A KNOWLEDGE-DRIVEN MULTI-LOCUS ANALYSIS OF MULTIPLE SCLEROSIS SUSCEPTIBILITY

By

William Scott Bush

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, 2009

Nashville, Tennessee

Approved:

Professor Douglas P. Mortlock

Professor Jonathan L. Haines

Professor Jay R. Snoddy

Professor James E. Crowe

Professor Marylyn D. Ritchie

In loving memory of **Anne Karpay**, my scientist in the stars…

ACKNOWLEDGEMENTS

This work was supported by grant HL65962 from the National Heart, Lung, and Blood Institute, grant AG20135 from the National Institute on Aging, and grant NS049477 from the National Institute of Neurological Disorders and Stroke. Thank you to the International Multiple Sclerosis Genetics Consortium for providing the opportunity to conduct a secondary analysis of their dataset. Thanks also to the Wellcome Trust Case Control Consortium for providing control samples. Dr. Phil De Jager generously provided a replication sample. Finally, a heartfelt thank you to all participants of these studies who entrust us with their DNA and their personal information – their selfless contributions drive our research.

Work presented in this dissertation was guided and greatly improved by input from the members of my thesis committee (Dr. Douglas P. Mortlock, Dr. Jonathan L. Haines, Dr. Jay R. Snoddy, Dr. James E. Crowe, and Dr. Marylyn D. Ritchie). Thanks to Dr. Haines for his expertise in "thinking through the data" and his extensive knowledge of disease gene mapping. Thanks to Dr. Mortlock for being an outstanding committee chair and for personal motivation and support.

Special acknowledgements are due to my Ph.D. mentor, scientific and professional advisor, and good friend Dr. Marylyn D. Ritchie. I have received the most outstanding academic training and have been exposed to numerous external ideas and viewpoints, both shaping and expanding my scientific mindset. Thanks most of all for understanding and teaching me the many priorities of life.

I thank the members of the Ritchie lab, especially Scott Dudek, Eric Torstenson, and Lance Hahn. Their expertise and guidance in computer science has greatly influenced my scientific thinking, in addition to providing software, code, and entire computational frameworks for this project. Special thanks to Todd Edwards for many insightful discussions (and arguments) that provided new perspectives. I thank Stephen Turner for several excellent discussions, for his statistical expertise, and for excellent project collaborations. Many thanks to

Ben Grady for poignant scientific and statistical questions, providing new ideas, and for great enthusiasm and support.

Many members of the Center for Human Genetics Research contributed to the development and implementation of ideas in this dissertation. Dr. Jacob McCauley and David Sexton provided expertise (and personal commentary) on numerous IMSGC projects. Jackie Bartlett generously helped to adapt and understand several scripts and statistical concepts. Justin Giles provided software, programming support, several innovative ideas, and many good laughs. Finally, Kylee Spencer consistently offers a level-headed approach to many scientific problems and has influenced multiple analysis and development decisions.

Several key individuals have contributed greatly to my broader professional development. I thank Dr. Dana Crawford and Dr. Tricia Thornton-Wells for many excellent career discussions, professional opportunities, and scientific insights. I also offer enormous gratitude to Maria Comer, not just for her excellent administrative support and guidance, but also for her personal encouragement and acumen.

Finally, the graduate students of the Program in Human Genetics have improved my work in enumerable ways. Special thanks to Digna Velez, Ryan Delehanty, Logan Dumitrescu, Kelli Ryckman, Sabrina Mitchell, Rebecca Zuvich, and Lauren Walters. By asking questions or requesting a detailed explanation of concepts, they have taught me more than lectures and exams ever could.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Technological revolution has transformed the field of disease gene discovery -- for better or worse – into a process of data collection, analysis, and interpretation of enormous proportions. Millions of genetic variations can be assessed in large samples, providing new analytical challenges and opportunities. Such advances were made to provide genetic insight into the development and progression of common diseases and conditions – phenotypes that, after decades of study, are described as *complex*.

There is a myriad of genetic, clinical, environmental, and biochemical components that interplay to influence common disease risk. In this work, I investigate the role that epistasis contributes to the complexity of a common human disease, multiple sclerosis. This disease was studied with a newly established tool in human genetics – genome-wide association – and I investigate the complex properties of these data and this disease by leveraging the vast amounts of biological knowledge now stored by the scientific community in elaborate database systems.

An exploration of the conceptual, analytical, and technological hurdles posed by the study of complex disease is presented in Chapter I. Definitions and examples of epistasis, complex disease, and genome-wide association studies are highlighted. The challenges of analytical complexity and the difficulties of processing and interpreting large-scale genetic data are also discussed.

Large-scale data is often stored using a database management system, and processing data within that system can provide great advantages. In Chapter II, a database procedure for mapping single genetic variations to larger genomic regions and genes is presented. This technique, called LD-Spline, is evaluated using a simulation study, and the results of applying this method to commonly used genome-wide association products are shown.

Once genomic regions and genes represented in a genome-wide association study are known, more sophisticated gene-centric analysis techniques can be applied. In Chapter III, a systematic method for incorporating knowledge about biological interactions among genes and gene products is discussed. This method, called Biofilter, provides a collection of structured knowledge-bases that can be used to synthesize biology-inspired genetic models. These models can then be tested in large-scale data. In this Chapter, I outline the theoretical basis for including biological knowledge into epistasis analysis, and I illustrate the utility of incorporating knowledge into the analysis of the genome-wide association study of multiple sclerosis.

In Chapter IV, the results of the multiple sclerosis interaction analysis are presented. The current state of multiple sclerosis genetics is discussed, along with examples and potential mechanisms for epistasis in this disease. This analysis serves as the functional application of the methods developed in Chapters II and III, and presents new findings and a potential role of a new biological mechanism in disease development and progression.

Finally, in Chapter V, the research conducted in this dissertation is concluded, providing perspective and a discussion of the benefits and difficulties of the approaches presented. The future directions of this work will be discussed, along with thoughts on systems biology and the ever-advancing pace of technology.

CHAPTER I

EPISTASIS IN GENOME-WIDE ASSOCIATION STUDIES OF COMPLEX DISEASE

Introduction

What is complex disease?

The central dogma of biology describes a hierarchical structure that governs all biological systems: DNA is transcribed to RNA, which is translated to amino acid sequences that fold into functional proteins. This dogma has developed and matured over many years to account for noted exceptions, such as functional RNA sequences, and the general hierarchy could be extended to describe higher level biological structure. Proteins work in concert to form the fundamental structural and chemical building blocks of the cell, providing channels, pores, receptors, and other molecular transport systems, and complex metabolic and regulatory pathways, eventually creating specialized organelles within the cell. The functional structure and output of cells delineate cell types. Aggregations of multiple cell types form tissues, and multiple tissues combine to form organs, organ systems, and ultimately an entire organism that interacts with the environment. In short, organisms are complex hierarchical systems with layers of interwoven components. A central goal of human genetics is to define relationships between DNA sequence variations and consequential changes in some level of this hierarchical system. Changes in the hierarchy that negatively impact how humans interact with the environment are generally called *disease*.

The earliest disease mechanisms identified by human genetic studies are now commonly called *Mendelian Disease*. The functional changes associated with these conditions in retrospect were easily detectable, as the disease trait is strongly and directly influenced by alterations of a

single gene that follow Mendelian inheritance. The loci associated with Mendelian disease phenotypes were identified by collecting affected families, genotyping a panel of genetic markers, and tracking the co-segregation of the genetic markers and the disease through pedigrees using linkage analysis. In the case of Mendelian disease, the strong negative impact of malfunctions in a single gene is so overwhelming that the complexity of the hierarchical system can be largely ignored when considering the condition.

Classic examples of Mendelian disease include cystic fibrosis (CF) and sickle cell anemia (Kerem et al., 1989; Pauling & Itano, 1949). CF is the most common lethal autosomal recessive disorder in the United States, affecting 30,000 individuals (Merlo & Boyle, 2003). Cystic fibrosis was linked to a 230 kb region on chromosome 7q31.3, identifying the CF transmembrane conductance regulator gene (*CFTR*). This protein is expressed in the apical membrane of exocrine epithelial cells lining the lungs, sinuses, pancreas, intestines, sweat ducts, and vas deferens. The CF phenotype is characterized by progressive bronchiectatic lung disease, pancreatic exocrine insufficiency, chronic sinusitis, and male infertility (Merlo & Boyle, 2003; Zielenski, 2000). The most common mutation in the *CFTR* gene is ΔF508, found in 70% of all CF chromosomes worldwide, and more than 50% of these are homozygous for the mutation. ΔF508 is a 3 bp deletion that results in the loss of a phenylalanine, affecting the ability of the protein to conduct chloride across the membrane, and while this mutation is by far the most common, the CF Genetic Analysis Consortium has reported more than 850 different mutant alleles.

Sickle cell anemia (SCA) is the first monogenic disease ever described (Nagel, 2001; Pauling & Itano, 1949). SCA is primarily due to an amino acid change in deoxyhemoglobin S (Ingram, 1957), causing alterations in polymerization leading to sickle shaped red blood cells with reduced capacity to carry oxygen (Eaton & Hofrichter, 1987). Lack of adequate oxygen flow induces periodic recurrant episodes of vasocclusive crisis causing progressively worsening injury to multiple organs. Much like CF, multiple distinct mutations can lead to the SCA phenotype, with differing severity of symptoms (Stuart & Nagel, 2004).

Table 1. Unresolved complexity of classic Mendelian genetic disorders.

Mendelian Phenotype	Primary Gene Mutation	Phenotypic Variation Explained by Mutation	Phenotypic Variation Presumably Influenced by Modifiers
Cystic Fibrosis	*CFTR*	Pancreatic exocrine insufficiency	Bronchiectatic lung disease severity Meconium ileuis (intestinal obstruction) Lung infection susceptibility
Sickle Cell Anemia	*HBB*	Dismorphic erythrocytes and reduced oxygen flow	Erythrocyte life span and density Erythrocyte adhesion to blood vessels Altered urine concentration Altered anion transport in erythrocyte cell membranes

While Mendelian disease typically has strong single gene effects, many instances are more complex than previously thought, with more subtle phenotypic aspects modulated by other genetic factors (Nagel, 2005; Zielenski, 2000), and commonly cited examples are shown in table 1. For example, variation in the CF lung phenotype cannot be explained by mutations in the *CFTR* gene alone. While the *CFTR* mutation class is predictive of pancreatic exocrine insufficiency (Merlo & Boyle, 2003), there is a full range of lung-disease severity among *CFTR* mutation classes I, II, and III which account for nearly 90% of CF in the US. The natural assumption is that environmental factors comprise the residual variance in respiratory severity, but studies of mucoid infection, tobacco use, and socioeconomic status suggest that these factors do not have a relatively significant effect. Furthermore, *CFTR* knockout mice are characterized by an intestinal obstruction that develops shortly after birth, similar to meconium ileus present in 20% of CF newborns. Genes in both humans (chromosome 19q13.2) and mice (chromosome 7) were found to modify the development and lethality of this condition, but severity of lung phenotype is significantly more difficult to study in model organisms. There are several genes affecting the CF lung phenotype independent of *CFTR* mutation (Collaco & Cutting, 2008), such as *TGFβ1* and *MBL2*. While none have been found to directly modulate CF, there is compelling evidence that alleles of these genes may contribute in combination to the lung phenotype.

In sickle cell anemia, several known epistatic modifiers alter severity and progression of the disease. α-thalassaemia mutations, the 158C->T mutation upstream of the β-globin-γ gene, and yet unknown genes in females all reduce disease severity (Stuart & Nagel, 2004). After accounting for the primary mutation, other remaining phenotypic variation includes the life span and density of red blood cells, sickle cell adhesion to blood vessel interiors, changes in urine concentration, and altered activity of anion transport proteins in red blood cell membranes (Nagel, 2001). These examples of unexplained variation present in Mendelian disease indicates that while the most pronounced trait follows a simple pattern of inheritance, the collection of all traits -- clinically defined as the disease phenotype -- is actually more complex in nature, involving subtle changes to multiple layers of the dogmatic hierarchal system.

In contrast to the comparatively rare Mendelian disease, more common conditions typically do not exhibit strong, highly penetrant effects from a single gene, and consequently do not follow consistent simple inheritance patterns. So-called *common complex diseases*, however, are generally still overrepresented in affected families indicating there is a genetic component and may be caused by genetic variants that are indistinguishable from normal human variation (Thomas & Kejariwal, 2004). The National Center for Health Statistics reports the top three leading causes of death in the U.S. for 2005 were heart disease, cancer, and stroke (National Center for Health and Statistics, 2008). Each of these broad disease conditions are thought to have genetic components that predispose or otherwise alter risk in the population. More highly penetrant familial forms of some common diseases exist, and have been studied to isolate important genes. In breast cancer, for example, variations in two tumor suppressor genes, Breast Cancer 1 early onset (*BRCA1*) and Breast Cancer 2 early onset (*BRCA2*) account for 30-40% of familial breast cancer, but explain only 2-3% of overall breast cancer prevalence (Wooster & Weber, 2003).

In the general case however, common complex diseases are characterized by the confounding influence of multiple genetic, environmental, and clinical factors. Type II diabetes

(T2D), for instance, has an increasing prevalence in the US population (Frayling et al., 2007), and has many well documented risk sources, including obesity, diet and exercise regimen, and genetic factors such as polymorphisms in the *PPAR-γ* gene (Deeb et al., 1998). Some population-specific genetic architectures incur a differential risk in response to similar environmental exposures, indicating a stronger role of gene-environment interaction. The Pima Indians, for example, have dramatically increased risk for T2D compared to Caucasian populations (Knowler et al., 1978), and this risk difference is thought to be attributed in part to major changes in diet and lifestyle over the last 100 years. The Pima Indians, and some Polynesian populations, have seen explosions in T2D prevalence as those communities shift from traditional diets to a "Western" diet (Steyn et al., 2004). Because other populations adapt to dietary changes without the subsequent increase in T2D risk, the Pima Indians must have a genetic architecture that interacts with dietary factors to influence T2D. In addition, obesity has also been a confounding factor for several genetic studies of T2D. The fat mass and obesity associated gene (*FTO*) is significantly associated to T2D, but the effect is due to increased body mass index (BMI), which subsequently increases T2D risk (Frayling, 2007). While these findings are still scientifically relevant for the etiology and understanding of obesity, they do not answer the larger question of why some obese individuals develop T2D while others do not.

As highlighted by the Pima Indians example, ethnicity is often a contributing risk factor for complex diseases. For example, multiple sclerosis is far more prevalent in the Caucasian population than in Asian or African populations (Lowis, 1990). Conversely, prostate cancer has much higher prevalence and mortality in African-American versus other populations (Powell, 1998). In most cases of complex disease where there is differential risk among ethnicities, carefully controlled studies have eliminated shared environment or shared cultural aspects as the source of this effect. To investigate population-specific genetic architectures, the International HapMap Project has documented differences in allele frequency and linkage disequilibrium across multiple human ethnic subpopulations by genotyping over four million single nucleotide

polymorphisms (SNPS) (Frazer et al., 2007; International HapMap Consortium, 2005). Caucasian, Yoruba, Han Chinese and Japanese populations were sampled initially, and in the most recent phase of the project, Toscans from Italy, Luhya and Maasai from Kenya, and US individuals with African and Mexican ancestry have been included. Data from Caucasian, Yoruba, Chinese, and Japanese populations have been analyzed by many groups to quantify population-specific differences (Nielsen et al., 2005; Weir et al., 2005) and to highlight differential regions of potential evolutionary selection (Sabeti et al., 2007; Voight et al., 2006). These ethnicity-specific differences in genetic architecture -- multiple sequence variations scattered throughout the human genome -- are likely related to the differences in disease risk.

In addition to confounding factors, intricate effects operating among multiple genetic variants are thought to be hallmarks of complex disease (Thornton-Wells, Moore, & Haines, 2004). Among these are various forms of heterogeneity that presumably complicate genetic studies. *Allelic heterogeneity* occurs when multiple different alleles at the same locus increase risk of disease, such as the numerous rare mutations of the *CFTR* gene that give rise to cystic fibrosis. Two or more distinct genetic loci can independently increase risk of disease in the case of *locus heterogeneity*, as in the case of tuberous sclerosis, where both *TSC1* and *TSC2* have been identified in families with the disorder (Povey et al., 1994; Young & Povey, 1998). *Phenotypic heterogeneity* occurs with an imprecise definition of the disease phenotype. Conditions that were previously called autism are now thought to be part of a larger complex mix of disease phenotypes called autism spectrum disorders (Collaborative Linkage Study of Autism, 2001; Shao et al., 2002), and using a more defined phenotypic subset may improve future genetic studies.

What is epistasis?

In addition to multiple forms of heterogeneity, there is also *epistasis* or gene-gene interaction at play in complex disease. Epistasis was first described by William Bateson as the effect of one gene masking (or literally standing upon) the effect of another (Bateson, 1909). The

Bateson view of epistasis has also been described as biological epistasis (Moore & Williams, 2005), similar to a biochemist's observation that variation in the physical interaction of biomolecules affects a phenotype (Moore, 2003). From a statistical perspective, epistasis was also observed as multi-allelic segregation patterns by R.A. Fisher who mathematically described the phenomenon as deviation from additivity in a linear model of genotypes (Fisher, 1918). The Fisher definition is more flexible, as it can describe how multiple variations can in concert influence a phenotype without the direct physical interaction of gene products. In the broader sense, statistical epistasis and biological epistasis should eventually converge as scientific understanding progresses and high level functional relationships between genes and gene products are elucidated.

Given the complexities of known biological pathways that involve numerous intra-molecular interactions, epistasis is presumed to be ubiquitous both statistically and biologically (Moore, 2003). This belief is driven largely by the notion that large networks of gene regulation and protein-protein interaction have a functional endpoint that may be influenced by the simultaneous presence of multiple variants in those genes (Moore, 2003; Moore & Williams, 2005). Functional epistasis has been well documented in model organisms, and was discovered early in the field of genetics. Lancefield described a two-locus inheritance pattern for the forked bristle phenotype in *Drosophila* (Lancefield, 1918). A few years later, Bridges discovered statistical epistasis in *Drosophila* eye color, where collections of several different alleles Mendelize with various eye color phenotypes (Bridges, 1919). These alleles influence a common set of biochemical pathways controlling eye pigmentation that was described many years later (Lloyd, Ramaswami, & Kramer, 1998). More recently, studies of mouse and rat chromosome substitution strains revealed substantial epistasis in over 140 quantitative trait loci (Shao, 2008).

Epistasis has been "rediscovered" by genetic epidemiologists in recent years. Candidate gene association studies, which attempt to implicate a single common variation as influencing disease, produced initial results that often fail to replicate in successive studies (Hirschhorn et al.,

2002). Epistasis was proposed as a potential reason for this non-replication of single-SNP effects. Suppose the effect of one allele is conditional on the presence of a second unknown allele not assayed in a study population or sample. If this second allele is of high frequency in the population or sample, the effect of allele one will be seen as a single main effect. In a new, replication sample, the second allele may be at a lower frequency, and the effect of the first allele will not be seen. This would be viewed as a failure to replicate, even if the joint effect of the two alleles is consistent across all samples and populations. Epistasis was also proposed as a reason for the more general failure of complex disease studies. Because complex diseases likely involve small effects from multiple genes that may interact, linkage and candidate gene studies may have failed to account for the full genetic architecture that is influencing risk.

As previously mentioned, epistatic modifiers of disease severity have been found for CF and SCA (Kerem et al., 1989; Pauling & Itano, 1949). In addition to the modulation of Mendelian disease, epistasis has been functionally demonstrated to play a role in common complex disease. Most notably, Hirschsprung's disease was found to be influenced by polymorphisms in *RET* and the *ERDB2* receptor in the Old Order Amish and was confirmed in a mouse model (Carrasquillo et al., 2002). Having both variants simultaneously increases risk of disease far beyond the combined risk of each independent variant.

What are genome-wide association studies?

In the face of the statistical and biological complications related to complex disease, traditional methods of study design and analysis have not fared well. Linkage analysis has produced few compelling findings in complex disease studies (Altmuller et al., 2001), especially in diseases with high sibling risk ratios. Autism and multiple sclerosis are both ostensibly great candidates for linkage analysis, but multiple studies of both phenotypes have yielded no new consistent genetic risk factors. Where linkage analysis was successful for complex disease, it was applied to rare familial forms of the phenotype, such as familial breast cancer (Wooster & Weber,

2003). Likewise, candidate gene studies produced an abundance of statistical associations, but only a scant few replicate (Hirschhorn et al., 2002). Over the last five years, cost-effective, high-throughput genotyping technologies have expanded our ability to explore the human genome. These advances have opened the door to a new study design paradigm -- the genome-wide association (GWA) study. In these studies, individuals are surveyed for 500,000 to over 1 million single nucleotide polymorphisms, capturing much of the common genetic variation across the genome (approximately 85%, depending on the platform) (Barrett & Cardon, 2006; Hirschhorn & Daly, 2005). The underlying principle of the GWA study is that blocks of linkage disequilibrium – contiguous regions of genomic sequence that flow through populations – can be marked or tagged by SNP markers. A disease-related variant that lies in one of these tagged genomic regions will be detectable as an association between the tagging SNP and the phenotype. This approach relies on the *common disease, common variant hypothesis* -- the idea that risk for common diseases is influenced primarily by common variations in the human genome (alleles in > 5% of the population) (Reich & Lander, 2001).

The first major "success" in GWA studies was the identification of complement factor H (*CFH*) as a causal factor for age-related macular degeneration (Klein et al., 2005). This study used a panel of just over 100,000 SNPs in a modest sample size (96 cases and 50 controls). While the *CFH* finding was discovered using the GWA approach, two concurrently published studies also identified the gene using traditional study designs -- Haines et al. using a family-based linkage approach (Haines et al., 2005), and Edwards et al. using a candidate SNP approach (Edwards et al., 2005). The *CFH* association has since been replicated in several other studies (Hageman et al., 2005; Thakkinstian et al., 2006; Zareparsi et al., 2005), and while this finding did not validate GWA studies as a successful alternative to linkage analysis or candidate gene studies, it does illustrate its generally utility.

The popularity of the GWA study design has increased dramatically, and over the past few years a tsunami of new genetic associations has been published. As of May 2008, 202 GWA

studies were published, reporting 436 novel SNP associations (Hindorff, Junkins, & Manolio, 2008). A summary of these results is shown in figure 1, reproduced from (Manolio, Brooks, & Collins, 2008).

For most of these studies, the influential genetic variants associated to the phenotype explain a very small proportion of the estimated overall disease heritability. This indicates that, while successful, these studies have only scratched the surface of common complex disease genetics. Many other heritable factors could be at play, such as methylation patterns or other epigenetic phenomenon, structural variations such as copy number polymorphisms, insertions, deletions, and inversions, or environmental-driven alterations in gene expression. Another possibility is that a large portion of the disease heritability is explained by epistasis, where the interaction of multiple alleles increases risk above and beyond the independent alleles. Epigenetics, structural variation, and gene-environment interactions all require additional data sets and new study designs to properly evaluate, but in general, epistasis can be evaluated with existing genotype data. It is therefore notable that to the author's knowledge, none of the 202 published GWA studies include a search for interactions.

