KNOWLEDGE-DRIVEN GENOME-WIDE ANALYSIS OF MULTIGENIC INTERACTIONS IMPACTING HDL CHOLESTEROL LEVEL

By

Stephen Dale Turner

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May, 2011

Nashville, Tennessee

Approved:

Professor Marylyn D. Ritchie

Professor Dana C. Crawford

Professor Jonathan L. Haines

Professor Erik M. Boczko

Professor Yu Shyr

For Elizabeth.

ACKNOWLEDGEMENTS

The work I present here was greatly improved by all the coauthors on the manuscripts from which this dissertation was adapted. This work was also guided and deeply enriched by input from my thesis committee (Dr. Dana C. Crawford, Dr. Erik M. Boczko, Dr. Jonathan L. Haines, Dr. Yu Shyr, and Dr. Marylyn D. Ritchie). I would like to especially thank my Ph.D. mentor, Dr. Marylyn D. Ritchie for her unwavering support in my scientific training and career development. Dr. Ritchie has been an outstanding personal role model and has provided an exceptional training climate through her scientific advice and personal motivation. I cannot be thankful enough for the opportunities Dr. Ritchie has given me to gain outside insight and share my work with other scientists both stateside and abroad.

I also thank the members of Dr. Ritchie's lab, especially Scott Dudek and Eric Torstenson. As elite computer scientists turned human geneticists, Scott and Eric have helped me grapple with numerous algorithmic concepts, and have provided support and innovation in the design and implementation of many of the methods presented throughout this dissertation. I also extend a special thanks to Ritchie lab alumni, now bioinformatics professor, Dr. William S. Bush. As a professional mentor, Dr. Bush showed me nearly everything I know of computer programming and relational databases, and provided insight to nearly every scientific problem I've dealt with. As a personal friend, Will has accompanied me through a marathon and over glaciers. Dr. Bush helped design the work that is presented in Chapter III of this dissertation.

I would also like to thank several key members of the Center for Human Genetics Research for their contributions and support. Yuki Bradford and Justin Giles contributed to the statistical analysis and the programming support for the work presented in Chapters IV-V. Nothing forces one to truly understand a concept as having to teach it to others. Logan Dumitrescu, Ben Grady, Carrie Buchanan, and Emily Holzinger have all helped me solidify my understanding of many scientific ideas simply by asking me to explain something. I also thank

Dr. Dana Crawford, chair of my thesis committee, for her timely and insightful commentary on all the manuscripts she co-authored with me, and for her support in my professional and career development.

Finally, this work was supported in part by federal grants F31-NS066638, R01-LM010040, U01-HG04603, U01-HG004608, and T32-GM080178 from the National Institutes of Health. Sincere thanks is due to all the unnamed study participants from the Marshfield Personalized Medicine Research Project and the Vanderbilt BioVU cohorts discussed in Chapters IV-V for their contribution and willingness to participate in our research.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

INTRODUCTION

On June 25, 2000, President Bill Clinton and Prime Minister Tony Blair jointly announced the completion of the draft sequence of the human genome. The National Human Genome Research Institute reported that the completion of the Human Genome Project cost approximately $2,700,000,000 (in 1991 dollars, approximately $4,330,000,000 in 2010 dollars). As of this writing, several commercial personal genomics companies offer complete genome sequencing to anyone willing to pay approximately $40,000 – nearly six orders of magnitude cheaper than what a genome would have cost about a decade ago. The potential for incalculable profits are a driving force behind the academic and commercial development of second and third generation sequencing instrumentation, which will no doubt enable the coveted $1,000 genome within the next 5 years or less. This relentless progression of technology in combination with the collection of extremely large cohorts has revolutionized the field of disease gene research. Although whole genome sequencing in thousands of samples is still a few years away, the last few years of disease gene research have been marked by the rise of the genome-wide association study (GWAS) as a preferred method for studying the genetic architecture of common diseases. In this dissertation I focus on study design principles, quality control procedures, and new analytical procedures that can be used in genome-wide association studies. I focus on the role that epistasis, or gene-gene interaction, plays in the genetic architecture of common disease, and how we might investigate epistasis in GWAS. I conclude this dissertation with a GWAS investigating the genetic and environmental factors that impact HDL-cholesterol levels in humans – a known heritable risk factor for cardiovascular disease and a potential therapeutic target of high public health interest.

Chapter I begins this dissertation with a review of study designs and analytical methods for genetic association studies partially adapted from Turner, S. D., Crawford, D. C., & Ritchie, M. D. (2009), "Methods for optimizing statistical analyses in pharmacogenomics research," *Expert*

Review of Clinical Pharmacology, 2(5), 559-570. This chapter explores principles of study design and statistical analysis for genetic association studies, including discussions of the importance of phenotyping and choosing an appropriate study population. Traditional statistical techniques for analyzing quantitative traits are reviewed, and the chapter concludes with an examination of new and emerging methodologies for analysis beyond simple single-locus effects, including an overview of the knowledge-driven evolutionary computing concept that forms the subject of the following chapter.

Chapter II discusses the motivation, development, and *in silico* assessment of ATHENA – the Analysis Tool for Heritable and Environmental Network Associations. ATHENA is an analytical platform that allows domain knowledge from publicly available biological repositories to be incorporated into a memetic algorithm for training neural networks which utilizes both backpropagation and grammatical evolution. This chapter is adapted in part from my series of three peer-reviewed articles on the method: [1] Turner, S. D., Dudek, S. M., & Ritchie, M. D, (2010), "Grammatical Evolution of Neural Networks for Discovering Epistasis among Quantitative Trait Loci," *Lecture Notes in Computer Science*, 6023, 86-97; [2] Turner, S. D., Dudek, S. M., & Ritchie, M. D, (2010), "Incorporating Domain Knowledge into Evolutionary Computing for Discovering Gene-Gene Interaction," *Lecture Notes in Computer Science*, 6238(I), 394-403; and [3] Turner, S. D., Dudek, S. M., & Ritchie, M. D, (2010), "ATHENA: A Knowledge-Based Hybrid Backpropagation-Grammatical Evolution Neural Network Algorithm for Discovering Epistasis among Quantitative Trait Loci," *BioData Mining*, 3:5.

In Chapter III, I use genome-wide transcriptome data and genome-wide SNP data from HapMap lymphoblastoid cell lines to examine a heretofore unexplored mechanism for epistasis to affect human traits. It is common knowledge that many human traits are driven by alterations in gene expression. It is also well accepted that common genetic variation affects the expression of nearby genes (Veyrieras et al., 2008). Furthermore, a common theme argued in this dissertation and elsewhere is that epistasis is ubiquitous and affects human traits (Manolio, 2010; Moore, 2003). Combining these three ideas, is it possible that genetic variation can interact epistatically to exert a *cis*-regulatory effect on the expression of nearby genes? If so, what is the genomic and statistical structure of these epistatically interacting multilocus models? Are genes which are affected by *cis*-epistasis associated with complex human disease or morphological phenotypes? If so, how might we use this knowledge to guide the reanalysis of existing datasets, and how might we incorporate this information into the ATHENA algorithm discussed in Chapter II? These questions are explored in Chapter III, which is partially adapted from a peer-reviewed manuscript, accepted for publication in early 2011 in the proceedings of the personal genomics session of the Pacific Symposium in Biocomputing: Turner, S.D., Bush, W.S., (2011), "Multivariate Analysis of Regulatory SNPs: Empowering Personal Genomics by Considering *cis*-Epistasis and Heterogeneity."

After discussing methodology in Chapters I-II and an alternative mechanism for epistasis in Chapter III, we move to quality control and analysis of natural, biological data in Chapters IV-V. Thorough quality control (QC) is of utmost importance when analyzing high-throughput GWAS datasets. Systematic biases that result in the deviation of test statistics from the truth are amplified with the extremely large sample sizes characteristic of contemporary GWA studies. For this reason I dedicate the entirety of Chapter IV to discussing the QC procedures that were used prior to the analyses that will be presented in Chapter V. I led the eMERGE genomics working group in developing this set of procedures and reporting them in a manuscript now accepted and in press at *Current Protocols in Human Genetics*: Turner S.D., Armstrong L., Bradford Y., Carlson C., Crawford D.C., Crenshaw A.T., de Andrede M., Doheny K., Haines J.L., Hayes G., Jarvik G., Jiang L., Ling H., Kullo I., Li R., Manolio T.A., Matsumoto M., McCarty C.A., McDavid A., Mirel D., Paschall J., Pugh E., Rasmussen L.V., Wilke R.A., Zuvich R.L., Ritchie M.D., (2011), "Quality Control procedures for Genome-Wide Association Studies." The eMERGE consortium has agreed on this set of QC procedures as a set of best practices guidelines for QC of GWAS data. In addition to serving as documentation of the QC procedures used prior to the analyses presented in the next chapter, this chapter may also be used as a set of guidelines for future investigators executing quality control on GWAS data.

In Chapter V, I perform a genome-wide association study investigating genetic and environmental factors that contribute to high density lipoprotein cholesterol (HDL-C) level. Although HDL-C levels are highly heritable, the genetic determinants identified through GWAS thus far only explain a small proportion of the variance in this trait. Reasons for this discrepancy may include rare variants, structural variants, gene-environment interaction, and gene-gene interaction. The world's growing clinical practice-based biobanks now allow investigators to address these challenges, by conducting GWAS in the context of comprehensive electronic medical records. Here I apply a novel EMR-based phenotyping approach, within the context of routine care, to replicate several known associations between HDL-C and previously characterized genetic variants. Through the application of additional analytical strategies incorporating biological knowledge similar to those presented in Chapter II, I further identified 11 significant gene-gene interaction models in a discovery cohort, 6 of which show evidence of replication in a second biobank cohort. The results presented here illustrate the power of linking electronic medical records to GWAS data from biobanked samples in genetic association studies and in the elucidation of gene-gene interactions for complex human traits. This work is adapted in part from Turner S.D., Berg R.L., Linneman J.G., Peissig P.L., Crawford, D.C., Denny J.C., Roden D.M., McCarty C.A., Ritchie M.D., Wilke R.A., Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks, *PLoS ONE*, in review.

In Chapter VI, I conclude this dissertation summarizing the work presented in Chapters I-V. I will also discuss the future directions of the work presented here in light of the challenges that must be faced as technology inexorably marches forward.

CHAPTER I

REVIEW OF DESIGN PRINCIPLES AND METHODS FOR STATISTICAL ANALYSIS OF COMPLEX HUMAN QUANTITATIVE TRAITS[1]

Introduction

Much evidence exists suggesting that individual variation in complex human traits can be attributed to genetic variation (Garrod, 1902; Altshuler, Daly, & Lander, 2008; Goldstein, 2009; Hirschhorn, 2009). The last several decades have ushered in technological advances that have allowed investigators to progress from coarse genomic coverage with linkage maps and candidate gene association studies, to very high resolution association studies using single nucleotide polymorphisms (SNPs) (Risch & Merikangas, 1996). The initial completion and ongoing development of the International HapMap Project (International hapmap consortium, 2003; International hapmap consortium, 2007) catalogs common human genetic variation at millions of polymorphic sites in several diverse human populations, facilitating more powerful and strategic association study designs. Several contemporary genotyping technologies enable rapid, highly accurate genotyping of up to millions of common SNPs at low cost per genotype (Spencer, Su, Donnelly, & Marchini, 2009). The genome-wide association study (GWAS) is a widely used technique in human genetics research to investigate DNA variations associated with common human diseases (Manolio, 2010; Hardy & Singleton, 2009a). In addition to SNPs, the latest generation of GeneChips (Affymetrix) and BeadChips (Illumina) contain thousands of probes targeting known copy number variations (CNV) based on the first generation CNV maps available for the human genome (Itsara et al., 2009; McCarroll et al., 2008; Jakobsson et al., 2008; Redon et al., 2006). Finally, much research is being pursued in both the academic and private sector to develop methods for inexpensive whole-genome next-generation sequencing (Mardis,

[1] Adapted in part from: Turner, S. D., Crawford, D. C., & Ritchie, M. D. (2009). “Methods for optimizing statistical analyses in pharmacogenomics research,” *Expert Review of Clinical Pharmacology*, 2(5), 559-570.

2008; Schuster, 2008; Von Bubnoff, 2008) allowing for complete examination of all human genetic sequence variation, which will capture rare variation that is currently missed in GWAS.

In addition to benefitting from the technological advances in the field, genetic association studies can benefit from much of the recent improvements in study design and optimization of statistical methodology that has recently transformed disease-gene association studies. Here, I review the study designs and statistical methodology that are commonly used in genetic studies of quantitative trait outcomes (quality control procedures for genome-wide association studies is the subject of extensive review in Chapter IV, and are not covered in detail in this chapter). I then discuss novel methodology research taking place in the human genetics community, and how these can be applied to optimize statistical analyses in genetic research. A flowchart representation of the main topics covered in this review is shown in Figure 1.

Phenotype selection and quantitative outcomes

The predominant study design for genetic association studies over the last decade has been the case-control design. In a disease gene association study, one would typically associate an allele or genotype frequency difference between affected and unaffected individuals (Haines & Pericak-Vance, 1998). The case-control design in disease gene studies has many advantages, namely the fact that relatively rare conditions can be ascertained after their onset, significantly reducing the costs of ascertainment and follow-up typically incurred by large prospective studies (Leon Gordis, 2008). The case-control design may be the most appropriate design in genetic studies examining a purely discrete outcome such as a cancer diagnosis or an extreme adverse

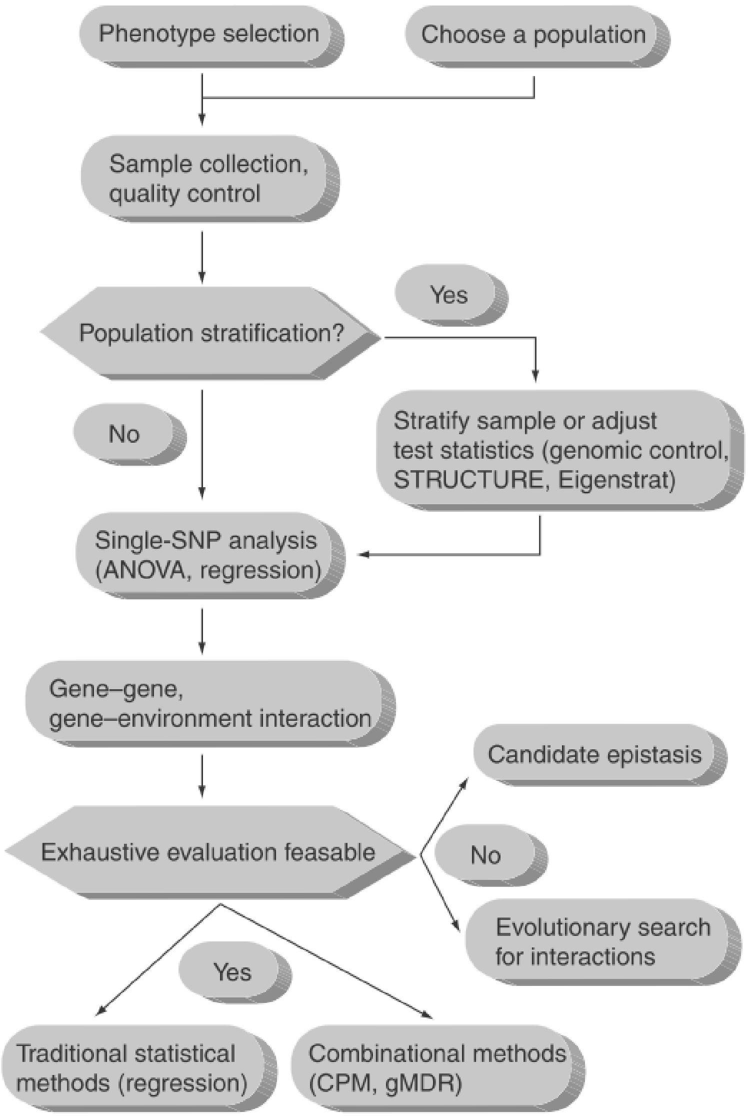

Phenotype selection
Choose a population
Sample collection, quality control
Population stratification?
Yes
No
Stratify sample or adjust test statistics (genomic control, STRUCTURE, Eigenstrat)
Single-SNP analysis (ANOVA, regression)
Gene–gene, gene–environment interaction
Candidate epistasis
Exhaustive evaluation feasable
No
Evolutionary search for interactions
Yes
Traditional statistical methods (regression)
Combinational methods (CPM, gMDR)

Figure 1. Design and statistical workflow of a genetic association study. This flowchart shows a general overview of a genetic association study with emphasis on analysis beyond single locus effects. ANOVA: Analysis of Variance; CPM: Combinatorial Partitioning Method; gMDR: Generalized Multifactor Dimensionality Reduction; SNP: Single Nucleotide Polymorphism.

drug reaction, where the phenotype neatly falls into one of two possible classes. However, analysis of a quantitative trait that varies continuously over a range of possible values may be more optimal in many cases. First, the phenotype of interest in many genetic studies naturally varies on a continuous scale. This is especially true in pharmacogenetic studies. Examples include determining the correct stable dosage for a drug with a narrow therapeutic range (Klein et al., 2009; Sills, 2005; Arranz et al., 2000), predicting treatment efficacy (Nagasubramanian, Innocenti, & Ratain, 2003; Wessels et al., 2007; Roses et al., 2007), and predicting drug resistance (Siddiqui et al., 2003). Furthermore, even many adverse drug reactions, which are typically thought of as discrete events, can be measured on a continuous scale, rather than discretized based on an often arbitrary threshold. Examples include analyzing blood iron content rather than anemic status yes/no, assaying liver enzyme activity rather than hepatotoxicity yes/no, or recording blood glucose concentration rather than hypoglycemia status yes/no. While artificially creating a discrete variable based on an arbitrary threshold in a naturally continuous trait can simplify analysis of the data, it can also be counterproductive as it comes with the cost of a dramatic decrease in statistical power (MacCallum, Zhang, Preacher, & Rucker, 2002). That is, the useful variance found in continuous outcome data is discarded in dichotomous outcome data. Therefore, creating categorical variables in such a way should be avoided as much as possible. Such genetic association study designs are common in cardiovascular genetics (Aulchenko et al., 2009; Sabatti et al., 2009; Kathiresan et al., 2009; Newton-Cheh et al., 2009; Willer et al., 2008b), genetic analysis of gene expression levels (eQTLs) (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009; Dixon et al., 2007a; Veyrieras et al., 2008; Pickrell et al., 2010), and psychiatric genetics (Meyer-Lindenberg & Weinberger, 2006). Finally, standard clinical chemistry and contemporary proteomics techniques have made collection of continuously varying biomarker traits relatively easy, accurate, and inexpensive (Cristea, Gaskell, & Whetton, 2004).

Regardless of whether the clinical outcome under investigation is a discrete or continuous endpoint, phenotype definition is crucial to optimizing statistical analysis for genetic association studies. This may be exceedingly difficult when a naturally continuous clinical outcome must be categorized into one class or another. However, even when a quantitative outcome is ascertained and recorded, care must be taken to select a continuously varying clinical feature that can be precisely defined and reliably measured. This is less challenging for some phenotypes than others. For instance, personality traits are exceedingly difficult to categorize and quantify, while circulating plasma lipid levels can easily be measured and recorded through medical record surveillance and by standard laboratory assays, respectively. Perhaps this is why the largest GWAS to date investigating personality traits found not a single significant result (Verweij et al., 2010), and why genetic factors underlying blood lipid levels have been studied more using GWAS than most other human phenotypes (Aulchenko et al., 2009; Chasman et al., 2008; Heid et al., 2008; Johansen et al., 2010; Kathiresan et al., 2007; Kathiresan et al., 2008; Kathiresan et al., 2009; Kooner et al., 2008; Sabatti et al., 2009; Sandhu et al., 2008; Saxena et al., 2007; Wallace et al., 2008; Willer et al., 2008b). Finally, thoughtful consideration must be given to choosing which clinical features will be used to represent complex disease endpoints. In addition to affecting the power of statistical analyses, the choice and specificity of phenotype definition have implications on the interpretation and reproducibility of one's results. Standard measures should be used so that others may follow up and replicate the analysis of an identical outcome in another dataset.

In summary, while the case-control design has dominated genetic association studies in recent years, continuous outcomes may be more readily ascertained in genetic association studies, and the methods designed to analyze this type of outcome (discussed below) tend to be statistically more powerful, taking into account the full range of phenotypic variability that may be influenced by genetic factors. It is important to choose an outcome that is reliably measured, and one that is a standard measure in the field so that others may easily follow up results in future studies.

Choosing a study population and methods for addressing stratification

One of the largest sources of confounding in association studies using unrelated individuals stems from population stratification, which occurs when the study population contains multiple subgroups of individuals with distinct genetic backgrounds (usually the result of including multiple racial or ethnic subgroups into a single study population). This becomes problematic and leads to systematic type I and type II errors (Marchini, Cardon, Phillips, & Donnelly, 2004) when two conditions apply: (1) The subgroups differ with respect to the frequency of an allele, and (2) The subgroups have different average values for the quantitative outcome of interest (or differ in the frequency of the occurrence of an event if the outcome under investigation is categorical). There are great differences in allele frequency and linkage disequilibrium patterns between populations worldwide (International hapmap consortium, 2003; International hapmap consortium, 2007). Others have furthermore shown that genetic variation mirrors geography within European populations alone (Novembre et al., 2008). If a truly irrelevant polymorphism was more common in the subgroup that, for some other genetic or non-genetic reason, has a higher or lower average value for the trait being studied, then the allele will appear associated with trait in this dataset. The association here would be purely artifactual, resulting from the failure to adjust for population stratification. Also, admixture and unknown and/or unintentional population stratification leads to artificially increased linkage disequilibrium across the genome or genomic regions being studied (Zhu, Tang, & Risch, 2008). While this phenomenon is used as a tool in admixture mapping, it adversely affects genetic association studies employing a tagSNP or other LD-based approach to identifying the causal genetic variant regardless of whether or not the trait of interest varies between or across the subgroups. Thus, in addition to causing excessive type I error inflation, population stratification could also obscure true genetic associations (Marchini et al., 2004).

Selecting a population to minimize stratification

One solution to this problem is to use family-based designs, which are robust to population stratification (Cardon & Bell, 2001; Cardon & Palmer, 2003). Rather than associating frequency of alleles across families, family based designs typically link or associate outcomes to regions of genetic variation by following alleles through meioses within families, which are robust to confounding by population substructure. While this has been an important design for disease gene association studies it is used less often than population-based designs because it is often more difficult to ascertain the extremely large sample sizes typically required in GWAS analysis using a family based design. Therefore, discussions of statistical procedures used for family-based study designs will not be summarized here, but they are the subject of an extensive review available in (Laird & Lange, 2006).

One strategy for avoiding bias introduced by population stratification is to ensure that study samples are drawn from a genetically homogenous population. For example, the Framingham Heart Study (Govindaraju et al., 2008) original and offspring cohorts are mainly comprised of Americans of European descent, where the most common self-reported ancestry was western European. As expected, this geographically and racially homogenous population does not display any evidence of population substructure (Wilk et al., 2005). However, it is likely that findings in one genetically homogenous population may not replicate or explain disease susceptibility variance in other ethnic groups or in the broader general population (Ioannidis, Thomas, & Daly, 2009). Furthermore, one of our goals as geneticists is to understand biology and enable personalized medicine in all humans, not just one genetically homogenous subset. Therefore, diverse population-based samples are desirable for genetic association studies focused on characterizing previous GWAS or candidate gene discoveries made in one population (Manolio, 2009). In addition to intentionally studying diverse populations, ascertainment of diverse samples may be unavoidable due to economics, ease of recruitment, and recruitment setting (such as an outpatient clinic in a diverse city). If sampling from a diverse population for a genetic association study, care should be taken to record ancestry and ethnic background of each individual in the study. While this is critical in candidate gene studies, it is less important in large-scale genetic studies because ancestry can be inferred genetically (discussed in the following section). In addition to self-report, other extensive questionnaire tools have been developed to aid in collecting information about ancestry (Lin & Kelsey, 2000). Analyses could then be carried out separately for each ethnic subgroup. However, self-reported race/ethnicity has been criticized for being an inaccurate assessment of genetic ancestry (Race Ethnicity and Genetics Working Group, 2005). Because this inaccuracy can lead to population stratification, investigators have advocated the genotyping of ancestry informative markers (AIMs) (Seldin & Price, 2008) to more accurately infer individual ancestry. Some studies have suggested that self-report can be equivalent to genetic ancestry determined by AIMs (Yaeger et al., 2008), but this is dependent on the specific markers genotyped as AIMs and the level of detectable substructure desired by the investigator.

Statistical methods for detecting and controlling for population stratification

Although the confounding effects of population stratification can be mitigated by carefully choosing a study population, it can rarely be completely eliminated. Furthermore, the confounding effects of population stratification becomes more severe as sample size increases (Pritchard & Rosenberg, 1999; Reich & Goldstein, 2001). Others have shown that even with a uniformly European sample of 3000 individuals, differences in genetic variation can be detected between populations centered in geographic areas as little as a few hundred kilometers apart (Novembre et al., 2008). Even slight differences in prevalence rates of the outcome of interest between these groups could cause spurious associations (McClellan & King, 2010). To deal with these challenges, statistical methodology has been developed and implemented into software to aid in detecting and adjusting for population stratification in genetic association studies. One method, genomic control (Devlin & Roeder, 1999; Reich et al., 2001), aims to control for population stratification by first estimating an inflation factor, then adjusting all of the test statistics downward by this factor. Several variations on genomic control have been developed, and a recent comprehensive review and critical evaluation of genomic control methods (Dadd, Weale, & Lewis, 2009) recommended genomic control F (GCF) (Devlin, Bacanu, & Roeder, 2004) as the most appropriate variation. GCF does not assume the inflation factor is measured without error, and refines this factor accordingly. Structured association (Pritchard, Stephens, & Donnelly, 2000), implemented in the STRUCTURE software (2009b), uses genotype data to infer population structure, and then performs tests of association within each inferred subpopulation. Investigators may also use STRUCTURE to identify individual samples that do not cluster with the majority of the samples. These samples may then be eliminated from the analysis. Because the risk of confounding by population stratification increases with sample size (Marchini et al., 2004), and because extremely large sample GWAS are becoming increasingly common, another method has been developed that utilizes large samples and thousands of markers throughout the genome to correct for population structure. Eigenstrat (Price et al., 2006; Patterson, Price, & Reich, 2006) uses principal components analysis (PCA) to explicitly detect and correct for population stratification on a genome-wide scale in large sample sizes in a computationally efficient manner. Eigenstrat was first described for case-control analysis but can be used for quantitative trait outcomes as well. EIGENSOFT is freely available open-source software for conducting Eigenstrat analyses, available online (2009a). A recent report using large-scale simulation studies to compare methods for correcting for population stratification examined all of the above-mentioned techniques, and found that PCA-based methods (as implemented in Eigensoft) outperformed both genomic control and structured association in terms of maximizing power, controlling type I error, maintaining in computational efficiency (Zhang, Wang, & Deng, 2008b).

Traditional statistical methods for genetic association analysis

Traditionally, the analysis of genetic factors that contribute to quantitative traits involves testing each marker individually for association to the phenotype. When the outcome of interest is a categorical trait or event, traditional analytical methods test for allele or genotype frequency differences between cases and controls. When the outcome is continuous, traditional approaches test for significant differences in the mean of the outcome of interest across different genotype classes at a given locus. Below, I outline several traditional statistical approaches for genetic association analysis when the outcome of interest varies continuously. I will discuss their strengths and weaknesses, where to find software implementations that make them accessible, and examples of how these techniques have been used in the analysis of genetic association data.

ANOVA

The analysis of variance (ANOVA) tests for significant differences in the mean value of a quantitative outcome between individuals in groups based on genotype. The theoretical justification of ANOVA has been demonstrated in numerous statistical texts (Maxwell & Delaney, 2004; Sokal & Rohlf, 1995), and its implementation is widely available in nearly every statistical computing software. ANOVA has a clear interpretation, and when its assumptions are met, it is uniformly the most powerful statistical procedure for detecting differences in a continuous outcome between groups. ANOVA also allows very specific hypotheses to be tested using linear contrasts, for example, testing the hypothesis that the homozygote for the minor allele has a decreased plasma concentration of high density lipoprotein when compared to individuals of both other genotypes.

Linear regression

Linear regression is a generalization of the analysis of variance - any analysis that can be performed in ANOVA can be performed equivalently in linear regression. Using the linear regression framework, a model can be fitted to test any specified mode of inheritance (dominant, additive, recessive). Linear regression also allows other clinical, genetic, or environmental components to be taken into account, or adjusted for, when assessing the unique effect of a genetic variant. Furthermore, linear regression allows very specific tests for both gene-gene and gene-environment interaction - a topic that I discuss at length later in this chapter. While adding more predictor variables and interaction terms to a regression equation will always improve the model fit, care should be taken to choose a model that has the added advantage of parsimony. A commonly used measure for aiding in model selection is the Akaike Information Criterion (AIC), which gauges how closely the predicted values fit the actual values, with a penalty for each predictor variable added to the model (Agresti, 1990). Once a well-fitting model has been developed, the linear regression equation can be used as a prediction equation for the value of the quantitative trait of interest as a function of genetic variants or other variables present in the equation. It is important, however, that the predictive ability of a model be tested in an independent dataset. In addition to being standard with almost any statistical computing software, linear regression is also available in PLINK (Purcell et al., 2007), an open-source software tailored specifically for the analysis of genetic data, freely available online (2009f).

Non-parametric or distribution-free methods

The analysis of variance and linear regression both have a very similar set of assumptions about the underlying distribution of data points that must hold for their estimates and standard errors to remain unbiased (Maxwell et al., 2004), namely that the outcome must follow a normal (Gaussian) distribution, and the variance of the outcome must be equal across groups with different genotypes (Maxwell et al., 2004; Sokal et al., 1995). Many quantitative traits often do not strictly adhere to these assumptions, often times being skewed, lognormal, or exponentially distributed (Kathiresan et al., 2009), or having a variance that differs dramatically across groups of subjects with different genotypes (Mushiroda et al., 2006). While the above methods are robust to small violations of these assumptions, substantial violations may warrant the use of non-parametric, or distribution-free methods. The Kruskal-Wallis procedure (Sokal et al., 1995) is a non-parametric alternative to ANOVA for testing differences in group means. The Kruskal-Wallis procedure and methods similar to it usually rely on rank statistics. While the Kruskal-Wallis procedure is robust to violations of the assumptions of ANOVA, it is not as powerful as ANOVA when its assumptions hold. The Kruskal-Wallis procedure is available in most statistical computing software, including the freely available open-source R statistical computing language (R Development Core Team, 2005).

All of the above-mentioned procedures share a feature in common: regardless of which of these is used, most analyses in GWA studies on quantitative outcomes test for differences in the average value of the trait between genotype groups one SNP at a time. Below I discuss the importance of gene-gene and gene-environment interaction, and recent developments for optimizing statistical analysis in genetic association studies that go beyond the traditional one-at-a-time approach that is most commonly used.

Rare Variation and Epistasis

Despite the dizzying pace of advances in genotyping technologies that have made GWAS accessible, we have not been able to fully take advantage of the wealth of data generated by these studies because our analytical strategies have not kept pace. As mentioned previously, the most commonly used analytical procedures for analyzing GWAS data are tests of association at a single genetic variant at a time. This approach has been arguably successful in identifying genetic variants associated with complex traits(McClellan et al., 2010; Goldstein, 2009;

Hirschhorn, 2009; Kraft & Hunter, 2009), but these variants collectively explain little of the genetic component expected based on family and twin studies (Maher, 2008; Manolio, 2010).

One potential explanation for this is the fact that current GWAS techniques are largely based on the "common disease common variant" (CDCV) hypothesis, which is largely unsupported by empirical evidence (Iles, 2008; McClellan et al., 2010). This is because most of our richest resource on human genetic variation is limited mostly to common variation (International hapmap consortium, 2003; International hapmap consortium, 2007), and because current GWAS technology is focused on providing assays for polymorphisms with high heterozygosity in populations of European descent. An alternative or perhaps supplemental hypothesis to the CDCV hypothesis is that the missing genetic component to complex traits may lie in rare variation, which is by and large not assayed by current GWAS techniques. Whole-genome sequencing technologies (Mardis, 2008; Schuster, 2008; Von Bubnoff, 2008; Metzker, 2010) are being developed as of this writing that will allow for inexpensive examination of rare variation within the next year.

In addition to rare variation, many investigators have speculated that the missing genetic component lies in gene-gene and gene-environment interactions (Maher, 2008; Manolio, 2010). Indeed, it is generally accepted that common traits are complex, and are influenced by an intricate interplay of multiple genetic and environmental exposure (Lander & Schork, 1994; Moore & Williams, 2002; Moore & Williams, 2005; Ritchie et al., 2001). This belief has been shared by biologists for over 70 years, when it was first emphasized by Sewall Wright that any biological or evolutionary endpoint is dependent on complex interactions between genes and environmental factor (Wright, 1932). It is still thought that gene-gene and gene-environment interactions are ubiquitous given the complex biomolecular interactions that are essential for regulation of gene expression and complex metabolic network (Gibson, 1996), and are likely to play a role in influencing human traits (Moore, 2003). Furthermore, while recent perspectives have emphasized the fact that most true genetic associations to complex traits carry a vanishingly small effect size (Hirschhorn, 2009; Goldstein, 2009; Hardy & Singleton, 2009b; Kraft et al., 2009), others have shown experimentally that gene-gene interactions are pervasive and often carry surprisingly large effects (Shao et al., 2008; He, Qian, Wang, Li, & Zhang, 2010).

Interactions and Quantitative Outcomes: New Approaches

Compelling evidence makes it clear that epistasis exists in humans and model organisms and influences human traits, yet there is no consensus on how to best optimize statistical analysis for investigating interactions in genetic association studies. One approach is to evaluate multi-marker combinations for potential interactive effects based on biological criteria (Carlson, Eberle, Kruglyak, & Nickerson, 2004). This may include, for instance, testing for interactions between genes that share a similar structure or function, or genes in the same pathway or biological process, such as a receptor and its ligand. Using this strategy would bias the analysis in favor of models with an established biological foundation in the literature, and novel interactions between SNPs would be missed. Furthermore, the entire analysis is conditional upon the quality of the biological information used. Another approach is to select SNPs based on the strength of their independent main effects, evaluating interactions only between SNPs that meet a certain effect size or significance threshold (Kooperberg & Leblanc, 2008). This strategy assumes that relevant interactions occur only between markers that independently have some major effect on the phenotype alone. This assumption is neither biologically nor statistically well-grounded. Biologically, compensatory mechanisms and redundancy at other loci can mitigate the effects of a devastating mutation or polymorphism at another locus, thus rendering its effect undetectable. This is evident in the many gene knockout mouse lines that show no apparent phenotype (Baba, Azuma, Kashiwabara, & Toyoda, 1994; Colucci-Guyon et al., 1994; Gorry et al., 1994; Gruda et al., 1996; Itohara et al., 1993; Killeen, Stuart, & Littman, 1992; Kneitz, Herrmann, Yonehara, & Schimpl, 1995; Zheng et al., 1996). Statistically, main effect components and interactions between them are mathematically independent effects (Maxwell et al., 2004). Furthermore, theoretical studies have shown that traits can be influenced exclusively through the interaction of two or more genetic variants (Culverhouse, Suarez, Lin, & Reich, 2002; Moore, Hahn, Ritchie, Thornton, & White, 2002), and filtering based on significant main effects would miss these types of discoveries.

Exhaustive Evaluation

A strategy to search for a gene-gene or gene-environment interaction that influences a complex trait without preconditioning on single SNP main effects is to exhaustively evaluate the relationship between the phenotypic outcome of interest and every possible combination of genetic and environmental exposures. While one may wish to fit ANOVA or linear regression models to every possible 2-, 3-, or n-way combination SNPs, this approach becomes problematic for several reasons. First, when interactions among multiple genetic and/or environmental components are considered, there are many combinations that are present in only a few individuals or perhaps none at all. This is known as the curse of dimensionality (Bellman, 1961), and results in unstable estimates of population parameters from large-sample based methods such as ANOVA and linear regression. Furthermore, while the interpretation of the statistical significance of models fit using traditional methods is fairly straightforward, correction must be made for multiple testing. Tests of interactions are large in number, and are not independent, each making multiple testing correction difficult. Also, as mentioned before, these methods are uniformly the most powerful technique for detecting differences in the mean value of an outcome, but this only holds when all assumptions are met. For many genetic studies of complex diseases, however, these assumptions are typically violated to some degree. Finally, these methods are typically the most efficient only when a mode of inheritance is specified (e.g.

dominant, recessive, additive, etc). One of the first methods proposed that would obviate some of these issues is the combinatorial partitioning method (CPM) (Nelson, Kardia, & Sing, 2000). CPM works by expanding multilocus genotype combinations and then partitioning these genotypes into groups that explain the largest proportion of variance in the quantitative trait outcome. A later improvement on this method was the restricted partitioning method (RPM) (Culverhouse, Klein, & Shannon, 2004), which does not spend valuable computing resources evaluating multilocus genotype partitions that explain little variance. A third similar approach is the generalized Multifactor Dimensionality Reduction (gMDR) (Lou et al., 2007), a variation on the widely used MDR case-control framework (Ritchie et al., 2001; Ritchie & Motsinger, 2005). An advantage of exhaustive approaches like CPM, RPM, and gMDR is that they will search through every possible multivariable model for a given dimensionality to find the optimal set of genes and environmental factors to most accurately model a quantitative outcome. These methods will report an optimal set of genes found and the amount of variance in the outcome explained by partitioning multilocus genotypes at these genes. As with any data mining technique however, care must be taken to avoid overfitting (Bishop, 2006), or "memorizing" each data point, rather than discovering the true underlying model. Cross-validation (Hastie, Tibshirani, & Friedman, 2001) is a widely employed and easily implemented technique for mitigating the risk of overfitting a model to a particular dataset. Furthermore, while the theoretical sampling distribution of test statistics for these methods is unknown, statistical significance of models discovered with these combinatorial procedures may be empirically estimated using permutation testing (Good P., 2000). Here, the null hypothesis will be empirically generated by shuffling the outcome variable values among individuals in the dataset, and running the modeling procedure many times, generating a null sampling distribution of the statistic reported by these methods (e.g. R^2). The statistic from the unpermuted analysis is compared against this null sampling distribution to estimate statistical significance. Because permutation testing requires running the modeling procedure usually thousands of times on permuted data, this can be computationally intensive even in small datasets. However, one group has recently reported a way to approximate permutation testing that requires as little as 1/50th of the time a full permutation test would require (Pattin et al., 2008).

Exhaustive approaches such as the ones mentioned above are ideally suited for exploring interactions in small genetic datasets comprised of only a few variables, such as in a candidate gene study. However, the computational resources required to exhaustively search for interactions in GWAS scale data is often prohibitive. For example, the number of two-way interactions that can be evaluated in a GWAS with 500,000 SNPs is 1.2×10^{11}. Memory issues aside, it would take many years on a desktop computer to run this analysis. This limitation is the motivation for developing techniques that still utilize the full dimensionality of the data without exhaustively searching all possible combinations of variables with the goal of discovering a well-fitting model that explains variance in a complex trait. Below I discuss three computational strategies for discovering gene-gene interactions in large-scale genetic data where an exhaustive approach would likely be computationally prohibitive.

