans* relationships would allow the incorporation of prior biological pathway information.

There may be a need to compare multiple groups instead of a binary outcome. For instance, in addition to comparing overweight subjects with favorable metabolic

profiles versus normal weight subjects with less favorable metabolic profiles, we could consider a third or fourth group with opposite metabolic and weight profiles. The permutation-based method (Z_1) can be expanded to allow for multiple comparisons. The background-corrected LD contrast and PLS methods introduced by Wang et al. [48] [47] are based in a regression framework, thus allowing for the adjustment of covariates.

LD is the consequence of selection on the basis of the disease variant and "background LD" due to various other factors [48]. These factors could include admixture. Because we are using correlation patterns to assess the difference between cases and controls, it will be invaluable to control for correlation not associated to disease. One obvious method is to stratify by race groups. Recently, many of the genome-wide association studies using HapMap-based principal components suggest removing subjects with genetically-identified racial groups rather than self-identified. The LD background-corrected method could be used to control admixture when estimating the random effects at the subject level.

6.3.2 GWAs and Gene Set Enrichment Analysis

Evans et al. investigated epistasis in genome-wide association using different research strategies [32]. This includes a single locus-search, an exhaustive two-locus search (pairwise), and two (two-stage) procedures in which a subset of loci initially identified with single locus tests are analyzed using a full two-locus model. The matrix-based measures will provide an additional tool to assess epistasis and may increase power by harnessing data from more SNPs compared to a two-loci approach. We would

also like to explore the importance of intergenic regions.

The LD-based statistics we have generated can be used in pathway ascertainment tools that require scores on a gene-level. We could compute pathway-level association measures by deriving an enrichment scores from GSEA. Providing a single statistical measure per gene using multiple SNPs will also enable the combination of multiple types of genomic data at a gene-level, including expression data.

6.3.3 The HapMap

Another future goal is to understand what level of pathway correlation patterns are observed in the presence of intra-gene LD using HapMap data. Using pathway data from KEGG, we will derive gene/SNP sets to query in HapMap. Given prior knowledge of a biological pathway, creating a matrix measures of correlation patterns of SNPs for genes constituting that pathway will provide insight into SNP-SNP and gene-gene interaction and the association to disease. I look forward to addressing these goals in my future research endeavors.