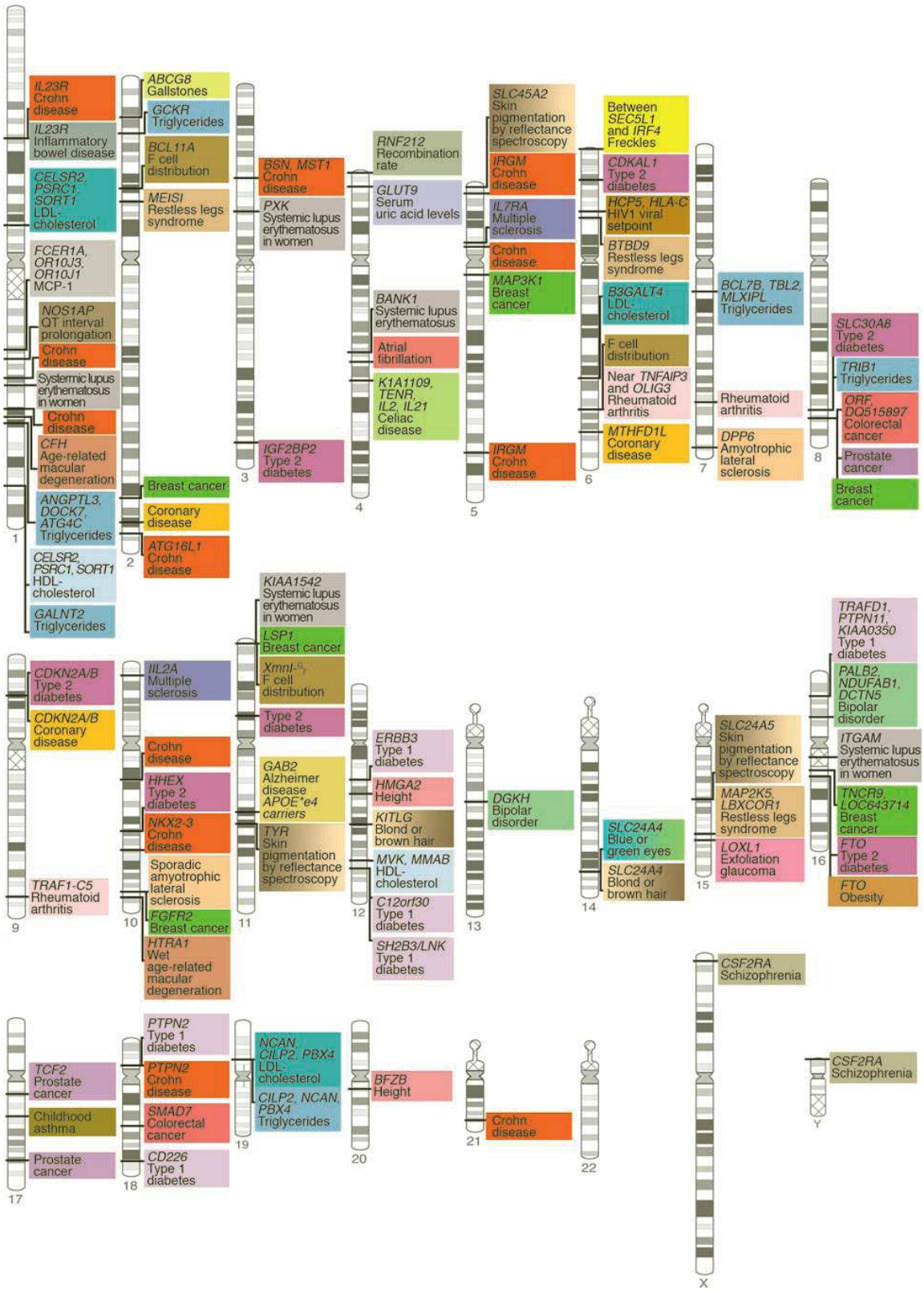

Figure 1. Significant SNP-trait associations detected in GWA studies to date.

Strategies for Epistasis Analysis

Analytical approaches

Searching for and characterizing epistasis even in small scale data is a challenge. In recent years, a number of computational and statistical techniques have been developed for or applied to the identification of epistasis in case/control data. An overview of these methods is shown in table 2. Automated Detection of Informative Combined Effects (DICE) (Tahri-Daizadeh et al., 2003) is a regression-based approach that systematically explores available variables and models based on a variation of the Akaike information criterion (AIC), similar to a step-wise regression approach. Regression models of increasing complexity are fitted to the data and the change in AIC is assessed to determine if adding a variable increases the model fit. DICE is computationally limited to three-locus models or less.

Classification and Regression Trees (CART) (Breiman et al., 1984) is a commonly used approach that partitions a dataset into subsets using the value of variables in the data, finding the data splitting procedure that best classifies categorical outcomes (or provides the best regression fit for continuous outcomes). For a categorical outcome, CART begins by selecting the variable that best classifies the examples in the dataset, typically using a measure of information gain such as the Gini index. The Gini index measures node "impurity" or the degree of classification error within each partitioned category. For each subset produced by the split based on this variable, the procedure is repeated, producing a second-level subset. This procedure continues recursively until the optimal partitioning of the data is produced. Numerous procedures have been developed to prune the tree and reduce model over-fitting, and various splitting measures have been applied to CART also.

One limitation of CART is that subsets are defined using values of a single variable. If the outcome is determined by a non-linear combination of variable values – such as a multi-locus genotype (Moore & Ritchie, 2004) – CART does not perform well (Ritchie et al., 2007). Patterning

and Recursive Partitioning (PRP) (Bastone et al., 2004) is an extension of CART that attempts to resolve this shortcoming by essentially pre-processing the input variables to generate a "dummy-encoded" set of multivariate states that can be used by the CART-like recursive partitioning procedure (RP). For example, if variable 1 has genotypes AA, Aa, and aa, and variable 2 has genotypes BB, Bb, and bb, the patterning procedure would produce nine new variables corresponding to the nine multi-locus genotypes (AABB, AABb, AAbb, AaBB, etc). RP could then split the data using non-linear combinations of multi-locus genotypes.

Table 2. Analytical approaches for epistasis analysis.

Acronym	Method	References	Description
DICE	Automated Detection of Informative Combined Effects	(Tahri-Daizadeh et al., 2003)	Step-wise search of regression model space using Akaike information criterion
CART	Classification and Regression Trees	(Breiman et al., 1984)	Iterative procedure to systematically split data into subsets that improve outcome classification
RF	Random Forests	(Breiman, 2001)	Induction of multiple classification or regression trees on bootstrap samples, which are averaged or voted on to generate a final model
PRP	Patterning and Recursive Partitioning	(Bastone et al., 2004)	Extension of CART to encode multivariate states based on original input variables
PLR	Penalized Logistic Regression	(Park & Hastie, 2008)	Step-wise regression modeling procedure with penalty functions that provide more stable coefficient estimates and reduce overfitting
	Logic Regression	(Kooperberg et al., 2001)	Regression modeling that induces combinations of attributes using logical operators
MARS	Multi-Adaptive Regression Splines	(Cook, Zee, & Ridker, 2004)	Recursive partitioning approach that uses regression splines to produce subset classifications
MDR	Multifactor Dimensionality Reduction	(Ritchie et al., 2001)	Partitioning approach that builds high and low risk multi-locus genotype combinations
MDR-PDT	Multifactor Dimensionality Reduction Pedigree Disequilibrium Test	(Martin et al., 2006)	MDR procedure looking for over-transmission of multi-locus genotype combinations in extended pedigrees
GMDR	Generalized Multifactor Dimensionality Reduction	(Lou et al., 2008)	Generalized linear modeling procedure using multi-locus genotype combinations to compare case/control frequencies or familial transmissions
	AMBIENCE	(Chanda et al., 2008)	Information theory-based evaluation of multivariate combinations

Random Forests (RF) (Breiman, 2001) are another extension of CART, where multiple classification (or regression) trees are induced, each on a bootstrap sample of the data. A final classification is derived by "voting" over all trees in the forest. Much like CART, many developments and modifications to RF have been made, including multiple measures of variable importance, and multiple voting systems or averaging over the trees in the forest (Chipman, George, & McCulloch, 1998; Chipman, George, & McCulloch, 2002).

Several regression-based approaches have been applied to multi-locus analysis. Penalized logistic regression (Park & Hastie, 2008) adds a penalty function to the traditional logistic regression likelihood equation to allow more stable estimates of regression coefficients. Logic regression (Kooperberg et al., 2001) uses logical operators (AND, OR, NOT, etc) to create complex variable patterns with an associated regression coefficient. Multi-Adaptive Regression Splines (MARS) (Cook, Zee, & Ridker, 2004) uses a recursive partitioning strategy like CART, but uses a basis function (typically a spline) for modeling, allowing more complex, non-linear relationships to be fitted.

Multifactor Dimensionality Reduction (MDR) (Ritchie et al., 2001) is a brute-force machine learning approach that exhaustively builds and classifies multi-locus combinations into high-risk or low-risk categories based on the ratio of cases to controls with each multi-locus genotype. The two-state risk variable derived for each set of genetic variables is compared to the case/control status to produce a classification error for that model. Models are then ranked by classification error to select the best multivariate combination. This procedure is implemented with cross-validation to estimate a prediction error and prevent over-fitting of data, and a Monte Carlo permutation test is applied to assess statistical significance.

More recently, several methods have been extended or developed to explore family-based data. Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (MDR-PDT) was developed by Martin et al. to examine multi-locus models in extended pedigree data (Martin et al., 2006). MDR-PDT applies the MDR procedure to the transmission of multiple loci to

affected offspring in pedigrees. Similarly, Generalized Multifactor Dimensionality Reduction (GMDR) is a model-based approach that uses MDR-based multi-locus genotype combinations to compare cases and controls or transmitted versus non-transmitted alleles in family data (Lou et al., 2008). Finally, AMBIENCE is a relatively new computational approach that applies common information theory measures to genotype data. Information entropy – a variance-like value for categorical data – is used to determine the degree of information (or certainty) about case/control status that a multivariate combination captures. These statistics are used to identify and model gene-gene and gene-environment interactions in both case/control and family data (Chanda et al., 2008).

Translating approaches for large-scale analysis

Adapting these approaches for genome-wide association studies has noted computational challenges (Moore & Ritchie, 2004). There are roughly 125 billion possible two-SNP models in a set of 500,000 SNPs, and the number of higher order models increases exponentially. Exhaustively analyzing all of these possible combinations with even the most basic of statistical tests is computationally costly and in some cases intractable. Applying the more elegant and powerful approaches outlined above only increases the computational complexity of the problem.

Some statistical methods are amenable to algorithm optimization and adaptation to multi-processor computing clusters. The MDR algorithm was retooled and parallelized (pMDR) (Bush, Dudek, & Ritchie, 2006), which reduced computation time roughly linearly with the number of processors used. Optimized forms of CART (Breiman et al., 1984), Brute (Segal, 1998), and other decision tree approaches have been used for several years as large-scale datamining tools, and may be useful for GWA analysis. In this spirit, a tree-based rule-mining approach called Apriori was applied to simulated data containing complex combinations of epistasis and genetic heterogeneity (Bush, Thornton-Wells, & Ritchie, 2007). This study found that

computationally efficient association rule mining has the same ability as MDR to identify two-locus and three-locus interaction effects, but loses power for higher order models. Association rules also had a modest ability to capture heterogeneity effects. Furthermore, many traditional statistical approaches such as logistic regression are available in highly optimized forms and could allow high-throughput analysis of genotype data (Marchini, Donnelly, & Cardon, 2005; Rouhani-Kalleh, 2007), and many non-traditional machine learning methods rely on permutation testing or bootstrapping to assess statistical significance, and those procedures are ideal for parallelization. As such, increases in computing power and volume, coupled with new algorithm development can make many established methods possible for GWA analysis.

In addition to brute-force datamining search procedures, evolutionary computing has been used to search for high-dimensional models in genetic association data. Techniques like genetic programming neural networks (GPNN) (Ritchie et al., 2007), grammatical evolution neural networks (GENN) (Motsinger-Reif & Ritchie, 2008), and symbolic discriminant analysis (SDA) (Moore et al., 2002; Moore et al., 2007) discover complex mathematical relationships between SNPs and a phenotype using principles of evolution and natural selection. With these approaches, a population of complex mathematical functions is initialized. Each function accepts genetic variables as input and classifies cases and controls as output. Each individual in this population of mathematical functions is potentially an optimal classifier, with an associated fitness value (generally a classification error). Solutions are mated, recombined, and mutated to produce a new population of solutions with presumably increased fitness. After a user-defined number of generations, the procedure is ended and the best solution is reported. While these search procedures cannot guarantee to find all relationships in the data, they routinely find functional signals in simulated genetic studies (Motsinger et al., 2006; Motsinger-Reif et al., 2008).

While algorithm optimization can decrease computation time for large-scale interaction analysis, the exponential trend of the problem and the ever increasing number of SNPs captured by modern genotyping technologies limits this solution. Search strategies provide an alternative

to exhaustive evaluation, but have many user-specified parameters that alter the speed and scope of the analysis. Even when such strategies are applicable, the biological interpretation of results from these procedures can be difficult. As such, new approaches for finding epistasis in GWA studies are needed.

New Approaches for Large-Scale Epistasis Analysis

Analytical filtering

One analysis strategy is to reduce or filter the set of genotyped SNPs, eliminating redundant or ostensibly useless information. A simple and common way to filter SNPs is to select a set of results from a single-SNP analysis based on an arbitrary significance threshold and exhaustively evaluate interactions in that subset. This can be perilous, however, as the most significant results from a single-SNP analysis aren't always the most likely to replicate (Zaykin & Zhivotovsky, 2005). Also, selecting SNPs to analyze based on main effects will prevent certain multi-locus models from being detected – so called "purely epistatic" models with vanishingly small, statistically undetectable marginal effects (Frankel & Schork, 1996; Moore et al., 2004). With these models, a large component of the heritability is concentrated in the interaction rather than in the main effects. In other words, a specific combination of markers (and only the combination of markers) incurs a significant change in disease risk. The benefits of this analysis are that it performs an unbiased analysis for interactions within the selected set of SNPs. It is also far more computationally and statistically tractable than analyzing all possible combinations of markers.

Analytical filters also can be applied to identify unknown structures within the data. Clustering (Falush, Stephens, & Pritchard, 2003) and principle components analysis (Price et al., 2006) are commonly used to identify and correct for population substructure in case/control

data. Similarly, various forms of cluster analysis have been applied to genotype data to identify more genetically homogenous sub-groups (Thornton-Wells, Moore, & Haines, 2006). These sub-groups are then analyzed separately to discover susceptibility variants within each unique genetic architecture. Moore et al. has conducted several studies using Relief-F and tuned Relief-F, statistics that perform clustering in genotype space to estimate the propensity of non-linear variable interactions (Moore & White, 2007). Relief-F has been proposed as a preprocessing step prior to the more computationally intensive MDR analysis (Pattin et al., 2008). There are also continued developments in machine learning, with hybrid learner-classifier systems used to capture multiple types of genetic models and effects in large-scale data (McKinney et al., 2007).

Analytical filtering can have disadvantages. Some analytical approaches use case/control status in the filtering process, and this likely impacts the false-positive rate of the overall statistical analysis in ways that are difficult to control for or adjust. Likewise, clustering and principle components analysis both produce subsets of the data which may under some circumstances reduce statistical power, so the effectiveness of these techniques are data and model dependent.

Knowledge-based filtering

Another approach to reducing the number of interaction tests is to generate multi-SNP models based on prior biological knowledge, for example testing for interactions only between SNPs that occur in the same biochemical pathway (Carlson et al., 2004). This approach attempts to take into account information about the known structure of biological systems to reduce the set of interactions that are analyzed (see Chapter III). There are many known biological mechanisms that may spawn the development of epistasis, such as gene regulatory networks, protein-protein interactions, and regulatory pathways. Providing a biology-based mechanism for why two SNPs might interact aids the interpretation of a multi-SNP statistical model, in that a functional relationship between the two genes in the model is already established.

As more and more GWA data becomes available, efficient and effective techniques for incorporating prior knowledge into multivariate analysis are needed. Borecki and Province have proposed a novel Bayesian modeling approach that incorporates pathway information (Province & Borecki, 2008), and can perform on a genome-wide scale. Work by David Conti uses expert knowledge ontologies to build hierarchical Bayesian models to analyze GWA data (Lewinger et al., 2007). Rather than performing filtering, these techniques leverage prior information when building genetic models while allowing all available SNPs to be included.

The disadvantage of knowledge-based filtering or analysis is that these strategies could suffer from the incomplete state of our knowledge of complex biological systems. Experimental science has dramatically advanced our understanding of many metabolic and regulatory pathways, but due to changes in funding or emphasis, certain processes are certainly better understood than others. Cell cycle pathways have been heavily emphasized in cancer research over the last 30 years, for example. There are also still many mechanisms and processes about which we understand very little, and if any of these were involved in complex disease etiology, knowledge-based approaches would certainly not fare well. It is quite possible, however, that there are subtle genetic effects embedded in many processes which we do understand, but have never investigated in the context of a particular disease, in which case, knowledge-based approaches would discover novel relationships between mechanism and disease. In short, the success of these techniques relies on the quality of information provided to them – something that is currently unknown.

General Issues with Epistasis in GWA Studies

Interpretation of results

Interpreting the wealth of statistical results that emerges from a GWA study is a challenge, even for a basic single-locus analysis (Pearson & Manolio, 2008). Even a small scale GWA study produces hundreds of thousands of statistical results that likely contain many false positive associations. Even if false positives could be eliminated, the remaining results may not have a clear biological meaning, such as an association to a SNP in a gene desert. In addition, some analysis approaches may pose a large computational burden when considering large-scale data. On top of the sheer number of results, a bevy of factors can influence statistical association, including clinical definition of the phenotype (Amos, 2007), assessment of population stratification (Falush, Stephens, & Pritchard, 2003), genotyping quality control, accounting for known clinical and environmental confounders, and appropriate replication (Chanock et al., 2007). All of these factors are amplified when conducting an epistasis analysis, along with a host of new issues that arise.

When assessing a significant multi-SNP model, a primary question is "What is the nature of the SNP-SNP interaction?" A significant multi-SNP model does not always imply a true statistical interaction of alleles or genotypes – this could simply imply the coupling of two strong independent effects, or a locus heterogeneity effect. Likelihood ratio tests using logistic regression are often employed to ask this specific question. If a true interactive effect is confirmed, the biological basis of the interaction should be explored. While statistical interactions do not imply biological interaction of molecules, those with a biological meaning are perhaps more compelling. If the two SNPs lie on a common haplotype background, the two-SNP model likely represents one signal resulting from the simultaneous over-transmission of the two SNPs together. In population-based studies, haplotype sizes are relatively small (approximately 35 KB), but in family-based studies, haplotypes can extend over several mega-base regions.

There may also be a biological relationship that implicates an interaction between two SNPs, such as the co-occurrence of two genes in a common biochemical pathway. To assess this possibility, the genomic context of each SNP should be investigated to determine if it lies within or near a gene. The location of SNPs within the genic region could also suggest a mechanism for interaction, as the two genes may have shared regulatory elements or binding sites.

Annotating results with biological information often reveals enrichment of significant associations within a pathway (Subramanian et al., 2007; Subramanian et al., 2005). This can potentially illuminate a new biological mechanism for disease pathology – a new *functional association*. This biological mechanism can serve as a platform for generating new testable statistical or biological hypotheses. These factors, among many others, should be considered when selecting multi-SNP models for follow-up in a replication set.

Challenges in replication

The NHGRI working group established criteria for bona fide replication of GWA study results (Chanock et al., 2007). Basic conditions for a positive replication include a sufficient sample size to replicate the detected effect, an independent replication set, the same outcome phenotype for both data sets, a similar study population, similar magnitude and 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. In addition to all these criteria, replication of a multi-locus model presents new challenges. As with a single-SNP association, the direction of all effects in the model should be consistent across the screening and replication stages, and ideally both the model fit and the interaction component should be statistically significant by a likelihood ratio test.

The notion of a pathway effect or other higher-level functional association is problematic in terms of replication. How to effectively test for pathway enrichment in a second data set is an

open question in the field of human genetics. Should the same set of pathway SNPs be associated in a replication, or can a different set of genes be represented? Should interactions among all genes in the pathway be considered? If not, could inconsistent direction of effects be attributable to interactions and heterogeneity within the pathway? Even if there is a consensus on how to reproduce a pathway-enrichment effect, it may be of limited usefulness without a concise set of risk factors to assess. Functional exploration of the numerous single SNP associations made to date will likely take years to complete, and experimentally studying the functional implications of multiple variants within a biological system is a combinatorial challenge of its own (Jansen, 2003). So, ideally statistical evidence of gene-gene interaction should be concise and compelling.

Conclusions

The initial wave of genome-wide association study findings represents a dramatic first step toward our understanding of common human diseases. However, the monumental task of adequately exploring the collected data has just begun, and new high-throughput methods of assessing genetic variation are on the horizon. As data quality improves, data quantity increases, and new types of data become available, it will become even more important to place that data in the context of current biological knowledge and explore all the ways that genetic, environmental, and clinical factors can combine to influence disease risk. The complex networks that constitute metabolic and regulatory function are so intricate and interwoven -- even at our current level of scientific understanding -- that there are a myriad of theoretical mechanisms by which epistatic segregation patterns can alter a phenotype. While the computational challenge of exploring epistasis in GWA studies is great, nature (along with 150 years of scientific endeavor) has provided a framework for beginning this exploration – a dogmatic hierarchical system of biological organization common to all known forms of life.

Acknowledgements

This work was motivated and inspired by the discussions and reviews of Scott Williams, Jason Moore, Marylyn Ritchie and Tricia Thornton-Wells.

CHAPTER II

LD-SPLINE: MAPPING SNPS ON GENOTYPING PLATFORMS TO GENOMIC REGIONS AND GENES USING PATTERNS OF LINKAGE DISEQUILIBRIUM

Introduction

What is linkage disequilibrium?

Recent advances in high-throughput genotyping technology have ushered in the era of genome-wide association (GWA) studies (Morton, 2008). The GWA approach has seen much success over the last few years, identifying many novel genetic effects for a multitude of human disease phenotypes (Manolio, Brooks, & Collins, 2008). The underlying philosophy of this research approach is that a dense panel of single nucleotide polymorphisms (SNPs) can mark broader genomic regions by exploiting patterns of linkage disequilibrium.

Linkage disequilibrium (LD) is a term first coined by Lewontin and Kojima in the field of population genetics (Lewontin & Kojima, 2001), and simply describes the non-random association of alleles at multiple loci. LD arises when a mutation occurs near a marker on a common haplotype background (Borecki & Province, 2008). If in subsequent generations there is no recombination between the marker and the mutation, the pair is passed together to offspring in the next generation. When assayed, the mutation and the marker always appear together in the population, and over time the haplotype carrying the mutation can become common. Eventually, through multiple generations and recombination events, in some individuals the marker and the mutation are separated by a recombination event. As this occurs more and more in the population over generations, the LD decays, or approaches *linkage equilibrium*, where the marker and the mutation appear independent in the population. The decay of LD is similar in concept to radioactive decay, and is directly related to the genetic distance between the two markers (the frequency of recombination events expected between the two).

Numerous phenomenon in population genetics and evolutionary biology can impact LD structure (Slatkin, 2008). Patterns of mating, geographic subdivision, natural selection, and mutation can all change LD. Genetic drift, for example, can create LD between nearby markers simply by oversampling a multi-marker haplotype. Similarly, population bottlenecks or subdivisions effectively resample an LD structure from the larger population, producing chance haplotype effects, thereby increasing LD (Schmegner et al., 2005; Zhang et al., 2004). Along these lines, various attributes of LD have been exploited to identify regions of positive selection (Sabeti et al., 2007).