Evolutionary Computing

Sharing many similarities with Darwinian evolution of biological organisms, evolutionary computing has been proposed as a way to discover gene-gene and gene-environment interactions that contribute to a human phenotype (Thornton-Wells, Moore, & Haines, 2004). Individuals in biological populations can be thought of as candidate solutions to a problem, where the problem in nature is to survive and reproduce. Individuals that are more fit will be selected, and their genes will be propagated in future generations in the population. By analogy, evolutionary computing commences by defining a population of candidate solutions to a problem, where the problem is to find a model containing influential genes that can explain a large proportion of variance in the outcome. "Individuals" are candidate solutions, i.e. mathematical models containing genetic and environmental variables attempting to explain variance in the outcome of interest in the study. The candidate solutions that explain more variance in the outcome are the models that contain combinations of variables that truly influence the phenotype, and these models are selected and reproduced in subsequent generations of evolutionary computing. In addition, after this phase of selecting the "most fit" individuals, models may be "mutated" (switching one genetic variant out for another in the dataset), or undergo "recombination" with another well-fitting model. Evolutionary computing can be thought of as a pattern recognition, or machine learning approach for discovering complex genetic models that influence a trait. Evolutionary computing has been used extensively in other disciplines to model complex processes (Hung & Adeli, 1994; Lee, 1996; Likartsis, Vlachavas, & Tsoukalas, 1997; Yang, Kao, & Horng, 1996; Zhang, Sankai, & Ohta, 1995; Belew, McInerney, & Schraudolph, 1990; Chen & O'Connell, 1997; Topchy & Lebedko, 1997; Cantu-Paz & Kamath, 2008; Skinner & Broughton, 1995; Yan, Zhu, & Hu, 1997). In addition to using evolutionary computing for genetic association studies (Motsinger, Lee, Mellick, & Ritchie, 2006; Motsinger, Dudek, Hahn, & Ritchie, 2006; Motsinger, Reif, Dudek, & Ritchie, 2006; Motsinger, Reif, Fanelli, Davis, & Ritchie, 2007; Motsinger-Reif, Dudek, Hahn, & Ritchie, 2008; Motsinger-Reif, Fanelli, Davis, & Ritchie, 2008; Ritchie, White, Parker, Hahn, & Moore, 2003; Ritchie & Coffey, 2004), evolutionary computing has been used in other biological applications including microarray analysis (Huang, Liu, & Xu, 2008), cancer classification (Mukhopadhyay, Maulik, & Bandyopadhyay, 2009), molecular docking (Tavares, Mesmoudi, & Talbi, 2009), and protein folding (Vullo, Passerini, Frasconi, Costa, & Pollastri, 2008). A team of leaders in the field have recently prepared a book (Poli, Langdon, & McPhee, 2008), giving an overview of genetic programming (a widely used type of evolutionary computing), available online (2009c).

Candidate Epistasis

Another recently described strategy, the Biofilter (Bush, Dudek, & Ritchie, 2009), combines a bioinformatics approach with traditional statistical hypothesis testing. Several years ago, when the genome was too large to fully interrogate with genotyping, disease gene mapping investigators relied on the candidate gene study design. Here, candidates were selected for genotyping based on their hypothesized biological function, and statistical tests were carried out on a SNP-by-SNP basis, as described previously. Now, with the advent of GWAS and the impending arrival of inexpensive whole-genome sequencing, assaying human variation across the entire genome is no longer the issue it was in the past. While millions of SNPs can be tested one-by-one for association to a trait, the interactome is too large for us to fully investigate, as described above. The approach taken in (Bush et al., 2009) reduces the interaction search space by assessing specific combinations of genetic variants based on prior statistical and biological knowledge. The method creates multi-SNP models based on information from publicly available bioinformatics data sources that can then be straightforwardly tested using logistic or linear regression.

Expert Knowledge guided Evolutionary Computation.

Finally, there is much interest in the computer science community to develop strategies for incorporating expert knowledge into evolutionary computation (Moore & White, 2007). As mentioned above, it is theoretically possible and likely in some cases that a trait is influenced exclusively by the interaction two or more genetic variants, with neither genetic variant having a main effect by itself. This represents the worst-case scenario for an evolutionary method. In fact, it has been shown that evolutionary methods perform little better than randomly testing models for association with the outcome when the underlying model is purely epistatic (White, Gilbert, Reif, & Moore, 2005). However, supplementing an evolutionary procedure with expert knowledge has been shown to increase the statistical sensitivity of evolutionary methods for finding these difficult-to-model interactions (Moore, Barney, & White, 2008; Greene, White, & Moore, 2007; Greene, White, & Moore, 2009). In these reports, the authors used a data-driven approach, relying upon prior statistical expert knowledge as a result of preprocessing the data.

The notion presented in (Bush et al., 2009) suggests a different approach, where expert knowledge is gleaned extrinsically, without any data analysis or preprocessing. Here, multi-gene groupings were created based on representation in publicly accessible biological databases, such as the Gene Ontology (Ashburner et al., 2000), or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000). Multi-SNP models in from these gene groupings were then prioritized in analysis. A very promising approach involves combining a bioinformatics approach such as this with evolutionary computation, allowing investigators to take advantage of the many decades of biomedical research to guide a machine learning procedure. Furthermore, while the above-mentioned bioinformatics approach is gene-centric, incorporation of these principles into a stochastic evolutionary procedure would allow for discovery of gene-gene interactions between genetic variants that may not be in gene regions (e.g. an interaction between variants in a micro-RNA and its target). This is the motivation for incorporating domain knowledge into grammatical evolution for discovering multi-locus genetic models that influence quantitative traits, as described in Chapter II.

Conclusions

I have presented here study design strategies and statistical methodologies for optimizing statistical analysis in genetic association studies. Careful consideration of the phenotype under study, the population in which the study is carried out, and the procedures used to model genetic influences are all equally important for achieving maximum statistical power and breadth of interpretation. Irrespective of the above-mentioned considerations, one of the most important aspects of genetic association studies is replication in an independent sample and/or functional studies. The NCI-NHGRI Working Group on Replication in Association Studies recently established recommendations for bona fide replication of GWAS result (Chanock et al., 2007a). Basic conditions for a successful replication include a sufficient sample size to replicate the genetic effect size estimated in the discovery dataset, an independent replication set, the same outcome phenotype for both data sets, a similar study population, similar direction of effect from the same SNP or a SNP in near perfect LD, a consistent genetic model, and adequate reporting of replication study design and analysis. Replication of a multi-SNP model presents new challenges, and how to effectively test for replication of higher-order models remains an open question in the field of human genetic epidemiology.

Several years have elapsed since the advent of genome-wide association studies, bringing several success stories as well as many disappointments. We must ensure that our strength in study design and optimal statistical methodology keeps pace with the relentless progression of genotyping and sequencing technology so that we may reap the benefits of the wealth of data we will soon face. A characteristic that most bleeding-edge statistical methodologies have in common is that they often abandon the simple approach of considering one genetic variant in isolation when modeling the etiology of complex phenotypes. Methods exploiting existing domain knowledge are likely one of many solutions required for the challenging task of properly mining large genomic datasets to identify all variation having an impact on human health.

Acknowledgements

The work presented in Chapter I was supported in part by National Institutes of Health (NIH) Pharmacogenetics Research Network (PGRN) Pharmacogenomics of Arrhythmia Therapy U01 (HL65962) and the Training Program on Genetic Variation and Human Phenotypes training grant (5T32GM080178-02).

CHAPTER II

ATHENA: A KNOWLEDGE-BASED HYBRID BACKPROPAGATION-GRAMMATICAL EVOLUTION NEURAL NETWORK ALGORITHM FOR DISCOVERING EPISTASIS AMONG QUANTITATIVE TRAIT LOCI[2]

Introduction

Genome-Wide Association Studies and Epistasis

As discussed in Chapter I, the genome-wide association study (GWAS) is a widely used technique in human genetics research to investigate DNA variations associated with common human diseases. Modern genotyping technology allows us to quickly and inexpensively assay millions of SNPs, but our analytical strategies have not kept pace with technological advances in the genotyping lab. The most commonly used analytical procedures for analyzing GWAS data are straightforward tests of association examining one SNP at a time. This approach has been somewhat successful in identifying genetic variants associated with complex traits, including age-related macular degeneration (Klein et al., 2005), type II diabetes (Frayling, 2007), hypertension (Newton-Cheh et al., 2009), and blood cholesterol levels (Kathiresan et al., 2009; Willer et al., 2008b), among others (Hindorff et al., 2009; Johnson & O'Donnell, 2009). With the possible exception of the CFH gene in AMD, however, these single SNPs collectively explain little of the genetic contribution to the trait variance that is expected based on family and twin studies (Maher, 2008; Goldstein, 2009). For instance, HDL-cholesterol level is highly under genetic control - up to 73% of variation in HDL can be explained by genetic factors (Pietilainen et al., 2009) - yet even the largest GWAS meta-analysis to date with over 100,000 samples found that

[2] Adapted in part from a series of three peer-reviewed articles on the method: [1] Turner, S. D., Dudek, S. M., & Ritchie, M. D, (2010), "Grammatical Evolution of Neural Networks for Discovering Epistasis among Quantitative Trait Loci," *Lecture Notes in Computer Science*, 6023, 86-97; [2] Turner, S. D., Dudek, S. M., & Ritchie, M. D, (2010), "Incorporating Domain Knowledge into Evolutionary Computing for Discovering Gene-Gene Interaction," *Lecture Notes in Computer Science*, 6238(I), 394-403; and [3] Turner, S. D., Dudek, S. M., & Ritchie, M. D, (2010), "ATHENA: A Knowledge-Based Hybrid Backpropagation-Grammatical Evolution Neural Network Algorithm for Discovering Epistasis among Quantitative Trait Loci," *BioData Mining*, 3:5.

collectively only ~12% of this variance could be accounted for by the additive effects of 58 highly significant single-SNPs (Teslovich et al., 2010). Many agree that a portion of this "missing heritability" likely lies in gene-gene and gene-environment interactions (Cordell, 2009; Maher, 2008; Manolio et al., 2009).

Grammatical Evolution Neural Networks (GENN) and Domain Knowledge: ATHENA

Neural networks (NNs) are a robust and flexible modelling technique that attempt to mimic the basic structure and function of biological neurons to solve complex problems. NNs have been applied to many research fields, including robotics, speech recognition, optical character recognition, task scheduling, and industrial processing among many others. NNs have also been widely applied to various problems in biological science, including microarray data analysis (Linder, Richards, & Wagner, 2007), genotype calling (Huang et al., 2004; Shen et al., 2005), human linkage analysis (Lucek, Hanke, Reich, Solla, & Ott, 1998), genetic association studies (Ott, 2001), medical expert systems (Porter & Crawford, 2003), survival analysis (Sato et al., 2005), and protein folding (Meiler & Baker, 2003). The conventional approach for applying NNs to a classification problem is to specify a network architecture, select which variables (SNPs) are included as inputs to the network, and fit network weights using a gradient-descent based approach such as backpropagation (BP) (Bishop, 1995). While BP is capable of quickly fine-tuning weights in a NN, variable selection and modelling are goals which cannot be accomplished using this traditional approach. Recently, numerous evolutionary search strategies have been applied to NN classification problems to reduce the issues associated with the traditional NN approach (Yao, 1999). Using evolutionary computing to model complex processes was discussed briefly in Chapter I. Genetic Programming Neural Networks (Ritchie et al., 2004) and Grammatical Evolution Neural Networks (GENN) (Motsinger-Reif et al., 2008) use genetic programming (Koza & Rice, 1991) or grammatical evolution (GE) (O'Neil & Ryan, 2003) to evolve populations of neural networks for human genetics classification problems. These populations are a heterogeneous mix of architectures, weights, and input variables that undergo mating, crossover, and recombination to ultimately identify an optimum NN solution, simultaneously finding influential SNPs and fitting networks weights. We have furthermore recently shown that certain features characteristic of human genetic data (linkage disequilibrium) may provide advantages to methods that evolve NNs to detect gene-gene interactions by transforming the fitness landscape from a "needle in a haystack" to a broader, smoother surface (Turner, Ritchie, & Bush, 2009).

The application of GE to find epistatic gene-gene interactions is still exceedingly difficult, especially when the underlying disease model is purely epistatic, where each variant has no independent effect on the phenotype (White et al., 2005). After demonstrating the critical need for expert knowledge when applying genetic programming to GWAS (Moore et al., 2007), others have shown that using expert knowledge guided mutation, selection, and crossover is highly beneficial, and dramatically improves the performance of evolutionary algorithms (Moore et al., 2008; Greene et al., 2007). In much of the previous work showing that expert knowledge increases the performance of natural computing algorithms for finding epistatically interacting SNPs, the statistical expert knowledge was gleaned intrinsically - typically using a data-driven approach using variants of the Relief algorithm for feature selection (Greene et al., 2007; Moore, Andrews, Barney, & White, 2008; Greene, Gilmore, Kiralis, Andrews, & Moore, 2009).

Here we extend our previous work with NN training (Turner, Dudek, & Ritchie, 2010b) to evaluate several fundamental modifications to the algorithm in a new tool, ATHENA (the Analysis Tool for Heritable and Environmental Network Associations). First, we implemented an alternative tree-based GE crossover strategy as previously described (Motsinger, Hahn, Dudek, Ryckman, & Ritchie, 2006; Turner et al., 2010b). A potential weakness of GE is the destructive single-point binary crossover (SPBXO) operator (O'Neil et al., 2003). Tree-based crossover (TBXO) instead swaps functionally analogous branches by first translating the grammar into functional neural network trees, identifying branches with identical root nodes, then initializing a crossover back at the genome level that would correspond to the crossover between the whole branches. This renders GE to be much more like genetic programming (GP), while still maintaining some of the key advantages of GE. We also evaluate the performance improvement when we combine GE with the traditional approach of fitting network weights with backpropagation. Finally, we evaluate with simulation whether utilizing available biological domain knowledge gleaned extrinsically would increase ATHENA's performance in discovering epistatic interactions between genetic variants contributing to a quantitative trait outcome. Here we present results of a simulation study showing that (1) using an alternative crossover strategy (TBXO) results in a considerable performance increase in some scenarios, (2) a hybrid backpropagation-GENN training algorithm has better performance than GE alone, and (3) incorporating biological knowledge from external sources results in an increase in ATHENA's ability to detect and model gene-gene interactions among a large pool of unassociated noise variables.

Methods

Genetic data simulation with genomeSIMLA

Simulated data where the true identity and size of the genetic or environmental effect in the population is known is a necessity for developing and testing novel methodology. It is also important that these true effects are embedded in a dataset containing many other nonfunctional polymorphisms and environmental factors, as is the case when real genetic data is collected. We developed genomeSIMLA (Edwards et al., 2008) for simulating genome-wide scale data in population based case-control samples with a categorical outcome. Here we use an extension of genomeSIMLA capable of simulating gene-gene interactions in the presence of main effects, all of which influence a quantitative trait at a desired effect size (Turner et al., 2010b). The genomeSIMLA source code and binaries can be downloaded freely online (2009d).

Whereas the common measure of effect size in genetic association studies employing a case-control design is the odds ratio, studies of continuously distributed outcomes, such as HDL cholesterol level, estimate effect size as the proportion of variance explained (Willer et al., 2008b; Boerwinkle & Sing, 1986), or R^2. This variance explained, or heritability, can be further divided into genetic and nongenetic components, and the genetic component can be further divided into additive, dominant, and epistatic variance components (Abney, McPeek, & Ober, 2001). The variance component explained uniquely by a single source of genetic variation (e.g. the main effect of one member of an interacting pair of variants, or the epistatic effect of the interaction term) is given by the semi-partial squared correlation coefficient (Cohen, Cohen, West, & Aiken, 2002):

$$sr_i^2 = R^2_{Y.x_1,x_2,...x_i,...x_k} - R^2_{Y.x_1,x_2,...(x_i),...x_k}$$

The first term on the right side of the equation is the overall variance explained by fitting a full model (regressing the outcome, *y*, on each main effect and the interaction term between them. The second term on the right-hand side is the proportion of variance explained by the model when a predictor variable of interest is omitted from the model – for instance, omitting the interaction term. The difference between these two quantities is the semi-partial squared correlation coefficient (Cohen et al., 2002), and describes the unique impact on the phenotype, *y*, for the particular variance component, x_i. These estimates do not take into account the bias corrections discussed by Boerwinkle and Sing (Boerwinkle et al., 1986). As these investigators showed, the bias in these estimators for the number of genotype classes represented here quickly approaches zero as sample size increases past n=100. Since our simulated datasets comprise 2000 samples, the bias discussed by these investigators is essentially zero.

Datasets were simulated as previously described (Turner et al., 2010b): Samples are drawn from a homoscedastic normal distribution with the mean being determined by the genotypes at the corresponding functional genetic variants. We simulated 500 SNPs in 2000 samples, where only two SNPs were functional and the other 498 SNPs were unassociated "noise" variables. We simulated a gene-gene interaction between these two SNPs that carried a narrow-sense heritability (h^2) of 0.05, meaning that only 5% of the variation in the quantitative trait could be explained by this gene-gene interaction. We simulated this interaction in the context of very small main effects at each locus (h^2=0.01). This low effect size is typical of most findings in human genetic epidemiology (Hirschhorn, 2009; Goldstein, 2009). Both main effects and the gene-gene interaction were additive. A scenario such as this where main effects explain little of the overall outcome variance represents a very difficult problem (Freitas, 2001) for an evolutionary search procedure to model.

Domain knowledge

A recently developed tool called Biofilter is capable of integrating information from several publicly available biological databases to assess specific combinations of genetic variations and their effect on the outcome based on prior statistical and biological knowledge (Bush et al., 2009). Specifically, this tool uses the Gene Ontology (Ashburner et al., 2000), the Database of Interacting Proteins (Xenarios et al., 2002), the Protein Families Database (Bateman et al., 2004; Finn et al., 2008), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2000), Reactome (Vastrik et al., 2007), NetPath (2008b), and the Genetic Association Database (GAD) (Becker, Barnes, Bright, & Wang, 2004) to construct two-SNP models that are supported by the biological literature. Their degree of support in the literature is characterized by an implication index - which is a count of how many times a relationship between a pair of two genes appears across multiple databases incorporated into Biofilter.

To determine whether incorporation of domain knowledge into NN training in ATHENA can improve its performance, simulated domain knowledge that mimics information obtained from Biofilter must be generated. Here, 4000 random undirected edges are drawn between a subset of the 500 SNPs simulated as described above. The implication index is the number of edges drawn between two models. This number typically ranges from 0 to 5, where implication index of zero indicates no support in the simulated knowledge pool, while an implication score of 5 indicates that this model is very well supported. The implication index corresponding to the functional two-SNP model where the true effect was embedded could be manually specified. Our specific goals were to determine if and to what degree ATHENA's performance would diminish if irrelevant domain knowledge were incorporated, and if and to what degree ATHENA's performance would increase if accurate domain knowledge were incorporated into the training process.

Alternative crossover, backpropagation, and incorporation of domain knowledge in ATHENA

NN training in ATHENA has been implemented as previously described (Motsinger-Reif et al., 2008; Turner et al., 2010b). Briefly, grammatical evolution (GE) is a variation of genetic programming (GP), an evolutionary algorithm originally proposed by Koza as a procedure to optimize NN architecture (Koza et al., 1991). In GE, randomly initialized binary strings are transcribed into an ordered list of integers that are used to select from production rules in a Backus-Naur form grammar. Our grammar applies GE to construct neural networks, and can simultaneously select important predictor variables and optimize network weights and architecture. We also implemented an alternative tree-based GE crossover strategy as previously described (Motsinger et al., 2006; Turner et al., 2010b). A potential weakness of GE is the destructive single-point binary crossover (SPBXO) operator (O'Neil et al., 2003). Tree-based crossover (TBXO) instead swaps functionally analogous branches by first translating the grammar into functional neural network trees, identifying branches with identical root nodes, then initializing a crossover back at the genome level which would correspond to the crossover between the whole branches. This renders GE to be much more like genetic programming (GP), while still maintaining some of the key advantages of GE. Representative NNs produced by GE, and the TBXO process are shown in Figure 2, under the "TBXO" panel. The NNs in this figure have either two or three inputs, corresponding to numerically coded values (-1, 0, 1) for SNP genotypes (Holzinger et al., 2010). A weight vector corresponds to each layer of weights in the NN. In TBXO, functionally analogous branches are crossed over, indicated by the asterisk, resulting in a 2-2-1 neural network (Sprinkhuizen-Kuyper & Boers, 1998) with SNPs 1 and 2 as inputs. If SNPs 1 and 2 are the functional SNPs responsible for the gene-gene interaction and if the weight vectors on this NN are favorable, then this NN should be capable of modelling a gene-gene interaction between these two SNPs.

In the first set of experiments, ATHENA was run for 100, 200, and 400 generations, in runs consisting of population sizes 100, 200 and 400, in each of 10 demes (populations), for a total NN population size of 1000, 2000, and 4000 respectively, on 100 simulated datasets. We varied the number of generations where tree-based crossover was used (TBXO). This could range from using single-point binary crossover (SPBXO) for every generation (i.e. no TBXO), TBXO for the first half of the total number of generations before switching back to SPBXO, or TBXO for the total number of generations run. This resulted in trials using 54 different combinations of ATHENA parameters, comprising 5,400 *in silico* datasets. The mean runtime per dataset was approximately 14 minutes spread across five 1.8 GHz Opteron PCs. The respective probability of a crossover and mutation were 0.9 and 0.01, typical values for these parameters in many genetic algorithms (Poli et al., 2008). Addition was the only production rule available for the arithmetic operator at each activation node, as described previously (Holzinger et al., 2010; Turner et al., 2010b). Both grammars are shown in Table 1. This allowed for the implementation and optional usage of backpropagation (BP), a local fitting procedure designed to optimize the weights in a neural network (Bishop, 1995). BP was either not used at all, or used at initialization and again at generations 100 and 200, using a learning rate of 0.3. BP was halted after either a maximum of 100 epochs (iterations) had been run, or when further BP showed no improvement (mean squared

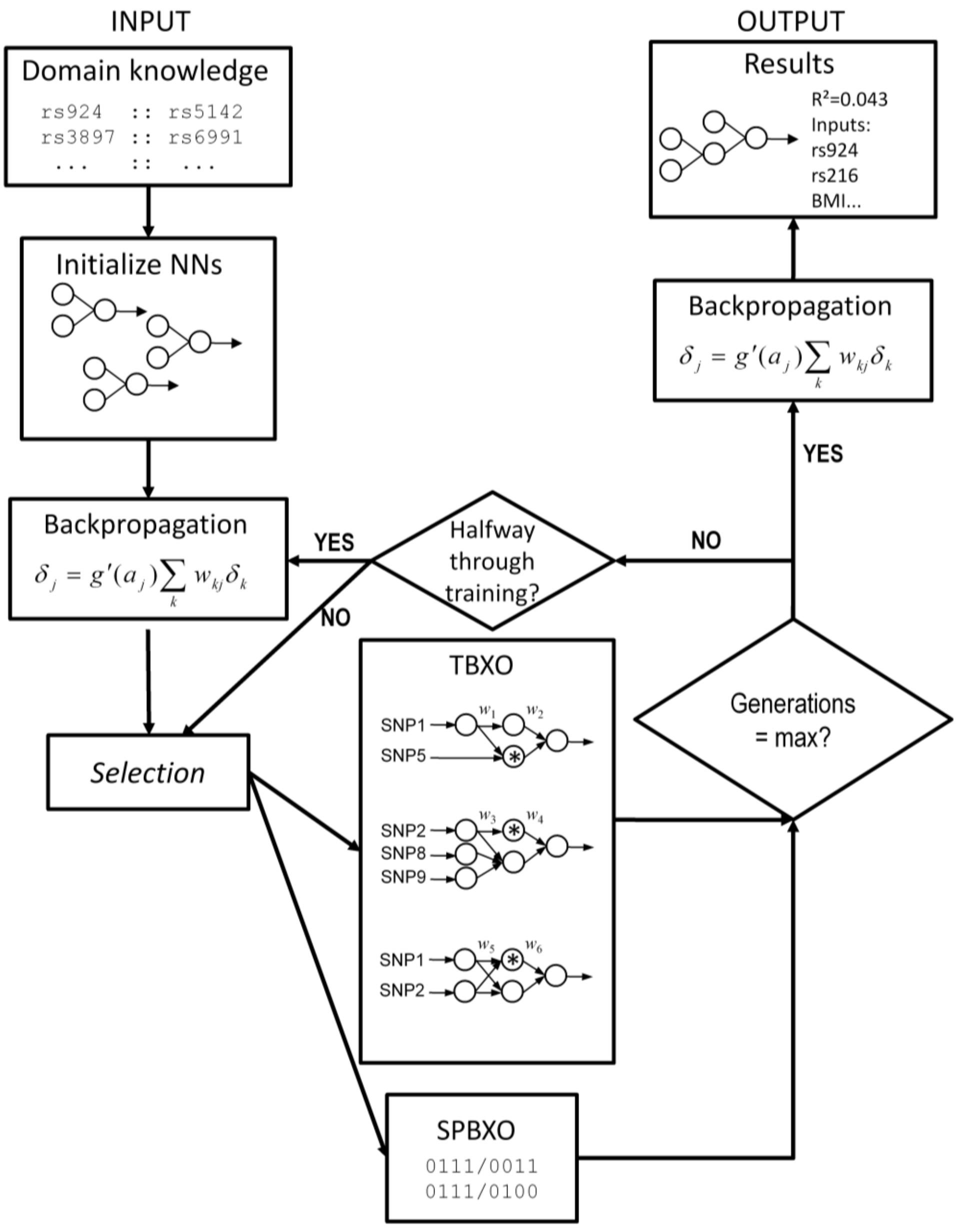


Figure 2. ATHENA algorithm. The ATHENA algorithm begins by optionally accepting a list of SNP-SNP models which are derived from biological knowledge sources. This domain knowledge is used to initialize a proportion of the NN population. BP is used to optimize the initial weights. After a round of selection, GE is used to simultaneously optimize variable selection, NN architecture, and weights. Another round of BP takes place midway through training, and at the end of training. Crossover can occur via single point binary (SPBXO) or tree-based crossover (TBXO). In SPBXO, crossover occurs at the binary string level, but in TBXO, NNs are first translated, and crossover occurs at the binary genome level that results in a crossover at functionally similar root nodes. The NNs in this figure have either two or three inputs, corresponding to numerically coded values (-1, 0, 1) for SNP genotypes. A weight vector corresponds to each layer of weights in the NN. In TBXO, functionally analogous branches are crossed over, indicated by the asterisk, resulting in a 2-2-1 neural network with SNPs 1 and 2 as inputs. If SNPs 1 and 2 are the functional SNPs responsible for the gene-gene interaction and if the weight vectors on this NN are favorable, then this NN should be capable of modelling a gene-gene interaction between these two SNPs.

Table 1. GENN Grammars. Abbreviated grammars for neural networks as currently implemented with all four arithmetic functions as production rules, and with addition as the only production rule.

Current Grammar	Addition only
<p> ::= <pn>(<pinput>)	<p> ::= <pn>(<pinput>)
<pn> ::= PA	***<pn> ::= PA***
| PS	
| PM	***# Protected addition only fn***
| PD	***available.***
<pinput> ::=	
<W>(<winput>)<,><W>(<winput>)<,>	<pinput> ::=
<winput> ::= <cop><,><v>	<W>(<winput>)<,><W>(<winput>)<,>
| <cop><,><p>	<winput> ::= <cop><,><v>
<cop> ::= (<cop><op><cop>)	| <cop><,><p>
| <Concat>(<num>)	<cop> ::= (<cop><op><cop>)
<Concat> ::= Concat	| <Concat>(<num>)
<,> ::= ,	<Concat> ::= Concat
<W> ::= W	<,> ::= ,
<op> ::= +	<W> ::= W
| -	<op> ::= +
| *	| -
| /	| *
<num> ::= <dig>1	| /
| <dig>.<dig>3	<num> ::= <dig>1
| .<dig><dig>3	| <dig>.<dig>3
| <dig>.<dig><dig>4	| .<dig><dig>3
|	| <dig>.<dig><dig>4
<dig><dig>.<dig><dig>5	|
<dig> ::= 0-9	<dig><dig>.<dig><dig>5
<v> ::= G1-G500	<dig> ::= 0-9
	<v> ::= G1-G00

error is reduced by less than 1×10^{-6}), after which the GE process continues. After every network had undergone BP, NNs were reverted back to a binary genome by marking blocks of codons (integer sequences) corresponding to a weight, which was then replaced with a block containing a grammar compatible block that generates the appropriate weight when GE continues after BP.

In the second set of experiments, domain knowledge was used to perform sensible initialization (O'Neil et al., 2003). Rather than initializing a population of NNs randomly, the initial generation is partially composed of NNs containing as input variables SNPs that are represented in a domain knowledge source. This source can be two-SNP models supported by biological literature derived from Biofilter (Bush et al., 2009) or simulated domain knowledge which mimics domain knowledge derived from Biofilter. Part of the population is still initialized randomly. Here the proportion of the initial population which is initialized from domain knowledge was varied from 0 to 99% in intervals between 1-10%. Two-SNP models from domain knowledge are prioritized for incorporation in the initial generation based upon implication index - models with higher implication index are initialized first. The implication index on the functional two-SNP model in these experiments ranged from 0 (negative control - all domain knowledge incorrect/irrelevant) to 3 (functional two-SNP model is somewhere in the top half of the implication index-ranked list of 4000 domain knowledge two-SNP models). Here, ATHENA was run for 200 generations using 10 demes with population sizes of either 50, 100, or 200 individual NNs. The mean runtime per dataset was approximately 6 minutes spread across five 1.8 GHz Opteron PCs. As above, probability of a crossover and mutation were 0.9 and 0.01, and the production rule for the activation node function was restricted to addition only, which allowed for the optional use of backpropagation. In addition to locally fitting weights to improve the model fit when NNs are initialized randomly, this hybrid algorithm allows for weight optimization in the event that sensible initialization from domain knowledge resulted in the inclusion of either of the two functional variables in the initial generation.

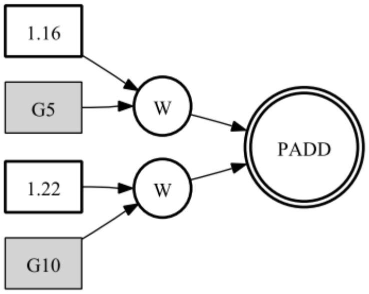
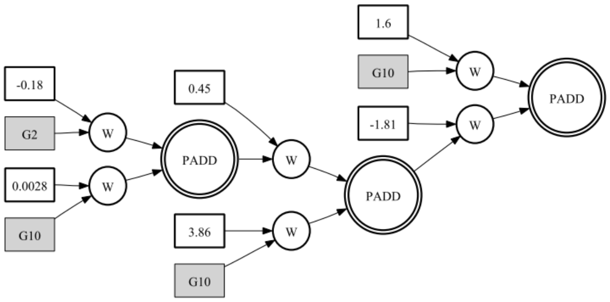
Results & Discussion

For the *in silico* studies described above, sensitivity was measured as the proportion of datasets out of 100 simulated datasets for each scenario where the best performing neural network model contained the two functional SNPs, with no other SNPs in the model, i.e. a perfect match (Turner et al., 2010b). For these experiments we were only interested in measuring performance in variable selection. Output from a neural network model using SNPs as the only inputs are necessarily discrete, as the output is a nonlinear transformation of a weighted sum of the inputs. Example neural network models are shown in Table 2. The output of each neural network model is the predicted value of the mean of the multilocus genotype. While constructing well-fitting models that can predict genotype means reliably is important, here we only assessed performance by examining variable selection. The best neural network model for each dataset was chosen using the following algorithm. First, 5-fold cross-validation (CV) was implemented. The data is divided into fifths, training initially occurs on four fifths of the data where a best model is chosen based on minimizing mean square error. The fit of this model to unseen data was tested on the fifth of the data initially left out using the standard coefficient of determination, R^2. This process was repeated for each CV interval, i.e., each 4/5-1/5 split of the data. At this point there are 5 models - one best model from each CV interval. The model that consistently appears most often across CV intervals is chosen as the best overall model for the entire dataset (Moore, Parker, Olsen, & Aune, 2002; Ritchie et al., 2001). In case of a tie (e.g. two different models replicated across two CV intervals), the model with the higher R^2 is chosen as the overall best model.

Tree based crossover (TBXO)

First we wanted to evaluate whether the alternative TBXO strategy described in the methods section resulted in increased performance in the context of GE alone or with the hybrid

Table 2. Example neural network models. The table below shows two representative neural network models that were produced by grammatical evolution, showing both text and graphical representations of the same models. PA/PADD=protected addition; W=weight, G=variable number. The first, simpler model weights variable number 5 by 1.16, variable number 10 by 1.22, and adds these weighted input variables prior to transforming the output with the nonlinear transfer function. The resulting output will be discrete if input variables are discrete, as with SNP genotypes.

Model (text representation)	Model (graphical)
`PA( W(1.16,G5), W(1.22,G10),2)`	
`PA( W(1.6,G10), W((0-1.81), PA( W(0.45, PA( W((0-0.18),G2), W((0.28*0.01),G10),2)), W(3.86,G10),2)),2)`	 1.6 G10 W -0.18 0.45 PADD -1.81 W G2 W PADD W 0.0028 W PADD G10 3.86 W G10

BP-GE algorithm in ATHENA which also used backpropagation (BP) in addition to GE. These results are summarized in Figure 3. Separate panels show the total number of generations and the size of the population in each deme. Dashed and solid lines show the performance (sensitivity) when BP was and was not used, respectively. The horizontal axis on each panel shows the proportion of the total number of generations in which TBXO was used. These results also show that our implementation of TBXO yields a modest yet notable increase in sensitivity, but when BP was not used, the performance increase is observed only when TBXO is used exclusively in the early generations of training (see the center point in the solid lines in each panel in Figure 3). When BP was used in addition to GE to locally fit NN weights, using TBXO for the first half of training resulted in increased performance that did not change when TBXO was used throughout the rest of training (dashed lines in Figure 3). This is in contrast to previous work where TBXO showed little improvement when the simulated model was an interaction contributing to a discrete trait in the complete absence of main effects (Motsinger et al., 2006). This difference may be due to the fact that here we are analyzing a continuous trait with small main effects rather than a case-control outcome under a purely epistatic model without main effects. We then statistically evaluated this performance increase, summarized in Figure 4. Here, boxplots show the distribution of sensitivity across all combinations of generations and population sizes, and P-values indicate whether there is a statistically significant increase in sensitivity gained by using TBXO (one-way analysis of variance). The top panel of Figure 4 shows the combined results from using and not using backpropagation. Bottom panel shows the results considering simulations using and not using backpropagation independently. This indicates that the benefit from using TBXO when concurrently using BP is highly statistically significant, but there is little evidence to suggest using TBXO with the standard GE crossover alone results in any appreciable performance gains.

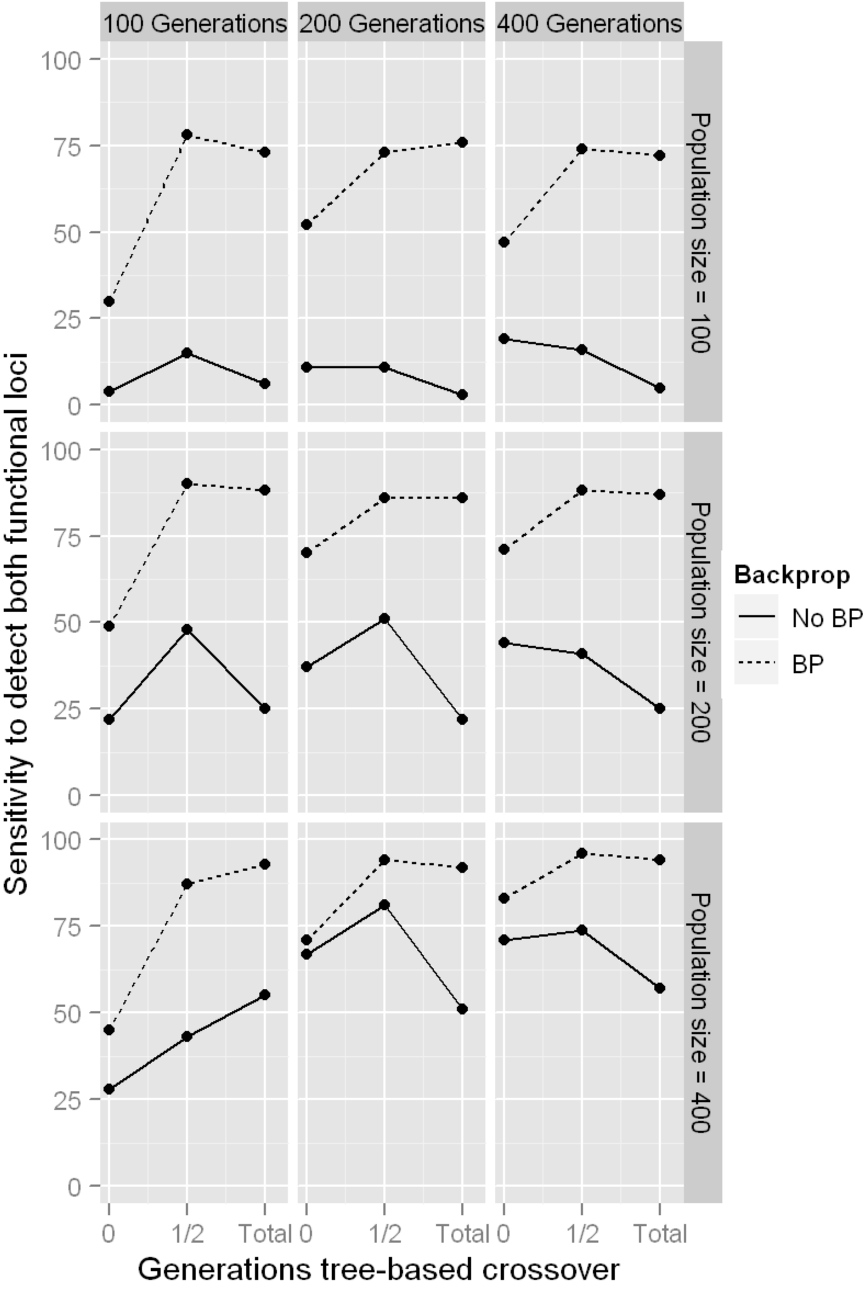


Figure 3. Sensitivity to detect both functional loci as the best GENN model. Each panel shows sensitivity over the proportion of total generations (none, half, all) where tree-based crossover was used instead of binary crossover. Solid line shows when GE alone was used to train NNs (no BP). Dashed line shows sensitivity when using the hybrid BP-GENN algorithm (see methods). Individual panels show combinations of the total number of generations GENN was run and the population size per deme.

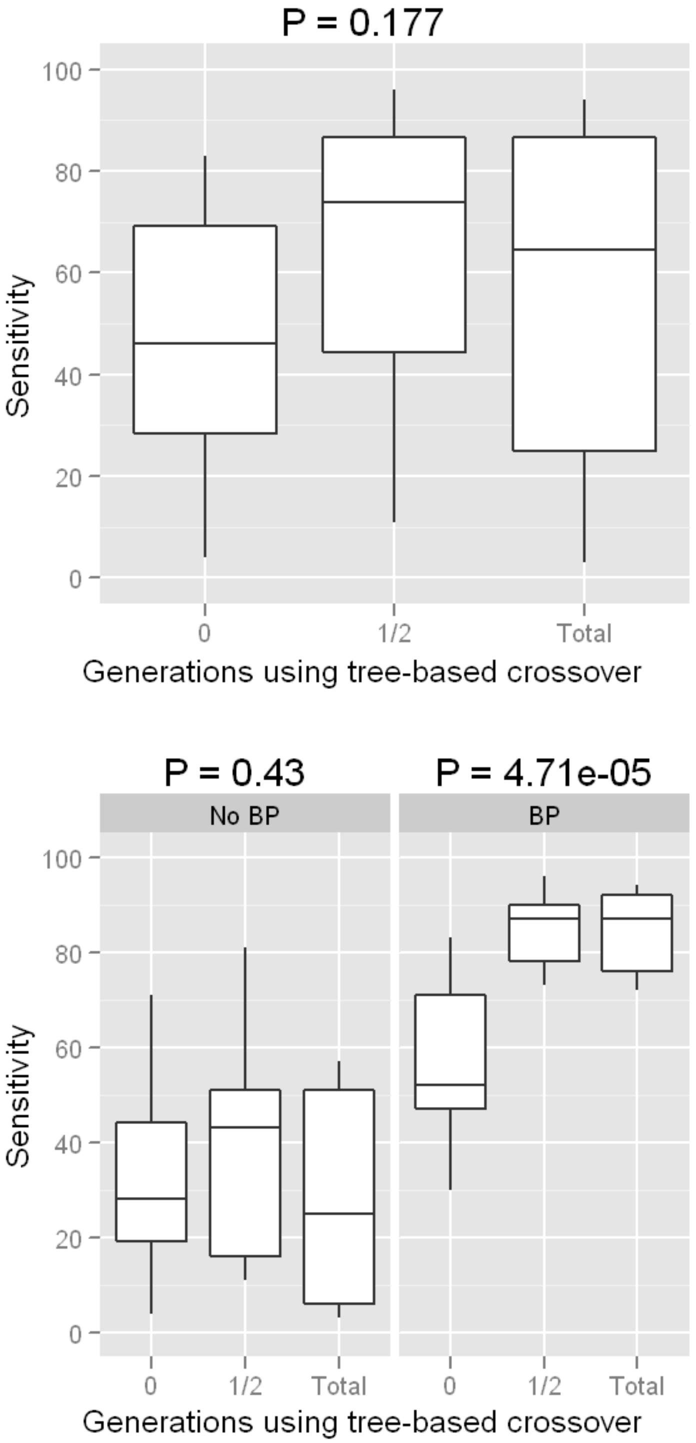


Figure 4. Statistical analysis of effectiveness of tree-based crossover. Boxplots show the distribution of sensitivity across all combinations of generations run and population sizes (see Figure 1). P-values indicate whether there is a statistically significant increase in sensitivity gained by using TBXO (one-way analysis of variance). Top panel shows the combined results from using and not using backpropagation. Bottom panel shows the results considering simulations using and not using backpropagation independently.