Appendix A

Sample Code

A.1 Perl

```
#!/usr/local/bin/perl
my @allele_freq = ('0.05','0.15','0.35','0.5');
my @prevelance = ('0.10'); #('0.10','0.25');
###  d RR
my @d_lrr = ( ); #this is for thei disease RR
my @d_lrr_inter = ( ); #this is for the RR interaction
#this is creating an array of arrays.....must dereference
my @disease_rr = (1.0,1.5,3.0);
my @disease_rr_inter = (1.0,3.0,10.0);
my @rrline2 = ( );
open(RRPARM,'<simla_parms_alternative.csv'); #open simla haplotype parameters file
my @rrlines = <RRPARM>;
```

```
shift(@rrlines);   #get rid of header
while (my $line = shift(@rrlines)) {
chomp($line);
push @rrline2 ,[split(/,/,$line)];
}
my $loop_index = 1;  # to connect simulations with statistics
for  my $prev (@prevelance) {
  for  my $a_freq_d1 (@allele_freq) {
    for  my $a_freq_d2 (@allele_freq) {
      for  my $d1 (@d_lrr) {
for  my $d2 (@d_lrr) {
  for  my $d1_d2 (@d_lrr_inter) {
    for my $row (@rrline2) {
###conditional test to see if paramters equal one another
if (  ("$a_freq_d1" eq "@$row[24]") and ("$a_freq_d2" eq "@$row[25]") and ("@$d1" eq "@$d2")  ) {
print "this is perld1: $a_freq_d1 this is rd1:@$row[24]  this is perld2:$a_freq_d2 this is rd2:@$row[25]\n";
my $allele2 = (1.00 - $a_freq_d1);
open(SIMPARM,'<simla_ctl.txt'); #READ
my @simlines = <SIMPARM>;
open(SIMPARM2,'>simla_ctl_final.txt'); #WRITE, CREATE, TRUNCATE
  while (my $line = shift(@simlines)) {
    chomp($line);
    if (substr($line,0,20) eq 'parameters_vector_RS') { #updating male parameter vector
       print SIMPARM2 "parameters_vector_RS:  $a_freq_d1 0.5 $a_freq_d2 0.5 0 0 @$d1\n";
     } #end of if
    elsif (substr($line,0,20) eq 'parameters_vector_RR') { #updating male parameter vector
       print SIMPARM2 "parameters_vector_RR:  $a_freq_d1 0.5 $a_freq_d2 0.5 0 0 @$d1\n";
     } #end of elsif
   elsif (substr($line,0,9) eq 'haplos:1:') {   #look for haplotype1
```

```
      print "updating $line\n";
    print SIMPARM2 "haplos:1:\n1-1 @$row[4] @$row[8]\n1-2 @$row[6] @$row[10]\n2-1]\n";
    } # end of elsif
   elsif (substr($line,0,9) eq 'haplos:2:') {   #look for haplotype2
     print "updating $line\n";
      print SIMPARM2 "haplos:2:\n1-1 @$row[12] @$row[16]\n1-2 @$row[14] @$row[18]\n2-1\n";
    } #end of elsif
    elsif ((substr($line,0,3) eq '1-1') or (substr($line,0,3) eq '1-2') or (substr($line,0,3) eq '2-1')  {
      next;
    } # end of elsif
#    elsif ((substr($line,0,6) eq '0 0.50') ) {
#     print SIMPARM2 "0 $a_freq_d1\n";
#     } # end of elsif
#     elsif ((substr($line,0,6) eq '1 0.50') ) {
#     print SIMPARM2 "1 $allele2\n";
#     } # end of elsif
    else  {
      print SIMPARM2 "$line\n";
    }  # end of else
  }  # end while
close SIMPARM2;
close SIMPARM;
print_control(0);  #print cases
wait ( );
print "running SIMLA for cases\n";
system('sim32 -r simla_ctl_final.txt');
wait ( );
system('mv 10001merstat1.ped temp_merlin.ped');
print_control(1);    #print controls
```

```
wait ( );
print "running SIMLA for controls\n";
system('sim32 -r simla_ctl_final.txt');
wait ( );
system('cat 10001merstat1.ped temp_merlin.ped > merlin.ped');
wait ( );
merge_simla_012 ( );
wait ( );
print "finish merging case/control \n";
print "running R \n";
system('R CMD BATCH blup_fast.R');
system('rm r_finished.csv');
print "finished R\n";
open(BLUP_P,'<r_output_blub.csv'); #READ
open(Z2_BLUP,'>>z2_blup.csv'); #APPEND
my @line_blup_p = <BLUP_P>;
#shift(@line_blup_p);   #get rid of header
while (my $line1 = shift(@line_blup_p)) {
chomp($line1);
my $rr_d1 = exp("@$d1");
my $rr_d2 = exp("@$d2");
my $rr_d1_d2 = exp("@$d1_d2");
print Z2_BLUP "$loop_index, $prev, $a_freq_d1, $a_freq_d2,@$row[0], @$row[1], @$row[2],
@$row[3], $rr_d1, $rr_d2, $rr_d1_d2,@$row[20], @$row[21], @$row[22], @$row[23], $line1,";
} #end of while 1
close Z2_BLUP;
close BLUP_P;
```

A.2 SAS IML

```
proc iml;
use temp2;
read all into praw;
t=ncol(praw-1);
/***********GET LD MATRIX ----------D' and CORR *************************/
start dprc(p,method);
if method='corr' then do;
k=ncol(p);
n=nrow(p);
dpmat=J(k,k,1);
corrmat=J(k,k,1);
do i = 1 to ncol(p)-1;
    do j =(i+1) to ncol(p);
q=p[,i]||p[,j];
na     = sum(q[,1]=-1);    nb   = sum(q[,2]=-1);
naa   = sum(q[,1]=0);     nbb  = sum(q[,2]=0);
naaa = sum(q[,1]=1);     nbbb = sum(q[,2]=1);
mx=q[+,1]/n;
my=q[+,2]/n;
nAABB_major_homo =      sum((q[,1]=1 & q[,2]=1));
naabb_minor_homo =        sum((q[,1]=-1 & q[,2]=-1));
nAA_bb_A_major_heter =  sum((q[,1]=-1 & q[,2]=1));
nAA_bb_B_major_heter =  sum((q[,1]=1 & q[,2]=-1));
delta  =((((nAABB_major_homo + naabb_minor_homo) -
nAA_bb_A_major_heter  - naa_BB_B_major_heter ) / ( 2*n)) - (((na - naaa)*(nb-nbbb))/(2*(n**2))));
pa_minor = 0.5-(sum(q[,1])/(2*n));   * this calculates the minor allele;
pb_minor = 0.5-(sum(q[,2])/(2*n));
```

```
PA_major = (1 - pa_minor);
PB_major = (1 - pb_minor);
PAA = (naaa /n);  * probability of PAA.....naaa = major allele....P(A) - .15 = -1;
PBB = (nbbb /n);  * probability of PAA;
DA = PAA - ((PA_major)**2);  * disequilibrium coefficient for A;
DB = PBB - ((PB_major)**2);   * disequilibrium coefficient for B;
composite_corr =  delta / sqrt(((PA_major*pa_minor)+DA)*((PB_major*pb_minor)+DB));
corrmat[i,j] = composite_corr ;   *fill the LD matrix;
corrmat[j,i] = composite_corr;
corr_function = corr(p);
           end;  *end of first do loop;
       end;  * end of second do loop;
   return(corrmat);
    end;
 /**********Z1 Z1 Z1 Z1 Z1 Z1 Z1 Z1 Z1 Z1‘ Z1 ************************************/
start z1_algorithm (u,v);
z1= trace(u*T(v)*v*T(u));
return (z1);
finish;
start z1_statistic(p,method,iter,k1,k2,k3)
global(eval_co,pre_evec_co, evec_co, eval_ca, pre_evec_ca,evec_ca,AvD);
k = ncol(p)-1; *number of SNPs p = SNPs plus outcome;
if iter = 0 then do;
create tempdata from p; *to subset for cases and controls;
append from p;
read all into p0 where (col1=0);
read all into p1 where (col1=1);
close tempdata;
call delete(work,tempdata);
```

```
co =p0[,2:ncol(p0)];
ca =p1[,2:ncol(p1)];
co2 = dprc(co,method);
ca2 = dprc(ca,method);
call eigen(eval_co,pre_evec_co,co2);
call eigen(eval_ca,pre_evec_ca,ca2);
/****weighting by eigenvalue******/
e_co_sum =  sum(eval_co);  /***sum the eigenvaluers***/ *controls;
eval_co2 =  eval_co/e_co_sum;   /***create eigenvaluer weigt by dividing by the sum***/
evec_co = pre_evec_co#T(eval_co2);  /***weight each vector by its eigenvalue wt.*****/
e_ca_sum =  sum(eval_ca); *cases;
eval_ca2 =  eval_ca/e_ca_sum;
evec_ca = pre_evec_ca#T(eval_ca2);
/****to print out whole matrix**********/
*controls on bottom  cases on top;
do i = 2 to k;
    do j =1 to (i-1);
         ca2[i,j]=co2[i,j];
      end;
   end;
AvD = ca2;
```