LD has recently become of great interest to genetic epidemiologists, as patterns of LD proved useful for fine mapping of disease genes, and later for large-scale surveys of much of the human genome. These patterns manifest in SNP data as correlations between genotypes of nearby SNPs in the panel, and is generally caused as these SNPs on a common genomic background are transmitted through human subpopulations. In gene mapping studies, there are indirect and direct associations (Carlson et al., 2004). An indirect association can be found if an influential polymorphism is located on the larger genomic region surveyed by genotyping other SNPs that mark the region. Any genotyped SNPs on the same genomic background as the influential polymorphism would appear associated to the disease in the study. If the influential variant itself is genotyped in the study, it would have a direct association to the phenotype. Generally when a SNP is associated and sufficiently replicated, the genomic region surrounding this SNP is re-sequenced to identify the true influential variation.

Measures of linkage disequilibrium

Many measures of LD have been proposed (Devlin & Risch, 1995), but all are ultimately related to the frequency difference between a two-marker haplotype and the frequency expected under the assumption that the two markers are independent. The two commonly used measures of linkage disequilibrium are D′ and r^2 (Devlin & Risch, 1995; International HapMap Consortium,

2005) shown in equations 1 and 2. In these equations, π_{12} is the frequency of the ab haplotype, $\pi_{1\bullet}$ is the frequency of the a allele, and $\pi_{2\bullet}$ is the frequency of the b allele.

Eq. 1

$$D' = \begin{cases} \dfrac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{1\bullet}\pi_{\bullet 2}, \pi_{\bullet 1}\pi_{2\bullet})} \; if \; \pi_{11}\pi_{22} - \pi_{12}\pi_{21} > 0 \\ \dfrac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{1\bullet}\pi_{\bullet 1}, \pi_{\bullet 2}\pi_{2\bullet})} if \; \pi_{11}\pi_{22} - \pi_{12}\pi_{21} < 0 \end{cases}$$

Eq. 2

$$r^2 = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{(\pi_{1\bullet}\pi_{2\bullet}\pi_{\bullet 1}\pi_{\bullet 2})^{1/2}}$$

D′ is a population genetics measure that is related to recombination events between markers, and is scaled between 0 and 1. A D′ value of 0 indicates complete linkage equilibrium, implying frequent recombination between the two markers, and statistical independence under principles of Hardy-Weinberg equilibrium. A D′ of 1 indicates complete linkage disequilibrium, implying no recombination between the two markers. Alternatively, r^2 is the square of the correlation coefficient, and is a more statistical measure of shared information between two markers. The r^2 measure is commonly used to determine how well one SNP can act as a surrogate for another. There are dependencies between these two statistics -- r^2 is sensitive to the allele frequencies of the two markers, and can only be high in regions of high D′.

One often forgotten issue associated with LD measures is that they are based (at some level) on a two-marker haplotype frequency. Current technology does not allow direct measurement of these frequencies from a sample – each SNP is genotyped independently, and the *phase,* or chromosome of origin for each allele, is unknown. Many well developed and documented methods for inferring haplotype phase and estimating the subsequent two-marker haplotype frequencies exist (Weir, 1979), and generally lead to reasonable results (Fallin & Schork, 2000).

The International HapMap Project cataloged distinct patterns of LD in four human sub-populations: Yoruba, Caucasian, Han Chinese, and Japanese (International HapMap Consortium,

2005). Phase I of this project examined 2.5 million SNPs across the human genome, and computed pair-wise D' and r^2 statistics in 500KB windows. These values were made publicly available as flat-file downloads from the HapMap project, release 21. Phase III of the Hapmap project expands the available populations to include Toscans from Italy, Luhya and Maasai from Kenya, and US individuals with African and Mexican ancestry.

Existing methods for SNP to gene mapping

Enrichment analysis of GWA single marker results is a common procedure to examine the functional relationships between genes in the significant marker set. In addition to single marker association methods, many new bioinformatics and statistical techniques take a gene-centric approach to analysis. Aubert et al. proposed a gene-based local false discovery rate (FDR) procedure (Aubert et al., 2004). Li et al. proposed prioritizing SNPs within candidate genes in genome-wide scans to improve power, using an FDR analysis on result subsets (Li et al., 2008). Lewinger et al. and Province and Borecki proposed elegant pathway-based Bayesian approaches to GWA analysis, incorporating gene information into SNP analysis (Lewinger et al., 2007; Province & Borecki, 2008). All these techniques require relating SNPs on a genotyping platform to genes in the genome. A gene-centric approach to GWA epistasis analysis is described in Chapter III. As such, a systematic and user-controlled method for mapping SNPs to the broader genomic regions they mark – and ultimately to genes – is needed.

The simplest approach for generating SNP-gene relationships is to determine if a SNP lies within the exonic or intronic region outlined by a genomic build. Some approaches pad the gene boundaries with a user-defined region upstream and downstream to account for possible linkage disequilibrium (see methods of (Torkamani, Topol, & Schork, 2008)). There are also several approaches for generating LD statistics that can then be used to partition genomic regions captured by genotyping platforms. The popular PLINK software has two options for generating LD information (Purcell,; Purcell et al., 2007). r^2 descriptive statistics can be computed quickly by

simply computing correlations between genotypes. Inferential statistics, population estimates of D' and r^2 can also be computed by PLINK but this procedure is much more computationally costly as it requires phasing haplotypes. Another approach is LdCompare, which can rapidly compute pair-wise r^2 values from genotype data, and can also generate multi-marker correlations when given phased data (Hao, Di, & Cawley, 2007). While these approaches provide valuable information about the redundancy of information captured by a genotyping platform, they do not readily relate a single SNP to a genomic region – that must be accomplished by a post-processing step to call haplotype blocks.

Currently, haplotype blocks are generally identified using two approaches, the Gabriel et al. method and the four gamete rule. These two approaches are implemented in Haploview software, and produce a global haplotype block partition for a given set of SNP genotypes. Both procedures are sequential, beginning with the first SNP in the dataset and defining non-overlapping blocks upstream. While these approaches provide the general haplotype structure of a given genomic region, they are global rather than SNP-centric procedures. These approaches could misrepresent the genomic region a particular SNP marks based on the global sequential nature of the partitioning strategy.

To the author's knowledge, there are no automated SNP-based procedures for systematically relating SNPs to genes or genomic regions using LD. In this work, we present an algorithm to accomplish this task by processing pair-wise LD statistics to identify the genomic region that a particular SNP putatively represents. This algorithm is implemented as a MySQL aggregate function and performs genomic region and gene assignments for collections of SNPs, such as GWA SNP marker lists, using locally stored LD information from the International HapMap Project.

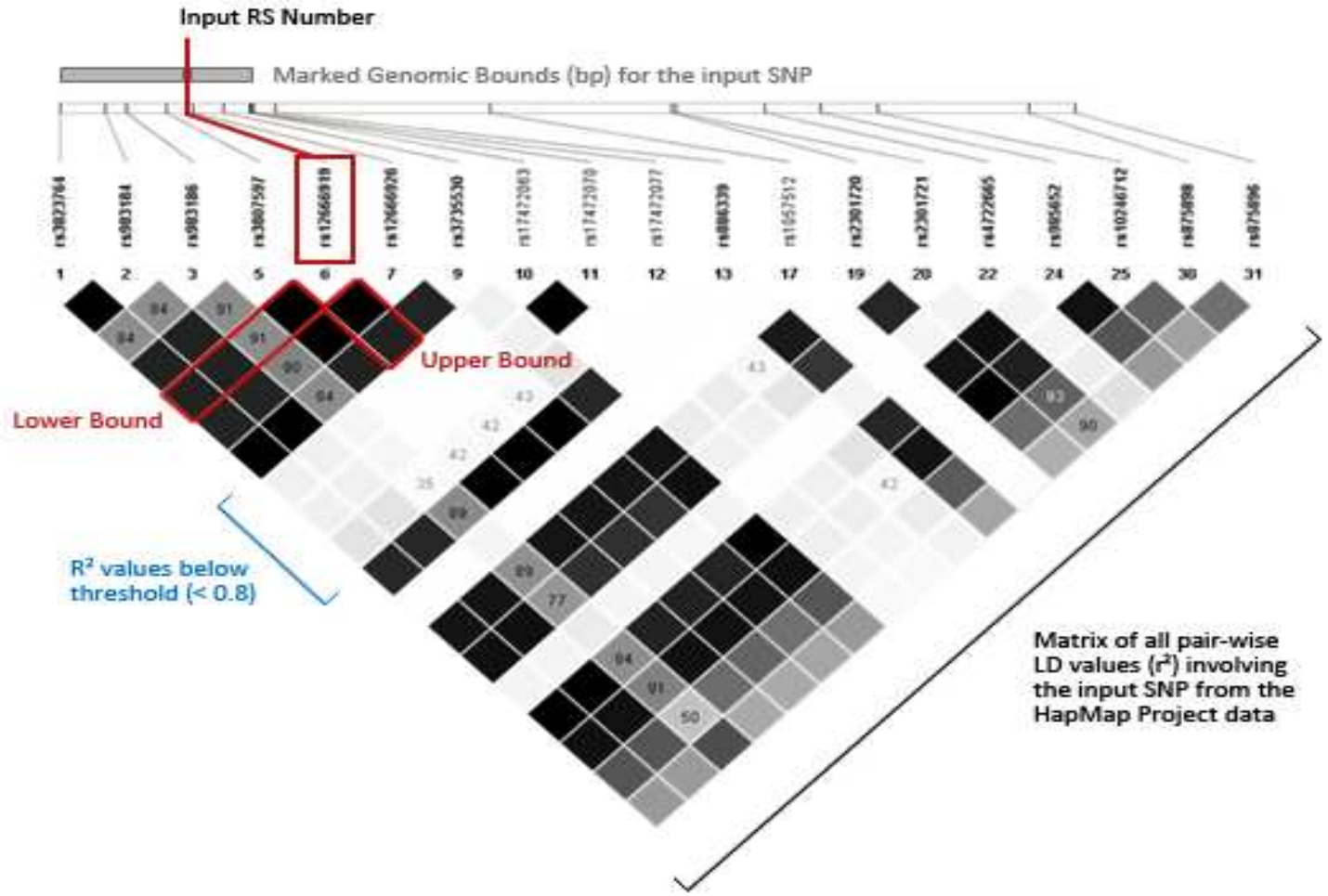

Figure 2. Overview of the LD-Spline Algorithm. A matrix of all HapMap-based pair-wise LD values (D′ or r²) is retrieved from the database. Using this matrix, the lower bound is incrementally extended to downstream SNPs while the pair-wise LD value between the downstream SNP and the input SNP is greater than the user-defined threshold (in this case $r^2 > 0.8$). The process is repeated for the upper bound to define the marked genomic bounds for the input SNP.

Methods

LD-Spline algorithm

To execute the LD-Spline function, a user specifies the following: the LD statistic to be used (D′ or r^2), an LD statistic threshold value (ranging between 0 and 1), and a reference sequence (RS) SNP identifier. The RS ID is used to query the specified LD statistic for all pair-wise values that exist in the Hapmap data that include the specified SNP. The procedure is illustrated in figure 2 and outlined in algorithm 1, and can be applied to LD values corresponding to any population.

Algorithm 1

Input: RS number of the SNP to map (rs_id), table or matrix of pair-wise LD values

1. Initialize the upper and lower bounds of the marked genomic region with the position of the input SNP.
2. Retrieve the value of the selected LD measure corresponding to the input SNP and the next downstream SNP, SNP X.
3. If the LD value is greater than the threshold value, change the lower bound of the marked genomic region to the position of SNP X.
4. Repeat 3 and 4 to extend the lower bound until the retrieved LD value is less than the threshold value.
5. Repeat 2 – 4 to define the upper bound.

The LD-Spline algorithm was implemented in C++ as an aggregate function for the MySQL database management system. The aggregate function, ldspline is executed twice ; once to define the upper bound and once to define the lower bound. These results are joined to produce the full mapped genomic region for a SNP or set of SNPs. The ldspline function accepts four arguments: a SNP index, LD measure (either dprime (D′) or rsquared (r^2), a threshold value (between 0 and 1), and a flag value to indicate an upper bound (0) or lower bound (1) search. Let

us define a table 'CEU' that contains pair-wise D' and r^2 statistics downloaded, inserted, and indexed by a composite key – a pair of indices that reference the two SNPs for which the LD values apply. Let us also define a table 'index_2_rs' that relates a SNP index to an RS number, and a single value, 'PlatformSNP' as the RS number of a SNP on a genotyping platform that we wish to relate to a genomic region. The SQL statement to map this SNP using an r^2 threshold of 0.8 would be:

```
SELECT lower_bound, position, upper_bound FROM
        (SELECT ldspline(A.pos2, A.rsquared, 0.8, 0, A.pos1) as upper_bound, A.pos1 AS position FROM
                (SELECT * FROM CEU inner join
                        (select pos from index_2_rs on rs_id = PlatformSNP) as f
                        where CEU.pos1 = f.pos
                ) AS A GROUP BY position) AS C
NATURAL JOIN
        (SELECT ldspline(B.pos1, B.rsquared, 0.8, 1, B.pos2) as lower_bound, B.pos2 AS position FROM
                 (SELECT * FROM LD.CEU inner join
                        (select pos from index_2_rs on rs_id = PlatformSNP) as g
                        where LD.CEU.pos2 = g.pos
                ) AS B GROUP BY position) AS D ;
```

Instead of mapping a single SNP, we could instead choose to map the entire platform of SNPs with one statement. In this case, let us define a table 'Genotyping_Platform' that contains an indexed set of RS IDs. The SQL statement to map the entire table of SNPs using a D' threshold of 0.9 would be:

```
SELECT lower_bound, position, upper_bound FROM
        (SELECT ldspline(A.pos2, A.rsquared, 0.8, 0, A.pos1) as upper_bound, A.pos1 AS position FROM
                (SELECT * FROM CEU inner join
                        (select pos from Genotyping_Platform a inner join LD.index_2_rs b on a.rs_id = b.rs_id) as f
                        where CEU.pos1 = f.pos
                ) AS A GROUP BY position) AS C
NATURAL JOIN
```

```
(SELECT ldspline(B.pos1, B.rsquared, 0.8, 1, B.pos2) as lower_bound, B.pos2 AS position FROM
        (SELECT * FROM LD.CEU inner join
                (select pos from Genotyping_Platform a inner join LD.index_2_rs b on a.rs_id = b.rs_id) as g
                where LD.CEU.pos2 = g.pos
        ) AS B GROUP BY position) AS D ;
```

Processing a table of approximately 600,000 SNPs using the user-defined aggregate function has a runtime of approximately 36 hours on a dual Xeon 3.06 GHz machine with 2 GB of RAM.

For ease of evaluation, we also produced a command-line version of this algorithm using the Perl scripting language. This version is functionally equivalent to the MySQL aggregate function, but rather than accessing a database table of pair-wise LD values, it reads LD values from a flat file.

Data simulations

We simulated realistic patterns of linkage disequilibrium to mimic two human CEU chromosomal regions using genomeSIMLA (genomeSIM version 2.0.4 software, functionally equivalent to genomeSIMLA 1.0 for LD generation) (Edwards et al., 2008). genomeSIMLA is a forward-time population simulator that uses random mating, genetic drift, recombination, and population growth to produce SNP genotype data with linkage disequilibrium. The general procedure for generational advancement is shown in figure 3.

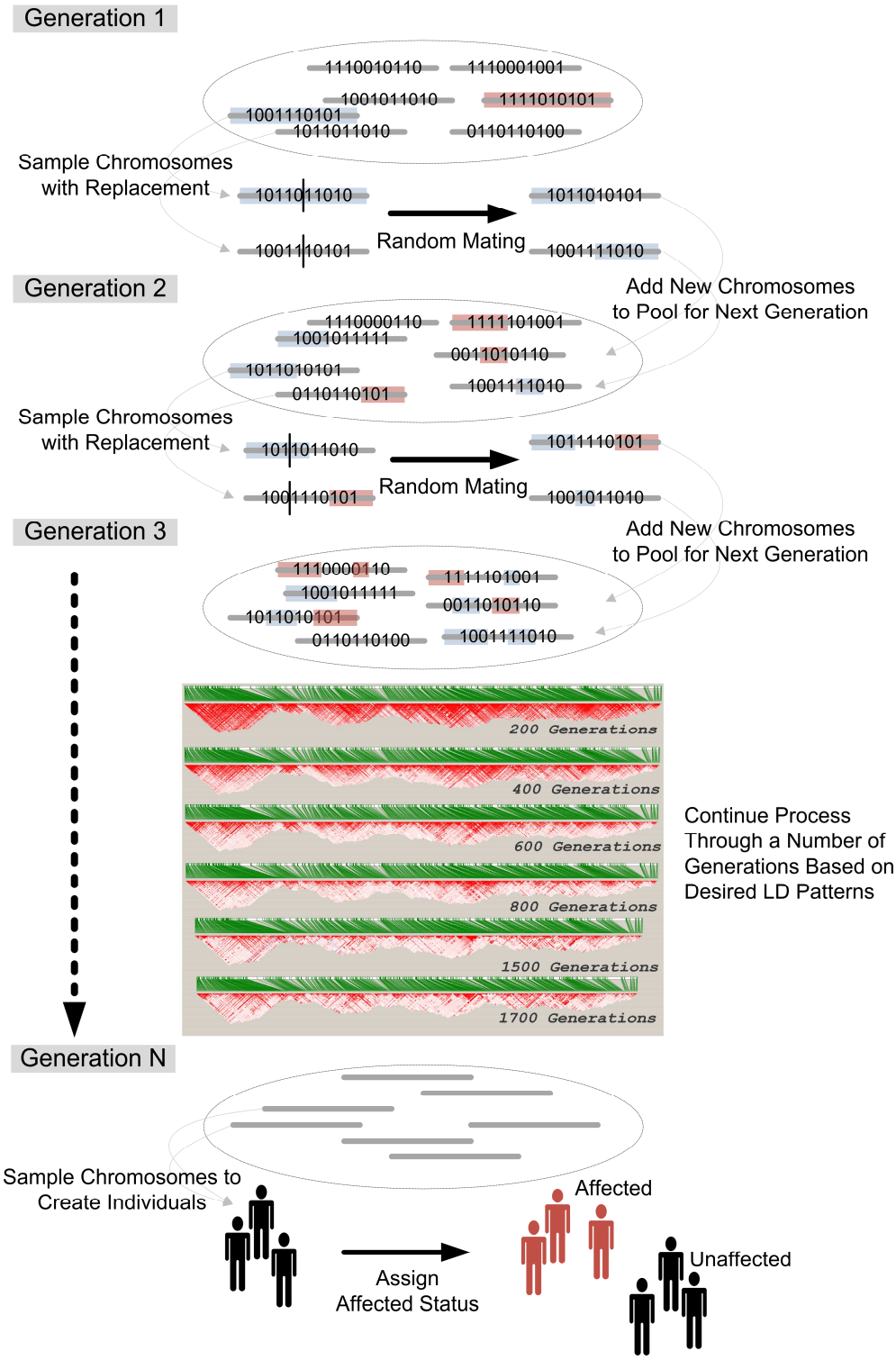

Figure 3. Overview of the genomeSIMLA process. Chromosomes are randomly initialized in the first generation, and then randomly sampled with replacement and crossed to produce the next generation. This process continues until the population has the desired LD patterns. Individuals are then sampled from this population for datasets.

Synthetic chromosomes were initialized using random allele frequencies. 1367 SNPs from chromosome 1 were selected from 792,429 bp to 9,965,572 bp, and 1146 SNPs from chromosome 18 were selected from 23,719,514 bp to 24,217,521 bp. All simulated SNPs were included in the HapMap CEU dataset, and HapMap build 35 positional information for each SNP was used. Recombinant gametes are created by sampling chromosomes with replacement from the population and crossing over based on intermarker recombination probabilities are determined by the Kosambi function map distance based on a 1 centimorgan per 1 million bases of physical distance. The number of recombination events per gamete is drawn from a Poisson distribution. Two gametes are combined to form a new individual for the next generation. This mating and recombination process continues for a user-specified number of generations, the size of each generation is determined by a logistic growth model.

The initial population size was 750, and was advanced over 454 generations using the Richard's growth curve (A = 750, B = 0.02, C = 1,200,000, M = 500, T = 0.01, Var = 0.03) to produce a final population size of 100,000 chromosomes. These parameters are a slight variation on an optimal set described in (Edwards et al., 2008). Once this population was generated, we produced 100 datasets consisting of 2,000 controls (a null genetic model was used). The random seed for these simulations was 2,225. For this simulated population, we manually selected 10 haplotype blocks and recorded their upper and lower bounding SNPs. genomeSIMLA tracks recombination events through generational advancement of a population, so the exact haplotype blocks are reported by the simulation. genomeSIMLA also reports exact D' and r^2 statistics computed for the entire population.

Block definition algorithms

In addition to the LD-Spline approach, we evaluated two block calling algorithms implemented in the popular Haploview software (Barrett et al., 2005): the Gabriel et. al approach (Gabriel et al., 2002) and the four-gamete rule (Barrett et al., 2005). Gabriel et. al used the 95%

confidence intervals of D′ estimates to establish stretches of "strong LD" (Gabriel et al., 2002). D′ estimates are unstable when sample size is small or allele frequency is low, so the confidence intervals of the statistic are used. If the D′ 95% confidence upper bound is > 0.98 and the lower bound is > 0.7, there is little statistical evidence of a historical recombination event between the two markers, and thus they form a haplotype block. Alternatively, the four-gamete rule is based on an algorithm described by Wang et al. where the frequency of the four possible two-marker haplotypes are computed for each pair of SNPs (Wang et al., 2002). Rather than estimating D′ confidence intervals, the four-gamete rule is similar to estimating a confidence interval on the two-marker haplotype frequencies. If all four haplotypes are observed with at least a frequency of 0.01, a recombination event between the two markers likely occurred. These two algorithms were applied to simulated unphased datasets, and the resulting haplotype block partitioning was recorded.

Block partitions were defined by these two algorithms, and compared to three parameterizations of the LD-Spline algorithm: D′ threshold of 0.6, D′ threshold of 0.8, and D′ threshold of 1. For each data simulation, a SNP that lies within each of the 10 selected haplotype blocks was randomly chosen. The LD-Spline approach used these SNPs as input for the algorithm, and haplotype blocks were defined around these SNPs. The Haploview-based algorithms were used to produce a full list of haplotype blocks for each dataset. This list was parsed to identify haplotype blocks that contain the randomly selected SNPs, and the bounds for those blocks were recorded.

Algorithm comparisons

The upper and lower bound SNP indices were compared to the true block boundaries for each block partitioning algorithm using weighted Kappa statistics to assess inter-rater (algorithm) agreement (Wickens, 1989). Weights for the Kappa statistic were calculated using a

standard weighting strategy shown in equation 3, incurring an increased penalty as the number of SNPs from the correct boundary edges increase.

Eq. 3 $$w_{ij} = 1 - \frac{|i - j|}{k - 1}$$

In equation 3, *i* is a row index and *j* is a column index of the boundaries specified by the two algorithms, and *k* is the maximum number of possible boundaries the algorithm could call.

The full weighted Kappa statistic is shown in equation 4 (Cohen, 1968). Agreement was evaluated within each of the 10 simulated haplotype blocks and for the overall block partitioning over 100 datasets. Kappa statistics were calculated using STATA 10.