These results indicate that when BP is used, GE with TBXO is more efficient at variable selection, while GE with normal crossover allows more variation in building architecture and fitting weights. We postulate that TBXO is preserving "building blocks" which are functionally useful to the resultant neural network models. Our simulations contained a modest interaction effect (h^2=0.05) in the presence of very small main effects (h^2=0.01) at each of the interacting genetic variants. These small main effects may provide the building blocks upon which TBXO can capitalize. Syntactic preservation of NN genomes coding for the inclusion of these variables in NN models while allowing the full variability and broader search capability of SPBXO in the latter generations of evolution appears to be more powerful than using SPBXO or TBXO exclusively. Furthermore, recent work has shown that linkage disequilibrium (correlation between genetic variants) may provide building blocks to an evolutionary algorithm which builds neural networks when the true underlying model is an interactive effect in the complete absence of any main effect at each of the two functional variables (Turner et al., 2009). It is expected that the TBXO strategy discussed here may be optimal in this situation as well. Because our TBXO procedure mimics the function of genetic programming (GP), further studies should compare this against GP or any hybrid GP-GE NN training algorithm.

Incorporation of domain knowledge

Next we evaluated whether initializing the NN population with two-SNP models from domain knowledge sources resulted in any changes in performance. These results are summarized in Figure 5. The results here show that sensitivity to detect both genetic variants contributing to the trait is always higher when BP was used in conjunction with GE, as also shown in Figure 3. When the implication index is 0 (i.e. all domain knowledge is irrelevant), the sensitivity when using BP decreases substantially as the proportion of the initial population initialized from domain knowledge increases (upper left panel of Figure 5, dashed line). This is likely due to the fact that as more NN models are initialized from a list of models from irrelevant

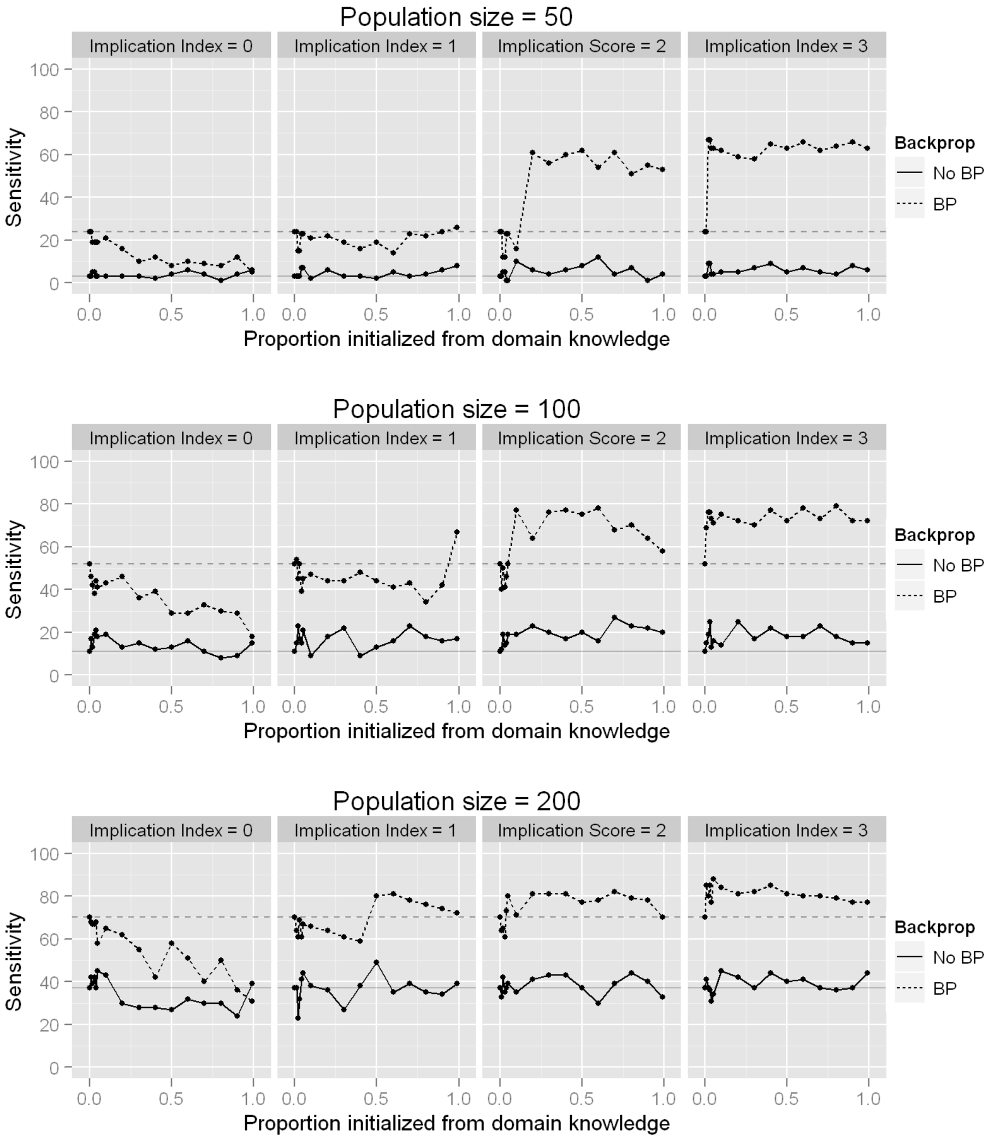


Figure 5. Sensitivity increases with the proportion initialized from domain knowledge. This figure illustrates the sensitivity of GENN to detect both functional SNPs as the proportion of the NN population initialized from domain knowledge increases from 0 to 99%. Panels going left to right show the increasing implication index of the model that includes both functional variables. Rows of panels show the population size per deme. The X-axis in each panel shows the proportion of the initial NN population which was seeded with two-SNP models from a domain knowledge source. Solid line shows when GE alone was used to train NNs (no BP). Dashed line shows sensitivity when using the hybrid BP-GENN algorithm (see methods). Faint horizontal solid and dashed lines show for reference the baseline sensitivity, for GENN and BP-GENN, when the population was initialized randomly, i.e. 0% initialized from domain knowledge. This figure indicates that sensitivity increases as the proportion of the NN population initialized from domain knowledge increases, and the increase is more notable in smaller population sizes.

domain knowledge, there is a smaller chance that either of the functional variables can be initialized by chance. When the implication index is at least 1 (meaning the functional two-SNP model is supported in our domain knowledge), as this proportion increases, sensitivity fluctuates around the baseline sensitivity (37%) at random initialization when BP is not used. This is not surprising, because even if a NN is initialized containing both functional variables that influence the trait, it is unlikely that by chance the NN would have suitable weights and architecture. An increase in performance can be seen when BP is then used to optimize the weights in the sensibly initialized NNs from relevant domain knowledge (dashed lines in panels in Figure 5 where implication index > 0). Furthermore, as the implication index for the domain knowledge model containing the functional variables increases from 1 to 3, this model is more likely to be incorporated into NNs in the initial generation. For instance, when the implication index of the functional model is 1, approximately 99% of the population must be initialized from domain knowledge to see any benefit. When the implication index is 2 or higher, it is very likely that the initial generation will contain a NN with the truly functional variables even when only a small proportion of the initial population is initialized from domain knowledge. Finally, looking down the rows of panels in Figure 5, it is clear that although the overall performance increases as the population size increases, as expected, the benefit of utilizing domain knowledge becomes less apparent. The benefits gained from utilizing domain knowledge to initialize a population of solutions is most apparent when the search space is large relative to the number of candidate solutions, as seen in the top right panel (implication=3, population size=50), dashed line.

These results demonstrate that the sensitivity of using GE to train NNs to find genes with a nonlinear influence on a quantitative outcome can be improved by effectively using extrinsic domain knowledge in conjunction with local weight fitting by BP. We showed that initializing a proportion of the NN population from two-SNP models incorporated from domain knowledge when BP is employed to locally optimize the weights in a NN can result in a performance improvement in ATHENA’s ability to detect and model SNPs influencing a quantitative trait. The performance increase was most notable when a smaller population size was used. This indicates that when the search space is small enough to be searched very thoroughly or exhaustively, using domain knowledge is less beneficial than when the search space is very large compared to the number of individual solutions being evolved. In this scenario (such is the case in genome-wide association studies), using domain knowledge to bias an evolutionary search in favor of important features will be critical for acceptable performance. While the benefits of using intrinsically obtained statistical expert knowledge (Moore et al., 2008; Greene et al., 2007), have not been explored in the ATHENA algorithm, using this framework to initialize an evolutionary search for disease genes based on domain knowledge obtained from public biological databases is another means to improve the performance of genetic algorithms for selection of important SNPs in a model.

Comparison with other methods

As discussed in Chapter I, other methods are available for probing the effect of gene-gene interactions on quantitative phenotypes. One exhaustive approach to testing gene-gene interaction among quantitative traits is the restricted partitioning method (RPM)(Culverhouse et al., 2004), an improvement over the combinatorial partitioning method (Nelson, Kardia, Ferrell, & Sing, 2001). RPM exhaustively evaluates all possible combinations of 2, 3, …, n-way combinations of SNPs, restricting the partitioning of each multilocus genotype into subsets that are likely to explain the most variation. While RPM should have high power and favorable computational performance in small datasets, as with any exhaustive approach to detecting interactions, its performance will decrease substantially as the number of SNPs in a dataset approaches that seen in genome-wide association studies. In addition to being extremely computationally intensive, exhaustive evaluation of all possible SNP-SNP interactions among GWAS data comes with an extraordinary loss of power due to the extremely large number of statistical tests being performed. Alternatively, parametric linear regression, when assumptions of normality and homoscedasticity are met, is uniformly the most powerful statistical method for ascertaining differences in group means (Maxwell et al., 2004; Cohen et al., 2002). In fact, when the functional variables are explicitly modelled, linear regression has >80% power to detect the gene-gene interaction effects simulated here (n=2000, sr^2_{main}=0.01, $sr^2_{interaction}$=0.05), determined using standard power calculation techniques for gene-gene interaction (Gauderman, 2002b; Gauderman, 2002a). This regression-based interaction-testing approach has been successfully used in a study of 13 SNPs in the APOE gene that influence ApoE protein levels in the blood (Hamon et al., 2004). Furthermore, regression models offer a very straightforward interpretation compared to NN models, which are often and unfortunately dubbed “black box” models (Dayhoff & DeLeo, 2001). However, in addition to the disadvantages discussed in Chapter I (curse of dimensionality, assumption violations, computational and multiple testing burdens with large datasets) that make exhaustive regression-based approaches impractical, it is also difficult to incorporate *a priori* information into a parametric regression analysis as it has been done here. Several knowledge-driven approaches have been applied to prioritize gene-gene interaction testing in large datasets (Baranzini et al., 2009; Bush et al., 2009; Peng et al., 2010; Ruano et al., 2010). These methods, however, limit statistical tests only to models supported by *a priori* knowledge. By contrast, the method proposed here only initializes a set of candidate solutions using domain knowledge – these solutions are then free to mutate and crossover, resulting in new and interesting combinations that may not be directly supported by the existing domain knowledge.

Conclusions

Here, we simulated a small effect size nonlinear interaction between two SNPs carrying minimal main effects and assessed the sensitivity of using GE to evolve NNs for detecting both functional SNPs out of a much larger set of unassociated variables. We showed that (1) using backpropagation, a fast NN weight optimization procedure, significantly improves ATHENA's performance, (2) using an alternative crossover strategy (TBXO) may allow for functional preservation of network information, and results in a statistically significant performance increase when used early in training in combination with backpropagation, and (3) incorporation of biological knowledge from the public domain can substantially improve ATHENA's performance at finding genes that interact to influence a trait. The general ATHENA algorithm is shown schematically in Figure 2.

Supplementing an evolutionary search using domain knowledge will be critical when using evolutionary procedures to find and model the effect of disease genes on complex human traits. Natural biological data will likely have many effects which will be enriched in knowledge sources, resulting in an improvement of the overall ability to find many members in the collection of influential loci. Genome-wide association studies offer very inexpensive measurement of over 1 million SNPs per sample. It is clear that there are more fruitful approaches for understanding the genetic architecture of common human phenotypes than ignoring the complexity of biology by testing single variants in isolation (Moore, Asselbergs, & Williams, 2010a). One of the strengths of the method presented here is that if any arbitrarily complex interaction of genetic and environmental exposures influences disease risk, a NN can approximate this function (Kurkova, 1991), given proper training. These experiments show that using a hybrid BP-GENN training algorithm, alternative crossover strategies, and incorporating domain knowledge into the search for genes related to disease can aid the variable selection and model fitting process of ATHENA. Ongoing efforts to incorporate other machine-learning algorithms, such as symbolic regression, into the ATHENA framework may also increase our ability to detect certain types of effects resistant to detection by BP-GENN.

One limitation in the current study is that these experiments make the assumption that loci involved in gene-gene interactions contributing to a heritable trait will carry with them some small main effect at either variant. This is a reasonable assumption to make, in that there are few, if any, examples of a consistently replicating, experimentally verified gene-gene interaction in the complete absence of main effects contributing to a complex quantitative trait in humans. Perhaps the reason for this, however, is the inadequacy of our methods for finding gene-gene interactions in the absence of main effects rather than the absence of such effects altogether. Biologically, redundancy and compensatory mechanisms at other loci can mitigate the effects of a devastating mutation or polymorphism at another locus, thus rendering its effect undetectable. This is evident in the many gene knockout mouse lines that show no apparent phenotype (Baba et al., 1994; Colucci-Guyon et al., 1994; Gorry et al., 1994; Gruda et al., 1996; Itohara et al., 1993; Killeen et al., 1992). Statistically, main effect components and interactions between them are mathematically independent effects (Maxwell et al., 2004). Furthermore, theoretical studies have shown that traits can be influenced exclusively through the interaction of two or more genetic variants (Culverhouse et al., 2002; Moore et al., 2002). Finally, one group has shown that main effects at variants involved in an epistatic interaction are highly dependent on the allele frequency in different populations at each locus, which may explain the lack of replication of many gene-gene interaction studies which rely on main effects (Penrod, Greene, & Moore, 2008). Future studies should aim to assess these and other extensions of ATHENA in their ability to detect and model epistatic interactions contributing to a quantitative trait in the absence of main effects, and should attempt to apply these methods in a natural biological data analysis.

Acknowledgements

The work presented in Chapter II was supported in part by NIH grants NS066638-01, LM010040, HG004608, and HL065962.

CHAPTER III

MULTIVARIATE ANALYSIS OF REGULATORY SNPS: EMPOWERING PERSONAL GENOMICS BY CONSIDERING CIS-EPISTASIS AND HETEROGENEITY[3]

Introduction

As discussed in the previous two chapters, epistasis is thought to be an important component of complex, multifactorial diseases due to the monumental complexity of biological systems (Tyler, Asselbergs, Williams, & Moore, 2009). Over the past 10 years, a wealth of data from model organisms has supported a role for epistasis (Shao et al., 2008; He et al., 2010). Furthermore, epistasis is one way to account for the problem of "missing heritability", where the analysis of single SNPs (single nucleotide polymorphisms) has explained very little of the heritability estimated from twin and adoption studies for complex traits (Eichler et al., 2010; Manolio et al., 2009). Accounting for interactions among SNPs may explain a larger portion of this heritability (Maher, 2008), expanding our understanding of the genomics of human disease and personalized medicine.

One often cited potentially causal mechanism of gene-gene interaction is due to variation in multiple genes in similar pathways, protein families, or genes with similar or redundant biological function (Aguilar et al., 2010; Costanzo et al., 2010). This generally implies that interaction occurs between genes scattered throughout the genome, implying a *trans*-epistasis effect. Several approaches, including the Biofilter approach described in Chapter II, have been applied to investigate these effects in genome-wide association studies (Baranzini et al., 2009; Bush et al., 2009; Peng et al., 2010; Ruano et al., 2010).

[3] Adapted in part from: Turner, S.D., Bush, W.S. (2011). "Multivariate Analysis of Regulatory SNPs: Empowering Personal Genomics by Considering cis-Epistasis and Heterogeneity." *Proceedings of the Pacific Symposium in Biocomputing*. In press.

"Epistasis" is often treated as synonymous with "gene-gene interaction" (Moore, 2003). Even I use these phrases interchangeably in other chapters in this thesis. The occurrence of epistatis, or statistically non-linear interactions (Fisher, 1918), however, need not be restricted to variation between distant genes. Epistatic interactions could also occur between genetic variants in close proximity that may impact transcriptional regulation. Recent work investigating the transcriptome of HapMap-based cell lines has led to the identification of expression quantitative trait loci (eQTLs) - genetic variants that influence the expression of a gene (Pickrell et al., 2010; Veyrieras et al., 2008). Veyrieras et al. published an analysis of gene expression for 11,446 genes from HapMap-based lymphoblastoid cell lines (Stranger et al., 2007) leveraging genotypes for roughly 3 million single nucleotide polymorphisms (SNPs) to identify eQTL SNPs in a 500 kilobase (kb) window both upstream of the transcription start site and downstream of the transcription end site (Veyrieras et al., 2008). This work discovered 744 genes containing at least one significant eQTL SNP ($p<7\times10^{-6}$). The single-SNP analysis, however, does not assess the variance in gene expression that can be explained by the interaction of multiple SNPs in regulatory regions of the gene. It has been shown that the underlying mechanisms of gene expression are incredibly complex, involving the binding of multiple factors to DNA to facilitate transcription and mRNA stability (Ravasi et al., 2010). Furthermore, polymorphisms within the binding sites of multiple factors may alter binding affinities to various degrees, exerting a non-linear influence on gene expression due to synergistic effects (Boj, Petrov, & Ferrer, 2010; Du, Thanos, & Maniatis, 1993). This principle has been demonstrated with multiple sclerosis where severity is impacted by functional effects of two alleles in close proximity in the MHC region (Gregersen et al., 2006). Despite the known complexity of gene regulation, multi-SNP interaction analysis has been previously examined only for genes having highly heritable expression but lacking single SNP associations (Dixon et al., 2007b). As a secondary analysis of eQTLs using lymphoblastoid lines isolated from children with asthma, the authors successfully explain some of the missing heritability from single SNP analysis using interactions. From this limited assessment, the authors conclude that genetic interactions may have an important role in the regulation of gene expression. From these points, we hypothesize that combinations of SNPs within the 500 kb window of potential transcriptional influence will alter gene expression in humans in a non-linear fashion, here dubbed *cis*-epistasis.

An analysis of gene expression phenotypes provides a unique opportunity to systematically assess the degree to which epistasis, or nonlinear interactions between genetic variants, might influence complex human traits. Linking the HapMap cell line expression data from (Veyrieras et al., 2008) with publicly available genotype data on the same cell lines gives us a dense collection of genetic variants in regions with strong biological plausibility for non-linear multi-SNP interaction within 11,466 quantitative expression outcomes with established main effects. Here we leverage these data to investigate the nature and degree to which *cis*-epistasis affects gene expression in humans. Furthermore, if epistasis plays an important role in influencing gene regulation, then it logically follows that epistasis is an important part of more complex downstream human disease phenotypes, as these traits are often associated to SNPs that alter gene expression (Gamazon et al., 2010). Finally, investigators could prioritize established combinations of eQTL SNPs to inform a SNP-SNP interaction analysis in complex human traits to reduce both the computational and multiple testing burdens that plague epistasis analysis in high-throughput genetic analysis. This would also motivate reanalysis of existing datasets for multi-SNP interactions that influence complex disease, many of which are publicly available at the database of genotypes and phenotypes (dbGaP) (Mailman et al., 2007). Put simply, if a study design that considers *cis*-epistasis can explain more heritability in gene expression, then genetic studies that account for *cis*-epistasis should be more fruitful.

Methods

Genotype and Gene Expression Data

As a starting point for these analyses, we retrieved the full eQTL results database and normalized gene expression data from the Veryrieras et al. analysis (available: http://eqtnminer.sourceforge.net/), containing 11,966,533 results (significant and non-significant) from 2,437,821 distinct SNPs and 11,466 distinct microarray probes (Veyrieras et al., 2008). We limited all of our analyses of the raw data to these SNPs and microarray probes. We analyzed SNP data from release #23 of the International HapMap project for 210 unrelated individuals, including 60 Yoruba (YRI) and 60 CEPH (CEU) parents, and 90 unrelated Chinese (CHB) and Japanese (JPT) samples (International hapmap consortium, 2007). We retrieved gene expression data that was normalized first by quantile normalization within replicates and then median normalized across all HapMap individuals. We then applied the normalization procedure from (Veyrieras et al., 2008), which is a Gaussian quantile normalization for each gene within each subpopulation separately to avoid results confounded by population stratification (the distribution of expression values within each subpopulation is now identical).

Statistical Analysis

From the single SNP results database, we extracted all SNPs with eQTL p-values <0.05 for each microarray probe – that is, all nominally significant SNPs falling within 500KB upstream of the transcription start site and 500KB downstream of the transcription end site. For each microarray probe, we generated all possible pair-wise combinations of SNPs, constructing 12,107,627 two-SNP models. For each model, we performed a multiple linear regression analysis fitting a model with main effect terms for the two individual SNPs and a multiplicative interaction term. We tested for significance of interaction via a student's T-test of the interaction term coefficient. All regression analyses were conducted using the 'rms' package for the R statistical computing environment (R Development Core Team, 2005). Statistical significance was determined by controlling the false discovery rate (FDR) at 0.20, using the 'qvalue' package available for R (Storey, Taylor, & Siegmund, 2004). Linkage disequilibrium was computed using PLINK software, analyzing the combined set of 210 HapMap samples without phasing using the '--r2' option (Purcell et al., 2007).

Annotation of Results Using GWAS Catalog

The National Human Genome Research Institute (NHGRI) actively maintains a catalog of all significant ($p<10^{-5}$) findings from published Genome-Wide Association Studies (GWAS) (Hindorff et al., 2009)(accessed March, 2010). The National Heart, Lung, and Blood Institute (NHLBI) also recently released comprehensive open access database of 118 GWAS studies containing 56,411 significant SNP-phenotype associations (Johnson et al., 2009). Illumina expression probe IDs were matched to transcripts within the Ensembl database (Release 49). Transcripts were matched to Ensembl Genes which have associated gene symbols within the Ensembl database. These symbols were matched to the "gene" fields in the GWAS catalogs to assess the number of matches. We also referenced the SNPs from our most significant results against these catalogs to determine if any single SNPs in the regions around our findings were known to influence any complex human phenotypes.

Results

Gene Expression in Humans is Influenced by Cis-Epistasis

After exhaustively fitting two-SNP models between known eQTL SNPs surrounding each microarray probe (12,107,627 two-SNP models in total), we examined the distribution of the p-values from the interaction term. The full results catalog from this analysis is available online at http://chgr.mc.vanderbilt.edu/bushlab/. Figure 6 is a quantile-quantile plot showing that the distribution of interaction term p-values deviates highly from the expected uniform distribution

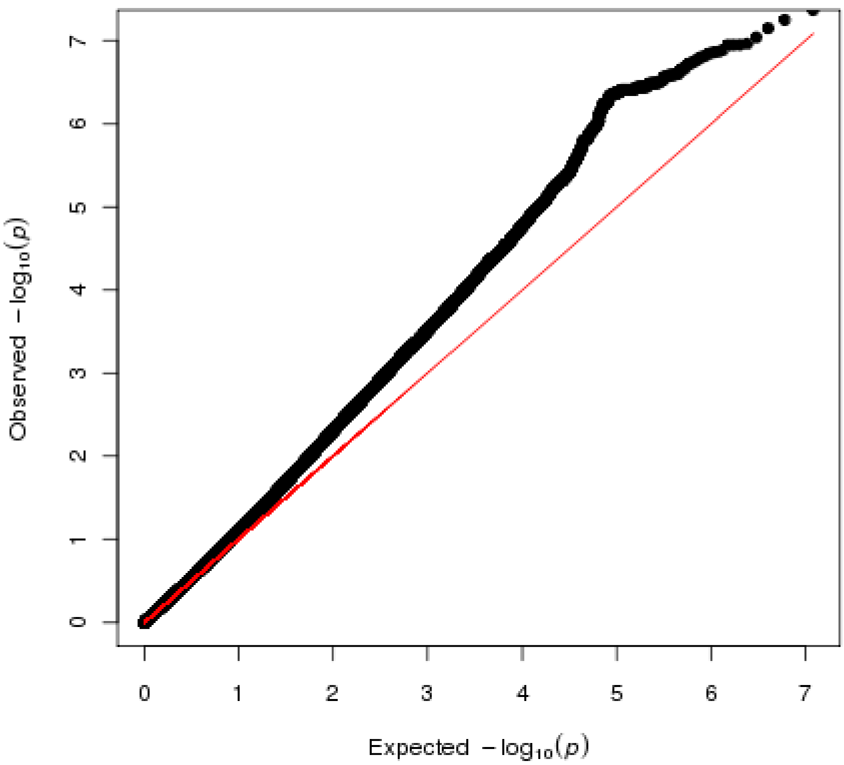


Figure 6. Quantile-quantile plot. This QQ plot shows the distribution of observed -log10(p-values) against the expected -log10(p-values) for the interaction term among 12,107,627 cis-epistasis models.

under the null hypothesis of no epistasis (diagonal line). This indicates that multi-SNP interaction may be common among eQTL SNPs that influence gene expression in humans.

Because a large number of statistical tests were performed, we corrected for multiple testing using the false discovery rate (FDR) method described in the methods section. Of the ~12 million two-SNP interaction models tested with multiple linear regression, 706 were still significant after correcting for multiple testing. It is of note that our multiple testing correction is extremely conservative because our tests of interaction are not independent of each other. The deviation from the null hypothesis of no interaction shown in Figure 6 suggests that there may be many more than 706 SNP-SNP interactions truly influencing gene expression that we are insufficiently powered to detect when applying our FDR correction. These 706 significant SNP-SNP interaction models influenced the expression of 79 unique probes, representative of 79 unique genes. All 706 interactions mapped to one of these 79 distinct genes. The reduction from 706 SNP-SNP interactions to 79 genes is due to correlation between statistical models, resulting from linkage disequilibrium (LD) between SNP 1 of model 1 and SNP 2 of model 2. However, there was relatively weak LD between the two SNPs participating in each interaction; i.e. SNP 1 and SNP 2 of model 1. The distribution of LD statistics (measured by r^2) between the SNPs in each interacting pair is shown in Figure 7. The median r^2 was 0.043, with a median distance between each pair of 108KB. Taken together, this suggests that the majority of the most significant results are indeed epistatic effects between independent SNPs, not simple haplotype effects. If SNPs participating in an interaction were in tight linkage disequilibrium it would be difficult to distinguish between a true multi-SNP interaction and a haplotype effect that is largely driven by the effect of a single SNP.

We then examined the most significant two-SNP models from each of these 79 genes, referencing each regulated gene to the GWAS results catalog described in the methods section. The GWAS results catalog contains SNPs that have been previously associated to a human phenotype, and the associated gene reported by the original GWAS publication. We matched the

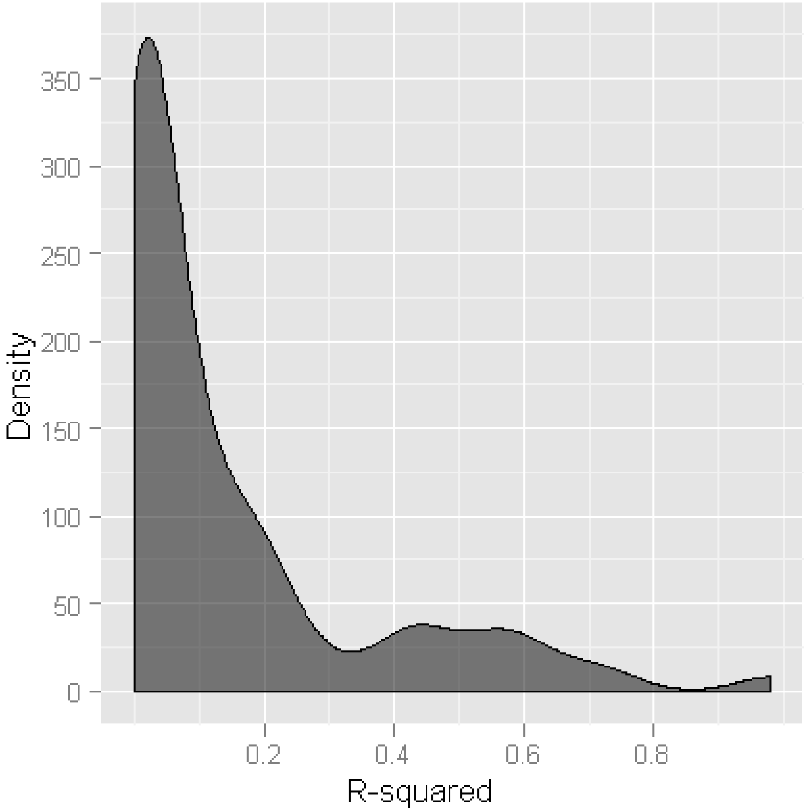


Figure 7. Linkage disequilibrium between *cis*-epistasis models. This shows the kernel density distribution of linkage disequilibrium values (r^2) between the most significant interacting SNP pair influencing expression of 79 genes after correcting for multiple testing.

significant *cis*-epistatic interactions to the GWAS results catalog in two ways: matching the 79 genes being regulated to the gene reported in the GWAS study, and matching SNPs participating in the 706 interactions to a SNP associated in a GWAS study. When matching by gene, we found that 20 of the 79 genes regulated by *cis*-epistasis have been previously reported in studies of approximately 20 human disease and morphological phenotypes (Table 3a). When matching by SNP, we found 10 additional *cis*-interactions where one of the specific SNPs has been associated to one or more disease or morphological phenotypes in humans (Table 3b).

For the majority the genes in Table 3, examining single SNP effects on expression only resulted in a nominal level of statistical significance (Table 3, columns "eQTL[1/2] P-value"). Examining the *cis*-epistasis interaction between the two SNPs allowed us to achieve a much greater degree of statistical significance (Table 3, columns "INT P-value" and "Model P-value"). Furthermore, accounting for *cis*-epistasis allows us to explain a much larger proportion (on average, 15 times more) of the heritability (variance) in gene expression (Table 3, column "R^2_{diff}", which is the difference in variance explained by the full model accounting for the interaction, "R^2_{full}", and the reduced model with main effects only, "R^2_{redu}").

Structural Characterization of Significant Two-SNP Interactions: Genomic Structure

Next we examined the genomic structural characteristics of the most significant two-SNP epistatic interactions that impact the expression of these 79 genes. Specifically, we examined the location of the two eQTL SNPs relative to each other and relative to the transcription start site (TSS) and transcription end site (TES) of the regulated gene. Based on structural characteristics, we defined four distinct classes of regulatory epistatic interactions: *upstream*, where both eQTL SNPs lie upstream of the TSS of the gene; *downstream,* where both eQTL SNPs lie downstream of the TES; *spanning*, where one eQTL SNP is upstream of the TSS and one eQTL SNP is downstream of the TES; and *intragenic*, where at least one eQTL SNP lies within the genic region, and the other may be either upstream, downstream, or also in the genic region.

Table 3a. Cis-epistasis models. Significant two-SNP interactions where the regulated gene has been previously associated to one or more complex human disease or morphological phenotypes. The specific SNPs which interact to regulate the gene were not necessarily reported as associated to the phenotype.

Assoc. Gene	eQTL1	eQTL2	R^2_{full}	R^2_{redu}	R^2_{diff}	β_1	β_2	β_{int}	LD (r^2)	eQTL1 Pvalue	eQTL2 Pvalue	INT Pvalue	Model Pvalue	GWAS Associated Phenotype
ABCA13	rs17132158	rs6945363	0.104	0.017	0.087	1.18	1.19	-0.77	0.00	0.0308	0.0012	1.3E-05	4.8E-05	Height,Triglycerides, Systolic Blood Pressure, Fasting glucose
AEBP2	rs7135885	rs11044945	0.100	0.013	0.087	-1.36	-0.80	0.58	0.74	0.0254	0.0222	1.4E-05	7.4E-05	Insulinogenic index
BLK	rs11986748	rs2572430	0.095	0.012	0.082	1.23	0.83	-0.77	0.02	0.0346	0.0461	2.4E-05	0.00013	Lupus*, Rheumatoid arthritis*
C10orf97	rs2883029	rs4748176	0.086	0.001	0.085	0.51	0.39	-0.47	0.02	0.0205	0.0188	1.9E-05	0.00032	Crohn's disease
CDKL1	rs1955926	rs7151406	0.115	0.037	0.078	-0.09	1.52	-0.58	0.56	1E-06	7E-09	3.3E-05	1.4E-05	Cognitive Performance
CPNE8	rs2387836	rs12818797	0.087	0.006	0.081	-1.19	-1.65	0.67	0.39	0.0114	0.0235	3.2E-05	0.00032	Waist circumference
DTNB	rs1369704	rs7607198	0.101	0.015	0.085	-0.02	2.94	-1.21	0.13	7E-15	0.0249	1.7E-05	7.2E-05	Type II Diabetes (T2D)
EEFSEC	rs2811484	rs2713590	0.099	0.013	0.086	-0.58	-0.28	1.24	0.00	0.0245	0.0307	1.4E-05	8E-05	Alzheimer's disease (AD)
FRMD3	rs10868025	rs11792634	0.124	0.042	0.082	-2.03	-1.30	0.70	0.06	0.0165	0.0307	2E-05	5.5E-06	HDL cholesterol
GNG2	rs1272117	rs3742536	0.087	0.005	0.082	-1.09	-1.23	0.90	0.01	0.0064	0.0171	2.6E-05	0.00031	AD, T2D, Crohn's disease
GRIP2	rs2607765	rs2607737	0.093	0.009	0.084	-0.19	-0.84	0.78	0.18	0.0262	0.0355	2E-05	0.00016	Cognitive performance
KIF7	rs17807856	rs3803530	0.081	0.003	0.078	-1.01	-0.88	0.58	0.01	0.0103	0.0012	4.4E-05	0.00057	LDL Cholesterol
MCOLN2	rs657309	rs6690583	0.095	0.003	0.092	0.42	1.02	-0.54	0.04	3E-13	0.0223	8E-06	0.00012	Fasting glucose
NMNAT3	rs10935317	rs7648532	0.121	0.033	0.088	1.54	1.40	-0.64	0.08	0.0273	2E-24	9.7E-06	7.1E-06	BMI, Fasting glucose
NPY	rs198723	rs16189	0.085	0.006	0.080	-0.63	-0.80	0.61	0.01	0.0461	0.0006	3.4E-05	0.00036	Early onset extreme obesity
NRN1	rs3763180	rs7763755	0.114	0.034	0.079	-1.10	-1.05	0.59	0.00	0.0169	0.0012	2.7E-05	1.6E-05	Waist/height ratio squared
OBFC1	rs2986059	rs3124	0.123	0.036	0.088	1.17	0.81	-0.70	0.01	0.002	0.0372	9.8E-06	5.5E-06	Parkinson's disease,brachial artery flow velocity, Height,Endothelial traits
PCM1	rs385139	rs7816561	0.101	0.019	0.082	0.76	1.17	-0.52	0.61	0.0377	0.0462	2.3E-05	6.9E-05	Triglyceride/HDL ratio
TYK2	rs10403787	rs4804480	0.166	0.095	0.071	1.48	0.43	-0.64	0.07	0.0004	0.0006	4.1E-05	3.7E-08	Type 1 Diabetes*,Lupus
ZBTB38	rs6802753	rs7626871	0.084	0.002	0.082	0.97	1.52	-0.76	0.08	0.0053	7E-10	2.8E-05	0.00042	Height*

*** indicates significant association of a gene to a complex human phenotype with p < 5E-8 (genome-wide significance)**

Table 3b Cis-epistasis models. Significant two-SNP interactions where one of the SNPs regulating a gene was previously associated to one or more human disease or morphological phenotypes. The involvement of the regulated gene in disease pathogenesis has not been investigated.

Gene	eQTL1	eQTL2	R^2_{full}	R^2_{redu}	R^2_{diff}	β_1	β_2	β_{int}	LD (r^2)	eQTL1 Pvalue	eQTL2 Pvalue	INT Pvalue	Model Pvalue	GWAS Associated Phenotype
SLIC1	rs2160683	rs9635531	0.087	0.009	0.079	1.17	1.17	-0.66	0.61	0.0461	0.0433	3.7E-05	0.00029	Crohn's disease
TMBIM1	rs7605980	rs12471773	0.087	0.011	0.076	0.37	0.64	-0.73	0.09	0.0002	0.0408	5E-05	0.00032	Type I Diabetes
PCM1	rs396462	rs2955427	0.091	0.003	0.088	0.95	0.92	-0.50	0.47	0.0326	0.0039	1.3E-05	0.00019	Crohn's disease
OTUB2	rs6575354	rs12433627	0.079	0.000	0.079	-1.00	-0.90	0.58	0.00	0.0346	0.0203	4E-05	0.00071	Type I Diabetes
DDX19A	rs929840	rs2303791	0.092	0.009	0.082	1.45	0.41	-0.80	0.32	0.042	0.0135	2.4E-05	0.00019	Alzheimer's disease
DDX19A	rs929840	rs4985534	0.087	0.008	0.079	1.43	0.39	-0.79	0.33	0.042	0.0104	3.7E-05	0.00032	Alzheimer's disease
ORMDL1	rs7568054	rs7568449	0.123	0.044	0.079	-0.24	-0.44	0.49	0.01	0.0263	3E-22	2.6E-05	5.5E-06	Amyotrophic Lateral Sclerosis
C3orf31	rs7615782	rs440746	0.115	0.020	0.095	1.56	0.16	-0.82	0.11	0.0031	4E-17	4.9E-06	1.4E-05	Waist circumference, Hypertension
XKR9	rs268625	rs7828552	0.137	0.051	0.085	-1.84	-1.29	0.66	0.23	0.0009	0.0002	1.1E-05	1.2E-06	Systolic blood pressure post-exercise
C17orf53	rs228769	rs2526021	0.086	0.007	0.079	0.95	1.02	-0.58	0.61	0.0103	0.0001	3.7E-05	0.00035	Bone mineral density

We observed 25 upstream interactions (32%), 18 downstream interactions (23%), 17 spanning interactions (21%), and 19 intragenic interactions (24%). Interestingly, all our significant results were evenly distributed among the four structural classes, as a *z*-test for population proportions revealed no significant difference from 25%. Figure 8 shows that the four structural classes are distributed evenly among these most significant 79 *cis*-epistatic interactions. Figure 8 also reveals that the distribution of structural class does not correlate with gene size, organized vertically along the figure.

Structural Characterization of Significant Two-SNP Interactions: Structure of the statistical model

Statistical epistasis is classically defined as the deviation from additivity in a linear model (Fisher, 1918). We have shown that there are significant nonadditive effects impacting gene expression throughout the genome. Next we examined the structure of the statistical models of the most significant interactions impacting the 79 unique genes discussed above. Specifically, we examined the direction of the coefficients of both main effect terms and the interaction term in each statistical model.