A.3 R statistical computing

```
library(MASS);
library(nlme);
####cross_prod function######
cross_prod <- function(blupdata) {
z <- ncol(blupdata)-1;
long <-  reshape(blupdata,idvar="ind",varying=list(names(blupdata)[1:z]), direction="long")
i <- order(long[,1],long[,2])   # provides an index
mix <- long[i,]
colnames(mix) <- c("ID","marker","y")
mix$marker<-as.factor(mix$marker)
lmeControl(returnObject=TRUE, maxIter=1000);   # lme control options
fm1 <- try(lme(y ~ marker, data = mix, random= ~1 | ID), silent=TRUE)
meancorrect <-cbind(mix,resid(fm1))
meancorrect2 <- cbind(meancorrect[,1:2],meancorrect[,4])
colnames(meancorrect2) <- c("ID","marker","residual")
wide <-  reshape(meancorrect2,idvar="ID", timevar="marker", v.names="residual", direction="wide")
k <- ncol(wide)-1; #hapk =10
pairwise_corrected <- matrix(1,k,k);  #initialize the matrix
for (v in (1:(k-1))) {
  for (w in ((v+1):k)) {
  vv <- (v+1)
  ww <- (w+1)
pairwise_corrected[v,w] <- pairwise_corrected[w,v] <- (wide[,vv]%*%wide[,ww])^2;
    } #for(j in (i+1):k)
} #end for(i in 1:(k-1))
return(pairwise_corrected);
  } # end else
```

```
}
#select case controls and compute statistic##
zz <- ncol(data);
case <- data[data$outcome==1,2:zz];  #will be case
control <- data[data$outcome==0,2:zz];
n <- nrow(data);
n2 <- nrow(data[data$outcome==0,]);
ca2 <- cross_prod(case);
co2 <- cross_prod(control);
t <- sum(diag(crossprod(ca2 - co2)));
ca2[row(ca2) > col(ca2)] <- co2[row(co2) > col(co2)];
mean_corrected = ca2;
stpermute <- numeric( )
for (i in 1:1000){
    ix0 <- sample(1:n, n2, replace=FALSE);
    ix1 <- c(1:n)[-ix0];
    co <- data[ix0,2:zz];
    ca <- data[ix1,2:zz];
    co2 <- cross_prod(co);
    ca2 <- cross_prod(ca);
tpermute <- sum(diag(crossprod(ca2 - co2)));
stpermute <-c(stpermute,tpermute);
       }  ## end else for empiracle p value
}  # end for i in
achsig <- sum(stpermute > t)/1000;
print (achsig)
sink (file="lta4h_hapk_results.csv");
cat ("blupstat","achsig", "\n",sep=',');
cat (t,achsig,sep=',', append = TRUE); sink ( );
```

Bibliography

[1] Chakravarti A. It's raining snps, hallelujah? *Nat. Genet.*, 19(3):216–217, 1998.