Eq. 4 $$\kappa_w = \frac{\sum_{cells} w_{ij}\pi_{ij} - \sum_{cells} w_{ij}\pi_{i\bullet}\pi_{\bullet j}}{1 - \sum_{cells} w_{ij}\pi_{i\bullet}\pi_{\bullet j}}$$

Results

Simulated data

An overview of the linkage disequilibrium present in our simulated population is shown in figure 4. The parameters used in this simulation recapitulate reasonable patterns of linkage disequilibrium, similar to those seen in Hapmap data (Edwards et al., 2008). A more detailed view of two simulated haplotype blocks on chromosome 1 is shown in figure 5. The blocks selected for evaluation ranged in SNP density from 5 SNPs to 2 SNPs.

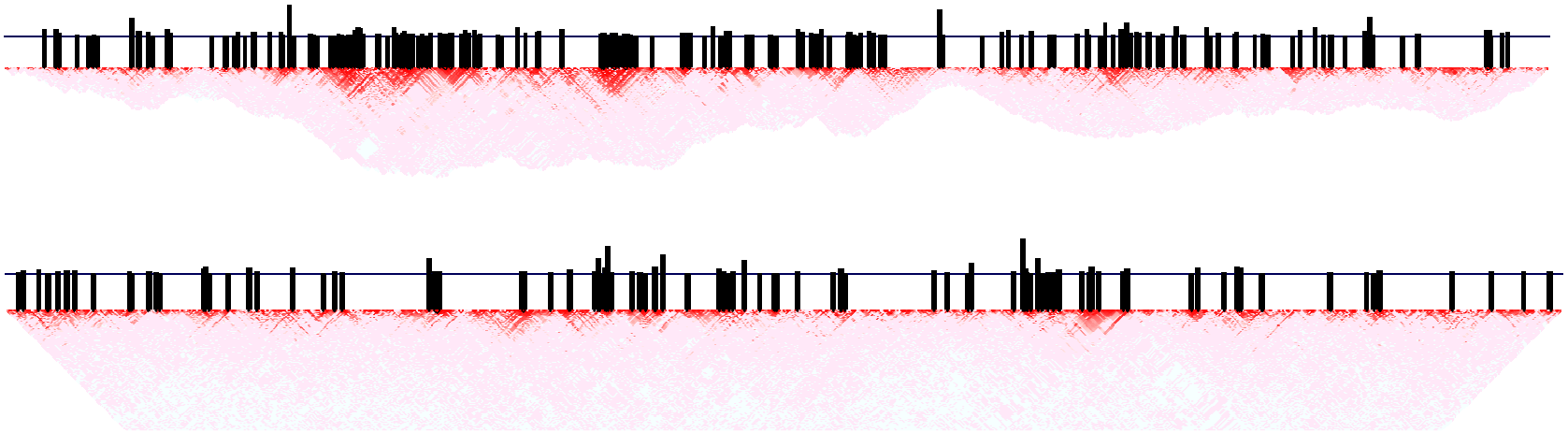

Figure 4. Linkage disequilibrium (D′) of chromosome 1 (top) and chromosome 18 (bottom) simulated using genomeSIMLA. Haploview-style correlation plots illustrate the LD structure (in D′). Each black line above the correlation plot indicates a haplotype block generated by the simulation, and the height of the bar above the horizontal line indicates SNP density.

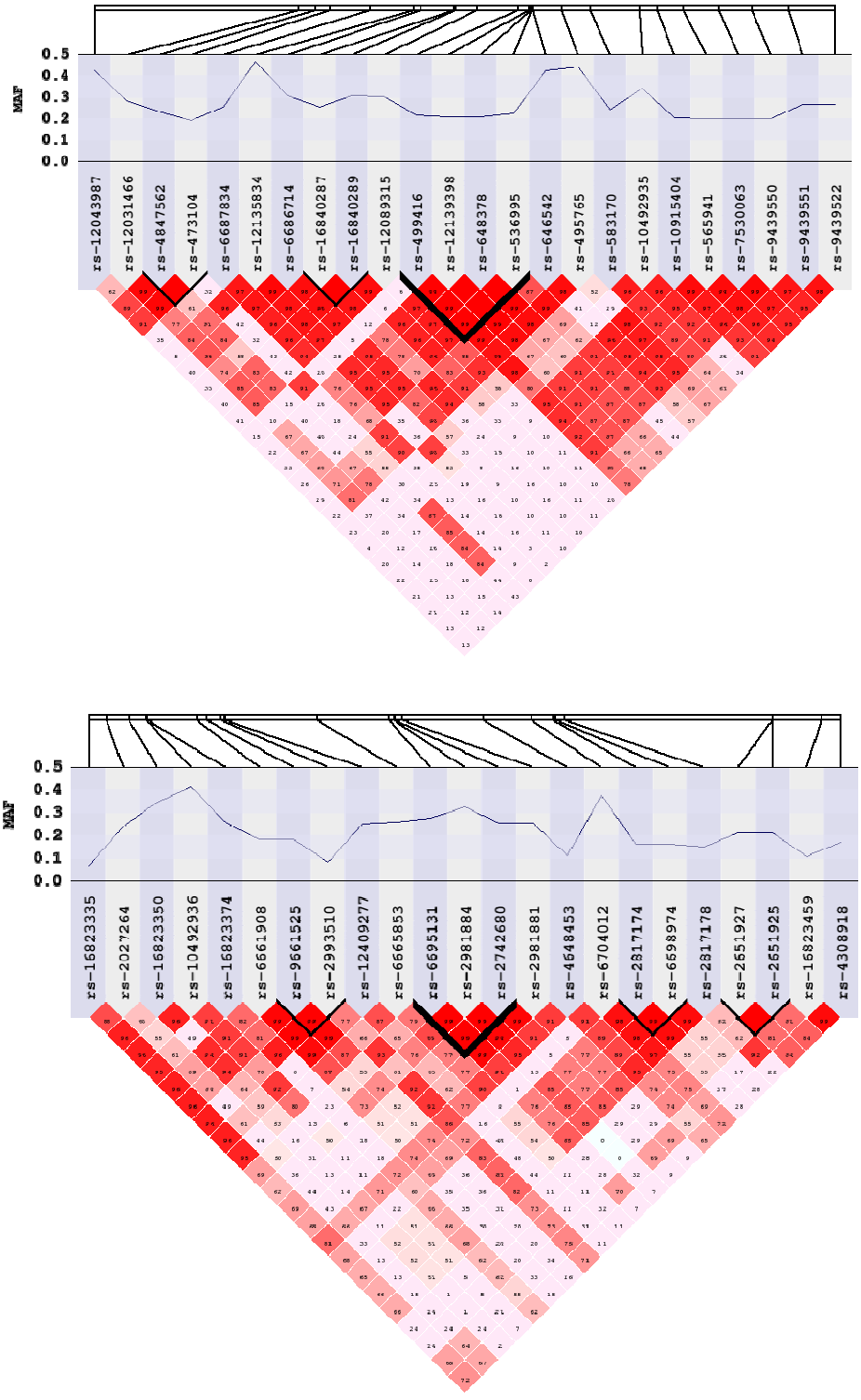

Figure 5. Regional haplotype structure for simulated block 7 (top) and 5 (bottom) on chromosome 1. The physical location and minor allele frequency of each simulated SNP is shown on the tracks along the top of the figure, and LD structure in D′ is shown in a Haploview-style correlation plot at the bottom. True haplotype blocks in the population are marked with dark lines in the correlation plot.

Algorithm agreement

Haplotype block partitioning of 100 datasets from the simulated region of chromosome 1 are shown in figures 6-10 and chromosome 18 in figures 11-15. Each horizontal line on these figures represents a called haplotype block, with the x-axis representing the index position of the SNP and the y-axis denoting the dataset for which the block partition was called. The ten gray vertical lines represent the true haplotype blocks simulated in the data (indexed across the top of the figure).

For chromosome 1, note the differences for blocks 3 and 4. The four gamete rule (figure 6) and Gabriel et al. (figure 7) call these two blocks as one larger block, and the four gamete rule seems more prone to produce a truncated block that does not include both the simulated blocks. LD-Spine (figure 8) does a better job of separating these two blocks, but is more likely to combine blocks 5 and 6 than Gabriel et al. and the four gamete rule. For chromosome 18, the general block calling from the four gamete rule (figure 11) and Gabriel et al. (figure 12) is sparse across datasets, indicating that for this particular simulated chromosome, sampling variability between datasets reduces the ability to find blocks consistently.

Weighted Kappa statistics for inter-rater agreement were calculated pair-wise to compare all algorithms to each other and to the true simulated block bounds. Results for chromosome 1 and chromosome 18 are shown in table 3. All algorithms had statistically significant agreement with each other and with true bounds by z-test (Cohen, 1968). The four-gamete rule performed best, with a weighted kappa near 0.95 for both simulations. The Gabriel et al. approach performed nearly as well. Of the three D' thresholds evaluated in this simulation (1, 0.8, and 0.6), using a threshold of 1 best matched the two established algorithms and the true block bounds in both simulated chromosomes (figures 8-10 and figure 13-15). While the LD-Spline approach does not outperform either of the established algorithms, it performs nearly as well, and still shows excellent agreement with true block bounds.

Table 3. Weighted kappa statistics for algorithm agreement.

Chromosome 1

	Four Gamete Rule	Gabriel et al.	LD-Spline 0.6	LD-Spline 0.8	LD-Spline 1
True Bounds	0.9512	0.9514	0.9092	0.9089	0.9383
Four Gamete Rule		0.9762	0.9123	0.9163	0.9498
Gabriel et al.			0.8931	0.9054	0.9412
LD-Spline 0.6				0.9681	0.9335
LD-Spline 0.8					0.9479

Chromosome 18

	Four Gamete Rule	Gabriel et al.	LD-Spline 0.6	LD-Spline 0.8	LD-Spline 1
True Bounds	0.9566	0.9271	0.9377	0.9153	0.9374
Four Gamete Rule		0.9740	0.9400	0.9379	0.9495
Gabriel et al.			0.9226	0.9208	0.9292
LD-Spline 0.6				0.9864	0.9635
LD-Spline 0.8					0.9671

Mapped block size distributions and captured genes

The LD-Spline algorithm using a D′ threshold of 1 was found to best recapitulate true haplotype block boundaries and best matches established algorithm block calls. We used these parameters and executed the LD-Spline procedure on two common GWA genotyping platforms, the Affymetrix Genome-Wide SNP Array 6.0 and the Illumina Human1M-Duo BeadChip. Block boundaries were mapped to NCBI genome build 36 using the Ensembl database (Hubbard et al., 2007). Frequency histograms of haplotype block sizes marked by each genotyping platform are shown in figure 16.

In each histogram, there is a notable increase in density near the 500 KB size. This is a procedural artifact due to computation restrictions used by the Hapmap during evaluation of LD statistics. Pair-wise LD is only calculated within a 500 KB window, and across the genome there are rare occurrences where LD extends beyond 500 KB – so the theoretical density of block sizes across the genome asymptotically approaches zero as block size increases. Therefore, density corresponding to block sizes larger than 500 KB is shifted into the 500 KB bin of the histogram.

The average block size captured by the Affymetrix 6.0 is 43 KB, and the average block size captured by the Illumina Human1M-Due is 38 KB. To quantify the number of genes captured by each platform, we used the Ensembl database to identify gene regions (defined as the start of the 5′ UTR to the end of the 3′ UTR), and to determine if SNPs lie within this region. Using this process, 17,418 genes were captured by the Affymetrix 6.0 platform, and 21,024 genes were captured by the Illumina Human 1M platform. Using the marked genomic regions generated by LD-Spline (using a D′ threshold of 1), we declare a gene "captured" if the marked region starts, ends or lies completely within the genic region, or alternatively, if the marked region completely encompasses the gene region. Using LD-Spline, the Affymetrix 6.0 captures 29,421 genes and the Illumina Human 1M captures 29,611 genes.

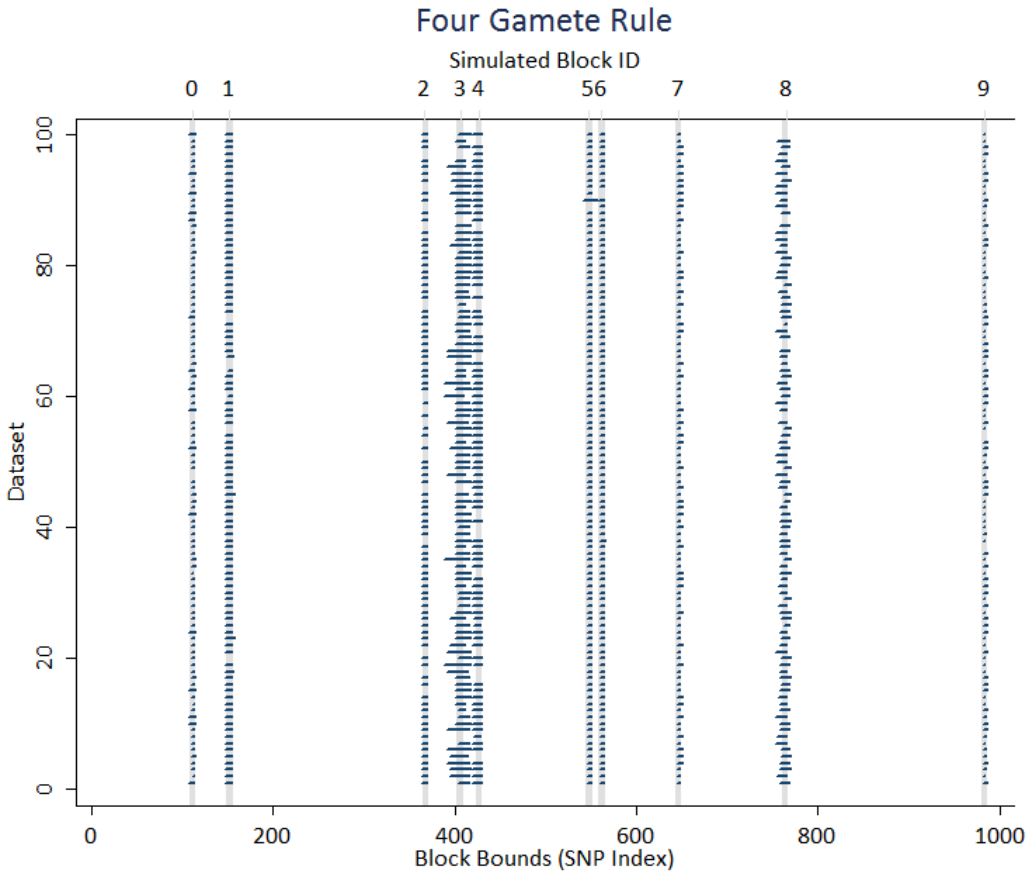

Figure 6. Four gamete rule haplotype block partitioning for simulated chromosome 1. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the four gamete rule, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

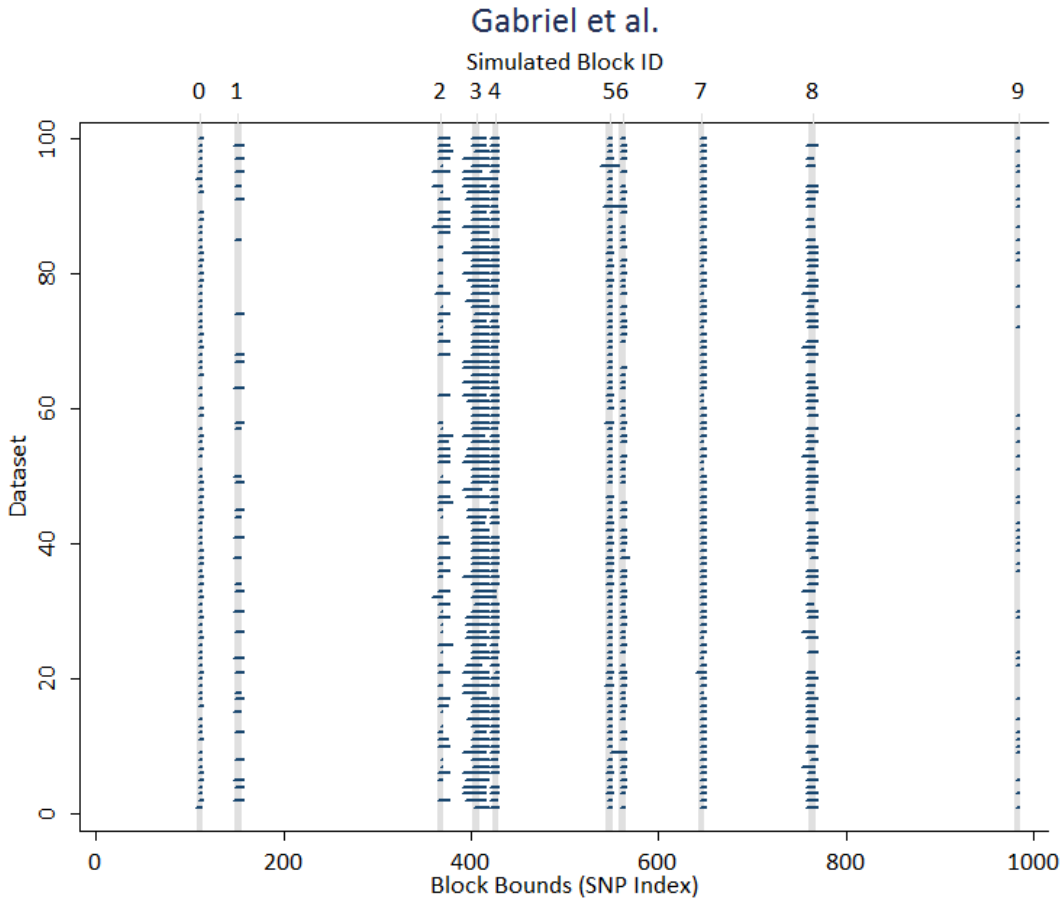

Figure 7. Gabriel et al. haplotype block partitioning for simulated chromosome 1. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the Gabriel et al. approach, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

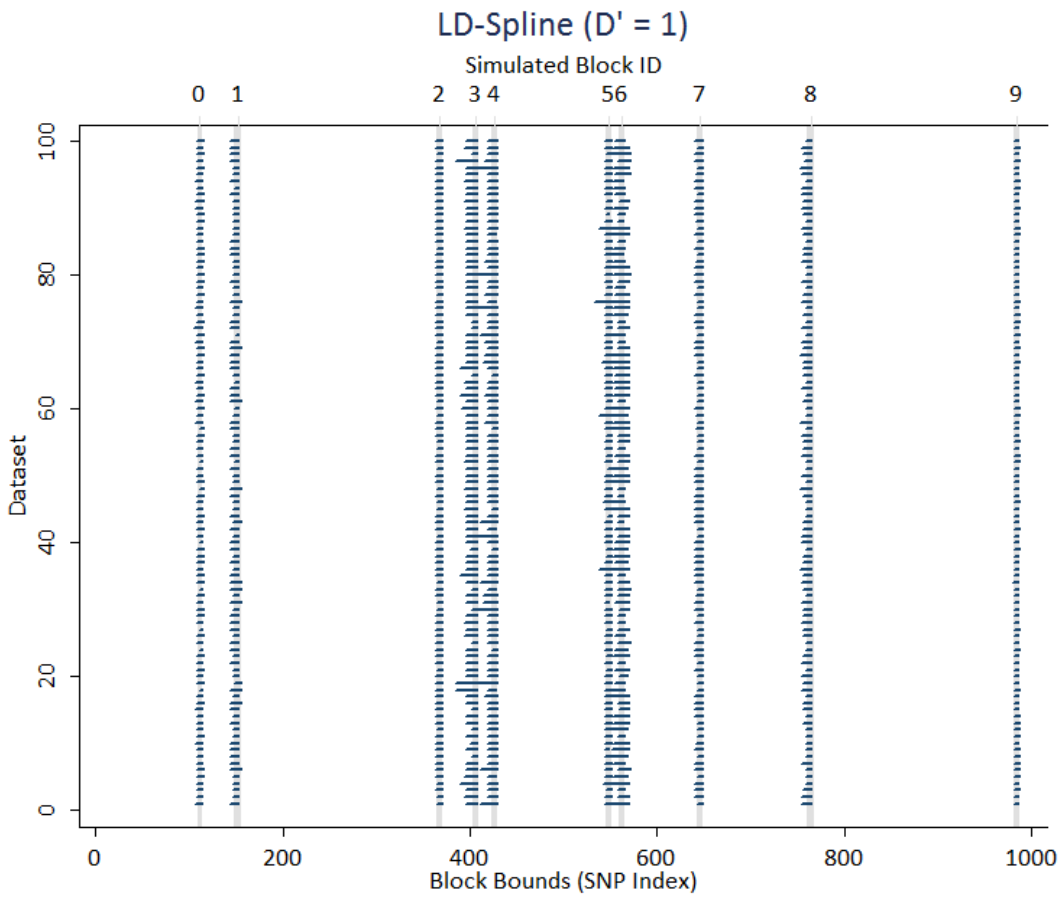

Figure 8. LD-Spline haplotype block partitioning for simulated chromosome 1. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the LD-Spline algorithm using a D' statistic of 1.0, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

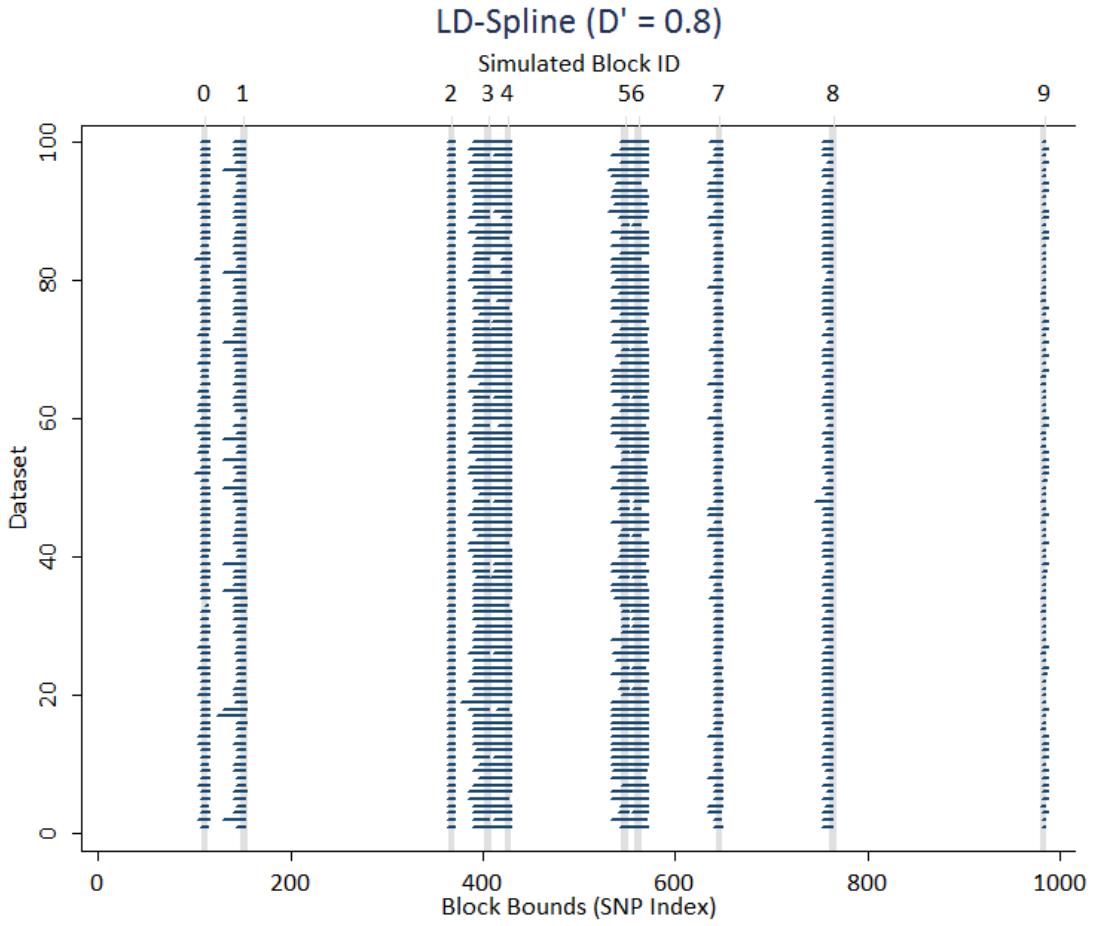

Figure 9. LD-Spline haplotype block partitioning for simulated chromosome 1. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the LD-Spline algorithm using a D′ statistic of 0.8, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