We found that of these 79 significant *cis*-epistasis interactions, the main effect coefficients in 75 of these models were in the same direction. That is, if inheriting one copy of the minor allele of a single variant caused an increase in expression, the main effect of the other SNP also resulted in an increase in expression. Recall that we only tested for SNP-SNP interactions among eQTL SNPs that had an established main effect. Interestingly, of these 75 *cis*-epistasis models where both main effects were in the same direction, the statistically significant interaction term coefficient was in the *opposite* direction. That is, if the main effect of each variant alone caused an *increase* in expression by *x* units, inheriting both variants resulted in an increase in expression that deviates significantly from the expected 2*x* increase. Of the remaining four significant *cis*-epistasis interactions, the main effects were in opposing directions. For three of these four, the main effect coefficient of one SNP in the model approached zero after accounting for the

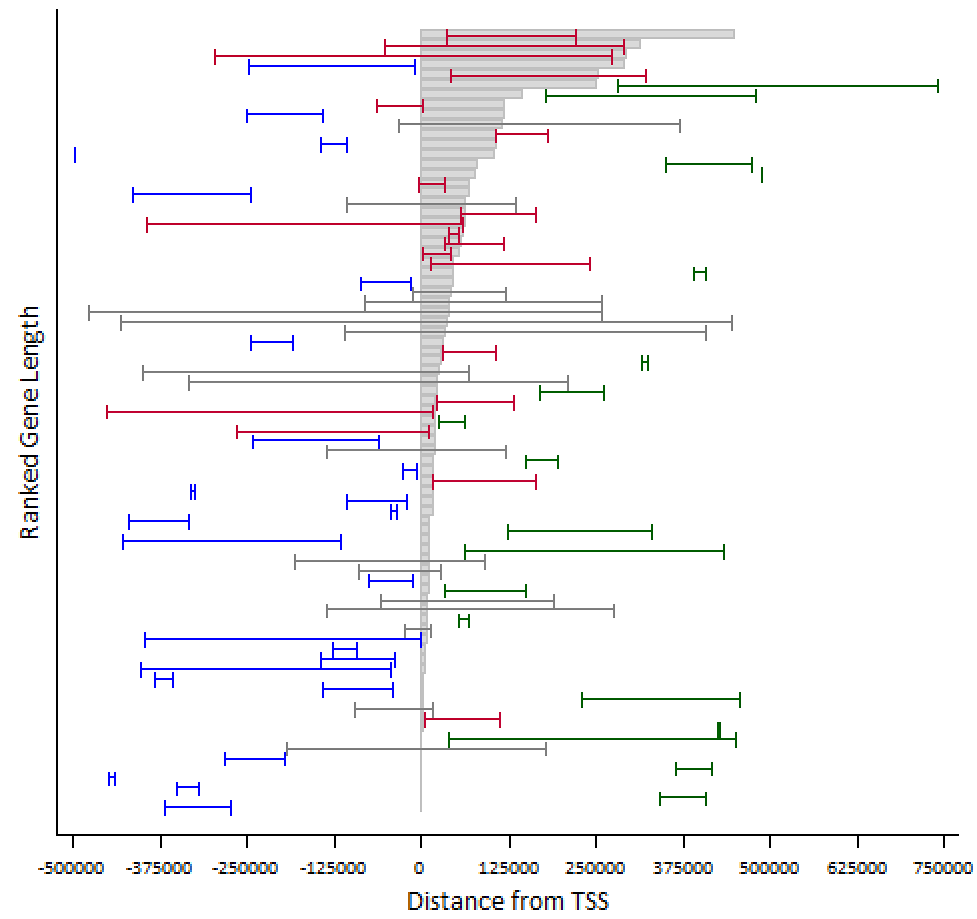


Figure 8. Genomic structure of significant cis-epistasis models. Transcribed regions of these 79 genes (gray boxes) are aligned by transcription start site, ordered by gene size. Epistatically interacting SNPs that influence the gene's expression are shown as connected hash marks, color coded by class: upstream (blue), downstream (green), spanning (gray) and intragenic (red). Analysis of the genomic structure of cis-epistatic interactions reveals that all four structural classes are evenly represented among the most significant cis-epistatic interactions, and that structural class does not correlate with gene size.

interaction. This suggests a classical modifier effect, where one variant only exerts an effect in the presence of another. In all three of these models, the presence of the "modifier SNP" ($\beta \approx 0$) results in a mitigation of the main effect of the other SNP.

The pattern of coefficients can be seen by examining $ß_1$, $ß_2$, and $ß_{int}$ for the models presented in Table 3 (showing only models related to a human phenotype from a GWAS). These results indicate that the overwhelming majority of significant non-additive two-SNP interactions influencing gene expression represent epistatic genetic heterogeneity rather than multiplicative effects. We consider this in greater detail in the discussion section below.

We also investigated the possibility that aspects of the genomic structure of the model might impact the statistical nature of the interaction. However these analyses revealed no significant relationships between genomic structure characteristics (such as class or the physical distance between the two SNPs) to the variance explained (R^2) or magnitude of the interaction coefficient.

Conclusions

In this work we examined eQTL SNPs known to impact gene expression in humans for non-additive epistatic effects by combining transcriptome-wide expression data from HapMap lymphoblastoid cell lines with genome-wide SNP data from the same cell lines. Specifically, we analyzed over 12 million potential two-SNP interactions for *cis*-epistasis among SNPs known to regulate transcription of a nearby gene, and found that multiple independent eQTL SNPs may often interact to influence gene expression non-additively. After correcting for multiple testing, we found 706 highly significant *cis*-epistasis interactions that influence the expression of 79 unique genes.

We characterized the genomic and statistical structure of the most significant *cis*-epistasis model corresponding to each of these 79 genes. Here we discovered that in the vast majority of *cis*-epistasis interactions (1) the main effects are in the *same* direction, and (2) the interaction was in the *opposite* direction. While still considered a nonlinear epistatic interaction, the structure of this type of model is referred to as a *heterogeneity model* (Cordell, 2002; Neuman & Rice, 1992) rather than a multiplicative model. Genetic heterogeneity is a serious concern with large-scale genetic studies, and is often cited as a reason for the widespread lack of replication in GWAS studies (McClellan et al., 2010; Sillanpaa & Auranen, 2004). Because epistatic genetic heterogeneity may commonly impact regulation of gene expression, and since SNPs associated with complex human phenotypes often result in changes of gene expression (Gamazon et al., 2010), it follows that *cis*-epistatic genetic heterogeneity could exert a significant influence over complex human traits and should be investigated as such. Others have recently argued that epistatic genetic heterogeneity should be considered when analyzing genomic data for association to disease (Moore, Asselbergs, & Williams, 2010b). Despite the fact that statistical tools have been available for some time now to accomplish this (Lunetta, Hayward, Segal, & Van, 2004; Thornton-Wells et al., 2004), analyses of genome-wide datasets accounting for the possibility of *cis*-epistasis is a task rarely undertaken. Accounting for genetic heterogeneity in gene expression may improve the replicability of existing genetic association studies.

By matching *cis*-epistasis interactions to the GWAS results catalogs by SNP, we discovered that of the 79 significant *cis*-epistasis interactions, 10 contained one SNP previously associated to a human phenotype via GWAS studies. Nearly all of these associations initially fell short of "genome-wide" statistical significance (Pe'er, Yelensky, Altshuler, & Daly, 2008) and thus would not be reported in the literature as a relevant gene for the phenotype. Furthermore, the statistical significance of each single SNP on the expression of a gene is weak. However, when we consider the joint effect of both SNPs involved in the *cis*-epistasis interaction, we see a dramatic improvement in the variance of gene expression explained. As such, we hypothesize that some of these reported associations from the GWAS catalog would show stronger associations to the phenotype if modeled with their *cis*-epistasis partner SNP. In light of the prevalence of *cis*-epistatic interactions, these examples provide motivation to re-examine existing datasets for *cis*-epistatic effects on human phenotypes. Our models provide a compelling set of specific regulatory hypotheses to examine in existing data.

Many new approaches have been recently used to examine epistasis in GWAS data (Baranzini et al., 2009; Bush et al., 2009; Peng et al., 2010; Ruano et al., 2010). All of these approaches focus on interactions among SNPs within genes related to a common biological mechanism, such as pathways, and structural or functional similarity. With these approaches, interaction models consist of SNPs from each of two distant genes – a *trans*-epistasis effect. In most cases, this precludes the possibility of capturing *cis*-epstasis effects. While *trans*-epistasis effects are likely to be important for complex disease etiology, we argue that *cis*-epistasis may be of equal or greater importance, and coupling *cis*- and *trans*-epistasis analysis methods may be more successful.

Furthermore, the available tools for the analysis of multi-locus interactions in genetic association studies (including the ATHENA platform discussed in Chapter II) are not likely to discover the *cis*-epstasis effects we describe here. Knowledge-based approaches generally test models of *trans*-epistasis (as discussed above). Sliding window-based haplotype association approaches typically use window sizes based on a fixed physical distance or number of SNPs (Lin, Chakravarti, & Cutler, 2004). These approaches would likely not discover *cis*-epstasis effects due to the variable and often large distances between the pairs of regulatory SNPs within the model (see Figure 8).

Additionally, any gene-based analysis approach that uses SNP data requires mapping SNPs to genes. This is exclusively done using either physical distance (base-pair proximity) or genetic distance (linkage disequilibrium). The genomic window generated using these approaches is typically conservative, including a small region upstream and downstream of the gene region. Others have shown in model organisms that regulatory elements exert effects from extremely long distances (Chandler, Chandler, McFarland, & Mortlock, 2007). Likewise, many of the single SNP eQTLs used in the examination of this study illustrate long range regulatory effects (Veyrieras et al., 2008). From our analysis, we provide additional evidence that SNPs can influence the regulation of a gene at great distances from the transcription start site, and existing SNP-to-gene mapping approaches should take this into account.

We therefore suggest that the re-analysis of existing datasets and the development of new analysis approaches take into account the possibility that long range regulatory interactions could alter gene expression and thus influence human phenotypes. By accounting for more variance in gene expression (thus increasing statistical power), this will improve performance of analytical methods and potentially improve the replicability of new GWAS findings. One basic approach would be to use the models we have generated as templates for the analysis of *cis*-epistasis in existing and future personal genomics studies. The 79 genes we identified after multiple testing correction suggest the most compelling cases of *cis*-epistasis. However, interaction models with less significant p-values may explain sufficient variance in a gene's expression to resolve an association with a phenotype.

One limitation of this study is that whole-transcriptome data was available for only 210 HapMap samples. However, dense genome-wide SNP data is available for 1397 individuals in 11 diverse human sub-populations through the HapMap project (International hapmap consortium, 2007), so if additional gene expression data were collected we could improve the statistical power of this analysis to detect *cis*-epistasis effects. Also, we only considered interactions among eQTL SNPs with a nominally significant regulatory effect ($p < 0.05$). A reanalysis of these data including all SNPs, (even those without a known regulatory effect) would be straightforward, perhaps revealing additional *cis*-epistasis effects; however this would cause a power loss from the increased burden of multiple testing correction.

In summary, we have shown that *cis*-epistasis is an important phenomenon regulating gene expression in humans. Using this information, we suggest ways in which the performance of existing and future analysis approaches can be improved, and how additional insights into human biology and disease pathogenesis could be gained from personal genomics studies. Likewise, future research and development on incorporating domain knowledge into the ATHENA algorithm introduced in Chapter II should consider *cis*-epistasis as a source of domain knowledge, rather than only considering potential *trans*-epistatic SNP-SNP interaction models from distant genes.

Acknowledgements

The work presented in Chapter III was supported in part by National Research Service Award F31-NS066638 from the National Institute of Neurological Disorders and Stroke.

CHAPTER IV

QUALITY CONTROL PROCEDURES FOR GENOME WIDE ASSOCIATION STUDIES[4]

Introduction

As discussed in previous chapters, genome-wide association studies (GWAS) are commonly used to identify common single nucleotide polymorphisms (SNPs) that influence human traits. GWAS have been conducted at increasing frequency using case-control, population-based prospective, and cross-sectional study designs (Klein et al., 2005; Frayling, 2007; Newton-Cheh et al., 2009; Kathiresan et al., 2009; Willer et al., 2008b; Hindorff et al., 2009). More recently, GWAS are being conducted in cohorts that are clinic-based (Barber et al., 2010; Daly et al., 2009; Link et al., 2008; Thompson et al., 2009). As a result, GWAS may soon move the field of genomics into clinical practice.

Whether the goal is to identify predictors of outcomes or to discover new biology underlying a trait of interest, the capability of GWAS to identify true genetic associations depends upon the overall quality of the data. Even simple statistical tests of association are compromised in the context of genome-wide SNP data that have not been properly cleaned, potentially leading to false-negatives and false-positive associations (Pluzhnikov et al., 2010). Additionally, problems with the overall data quality will likely affect downstream analyses and studies beyond the initial GWAS. For example, the National Human Genome Research Institute (NHGRI) actively maintains an online catalog of GWAS results and associated publications (Hindorff et al., 2009), which stimulates downstream studies of replication and characterization in independent populations. Compromised data quality in the discovery phase may lead to false

[4] Adapted in part from: Turner S.D., Armstrong L., Bradford Y., Carlson C., Crawford D.C., Crenshaw A.T., de Andrede M., Doheny K., Haines J.L., Hayes G., Jarvik G., Jiang L., Ling H., Kullo I., Li R., Manolio T.A., Matsumoto M., McCarty C.A., McDavid A., Mirel D., Paschall J., Pugh E., Rasmussen L.V., Wilke R.A., Zuvich R.L., Ritchie M.D. (2011). "Quality Control procedures for Genome-Wide Association Studies." *Current Protocols in Human Genetics*. In press.

positives that are carried forward into replication studies at great cost both in time and expense. Also, the National Institutes of Health (NIH) now mandates that secure, encrypted copies of primary GWAS data funded by NIH be made publicly available (with controlled access) for secondary analyses. These accessible datasets are maintained by the National Center for Biotechnology Information (NCBI) in the database of Genotypes and Phenotypes (dbGaP). dbGaP provides both open and controlled access, which allow for both broad release of non-sensitive information, and restricted access to datasets involving genomic data and phenotypic information, respectively (Mailman et al., 2007). Data access through dbGaP is commonly used for replication and meta-analysis, both of which will be compromised by poor quality data.

Genotyping technology and allele calling algorithms continue to improve and quality-improvement strategies continue to ensure that only reliable, rigorously scrutinized markers and samples are used for analysis. Reconciling genetic data with clinical and self-reported data (e.g., sex or familial relationships) can potentially identify sample identity problems caused by sample handling mishaps. Batch effects, population stratification, and sample relatedness can confound genetic association analyses and can lead to excessive type I and type II errors. Here I discuss methods that can be used to detect and account for various data quality issues to better ensure the integrity of the primary GWAS as well as its downstream applications.

The eMERGE (electronic MEdical Records and GEnomics) Network is an NHGRI-supported consortium of five institutions charged with exploring the utility of DNA repositories coupled to Electronic Medical Record (EMR) systems for advancing discovery in genome science (McCarty et al., 2010a). eMERGE also includes a special emphasis on the ethical, legal and social issues related to these endeavors. The five sites in eMERGE include Marshfield Clinic, Group Health Seattle, Vanderbilt University, Mayo Clinic, and Northwestern University. Using an algorithm for a specific phenotype, each of the participating sites extracted study samples for a specific disease or phenotype from their EMR. Genome-wide genotyping has been performed on ~17,000 samples across the eMERGE network at the Broad Institute and at the Center for

Inherited Disease Research (CIDR) using the Illumina 660W-Quad or 1M-Duo Beadchips. Each study site is conducting a GWAS, in addition to a number of cross-network analyses. These studies adhere to NIH's data sharing policies, and all data generated in this study will be available on dbGaP (Mailman et al., 2007). Due to the complexity involved in a single site GWAS, in addition to the combining of data and results across study sites, it became clear that a unified QC pipeline was imperative. To facilitate this process, the eMERGE Coordinating Center at Vanderbilt University is leading the genomics group in the development of the eMERGE QC pipeline.

Others have discussed quality control procedures for genotypic data (Laurie et al., 2010; Miyagawa et al., 2008; Chanock et al., 2007b; Broman, 1999). This chapter documents the QC procedures that I used prior to analyzing the GWAS data presented in Chapter V using the Marshfield Personalized Medicine Research Project and Vanderbilt BioVU cohorts. This chapter may also serve to instruct future investigators in QC procedures that should be implemented prior to analyzing GWAS data. The procedures discussed here were developed by the genomics group of the eMERGE network, where phenotyping and other sample information is obtained through sophisticated mining of the EMR. This protocol can be applied to many GWAS studies, regardless of phenotyping strategy. Given that most of the genotyping data available for GWAS is currently SNP-based, I will limit our discussion to these biallelic markers and QC procedures for CNV analysis will not be discussed here.

Figure 9 shows a flowchart overview of the entire QC process, where each step is discussed in detail in the following sections.

GWAS Data Format

Regardless of the underlying study design (such as family-based or population-based), the most commonly used format for genetic data is the linkage, or pedigree file format (pedfile).

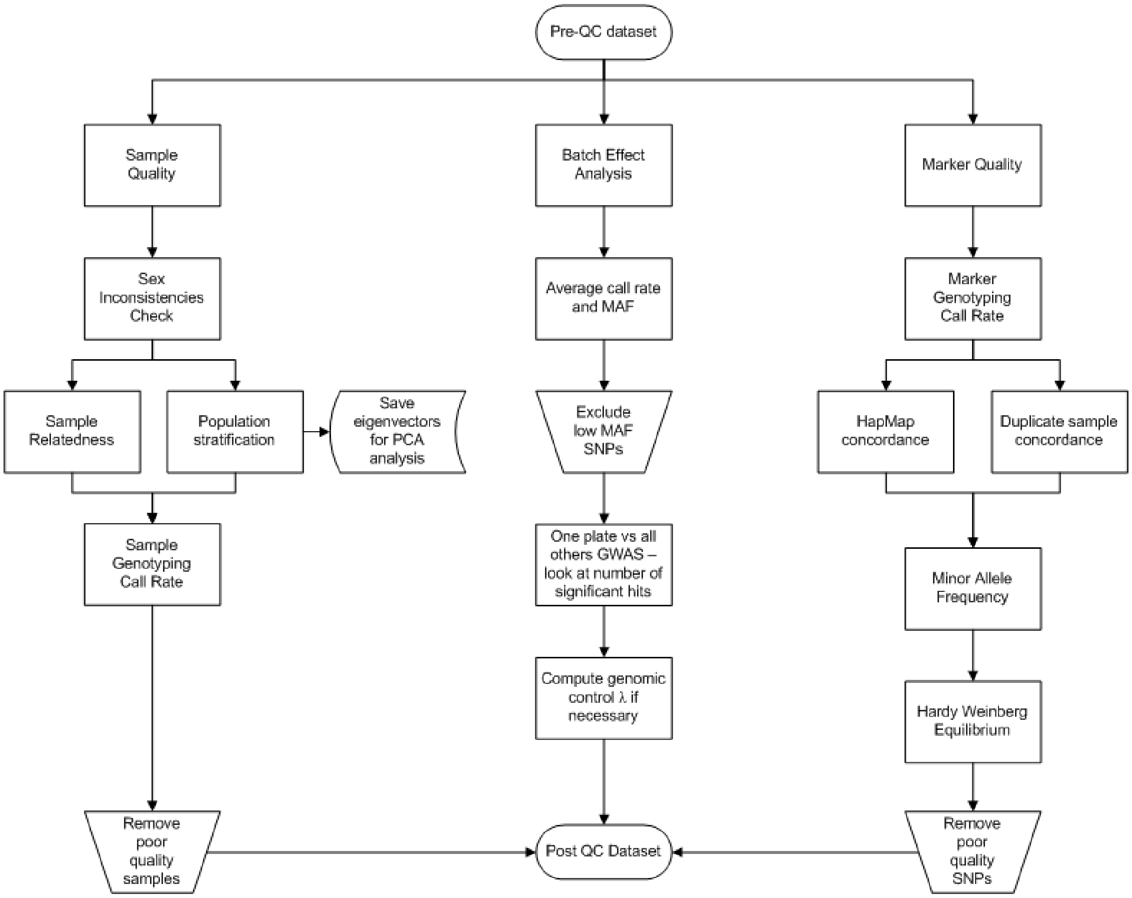

Pre-QC dataset
Sample Quality
Batch Effect Analysis
Marker Quality
Sex Inconsistencies Check
Average call rate and MAF
Marker Genotyping Call Rate
Sample Relatedness
Population stratification
Save eigenvectors for PCA analysis
Exclude low MAF SNPs
HapMap concordance
Duplicate sample concordance
Sample Genotyping Call Rate
One plate vs all others GWAS – look at number of significant hits
Minor Allele Frequency
Compute genomic control λ if necessary
Hardy Weinberg Equilibrium
Remove poor quality samples
Post QC Dataset
Remove poor quality SNPs

Figure 9. A flowchart overview of the entire GWAS QC process. Each topic is discussed in detail in the corresponding section in the text. Squares represent steps, ovals represent input or output data, and trapezoids represent filtering of data.

This file contains one individual per row, where the first six columns are identifying information (family ID, individual ID, father ID, mother ID, sex, phenotype), and the remaining columns are genotypes (2 columns per genotype; one for each allele). The genotype column-pairs correspond to an ordered set of SNP markers present in an associated file (.map or .bim). Additional phenotypes can also be stored in separate files consisting of family ID, individual ID, then extra columns representing additional phenotypes. There are several variations on pedfile format, including transposed (long) formats (tped), and compressed (binary) formats. Descriptions of these file formats can be found on the PLINK homepage (Table 4). PLINK is a freely available, open source, cross platform application for QC and analysis of GWAS data (Purcell et al., 2007). We used PLINK for implementing most of the eMERGE network's QC pipeline.

An important issue when creating a pedfile for QC analysis is the choice of strand orientation to use for allele calls (i.e. forward or reverse complement). While forward strand is a commonly used allele coding scheme, Illumina has developed a consistent and simple method to ensure uniformity in genotype call reporting that uses the polymorphism itself and the contextual surrounding sequence ("TOP/BOT" strand and "A/B" allele coding) (2009e). Since 2005, the database of genetic variation (dbSNP) (Sherry et al., 2001) has used this designation for all SNP entries. We used forward strand orientation for eMERGE. Choice of strand orientation might depend on the strand orientation of other data used in a combined analysis or of a reference set used for imputation. The goal is to ensure uniformity in genotype call reporting that is critically important in downstream analyses, reporting, and annotation.

Sample quality

Sex inconsistencies and chromosomal anomalies

One of the first procedures that should be implemented in any GWAS QC protocol is checking for potential sample identity problems that typically result from sample handling

Table 4. Software for data analysis. Useful software packages for data management, quality control, and statistical analysis in genome-wide association studies.

Software package	URL	Purpose
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/	Free, open-source GWAS analysis software package. Contains many tools for data management, quality control, and statistical analysis. (PC, Mac, Linux).
PLATO	https://chgr.mc.vanderbilt.edu/plato	*PL*atform for the *A*nalysis, *T*ranslation, and *O*rganization of large-scale data – software for GWAS analysis similar to PLINK.
R	http://www.r-project.org/	Free, open-source statistical computing software with excellent graphical capabilities. (PC, Mac, Linux).
Eigensoft	http://genepath.med.harvard.edu/~reich/Software.htm	Free, open-source software for performing principal components analysis based method for detecting and correcting for population stratification in GWAS. (Linux only).
Structure	http://pritch.bsd.uchicago.edu/structure.html	Free, open-source software for inferring the presence of distinct populations and assigning individuals to those populations for a stratified analysis. (Windows, Mac, Linux)
MySQL Workbench	http://wb.mysql.com/	Free, open-source software for creating, administering and querying relational databases. This is helpful for subsetting data, merging results, and joining QC metrics (e.g. HWE) to final association results. (Windows, Mac, Linux).

errors. One of the easiest ways to discover potential sample handling issues that result in mix-ups is by checking the reported sex of each individual against that predicted by the genetic data. The *check-sex* option in PLINK uses X chromosome heterozygosity rates to determine sex empirically, then reports individuals for whom the sex recorded in the pedfile does not match the predicted sex based on genetic data (example output and explanation shown in Table 5). If discrepancies are found (e.g. an individual is recorded being female but appears homozygous for every X chromosome marker), the EMR or any available study questionnaires should be reviewed to make a determination whether there was a sample handling mistake that caused a sample mix-up. Checking X chromosome heterozygosity may also reveal sex chromosome anomalies such as Turner syndrome (females having karyotype XO), Kleinfelter syndrome (males having karyotype XXY), mosaic individuals (XX/XO, XX/XXY), or females with large stretches of loss-of-heterozygosity on the X chromosome who are otherwise phenotypically normal.

X chromosome heterozygosity is a fairly sensitive heuristic to detect sample swaps, but not very specific. A variety of factors besides a crude sample mix-up will affect heterozygosity. Furthermore, if the goal is to enumerate as many samples with atypical sex karyotypes as possible, then X heterozygosity alone will not detect abnormalities such as triple X or XYY or homozygous X Kleinfelter syndrome. Examining the intensity of probe binding on the sex chromosomes will better resolve these cases. Illumina calls this intensity LogR ratio. On Affymetrix systems, it is simply known as probe intensity. These metrics, once suitably normalized, are roughly linear in copy number. Because there are tens of thousands of loci on the X chromosome on modern platforms, it is appropriate to examine a subsample of markers, and then take a measure of central tendency of each sample such as the median or mean intensity. The intensity plot provides a visualization of the intensity of X and Y probes (Figure 10). It is expected that females should have low Y intensity and high X intensity (bottom right corner), and the males show have similar level of X and Y intensities (top left corner). We also observed two individuals with mislabeled sex as well several individuals with XXY. Structural

Table 5. Example table showing output from --check-sex routine using PLINK. IID=individual id; PEDSEX=sex as recorded in pedfile (1=male, 2=female); SNPSEX=sex as predicted based on genetic data (1=male, 2=female, 0=unknown); F=X chromosome inbreeding (homozygosity) estimate.

IID	PEDSEX	SNPSEX	STATUS	F	Explanation
1	1	1	OK	0.98	Male
2	2	2	OK	0.03	Female
3	2	1	PROBLEM	0.99	Recorded female, genetically male
4	1	2	PROBLEM	0.02	Recorded male, genetically female
5	2	0	PROBLEM	0.28	Likely a female with sex chromosome anomaly (e.g. XX/XO mosaic, loss-of-heterozygosity on X)
6	1	0	PROBLEM	0.35	Likely a male with sex chromosome anomaly (e.g. XXY or XX/XY mosaic)

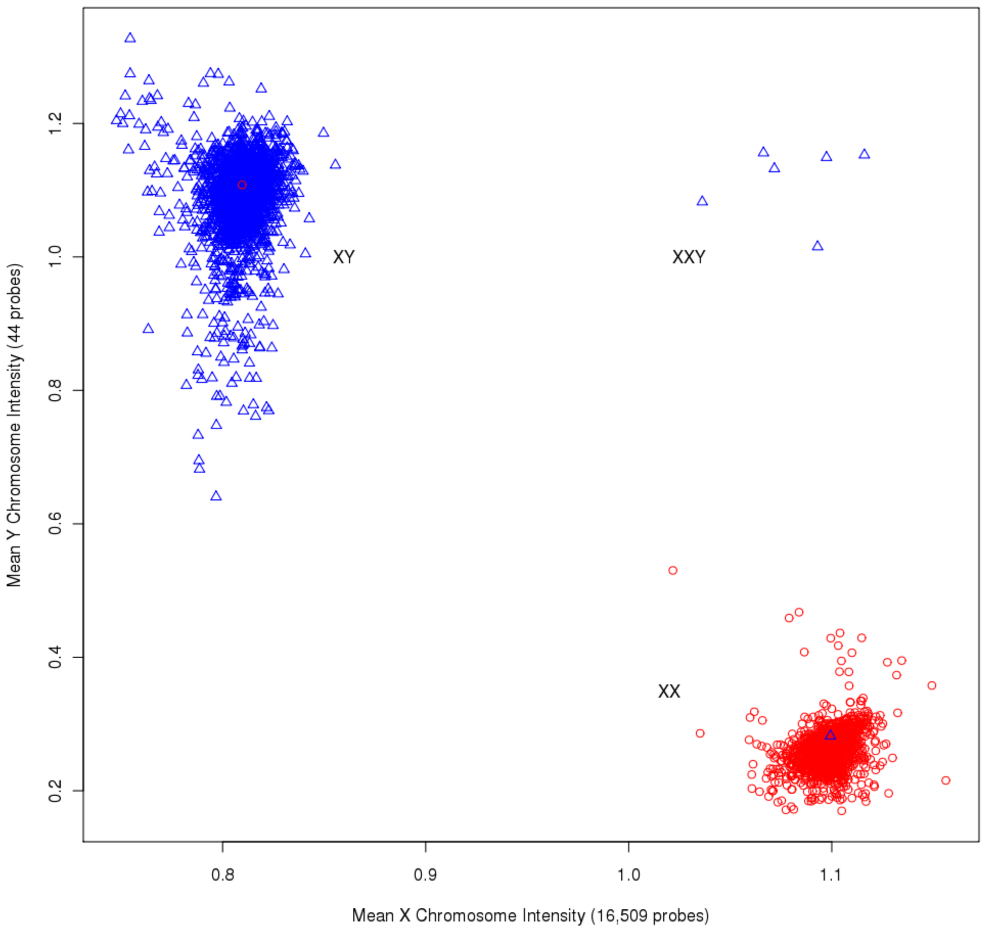


Figure 10. Visualization of X and Y probe intensities. The x-axis and y-axis represent the sum of the average over all probes for the normalized Cartesian intensity for allele A and the average over all probes for the normalized Cartesian intensity for allele B using all probes available on X chromosome and Y chromosome, respectively. The XX (female, red circles) and XY (male, blue triangles) subjects are shown on the bottom right corner and on the top left corner, respectively. The plot reveals two mislabeled individuals (one male with the female cluster, and one female with the male cluster). Several XXY individuals are also clearly visible (upper right corner).

chromosomal variation can be identified using intensity only probes to calculate loss of heterozygosity and abnormal copy numbers using B allele frequency and Log R Ratio plots (Figure 11). The B allele frequency plot is the amount of B allele observed in a probe that should concentrate at zero for zero copy, at 0.5 for one copy and at 1 for two copies. Log R ratio is the ratio of a particular sample overall all samples. It is expected to concentrate at zero; sometimes an upward or downward bump is observed meaning amplification or deletion, respectively. These two plots can be obtained using Illumina BeadStudio/GenomeStudio.

Depending on the aims of the study, often these individuals are not eliminated from the study due to sex chromosome anomalies alone. Even in carefully collected samples, the numbers of samples with discrepant self-reported sex having a normal karyotype is appreciable, and sample processing pipelines need to have these checks in place to detect such potential sample swaps. While it may be possible to go back to the EMR to reconcile chromosomal anomalies found using genetic data, researchers need to be aware of any ethical issues that may arise concerning return of results, and these issues should be considered and resolved prior to revisiting the EMR.

Sample relatedness

Another way to simultaneously examine both sample identity and pedigree integrity is by reconciling genomic data with self-reported relationships between individuals (if available). The Marshfield Personalized Medicine Research Project (PMRP) cohort enrolled many participants who were related to other individuals in the dataset. Although this has consequences that impact which analytical approaches are appropriate for the downstream association study, having related samples in the dataset makes it possible to further investigate potential DNA sample mix-ups. Using dense marker data obtained in GWAS it is easy to compute pairwise kinship estimates between every individual in the study using the *genome* option in PLINK. This procedure was not performed on the entire GWAS dataset – using only 100,000 markers yielded

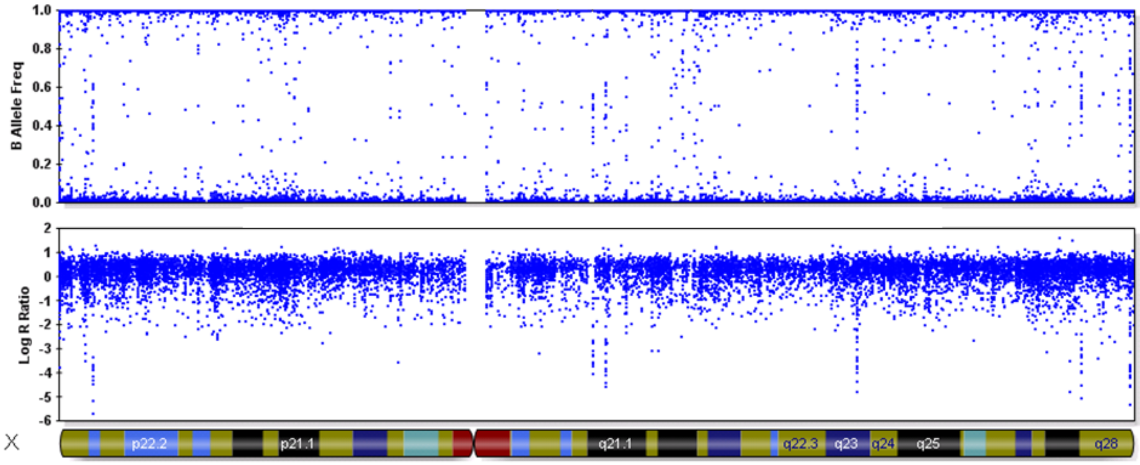


Figure 11. Copy Number and allelic variation to detect anomalies on X chromosome. The top plot shows the B-Allele frequencies for all probes for one sample with total loss of heterozygosity (LOH) on X chromosome. The bottom plot shows the copy number variation from the same sample on X chromosome. Both plots are helpful to detect regions of LOH and/or copy number variation such as deletion and amplification. Log R ratio is expected to concentrate at zero; sometimes an upward or downward bump is observed meaning amplification or deletion, respectively.

stable estimates of kinship coefficients. In addition to reporting the relationship type as reported using pedigree data (e.g. siblings, parent-child, unrelated), this procedure also calculated the proportion of loci where two individuals share zero, one, or two alleles identical by descent (IBD). Individuals sharing two alleles IBD at every locus are either monozygotic twins, or the pair is actually a single sample processed twice. Individuals sharing zero alleles IBD at every locus are unrelated. Individuals sharing one allele IBD at every locus are parent-child pairs. On average, siblings share zero, one, and two alleles IBD at 25%, 50%, and 25% of the genome, respectively. Using these data, the proportion of loci sharing one allele IBD (the Z1 column) by the proportion of loci where individuals share zero alleles IBD (column Z0) was plotted and points color coded by the relationship type. For clarity, this plot was restricted to points where the overall kinship coefficient is ≥ 0.05, as most of the individuals where kinship ≤ 0.05 will be unrelated. This relatedness plot for the Marshfield PMRP cohort is shown in Figure 12. Detailed information on producing this graphic using R (R Development Core Team, 2005) can be found online (Turner, 2009). If it is believed that pedigree records obtained through ascertainment or through the EMR are accurate, then a point out of place (e.g. points colored as unrelated showing up where most of the parent-offspring pairs cluster) would be indicative of either nonpaternity, adoption, sample mix-up, or duplicate processing of a single individual. Further investigation using the EMR or ascertainment records can be used to attempt to identify the problem. It is also worth noting in studies where datasets from multiple sites are combined that it is possible that the same participant is present in more than one study. These two data points would appear genetically identical across sites.

In addition to potentially discovering sample handling issues, visualizing sample relatedness as shown in Figure 12 also reveals any cryptic relatedness that may be present in the study sample. Figure 12 shows that many individuals in the Marshfield PMRP cohort who indicated that they were unrelated (black points) or distantly related (blue points) line up along the diagonal in this plot. These individuals represent second, third, fourth, and fifth degree

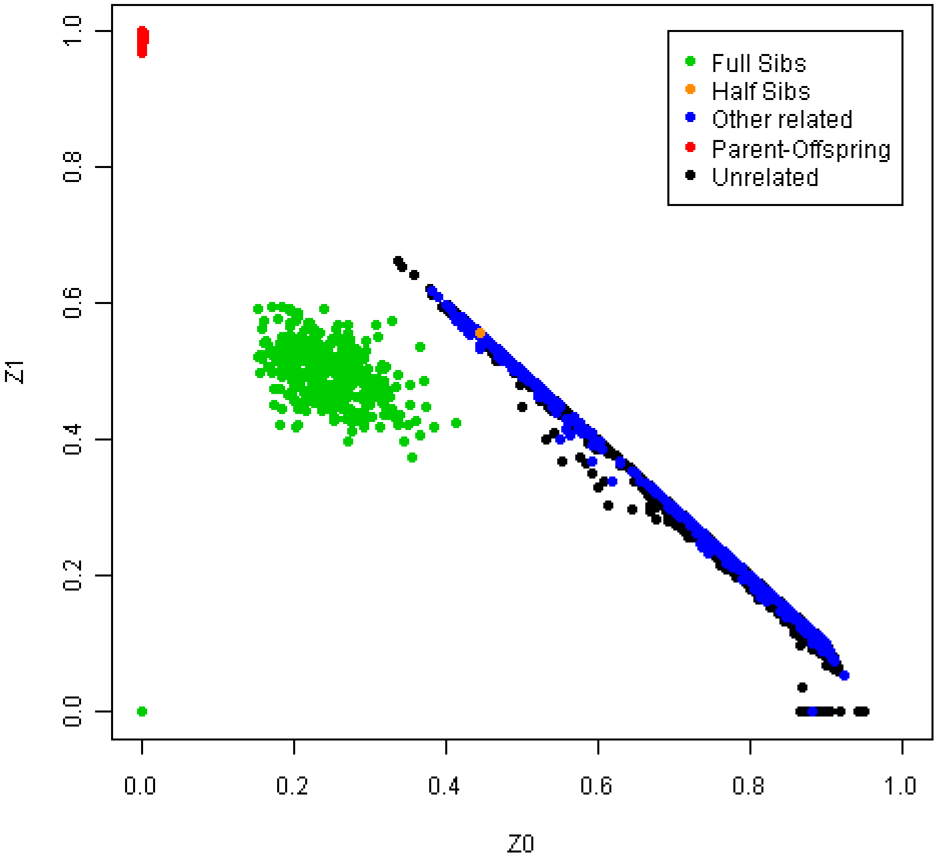


Figure 12. Relatedness plot for Marshfield PMRP cohort. Points in this plot show pairs of individuals plotted by their degree of relatedness: the proportion of loci where the pair shares one allele IBD (Z1) by the proportion of loci where the pair shares zero alleles IBD (Z0). These values are obtained from PLINK using the *genome* option. Pairs are color-coded by the type of relationship determined by the pedigree information embedded in the pedfile (also reported by PLINK). This plot omits pairs of individuals having an overall kinship coefficient < 0.05 for clarity. There is a pair of monozygotic twins represented by a point in the lower left at (0,0), because they share two alleles IBD at every locus across the genome.

relatives. If treated as independent samples in the downstream analyses, having many related samples in the dataset would result in increased type I and type II errors, thus analytical methods such as mixed model regression (Aulchenko, de Koning, & Haley, 2007) must be used in place of simple linear or logistic regression. Figure 13 shows another way to visualize the degree of relatedness in the Marshfield PMRP cohort by plotting a histogram of the distribution of kinship coefficients over 0.05 between all pairs of individuals in the dataset.

Population substructure

Population stratification occurs when the study samples comprise multiple groups of individuals who differ systematically in both genetic ancestry and the phenotype under investigation. Spurious apparent associations would be due to differences in ancestry rather than true association of alleles to disease (Cardon et al., 2003). Thus it is critical to check for population stratification within the study samples and leverage this information to inform the downstream analyses.