[2] Spanbroek R.; Grabner R.; Lotzer K.; Hildner M.; Urbach A.; Ruhling K.; Moos M.P.; Kaiser B.; Cohnert T.U.; Wahlers T.; Zieske A.; Plenz G.; Robenek H.; Salbach P.; Kuhn H.; Radmark O.; Samuelsson B.; Habenicht A.J. Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. *Proc. Natl. Acad. Sci.*, 100(3):1238–1243, 2003.

[3] Nielsen D.M.; Ehm M.G.; Zaykin D.V.; Weir B.S. Effect of two and three locus linkage disequilibrium on the power to detect marker phenotype associations. *Genetics*, 168(2):1029–1040, 2004.

[4] Weir B.S. *Genetic Data Analysis II.* Sinauer Associates, Inc., Sunderland , MA, 2nd edition, 1996.

[5] Weir B.S.; Cockerham C.C. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 4(1):105–111, 1979.

[6] Funk C.D. Leukotriene modifiers as potential therapeutics for cardiovascular disease. *Nat. Rev. Drug Discov.*, 4(8):664–672, 2005.

[7] Zhao L.; Funk C.D. Lipoxygenase pathways in atherogenesis. *Trends in Cardiovascular Medicine*, 14(5):191–195, 2004.

[8] Nielsen D.M.; Suchindran S.; Smith C.P. Does strong linkage disequilibrium guarantee redundant association results? *Genet. Epidemiol.*, 32(6):546–552, 2008.

[9] Price A.L.; Patterson N.J.; Plenge R.M.; Weinblatt M.E.; Shadick N.A.; Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, 2008.

[10] Schaid D.J. Linkage disequilibrium testing when linkage phase is unknown. *Genetics*, 166(1):505–512, 2004.

[11] Cornhill J.F.; Barrett W.A.; Herderick E.E.; Mahley R.W.; Fry D.L. Topographic study of sudanophilic lesions in cholesterol-fed minipigs by image analysis. *Arteriosclerosis*, 5(5):415–426, 1985.

[12] Cipollone F.; Mezzetti A.; Fazia M.L.; Cuccurullo C.; Iezzi A.; Ucchino S.; Spigonardo F.; Bucci M.; Cuccurullo F.; Prescott S.M.; Stafforini D.M. Association between 5-lipoxygenase expression and plaque instability in humans. *Arterioscler.Thromb.Vasc.Biol.*, 25(8):1665–1670, 2005.

[13] Zaykin D.V. Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet. Epidemiol.*, 27(3):252–257, 2004.

[14] Chung R.H.; Hauser E.R.; Martin E.R. The apl test: extension to general nuclear families and haplotypes and examination of its robustness. *Hum.Hered.*, 61(4):189–199, 2006.

[15] Crosslin D.R.; Shah S.H.; Nelson S.C.; Haynes C.S.; Connelly J.J.; Gadson S.; Goldschmidt-Clermont P.J.; Vance J.M.; Rose J.; Granger C.B.; Seo D.; Gregory S.G.; Kraus W.E.; Hauser E.R. Genetic effects in the leukotriene biosynthesis pathway and association with atherosclerosis. *Hum.Genet.*, 125(2):217–229, 2009.

[16] Kwee L.C.; Epstein M.P.; Manatunga A.K.; Duncan R.; Allen A.S.; Satten G.A. Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology*, 31:75–90, 2007.

[17] Lee M.L.; Whitmore G.A. Power and sample size for dna microarray studies. *Stat. Med.*, 21(23):3543–3570, 2002.

[18] Schaid D.J.; Rowland C.M.; Tines D.E.; Jacobson R.M.; Poland G.A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, 70(2):425–434, 2002.

[19] Clair D.; Manolescu A.; Cheung J.; Thorleifsson G.; Pasdar A.; Grant S.F.; Whalley L.J.; Hakonarson H.; Thorsteinsdottir U.; Kong A.; Gulcher J.; Stefansson K.; MacLeod M.J. Helgadottir A.; Gretarsdottir S.; St. Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a scottish population. *Nat. Genet.*, 76(3):505–509, 2005.

[20] Smith L.R.; Harrell F.E. Jr.; Rankin J.S.; Califf R.M.; Pryor D.B.; Muhlbaier L.H.; Lee K.L.; Mark D.B.; Jones R.H.; Oldham H.N. Determinants of early versus late cardiac death in patients undergoing coronary artery bypass graft surgery. *Circulation*, 84(5 Suppl.):III245–III253, 1991.

[21] Hauser M.A.; Li Y.J.; Takeuchi S.; Walters R.; Noureddine M.; Maready M.; Darden T.; Hulette C.; Martin E.; Hauser E.; Xu H.; Schmechel D.; Stenger

J.E.; Dietrich F.; Vance J. Genomic convergence: identifying candidate genes for parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum. Mol. Genet.*, 12(6):671–677, 2003.