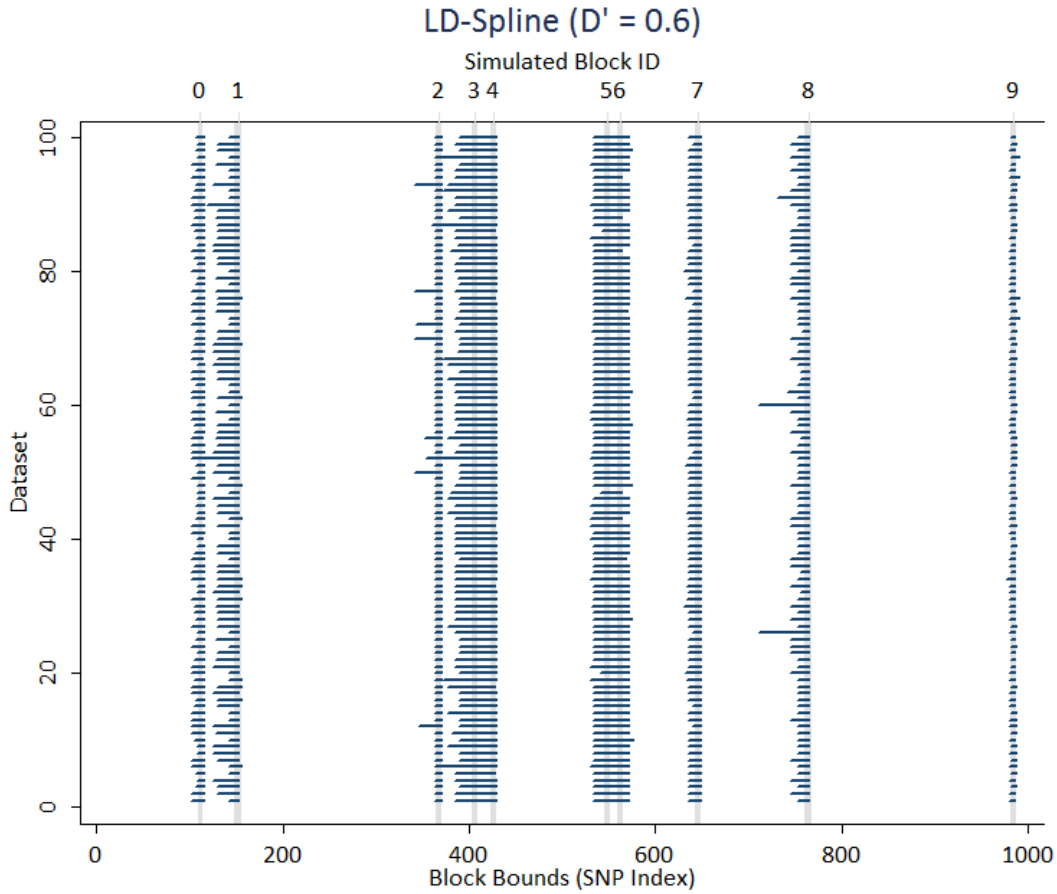

Figure 10. LD-Spline haplotype block partitioning for simulated chromosome 1. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the LD-Spline algorithm using a D′ statistic of 0.6, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

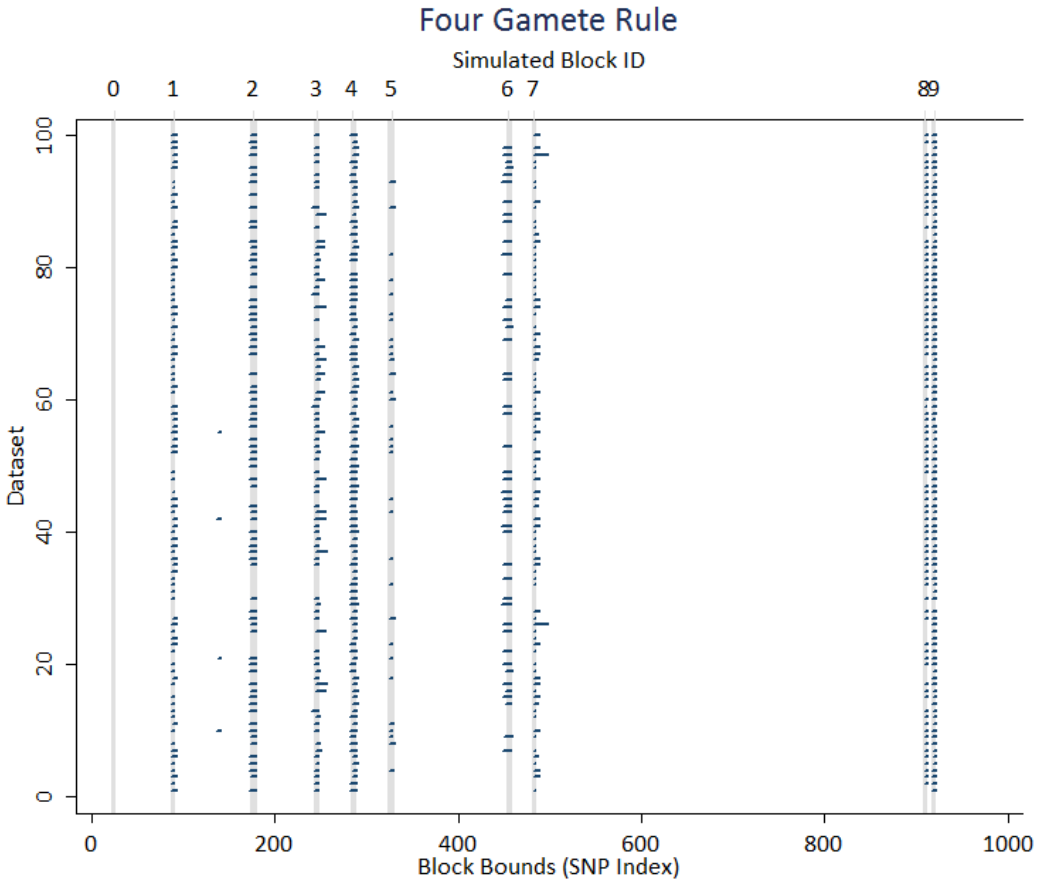

Figure 11. Four gamete rule haplotype block partitioning for simulated chromosome 18. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the four gamete rule with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

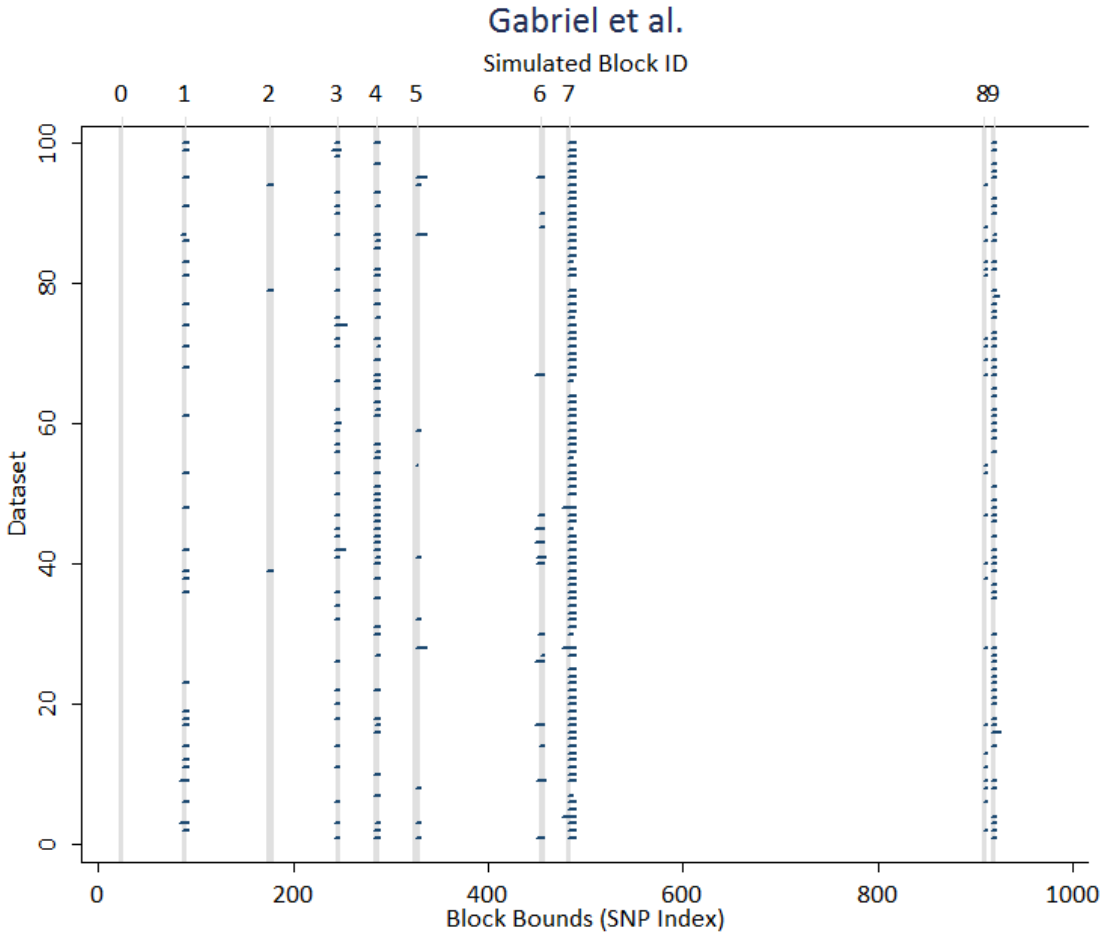

Figure 12. Gabriel et al. haplotype block partitioning for simulated chromosome 18. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the Gabriel et al. approach with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

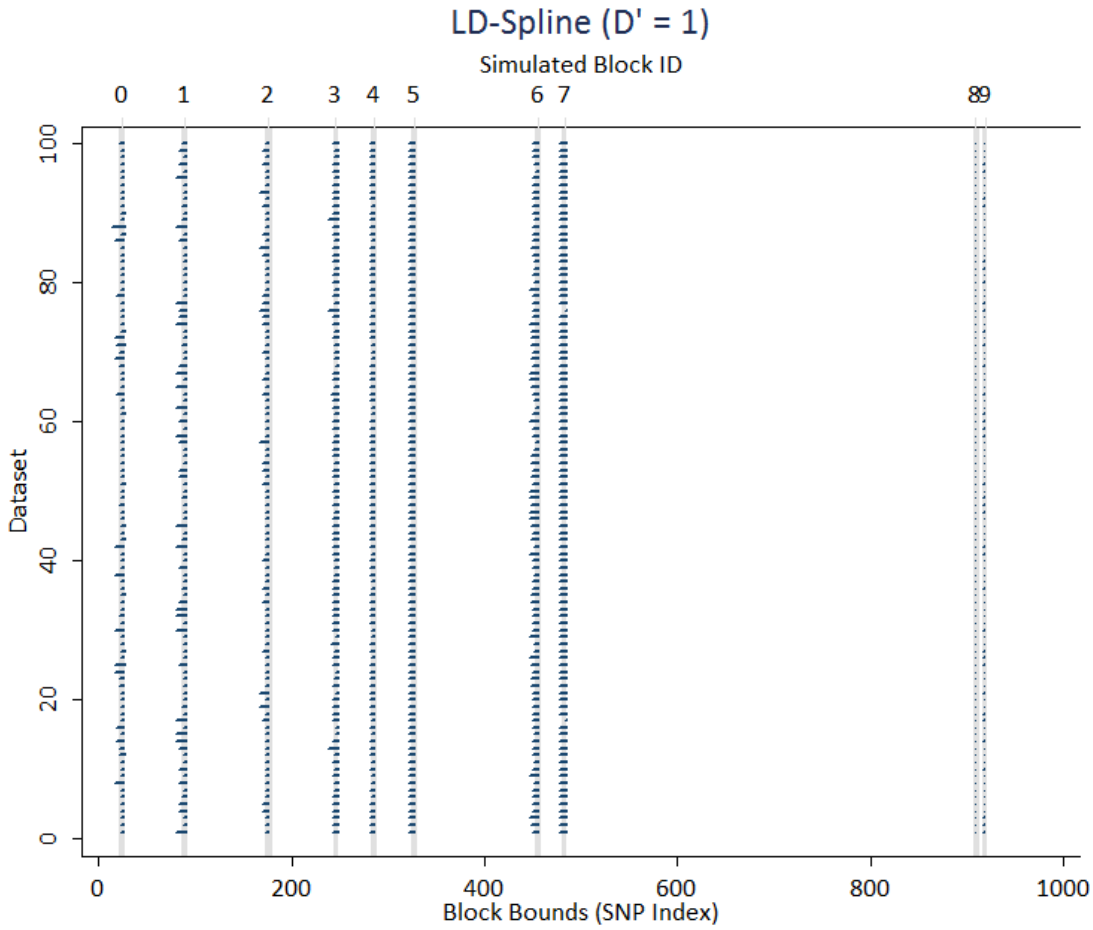

Figure 13. LD-Spline haplotype block partitioning for simulated chromosome 18. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the LD-Spline algorithm using a D′ statistic of 1.0, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

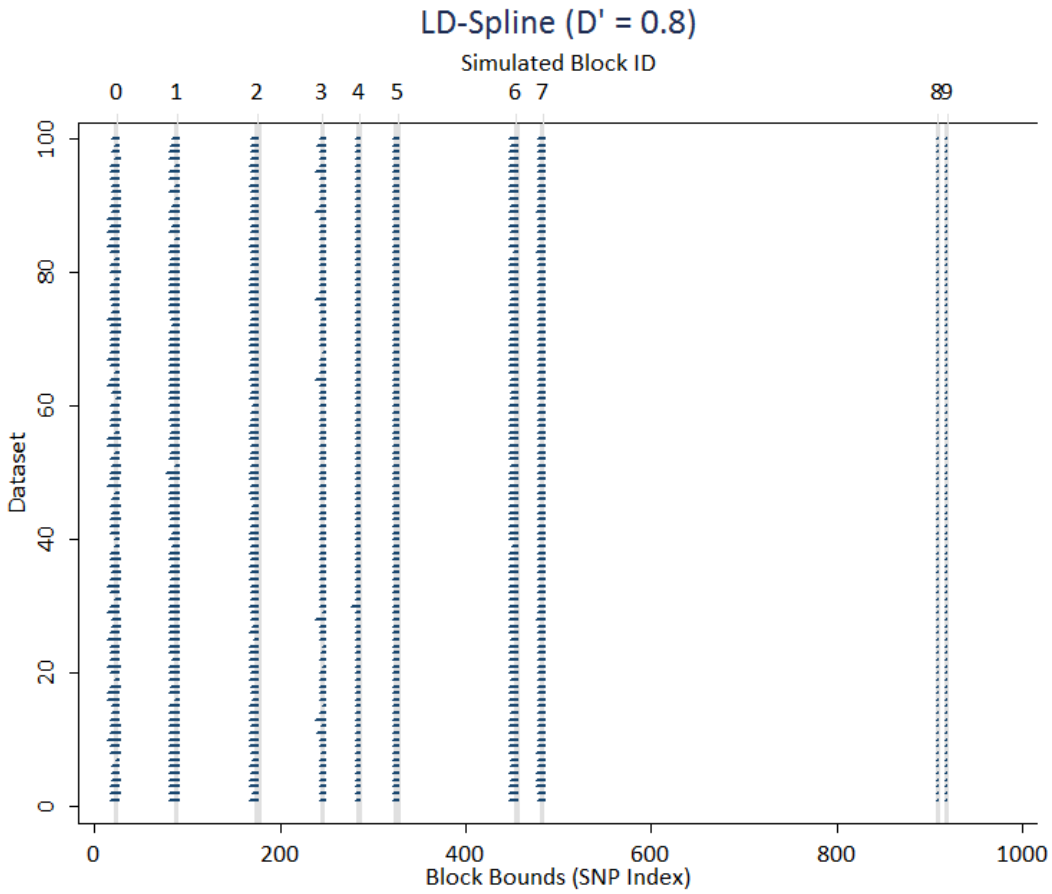

Figure 14. LD-Spline haplotype block partitioning for simulated chromosome 18. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the LD-Spline algorithm using a D' statistic of 0.8, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

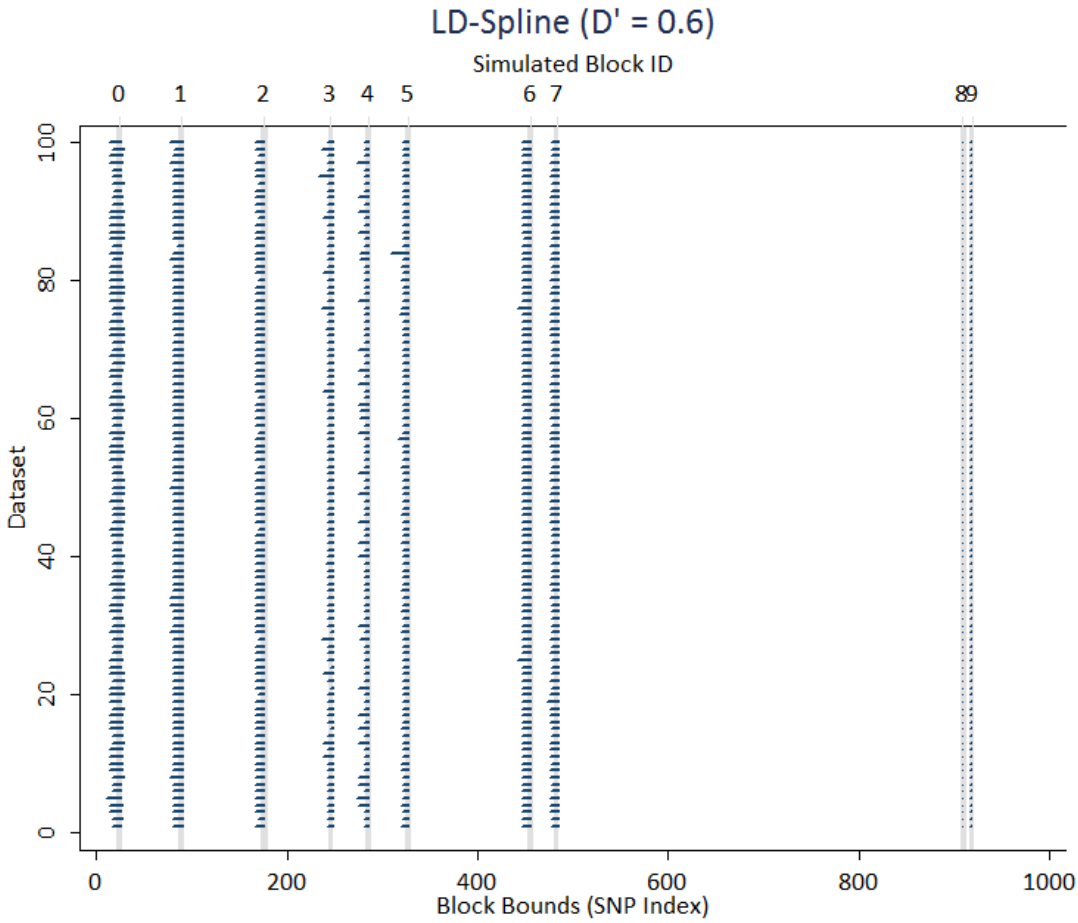

Figure 15. LD-Spline haplotype block partitioning for simulated chromosome 18. Ten haplotype blocks were selected from the simulation for algorithm assessment. Blocks are identified by an integer ID shown across the top of the figure, indicating the position within the 1000 SNPs simulated. The true bounds for each block are shown as gray vertical lines, with the thickness of the line indicating the block size. Each horizontal line represents a haplotype block called by the LD-Spline algorithm using a D′ statistic of 1.0, with the length of the line representing the number of SNPs included in the haplotype block call. The x-axis illustrates the upper and lower SNP index in the dataset for each block, and the y-axis indicates the dataset for which each block is called.

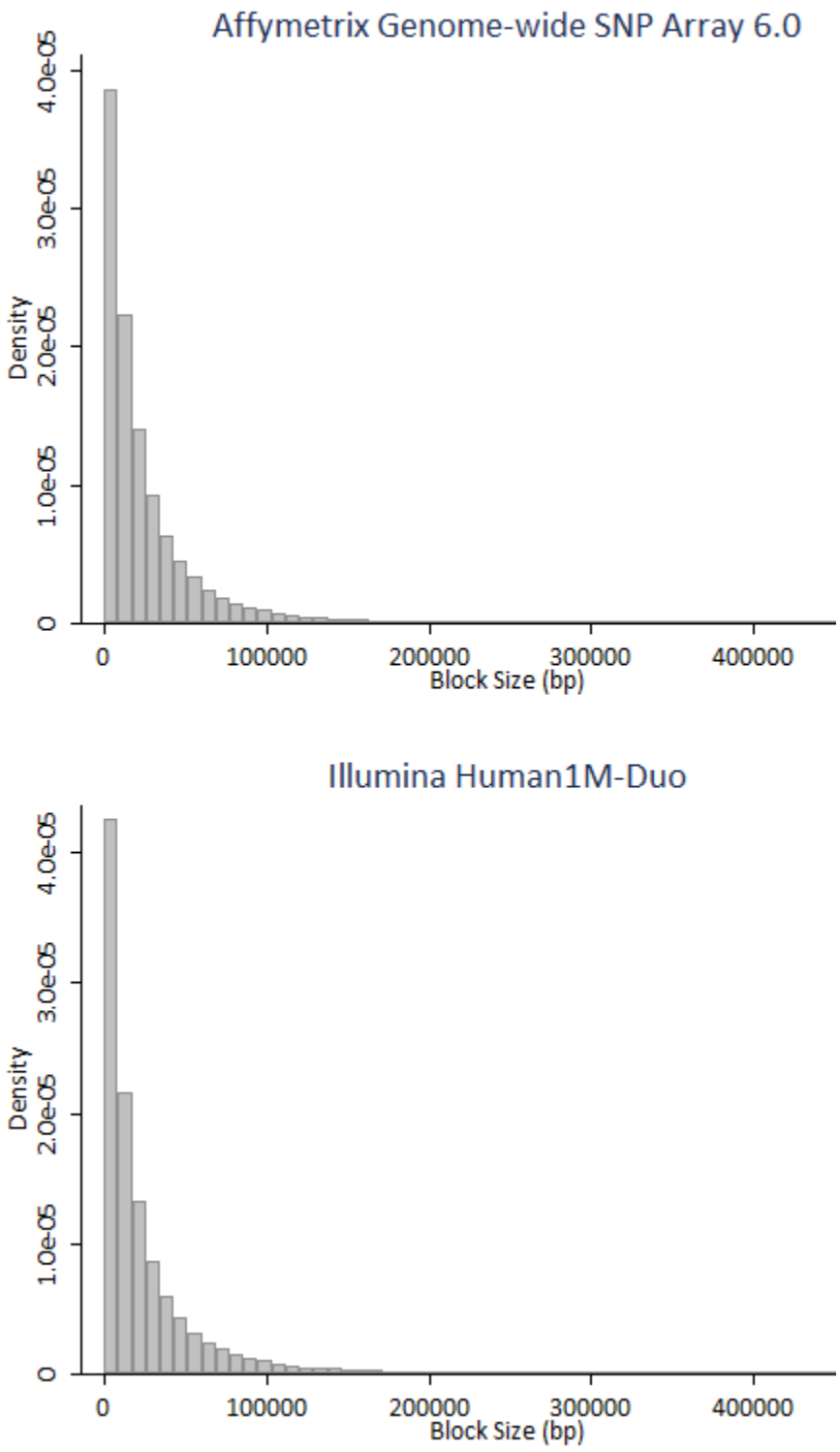

Figure 16. Frequency histogram of LD-Spline called haplotype block sizes. The Affymetrix Genome-wide SNP Array 6.0 (top) and the Illumina Human 1M -Duo (bottom) genotyping platforms are shown.