One strategy for avoiding bias induced by population stratification is to ensure that study samples are drawn from a relatively homogenous population as discussed in Chapter I. One of the sites in the eMERGE network represents such a sample, as over 98% of the study sample self-reported "Caucasian," on a study questionnaire. This percentage is consistent with data from the 2000 Census (2000), and self-report often shows very high correspondence with genetically inferred ancestry (Tang et al., 2005). Some clinics, such as the Vanderbilt BioVU dataset discussed in Chapter V, record ethnicity via observer-report (typically a clerk or nurse's aide). Even in this settings, observer-reported ancestry closely matches genetically inferred ancestry, especially for populations of European descent (Dumitrescu et al., 2010). However, as discussed in Chapter I, population-based diverse samples are often desirable for genetic association studies focused on characterizing previous GWAS or candidate gene discoveries made in one population (Manolio, 2009). The eMERGE network is currently combining samples

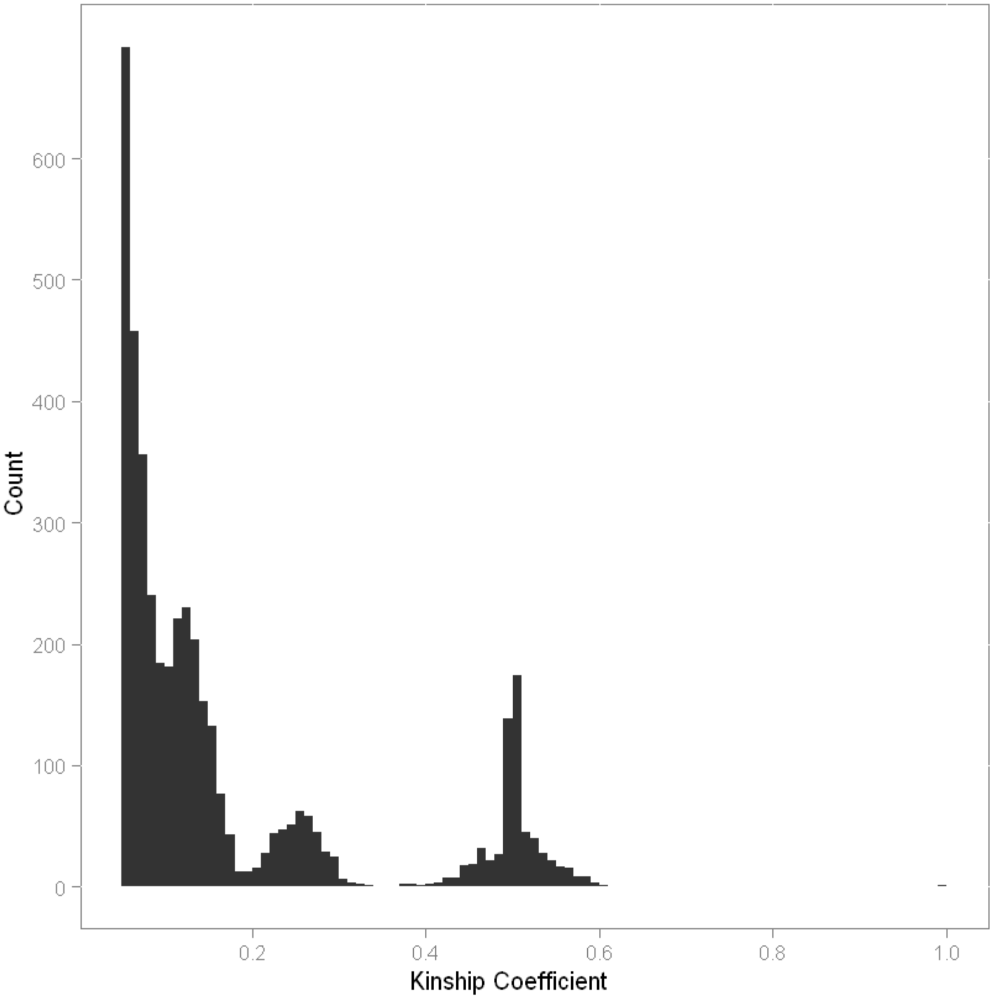


Figure 13. Kinship coefficient distribution. This histogram shows the distribution of pairwise kinship coefficients (where kinship coefficient is greater than 0.05). The peak over 0.5 represents first degree relatives (parent-offspring, full siblings). The peak over 0.25 represents second degree relatives (half siblings, avuncular, grandparent-grandchild). Third and fourth degree relatives begin to blend into more distantly related samples between zero and 0.125.

from all five sites for a joint analysis, which may result in population stratification in the combined sample, if both allele frequencies and outcomes differ between sites.

Statistical methodology for detecting and adjusting for population stratification in GWAS was covered in great detail in Chapter I. Briefly, genomic control (Devlin et al., 1999; Reich et al., 2001), aims to control for population stratification by first estimating an inflation factor, then adjusting all of the test statistics downward by this factor. Structured association (Pritchard et al., 2000), uses genotype data to infer population structure and subsequently performs tests of association within each inferred subpopulation. STRUCTURE may also be used to identify individual samples that do not cluster with the majority of the samples. These samples may then be eliminated from the analysis. Eigenstrat (Price et al., 2006; Patterson et al., 2006) uses principal components analysis to explicitly detect and adjust for population stratification on a genome-wide scale in large sample sizes in a computationally efficient manner. This method may be preferred over a stratified analysis because the combined sample often yields more powerful statistical tests, even after adjusting for significant eigenvectors (Zhang, Wang, & Deng, 2008a). Using Eigensoft requires dense genotyping coverage. The eMERGE genomics working group recommends using all the default options, including 100,000 randomly chosen high-quality markers. There are several SNPs in the HLA region on chromosome 6, in the lactase locus on chromosome 2, and in the inversion regions on 8p23 and 17q21.31 common in populations of European ancestry (Novembre et al., 2008) that are sources of stratification that will often appear in the top principal components. While one may exclude these SNPs from such an analysis, it is unknown if any similar inversions exist at appreciable frequency in non-European populations. The Eigensoft analysis will result in the computation of 10 principal components. If any of these eigenvectors are significantly associated with the phenotype under study, it is recommended that these eigenvectors be used as covariates in a multiple regression model in any downstream analysis to correct for any bias due to population stratification. Alternatively, if it is expected that only a very small number of samples represent ethnic outliers in the study population, using

Eigenstrat with iterative outlier removal, and reconciling these individuals with other ancestry information such as self-report could be used to identify a coherent set of ethnic outliers (which are identified statistically as outliers *and* self-report as outliers) to potentially exclude from the analysis, rather than adjusting for many eigenvectors during the analysis only to retain a very small number of samples.

Sample genotyping efficiency / call rate

Genotyping efficiency, or call rate, is an issue which will be discussed in greater depth in the Marker Quality section below. A large proportion of SNP assays failing on an individual DNA sample may be indicative of a poor quality DNA sample, which could lead to aberrant genotype calling. Samples with low genotyping efficiency, or call rate, should be eliminated from further analysis. A recommended threshold is 98-99% efficiency, after first removing markers which have a low genotype call rate across samples. The suggested 98-99% threshold is an approximate threshold – the exact threshold may vary from study to study depending on the genotyping platform used, quality of the DNA samples used, and the variability in human and equipment error in genotyping. The threshold should be determined based on a goal whereby a balance minimizing the number of samples dropped and maximizing genotyping efficiency is attained. As discussed in Chapter V, we used a strict 99% call rate threshold in both the Marshfield PMRP and Vanderbilt BioVU cohorts described therein. Figure 14 shows the proportion of samples (red and blue lines) or SNPs (green line) remaining at different call rate thresholds using the Marshfield PMRP cohort. Genotyping efficiency can be checked using the *missing* option in PLINK. This will produce a file showing genotype missingness rate (1-efficiency) for each individual (proportion of SNPs which failed on each sample), and for each SNP (proportion of individuals for which no genotype was called). Samples below a desired threshold can be eliminated from any downstream analyses by using the *mind* option in PLINK. Genotyping efficiency is also an important marker QC step, and is discussed below.

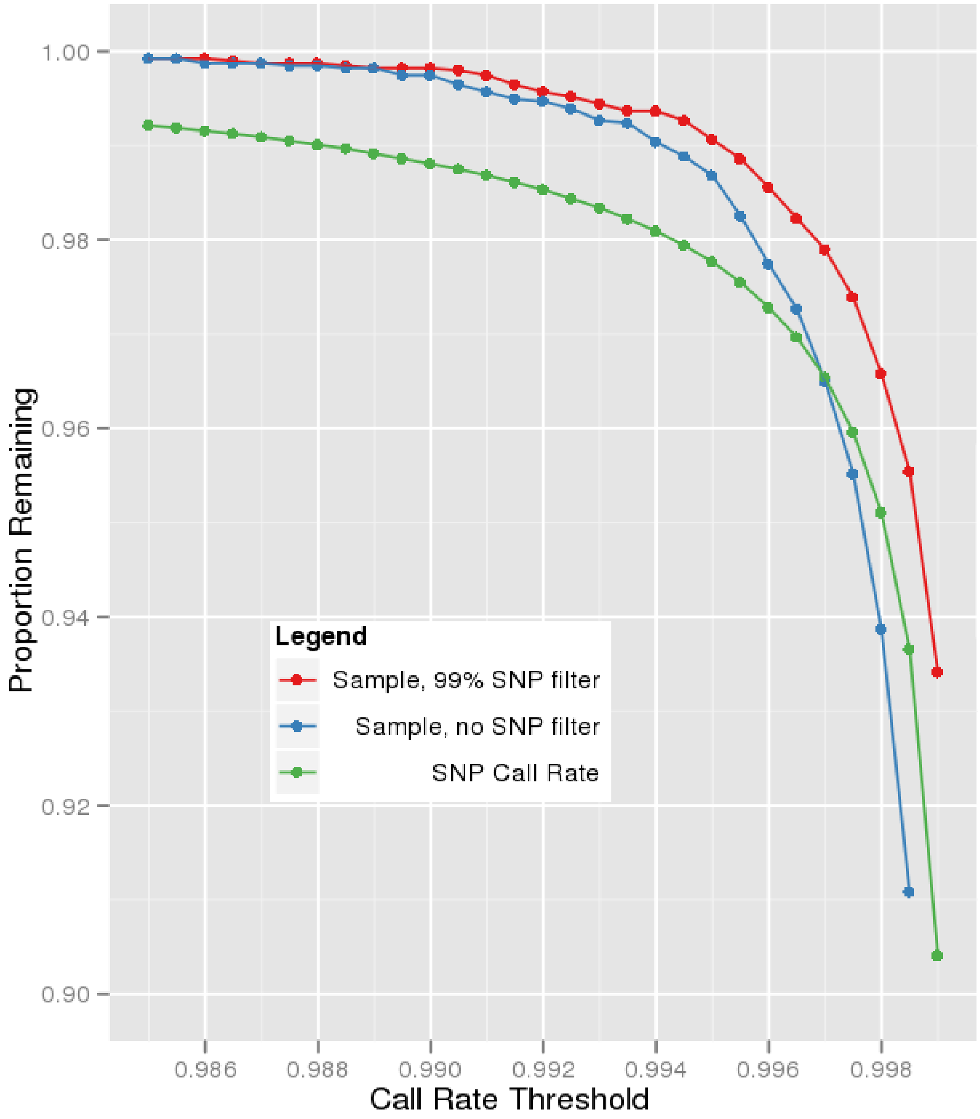


Figure 14. Proportion of SNPs or samples remaining as call rate threshold increases. The green line shows the propotion of SNPs remaining when SNPs are discarded if they fall below the given genotyping efficiency threshold. The blue line shows the proportion of samples remaining, while the red line shows the proportion of samples remaining if a 99% call rate threshold is applied to eliminate poor quality markers first.

Marker quality

<u>*Marker genotyping efficiency / call rate*</u>

As mentioned in the sample genotyping efficiency section above, marker genotyping efficiency (the proportion of samples with a genotype call for each marker) is a good indicator of marker quality. SNP assays that failed on a large number of samples are poor assays, and are likely to result in spurious calls. A recommended threshold for removing SNPs with low call rate is approximately 98-99%, although as mentioned in the sample genotyping efficiency section, this threshold may vary from study to study. Marker genotyping efficiency can be reviewed using the *missing* option in PLINK. We recommend removing poor quality SNPs before running the sample genotyping efficiency check discussed above, so that fewer samples will be dropped from the analysis simply because they were genotyped with SNP assays that had poor performance (see Figure 14). Markers can be removed based on call rate by using the *geno* option, followed by a threshold for a lower limit of missingness (e.g., *geno* 0.02 would remove SNPs with more than 2% missing, i.e. less than a 98% call rate).

<u>*Control sample reproducibility / HapMap concordance*</u>

It is advantageous to incorporate internal controls in the genotyping pipeline to estimate genotyping reproducibility rate and for selecting which markers to eliminate based on poor reproducibility. The eMERGE network routinely genotyped DNA samples from the HapMap cell lines (International hapmap consortium, 2003; International hapmap consortium, 2007). In addition to providing samples of known ancestry to anchor the Structure analysis discussed in the *Population Substructure* section above, genotype calls on HapMap samples can be compared to the corresponding reference genotypes to estimate the degree of concordance. Genotyping for the Marshfield PMRP and Group Health was performed by CIDR, which considered any SNP having more than one replicate error on HapMap samples run with the study samples to be a technical failure, and only intensity data were released for these markers. CIDR also considered SNPs technical failures if the SNP had a call rate <85%, if the absolute difference in call rate between sexes is greater than 2.5%, if the absolute difference in heterozygosity between sexes is greater than 7%, or if cluster separation <0.20. Vanderbilt BioVU, Mayo, and Northwestern NUGene samples were genotyped at the Broad Institute, where technical failure was determined by call rate 95%, GenTrain score <0.6 (a statistical measure from Illumina's clustering algorithm(2008a)), cluster separation <0.4, or more than one replicate error. It is also advantageous to build in duplicate samples to estimate the reproducibility rate within study samples. By design, both CIDR and Broad include HapMap control samples and duplicate samples across all plates in the study. It is anticipated that for accurate genotyping data, duplicate reproducibility and HapMap concordance of >99% are expected. We removed any SNPs which had one or more discordant calls on duplicate samples. Both HapMap concordance and replicate sample concordance can be checked using the concordance procedure in the PLATO software (Grady et al., 2010) (Table 4) or by using the *genome rel-check* options in PLINK. By using HapMap trio samples it is also possible to inspect each SNP for Mendelian inconsistencies, which indicate genotyping errors if pedigree information is correct. Mendelian inconsistencies can be assessed using the `--mendel` option in PLINK. PLINK only detects Mendelian inconsistencies in full trios. The Mendelian-error procedure in PLATO will also evaluate Mendelian consistency in parent-offspring pairs or sib-pairs with missing parental genotypes. We recommend removing or flagging any SNPs that have one or more Mendelian errors on HapMap control samples. While it may be possible to look for Mendelian inconsistencies using study samples, removal of these SNPs could potentially be filtering out a phenotype specific copy number variant. If this is the case, there will likely be more than three genotype clusters. The extra clusters or parts of them will be missing or miscalled. For instance, for a locus with alleles A and B, A-, AA, and AAB may all cluster together unless the SNP is re-called with a specific model pre-specified.

Minor allele frequency

It is also important to filter SNPs based on minor allele frequency because statistical power is extremely low for rare SNPs. Figure 15 shows that the power to detect an association in a large dataset (n=10,000) with a relatively large effects (OR between 1.3 and 1.7) is extremely low for rare SNPs (<1% frequency). Because power to detect an association (OR=1.5, n=10,000) at a genome-wide significance threshold is nearly zero for a SNP with frequency 1% or less, we recommend removing any extremely rare SNPs (including any monomorphic SNPs, which are completely uninformative). The threshold chosen depends on the size of the study and the effect sizes expected. Power calculation software such as CaTS Power (Skol, Scott, Abecasis, & Boehnke, 2006) or Quanto (Gauderman, 2002c) can simplify power calculations for genetic association studies and inform the investigator of the allele frequency below which the study becomes severely underpowered. Minor allele frequency can be reported for each SNP using the *freq* option in PLINK, and SNPs can be removed from the analysis using the *maf* option, followed by a lower limit threshold. SNPs with frequency too low to yield reasonable statistical power (e.g. below 1%) may be removed from the analysis to lighten the computational and multiple testing correction burden. However, in studies with very large sample sizes it may be beneficial to avoid removing these rare SNPs. Others have shown that nonsynonymous, possibly deleterious SNPs are on average rarer than synonymous SNPs that likely do not cause any adverse phenotypes (Gorlov, Gorlova, Sunyaev, Spitz, & Amos, 2008).

Hardy-Weinberg Equilibrium

Checking for Hardy-Weinberg Equilibrium (HWE) is one final step in the quality control analysis of markers in GWAS data. Under Hardy-Weinberg assumptions, allele and genotype frequencies can be estimated from one generation to the next. Departure from this equilibrium can be indicative of potential genotyping errors, population stratification, or even actual association to the trait under study (Wittke-Thompson, Pluzhnikov, & Cox, 2005). A very detailed

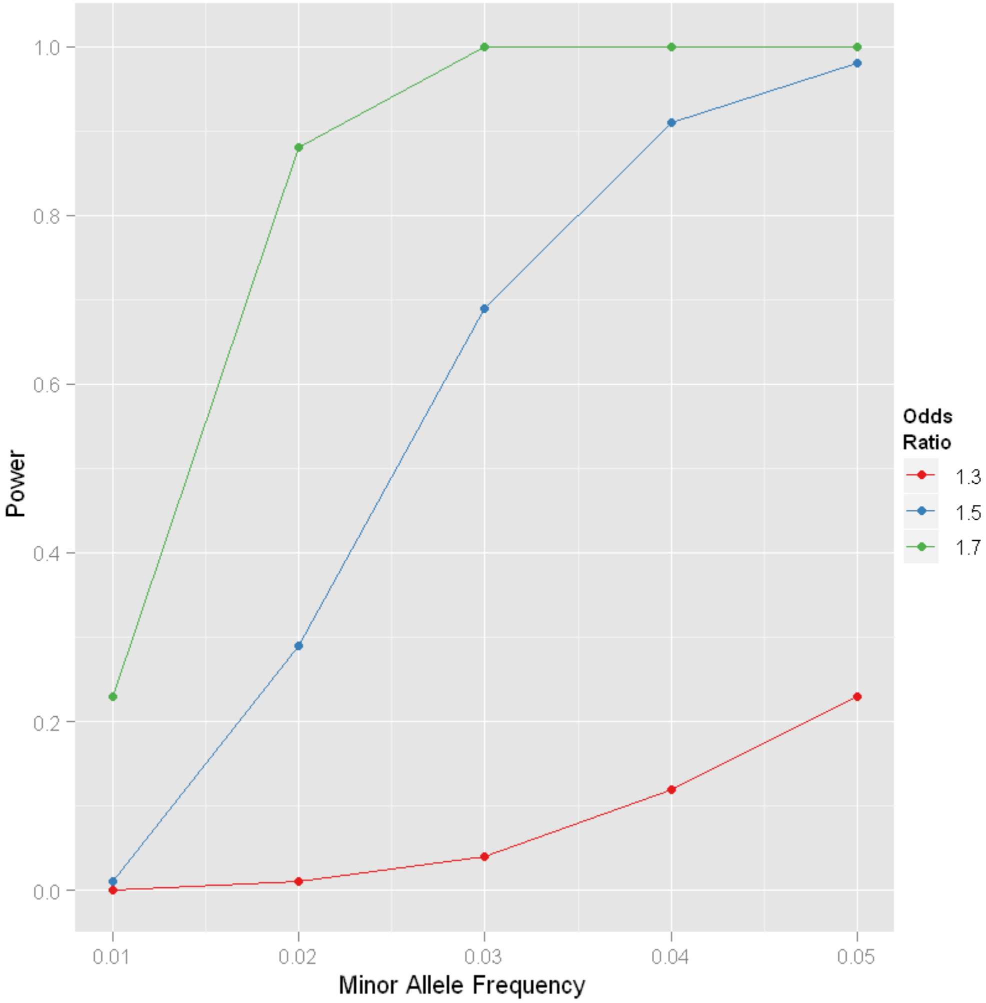


Figure 15. Power over minor allele frequency by odds ratio. This shows the power to detect an association at genome-wide significance (p<5×10-8), assuming the actual causal SNP is genotyped in a case-control study consisting of 5000 cases and 5000 controls of a common disease with 10% prevalence under an additive model at three different odds ratios. Note that when the MAF is low, power is extremely low even for very large effects (OR=1.7).

description of the key principles and assumptions of HWE and how HWE is tested and applied in genetic association studies is the subject of a review elsewhere (Ryckman & Williams, 2008). HWE can be assessed using the *hardy* option in PLINK. While departure from HWE can indicate potential genotyping error, disequilibrium can also result from a true association. The eMERGE sites and others have consistently noted that many more SNPs are out of HWE at any given significance threshold than would be expected by chance. SNPs severely out of HWE should therefore not be eliminated from the analysis, but flagged for further analysis after the association analyses are performed. Databases such as MySQL (see Table 4) can be very useful for joining association statistics with HWE statistics for easy reporting. It is also beneficial to examine HWE in controls separately, as disease-free controls should more closely follow the assumptions that lead to HWE than cases, and because some true associations are expected to be out of HWE. SNPs that are highly associated with the trait of interest which also show highly significant departures from HWE, especially in controls if the outcome is discrete, should be closely scrutinized. Typically HWE deviations toward an excess of heterozygotes reflect a technical problem in the assay, such as non-specific amplification of the target region. On many GWAS platforms the quantitative allelic signals at a marker, i.e. the intensity plot for the SNP, can be used to screen for a technical origin of the HWE deviation: null alleles can produce multimodal genotype clusters in the heterozygote clusters and one of the homozygote clusters (Figure 16) or can produce an unexpected number of samples with no signal (Figure 17), and SNPs within CNVs or segmental duplications can produce clusters of genotypes intermediate between the three expected clusters of genotype (Figure 18) (Carlson, Smith, Stanaway, Rieder, & Nickerson, 2006). In the loci depicted in figures 8, 9 and 10, chi-square tests for HWE are rejected at P-values less than 10^{-80}, so these represent the most egregious examples of the aforementioned behavior. If no technical errors are detected then a number of biologically plausible explanations exist for HWE deviations toward an excess of homozygotes: population stratification, assortative mating and inbreeding, to name a few.

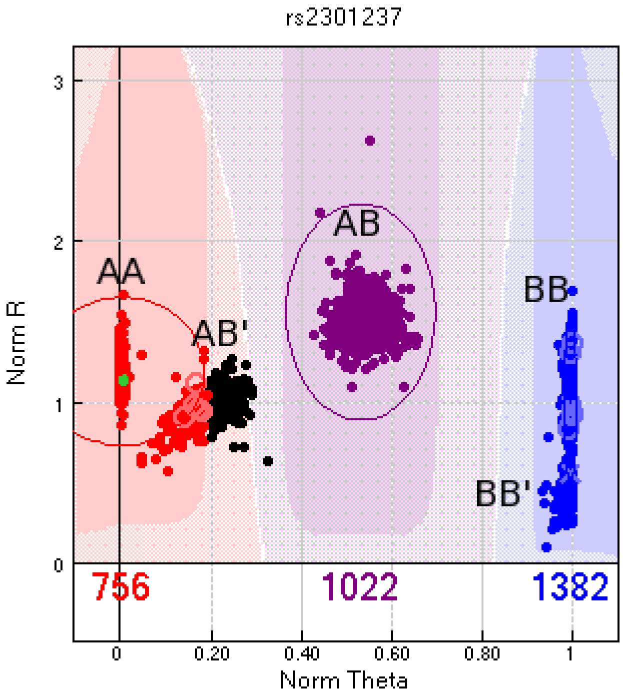


Figure 16. Genotype intensity plot showing multimodal genotype clusters. For each biallelic SNP, an Illumina assay gives raw intensity values for each of the two alleles. Polar transformation of these intensities results provides normalized intensity values (Norm R on the y-axis) and allelic intensity ratios (Norm Theta on the x-axis). These values are used to determine SNP genotypes and copy number estimates. In this example, AB and BB individuals are split into sub-clusters AB and AB', BB and BB', while AA cluster is unaffected. The AB/AB' split results in some AB samples miscalled as AA (diagnosed by Mendelian inconsistencies in the genotypes), as well deviation from HWE due to excess homozygosity. Since only samples with at least one B allele demonstrate the splitting, one consistent explanation is the presence of a cryptic polymorphism near rs2301237 on a haplotype that contains the B allele. In this case, a second polymorphism (rs3114267) lies eight bases upstream from the typed polymorphism, and is in complete LD (D'=1, r^2=.2) with rs2301237.

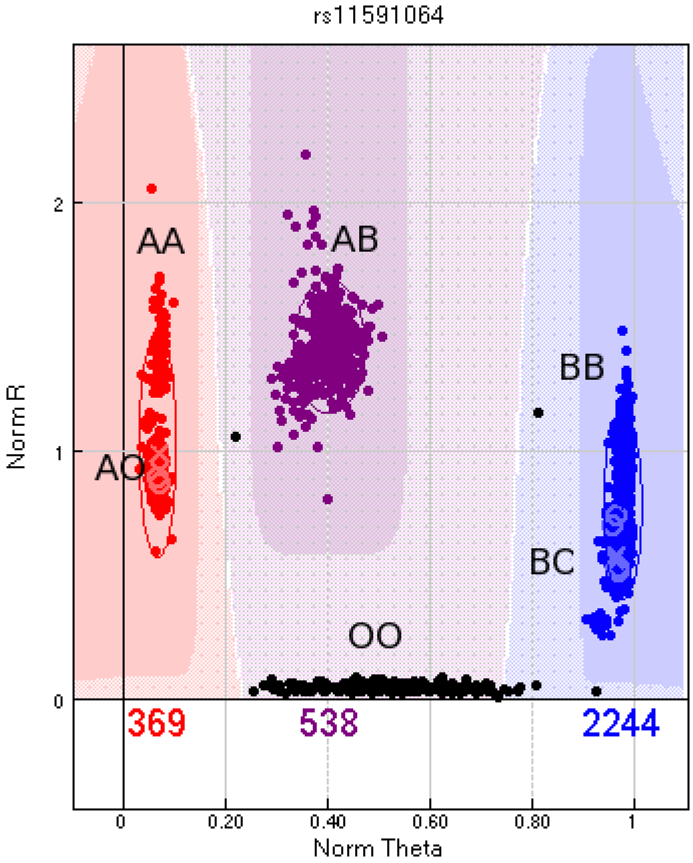


Figure 17. Genotype intensity plot for samples with no signal. Unexpected number of clusters resulting in departure from HWE consistent with copy loss. Hemizygous individuals cluster at AO and BO. Individuals with homozygous deletions cluster at OO and their genotype calls are missing. The AB cluster remains intact, since these individuals are ipso facto diploid at the locus. Parent-parent-child Mendelian errors are present when at least one parent is hemizygous and produces hemizygous offspring. The deletion results in excess homozygosity. In this case, the "copy loss" appears to be a six-nucleotide insertion (rs71578153) coincident with rs11591064 that disrupts both A and B probes.

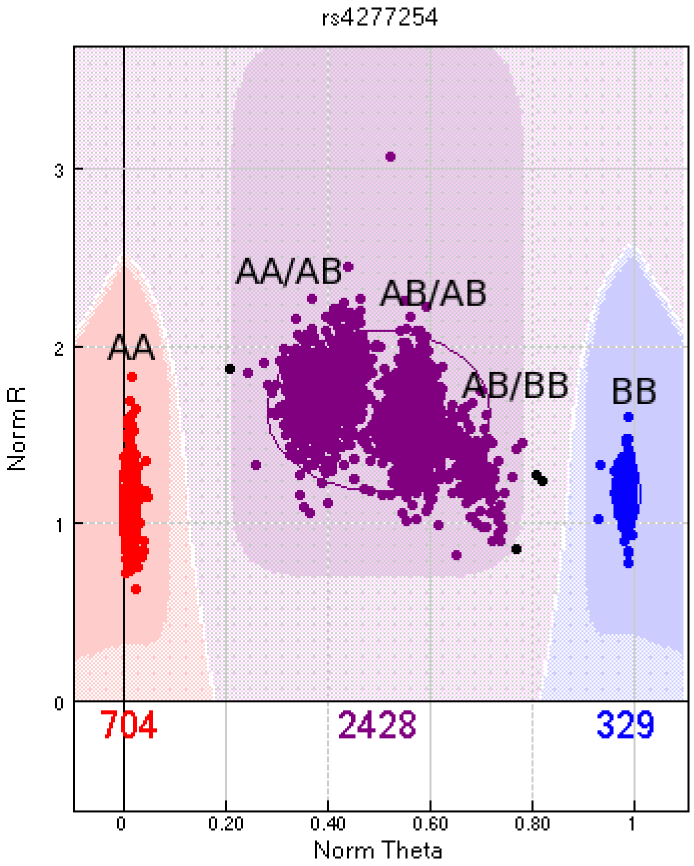


Figure 18. Genotype intensity plot showing CNVs or segmental duplications. The five observed clusters are most consistent with a segmental duplication, although none is curated around the locus. A copy number variant would be expected to produce additional clusters above the AA and BB clusters (ie, AAA and BBB), as opposed to the splits being confined to strictly the heterozygous clusters. Regardless, the artifact results in excess heterozygosity.

Batch effects

Thousands of DNA samples are typically genotyped in a GWAS, which necessitates partitioning samples into small batches of samples processed in the lab together for genotyping (e.g. the set of samples on a 96 well plate). The precise size and composition of the sample batch depend on the array and lab process used. Systematic differences among the composition of individuals in a batch (i.e. the case to control ratio or race/ethnicity of individuals on plates) and the within-plate accuracy and efficiency can result in batch effects – apparent associations confounded by batch. The problem is in essence the same problem observed with population stratification – namely that if there is an imbalance of cases and controls on a plate, and there are nonrandom (unknown) biases or inaccuracies in genotyping that differ from plate to plate – spurious associations will result.

Ideally, no batch effect will be present because individuals with different phenotypes, sex, race, and other confounders should be plated randomly, and because modern high-throughput genotyping technology is much more accurate, efficient, and consistent than earlier generations of GWAS assays. There are several approaches for examining a dataset for potential batch effects. One simple approach is to calculate the average minor allele frequency and average genotyping call rate across all SNPs for each plate. Gross differences in either of these on any plates can easily be identified. Another method involves coding case/control status by plate followed by running the GWAS analysis testing each plate against all other plates. For example, the status of all samples on plate or batch 1 will be coded as case, while the status of every other sample is to be coded control. A GWAS analysis is to be performed (e.g. using the *assoc* option in PLINK), and both the average p-value and the number of results significant at a certain threshold (e.g. $p<1\times10^{-4}$) can be recorded. SNPs with low minor allele frequency (i.e. <5%) should be removed before this analysis is performed to improve the stability of test statistics. This procedure should be repeated for each plate or batch in the study. If any single plate has many more or many fewer significant results, or has an average p-value that deviates from 0.5 (under the null the average p-value will be 0.5 over many tests), then this batch should be further investigated for genotyping or composition problems. If batch effects are present, methods similar to those used for population stratification (e.g. genomic control) may be used to mitigate the confounding effects.

Evaluation of QC after association analysis

After phenotypic association analysis, the quality control measures used should be evaluated. One method is to compare the observed number of statistically significant results with a phenotype to expectation to produce a QQ-plot as shown in Figure 6 in Chapter III. Too many significant results may indicate insufficient QC. Also because no QC will catch all problematic SNPs, the intensity plots for statistically significant SNPs must be reviewed to make sure there are no obvious clustering problems. Replication of results using different genotyping technology (such as TaqMan) and/or in another sample may be needed as well.

Future directions

The QC pipeline developed by the eMERGE network has enabled a thorough analysis of the quality of the genome-wide genotype data generated on the ~17,000 samples in the eMERGE network (this includes the samples used in the Marshfield PMRP and Vanderbilt BioVU samples discussed in Chapter V). All of these data have been deposited in dbGaP along with corresponding quality control documents that describe all of the QC details for each dataset individually. Conducting QC in parallel at the eMERGE coordinating center and study sites has been a tremendously valuable experience, as it led to a more thorough understanding since each group had to reconcile its results with others. Additionally, the genomics group is currently performing a second round of QC, whereby the 17,000 subjects are being merged into one large dataset, and all of the previous QC steps will be performed again on the merged dataset. It is not clear what novel characteristics of the genotype data will be revealed as we explore this very large combined dataset. We will compare results of various QC procedures in the individual study sites and the merged eMERGE dataset to determine what, if anything, is learned in the process of merging raw data as opposed to performing a meta-analysis of the QC results.

Acknowledgements

This research was supported in part by NIH grants U01HG004608, U01HG004609, U01HG004610, U01HG04599, U01HG04603, U01HG004438, R01LM010040, and by the Intramural Research Program of the NIH, National Library of Medicine.

CHAPTER V

KNOWLEDGE-DRIVEN MULTI-LOCUS ANALYSIS REVEALS GENE-GENE INTERACTIONS INFLUENCING HDL CHOLESTEROL LEVEL IN TWO INDEPENDENT EMR-LINKED BIOBANKS[5]

Introduction

To date, nearly 600 genome-wide association studies (GWAS) have been completed, investigating 150 distinct complex human traits (Hindorff et al., 2009; Manolio, 2010). Circulating levels of high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and triglycerides (TG) are quantitative traits commonly measured in clinical practice and strongly associated with vascular disease, making them appealing traits to investigate from a statistical, clinical, and practical standpoint (Edmondson & Rader, 2008). The genetic factors underlying variability in blood lipid levels have been extensively studied using GWAS (Aulchenko et al., 2009; Chasman et al., 2008; Heid et al., 2008; Johansen et al., 2010; Kathiresan et al., 2007; Kathiresan et al., 2008; Kathiresan et al., 2009; Kooner et al., 2008; Sabatti et al., 2009; Sandhu et al., 2008; Saxena et al., 2007; Wallace et al., 2008; Willer et al., 2008b).

There is particularly strong interest in the characterization of genetic factors underlying population variability in HDL-C (Wilke et al., 2010). In human populations, every 1 mg/dl decrease in HDL-C is associated with a 6% increase in cardiovascular risk (Ashen & Blumenthal, 2005). HDL particles also appear to have direct anti-atherogenic properties in animal models (Rubin, Krauss, Spangler, Verstuyft, & Clift, 1991). Therefore, while these smaller particles may serve as a source of cholesterol esters for the larger, more atherogenic LDL particles, the HDL particles themselves actually appear to attenuate the development of cardiovascular disease (2002). HDL is under tight genetic control (h^2 up to ~70% as determined by family studies)

[5] Adapted in part from: Turner S.D., Berg R.L., Linneman J.G., Peissig P.L., Crawford, D.C., Denny J.C., Roden D.M., McCarty C.A., Ritchie M.D., Wilke R.A., Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks, *PLoS ONE*, submitted.

(Zhang et al., 2009), yet despite HDL's high heritability, even some of the most well powered GWAS studies have only explained a very small proportion of HDL variation using common SNPs (Kathiresan et al., 2008; Willer et al., 2008a; Sabatti et al., 2009). This net unexplained variation due to genetics, often termed the "missing heritability" problem (Maher, 2008), has challenged GWAS studies for many complex traits beyond circulating lipid levels (Goldstein, 2009; Manolio, 2010).

A general explanation for this missing heritability is that it reflects forms of genetic variation that are not captured in the GWAS paradigm; these include rare genetic variation and copy number polymorphisms (Maher, 2008; Manolio, 2010), and gene-gene (GxG) and gene-environment (GxE) interactions (Eichler et al., 2010; Manolio et al., 2009; Maher, 2008; Manolio, 2010). GxG interaction (epistasis) is thought to be an important component of complex, multifactorial diseases due to the complexity of biological systems (Tyler et al., 2009). Data from animal models provide compelling support for the role of GxG interaction in the control of complex traits (Shao et al., 2008; He et al., 2010). Exploration of GxG in GWAS is often limited by lack of large sample sizes and statistical methods. One possible solution to the sample size problem is presented by the growing number of DNA repositories linked to electronic health records. These resources can provide cohorts of sufficient size for the characterization of GxG interaction. In parallel, computational capacity and novel methodologies have emerged to make the search for epistasis in GWAS feasible (Cordell, 2009; Turner, Crawford, & Ritchie, 2009).

Here we present data from a GWAS analyzing HDL-C using the Marshfield Clinic Personalized Medicine Research Project (PMRP) database (McCarty et al., 2010b), a node of the NHGRI-funded eMERGE network (*e*lectronic *M*edical *R*ecords and *Ge*nomics) [www.gwas.org]. We first conducted a genome-wide scan for SNPs associated with median HDL-C level using the comprehensive electronic medical records (EMRs) of 3947 PMRP participants. We also constructed a modeled HDL phenotype that accounts for environmental effects such as population trends in age, body mass index (BMI), and relevant co-morbidities (Wilke et al., 2010).

We next investigated GxG interaction in this same dataset. One frequently-cited causal mechanism underlying GxG interaction has been variability within multiple genes in similar pathways, protein families, or genes with similar or redundant biological function (Aguilar et al., 2010; Costanzo et al., 2010). We applied the Biofilter (Bush et al., 2009), a new bioinformatics technique that leverages domain knowledge from publicly available biological databases, to systematically investigate GxG interactions in this cohort. We required that any GxG interactions in this cohort (n = 3740 PMRP participants) showed evidence of replication in a second cohort from the eMERGE network (n = 1858 records from the Vanderbilt BioVU project (Ritchie et al., 2010)). This resulted in replicated GxG interactions associated with variation in HDL-C; all of which have potential biological relevance.

Methods

For the discovery cohort in the Marshfield PMRP, the study was approved by the Institutional Review Board of the Marshfield Clinic, and conducted in accordance with the basic principles of the Declaration of Helsinki. All study subjects in the discovery cohort provided informed consent allowing access to their entire electronic medical record, through their participation in the Marshfield Clinic Personalized Medicine Research Project (PMRP). The replication cohort was constructed from a de-identified mirror image of the electronic medical record at Vanderbilt University (BioVU). BioVU accrues DNA samples extracted from discarded blood samples from routine clinical care. The project has been reviewed and approved at multiple levels, including the Institutional Review Board, internal and external ethics committees, Community Advisory Board, legal department, and the Federal Office of Human Research Protection, and this oversight is ongoing (Roden et al., 2008). The PMRP database and BioVU are both eMERGE nodes, and are two of the largest practice-based biobanks in the U.S. with over 20,000 in the PMRP and over 95,000 in BioVU as of September, 2010 (McCarty, Wilke, Giampietro, Wesbrook S., & Caldwell, 2005; McCarty, Mukesh, Giampietro, & Wilke, 2007) (Ritchie et al., 2010).

Phenotyping

To facilitate the construction of accurate prediction models for cardiometabolic risk, the eMERGE network has begun extracting clinical lipid data from multiple participating sites. The Marshfield Clinic in Central Wisconsin has one of the oldest internally developed EMRs in the US, with coded diagnoses dating back to the early 1960's and laboratory observations dating back to 1985. The EMR data collected for clinical care is transferred daily into the Marshfield Clinic Data Warehouse where it is made available for research. We modeled lipid variables using this data source for the participants in the Marshfield PMRP cohort (Wilke et al., 2010).

The PMRP demographics reflect the composition of the Central Wisconsin community, and the corresponding dataset has a distribution of fasting lipid levels similar to that reported by NHANES III (Carroll et al., 2005) for European Americans (non-Hispanic whites). At present, the PMRP Biobank contains data from over 20,000 adult participants; more than 10,000 individuals (54%) have impaired fasting glucose; >8,000 (41%) have hypertriglyceridemia; and >9,000 (48%) have reduced levels of HDL cholesterol according to criteria published by NCEP ATP-III (2001; Johnson & Weinstock, 2006).

The eMERGE network design includes selection of ~3,000 subjects with a predesignated phenotype at each node for genome-wide genotyping. In PRMP, the primary phenotype was cataract (n=3,947 subjects). This phenotype resulted in a set enriched for older adults (age range 52 to 90 years, mean age 72). The set included 3,740 individuals with at least two HDL measurements available for use in the present analysis. Due to the longitudinal nature of these data (2 to 78 lipid data points per individual; mean = 14.4 ± 10.1 data points), we defined two phenotypes to be used in the analyses below: (1) median HDL and (2) modeled HDL as previously described (Wilke et al., 2010). All HDL measurements for every individual were extracted from the EMR.

For the single-locus analysis of HDL level in the Marshfield PMRP, a simple median HDL level was computed for every individual who had at least two or more HDL datapoints in the record. Analyses using median HDL included adjustments for smoking, age, age^2, BMI, BMI^2, and gender, as all of these factors showed highly significant associations with median HDL. A second HDL measurement, termed here as modeled HDL, was also calculated for individuals in the Marshfield dataset. To calculate a modeled HDL for an individual, we extracted all lipid data and censored any HDL data acquired after lipid treatment or the onset of a relevant co-morbidity, and then used population-based trends in age and BMI to adjust individual estimates (Wilke et al., 2010). Records were censored at the first date of diagnosis for the following clinical co-morbidities known to influence circulating lipid levels: cancer, diabetes mellitus, and thyroid disease. Cancer (any malignancy except two common skin cancers - basal cell carcinoma and squamous cell carcinoma of the skin) was censored based on ICD-9 codes in the EMR. Thyroid disease and diabetes were assessed from the EMR using an electronic phenotyping algorithm available at the eMERGE website (www.gwas.org). Smoking status was acquired via information provided on a questionnaire provided at study entry, later confirmed by interview. Records were also censored at the first date of prescription for medications known to alter circulating lipid levels (either therapeutically or through an indirect interaction), including statins, fibric acid derivatives, niacin, and exogenous gonadal steroids. These data were assessed from the EMR using natural language processing (NLP) (Peissig et al., 2007; Wilke et al., 2008; Xu et al., 2010; Chen, Hripcsak, Xu, Markatou, & Friedman, 2008). These algorithms have been validated and published (Peissig et al., 2007), and the programming pseudocode is freely available through the eMERGE network (www.gwas.org).