[22] Weiss K.M.;Terwilliger J.D. How many diseases does it take to map a gene with snps? *Nat. Genet.*, 26(2):151–157, 2000.

[23] Subramanian A.; Tamayo P.; Mootha V.K.; Mukherjee S.; Ebert B.L.; Gillette M.A.; Paulovich A.; Pomeroy S.L.; Golub T.R.; Lander E.S.; Mesirov J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, 102(43):15545–15550, 2005.

[24] Helgadottir A.; Manolescu A.; Helgason A.; Thorleifsson G.; Thorsteinsdottir U.; Gudbjartsson D.F.; Gretarsdottir S.; Magnusson K.P.; Gudmundsson G.; Hicks A.; Jonsson T.; Grant S.F.; Sainz J.; O'brien S.J.; Sveinbjornsdottir S.; Valdimarsson E.M.; Matthiasson S.E.; Levey A.I.; Abramson J.L.; Reilly M.P.; Vaccarino V.; Wolfe M.L.; Gudnason V.; Quyyumi A.A.; Topol E.J.; Rader D.J.; Thorgeirsson G.; Gulcher J.R.; Hakonarson H.; Kong A.; Stefansson K. A variant of the gene encoding leukotriene a4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat. Genet.*, pages 1–7, 2005.

[25] Helgadottir A.; Manolescu A.; Thorleifsson G.; Gretarsdottir S.; Jonsdottir H.; Thorsteinsdottir U.; Samani N.J.; Gudmundsson G.; Grant S.F.A.; Thorgeirsson G.; Sveinbjornsdottir S.; Valdimarsson E.M.; Matthiasson S.E.; Johannsson H.; Gudmundsdottir O.; Gurney M.E.; Sainz J.; Thorhallsdottir M.; Andresdottir M.; Frigge M.L.; Topol E.J.; Kong A.; Gudnason V.; Hakonarson H.; Gulcher J.R.; Stefansson K. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nature Genetics*, 36(3):233–239, 2004.

[26] Coombes K.R.; Wang J; Baggerly K.A. Microarrays: retracing steps. *Nature Medicine*, 13(11):1276–1277, 2007.

[27] Gaulton K.J.; Willer C.J.; Li Y.; Scott L.J.; Conneely K.N.; Jackson A.U.; Duren W.L.; Chines P.S.; Narisu N.; Bonnycastle L.L.; Luo J.; Tong M.; Sprau A.G.; Pugh E.W.; Doheny K.F.; Valle T.T.; Abecasis G.R.; Tuomilehto J.; Bergman R.N.; Collins F.S.; Boehnke M.; Mohlke K.L. Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. *Diabetes*, 2008.

[28] Loos R.J.; Lindgren C.M.; Li S.; Wheeler E.; Zhao J.H.; Prokopenko I.; Inouye M.; Freathy R.M.; Attwood A.P.; Beckmann J.S.; Berndt S.I.; Jacobs K.B.;

Chanock S.J.; Hayes R.B.; Bergmann S.; Bennett A.J.; Bingham S.A.; Bochud M.; Brown M.; Cauchi S.; Connell J.M.; Cooper C.; Smith G.D.; Day I.; Dina C.; De S.; Dermitzakis E.T.; Doney A.S.; Elliott K.S.; Elliott P.; Evans D.M.; Sadaf Farooqi I; Froguel P.; Ghori J.; Groves C.J.; Gwilliam R.; Hadley D.; Hall A.S.; Hattersley A.T.; Hebebrand J.; Heid I.M.; Lamina C.; Gieger C.; Illig T.; Meitinger T.; Wichmann H.E.; Herrera B.; Hinney A.; Hunt S.E.; Jarvelin M.R.; Johnson T.; Jolley J.D.; Karpe F.; Keniry A.; Khaw K.T.; Luben R.N.; Mangino M.; Marchini J.; McArdle W.L.; McGinnis R.; Meyre D.; Munroe P.B.; Morris A.D.; Ness A.R.; Neville M.J.; Nica A.C.; Ong K.K.; O'Rahilly S.; Owen K.R.; Palmer C.N.; Papadakis K.; Potter S.; Pouta A.; Qi L.; Randall J.C.; Rayner N.W.; Ring S.M.; Sandhu M.S.; Scherag A.; Sims M.A.; Song K.; Soranzo N.; Speliotes E.K.; Syddall H.E.; Teichmann S.A.; Timpson N.J.; Tobias J.H.; Uda M.; Vogel C.I.; Wallace C.; Waterworth D.M.; Weedon M.N.; Willer C.J.; Wraight; Yuan X.; Zeggini E.; Hirschhorn J.N.; Strachan D.P.; Ouwehand W.H.; Caulfield M.J.; Samani N.J.; Frayling T.M.; Vollenweider P.; Waeber G.; Mooser V.; Deloukas P.; McCarthy M.I.; Wareham N.J.; Barroso I.; Jacobs K.B.; Chanock S.J.; Hayes R.B.; Lamina C.; Gieger C.; Illig T.; Meitinger T.; Wichmann H.E.; Kraft P.; Hankinson S.E.; Hunter D.J.; Hu F.B.; Lyon H.N.; Voight B.F.; Ridderstrale M.; Groop L.; Scheet P.; Sanna S.; Abecasis G.R.; Albai G.; Nagaraja R.; Schlessinger D.; Jackson A.U.; Tuomilehto J.; Collins F.S.; Boehnke M.; Mohlke K.L. Common variants near mc4r are associated with fat mass, weight and risk of obesity. *Nat. Genet.*, 40(6):768–775, 2008.