Conclusions

In this work, we introduce LD-Spline, an efficient database procedure for establishing genomic regions that a SNP potentially represents by mining linkage disequilibrium statistics available from the International Hapmap Project. The two established block-calling algorithms (Gabriel et al. and four gamete rule) function by producing a global haplotype block partitioning, starting at the first SNP and sequentially defining blocks upstream. The LD-Spline approach is SNP-centric, in that it uses LD statistics between a user-provided SNP (such as one from a genotyping platform) and surrounding SNPs in the genome to define the region the specified SNP marks. This SNP-centric approach also has a computational advantage, since only relevant haplotype blocks (the region surrounding SNPs of interest) are called by the algorithm. Gabriel et al. and the four gamete rule would require processing and partitioning the entire human genome to determine the regions marked by a genotyping platform. LD-Spline also has the great advantage of running as a fast and efficient self-contained procedure within the database management system, allowing seamless integration with existing database queries and operations

We compared the LD-Spline algorithm to the Gabriel et al. and four gamete rule methods, and compared all methods to the true simulated haplotype block boundaries. Weighted kappa agreement statistics between LD-Spline, traditional block calling algorithms, and the true block partition in simulated data were rather good (> 0.90 in most cases). While none of the block partitioning algorithms perfectly identify true block boundaries, the LD-Spline approach using a D′ threshold of 1 appears to work as well as other established algorithms.

As a SNP-centric approach, LD-Spline has the added advantage of consistently marking a genomic region for each SNP. The sequential partitioning achieved by the Gabriel et al. and four gamete rule approaches do not consistently identify a haplotype block for each dataset. For example with chromosome 18, the four gamete rule and Gabriel et al. did not define haplotypes for simulated blocks 0, 2, 5, or 6. For the specific application of determining what genomic region a typed SNP likely represents, a SNP-centric approach is advantageous, as the long-range LD patterns specifically related to the typed

SNP are exploited. Sequential partitioning approaches generally use a two-SNP sliding window to define haplotype blocks, and as such are not robust to situations where short-range LD is weaker than long-range LD. It is important to note that the weighted Kappa statistics for algorithm agreement do not take into account the number of uncalled haplotype blocks, but do indicate that the boundaries for the called haplotype blocks are similar. With this in mind, LD-Spline provides superior performance when assigning genomic regions to typed SNPs because of its SNP-centric nature.

Another explanation for the lack of block identification for chromosome 18 and the general small degree of disagreement with the true block boundaries is sampling variability. In our data simulations, we empirically track recombination events to produce exact LD statistics and LD block boundaries on the population level. Each simulated dataset was drawn from that population, and sampling variability could lead to biased LD estimates and subsequent block partitions. Also, to more closely mimic real data collection in the Hapmap project, datasets were produced as unphased genotype data. We then used Haploview software to estimate two-marker haplotypes using the EM algorithm to calculate D′ and r^2 LD statistics. This procedure could also introduce bias and error into the haplotype block calling procedures.

When applied to GWA genotyping platforms, block sizes follow the pattern expected based on previous estimates of block size by the Hapmap project (International HapMap Consortium, 2005). The average block size differs slightly between platforms. This could be because of bias in SNP selection by the genotyping platform manufacturers, particularly Illumina (Eberle et al., 2007). If SNPs are specifically selected that tag larger genomic regions, and avoided SNPs in regions of sparse LD, this could inflate the average block size. Also, if SNPs in genic regions are overrepresented by genotyping platforms, this could also cause inflation, as r^2 measures of LD have been found to be higher in genic versus inter-genic regions (Eberle et al., 2006).

Overall, we have illustrated the performance of the LD-Spline routine, and the utility of applying this database-centric procedure to GWA platforms. One key advantage of the database-centric nature of the LD-Spline user-defined function is that it can easily be incorporated into more sophisticated queries

for information retrieval. Once established, the database routine can seamlessly extend the range of data queries to include statistics based on a broader genomic region, rather than a single base-pair location.

Acknowledgements

Eric Torstenson implemented and tested the LD-Spline user-defined function for the MySQL database system. Portions of the simulation study were conducted by Guanhua Chen. Our thanks to the International HapMap Project for making population-based collections of LD statistics freely available.

CHAPTER III

INTEGRATING BIOLOGICAL KNOWLEDGE INTO DATA ANALYSIS FOR GENOME-WIDE ASSOCIATION STUDIES[1]

Introduction

Genome-wide association (GWA) studies

Over the last five years, genome-wide association (GWA) has become a very popular study design for identifying genetic variants that incur disease risk in human populations. As described in Chapter I, the overall strategy of the GWA approach is inherently high-throughput, allowing investigators to blanket the genome with hundreds of thousands of single nucleotide polymorphisms (SNPs) in many individuals with the general goal of elucidating genetic causes of common human phenotypes – complex diseases in particular. The development of methods to effectively analyze the wealth of information produced from these studies has not kept pace with the technological advances that produce the data. Using mostly basic analysis techniques, GWA studies have produced many novel genetic associations to multiple phenotypes, but in general these findings explain only a small portion of the overall genetic risk for those phenotypes. Because the common diseases studied with the GWA approach presumably involve multiple interacting genetic and environmental factors, basic single-SNP analyses have certainly missed important genetic effects. More complex analysis techniques are challenging to apply to GWA data (see Chapter I), and incorporating prior biological knowledge into the analysis is one approach to reduce the subsequent computational and statistical burden. Also, placing SNPs and their related genes into their larger genomic and functional context will be key in fully understanding and interpreting GWA findings.

[1] Adapted from Bush W.S., Dudek S.M., Ritchie M.D. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Proceedings of the 2009 Pacific Symposium on Biocomputing.* Accepted September 2008.

The utility of prior knowledge

The last 15 years of biological science have been aided dramatically by the advent and introduction of internet-based technologies (such as genome browsers) and the databases of information that drive them. Rather than relying solely on the publication system, investigators and publishers began to deposit scientific findings about protein structure, biochemical systems, and gene regulatory networks (among many others) into a collection of highly structured and cross-referenced database systems. These interoperable and highly accessible systems allow easy search and comparison of many biologically relevant information types, and the information contained in these databases is just now beginning to be exploited for interpreting and processing GWA results. As discussed in Chapter I, epistasis is likely to play a role in the complex diseases studied with the GWA approach. The statistical and biological interaction of alleles and genotypes has been observed in several biochemical systems and pathways, and is presumed to be a ubiquitous component of variability for a broad range of phenotypes (Moore, 2003). As such, a logical application of established biological knowledge in GWA analysis is in the search for epistatic interactions.

Many types of structured biological knowledge could prove useful for identifying epistasis. For example, protein-protein interaction is suggestive of potential gene-gene interaction. In tuberous sclerosis, both *TSC1* and *TSC2* bind to form a protein complex that functions in a tumor suppression pathway (Huang & Manning, 2008). Mutations in both *TSC1* and *TSC2* can disrupt this binding, abolishing the function of the complex, ultimately leading to the scattered and widespread formation of tumerous nodules. While in this case the two mutations have strong independent effects, it is logical to assume that subtle changes in *TSC1* or *TSC2* expression or structure could alter binding or other functional properties to modestly modulate risk for some cancer phenotypes. Similarly, pathway information can have great utility for identifying epistasis and elucidating functional causes of disease. Nephrogenic diabetes insipidus (NDI) can be caused by numerous mutations throughout the vasopressin type 2 receptor activated, calcium-signaled insertion of the aquaporin-2 channel into the apical membrane

of the collecting ducts of the kidneys, and results in the inability to concentrate urine (Spanakis, Milord, & Gragnoli, 2008). Again, strong mutations in this pathway abolish function, but more subtle variations could alter function, increasing risk for conditions related to blood-fluid homeostasis, such as essential hypertension. Protein structural and sequence information could also be used to identify common functional domains that may in concert be relevant to a disease process. For example, many downstream pathways are triggered from the cellular surface via G-protein-coupled receptors (GPCRs). These receptors are often tissue-specific and can have some degree of redundancy, where multiple different receptors that bind the same ligand trigger the same physiological process (Amisten et al., 2008). GPCRs also account for more than 30% of all pharmacological targets (Wise, Gearing, & Rees, 2002), and are thus likely candidates for pharmacogenomic phenotypes. It is therefore plausible that variants in multiple similar GPCR genes could aggregate to dramatically influence risk beyond the additive effects of any single variation. High resolution structural information of this important class of proteins will help elucidate ways that GPCR functional variation can impact downstream signaling (Tian et al., 2005) and other disease phenotypes may be related to subtle changes in protein structure (Myers, Beihoffer, & Sanders, 2005). Ultimately, protein structural and sequence information will be incorporated with pathway information to build sophisticated kinetic models predicting functional consequences of amino acid sequence substitutions and/or gene expression changes on a systems biology level (Beltrao, Kiel, & Serrano, 2007). This modest set of examples only begins to illustrate the innumerable data sources that can be applied to multivariate genetic analysis of disease phenotypes.

Several new tools have recently been developed to incorporate biological information with analytical approaches for GWA data. Prioritizer is a Bayesian approach to synthesize multiple sources of gene interrelationships into a global "functional gene network". This network can be used to prioritize significant single-SNP results by gene function (Franke et al., 2006). For a sequence-oriented approach, PROSPECTR is a machine learning method that prioritizes genes in candidate regions based on sequence features, such as transmembrane regions, GC content, number of exons, and homology to other organisms, and could be applied to GWA data as well (Adie et al., 2005).

Some methods use structured knowledge as a way to guide (but not restrict) variable selection for regression-based modeling. Province and Borecki propose a Bayesian re-sampling approach to select collections of SNPs that may have very small independent effects but function in aggregate to explain a more substantial portion of trait variance (Province & Borecki, 2008). Lewinger et al. proposes a hierarchical modeling approach that uses an expert knowledge ontology to search for and test complex multi-SNP models. This Bayesian modeling process is flexible, allowing SNPs outside the knowledge-base to also be used in models (Lewinger et al., 2007).

Additionally, some approaches annotate or weight single-SNP results. Gene Set Enrichment Analysis (GSEA) has been used extensively for gene expression data, and has been modified to investigate enrichment of gene categories in significant GWA SNP associations (Wang, Li, & Bucan, 2007). Curtis et al. describe a method for weighting SNP p-values based on prior candidate information, such as previous associations or linkage peaks (Curtis, Vine, & Knight, 2007), and similarly Li et al. illustrates power improvement when assessing false discovery rate (FDR) separately for prioritized candidate SNPs versus non-candidate SNPs (Li et al., 2008). Finally, Pan proposes weighting the p-values of multi-locus models based on putative protein-protein interactions between genes included in the model (Pan, 2008).

We propose a strategy that steps beyond the annotation, grouping, and weighting of independent SNP effects, but does not attempt to jointly model large numbers of SNPs simultaneously. Also, we believe that ultimately data from multiple sources will better facilitate a comprehensive analysis, providing a biological foundation for testing specific multi-SNP association models in GWA data. In this work, we present the Biofilter, a tool for knowledge-driven multi-SNP analysis of large scale SNP data. The Biofilter fundamentally differs from other methods in the way knowledge is incorporated into the analysis pipeline. The Biofilter uses biological information about gene-gene relationships and gene-disease relationships to construct multi-SNP models before conducting any statistical analysis. Rather than annotating the independent effect of each SNP in a GWA dataset, the Biofilter allows the explicit detection and modeling of interactions between a set of SNPs. In this manner, the Biofilter process provides a tool to discover significant multi-SNP models (regardless of main effects) that have established

biological plausibility. This approach has the added benefit of reducing both the computational and statistical burden of exhaustively evaluating all possible multi-SNP models.

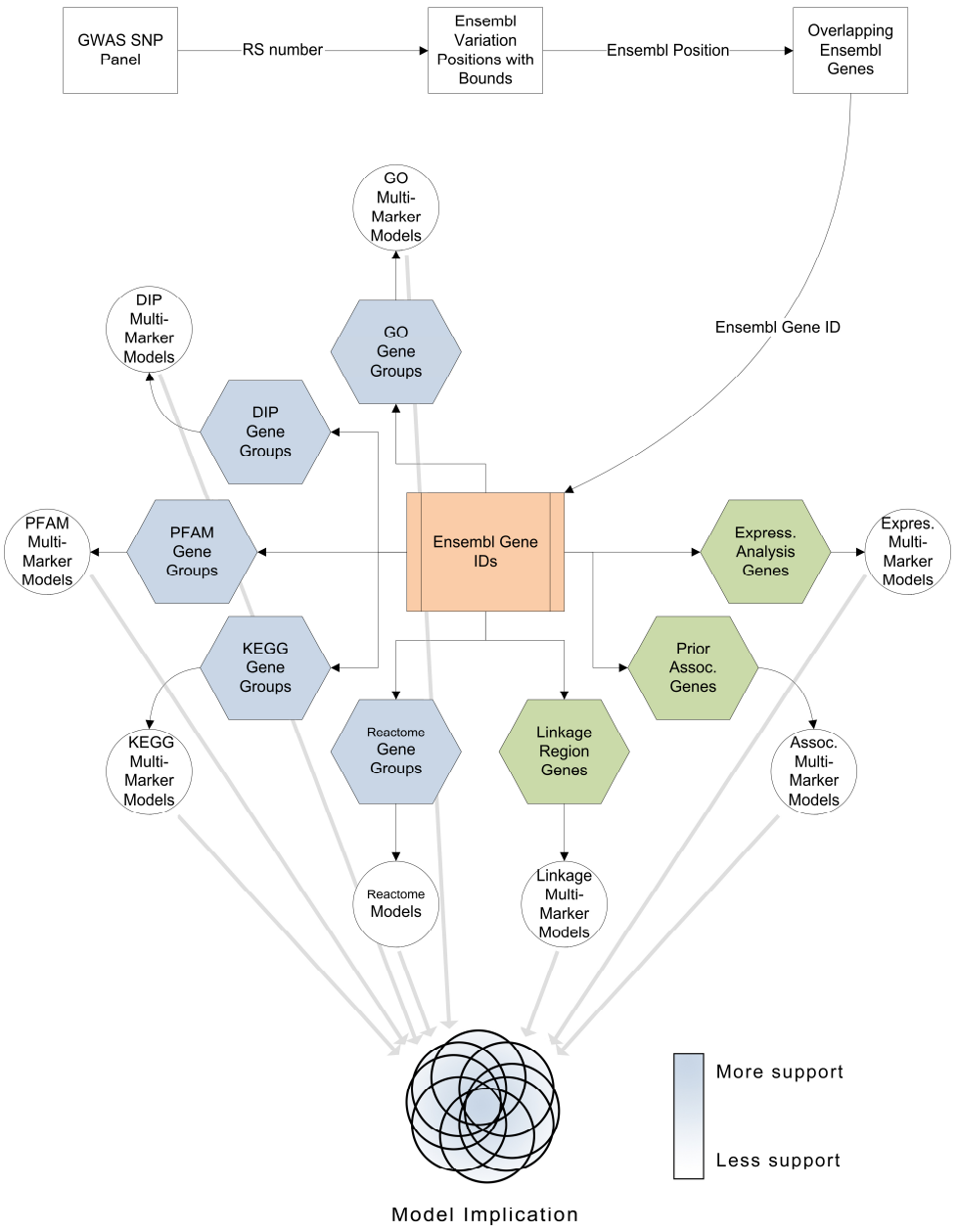

Figure 17. Overview of the Biofilter process. GWA platform SNPs are mapped to Ensembl gene IDs and related to disease-independent sources (left) and to disease-dependent sources (right). Multi-marker models are generated from SNPs within knowledge-related genes. Derived models are overlaid to assess overall model implication.

Methods

Overview

An overview of the Biofilter method is show in figure 17. The Biofilter model generation process is gene-centric, and as such, SNPs from GWA genotyping platforms must first be assigned to genes. SNPs can be evaluated to determine if they fall directly within gene boundaries using a genomic resource like Ensembl, or a mapping routine like the LD-Spline approach described in Chapter II can be applied to use linkage disequilibrium to better define SNP-gene representation. Relationships between the genes represented by a genotyping platform can then be translated to multi-SNP models. Structured biological knowledge relevant to GWA interaction analysis can come from various sources. We have partitioned relevant knowledge into two basic types: disease-dependent and disease-independent. *Disease-dependent* knowledge is information that relates a gene to the disease phenotype being studied, such as a previously associated SNP or a gene that is over-expressed in cases. *Disease-independent* knowledge is information that relates genes to one another, or defines collections of genes, such as a metabolic pathway or a common structural motif. These two types of information can be combined to form different classes of multi-SNP models and provide a measure of how strongly implicated a given model is based on the current available knowledge.

Database integration

GWA studies use a large collection of SNP probes to capture genetic variation in the study population. Often, vendors of GWA genotyping platforms provide some annotation of the probes used in their products, including SNP identifier, genomic position, and nearby genes. Documentation on how this annotation was generated, particularly what genomic build or version, is often sparse or incomplete, and so a consistent and updatable system for annotating GWA platforms was needed. We used Ensembl as our source of gene and SNP positional information.

Ensembl is a well-established, open database system providing extensive annotation of the human genome, from raw genomic sequence to cross-references for protein structure and functional biochemical systems (Hubbard et al., 2007). In addition to its extensive information store, Ensembl is easy to access and reconstitute, and its contents follow a clearly defined and documented database schema. The components of the Ensembl core database and the Ensembl variation database (Release 49) were established on a local machine. Using this local copy, RS numbers for SNPs used on genotyping platforms are matched to records within the Ensembl variation database to retrieve position information in the current genomic build. To establish SNP-to-gene relationships for each platform, we used the Ensembl database to identify gene regions (defined as the start of the 5′ UTR to the end of the 3′ UTR), and assigned SNPs from each genotyping platform to a gene if the SNP lies within a genic region in Ensembl. Using this process, 17,418 genes were captured by the Affymetrix 6.0 platform, and 21,024 genes were captured by the Illumina Human 1M platform. X and Y chromosomes were excluded from this analysis. With SNP-gene relationships established, we can now apply gene-gene information to build multi-SNP models.

Disease-dependent knowledge sources link individual genes to a disease phenotype. The goal of using disease-dependent knowledge is to identify genes that have some prior evidence of putative influence on the phenotype. One systematic source of disease-dependent knowledge is the Genetic Association Database (GAD). GAD is an archive of human genetic association studies of complex diseases and disorders established in 2004 by the National Institutes of Health (Becker et al., 2004). GAD has a hierarchical arrangement of disease phenotypes, with a top level "disease class", such as immune or psychiatric, and more narrow phenotype identifiers, such as asthma or schizophrenia. Each GAD entry is a polymorphism-phenotype association, annotated with the gene, p-value, chromosomal band, and author list and Pubmed ID of the reporting study. GAD is easily searchable to find prior associated genes for a given phenotype, and sets of these genes could be tested jointly in GWA data.

Other types of disease-dependent knowledge may require manual selection from literature. Published genome-wide linkage screens often report chromosomal bands which could be collected to

identify genes in regions of linkage. Likewise, gene expression studies can highlight groups of genes that are differentially regulated or active between cases and controls or discordant siblings. This collection of genes could highlight potential disease mechanisms, and could be evaluated in GWA data. Disease review articles often postulate hypothetical disease etiologies and may provide candidate gene lists that could be incorporated.

Disease-dependent knowledge leverages information collected by prior studies of the phenotype, and since inconsistent replication of previously associated or linked genes could be an indicator of multi-locus interactions, these sources have utility in constructing epistasis models in GWA data. Disease-dependent sources are susceptible to publication bias however, and applying these sources exclusively would bias the GWA analysis toward replicating known potential effects rather than exploring novel disease mechanisms.

Disease-independent sources link two or more genes together, irrespective of the phenotype. The goal of using disease-independent knowledge is to identify gene sets with some prior evidence of putative epistasis. The Gene Ontology project (GO, accessed on 3/16/08) is a collaborative effort to characterize and describe gene products in a collection of three hierarchical ontologies: cellular component, biological process, and molecular function. Cellular component categories describe the location of gene product activity within the cell, such as "ribosome" or "nuclear pore". Biological process categories describe a chemical or mechanical process, such as "flagellar cell motility" or "nucleic acid biosynthesis". Finally molecular function categories describe a molecule-specific activity, such as "DNA binding". Because of its hierarchical structure, some broad ontology categories contain many hundreds of genes. For this analysis, smaller, more precisely defined gene categories (< 30 genes) were used as these presumably contain stronger, more precise gene relationships. The working hypothesis when using Gene Ontology groups is that genes participating in a common function, common cellular component, or with a common molecular feature are more likely to contain epistatic alleles.

The Database of Interacting Proteins (DIP, 1/14/08 update) documents experimentally determined protein-protein interactions from more than 80 organisms (Xenarios et al., 2002). These data

were collected primarily from yeast-two-hybrid experiments, where the co-localization of tagged gene-products triggers a fluorescent molecule (Suter, Kittanakom, & Stagljar, 2008), and two proteins seen in close cellular proximity are said to "interact". The genes for the corresponding interacting proteins are then mapped to other organisms, including humans, using sequence homology. We used the pair-wise human protein-protein interaction set contained in DIP to produce gene-gene pairs, since cellular co-localization likely increases the probability of epistasis.

The Protein Families Database (PFAM, Release 22) uses multiple sequence alignments and hidden Markov models to identify common protein domains and families based on structural and functional sequence patterns (Finn et al., 2008). Generating pairs of genes using these data relies on the hypothesis that genetic variants in proteins with similar structural elements are more likely to interact (biologically or statistically) to influence disease risk. As such, we generated gene-gene pairs within proteins having the same domain, the same protein family, the same structural motif, or the same sequence repeat.