For the gene-gene interaction analysis, we replicated findings using independent samples from BioVU, the EMR-derived DNA databank at Vanderbilt University Medical Center (Ritchie et al., 2010; Denny et al., 2010b). The primary eMERGE phenotype at this site is variability in the duration of the QRS complex on the normal electrocardiogram. Replication of the gene-gene interaction analysis in the Marshfield PMRP was conducted in 1,858 individuals from the Vanderbilt BioVU dataset having at least two HDL measurements extracted from a de-identified synthetic derivative of the EMR. Because the additional clinical data mentioned above were not available for samples in this dataset, no modeled HDL could be computed on these data, and no adjustments were made in the gene-gene interaction analysis. It is, however, unlikely that any of these variables will confound any association to a gene-gene interaction replicating across both the Marshfield PMRP and Vanderbilt BioVU samples.

Genotyping and Quality Control

Genotyping for the Marshfield PMRP samples was performed as part of the eMERGE network at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. The Illumina Human660W-Quadv1_A genotyping platform was used for this study. This platform consists of 560,635 SNPs and 96,731 intensity-only probes. Genotyping calls were made at CIDR using BeadStudio version 3.3.7. Our discovery cohort includes 3,947 samples from the Marshfield PMRP, 21 blind duplicates, and 85 HapMap controls. The HapMap controls include 44 CEPH, 32 Yoruba, 5 Japanese, and 4 Han Chinese; 40 independent HapMap replicate pair experiments, 19 independent HapMap trio experiments and 8 independent parent-child pairs were defined. The HapMap concordance rate was 99.8%. The blind duplicates reproducibility rate was 99.99%.

Genotyping for the Vanderbilt BioVU samples was performed as part of eMERGE primarily at the Broad Institute. A small number of BioVU samples, where the primary phenotype of interest was dementia were genotyped at CIDR. Similar to the PMRP samples, most samples were genotyped for SNPs on the Illumina Human660W-Quadv1_A platform, although a subset of these samples were genotyped on the Illumina 1M-Duo BeadChip (those who were known to be African American), which included all the SNPs on the 660W platform in addition to extra SNPs that were not considered in this analysis.

Genome-wide SNP data were cleaned using the eMERGE quality control (QC) pipeline developed by the eMERGE Genomics Working Group (Turner et al., 2010a). This process includes evaluation of sample and marker call rate, gender anomalies, duplicate and HapMap concordance, batch effects, Hardy-Weinberg equilibrium, sample relatedness, and population stratification. We also removed any SNPs with a minor allele frequency less than 1%, as power to detect associations with these variants was low. After QC and minor allele frequency filtering, 522,204 SNPs were used for the single locus analysis in the Marshfield PMRP dataset. Seven individuals were removed due to low (<99%) call rate. Reconciling ancestry using self-report, principle components (Price et al., 2006), and structured analysis resulted in exclsuion of a further 37 genetic ancestry outliers. After excluding these samples, none of the top ten principal components were significant (Price et al., 2006) or explained more than one tenth of one percent of HDL variation, and so no adjustment for principal components of ancestry was used in this sample. For the gene-gene interaction replication set using Vanderbilt BioVU samples, we identified and excluded 5 ethnic outliers and adjusted analyses for two marginally significant principle components using Eigensoft (Price et al., 2006). All genotype data and detailed documentation of the quality control procedures summarized here have been deposited and are available at dbGaP (Mailman et al., 2007).

Statistical analysis

More than half of the individuals in the Marshfield dataset had a first, second, or third degree relative also in the dataset. To allow for the inclusion of related individuals without inflating the type I error rate in the single-locus analysis, we performed the GWAS analysis using a linear mixed effects analysis (Aulchenko et al., 2007) with GenABEL (Aulchenko, Ripke, Isaacs, & van Duijn, 2007) implemented in the R statistical computing environment (R Development

Core Team, 2005). Residuals from this model are essentially free from familial correlations, and can be used in a simple linear regression model for each SNP, which was performed using PLINK (Purcell et al., 2007). Genome-wide statistical significance was determined using a Bonferroni correction. Figure 23 and Figure 24 were produced using the LocusZoom software (Willer et al., 2010).

For analyses of gene-gene interaction, the Biofilter is a bioinformatics algorithm that leverages domain knowledge from publicly available biological databases among SNPs from biologically plausible gene sets sharing physiological or biochemical similarity (or that have been previously associated with the phenotype under investigation) (Bush et al., 2009). The rationale for using the Biofilter is to reduce both the computational and multiple testing burdens inherent in testing for gene-gene interactions. The Biofilter quantifies evidence in support of a particular gene-gene interaction model by counting the number of public database sources that independently link the two genes within a similar biological mechanism. Choosing to be conservative, we required that any gene-gene model be independently supported by four sources of biological knowledge among the following six: the Gene Ontology(Ashburner et al., 2000); the Database of Interacting Proteins(Xenarios et al., 2002); the Protein Families Database(Bateman et al., 2004; Finn et al., 2008); the Kyoto Encyclopedia of Genes and Genomes (KEGG)(Kanehisa et al., 2000); Reactome(Vastrik et al., 2007); or NetPath(2008b). This criterion identified 22,769 SNP-SNP interactions that were tested for association to HDL-C levels in the Marshfield PMRP, and significant results were followed up using the Vanderbilt BioVU cohort. Figure 19 is a flow diagram illustrating an overview of the analysis plan.

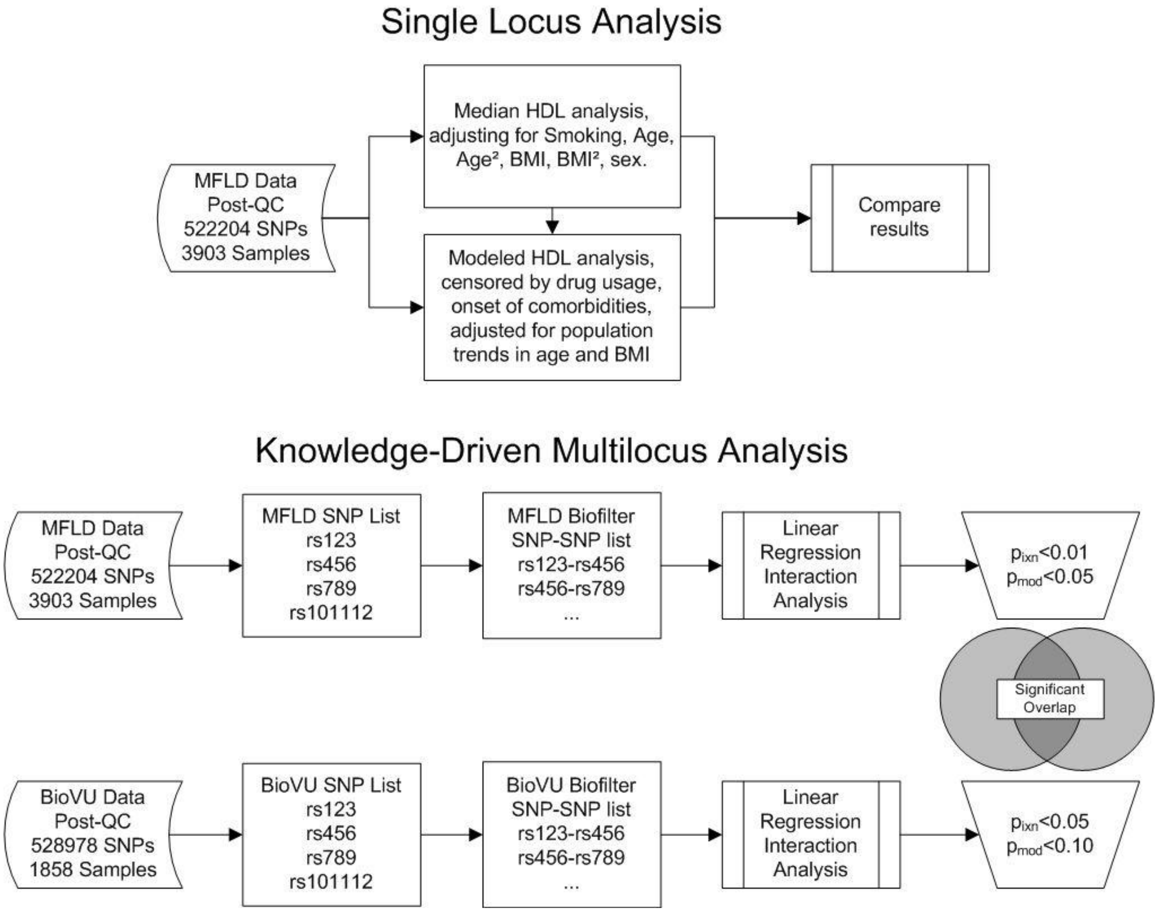


Figure 19. Flow diagram overview of the analysis plan. For the single locus analysis in the Marshfield PMRP cohort, a genome-wide association study was performed for both median and modeled HDL-C (see Phenotyping section in Methods). For the multilocus analysis, the Biofilter was used to generate putative multilocus interaction models that were tested for association with median HDL-C in the Marshfield PMRP cohort. The Vanderbilt BioVU cohort was used for replication. See the Methods and Results sections.

Results

Sample Summary Statistics

Genome-wide association analyses were conducted in the Marshfield PMRP cohort using 522,204 SNPs which passed rigorous quality control procedures established by the eMERGE network. This analysis was performed for both median adjusted HDL-C and the modeled HDL-C phenotype (censored according to medication exposure and relevant co-morbidities). Table 6 summarizes the clinical characteristics of the Marshfield population. There were 1,541 male (41%) and 2,199 female study subjects (total 3,740) with clinical laboratory records containing at least two fasting HDL measurements that could be used in the median HDL-C analysis. Of these 3,740 samples, 2,190 had used statins, 883 have had hypothyroidism or hyperthyroidism, 733 have had cancer, 1,515 reported having ever smoked, and 301 are current smokers.

Nested within this group of 3,740 unique individuals, there were 1,142 male and 1,386 female study subjects (total 2,528) with a clinical laboratory record complete enough to construct a more rigorous modeled HDL-C phenotype (censored according to medication exposure and relevant co-morbidities). Although the sample size for modeled HDL-C was lower than the sample size for median adjusted HDL-C (because censoring left no analyzable data for some individuals), we performed the analyses using both phenotypes. Our reasoning was that modeled HDL-C has the potential of revealing novel associations otherwise masked by the presence of anabolic/catabolic disorders and/or clinical intervention.

Gender was the most reliable clinical predictor of median HDL-C level in the Marshfield PMRP cohort. The average HDL-C concentration (± standard deviation) was 45.9±10.9 mg/dl in males, and 58.5±14.3 mg/dl in females (Figure 20), p=2e-163. This observation is consistent with the established literature (Johnson et al., 2006; 2002). We therefore also included gender as a covariate in the model for each SNP association. We fit a linear regression model for each SNP assuming an additive allelic model. To allow for related samples without inflating the type I

Table 6. Descriptive statistics of clinical variables in the Marshfield PMRP cohort.

	N	Median	Mean	SD	Min	Max
Age (years)	3,964	72	72.16	11.03	52	90
Weight (kilograms)	3,925	80.7	82.92	18.63	34.5	186
Height (centimeters)	3,927	165.1	166.88	9.63	121.9	195.6
BMI (kg/m^2)	3,923	28.8	29.68	5.79	16.1	64.5
Modeled HDL-C (mg/dl)	2,528	49.1	50.8	13.44	19.8	114.7
Median HDL-C (mg/dl)	3,740	51	53.4	14.43	22	127.5

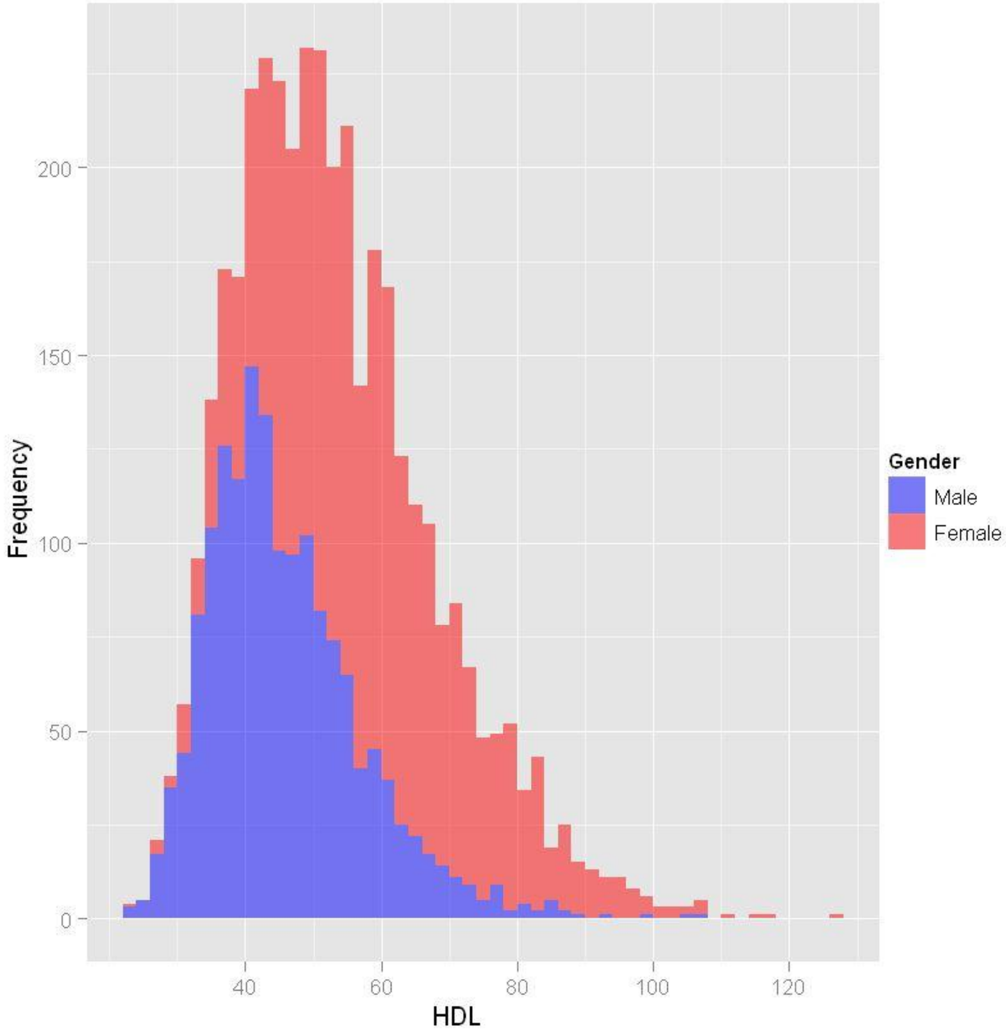


Figure 20. Distribution of HDL-C concentration. Distribution of HDL-C concentration (mg/dL) is shown for males (blue) and females (red) in the Marshfield PMRP dataset.

error rate, regression on each SNP was performed using residuals from the linear mixed effects model that allowed for a random polygenic effect (see methods section).

Single Locus GWAS Analysis: Median Adjusted HDL-C

Figure 21 summarizes the results from our genome-wide association study, showing the –log10(P-value) for each SNP from the analysis of median HDL-C level, adjusted for smoking, age, age^2, BMI, BMI^2, and gender. Figure 22 shows the quantile-quantile plot of the –log10(P-values) from this analysis plotted against the expected distribution of P-values under the null hypothesis. Median adjusted HDL-C level in the Marshfield PMRP cohort was very strongly associated with SNPs in cholesterol ester transfer protein (*CETP*) on chromosome 16, and strongly associated with hepatic lipase (*LIPC*) on chromosome 15. Both *CETP* and *LIPC* have previously been implicated in cholesterol homeostasis in other genome-wide association studies (Aulchenko et al., 2009; Kathiresan et al., 2008; Kathiresan et al., 2009; Sabatti et al., 2009; Willer et al., 2008b). The SNP with the strongest evidence for association with median HDL-C was rs3764261 (p=1.22e-25), 2.5kb upstream of the *CETP* transcription start site. This SNP alone accounted for approximately 3% of the variance in median adjusted HDL-C level. This SNP is in LD (r^2>0.8 in HapMap CEU) with rs247616 (not genotyped here), an eQTL that contributes to cis-regulation of *CETP* mRNA levels in HapMap lymphoblastoid cell lines (Veyrieras et al., 2008). There were several other strongly associated variants upstream of the *CETP* transcription start site, and we observed association between HDL-C and a well-characterized nonsynonymous coding SNP, rs5882 (p=4.1e-7), known to encode a valine to isoleucine change at amino acid position 422 in the *CETP* gene product. A previous study in Ashkenazi Jews with exceptional longevity has revealed that individuals with the V/V genotype at this site demonstrate increased lipoprotein sizes and lower serum *CETP* concentration, both of which are heritable and promote successful aging (Barzilai et al., 2003).

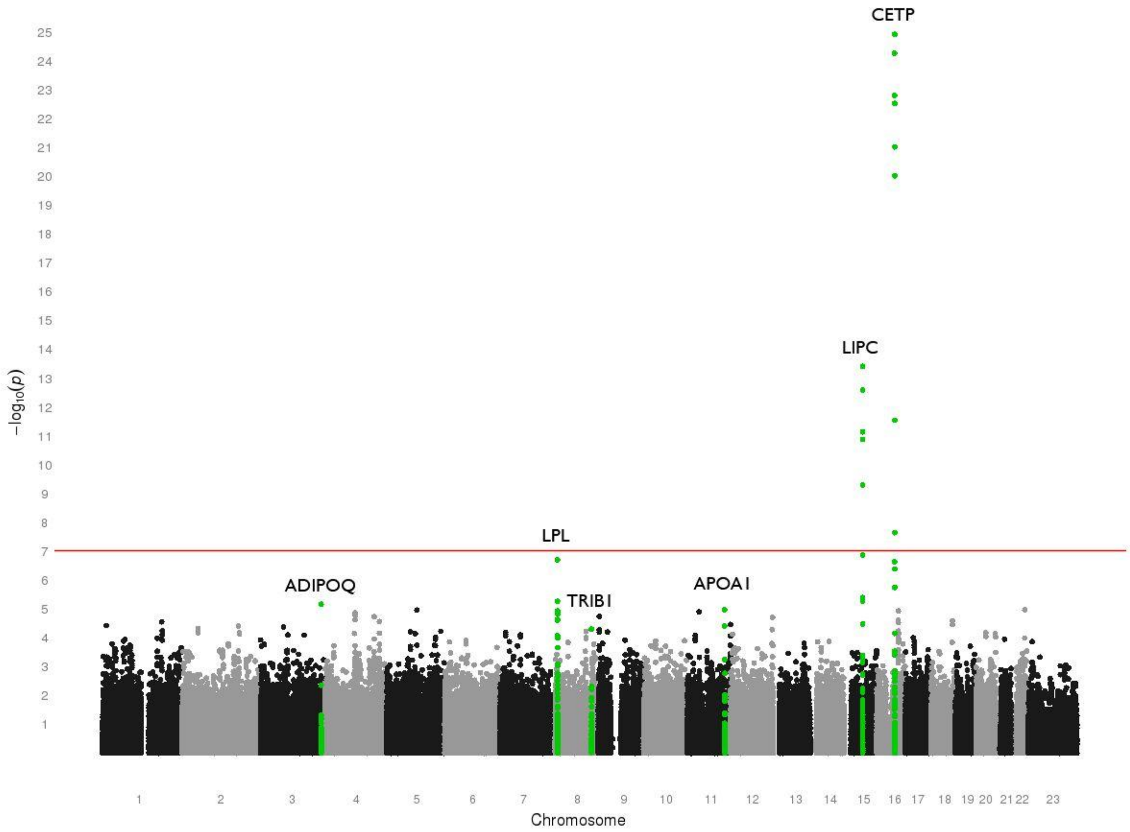


Figure 21. Summary of genome-wide association results. This manhattan plot shows the –log10(P-values) from linear mixed effects regression model using the natural log tranformed median HDL-C level in the Marshfield PMRP sample, adjusted for smoking, age, age², BMI, BMI², and gender. The red line indicates a Bonferroni-corrected significance threshold. Genes described in the results and the discussion sections are highlighted in green: *ADIPOQ* (chromosome 3), *LPL* (chromosome 8), *TRIB1* (chromosome 8), *APOA1/C3/A4/A5* (chromosome 11), *LIPC* (chromosome 15), and *CETP* (chromosome 16).

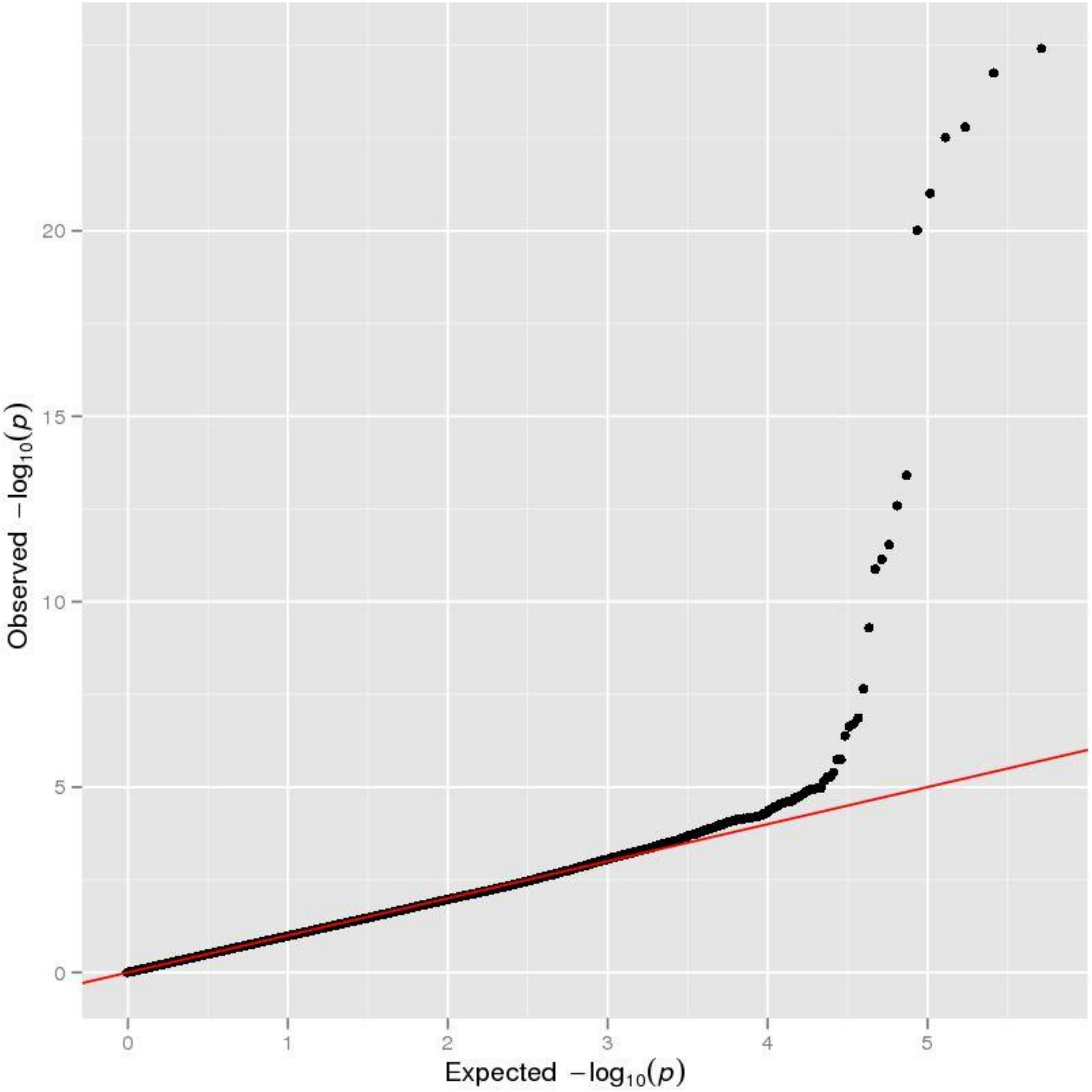


Figure 22. Q-Q plot of P-values. Quantile-quantile plot of the –log10(P-values) from the median adjusted HDL-C analysis in the Marshfield PMRP cohort plotted against the expected null distribution.

The other signal observed at a level of genome-wide significance was in hepatic lipase (*LIPC*). Variants in this gene have also been associated with HDL-C level in several previous studies (Aulchenko et al., 2009; Kathiresan et al., 2008; Kathiresan et al., 2009; Sabatti et al., 2009; Willer et al., 2008b). The SNP in the *LIPC* locus with the strongest association in our study cohort was rs11855284 (p=3.92e-14), residing 13.5kb upstream of the *LIPC* transcription start site on chromosome 15. The most significantly associated SNP within the *LIPC* genic region was intronic SNP rs261336 (p=5.25e-6). These two SNPs explained 1.6% and 0.5% of the variance in HDL-C cholesterol, respectively.

Our most significant signal that did not meet genome-wide significance was rs12678919 (p=1.99e-7), a SNP in an intergenic region 19kb downstream of lipoprotein lipase (*LPL*) on chromosome 8. This SNP explained 0.7% of the variance in HDL cholesterol. It is noteworthy that a different variant at the *LPL* locus (rs253) was identified in subsequent gene-gene interaction analyses here. Details are discussed further below.

Single Locus GWAS Analysis: Modeled HDL-C

Our data were derived from an electronic medical record. Therefore, we also tested each SNP for genome-wide association with modeled HDL-C as the outcome, a trait which censors by onset of relevant co-morbidities or usage of lipid-modifying drugs and adjusts for population trends in age and BMI (Wilke et al., 2010). Because this more rigorous phenotype required data for several covariates, complete data was only available on 2,528 samples in the Marshfield PMRP cohort. Using 1,156 fewer samples we still reproduced the genome-wide significant association to *CETP* (rs3764261, p=2.63e-13), and a highly significant association to both *LPL* (rs1441762, p=1.53e-6) and *LIPC* (rs11856159, p=1.59e-6). Interestingly, even though using modeled HDL-C as the phenotype resulted in a decreased sample size, the association signal was stronger for certain regions of the genome than when using median adjusted HDL-C. For example, rs964184 near *APOA1-APOC3-APOA4-APOA5* showed a stronger association (p=8.37e-

7) when modeled HDL-C was used instead of median HDL-C (p=1.06e-5), as shown in Figure 23, top panels. This emphasizes the need for evaulating data with and without strategies that censor the longitudinal strings of lipid data based on medication (e.g., niacin) and relevant co-morbidity (e.g., diabetes). Both are known to alter *APOA1* expression. Our ability to resolve this association strengthened when these variables were considered.

Conversely, some associations were attenuated in the modeled data. For example, rs7627293 on chromosome 3, upstream of *ST6GAL1* (a sialyltransferase) and the adiponectin gene (*ADIPOQ*), was associated with median adjusted HDL-C (p=6.89e-6) but not with modeled HDL-C (p=0.11). This is shown in the bottom panels of Figure 23. Adiponectin is released by fat cells, and modulates insulin sensitivity in a variety of tissues. As weight increases, adiponectin levels decrease. Subjects then become insulin resistant and develop secondary defects in lipid homeostasis. Because our modeled HDL-C trait censors lipid observations at the first diagnoses of diabetes, and adjusts for BMI, the adiponectin effect on HDL-C would be obscured. In fact, it has been shown in a study characterizing the relationship between adiponectin levels and coronary artery disease (CAD) that - after adjusting for HDL-C - the effect of adiponectin on CAD is no longer significant (Toth, 2005).

In general, despite using lower numbers, we observed improvement in the strength of association by more than two orders of magnitude, for many variants included in this genome-wide SNP scan, when the HDL-C trait was modeled (e.g., *CENTG2*, *COL23A1*, *CACNA2D1*, *CYB5B*, *FHOD3*, *GBX2*, *TRIB1*, *TLE4*, *TMEM135*, *UBE3A*). As shown in Figure 24, this effect was most pronounced for SNPs in or near *TRIB1*, the Tribbles homolog 1 gene (rs2385114, an intronic SNP, p=8.96e-5; rs4871603, 30kb downstream, p=2.61e-6). *TRIB1* was also associated at the genome-wide level with triglyceride concentration in our data (most significant SNP, rs6982502, p=3.7e-9, data not shown). Because we observed a very strong logarithmic correlation between median triglyceride level and median HDL-C cholesterol level (Figure 25), we further assessed the *TRIB1*-HDL-C relationship using multiple linear regression, adjusting for triglyceride

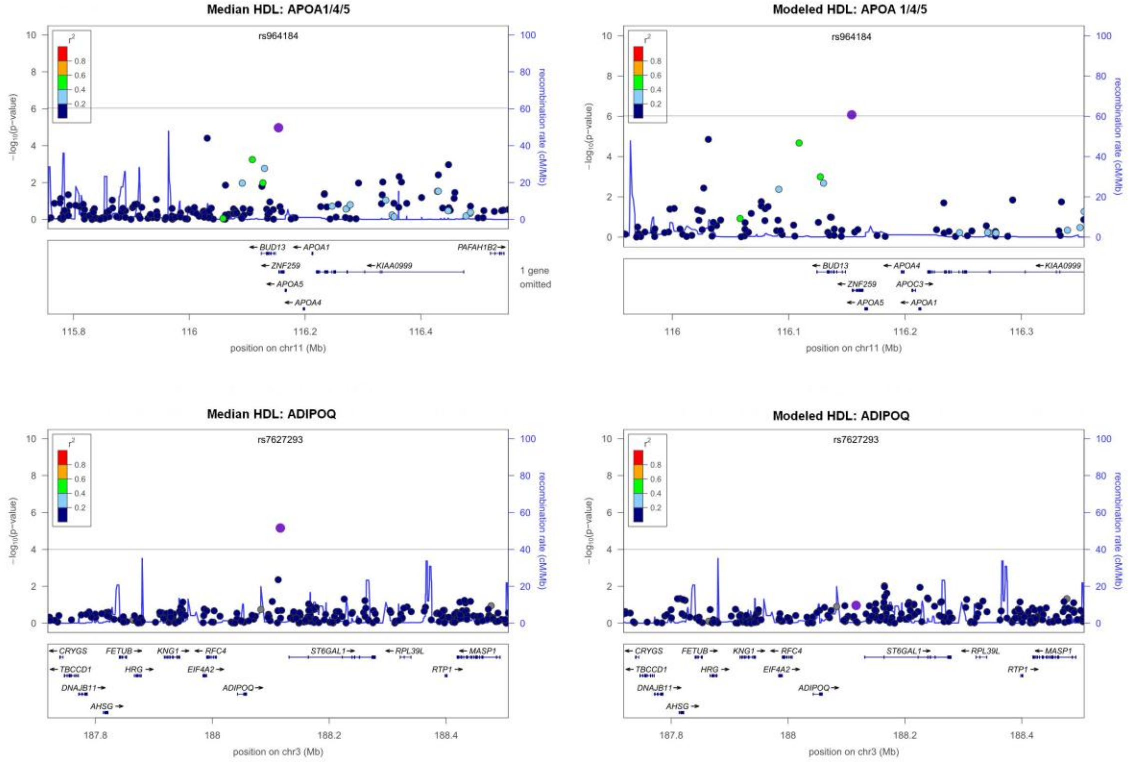


Figure 23. Association results from the *APOA1/C3/A4/A5* and *ADIPOQ* loci. This shows the association results for two regions, *APOA1-APOC3-APOA4-APOA5* (top row) and *ADIPOQ* (bottom row) for median adjusted HDL-C (left columns) and modeled HDL-C (right columns) in the Marshfield PMRP cohort. Color scale displays the degree of linkage disequilibrium (r²) between markers. Blue line shows recombination rate from HapMap CEU. Gene location is shown along the horizontal axis of each panel. Here, both the *APOA* gene cluster region is more strongly associated with modeled HDL-C. A SNP near *APIPOQ* is associated (p<1e-4) with median HDL-C, but this association disappears when taking other factors such as diabetes and medication into account with the model.

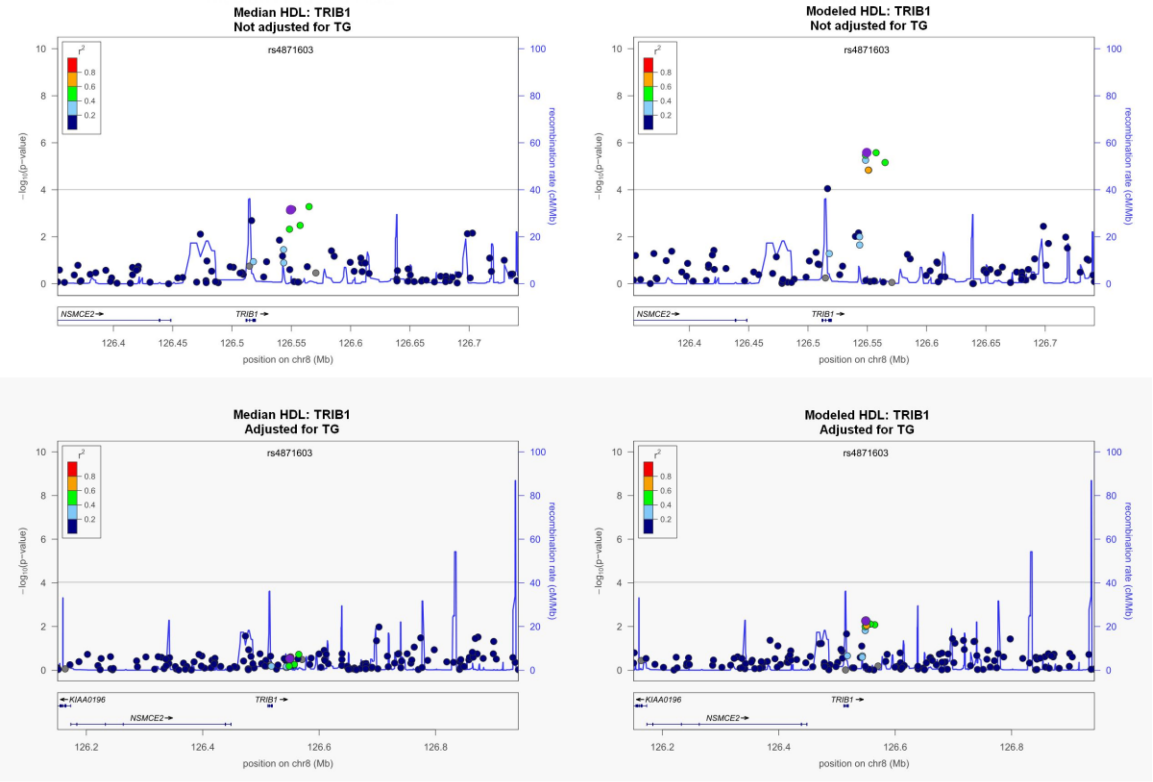


Figure 24. Association results for the *TRIB1* region. This shows the association results for the Tribbles 1 Homolog (*TRIB1*) region, without controlling for triglyceride concentration (top row) and after controlling for triglycerides (bottom row), for median adjusted HDL-C (left columns) and modeled HDL-C (right columns) in the Marshfield PMRP cohort. Color scale displays the degree of linkage disequilibrium (r^2) between markers. Blue line shows recombination rate from HapMap CEU. Gene location is shown along the horizontal axis of each panel. This plot shows that while the statistical significance of the effect of *TRIB1* on HDL-C levels is less compelling when adjusting for triglyceride concentration, the association is much stronger when using the modeled HDL-C phenotype (even though this phenotype has far fewer samples than the median adjusted HDL-C phenotype). SNP rs4871603 in *TRIB1* was associated with median HDL-C at p=7.06e-4 and with modeled HDL-C at p=2.61e-6 without adjusting for triglyceride concentration. After adjusting for triglycerides, the p-values for median and modeled HDL-C become less significant (p=0.296 and p=0.0056, respectively).

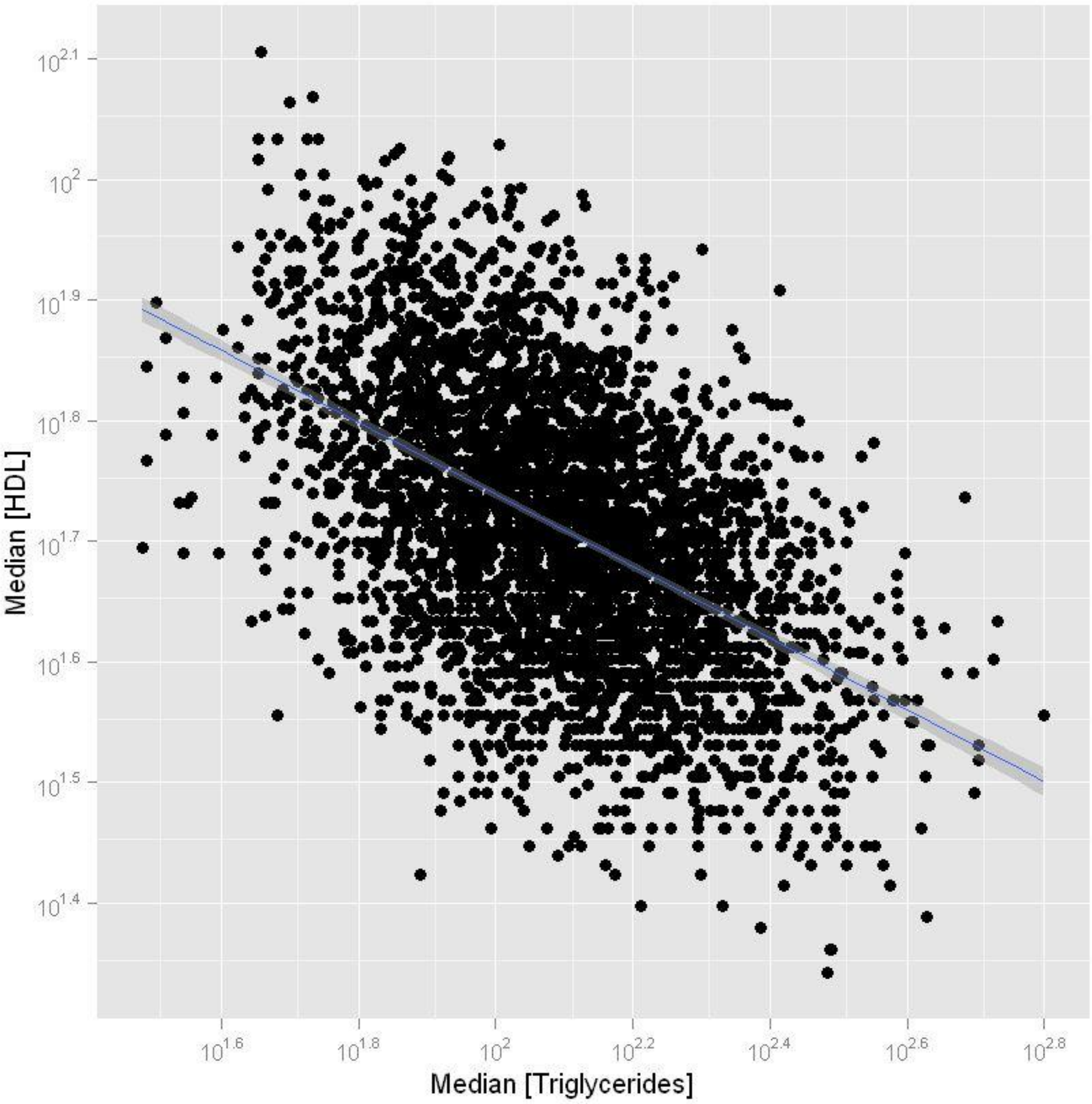


Figure 25. HDL-C vs. TG concentration. Median HDL-C and median triglyceride concentrations are highly logarithmically correlated (r^2=.258). Trend line ± 95% confidence interval is shown.

concentration in addition to the other clinical variables (e.g., age, gender, BMI and smoking status). As shown in the bottom panels of Figure 24, these associations are attenuated after adjustment for triglyceride levels.