[29] Terwilliger J.D.;Weiss K.M. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann.Med.*, 35(7):532–544, 2004.

[30] Shah S.H.; Hauser E.R.; Crosslin D.; Wang L.; Haynes C.; Connelly J.; Nelson S.; Johnson J.; Gadson S.; Nelson C.L.; Seo D.; Gregory S.; Kraus W.E.; Granger C.B.; Goldschmidt-Clermont P.; Newby L.K. Alox5ap variants are associated with in-stent restenosis after percutaneous coronary intervention. *Atherosclerosis*, 201:148–154, 2008.

[31] Abecasis G.R.; Cherny S.S.; Cookson W.O.; Cardon L.R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat.Genet.*, 30(1):97–101, 2002.

[32] Evans D.M.; Marchini J.; Morris A.P.; Cardon L.R. Two-stage two-locus models in genome-wide association. *PLoS.Genet.*, 2(9):e157, 2006.

[33] Marchini J.; Donnelly P.; Cardon L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat.Genet.*, 37(4):413–417, 2005.

[34] Zondervan K.T.; Cardon L.R. The complex interplay among factors that influence allelic association. *Nature Reviews Genetics*, 5(2):89–100, 2004.

[35] Davidian M. *Simulation Studies in Statistics; ST 810A*, Spring 2005.

[36] Dwyer J.H.; Allayee H.; Dwyer K.M.; Fan J.; Wu H.; Mar R.; Lusis A.J.; Mehrabian M. Arachidonate 5-lipoxygenase promoter genotype, dietary arachidonic acid, and atherosclerosis. *N. Engl. J. Med.*, 350(1):29–37, 2004.

[37] Lohmussaar E.; Gschwendtner A.; Mueller J.C.; Org T.; Wichmann E.; Hamann G.; Meitinger T.; Dichgans M. Alox5ap gene and the pde4d gene in a central european population of stroke patients. *Stroke*, 36(4):731–736, 2005.

[38] Ogata H.; Goto S.; Sato K.; Fujibuchi W.; Bono H.; Kanehisa M. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 27(1):29–34, 1999.

[39] Zhao J.; Jin L.; Xiong M. Test for interaction between two unlinked loci. *Am. J. Hum. Genet.*, 79(5):831–845, 2006.

[40] Xu H.; Gregory S.G.; Hauser E.R.; Stenger J.E.; Pericak-Vance M.A.; Vance J.M.; Zuchner S.; Hauser M.A. Snpselector: a web tool for selecting snps for genetic association studies. *Bioinformatics*, 21(22):4181–4186, 2005.

[41] Zaykin D.V.; Meng Z.; Ehm M.G. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am. J. Hum. Genet.*, 78(5):737–746, 2006.

[42] Barrett J.C.; Fry B.; Maller J.; Daly M.J. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[43] Martin E.R.; Bass M.P.; Kaplan N.L. Correcting for a potential bias in the pedigree disequilibrium test. *Am. J. Hum. Genet.*, 68(4):1065–1067, 2001.

[44] Littell R.C.; Milliken G.A.; Stroup W.W.; Wolfinger R.D.; Schabenberger O. *SAS for Mixed Models*. SAS Institute Inc., Cary, NC, 2nd edition, 2006.

[45] Stephens M.; Smith N.J.; Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68(4):978–989, 2001.

[46] Seo D.; Wang T.; Dressman H.; Herderick E.E.; Iversen E.S.; Dong C.; Vata K.; Milano C.A.; Rigat F.; Pittman J.; Nevins J.R.; West M.; Goldschmidt-

Clermont P.J. Gene expression phenotypes of atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.*, 24(10):1922–1927, 2004.

[47] Wang T.; Ho G.; Ye K.; Strickler H.; Elston R.C. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genet. Epidemiol.*, 2008.