The Kyoto Encyclopedia of Genes and Genomes (KEGG, 3/6/08 update) pathway set is a collection of manually drawn maps for a variety of metabolic and signaling pathways (Kanehisa et al., 2008). These pathways are well established in the literature, and contain links to original publications proposing the pathway structure. KEGG loosely defines a pathway as a simple collection of genes, with no electronically stored information about metabolic or physical relationships between genes. Reactome (Version 24) is also a pathway database, containing curated core pathways and reactions in human biology (Vastrik et al., 2007). A fundamental difference between Reactome and KEGG is that Reactome electronically stores the direct inter-relationships between genes in the pathway, rather than storing a simple gene set. Netpath is a relatively new source of curated immune signaling and cancer pathways provided by the Pandey Lab at Johns Hopkins University and the Institute of Bioinformatics (Pandey & Institute of Bioinformatics, 2008). This is a simple set of gene groups based on literature and experimental-based pathway information, but with exclusive emphasis on immune and cancer pathways.

Supporting gene expression and co-regulation data are also provided. With these pathway collections, all possible gene-gene pairs were generated within each pathway-based gene group.

Relational data sources were downloaded and reconstituted in their original form within a MySQL database using Perl scripts. Using the schema for each data source, proteins and/or genes were translated to Ensembl gene IDs, and derivative tables containing gene groupings (such as protein families) were generated within the Biofilter database. Non-structured data sources, such as gene lists from publications, were manually imported. These gene lists were then translated to Ensembl gene IDs and used to establish gene groupings.

Model types and generation

Using both disease-dependent and disease-independent data sources, there are four types of two-SNP models possible: disease-independent, disease-dependent, hybrid with one disease-dependent gene, and hybrid with two disease-dependent genes. Figure 18 illustrates each of these model types. Disease-dependent information is based on a set of genes that are related to disease, visualized in the figure as a collection of dashed boxes, or unconnected nodes of a graph. Disease-dependent models are generated by exhaustively pairing all possible combinations of disease-related genes. Disease-independent information is based on relationships between sets of genes, visualized in the figure as a set of lines, or edges in a graph. To build disease-independent models, we generate pair-wise combinations of SNPs located in genes that are related, illustrated as edges in figure 18. Hybrid models blend disease-dependent and disease-independent information, and can contain either one or two disease-related genes. In the figure, one-gene hybrids must be connected by an edge (disease-independent connection) and contain at least one dashed square (disease-dependent gene). Two-gene hybrids must contain two dashed squares connected by an edge, meaning that there is evidence for biological interaction of two disease-related genes.

An illustration of two-SNP model generation for the Urea Cycle Gene Ontology category is shown in figure 19. For each pair-wise combination of genes, all possible two-SNP models across the two

genes are built. In the urea cycle category of the Gene Ontology, there are two SNPs (rs160648 and rs313830) in the *ASL* gene and three SNPs (rs3770684, rs16844641, and rs6714124) in the *CPS1* gene. Six models are generated by pairing SNPs across genes, with each SNP from *ASL* paired with each SNP from *CPS1*.

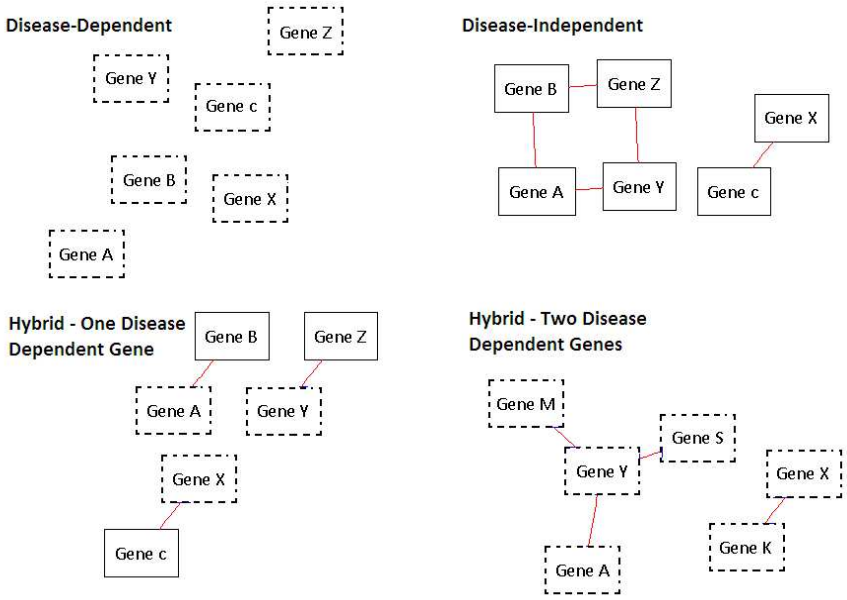

Figure 18. Biofilter two-gene model types. Each box represents a gene, and each line a connection between genes. Boxes that are dashed have been previously linked to disease by at least one data source.

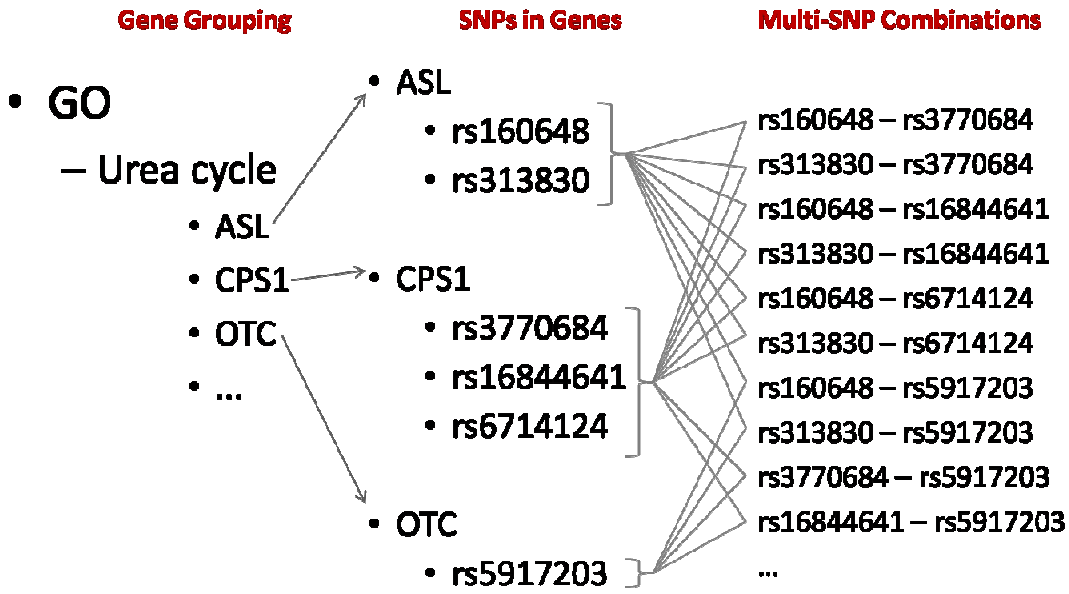

Figure 19. Biofilter two-SNP model generation process. A gene grouping (urea cycle) is selected from a data source (Gene Ontology), and SNPs from a genotyping platform mapped to those genes are retrieved. SNPs are paired into models *across* genes (one SNP from one gene paired with one SNP from another).

Model implication

Each constructed model has a set of Biofilter data sources that support it. If a combination of genes is supported by multiple data sources, it is likely more accepted by the scientific community and therefore may be more biologically plausible. We quantify the degree of knowledge-based support for a model with an *implication index*. The implication index is a crude measure of the strength of gene-gene interaction or gene-disease relationship, and is calculated simply by summing the number of data sources supporting each of the two genes and the connection between them. An example is shown in figure 20. One disease-dependent gene (*ASL*) is supported by two data sources, and is connected to another disease-dependent gene (*CPS1*) that is supported by one source. The connection between these two genes is supported by three data disease independent sources, so the implication index of the model is six. In this manner, the implication index provides an ordinal value representing the degree of evidence supporting the biological plausibility of a multi-SNP model.

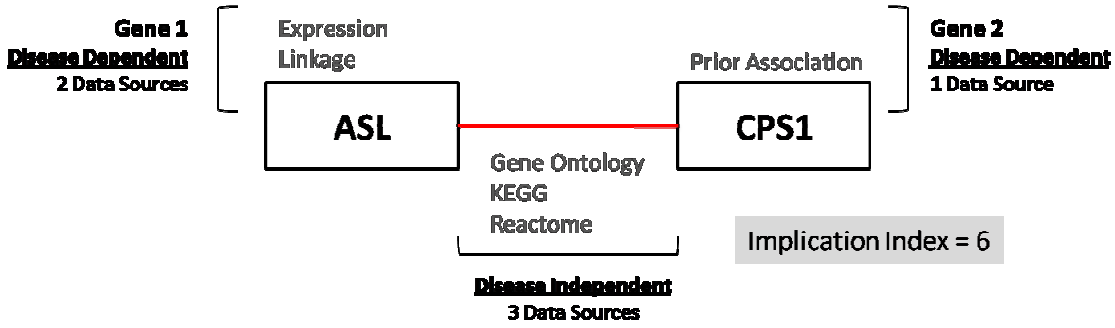

Figure 20. Implication index calculation. A two-gene model has two disease dependent sources that support gene 1, one disease dependent source that supports gene 2, and three disease independent sources that support the connection of gene 1 and gene 2. Thus, the implication index for the model is six.

Table 4. General GWA platform statistics

	Illm1M	Affy60
Total SNPs (with RS IDs)	1,055,373	924,689
SNPs within genes	493,854	353,913
Genes directly represented	21,024	17,418
Common SNPs	267,900	
Common SNPs within genes	118,355	
Common genes represented	16,908	

Results

GWA platform representation

Two commonly used large-scale genotyping platforms were assessed in this study: the Illumina Human1M-Duo BeadChip (Ilmn1M) and the Affymetrix Genome-Wide Human SNP Array 6.0 (Affy60). For the purposes of our assessment, only probes with vendor-specified Reference Sequence (RS) numbers were used to assure continuity of genomic position. General statistics for these two platforms are shown in Table 4. The Affymetrix and Illumina platforms are largely comparable, with more genes directly

represented on the Illumina Human1M. A core set of nearly 17,000 genes are well-represented by both platforms.

As described in the methods section, genes can belong to disease-dependent and disease-independent models. Because disease-independent models can be used in an analysis of any phenotype, we focused on how genes from those models are represented in different data sources. Table 5 shows the number of gene pairs represented from each platform in each disease-independent data source. PFAM has the highest number of gene pairs for both platforms, and DIP has the lowest number – this is directly related to the difficulty in collecting data for each of these respective sources. PFAM generates data by sequence analysis – a purely computational procedure. DIP on the other hand is based entirely on experimental data and so has far fewer gene pairs than other sources. GO is the oldest and consequently has the highest gene-pair representation of the remaining data sources. Because the Illumina platform captures more genes than the Affymetrix platform, Illm1M has a consistently higher number of gene pairs across all data sources.

Table 6 shows the pair-wise overlap across the six selected public databases. In this table, the amount of redundant gene representation is shown, with PFAM containing more than 90% of genes included in all other data sources. GO contains more than half the genes represented in each of the pathway data sources, but the three pathway sources appear to contain mostly independent gene sets (< 50% overlap).

Generalized disease independent models

Disease-independent models are universally applicable to any phenotype, so we generated all derived gene-gene pairs, and platform specific two-SNP models for the Illm1M and Affy60. Counts of these two-SNP models by implication index are shown in table 7. The Illm1M platform covers over a million more gene pairs and over a billion more two-SNP models than the Affy60. As such, the Illm1M platform has far superior coverage of the interaction models generated using these data sources.

Table 5. Gene pairs produced from Biofilter data sources by platform.

Data Source	Illm1M	Affy60
PFAM	14911	12837
DIP	747	638
GO	6129	5359
KEGG	4058	3543
Reactome	1799	1610
Netpath	3704	3246

Table 6. Pair-wise overlap of all genes in disease-independent Biofilter data sources. The data source listed on each row contains genes that overlap the data source listed on each column. Cell values indicate the proportion of the genes in the column source that are represented in the row source.

	PFAM	DIP	GO	KEGG	Reactome	Netpath
PFAM	1.00	0.95	0.92	0.95	0.95	0.92
DIP	0.05	1.00	0.09	0.08	0.15	0.12
GO	0.37	0.73	1.00	0.63	0.64	0.56
KEGG	0.01	0.02	0.01	1.00	0.02	0.02
Reactome	0.12	0.36	0.19	0.25	1.00	0.21
Netpath	0.22	0.6	0.34	0.41	0.41	1.00

Performing an exhaustive analysis of all possible two-SNP models within genes represented by these two platforms would result in 1.22e11 models for the Illumina 1M and 6.26e10 models for the Affymetrix 1M. By reducing the interaction search space to only models with established biological plausibility via the disease-independent data sources, only 2.23e9 (Illm1M) and 1.2e9 (Affy60) model evaluations are required. Applying a Bonferroni correction to the exhaustive approach would require a model fit p-value of 4.10e-13 (Illm1M) and 7.98e-13 (Affy60) to be statistically significant. In contrast, using the knowledge-based approach, a Bonferroni correction of 2.25e-11 (Illm1M) or 4.16e-11 (Affy60) is required. In this manner, reducing the search space not only improves computation time, but also reduces the statistical burden of conducting biologically non-relevant statistical tests. Further model restriction (such as using models with an implication index > 1) would further reduce the Bonferroni adjusted significance threshold.

Table 7. Disease-independent gene pairs and model counts by implication index.

Implication Index	Illm1M Gene Pairs	Illm1M Two-SNP Models	Affy60 Gene Pairs	Affy60 Two-SNP Models
1	4,679,363	2,174,328,700	3,505,773	1,162,090,222
2	87,163	44,960,600	67,341	36,825,703
3	8,065	6,425,788	6,094	4,102,173
4	715	397,966	546	171,075
5	45	11,033	40	4,122
6	1	569	1	757
Total	4,775,352	2,226,124,656	3,579,795	1,203,194,052

Implication index in a GWA study of multiple sclerosis

To investigate the benefits of the Biofilter method, we applied it to a genome-wide association study of multiple sclerosis (see Chapter IV for full details and results). In addition to the six disease-independent data sources outlined previously, we included disease-dependent sources – linkage regions, prior association studies, candidate pathways, and gene expression studies (see Chapter IV). Genotype data were collected on 931 MS affected trios using an Affymetrix Mapping 500K SNP Chip, with 334,923 SNPs passing quality control thresholds (see Chapter IV). Roughly twenty million two-SNP models were constructed from 334,923 SNPs, and evaluated using conditional logistic regression in 931 case/pseudo-control pairs generated from the transmitted and untransmitted alleles for each trio. The model contained a term for the SNP from gene 1, a term for the SNP from gene 2, and an interaction term (SNP 1 X SNP 2).

Two statistics were computed for each model evaluated in the data: a model fit p-value and an interaction term p-value. The model fit p-value describes how well the statistical model fits or explains the case/pseudo-control data, and the interaction term p-value describes the decline in model fit when the interaction term is removed from the model. Because each of the twenty million models has a corresponding implication index, we examined the relationship between the number of data sources supporting a model and the model statistics.

Figure 21 shows the proportion of models with significant model fit statistics ($p < 0.05$) by relative implication index, displayed by model type. *Relative implication index* only implies that there is a baseline number of data sources required for some model types, for example a hybrid two gene model has a minimum of three supporting data sources and thus an implication index of three.

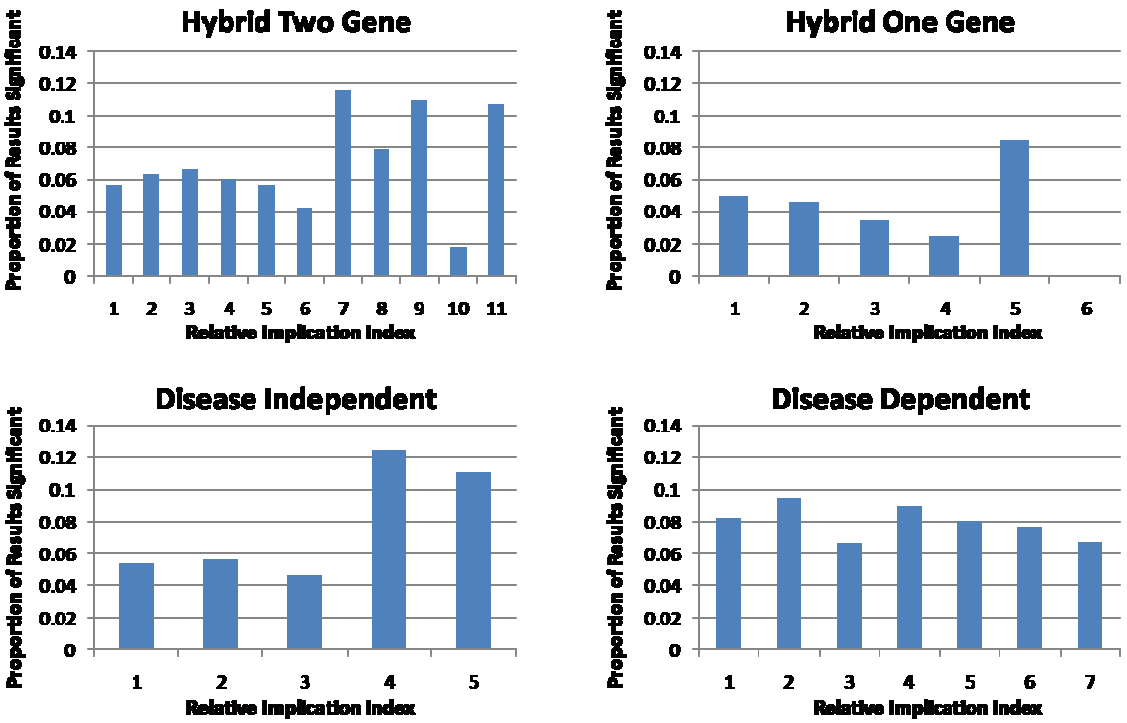

Figure 21. Proportion of significant model fit statistics by relative implication index. The x-axis indicates the relative number of knowledge sources supporting the statistical model, and the y-axis indicates the proportion of two-SNP models significant ($p < 0.05$) for each model type.

For disease-independent and hybrid model types, the proportion of significant model fit statistics is, in general, greater for when more knowledge sources support a given model. Disease-dependent models do not show this trend, indicating that using disease-gene relationships alone is not as beneficial for discovering multi-SNP models.

Figure 22 illustrates a similar effect for interaction term significance. The hybrid model types in general show increased interaction term significance with more supporting knowledge sources. This effect is not nearly as pronounced in the disease-dependent or disease-independent models.

To formalize this concept, we conducted a logistic regression analysis to assess the relationship between the relative implication index and the significance of model statistics. The outcome of the regression model was a binary significance indicator (0 if $p > 0.05$, 1 if $p <= 0.05$) and the relative implication index was used as the dependent variable. Results from the regression analysis are shown in table 8. The regression results demonstrate significant relationships ($p < 0.05$) between the relative implication index and the significance of model fit and interaction term statistics for all model types. Odds ratios were calculated to determine the direction of the effect -- and odds ratio greater than 1 indicates that increasing the number of supporting knowledge sources also increases the probability of a statistic being significant, while an odds ratio less than 1 indicates a decreasing probability. The only consistent effect is from the disease-independent models, where a higher implication index increases the odds of a significant model fit, a significant interaction term, and both significant model fit and interaction terms. The odds of model significance by implication index and model type are shown in table 9. These regression results are descriptive statistics for this dataset only, and may not generalize to other datasets or phenotypes. In this GWA study, however, they indicate that disease-dependent knowledge is not as useful as disease-independent knowledge for finding significant interactions.

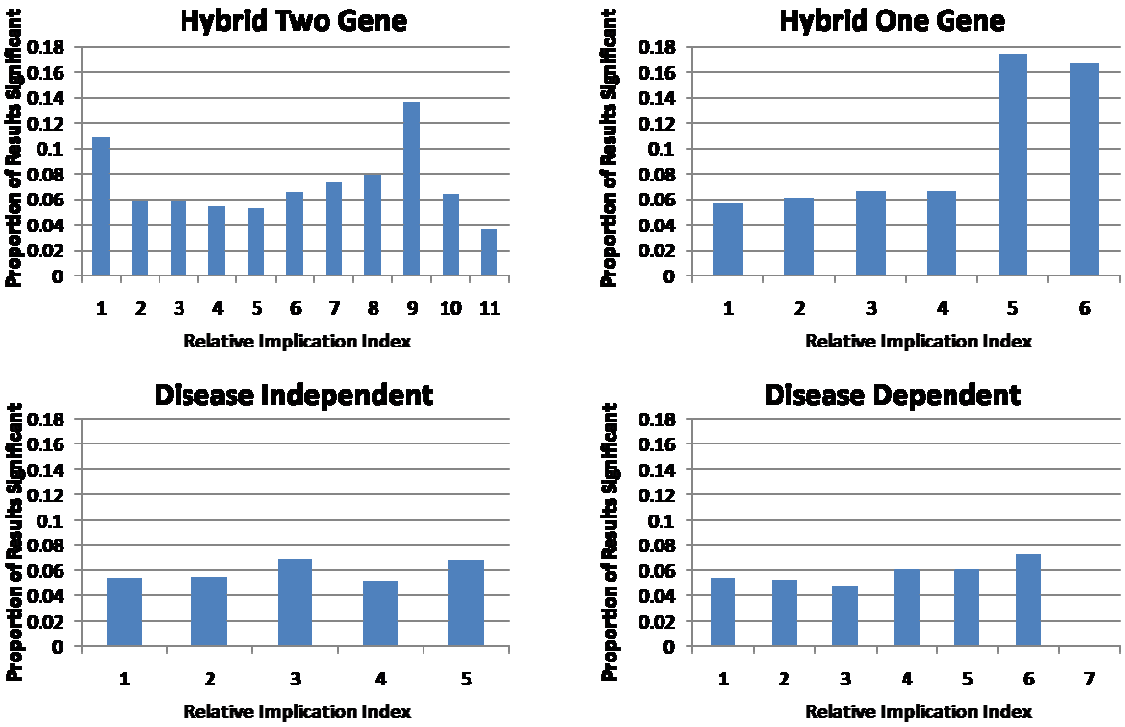

Figure 22. Proportion of significant interaction term statistics by relative implication index. The x-axis indicates the relative number of knowledge sources supporting the statistical model, and the y-axis indicates the proportion of two-SNP models with significant interaction terms by likelihood ratio test ($p < 0.05$) for each model type.

Table 8. Logistic regression of implication index on model statistics.

	Interaction Term		Model Fit		Both	
	p-value	OR	p-value	OR	p-value	OR
Disease Dependent	0.008	0.986	0	0.962	0	0.9227
Disease Independent	0.003	1.018	0	1.039	0.027	1.014
Hybrid One	0	1.032	0	0.887	0.137	1.016
Hybrid Two	0	1.012	0	1.024	0.098	0.995

Table 9. Odds of significance by implication index. Blue shaded cells indicate an extrapolated value.