Gene-Gene Interaction Analysis and Replication

We have demonstrated that biobanks linked to EMR data can provide a unique resource for robust replication of previous GWAS findings, and they have the capability of uncovering novel genetic associations. However, as in other studies, even our most significant findings individually explain only a small proportion of the variance in HDL-C level. We next leverage both the PMRP cohort and a similarly phenotyped cohort from the Vanderbilt BioVU EMR-linked biobank to investigate gene-gene interactions and HDL-C. Table 7 summarizes the clinical characteristics of the Vanderbilt BioVU population. There were 659 male (35%) and 1,199 female study subjects (total 1,858) with clinical laboratory records containing at least two fasting HDL measurements that could be used in the median HDL-C analysis.

The number of possible SNP-SNP interactions among 522,204 SNPs is over 1.36×10^{11}. In addition to being extremely computationally intensive, exhaustive evaluation of all possible pairwise (SNP-SNP) interactions among GWAS data comes with an extraordinary loss of power due to the extremely large number of statistical tests being performed. This mandates prioritization of which interactions to test based on some intrinsic aspect of the data or other extrinsic domain knowledge. One data-driven approach is to select SNPs based on the strength and statistical significance of their independent main effects, evaluating interactions only between SNPs that meet a certain effect size or significance threshold (Kooperberg et al., 2008). This strategy makes the simplifying but unnecessary assumption that gene-gene interactions affecting the phenotype can only occur between SNPs that independently have a detectable effect on the trait. Instead, we used the Biofilter (described in the methods section) to prioritize a small subset of SNP-SNP interactions to test for association with HDL-C level using existing extrinsic

Table 7. Descriptive statistics of clinical variables in the Vanderbilt BioVU cohort.

	N	Median	Mean	SD	Min	Max
Age (years)	2,576	57	56.56	15.85	19	90
Weight (kilograms)	2,522	85.16	88.05	24.62	36.29	251.74
Height (centimeters)	2,294	167.64	168.5	10.01	129.54	203.2
BMI (kg/m^2)	2,292	29.44	30.98	7.94	14.63	72.5
Median HDL-C (mg/dl)	2,576	50.0	53.20	17.320	9.0	146.0

biological knowledge. It is of note that this method does not require that any SNP have an independent statistically significant main effect.

We ran a gene-gene interaction analysis using multiple linear regression allowing for a multiplicative interaction term between the two additive-encoded SNPs. Requiring that SNP-SNP models be supported by at least four sources of extrinsic domain knowledge (see *Statistical Analysis* in Methods section), we tested 22,769 SNP-SNP interactions in the Marshfield PMRP cohort. Two test statistics were generated for each model tested: a t-test on the multiplicative interaction term (P_{ixn}), and an F-test on the overall model (P_{mod}). A significant P_{ixn} indicates a significant non-zero multiplicative gene-gene interaction between the two SNPs while a significant P_{mod} indicates a significant overall model (i.e. $R^2>0$). We required both statistics to be significant in both datasets, further constraining our results to well-fitting models with strong evidence of non-additive gene-gene interaction.

Using this approach, we found 11 models with a significant interaction term and ANOVA p-values ($P_{ixn}<0.01$ and $P_{mod}<0.05$), indicative of non-additive gene-gene interaction contributing to median HDL-C level. These 11 SNP-SNP interaction models were then tested in the BioVU cohort, adjusting for two significant principal components to avoid any potential confounding by population stratification. Of the 11 models that were significant in the initial screen, six replicated in the BioVU replication cohort. Statistical significance thresholds for the replication cohort were slightly more liberal ($P_{ixn}<0.05$, $P_{mod}<0.1$) to avoid excessive type II errors due to the smaller sample size.

The results highlighting these 11 significant gene-gene interaction models are summarized in Table 8 (models indicated by one or more stars). The six models that show evidence for replication in the BioVU cohort are indicated by two or more stars. These six models are representative of four distinct gene-gene interactions: *GALNT1-GALNT2*, *GALNT2-GALNT3* (members of the GalNAc-transferases family), *LPL-ABCA1* (lipoprotein lipase and ATP Binding Casette A1), and *RPA2-RPA3* (Replication Protein A 2/3).

Table 8. Gene-gene interaction models. This table shows 11 significant gene-gene interaction models discovered in the Marshfield PMRP cohort (M $P_{ixn}<0.01$ and M $P_{mod}<0.05$, indicated by one to three stars). Six of these models show evidence for replication in the Vanderbilt BioVU cohort (M $P_{ixn}<0.05$, M $P_{mod}<0.1$, indicated by two to three stars). In two of these replicating models, all three pairs of coefficients were in the same direction in both datasets (indicated by three stars). The table shows the two SNPs and their corresponding genes involved in the gene-gene interaction. All SNPs here were intronic SNPs. The coefficients for the main effects (β_1 and β_2) and the interaction term (β_3) are shown for both the Marshfield cohort (prefixed by “M”), and the Vanderbilt cohort (prefixed by “V”). Also shown are the interaction term p-values (P_{ixn}), ANOVA model fit p-values (P_{mod}), and overall R^2 statistics for both the Marshfield and Vanderbilt cohorts (prefixed by “M” and “V” respectively).

REP	SNP 1	Gene 1	SNP 2	Gene 2	M β_1	M β_2	M β_3	M P_{ixn}	M P_{mod}	M R^2	V β_1	V β_2	V β_3	V P_{ixn}	V P_{mod}	V R^2
*	rs3927911	*BCL2*	rs4645900	*BAX*	0.213	3.901	-3.890	0.004	0.018	0.003	0.805	5.397	-5.808	0.042	0.154	0.003
*	rs2271709	*C7*	rs6699859	*C8A*	1.203	1.068	-1.776	0.005	0.028	0.002	-1.173	-1.176	2.433	0.020	0.138	0.003
*	rs910497	*GALNT2*	rs4621175	*GALNT3*	-0.727	-1.250	2.347	0.003	0.013	0.003	-0.890	-1.976	2.148	0.024	0.129	0.003
*	rs4621175	*GALNT3*	rs4846930	*GALNT2*	-1.213	-0.726	2.291	0.004	0.014	0.003	-1.750	-0.955	2.261	0.017	0.100	0.003
*	rs4621175	*GALNT3*	rs10864732	*GALNT2*	-1.179	-0.726	2.243	0.004	0.017	0.003	-1.641	-0.985	2.245	0.019	0.106	0.003
**	rs886724	*RPA3*	rs7536088	*RPA2*	1.493	1.713	-1.818	0.000	0.002	0.004	-2.064	-1.266	1.995	0.019	0.099	0.003
**	rs886724	*RPA3*	rs17257252	*RPA2*	0.890	1.182	-1.703	0.003	0.029	0.002	-2.035	-1.938	2.795	0.007	0.046	0.004
**	rs901675	*GALNT2*	rs4621175	*GALNT3*	1.216	2.109	-2.521	0.004	0.004	0.004	-2.114	-1.512	2.535	0.037	0.077	0.004
**	rs1471915	*GALNT2*	rs12963790	*GALNT1*	-0.410	-0.447	2.778	0.004	0.020	0.003	-2.114	0.098	-3.487	0.037	0.002	0.008
***	rs253	*LPL*	rs2515614	*ABCA1*	-0.340	-1.098	1.441	0.006	0.011	0.003	-0.618	-2.797	2.790	0.001	0.006	0.007
***	rs253	*LPL*	rs2472509	*ABCA1*	-0.338	-1.113	1.438	0.006	0.011	0.003	-0.399	-2.797	2.790	0.001	0.006	0.007

We then refined our results highlighting models that were significant in the Marshfield cohort, replicate in the BioVU cohort, and where all coefficients were in the same direction – that is, if the coefficient for either SNP or the interaction term is positive in the Marshfield cohort, the corresponding coefficient must also be positive in the BioVU dataset. This stringent criterion further reduced our replicating models to two similar interaction models involving *LPL* and *ABCA1* (models indicated by three stars in Table 8): rs253×rs2515614 and rs253×rs2472509. These two models were statistically redundant - the *ABCA1* SNPs (rs2515614 and rs2472509) were in extremely high linkage disequilibrium (r^2=1 in Marshfield, r^2=0.99 in BioVU), resulting in nearly identical coefficients and test statistics. One characteristic of this *LPL-ABCA1* interaction warrants special emphasis. In this model the direction of the main effects (β_1 and β_2) were all in the same direction, while the interaction effects were in the opposite direction. That is, in both datasets, inheriting a minor allele at either locus (but not both) results in a dosage-dependent decrease in HDL-C level, while inheriting a minor allele at both variants results in change that is significantly higher than the expected decrease caused by the additive effects of both variants alone.

Conclusions

We performed a genome-wide association analysis of HDL cholesterol level in two large clinical practice-based biobanks. We observed a number of previously reported associations between HDL-C level and genes impacting lipoprotein homeostasis (e.g., *CETP*, *LIPC*, *LPL*). Furthermore, our approach using an electronic phenotyping algorithm to censor based on clinical factors known to influence HDL-C cholesterol levels, allowed us to identify genes that are *less* strongly associated in the context of clinical covariates (e.g., *ADIPOQ*), and genes that are *more* strongly associated in the context of clinical covariates (e.g.,*TRIB1*). Finally, we investigated gene-gene interaction using a novel bioinformatics approach that restricts the number of pairwise tests based on existing biological knowledge.

The primary phenotype utilized in this study was median HDL cholesterol level, derived from longitudinal clinical data. This trait was determined from laboratory data obtained during routine clinical care, and available within each subject's individual electronic medical record. In the single-locus analysis, the SNP most strongly associated with median HDL-C level was rs3764261 (p=1.22e-25), 2.5kb upstream of the *CETP* transcription start site on chromosome 16. This SNP alone accounted for approximately 3% of the variance in median adjusted HDL-C level. This same variant had also been reported in another genome-wide association study in the Northern Finland Birth Cohort (p=6.97e-29) (Sabatti et al., 2009), and a meta analysis of three genome-wide association studies initially comprising 8,656 individuals and ~2,261,000 imputed and/or genotyped SNPs (p=2.8e-19) (Willer et al., 2008b), followed by validation in six European cohorts totaling 11,569 individuals (p=6.4e-43) (Willer et al., 2008b). Other studies have found different SNPs in the region upstream of *CETP* to be even more highly associated with HDL cholesterol levels: rs1800775 (p=1e-73) in (Kathiresan et al., 2008), rs173539 (p=4e-75) in (Kathiresan et al., 2009) where samples were combined from the two previously cited studies, and rs1532624 (p=9.4e-94) in (Aulchenko et al., 2009). The SNP with the strongest association in our dataset (rs3764261) is in LD with an eQTL SNP which affects mRNA levels of *CETP* in HapMap lymphoblastoid cell lines (Veyrieras et al., 2008). The *CETP* gene product has known biological relevance, redistributing cholesterol esters between HDL-C particles and the larger, more atherogenic lipoproteins.

Free fatty acids and TG are liberated from HDL-C particles through the activity of three well-characterized lipolytic enzymes (*LIPC*, *LIPG*, and *LPL*). In our dataset, a SNP in the first gene reaching genome-wide significance for association with HDL-C level was rs11855284 (p=3.90e-14), 13.5kb upstream of the hepatic lipase (*LIPC*) transcription start site on chromosome 15. Further, our most significant signal that did not meet genome-wide significance was rs12678919 (p=1.99e-7), a SNP in an intergenic region 19kb downstream of lipoprotein lipase (*LPL*) on chromosome 8. Thus, we observed association with *LPL* and *LIPC* but not *LIPG*. These observations are consistent with existing biological knowledge. The enzymatic activity of LPL favors lipolysis of TG (i.e., phospholipase activity is relatively minor). TG rich HDL-C particles are less stable, and are quickly shuttled back to the liver for elimination, as SRB1 mediated removal is a function of TG enrichment in HDL-C particles. Conversely, *LIPG* has relatively little TG-lipase activity, (i.e., primarily a phospholipase), and any common variants that affect expression or function of *LIPG* would have small effect on HDL-C stability, requiring a very large sample size to detect an association (~9,800 samples to achieve 80% power to detect an effect of similar magnitude as has been previously observed for *LIPG* (Willer et al., 2008b) at p=1e-7).

Nascent HDL-C particles contain two copies of Apo-A1 and very little lipid (less than 10%) (Lewis & Rader, 2005). In our data, HDL-C level showed suggestive association with the *APOA1/C3/A4/A5* locus (rs964184, p = 1.06e-5), adjusted for age, gender, BMI, smoking status, and family relatedness. As shown in Figure 23, the HDL-C association with *APOA1/C3/A4/A5* locus was strengthened (p=8.37e-7) when we used our previously published approach to modeling HDL-C within an electronic record, which censors data based on relevant co-morbidities (e.g., diabetes mellitus) and lipid modifying medications (e.g., niacin). Medications such as niacin are highly efficacious at increasing HDL-C level. Niacin increases HDL-C level approximately 30% (Joy & Hegele, 2008). While fibric acid derivatives also increase HDL-C, they are far more efficacious at reducing TG (Joy et al., 2008). HMG CoA reductase inhibitors (statins) have modest but reproducible beneficial effects on both HDL-C and TG levels (Yee & Fong, 1998; 1984; Shepherd et al., 1995; 1994). When our data were censored based on first exposure to a lipid modifying medication, or first co-morbidity known to alter lipid homeostasis (Wilke et al., 2010), the association between HDL-C and *APOA* gene cluster variants increased by two orders of magnitude. We therefore scanned our genome-wide SNP data to identify other variants that were more strongly associated using this approach. SNPs in or near *TRIB1* were also strongly associated with modeled HDL-C level. In our data, HDL-C and TG levels were logarithmically correlated (Figure 25), and the association between rs4871603 in *TRIB1* and HDL-C (p=2.61e-6) was markedly reduced after adjusting for TG (p=0.0056). Further study is warranted to fully characterize the biology underlying this interaction. *TRIB1* may serve as a clinical surrogate for the triglyceride effect on HDL-C concentration. *TRIB1* encodes a G-protein-coupled-receptor induced protein involved in the function of mitogen-activated protein kinases (Kiss-Toth et al., 2004), and the role of HDL-C in reverse cholesterol transport may be modulated through such an interaction (Ghosh, Ghosh, Gehr, & Sica, 2004; Miller et al., 2007; Grewal et al., 2003). Because *TRIB1* has previously been associated with coronary artery disease (Aulchenko et al., 2009; Kathiresan et al., 2008; Kathiresan et al., 2009; Willer et al., 2008b), the *TRIB1*-HDL-C relationship observed within our data may have a profound impact on public health.

Our most significant findings, however, still explain less than 3% of the variance in HDL-C level, a trait that is up to 70% heritable. We therefore examined a small subset of all the possible gene-gene interactions in our GWAS data. We found 11 interactions that were nominally significant in our discovery cohort (the Marshfield PMRP biobank). After evaluating these models in a validation sample (Vanderbilt BioVU cohort), we found that six of the 11 models replicated. We required both the p-value on the interaction term and the p-value on the ANOVA F-test for the regression model to be significant in both datasets, further constraining our results to models with evidence of nonlinear gene-gene interaction in a well-fitting model. These models individually explained 0.2-0.8% of the variation in HDL-C in the Marshfield and BioVU cohorts (see Table 8). Only one set of observed gene-gene interactions showed evidence of replication in both datasets with consistent directionality in all three coefficients: *LPL* and *ABCA1*. Because we required evidence of replication with consistent direction of effect in a second dataset, we did not require a stringent multiple testing correction. While it would be possible to randomize the outcome data and permute this entire procedure, this method of permutation testing would be computationally expensive, and thus was not performed here. *LPL* mediates the release of free fatty acids and TG from HDL-C particles, while *ABCA1* moves free cholesterol into HDL-C particles as they undergo intravascular remodeling. Thus, it is not surprising that we observed a statistically meaningful interaction between variants in these two genes. It is however surprising that the direction of the main effect coefficients (β_1 and β_2) were in the same direction while the interaction effect was in the opposite direction. While still considered a nonlinear epistatic interaction, the structure of the model in Table 8 is referred to as a *heterogeneity model* (Cordell, 2002; Neuman et al., 1992) rather than a synergistic multiplicative model. In a separate study of type I diabetes, investigators found that four out of five statistically significant gene-gene interactions were also of this type (Barrett et al., 2009). Genetic heterogeneity is often blamed for the lack of replication in GWAS studies (McClellan et al., 2010; Sillanpaa et al., 2004). Others have recently argued that epistatic genetic heterogeneity should be considered when analyzing genetic data for association to complex human traits (Moore et al., 2010b). Despite the fact that statistical tools, such as random forests, have been available for some time now to accomplish this (Lunetta et al., 2004; Thornton-Wells et al., 2004), analyses of GWAS data accounting for the possibility of epistatic heterogeneity is a task rarely undertaken. Accounting for genetic heterogeneity in genetic studies of complex disease may improve the replicability of findings in genome-wide studies of gene-gene interactions in lipid and other complex human phenotypes.

Here we have used EMR data from genotyped biobanked samples to perform the first knowledge-based gene-gene interaction analysis for HDL-C. Using a second EMR-linked biobank cohort we demonstrated evidence for replication of a gene-gene interaction between *LPL* and *ABCA1*. As demonstrated here and elsewhere (Ritchie et al., 2010; Denny et al., 2010a), biobank-linked EMRs provide an excellent resource for genetic studies of complex traits. By utilizing the EMR to construct a rigorous phenotype and by accounting for gene-gene interaction as presented here, perhaps more variation can be explained and new biology discovered in complex traits like HDL-C level.

Acknowledgements

The work presented in Chapter V was supported in part by NIH grants F31-NS066638, R01-LM010040, U01-HG004608, and U01-HG04603.

CONCLUSION

The “missing heritability” problem (Maher, 2008) discussed in the previous chapter is a familiar phenomenon that is shared among nearly every complex human disease or morphological phenotype. Twin and adoption studies show that many complex traits are heritable, yet GWAS, dispite its many successes (Hindorff et al., 2009), has consistently failed to explain any appreciable amount of this heritability for most complex traits. This does not mean that GWAS has failed; on the contrary, many GWAS studies have produced highly replicable results and have led to the discovery of new biology (Manolio et al., 2009). A notable example was the discovery of the involvement of complement factor H (CFH) in age-related macular degeneration (AMD). Four years after the CFH association was discovered using GWAS and other study designs (Klein et al., 2005; Haines et al., 2005; Edwards et al., 2005), researchers have shown that inhibiting the alternative complement pathway can reduce vascularization of the retina (a symptom of the “wet” form of AMD) in a mouse model of AMD.

GWAS home-runs like the CFH association with AMD – where the first odds ratio reported was 7.4 – are very seldom found for most complex human traits. As discussed throughout this dissertation, the standard single-SNP-at-a-time approach to GWAS ignores the complexity of biological systems. Accounting for this complexity in an analysis may enable the discovery of disease-predisposing genetic variation that would otherwise remain elusive.

After reviewing the state-of-the-art for analysis of GWAS data in Chapter I, I presented in Chapter II a series of improvements on an existing algorithm for finding multi-locus genetic associations to complex traits that uses grammatical evolution to train neural networks (GENN). This chapter drew from several previously published works showing that using a memetic algorithm by incorporating backpropagation with grammatical evolution, in combination with initializing the neural network population in part from domain knowledge can greatly improve the ability of GENN to discover gene-gene interactions impacting complex diseases (Turner et al., 2010b; Turner, Dudek, & Ritchie, 2010c; Turner, Dudek, & Ritchie, 2010a).

"Epistasis" is often used interchangeably with "gene-gene interaction," implying that variation in two or more different genes are involved. Chapter III explored a different kind of epistasis, one that involves variation in regulatory sequences in or around a *single* gene. Here we defined *cis*-epistasis as the non-linear effect on a gene's expression of multiple variants within a 500kb region of potential transcriptional influence. Here we found 706 significant *cis*-epistasis interactions that influence the expression of 79 unique genes. We then investigated the genomic and statistical structure of these interactions, mapped these genes and SNPs to previously reported associations to human disease and morphological phenotypes, and suggested ways in which analysis of future datasets or reanalysis of existing datasets may account for *cis*-epistasis in an interaction analysis.

In Chapter IV I discuss quality control procedures and best practices for GWAS data. These procedures were used prior to the analysis that was presented in Chapter V – a GWAS investigating both single-locus association with HDL-C in the context of environmental variables, and interrogating gene-gene interactions that affect HDL-C using a knowledge based-approach. Using phenotypes abstracted from the electronic medical record linked to DNA from biobanked samples we discovered a potentially new HDL-C gene and replicated several known genetic associations with HDL-C. Using a knowledge-based approach we identified 11 gene-gene interactions in one biobank cohort, with 6 of these showing evidence for replication in a second similarly phenotyped cohort. These results demonstrate the power of linking electronic medical records to GWAS data from biobanked samples and using prior biological knowledge to guide a gene-gene interaction analysis for complex human traits.

The studies that comprise Chapter II (Turner et al., 2010b; Turner et al., 2010c; Turner et al., 2010a) evaluated improvements in the GENN/ATHENA training algorithm using simulated datasets containing 500 SNPs, where the amount of variance, or heritability, explained uniquely by the main effect and interaction variance components was 1% and 5% respectively. During the planning phases of these works, it was thought that while single SNPs individually could only explain a small portion (1-3%) of the variance of complex traits, perhaps gene-gene interactions could individually explain a larger portion of complex trait variance (up to 5%). Pervasive epistasis with large effects exist in animal models (Shao et al., 2008), and could theoretically exist in humans as well (Li & Reich, 2000). However, effects this large are rarely found for common variants, whether alone or in combination with other variants. Our most significant, replicating two-SNP models affecting HDL-C levels presented in Chapter V explain less than half a percent of the variance in HDL level each.

Although we have made tangible improvements to the GENN/ATHENA training algorithm (Chapter II), many more improvements are necessary before such an approach is adequately powered to detect effects in GWAS or sequencing data. In Chapter II I presented a method for sensibly initializing GENN/ATHENA with genes implicated in a shared biological mechanism. After initialization, stochastic computational evolution rearranged solutions, fitted weights, and performed variable selection. A potential algorithmic improvement would utilize our wealth of biological knowledge *throughout* the training process, rather than at *initialization only*. Here, genes having SNPs as inputs to well-performing NN solutions could be mapped back onto known biological pathways, protein families, gene ontologies, or signal transduction pathways, midway through training. NNs with poor performance could be replaced with NNs containing SNPs from other genes in these known biological gene groupings, and stochastic computational evolution would resume. In this two-step procedure, the first iteration of training would find promising regions of the solution space, while the second iteration would hone in on the best solutions in a region. This process could be repeated iteratively, rather than a two-step procedure. From a software engineering perspective, massive performance improvements could potentially be made by rewriting code that would allow ATHENA to run on modern PC graphics cards, called *graphics processing units* (GPUs). A single typical GPUs can have hundreds of processors running in parallel, and GPU floating point performance doubles every year, rather than every 18-24 months for personal computer CPUs (Poli et al., 2008). Another group has recently ported the Multifactor Dimensionality Reduction (MDR) algorithm to a GPU implementation, and found that the GPU implementation running on the graphics card of a single workstation outperformed an optimized C++ implementation running on 150 cores on a Beowulf cluster (Sinnott-Armstrong, Greene, Cancare, & Moore, 2009). The faster-than-Moore's-law (Moore, 1965) performance increases of GPUs, and the parallel nature of GENN makes GPU implementation a very promising avenue of future development.

In addition to the methodological research discussed above, there are several additional analytical approaches that future research should focus on to gain the most information possible from existing data. As discussed in the introduction, the pace of technological innovation in the genotyping laboratory is staggering. The completion of the pilot phase of the 1000 Genomes Project has shed light on patterns of genome-wide genetic variation in multiple populations (Durbin et al., 2010). Although whole-genome sequencing in thousands of samples is still prohibitively expensive, this resource will facilitate imputation, or inference, of alleles at untyped loci. As demonstrated with an eQTL locus in the pilot project, using 1000 Genomes Project reference haplotypes for imputation will allow more fine-grained mapping of GWAS hits than using HapMap II imputation or genotyped markers alone (Durbin et al., 2010). Finally, a recently published perspective argued that performing GWAS with increasingly larger sample sizes will not yield meaningful discoveries, especially if the newly detected loci explain such a small proportion of the heritability of the trait being studied (Goldstein, 2009). A recent lipids meta-analysis to date, using over 100,000 samples, recently provided the first empirical evidence to refute such claims. Teslovich et al. analyzed over 100,000 samples of European descent and identified 95 loci (59 novel) associated with either total cholesterol, LDL, HDL, or triglycerides, several of which demonstrated clear clinical and/or biological significance, albeit with small effect sizes (Teslovich et al., 2010). Efforts are ongoing to impute data from the cohorts presented in Chapter V using 1000 Genomes Project reference haplotypes, and to contribute to the next lipids meta-analysis, which will comprise the largest lipids genetic analysis to date. Finally, future research should aim to unravel the influences of genes and environment by performing cross-cultural studies of disease risk using diverse, non-European cohorts. Because rates of genetic and environmental exposures and disease prevalence differ between populations around the world, it is possible to unravel the relative contributions of environmental and inborn risks (Kolonel, Altshuler, & Henderson, 2004). With the maturation of new genotyping and sequencing technology, acquisition of increasingly large and diverse cohorts, and development of powerful and robust statistical methodology, we are well positioned to begin accounting for nonlinear joint effects between multiple SNPs and environmental exposures which will explain more variation in complex traits, driving personalized medicine and enabling the discovery of new biology.

BIBLIOGRAPHY

Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S) (1994). *Lancet, 344,* 1383-1389.

Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report (2002). *Circulation, 106,* 3143-3421.

Illumina GenCall Data Analysis Software. http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf. (2008a).

NetPath. http://www.netpath.org/. (2008b).

EIGENSOFT. http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm. (2009a).

GenomeSIMLA Software. http://chgr.mc.vanderbilt.edu/ritchielab/subscriptions. (2009d).

Illumina Technical Note: "TOP/BOT" Strand and "A/B" Allele. http://pngu.mgh.harvard.edu/~purcell/plink/. (2009e).

Census 2000: Profile of Demographic Characteristics, Marshfield WI. http://censtats.census.gov/data/WI/1605549675.pdf. (2000).

A Field Guide to Genetic Programming. http://www.gp-field-guide.org.uk/. (2009c).

Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In

Adults (Adult Treatment Panel III) (2001). *Journal of the American Medical Association, 285,* 2486-2497.

The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease (1984). *Journal of the American Medical Association, 251,* 351-364.

PLINK. http://pngu.mgh.harvard.edu/~purcell/plink/. (2009f).

STRUCTURE. http://pritch.bsd.uchicago.edu/structure.html. (2009b).

Abney, M., McPeek, M. S., & Ober, C. (2001). Broad and narrow heritabilities of quantitative traits in a founder population. *Am.J Hum.Genet., 68,* 1302-1307.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Aguilar, P. S., Frohlich, F., Rehman, M., Shales, M., Ulitsky, I., Olivera-Couto, A., Braberg, H., Shamir, R., Walter, P., Mann, M., Ejsing, C. S., Krogan, N. J., & Walther, T. C. (2010). A plasma-membrane E-MAP reveals links of the eisosome with sphingolipid metabolism and endosomal trafficking. *Nat.Struct.Mol.Biol, 17,* 901-908.

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease. *Science, 322,* 881-888.

Arranz, M. J., Munro, J., Birkett, J., Bolonna, A., Mancama, D., Sodhi, M., Lesch, K. P., Meyer, J. F., Sham, P., Collier, D. A., Murray, R. M., & Kerwin, R. W. (2000). Pharmacogenetic prediction of clozapine response. *Lancet, 355,* 1615-1616.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat.Genet., 25,* 25-29.

Ashen, M. D. & Blumenthal, R. S. (2005). Clinical practice. Low HDL cholesterol levels. *N Engl.J.Med., 353,* 1252-1260.

Aulchenko, Y. S., de Koning, D. J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics, 177,* 577-585.

Aulchenko, Y. S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I. M., Pramstaller, P. P., Penninx, B. W., Janssens, A. C., Wilson, J. F., Spector, T., Martin, N. G., Pedersen, N. L., Kyvik, K. O., Kaprio, J., Hofman, A., Freimer, N. B., Jarvelin, M. R., Gyllensten, U., Campbell, H., Rudan, I., Johansson, A., Marroni, F., Hayward, C., Vitart, V., Jonasson, I., Pattaro, C., Wright, A., Hastie, N., Pichler, I., Hicks, A. A., Falchi, M., Willemsen, G., Hottenga, J. J., de Geus, E. J., Montgomery, G. W., Whitfield, J., Magnusson, P., Saharinen, J., Perola, M., Silander, K., Isaacs, A., Sijbrands, E. J., Uitterlinden, A. G., Witteman, J. C., Oostra, B. A., Elliott, P., Ruokonen, A., Sabatti, C., Gieger, C., Meitinger, T., Kronenberg, F., Doring, A., Wichmann, H. E., Smit, J. H., McCarthy, M. I., van Duijn, C. M., & Peltonen, L. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat.Genet., 41,* 47-55.

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23,* 1294-1296.

Baba, T., Azuma, S., Kashiwabara, S., & Toyoda, Y. (1994). Sperm from mice carrying a targeted mutation of the acrosin gene can penetrate the oocyte zona pellucida and effect fertilization. *J.Biol.Chem., 269,* 31845-31849.

Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B. M., Kappos, L., Polman, C. H., Matthews, P. M., Hauser, S. L., Gibson, R. A., Oksenberg, J. R., & Barnes, M. R. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum.Mol.Genet., 18,* 2078-2090.

Barber, M. J., Mangravite, L. M., Hyde, C. L., Chasman, D. I., Smith, J. D., McCarty, C. A., Li, X., Wilke, R. A., Rieder, M. J., Williams, P. T., Ridker, P. M., Chatterjee, A., Rotter, J. I., Nickerson, D. A., Stephens, M., & Krauss, R. M. (2010). Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS.One., 5,* e9763.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., & Rich, S. S. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet, 41,* 703-707.

Barzilai, N., Atzmon, G., Schechter, C., Schaefer, E. J., Cupples, A. L., Lipton, R., Cheng, S., & Shuldiner, A. R. (2003). Unique lipoprotein phenotype and genotype associated with exceptional longevity. *Journal of the American Medical Association, 290,* 2030-2040.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res., 32,* D138-D141.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nat.Genet., 36,* 431-432.

Belew, R. K., McInerney, J., & Schraudolph, N. N. (1990). Evolving Networks: Using the Genetic Algorithm with Connectionist Learning. *Computer Science & Engineering Department Technical Report*.

Bellman, R. (1961). Adaptive control processes. In (Princeton: Princeton University Press.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. London: Oxford University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Boerwinkle, E. & Sing, C. F. (1986). Bias of the contribution of single-locus effects to the variance of a quantitative trait. *Am.J Hum.Genet., 39,* 137-144.

Boj, S. F., Petrov, D., & Ferrer, J. (2010). Epistasis of transcriptomes reveals synergism between transcriptional activators Hnf1alpha and Hnf4alpha. *PLoS.Genet., 6,* e1000970.

Broman, K. W. (1999). Cleaning genotype data. *Genet.Epidemiol., 17 Suppl 1,* S79-S83.

Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput, 14,* 368-379.

Cantu-Paz, E. & Kamath, C. (2008). Evolving neural networks to identify bent-double galaxies in the FIRST survey. *Neural Networks, 16,* 507-517.

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

Cardon, L. R. & Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet, 361,* 598-604.

Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature, 429,* 446-452.

Carlson, C. S., Smith, J. D., Stanaway, I. B., Rieder, M. J., & Nickerson, D. A. (2006). Direct detection of null alleles in SNP genotyping data. *Hum.Mol.Genet., 15,* 1931-1937.

Carroll, M. D., Lacher, D. A., Sorlie, P. D., Cleeman, J. I., Gordon, D. J., Wolz, M., Grundy, S. M., & Johnson, C. L. (2005). Trends in serum lipids and lipoproteins of adults, 1960-2002. *Journal of the American Medical Association, 294,* 1773-1781.

Chandler, R. L., Chandler, K. J., McFarland, K. A., & Mortlock, D. P. (2007). Bmp2 transcription in osteoblast progenitors is regulated by a distant 3' enhancer located 156.3 kilobases from the promoter. *Mol.Cell Biol., 27,* 2934-2951.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni, J. F., Jr., Freimer, N. B., Gerhard, D. S., Gunter, C., Guttmacher, A. E., Guyer, M. S., Harris, E. L., Hoh, J., Hoover, R., Kong, C. A., Merikangas, K. R., Morton, C. C., Palmer, L. J., Phimister, E. G., Rice, J. P., Roberts, J., Rotimi, C., Tucker, M. A., Vogan, K. J., Wacholder, S., Wijsman, E. M., Winn, D. M., & Collins, F. S. (2007a). Replicating genotype-phenotype associations. *Nature, 447,* 655-660.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni, J. F., Jr., Freimer, N. B., Gerhard, D. S., Gunter, C., Guttmacher, A. E., Guyer, M. S., Harris, E. L., Hoh, J., Hoover, R., Kong, C. A., Merikangas, K. R., Morton, C. C., Palmer, L. J., Phimister, E. G., Rice, J. P., Roberts, J., Rotimi, C., Tucker, M. A., Vogan, K. J., Wacholder, S., Wijsman, E. M., Winn, D. M., & Collins, F. S. (2007b). Replicating genotype-phenotype associations. *Nature, 447,* 655-660.

Chasman, D. I., Pare, G., Zee, R. Y., Parker, A. N., Cook, N. R., Buring, J. E., Kwiatkowski, D. J., Rose, L. M., Smith, J. D., Williams, P. T., Rieder, M. J., Rotter, J. I., Nickerson, D. A., Krauss,

R. M., Miletich, J. P., & Ridker, P. M. (2008). Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication. *Circ.Cardiovasc.Genet., 1,* 21-30.

Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., & Friedman, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J.Am.Med.Inform.Assoc., 15,* 87-98.

Chen, Y. M. & O'Connell, R. M. (1997). Active power line conditioner with a neural network control. *IEEE Transactions on Industry Applications, 33,* 1131-1136.

Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. (3rd ed.) Philadelphia: Lawrence Erlbaum.

Colucci-Guyon, E., Portier, M. M., Dunia, I., Paulin, D., Pournin, S., & Babinet, C. (1994). Mice lacking vimentin develop and reproduce without an obvious phenotype. *Cell, 79,* 679-694.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat.Rev.Genet., 10,* 184-194.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum.Mol.Genet., 11,* 2463-2468.

Cordell, H. J. (2009). Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. *Nat.Rev.Genet., 10,* 392-404.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z. Y., Liang, W., Marback, M., Paw, J., San Luis, B. J., Shuteriqi, E., Tong, A. H., van, D. N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A. C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., & Boone, C. (2010). The genetic landscape of a cell. *Science, 327,* 425-431.

Cristea, I. M., Gaskell, S. J., & Whetton, A. D. (2004). Proteomics techniques and their application to hematology. *Blood, 103,* 3624-3634.

Culverhouse, R., Klein, T., & Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genet.Epidemiol., 27,* 141-152.

Culverhouse, R., Suarez, B. K., Lin, J., & Reich, T. (2002). A perspective on epistasis: limits of models displaying no main effect. *Am.J Hum.Genet, 70,* 461-471.

Dadd, T., Weale, M. E., & Lewis, C. M. (2009). A critical evaluation of genomic control methods for genetic association studies. *Genet.Epidemiol., 33,* 290-298.

Daly, A. K., Donaldson, P. T., Bhatnagar, P., Shen, Y., Pe'er, I., Floratos, A., Daly, M. J., Goldstein, D. B., John, S., Nelson, M. R., Graham, J., Park, B. K., Dillon, J. F., Bernal, W., Cordell, H. J., Pirmohamed, M., Aithal, G. P., & Day, C. P. (2009). HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat.Genet., 41,* 816-819.

Dayhoff, J. E. & DeLeo, J. M. (2001). Artificial neural networks: opening the black box. *Cancer, 91,* 1615-1635.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., & Crawford, D. C. (2010a). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics., 26,* 1205-1210.

Denny, J. C., Ritchie, M. D., Crawford, D. C., Schildcourt, J. S., Ramirez, A. H., Pulley, J. M., Basford, M. A., Masys, D. R., Haines, J. L., & Roden, D. M. (2010b). Identification of genomic predictors of atrioventricular conduction: Using electronic medical records as a tool for genome science. *Circulation, In press*.

Devlin, B., Bacanu, S. A., & Roeder, K. (2004). Genomic Control to the extreme. *Nat.Genet., 36,* 1129-1130.

Devlin, B. & Roeder, K. (1999). Genomic control for association studies. *Biometrics, 55,* 997-1004.

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., & Cookson, W. O. (2007b). A genome-wide association study of global gene expression. *Nat.Genet., 39,* 1202-1207.

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., & Cookson, W. O. (2007a). A genome-wide association study of global gene expression. *Nat.Genet., 39,* 1202-1207.

Du, W., Thanos, D., & Maniatis, T. (1993). Mechanisms of transcriptional synergism between distinct virus-inducible enhancer elements. *Cell, 74,* 887-898.

Dumitrescu, L. C., Ritchie, M. D., Brown-Gentry, K., Pulley, J. J., Basford, M., Denny, J., Oksenberg, J. R., Roden, D. M., Haines, J. L., & Crawford, D. C. (2010). Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genetics in Medicine, In Press*.

Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467,* 1061-1073.

Edmondson, A. C. & Rader, D. J. (2008). Genome-wide approaches to finding novel genes for lipid traits: the start of a long road. *Circ.Cardiovasc.Genet., 1,* 3-6.

Edwards, A. O., Ritter, R., III, Abel, K. J., Manning, A., Panhuysen, C., & Farrer, L. A. (2005). Complement factor H polymorphism and age-related macular degeneration. *Science, 308,* 421-424.

Edwards, T. L., Bush, W. S., Turner, S. D., Dudek, S. M., Torstenson, E. S., Schmidt, M., Martin, E., & Ritchie, M. D. (2008). Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA. *Lecture Notes in Computer Science, 4793,* 24-35.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat.Rev.Genet., 11,* 446-450.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., & Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res., 36,* D281-D288.

Fisher, R. A. (1918). The correlations between relatives on the supposition of Mendelian inheritance. *Trans.R.Soc.Edinb., 52,* 399-433.

Frayling, T. M. (2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat.Rev.Genet., 8,* 657-662.

Freitas, A. (2001). Understand the Crucial Role of Attribute Interactions in Data Mining. Artif Intel Rev 16, 177-199.

Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E. O., Nicolae, D. L., Dolan, M. E., & Cox, N. J. (2010). SCAN: SNP and copy number annotation. *Bioinformatics, 26,* 259-262.

Garrod, A. E. (1902). The incidence of alkaptonuria: a study in clinical individuality. *Lancet, 2,* 1616-1620.

Gauderman, W. J. (2002c). Sample size requirements for matched case-control studies of gene-environment interaction. *Stat.Med., 21,* 35-50.

Gauderman, W. J. (2002b). Sample size requirements for association studies of gene-gene interaction. *Am.J Epidemiol., 155,* 478-484.

Gauderman, W. J. (2002a). Sample size requirements for matched case-control studies of gene-environment interaction. *Stat.Med., 21,* 35-50.

Ghosh, S. S., Ghosh, S., Gehr, T. W. B., & Sica, D. A. (2004). HDL mediates reverse cholesterol transport from mesangial cells via map kinase. *American Journal of Hypertension, 17,* S91.

Gibson, G. (1996). Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor.Popul.Biol., 49,* 58-89.

Goldstein, D. B. (2009). Common Genetic Variation and Human Traits. *N Engl.J.Med., 360,* 1696-1698.

Good P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R., & Amos, C. I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am.J Hum.Genet., 82,* 100-112.

Gorry, P., Lufkin, T., Dierich, A., Rochette-Egly, C., Decimo, D., Dolle, P., Mark, M., Durand, B., & Chambon, P. (1994). The cellular retinoic acid binding protein I is dispensable. *Proc.Natl.Acad.Sci.U.S.A, 91,* 9032-9036.

Govindaraju, D. R., Cupples, L. A., Kannel, W. B., O'Donnell, C. J., Atwood, L. D., D'Agostino, R. B., Sr., Fox, C. S., Larson, M., Levy, D., Murabito, J., Vasan, R. S., Splansky, G. L., Wolf, P. A., & Benjamin, E. J. (2008). Genetics of the Framingham Heart Study population. *Adv.Genet., 62,* 33-65.