[48] Wang T.; Zhu X.; Elston R.C. Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am. J. Hum. Genet.*, 80(5):911–920, 2007.

[49] Chen W.M.; Erdos M.R.; Jackson A.U.; Saxena R.; Sanna S.; Silver K.D.; Timpson N.J.; Hansen T.; Orru M.; Grazia Piras M.; Bonnycastle L.L.; Willer C.J.; Lyssenko V.; Shen H.; Kuusisto J.; Ebrahim S.; Sestu N.; Duren W.L.; Spada M.C.; Stringham H.M.; Scott L.J.; Olla N.; Swift A.J.; Najjar S.; Mitchell B.D.; Lawlor D.A.; Smith G.D.; Ben-Shlomo Y.; Andersen G.; Borch-Johnsen K.; Jorgensen T.; Saramies J.; Valle T.T.; Buchanan T.A.; Shuldiner A.R.; Lakatta E.; Bergman R.N.; Uda M.; Tuomilehto J.; Pedersen O.; Cao A.; Groop L.; Mohlke K.L.; Laakso M.; Schlessinger D.; Collins F.S.; Altshuler D.; Abecasis G.R.; Boehnke M.; Scuteri A.; Watanabe R.M. Variations in the g6pc2/abcb11 genomic region are associated with fasting glucose levels. *J. Clin. Invest.*, 18(7):2620–2628, 2008.

[50] Schmidt M.; Hauser E.R.; Martin E.R.; Schmidt S. Extension of the simla package for generating pedigrees with complex inheritance patterns: Environmental covariates, gene-gene and gene-environment interaction. *Stat. Appl. Genet. Mol. Biol.*, 4(1):Article15, 2005.

[51] Hauser E.R.; Crossman D.C.; Granger C.B.; Haines J.L.; Jones C.J.; Mooser V.; McAdam B.; Winkelmann B.R.; Wiseman A.H.; Muhlestein J.B.; Bartel A.G.; Dennis C.A.; Dowdy E.; Estabrooks S.; Eggleston K.; Francis S.; Roche K.; Clevenger P.W.; Huang L.; Pedersen B.; Shah S.; Schmidt S.; Haynes C.; West S.; Asper D.; Booze M.; Sharma S.; Sundseth S.; Middleton L.; Roses A.D.; Hauser M.A.; Vance J.M.; Pericak-Vance M.A.; Kraus W.E. A genomewide scan for early-onset coronary artery disease in 438 families: the genecard study. *Am. J. Hum. Genet.*, 75(3):436–447, 2004.

[52] Hauser E.R.; Mooser V.; Crossman D.C.; Haines J.L.; Jones C.H.; Winkelmann B.R.; Schmidt S.; Scott W.K.; Roses A.D.; Pericak-Vance M.A.; Granger C.B.; Kraus W.E. Design of the genetics of early onset cardiovascular disease (genecard) study. *Am. Heart J.*, 145(4):602–613, 2003.

[53] Krzanowski WJ. Between-groups comparison of principal components. *Diabetes*, 74(367):703–707, 1979.

[54] Krzanowski WJ. Permutational tests for correlation matrices. *Statistics and Computing*, 3(367):37–44, 1993.

[55] Laird N.M.; Horvath S.; Xu X. Implementing a unified approach to family-based tests of association. *Genet.Epidemiol.*, 19(Suppl. 1):S36–S42, 2000.

[56] A. Ziegler, C. Kastner, U. Gromping, and M. Blettner. The generalized estimating equations in the past ten years: An overview and a biomedical application.

Biography

David R. Crosslin was born on August 8, 1969 in Winston-Salem, North Carolina. David and his wife (Rachel) of 13 years are blessed with two daughters - Meredith (6) and Caroline (4). He received a Bachelor of Science degree in Zoology and a minor in Genetics from North Carolina State University in December 2002. In December 1998, he received a Masters of Science degree in Comparative Biomedical Sciences and a minor in Biotechnology from North Carolina State University. After years in clinical research, David joined the Duke Clinical Research Institute in 2003. Two years later he enrolled in Duke's Computational Biology and Bioinformatics program under the direction of Dr. Elizabeth Hauser. Soon after, he joined her group as a Statistical Geneticist at the Center for Human Genetics. David will continue his research in statistical genetics with Dr. Bruce Weir at the University of Washington as a Research Scientist. Through numerous collaborations, his research has resulted in the following publications:

1. Crosslin DR, Shah SH, Nelson SC, Hayes C, Connelly JJ, Gadson S, Goldschmidt-Clermont PJ, Vance JM, Rose J, Granger CB, Seo D, Gregory SG, Kraus WE, Hauser ER. (2008) Genetic Effects in the Leukotriene biosynthesis pathway and association with atherosclerosis. *Human Genetics*, 125:217-229.