Grady, B. J., Torstenson, E., Dudek, S. M., Giles, J., Sexton, D., & Ritchie, M. D. (2010). Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *Pac.Symp.Biocomput.,* 315-326.

Greene, C. S., Gilmore, J., Kiralis, J., Andrews, P. C., & Moore, J. H. (2009). Optimal Use of Expert Knowledge in Ant Colony Optimization for the Analysis of Epistasis in Human Disease. *Lect.Notes Comput.Sci., 5483/2009,* 92-103.

Greene, C. S., White, B. C., & Moore, J. H. (2007). An expert knowledge-guided mutation operator for genome-wide genetic analysis using genetic programming. *Lecture Notes in Bioinformatics, 4774,* 30-40.

Greene, C. S., White, B. C., & Moore, J. H. (2009). Sensible initialization using expert knowledge for genome-wide analysis of epistasis using genetic programming (In Press). *Proceedings of the IEEE Congress on Evolutionary Computing,* 676-682.

Gregersen, J. W., Kranc, K. R., Ke, X., Svendsen, P., Madsen, L. S., Thomsen, A. R., Cardon, L. R., Bell, J. I., & Fugger, L. (2006). Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature, 443,* 574-577.

Grewal, T., de, D., I, Kirchhoff, M. F., Tebar, F., Heeren, J., Rinninger, F., & Enrich, C. (2003). High density lipoprotein-induced signaling of the MAPK pathway involves scavenger receptor type BI-mediated activation of Ras. *J Biol Chem, 278,* 16478-16481.

Gruda, M. C., van, A. J., Rizzo, C. A., Durham, S. K., Lira, S., & Bravo, R. (1996). Expression of FosB during mouse development: normal development of FosB knockout mice. *Oncogene, 12,* 2177-2185.

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., Schnetz-Boutaud, N., Agarwal, A., Postel, E. A., & Pericak-Vance, M. A. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science, 308,* 419-421.

Haines, J. L. & Pericak-Vance, M. A. (1998). *Approaches to gene mapping in complex human diseases*. New York: John Wiley and Sons.

Hamon, S. C., Stengard, J. H., Clark, A. G., Salomaa, V., Boerwinkle, E., & Sing, C. F. (2004). Evidence for non-additive influence of single nucleotide polymorphisms within the apolipoprotein E gene. *Ann.Hum.Genet., 68,* 521-535.

Hardy, J. & Singleton, A. (2009b). Genomewide Association Studies and Human Disease. *N Engl.J.Med., 360,* 1759-1768.

Hardy, J. & Singleton, A. (2009a). Genomewide association studies and human disease. *N.Engl.J.Med., 360,* 1759-1768.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

He, X., Qian, W., Wang, Z., Li, Y., & Zhang, J. (2010). Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat.Genet., 42,* 272-276.

Heid, I. M., Boes, E., Muller, M., Kollerits, B., Lamina, C., Coassin, S., Gieger, C., Doring, A., Klopp, N., Frikke-Schmidt, R., Tybjaerg-Hansen, A., Brandstatter, A., Luchner, A., Meitinger, T., Wichmann, H. E., & Kronenberg, F. (2008). Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. *Circ.Cardiovasc.Genet., 1,* 10-20.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc.Natl.Acad.Sci.U.S.A, 106,* 9362-9367.

Hirschhorn, J. N. (2009). Genomewide Association Studies -- Illuminating Biologic Pathways. *N Engl.J.Med., 360,* 1699-1701.

Holzinger, E. R., Buchanan, C., Turner, S. D., Dudek, S. M., Torstenson, E. S., & Ritchie, M. D. (2010). Optimizing Neural Networks for Detecting Gene-Gene Interactions in the Presence of Small Main Effects. *Genetic and Evolutionary Computation Conference - GECCO 2010, ACM Press., 2010,* 203-210.

Huang, C. Y., Studebaker, J., Yuryev, A., Huang, J., Scott, K. E., Kuebler, J., Varde, S., Alfisi, S., Gelfand, C. A., Pohl, M., & Boyce-Jacino, M. T. (2004). Auto-validation of fluorescent primer extension genotyping assay using signal clustering and neural networks. *BMC Bioinformatics, 5,* 36.

Huang, D. S., Liu, K. H., & Xu, C. G. (2008). A Genetic Programming Based Approach to the Classification of Multiclass Microarray Datasets. *Bioinformatics,* btn644.

Hung, S. L. & Adeli, H. (1994). A parallel genetic/neural network learning algorithm for MIMD shared memory machines. *IEEE Trans.Neural Netw., 5,* 900-909.

Iles, M. M. (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS.Genet., 4,* e33.

International hapmap consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449,* 851-861.

International hapmap consortium (2003). The International HapMap Project. *Nature, 426,* 789-796.

Ioannidis, J. P., Thomas, G., & Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nat.Rev.Genet., 10,* 318-329.

Itohara, S., Mombaerts, P., Lafaille, J., Iacomini, J., Nelson, A., Clarke, A. R., Hooper, M. L., Farr, A., & Tonegawa, S. (1993). T cell receptor delta gene mutant mice: independent generation of alpha beta T cells and programmed rearrangements of gamma delta TCR genes. *Cell, 72,* 337-348.

Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R. M., Myers, R. M., Ridker, P. M., Chasman, D. I., Mefford, H., Ying, P., Nickerson, D. A., & Eichler, E. E. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *Am.J.Hum.Genet., 84,* 148-161.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de, L. J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., & Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature, 451,* 998-1003.

Johansen, C. T., Wang, J., Lanktree, M. B., Cao, H., McIntyre, A. D., Ban, M. R., Martins, R. A., Kennedy, B. A., Hassell, R. G., Visser, M. E., Schwartz, S. M., Voight, B. F., Elosua, R., Salomaa, V., O'Donnell, C. J., linga-Thie, G. M., Anand, S. S., Yusuf, S., Huff, M. W., Kathiresan, S., & Hegele, R. A. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat.Genet.*.

Johnson, A. D. & O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC Med.Genet., 10,* 6.

Johnson, L. W. & Weinstock, R. S. (2006). The metabolic syndrome: concepts and controversy. *Mayo Clin.Proc., 81,* 1615-1620.

Joy, T. & Hegele, R. A. (2008). Is raising HDL a futile strategy for atheroprotection? *Nat.Rev.Drug Discov., 7,* 143-155.

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res., 28,* 27-30.

Kathiresan, S., Manning, A. K., Demissie, S., D'Agostino, R. B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N. P., Melander, O., Orho-Melander, M., Arnett, D. K., Peloso, G. M., Ordovas, J. M., & Cupples, L. A. (2007). A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC.Med.Genet., 8 Suppl 1,* S17.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., Cooper, G. M., Roos, C., Voight, B. F., Havulinna, A. S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E. S., Ordovas, J. M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M. R., Newton-

Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D. M., & Orho-Melander, M. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat.Genet., 40,* 189-197.

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B. F., Bonnycastle, L. L., Jackson, A. U., Crawford, G., Surti, A., Guiducci, C., Burtt, N. P., Parish, S., Clarke, R., Zelenika, D., Kubalanza, K. A., Morken, M. A., Scott, L. J., Stringham, H. M., Galan, P., Swift, A. J., Kuusisto, J., Bergman, R. N., Sundvall, J., Laakso, M., Ferrucci, L., Scheet, P., Sanna, S., Uda, M., Yang, Q., Lunetta, K. L., Dupuis, J., de Bakker, P. I., O'Donnell, C. J., Chambers, J. C., Kooner, J. S., Hercberg, S., Meneton, P., Lakatta, E. G., Scuteri, A., Schlessinger, D., Tuomilehto, J., Collins, F. S., Groop, L., Altshuler, D., Collins, R., Lathrop, G. M., Melander, O., Salomaa, V., Peltonen, L., Orho-Melander, M., Ordovas, J. M., Boehnke, M., Abecasis, G. R., Mohlke, K. L., & Cupples, L. A. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat.Genet., 41,* 56-65.

Killeen, N., Stuart, S. G., & Littman, D. R. (1992). Development and function of T cells in mice with a disrupted CD2 gene. *EMBO J., 11,* 4329-4336.

Kiss-Toth, E., Bagstaff, S. M., Sung, H. Y., Jozsa, V., Dempsey, C., Caunt, J. C., Oxley, K. M., Wyllie, D. H., Polgar, T., Harte, M., O'neill, L. A., Qwarnstrom, E. E., & Dower, S. K. (2004). Human tribbles, a protein family controlling mitogen-activated protein kinase cascades. *J Biol Chem, 279,* 42703-42708.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., Sangiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308,* 385-389.

Klein, T. E., Altman, R. B., Eriksson, N., Gage, B. F., Kimmel, S. E., Lee, M. T., Limdi, N. A., Page, D., Roden, D. M., Wagner, M. J., Caldwell, M. D., & Johnson, J. A. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl.J.Med., 360,* 753-764.

Kneitz, B., Herrmann, T., Yonehara, S., & Schimpl, A. (1995). Normal clonal expansion but impaired Fas-mediated cell death and anergy induction in interleukin-2-deficient mice. *Eur.J.Immunol., 25,* 2572-2577.

Kolonel, L. N., Altshuler, D., & Henderson, B. E. (2004). The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat Rev.Cancer, 4,* 519-527.

Kooner, J. S., Chambers, J. C., guilar-Salinas, C. A., Hinds, D. A., Hyde, C. L., Warnes, G. R., Gomez Perez, F. J., Frazer, K. A., Elliott, P., Scott, J., Milos, P. M., Cox, D. R., & Thompson, J. F. (2008). Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat.Genet., 40,* 149-151.

Kooperberg, C. & Leblanc, M. (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet.Epidemiol., 32,* 255-263.

Koza, J. & Rice, J. (1991). Genetic generation of both the weights and architecture for a neural network. *IEEE Transactions, II*.

Kraft, P. & Hunter, D. J. (2009). Genetic Risk Prediction -- Are We There Yet? *N Engl.J.Med., 360,* 1701-1703.

Kurkova, V. (1991). Kolmogorov's Theorem is Relevant. *Neural Computation, 3,* 617-622.

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

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

Laurie, C., Mirel, D., Pugh, E., Bierut, L., Bhangale, T., Boehm, F., Caporaso, N., Edenburgh, H., Gabriel, S., Harris, E., Hu, F., Jacobs, K., Kraft, P., Landi, M., Lumley, T., Manolio, T., McHugh, C., Painter, I., Paschall, J., Rice, J., Zheng, X., & Weir, B. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology, In Press*.

Lee, S. W. (1996). Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18,* 648-652.

Leon Gordis (2008). *Epidemiology*. (4 ed.) Philadelphia: Saunders.

Lewis, G. F. & Rader, D. J. (2005). New insights into the regulation of HDL metabolism and reverse cholesterol transport. *Circ.Res., 96,* 1221-1232.

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

Likartsis, A., Vlachavas, I., & Tsoukalas, L. H. (1997). A new hybrid neural-genetic methodology for improving learning. *Ninth IEEE International Conference on Tools with Artificial Intelligence Proceedings,* 32-36.

Lin, S., Chakravarti, A., & Cutler, D. J. (2004). Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat.Genet., 36,* 1181-1188.

Lin, S. S. & Kelsey, J. L. (2000). Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research. *Epidemiol.Rev., 22,* 187-202.

Linder, R., Richards, T., & Wagner, M. (2007). Microarray data classified by artificial neural networks. *Methods Mol.Biol., 382,* 345-372.

Link, E., Parish, S., Armitage, J., Bowman, L., Heath, S., Matsuda, F., Gut, I., Lathrop, M., & Collins, R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N Engl.J.Med., 359,* 789-799.

Lou, X. Y., Chen, G. B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., & Li, M. D. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *American Journal of Human Genetics, 80,* 1125-1137.

Lucek, P., Hanke, J., Reich, J., Solla, S. A., & Ott, J. (1998). Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Hum.Hered., 48,* 275-284.

Lunetta, K. L., Hayward, L. B., Segal, J., & Van, E. P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet., 5,* 32.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychol.Methods, 7,* 19-40.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature, 456,* 18-21.

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., & Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat.Genet., 39,* 1181-1186.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N Engl.J Med., 363,* 166-176.

Manolio, T. A. (2009). Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. *Pharmacogenomics., 10,* 235-241.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461,* 747-753.

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat.Genet., 36,* 512-517.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet., 24,* 133-141.

Maxwell, S. E. & Delaney, H. D. (2004). *Designing Experiments and Analyzing Data*. (2nd ed.) Mahwah, New Jersey: Lawrence Erlbaum Associates.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., & Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat.Genet., 40,* 1166-1174.

McCarty, C., Chrisolm, R., Chute, C., Kullo, I., Jarvik, G., Larson, E., Li, R., Masys, D., Ritchie, M., Roden, D., Struewing, J., & Wolf, W. (2010a). The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics, In Revision*.

McCarty, C. A., Chrisholm, R. L., Chute, C. G., Kullo, I., Jarvik, G., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., Struewing, J., Wolf, W. A., & The eMERGE Team (2010b). The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Genomics (Submitted)*.

McCarty, C. A., Mukesh, B. N., Giampietro, P. F., & Wilke, R. A. (2007). Healthy People 2010 disease prevalence in the Marshfield Clinic Personalized Medicine Research Project: Opportunities for public health genomic research. *Personalized Medicine, 4,* 183-190.

McCarty, C. A., Wilke, R. A., Giampietro, P., Wesbrook S., & Caldwell, M. D. (2005). The Marshfield Clinic Personalized Medicine Research Project (PMRP) - design, methods and initial recruitment results for a population-based DNA Biobank. *Personalized Medicine, 2,* 49-79.

McClellan, J. & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell, 141,* 210-217.

Meiler, J. & Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. *Proc.Natl.Acad.Sci.U.S.A, 100,* 12105-12110.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat.Rev.Genet., 11,* 31-46.

Meyer-Lindenberg, A. & Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nat.Rev.Neurosci., 7,* 818-827.

Miller, S. G., Crowley, C., Lundstrom, J., Larson, C. J., Prior, K. E., & King, B. D. (2007). KC706, an Oral p38 MAP Kinse Inhibitor, Increases HDL-C. *Circulation, 116,* II_126.

Miyagawa, T., Nishida, N., Ohashi, J., Kimura, R., Fujimoto, A., Kawashima, M., Koike, A., Sasaki, T., Tanii, H., Otowa, T., Momose, Y., Nakahara, Y., Gotoh, J., Okazaki, Y., Tsuji, S., & Tokunaga, K. (2008). Appropriate data cleaning methods for genome-wide association study. *J Hum.Genet., 53,* 886-893.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics, 38,* 114-117.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum.Hered., 56,* 73-82.

Moore, J. H., Andrews, P. C., Barney, N., & White, B. C. (2008). Development and Evaluation of an Open-Ended Computational Evolution System for the Genetic Analysis of Susceptibility to Common Human Diseases. *Lecture Notes in Computer Science, 4973,* 129-140.

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010a). Bioinformatics challenges for genome-wide association studies. *Bioinformatics, 26,* 445-455.

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010b). Bioinformatics challenges for genome-wide association studies. *Bioinformatics, 26,* 445-455.

Moore, J. H., Barney, N., & White, B. C. (2008). Solving complex problems in human genetics using genetic programming: The importance of theorist-practitioner-computer interaction. *Genetic Programming Theory and Practice, 5,* 69-85.

Moore, J. H. & White, B. C. (2007). Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. *Genetic Programming Theory and Practice, 4,* 11-28.

Moore, J. H. & Williams, S. M. (2002). New strategies for identifying gene-gene interactions in hypertension. *Ann.Med., 34,* 88-95.

Moore, J. H. & Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays, 27,* 637-646.

Moore, J., Hahn, L., Ritchie, M., Thornton, T., & White, B. (2002). Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics. Langdon, WB, Cantu-Paz, E, Mathias, K, Roy, R, Davis, D, Poli, R, Balakrishnan, K, Honavar, V, Rudolph, G, Wegener, J, Bull, L, Potter, MA, Schultz, AC, Miller, JF, Burke, E, and Jonoska, N. Proceedings of the Genetic and Evolutionary Algorithm Conference. 1150-1155. San Francisco, Morgan Kaufman Publishers.

Moore, J., Parker, J., Olsen, N., & Aune, T. (2002). Symbolic discriminant analysis of microarray data in autoimmune disease. *Genet Epidemiol, 23,* 57-69.

Motsinger, A. A., Dudek, S. M., Hahn, L. W., & Ritchie, M. D. (2006). Grammatical evolution for the optimization of neural networks for genetic association studies. *Bioinformatics (In Submission)*.

Motsinger, A. A., Hahn, L. W., Dudek, S. M., Ryckman, K. K., & Ritchie, M. D. (2006). Alternative cross-over strategies and selection techniques for grammatical evolution optimized neural networks. *Proceedings of the 8th annual Genetic and Evolutionary Computation Conference (GECCO), 8,* 947-948.

Motsinger, A. A., Lee, S. L., Mellick, G., & Ritchie, M. D. (2006). GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics., 7,* 39.

Motsinger, A. A., Reif, D. M., Dudek, S. M., & Ritchie, M. D. (2006). Understanding the Evolutionary Process of Grammatical Evolution Neural Networks for Feature Selection in Genetic Epidemiology. *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology,* 1-8.

Motsinger, A. A., Reif, D. M., Fanelli, T. J., Davis, A. C., & Ritchie, M. D. (2007). Linkage Disequilibrium in Genetic Association Studies Improves the Performance of Grammatical Evolution Neural Networks. *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology,* 1-8.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., & Ritchie, M. D. (2008). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology, 32,* 325-340.

Motsinger-Reif, A. A., Fanelli, T. J., Davis, A. C., & Ritchie, M. D. (2008). Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. *BMC.Res.Notes, 1,* 65.

Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2009). Refining Genetic Algorithm Based Fuzzy Clustering through Supervised Learning for Unsupervised Cancer Classification. *Lecture Notes in Computer Science, 5483,* 191-202.

Mushiroda, T., Ohnishi, Y., Saito, S., Takahashi, A., Kikuchi, Y., Saito, S., Shimomura, H., Wanibuchi, Y., Suzuki, T., Kamatani, N., & Nakamura, Y. (2006). Association of VKORC1 and

CYP2C9 polymorphisms with warfarin dose requirements in Japanese patients. *J.Hum.Genet., 51,* 249-253.

Nagasubramanian, R., Innocenti, F., & Ratain, M. J. (2003). Pharmacogenetics in cancer treatment. *Annu.Rev.Med., 54,* 437-452.

Nelson, M. R., Kardia, S. L., Ferrell, R. E., & Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res., 11,* 458-470.

Nelson, M. R., Kardia, S. L. R., & Sing, C. F. (2000). The Combinatorial Partitioning Method. *Lecture Notes in Computer Science, 1848,* 293-304.

Neuman, R. J. & Rice, J. P. (1992). Two-locus models of disease. *Genet.Epidemiol., 9,* 347-365.

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., Papadakis, K., Voight, B. F., Scott, L. J., Zhang, F., Farrall, M., Tanaka, T., Wallace, C., Chambers, J. C., Khaw, K. T., Nilsson, P., van der, H. P., Polidoro, S., Grobbee, D. E., Onland-Moret, N. C., Bots, M. L., Wain, L. V., Elliott, K. S., Teumer, A., Luan, J., Lucas, G., Kuusisto, J., Burton, P. R., Hadley, D., McArdle, W. L., Brown, M., Dominiczak, A., Newhouse, S. J., Samani, N. J., Webster, J., Zeggini, E., Beckmann, J. S., Bergmann, S., Lim, N., Song, K., Vollenweider, P., Waeber, G., Waterworth, D. M., Yuan, X., Groop, L., Orho-Melander, M., Allione, A., Di, G. A., Guarrera, S., Panico, S., Ricceri, F., Romanazzi, V., Sacerdote, C., Vineis, P., Barroso, I., Sandhu, M. S., Luben, R. N., Crawford, G. J., Jousilahti, P., Perola, M., Boehnke, M., Bonnycastle, L. L., Collins, F. S., Jackson, A. U., Mohlke, K. L., Stringham, H. M., Valle, T. T., Willer, C. J., Bergman, R. N., Morken, M. A., Doring, A., Gieger, C., Illig, T., Meitinger, T., Org, E., Pfeufer, A., Wichmann, H. E., Kathiresan, S., Marrugat, J., O'Donnell, C. J., Schwartz, S. M., Siscovick, D. S., Subirana, I., Freimer, N. B., Hartikainen, A. L., McCarthy, M. I., O'Reilly, P. F., Peltonen, L., Pouta, A., de Jong, P. E., Snieder, H., van Gilst, W. H., Clarke, R., Goel, A., Hamsten, A., Peden, J. F., Seedorf, U., Syvanen, A. C., Tognoni, G., Lakatta, E. G., Sanna, S., Scheet, P., Schlessinger, D., Scuteri, A., Dorr, M., Ernst, F., Felix, S. B., Homuth, G., Lorbeer, R., Reffelmann, T., Rettig, R., Volker, U., Galan, P., Gut, I. G., Hercberg, S., Lathrop, G. M., Zelenika, D., Deloukas, P., Soranzo, N., Williams, F. M., Zhai, G., Salomaa, V., Laakso, M., Elosua, R., Forouhi, N. G., Volzke, H., Uiterwaal, C. S., van der Schouw, Y. T., Numans, M. E., Matullo, G., Navis, G., Berglund, G., Bingham, S. A., Kooner, J. S., Connell, J. M., Bandinelli, S., Ferrucci, L., Watkins, H., Spector, T. D., Tuomilehto, J., Altshuler, D., Strachan, D. P., Laan, M., Meneton, P., Wareham, N. J., Uda, M., Jarvelin, M. R., Mooser, V., Melander, O., Loos, R. J., Elliott, P., Abecasis, G. R., Caulfield, M., & Munroe, P. B. (2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nat.Genet., 41,* 666-676.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature, 456,* 98-101.

O'Neil, M. & Ryan, C. (2003). *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. (First ed.) Norwell, MA: Kluwer Academic Publishers.

Ott, J. (2001). Neural networks and disease association studies. *American Journal of Medical Genetics (Neuropsychiatric Genetics), 105,* 61.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS.Genet., 2,* e190.

Pattin, K. A., White, B. C., Barney, N., Gui, J., Nelson, H. H., Kelsey, K. T., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2008). A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genet.Epidemiol., 33,* 87-94.

Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet.Epidemiol., 32,* 381-385.

Peissig, P., Sirohi, E., Berg, R. L., Brown-Switzer, C., Ghebranious, N., McCarty, C. A., & Wilke, R. A. (2007). Construction of atorvastatin dose-response relationships using data from a large population-based DNA biobank. *Basic Clin.Pharmacol.Toxicol., 100,* 286-288.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., Amos, C. I., & Xiong, M. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur.J Hum.Genet., 18,* 111-117.

Penrod, N., Greene, C., & Moore, J. (2008). Failure to replicate a genetic association may provide important clues about genetic architecture. *Presented at the annual meeting of The American Society of Human Genetics, November 14 2008, Philadelphia PA.*.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature, 464,* 768-772.

Pietilainen, K. H., Soderlund, S., Rissanen, A., Nakanishi, S., Jauhiainen, M., Taskinen, M. R., & Kaprio, J. (2009). HDL subspecies in young adult twins: heritability and impact of overweight. *Obesity.(Silver.Spring), 17,* 1208-1214.

Pluzhnikov, A., Below, J. E., Konkashbaev, A., Tikhomirov, A., Kistner-Griffin, E., Roe, C. A., Nicolae, D. L., & Cox, N. J. (2010). Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am.J Hum.Genet, 87,* 123-128.

Poli, R., Langdon, W. B., & McPhee, N. F. (2008). *A Field Guide to Genetic Programming*. United Kingdom: Lulu Enterprises.

Porter, C. R. & Crawford, E. D. (2003). Combining artificial neural networks and transrectal ultrasound in the diagnosis of prostate cancer. *Oncology (Williston.Park), 17,* 1395-1399.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat.Genet., 38,* 904-909.

Pritchard, J. K. & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am.J.Hum.Genet., 65,* 220-228.

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am.J.Hum.Genet., 81,* 559-575.

R Development Core Team (2005). *R: A language and environment for statistical computing. ISBN 3900051070, URL http://www.R-project.org.* Vienna, Austria: R Foundation for Statistical Computing.

Race Ethnicity and Genetics Working Group (2005). The use of racial, ethnic, and ancestral categories in human genetics research. *Am.J.Hum.Genet., 77,* 519-532.

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R., Gough, J., Grimmond, S., Han, J. H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van, N. E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegner, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T., & Hayashizaki, Y. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell, 140,* 744-752.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., & Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444,* 444-454.

Reich, D. E. & Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genet.Epidemiol., 20,* 4-16.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human disorders. *Science, 273,* 1516-1517.

Ritchie, M. D. & Coffey, C. S. M. J. H. (2004). Genetic programming neural networks: A bioinformatics tool for human genetics. *Lecture Notes in Computer Science, 3102,* 438-448.

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., Basford, M. A., Brown-Gentry, K., Balser, J. R., Masys, D. R., Haines, J. L., & Roden, D. M. (2010). Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am.J Hum.Genet., 86,* 560-572.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am.J.Hum.Genet., 69,* 138-147.

Ritchie, M. D. & Motsinger, A. A. (2005). Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. *Pharmacogenomics., 6,* 823-834.

Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., & Moore, J. H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC.Bioinformatics., 4,* 28.

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., & Masys, D. R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin.Pharmacol.Ther., 84,* 362-369.

Roses, A. D., Saunders, A. M., Huang, Y., Strum, J., Weisgraber, K. H., & Mahley, R. W. (2007). Complex disease-associated pharmacogenetics: drug efficacy, drug safety, and confirmation of a pathogenetic hypothesis (Alzheimer's disease). *Pharmacogenomics.J., 7,* 10-28.

Ruano, D., Abecasis, G. R., Glaser, B., Lips, E. S., Cornelisse, L. N., de Jong, A. P., Evans, D. M., Davey, S. G., Timpson, N. J., Smit, A. B., Heutink, P., Verhage, M., & Posthuma, D. (2010). Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *Am.J Hum.Genet., 86,* 113-125.

Rubin, E. M., Krauss, R. M., Spangler, E. A., Verstuyft, J. G., & Clift, S. M. (1991). Inhibition of early atherogenesis in transgenic mice by human apolipoprotein AI. *Nature, 353,* 265-267.

Ryckman, K. & Williams, S. M. (2008). Calculation and use of the Hardy-Weinberg model in association studies. *Curr.Protoc.Hum.Genet., Chapter 1,* Unit.

Sabatti, C., Service, S. K., Hartikainen, A. L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M. I., Daly, M. J., Jarvelin, M. R., Freimer, N. B., & Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat.Genet., 41,* 35-46.

Sandhu, M. S., Waterworth, D. M., Debenham, S. L., Wheeler, E., Papadakis, K., Zhao, J. H., Song, K., Yuan, X., Johnson, T., Ashford, S., Inouye, M., Luben, R., Sims, M., Hadley, D., McArdle, W., Barter, P., Kesaniemi, Y. A., Mahley, R. W., McPherson, R., Grundy, S. M., Bingham, S. A., Khaw, K. T., Loos, R. J., Waeber, G., Barroso, I., Strachan, D. P., Deloukas, P., Vollenweider, P., Wareham, N. J., & Mooser, V. (2008). LDL-cholesterol concentrations: a genome-wide association study. *Lancet, 371,* 483-491.

Sato, F., Shimada, Y., Selaru, F. M., Shibata, D., Maeda, M., Watanabe, G., Mori, Y., Stass, S. A., Imamura, M., & Meltzer, S. J. (2005). Prediction of survival in patients with esophageal carcinoma using artificial neural networks. *Cancer, 103,* 1596-1605.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Bengtsson, B. K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rastam, L., Speliotes, E. K., Taskinen, M. R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., DeFelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G. W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., & Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science, 316,* 1331-1336.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat.Methods, 5,* 16-18.

Seldin, M. F. & Price, A. L. (2008). Application of ancestry informative markers to association studies in European Americans. *PLoS.Genet., 4,* e5.

Shao, H., Burrage, L. C., Sinasac, D. S., Hill, A. E., Ernest, S. R., O'Brien, W., Courtland, H. W., Jepsen, K. J., Kirby, A., Kulbokas, E. J., Daly, M. J., Broman, K. W., Lander, E. S., & Nadeau, J. H. (2008). Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc.Natl.Acad.Sci.U.S.A, 105,* 19910-19914.

Shen, R., Fan, J. B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Wickham, G. E., McBride, C., Steemers, F., Garcia, F., Kermani, B. G., Gunderson, K., & Oliphant, A. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutat.Res., 573,* 70-82.

Shepherd, J., Cobbe, S. M., Ford, I., Isles, C. G., Lorimer, A. R., MacFarlane, P. W., McKillop, J. H., & Packard, C. J. (1995). Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group. *N Engl.J.Med., 333,* 1301-1307.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res., 29,* 308-311.

Siddiqui, A., Kerb, R., Weale, M. E., Brinkmann, U., Smith, A., Goldstein, D. B., Wood, N. W., & Sisodiya, S. M. (2003). Association of multidrug resistance in epilepsy with a polymorphism in the drug-transporter gene ABCB1. *N Engl.J.Med., 348,* 1442-1448.

Sillanpaa, M. J. & Auranen, K. (2004). Replication in genetic studies of complex traits. *Ann.Hum.Genet., 68,* 646-657.

Sills, G. J. (2005). Pharmacogenetics of epilepsy: one step forward? *Epilepsy Curr., 5,* 236-238.

Sinnott-Armstrong, N. A., Greene, C. S., Cancare, F., & Moore, J. H. (2009). Accelerating epistasis analysis in human genetics with consumer graphics hardware. *BMC Res.Notes, 2,* 149.

Skinner, A. J. & Broughton, J. Q. (1995). Neural networks in computational materials science: training algorithms. *Modelling and Simulation in Materials Science and Engineering, 3,* 371-390.

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat.Genet, 38,* 209-213.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry*. (3rd ed.) New York: Freeman.

Spencer, C. C., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS.Genet., 5,* e1000477.

Sprinkhuizen-Kuyper, I. G. & Boers, E. J. (1998). The error surface of the 2-2-1 XOR network: The finite stationary points. *Neural Netw., 11,* 683-690.

Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 66,* 187-205.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de, G. A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavare, S., Deloukas, P., Hurles, M. E., & Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science, 315,* 848-853.

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L., Zhu, X., Brown, A., Pankow, J. S., Province, M. A., Hunt, S. C., Boerwinkle, E., Schork, N. J., & Risch, N. J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am.J.Hum.Genet., 76,* 268-275.

Tavares, J., Mesmoudi, S., & Talbi, E. (2009). On the Efficiency of Local Search Methods for the Molecular Docking Problem. *Lecture Notes in Computer Science, 5483,* 104-115.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin, C. Y., Jin, G. M., Jin, K. Y., Lee, J. Y., Park, T., Kim, K., Sim, X., Twee-Hee, O. R., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Hua, Z. J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y., Wright, A. F., Witteman, J. C., Wilson, J. F., Willemsen, G., Wichmann, H. E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K., Lucas, G., Luben, R., Loos, R. J., Lokki, M. L., Lettre, G., Langenberg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., Konig, I. R., Khaw, K. T., Kaprio, J., Kaplan, L. M., Johansson, A., Jarvelin, M. R., Cecile, J. W. J., Ingelsson, E., Igl, W., Kees, H. G., Hottenga, J. J., Hofman, A., Hicks, A. A., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Doring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J., de, F. U., Crawford, G., Collins, F. S., Chen, Y. D., Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, S. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E. S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor, H. A., Jr., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M., & Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466,* 707-713.

Thompson, J. F., Hyde, C. L., Wood, L. S., Paciga, S. A., Hinds, D. A., Cox, D. R., Hovingh, G. K., & Kastelein, J. J. (2009). Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. *Circ.Cardiovasc.Genet., 2,* 173-181.

Thornton-Wells, T. A., Moore, J. H., & Haines, J. L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet., 20,* 640-647.

Topchy, A. & Lebedko, O. A. (1997). Evolving Networks: Using the Genetic Algorithm with Connectionist Learning. *Nuclear Instruments and Methods in Physics Research, 389,* 240-241.

Toth, P. P. (2005). Adiponectin and high-density lipoprotein: a metabolic association through thick and thin. *Eur.Heart J, 26,* 1579-1581.

Turner, S. D. (2009). Visualizing sample relatedness in a GWAS using PLINK and R. https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Visualizing_relatedness.

Turner, S. D., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., Mirel, D. B., Paschall, J. E., Pugh, E. W., Rasmussen, L. V., Wilke, R. A., Zuvich, R. L., & Ritchie, M. D. (2010a). Quality Control Procedures for Genome-Wide Association Studies. *Current Protocols in Human Genetics, In press*.

Turner, S. D., Crawford, D. C., & Ritchie, M. D. (2009). Methods for optimizing statistical analyses in pharmacogenomics research. *Expert Reviews in Clinical Pharmacology, 2,* 559-570.

Turner, S. D., Dudek, S. M., & Ritchie, M. D. (2010b). Grammatical Evolution of Neural Networks for Discovering Epistasis among Quantitative Trait Loci. *Lecture Notes in Computer Science, 6023,* 86-97.

Turner, S. D., Dudek, S. M., & Ritchie, M. D. (2010a). ATHENA: A Knowledge-Based Hybrid Backpropagation-Grammatical Evolution Neural Network Algorithm for Discovering Epistasis among Quantitative Trait Loci. *BioData Mining, 3,* 5.

Turner, S. D., Dudek, S. M., & Ritchie, M. D. (2010c). Incorporating Domain Knowledge into Evolutionary Computing for Discovering Gene-Gene Interaction. *Lecture Notes in Computer Science, 6238,* 394-403.

Turner, S. D., Ritchie, M. D., & Bush, W. S. (2009). Conquering the Needle-in-a-Haystack: How Correlated Input Variables Beneficially Alter the Fitness Landscape for Neural Networks. *Lecture Notes in Computer Science, 5483,* 80-91.

Tyler, A. L., Asselbergs, F. W., Williams, S. M., & Moore, J. H. (2009). Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays, 31,* 220-227.

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de, B. B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., & Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol., 8,* R39.

Verweij, K. J., Zietsch, B. P., Medland, S. E., Gordon, S. D., Benyamin, B., Nyholt, D. R., McEvoy, B. P., Sullivan, P. F., Heath, A. C., Madden, P. A., Henders, A. K., Montgomery, G. W., Martin, N. G., & Wray, N. R. (2010). A genome-wide association study of Cloninger's temperament scales: Implications for the evolutionary genetics of personality. *Biol.Psychol.*.

Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., & Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS.Genet., 4,* e1000214.

Von Bubnoff, A. (2008). Next-generation sequencing: the race is on. *Cell, 132,* 721-723.

Vullo, A., Passerini, A., Frasconi, P., Costa, F., & Pollastri, G. (2008). On the Convergence of Protein Structure and Dynamics. Statistical Learning Studies of Pseudo Folding Pathways. *Lecture Notes in Computer Science, 4973,* 200-211.

Wallace, C., Newhouse, S. J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R. J., Marcano, A. C., Hajat, C., Burton, P., Deloukas, P., Brown, M., Connell, J. M., Dominiczak, A., Lathrop, G. M., Webster, J., Farrall, M., Spector, T., Samani, N. J., Caulfield, M. J.,

& Munroe, P. B. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am.J Hum.Genet., 82,* 139-149.

Wessels, J. A., van der Kooij, S. M., le, C. S., Kievit, W., Barerra, P., Allaart, C. F., Huizinga, T. W., & Guchelaar, H. J. (2007). A clinical pharmacogenetic model to predict the efficacy of methotrexate monotherapy in recent-onset rheumatoid arthritis. *Arthritis Rheum., 56,* 1765-1775.

White, B. C., Gilbert, J. C., Reif, D. M., & Moore, J. H. (2005). A statistical comparison of grammatical evolution strategies in the domain of human genetics. *Proceedings of the IEEE Congress on Evolutionary Computing,* 676-682.

Wilk, J. B., Manning, A. K., Dupuis, J., Cupples, L. A., Larson, M. G., Newton-Cheh, C., Semissie, S., DeStefano, A. L., Hwang, S. J., Liu, C., Yang, Q., & Lunetta, K. L. (2005). No evidence of major population substructure in the Framingham Heart Study. *Genetic Epidemiology, 29,* 234-292.

Wilke, R. A., Berg, R. L., Linneman, J. G., Peissig, P., Starren, J., Ritchie, M. D., & McCarty, C. A. (2010). Quantification of the clinical modifiers impacting high density lipoprotein (HDL) cholesterol in the community - Personalized Medicine Research Project (PMRP). *Preventive Cardiology, E-pub ahead of print*.

Wilke, R. A., Berg, R. L., Linneman, J. G., Zhao, C., McCarty, C. A., & Krauss, R. M. (2008). Characterization of low-density lipoprotein cholesterol-lowering efficacy for atorvastatin in a population-based DNA biorepository. *Basic Clin.Pharmacol.Toxicol., 103,* 354-359.

Willer, C. J., Pruim, R. J., Sanna, S., Welch, R. P., Teslovich, T. M., Gliedt, T. P., Boehnke, M., & Abecasis, G. R. (2010). LocusZoom: A fast web-based method for visual display of re-gional association results from genome-wide association scans. *Bioinformatics, In review.*

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., vey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., & Abecasis, G. R. (2008b). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat.Genet., 40,* 161-169.

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., vey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., & Abecasis, G. R. (2008a). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat.Genet., 40,* 161-169.

Wittke-Thompson, J. K., Pluzhnikov, A., & Cox, N. J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *Am.J Hum.Genet., 76,* 967-986.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc 6th Intl.Congress of Genetics, 1,* 356-366.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res., 30,* 303-305.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *J.Am.Med.Inform.Assoc., 17,* 19-24.

Yaeger, R., vila-Bront, A., Abdul, K., Nolan, P. C., Grann, V. R., Birchette, M. G., Choudhry, S., Burchard, E. G., Beckman, K. B., Gorroochurn, P., Ziv, E., Consedine, N. S., & Joe, A. K. (2008). Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa. *Cancer Epidemiol.Biomarkers Prev., 17,* 1329-1338.

Yan, W., Zhu, Z., & Hu, R. (1997). A hybrid genetic/BP algorithm and its application for radar target classification. *Proceedings of the IEEE 1997 National Aerospace and Electronics Conference,* 981-984.

Yang, J. M., Kao, C. Y., & Horng, J. T. (1996). Evolving neural induction regular language using combined evolutionary algorithms. *ISAI/IFIS 1996.Mexico-USA Collaboration in Intelligent Systems Technologies.Proceedings,* 162-169.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE, 87,* 1423-1447.

Yee, H. S. & Fong, N. T. (1998). Atorvastatin in the treatment of primary hypercholesterolemia and mixed dyslipidemias. *Ann.Pharmacother., 32,* 1030-1043.

Zhang, F., Wang, Y., & Deng, H. W. (2008a). Comparison of population-based association study methods correcting for population stratification. *PLoS.One., 3,* e3392.

Zhang, F., Wang, Y., & Deng, H. W. (2008b). Comparison of population-based association study methods correcting for population stratification. *PLoS.One., 3,* e3392.

Zhang, P., Sankai, Y., & Ohta, M. (1995). Hybrid adaptive learning control of nonlinear system. *Proceedings of the 1995 American Control Conference,* 2744-2748.

Zhang, Y., Sonnenberg, G. E., Baye, T. M., Littrell, J., Gunnell, J., DeLaForest, A., MacKinney, E., Hillard, C. J., Kissebah, A. H., Olivier, M., & Wilke, R. A. (2009). Obesity-related dyslipidemia associated with FAAH, independent of insulin response, in multigenerational families of Northern European descent. *Pharmacogenomics, 10,* 1929-1939.

Zheng, H., Jiang, M., Trumbauer, M. E., Hopkins, R., Sirinathsinghji, D. J., Stevens, K. A., Conner, M. W., Slunt, H. H., Sisodia, S. S., Chen, H. Y., & Van der Ploeg, L. H. (1996). Mice deficient for the amyloid precursor protein gene. *Ann.N.Y.Acad.Sci., 777,* 421-426.

Zhu, X., Tang, H., & Risch, N. (2008). Admixture mapping and the role of population structure for localizing disease genes. *Adv.Genet., 60,* 547-569.