2. Shah SH, Freedman JN, Zhang L, Crosslin DR, Stone DH, Haynes C, Johnson J, Nelson D, Wang L, Connelly JJ, Muehlbauer M, Ginsburg GS, Crossman DC, Jones CJH, Vance J, Sketch MH, Granger CB, Newgard CB, Gregory SG, Goldschmidt-Clermont PJ, Kraus WE, Hauser ER (2008) Neuropeptide Y gene polymorphisms confer risk of early-onset atherosclerosis. *PLoS Genetics*

3. Voora D, Shah SH, Reed CR, Zhai J, Crosslin DR, Messer CJ, Salisbury BA, Ginsburg GS (2008) Pharmacogenetic predictors of statin mediated LDLc reduction and dose response. *Circ Cardiovasc Genet*, 1: 100-106

4. Connelly JJ, Shah SH, Doss JF, Gadson S, Nelson S, Crosslin DR, Hale AB, Lou X, Wang T, Haynes C, Seo D, Crossman DC, Mooser V, Granger CB, Jones CJ, Kraus WE, Hauser ER, Gregory SG (2008) Genetic and functional association of FAM5C with myocardial infarction. *BMC Med Genet*, :9-33

5. Shah SH, Hauser ER, Crosslin D, Wang L, Haynes C, Connelly J, Nelson S, Johnson J, Gadson S, Nelson CL, Seo D, Gregory S, Kraus WE, Granger CB, Goldschmidt-Clermont P, Newby LK (2008) ALOX5AP variants are associated with in-stent restenosis after percutaneous coronary intervention. *Atherosclerosis*, 201:148-154

6. Sutton BS, Crosslin DR, Shah SH, Nelson SC, Bassil A, Hale AB, Haynes C, Goldschmidt-Clermont PJ, Vance JM, Seo D, Kraus WE, Gregory SG, Hauser ER (2008) Comprehensive genetic analysis of the platelet activating factor acetylhydrolase (PLA2G7) gene and cardiovascular disease in case-control and family datasets. *Hum Mol Genet*, 9:1318-1328

7. Wang L, Hauser ER, Shah SH, Pericak-Vance MA, Haynes C, Crosslin D, Harris M, Nelson S, Hale AB, Granger CB, Haines JL, Jones CJ, Crossman D, Seo D, Gregory SG, Kraus WE, Goldschmidt-Clermont PJ, Vance JM (2007) Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. *Am J Hum Genet*, 4:650-663

8. Eisenstein EL, Ortiz M, Anstrom KJ, Crosslin DR, Lobach DF (2006) Economic evaluation in medical information technology: why the numbers don't add up. *AMIA Annu Symp Proc*: 914

9. Eisenstein EL, Ortiz M, Anstrom KJ, Crosslin DR, Lobach DF (2006) Assessing the quality of medical information technology economic evaluations: room for improvement. *AMIA Annu Symp Proc*: 234-238

10. Connelly JJ, Wang T, Cox JE, Haynes C, Wang L, Shah SH, Crosslin DR, Hale AB, Nelson S, Crossman DC, Granger CB, Haines JL, Jones CJ, Vance JM, Goldschmidt-Clermont PJ, Kraus WE, Hauser ER, Gregory SG (2006) GATA2 is associated with familial early-onset coronary artery disease. *PLoS Genet* 8:1265-1273

11. Eisenstein EL, Anstrom KJ, Macri JM, Crosslin DR, Johnson FS, Kawamoto K, Lobach DF (2005) Developing a framework for conducting economic evaluations of community-based health information technology interventions. *AMIA Annu Symp Proc*: 948

12. Eisenstein EL, Anstrom KJ, Macri JM, Crosslin DR, Johnson FS, Kawamoto K, Lobach DF (2005) Assessing the potential economic value of health infor-

mation technology interventions in a community-based health network. *AMIA Annu Symp Proc*: 221-225

13. Chu VH, Crosslin DR, Friedman JY, Reed SD, Cabell CH, Griffiths RI, Masselink LE, Kaye KS, Corey GR, Reller LB, Stryjewski ME, Schulman KA, Fowler VG, Jr. (2005) Staphylococcus aureus bacteremia in patients with prosthetic devices: costs and outcomes. *Am J Med* 12:1416e19-e24

14. Kauf TL, Velazquez EJ, Crosslin DR, Weaver WD, Diaz R, Granger CB, McMurray JJ, Rouleau JL, Aylward PE, White HD, Califf RM, Schulman KA (2006) The cost of acute myocardial infarction in the new millennium: evidence from a multinational registry. *Am Heart J* 1:206-212

15. Shelton CB, Crosslin DR, Casey JL, Ng S, Temple LM, Orndorff PE (2000) Discovery, purification, and characterization of a temperate transducing bacteriophage for Bordetella avium. *J Bacteriol* 21:6130-6136