Implication	Disease Dependent	Disease Independent	Hybrid One	Hybrid Two
Odds of Model Fit Significant ($p < 0.05$)				
1	0.96	1.04	0.89	1.02
2	0.93	1.08	0.79	1.05
3	0.89	1.12	0.70	1.07
4	0.86	1.17	0.62	1.10
5	0.82	1.21	0.55	1.13
6	0.79	1.26	0.49	1.15
7	0.76	1.31	0.43	1.18
8	0.73	1.36	0.38	1.21
9	0.71	1.41	0.34	1.24
10	0.68	1.47	0.30	1.27
Odds of Interaction Term Significant ($p < 0.05$)				
1	0.99	1.02	1.03	1.01
2	0.97	1.04	1.07	1.02
3	0.96	1.05	1.10	1.04
4	0.95	1.07	1.13	1.05
5	0.93	1.09	1.17	1.06
6	0.92	1.11	1.21	1.07
7	0.91	1.13	1.25	1.09
8	0.89	1.15	1.29	1.10
9	0.88	1.17	1.33	1.11
10	0.87	1.20	1.37	1.13
Odds of Model Fit and Interaction Term Significant ($p < 0.05$)				
1	0.92	1.01	1.02	1.00
2	0.85	1.03	1.03	0.99
3	0.79	1.04	1.05	0.99
4	0.72	1.06	1.07	0.98
5	0.67	1.07	1.08	0.98
6	0.62	1.09	1.10	0.97
7	0.57	1.10	1.12	0.97
8	0.53	1.12	1.14	0.96
9	0.48	1.13	1.15	0.96
10	0.45	1.15	1.17	0.95

Conclusions

When examining epistasis in genome-wide association studies, there are several variable selection strategies. Exhaustive evaluation of all multi-SNP models is generally computationally impractical. Exploring epistasis within a set of SNPs with detectable main effects may prevent the discovery of complex genetic models where trait variance is explained largely by the interaction of SNPs. Using biological knowledge to perform SNP selection provides two key benefits simultaneously: it reduces the multi-SNP model search space, and it provides a biologically plausible foundation for the models to be evaluated. Also, in common practice biological information is applied post-analysis to identify biologically compelling or relevant findings, and these results may even be exclusively selected for follow-up. Applying this knowledge pre-analysis prevents the excess computation and statistical interpretation of results that are not immediately relevant. As such, we developed the Biofilter to systematically reduce model search space based on multiple sources of structured biological knowledge.

We mapped the disease-independent models generated by the Biofilter to two GWA genotyping platforms, the Affymetrix 1M and the Illumina 1M. The final evaluated model search space was 0.241% of the exhaustive model space for Affy60 and 0.40% of the exhaustive model space for Illm1M when requiring at least one source of structured biological knowledge connecting the two genes in a two-SNP model, with further reductions possible by adjusting the number of required knowledge sources implicating the model.

The Biofilter method of variable selection can be implemented with a variety of analysis techniques, including logistic regression, classification and regression trees, and basic categorical statistics, among many others. To this end, the Biofilter is being developed as a knowledge-based filter for the PLATO analysis framework (PLATO REF?). Using PLATO, the collection of multi-SNP models generated by the Biofilter can be passed seamlessly to several other filters, including quality control checks for genotyping, analysis filters, and other knowledge-based processing and annotation tools. Results or other properties of each multi-SNP model are then stored, allowing retrieval of results with complete annotation of the SNPs, genes, gene grouping information, and in some cases, PubMed

references to the original articles implicating the model. The end result of a Biofilter style PLATO analysis is a set of biologically plausible, statistically relevant multi-SNP genetic models. Other modular whole-genome analysis platforms have been proposed as well (Galaxy ref - 16169926, Bioconductor ref -- Bioinformatics and Computational Biology Solutions Using R and Bioconductor -- Series: Statistics for Biology and Health -- Gentleman, R. ; Carey, V. ; Huber, W. ; Irizarry, R. ; Dudoit, S. (Eds.) -- 2005, XIX, 473 p. 128 illus. in color., Hardcover)

Some approaches may be adapted to incorporate the implication index into the analysis plan. Prioritized subset analysis, for example, partitions statistical results based on prior biological knowledge. The false discovery rate (FDR) for the "prioritized subset" is estimated separately, improving power when the prior knowledge is accurate (Li et al., 2008). Applying this strategy to subsets defined by the implication index could improve statistical power via p-value correction. Ranking models based on the number of supporting data sources may introduce unknown literature-based biases. Some data sources may have inter-dependencies, where one source was referenced in the creation of another. The breakdown of gene overlap for the 6 disease-independent data sources shows the diversity of gene-pairs represented, though notably PFAM contains nearly all of the gene-pairs established by the other sources. This is likely because PFAM contains the largest number of genes. When using disease-dependent data sources, there are certainly many factors that influence the inclusion and promotion of specific genes in relation to a phenotype, such as reporting bias.

Overall, the Biofilter provides a systematic way to assess the level of knowledge-based support for a given genetic model, provide a ranked list of all possible knowledge-based models, and to statistically test each of these hypotheses in genome-wide association data.

Acknowledgments

This work was supported by National Institutes of Health grants HL65962, and AG20135. Thanks to the authors, curators, and developers of the various database systems used in this project.

CHAPTER IV

A KNOWLEDGE-DRIVEN GENOME-WIDE MULTI-LOCUS ANALYSIS REVEALS A POTENTIAL ROLE FOR INOSITOL-BASED SIGNALING IN MULTIPLE SCLEROSIS

Introduction

Genetic epidemiology of multiple sclerosis

Multiple sclerosis (MS) is a complex, presumably autoimmune disorder characterized by demyelination and neurodegeneration within the central nervous system. Demyelination within multiple regions of the brain and spinal cord results in the formation of hardened scar tissues, which over time leads to neurodegeneration as the sclerosis impairs salutatory conduction of signals along axons causing reduced nerve function. MS has several characteristics in common with the larger class of demyelinating diseases, including relative high frequency, tendency to strike young adults, and diversity of disease manifestations (Oksenberg & Barcellos, 2005). The prevalence of MS is approximately 1 in 1000 in the US population (Anderson et al., 1992), and while the life expectancy of MS patients is not greatly decreased, the reduced quality of life and associated health care costs incur enormous personal burdens.

The etiology of MS is not well understood, but the role of both genetic and environmental factors has been established by multiple studies. Twin studies show concordance rates of 25-30% among monozygotic twins compared to 2-5% in dizygotic twins with MS (Sadovnick & Ebers, 1995). Furthermore, familial aggregation studies have estimated an increased relative risk (λ) of 20-40 for full siblings, 7-13 for half-siblings, and 5.5 for the offspring of an MS affected parent (Kenealy, Pericak-Vance, & Haines, 2003; Mumford et al., 1994; Robertson et al., 1996; Sadovnick & Ebers, 1995). Together, these studies suggest a complex genetic component to the disease.

The largest known genetic risk factor for MS is the *HLA-DRB1*1501* allele in the MHC region of chromosome 6. Individuals with this allele have twice the risk of developing MS versus the general population, and while this variant explains a significant proportion of genetic risk, it accounts for less

than 50% of the total genetic basis of MS (Haines et al., 1998). Linkage to the MHC region was identified and confirmed in several family studies (Chataway et al., 1998; Ebers, 1996; Oksenberg et al., 2001; Oturai et al., 1999), but other regions of increased linkage have failed to consistently replicate across studies.

In addition to *HLA-DRB1*, a single nucleotide polymorphism (SNP) in the interleukin-7 receptor α chain (*IL-7R*) was associated with MS in four independent datasets (Gregory et al., 2007; Lundmark et al., 2007; The International Multiple Sclerosis Genetics Consortium, 2007). The SNP, rs6897932, was shown to functionally influence the ratio of soluble and membrane-bound isoforms of the IL-7R protein, likely due to altered splicing (Gregory et al., 2007).

Concurrently, the International Multiple Sclerosis Genetics Consortium (IMSGC) conducted a genome-wide association study using 931 affected family trios (The International Multiple Sclerosis Genetics Consortium, 2007). This study confirmed the IL-7R association, and also identified the interleukin-2 receptor α chain (*IL-2R*) as associated with MS. These findings were confirmed in a large replication set. The *IL-7R* and *IL-2R* findings have very small effect sizes, with odds ratios of 1.18 for *IL-7R* and 1.25 for *IL-2R*. Together, the *IL-7R* and *IL-2R* SNPs explain less than 0.2% of the variance in MS risk (The International Multiple Sclerosis Genetics Consortium, 2007), and including the *HLA-DRB1* allele, roughly half of the genetic component at play in MS susceptibility remains undiscovered.

There are several ways to account for the remaining genetic effect that is unexplained. One possibility is that there are many small independent effects that influence MS risk -- small effects that are difficult to detect without very large sample sizes. Another possibility is that there are heritable factors that are not detected using the linkage or association studies that have been applied to MS to date. These factors could be copy number variants (CNVs) or other structural variations, methylation patterns, or other epigenetic components. Environmental factors may combine with genetic factors to influence risk, with a particular genetic architecture inducing an autoimmune response after a common infection or other external stimulus. As discussed in Chapter I, another possibility is epistasis, where combinations of genetic factors influence disease risk beyond their independent additive effects. In this case, the genetic

effect would remain largely hidden unless combinations of markers are examined together. Of these possibilities, we can examine epistasis using available data.

Epistasis in multiple sclerosis

Epistasis may play an important role in MS. Both *IL-7R* and *IL-2R* mediate downstream immune response pathways. These pathways constitute complex systems with interdependencies that provide the evolutionary opportunity for unlinked alleles to co-segregate in a population. Several examples of epistasis in complex disease can be found in Chapter I, but notably, epistasis has been functionally demonstrated in multiple sclerosis as an interaction among alleles of the MHC region (Gregersen et al., 2006). These alleles fall on a common haplotype background, and functionally interact to alter T-cell response, reducing the severity of the disease. It is proposed that these alleles remain in linkage disequilibrium due to positive selection.

Evidence suggesting epistasis was also presented in two studies using Multifactor Dimensionality Reduction on a set of inflammatory candidate genes. In a collection of 442 African-American MS cases and 293 controls, significant interactions were found between interleukin receptors 4 and 5 (*IL4-R*, *IL5-RA*) and *CD14*, indicating a potential role for dendritic cell antigen presentation in MS etiology (Brassat et al., 2006). Also, in a study of multiple family-based and case-control data sets, several epistatic models are proposed, including adrenergic beta-2 receptor (*ADRB2*), nitric oxide synthase 2A (*NOS2A*), nuclear factor kappa-B (*NFKB*), CD14, complement component 5 (*C5*), and uterglobulin (*UGB*) (Motsinger et al., 2007).

Examining epistasis in genome-wide association studies has noted challenges (Moore & Ritchie, 2004). There are roughly 125 billion possible two-SNP models in a set of 500,000 SNPs. Exhaustively analyzing all of these possible combinations is computationally costly, and dramatically amplifies the problem of multiple testing. One approach to reducing the number of interaction tests is to select a set of most significant results from a single-SNP analysis and exhaustively evaluate interactions in that set.

Another approach to reducing the number of interaction tests is to generate multi-SNP models based on prior biological knowledge, for example testing for interactions only between SNPs that occur in the same biochemical pathway (Carlson et al., 2004). The drawback of this approach is that it is dependent on the current, incomplete state of biological knowledge, and thus will not discover novel interactions between SNPs that may influence disease risk. However, the benefits of this approach are that it takes into account the vast information known about the structure of biological systems, effectively reducing the set of interactions that are analyzed to those relating to a common functional endpoint (see Chapter III). Using biological knowledge also provides a mechanistic rationale for why two SNPs might interact, or perhaps how they are related to the disease being studied. This approach is a biased search, in that it would not discover the novel interaction of genes, but instead would discover novel relationships between the phenotype and functional components containing those interacting genes. As such, this knowledge-based analysis examines the influence that a known set of system-level components might have on disease risk.

Here we present a knowledge-driven multi-locus analysis of MS susceptibility using the 931 affected trios from the 2007 IMSGC genome-wide association study (The International Multiple Sclerosis Genetics Consortium, 2007). Each individual in the study was genotyped using the Affymetrix Mapping 500K SNP chip. 334,923 SNPs passed quality control procedures outlined in (The International Multiple Sclerosis Genetics Consortium, 2007). These 334,923 SNPs from the study were assigned to 13,425 genes using patterns of linkage disequilibrium from the International Hapmap Project (International HapMap Consortium, 2005) by applying the LD-Spline approach outlined in Chapter II (142,814 SNPs were used in the mapping). Using a collection of data sources that suggest putative gene-gene interaction or prior disease gene implication, we generated a set of multi-SNP models and evaluated them using conditional logistic regression. Sources of evidence for potential gene-gene interaction are KEGG (Kanehisa et al., 2006; Kanehisa et al., 2008; Kanehisa & Goto, 2000), Reactome (Vastrik et al., 2007), DIP (Xenarios et al., 2002; Xenarios et al., 2000; Xenarios et al., 2001), PFAM (Finn et al., 2006; Finn et al., 2008), GO (Ashburner et al., 2000), and Netpath (Pandey & Institute of Bioinformatics, 2008). Sources of evidence for prior

disease gene implication are GAD(Becker et al., 2004), previous linkage screens by the IMSGC (Sawcer et al., 2005) and Genetic Analysis of Multiple sclerosis in EuropeanS (GAMES & Transatlantic Multiple Sclerosis Genetics Cooperative, 2003), three studies of MS gene expression (Bomprezzi et al., 2003; Comabella & Martin, 2007; Sarkijarvi et al., 2006), a hand-curated set of IL-7 pathway genes (Rebecca Zuvich, personal communication), and other literature-based candidates (Frohman et al., 2005; Lock et al., 2002; Saarela et al., 2006; Sinclair et al., 2007; Swanberg et al., 2005). By incorporating these sources of prior information into our analysis, we have identified two replicating multi-locus associations to MS within biological pathways.

Materials and Methods

Samples

The data used for the screening analysis was a collection of 931 family trios, consisting of an affected child and both parents (IMSGC). There were 478 samples ascertained from the US (356 female, 122 male, 2.92:1) and 453 from the UK (339 female, 114 male, 2.97:1). US and UK based samples were homogenous with respect to age at analysis (US mean 39.0, range 17-57 ; UK mean 38.2, range 19-59), and age at disease onset (US mean 29.0, range 11-51 ; UK mean 26.5, range 10-48).

After review of clinical data, the diagnosis of MS was made according to the McDonald criteria (McDonald et al., 2001; Polman et al., 2005). Some clinical heterogeneity was present between US and UK samples with respect to disease course, with differences in frequency of relapsing remitting (US 374, 78.2% ; UK 269, 59.4%) and secondary progressive (US 77, 16.1% ; UK 152, 33.6%) clinical subtypes. The small number of primary progressive cases had similar frequencies (US 25, 5.2% ; UK 32, 7.2%).

A set of population-based controls from the 1958 UK birth cohort (1485 individuals) and the UK blood service (1465 individuals) were generously provided by the Wellcome Trust Case Control Consortium (WTCCC). Female to male ratio was roughly 1:1 (1503 female, 1447 male), and age was reported by range, with over 50% of controls between 40-49.

The replication sample was a collection of 808 cases and 1720 controls ascertained at Brigham and Women's Hospital, Boston (BWH). Female to male ratio was similar to the screening sample (595 female, 213 male, 2.8:1). Mean age at disease onset was also comparable (mean 33, range 8-59).

IMSGC and WTCCC samples were genotyped using the Affymetrix GeneChip Human Mapping 500K Array Set, and BWH samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Extensive quality control procedures were applied to IMSGC and BWH samples to assess genotyping efficiency, monomorphic SNPs, multi-hit SNPs, Mendelian errors (for trios), sex assignment discrepancies, Hardy-Weinberg Equilibrium, population stratification, low minor allele frequency, and plate effects. In the IMSGC sample, genotyping was validated using Sequenom iPLEX™ and iPLEX Gold™ MassARRAY®11 genotyping platform. For the IMSGC analysis, 334,923 SNP markers were used. From the BWH genome-wide panel, 453 SNPs were used for analysis. Futher details of IMSGC sample and SNP quality assessment is available in supplemental data of (The International Multiple Sclerosis Genetics Consortium, 2007).

Linkage Disequilibrium Mapping of Markers to Genes

Pair-wise linkage disequilibrium (LD) statistics computed for over two million SNPs by the International HapMap Project (Frazer et al., 2007; International HapMap Consortium, 2005) (posted June, 2006) were used to establish the Caucasian-specific haplotype block boundary for each of the 334,923 SNPs in the IMSGC data set. Using the LD-Spline procedure described in Chapter II, we defined the boundaries of the haplotype block represented by each IMSGC SNP. Because the IMSGC SNPs are a subset of all known genomic variants, using HapMap LD statistics in this way provides the larger genomic region (which may harbor susceptibility variants) represented by each IMSGC SNP. 5,137 markers in the IMSGC data set were not represented in the HapMap LD data, and the nearest HapMap marker was used as a surrogate to assess haplotype block boundaries. Marker-gene mappings were generated if a haplotype block that overlaps with any portion of a gene as described by the Ensembl

database (Flicek et al., 2008). IMSGC markers capture 14,236 genes using LD, compared to 13,425 using the markers without accounting for LD.

Incorporating Biological Knowledge Sources

We used two types of biological knowledge sources for this analysis: sources that contain putative gene relationships to disease (disease dependent sources), and sources that contain putative gene-gene relationships (disease independent sources).

Linkage studies, prior association studies, candidate pathways, and differential gene expression studies all serve to implicate sets of genes as being important for multiple sclerosis susceptibility. The multi-point LOD scores for 5,282 markers used in a study of 730 multiplex families (Sawcer et al., 2005) were used to identify genes under linkage peaks of 1.8 or greater (139 genes) and between 1.5 and 1.8 (502 genes). Three candidate regions (17q21, 19q13, and 22q13) identified by multiple linkage studies were also included (2021 genes) (GAMES & Transatlantic Multiple Sclerosis Genetics Cooperative, 2003; Pericak-Vance et al., 2001). The Genetic Association Database (GAD) was used to identify genes previously associated to multiple sclerosis (183 genes), and to the larger class of immune diseases (776 genes) (Becker et al., 2004). Three studies highlighting differences in gene expression between cases and controls (50 genes) (Bomprezzi et al., 2003), and between discordant monozygotic twins (35 genes) (Sarkijarvi et al., 2006) were used, along with common expression differences in MS tissues and experimental allergic encephalomyelitis (EAE) mouse models (47 genes) (Comabella & Martin, 2007). We also included a list of 73 candidate genes based on the IL-7 pathway (Rebecca Zuvich, personal communication) and 19 genes implicated by population-based and functional studies (Frohman et al., 2005; Lock et al., 2002; Saarela et al., 2006; Sinclair et al., 2007; Swanberg et al., 2005).

Biochemical and regulatory pathways, protein families and ontologies, and protein interaction networks are all sources of putative gene-gene relationships, as described in Chapter III. For this application, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) (4238 genes) (Kanehisa et al., 2006; Kanehisa et al., 2008; Kanehisa & Goto, 2000), Reactome (1931 genes) (Joshi-Tope et al., 2005;

Vastrik et al., 2007), and Netpath (3855 genes) (Pandey & Institute of Bioinformatics, 2008) as sources of pathway information. The Database of Interacting Proteins (DIP) (791 genes) (Xenarios et al., 2002; Xenarios et al., 2000; Xenarios et al., 2001) was used as a source of protein interactions. The Protein Families Database (PFAM) (15969 genes) (Finn et al., 2006; Finn et al., 2008) was used to group genes by protein sequence and functional similarity. The Gene Ontology (GO) (6391 genes) (Ashburner et al., 2000) was used to group genes by ontological categories related to cellular components, biological processes, and molecular functions.

Each data source contained a form of gene or protein identifier (i.e. Entrezgene ID, Unigene ID, Uniprot ID, etc) that was translated to an Ensembl gene ID using the Ensembl database. Using both putative gene-disease and putative gene-gene relationship sources, we used the procedure described in Chapter III to generate disease-independent, disease-dependent, hybrid one disease-dependent gene, and hybrid two disease-dependent gene models. Overlap between each model type is removed, such that there are no two gene hybrid models in the disease-independent set.

Statistical Analysis

Multi-SNP models are evaluated using case/pseudo-control pairs in a conditional logistic regression analysis as described in (Cordell, Barratt, & Clayton, 2004; Siegmund et al., 2000). The conditional regression algorithm was adapted from a Fortran routine (Krailo & Pike, 1984), implemented in C++, and integrated into the PLATO computational framework. Implemented code was validated using routines available in SAS 9.1.3. Each multi-SNP model contains main effect terms for the two SNPs and an interaction term for the joint effect of the two SNPs. SNP genotypes were encoded in an additive manner, assuming a multiplicative interaction model as in (Marchini, Donnelly, & Cardon, 2005). Model significance was assigned by a likelihood ratio test of the fitted model to a null model (Hosmer & Lemeshow, 2000). Significance of the interaction term was assessed by a likelihood ratio test of the fitted model with an interaction term to the fitted model with no interaction term (main effect terms of the two SNPs only).

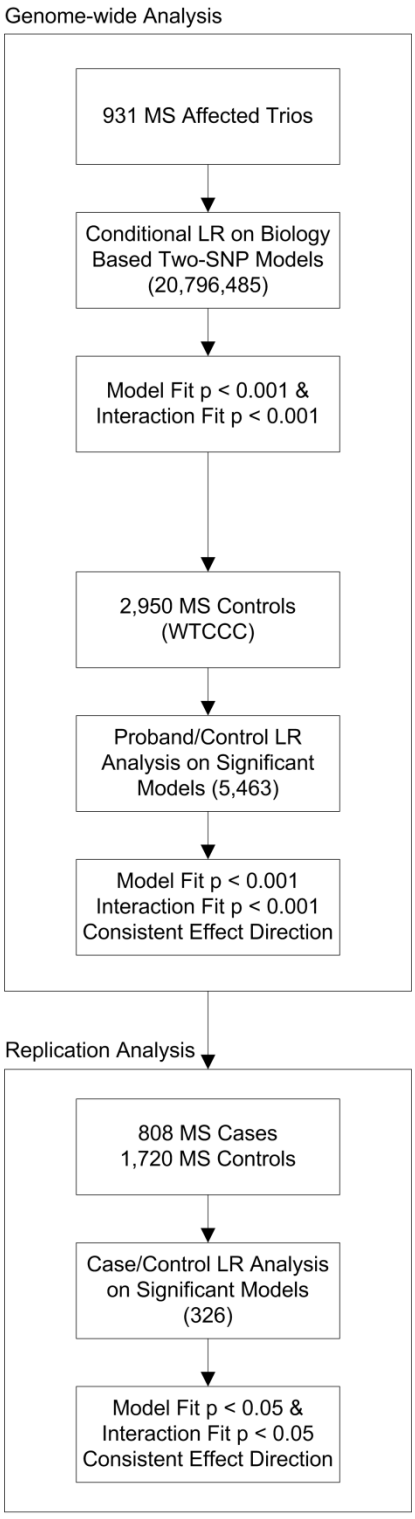

Figure 23. Analysis Plan Overview

Results

Screening analysis of genome-wide association data

Figure 23 shows an outline of the analysis. To identify statistically interacting genetic factors that influence multiple sclerosis susceptibility, we evaluated 334,923 quality control positive genotypes from 931 affected MS trios collected by the International Multiple Sclerosis Genetics Consortium (IMSGC). Using this panel of genotypes, we constructed 20,796,485 two-SNP models using prior biological information. These models were analyzed using conditional logistic regression on case/pseudo-control pairs generated from the transmitted and non-transmitted alleles of the original affected trios (Cordell, Barratt, & Clayton, 2004; Siegmund et al., 2000). The conditional logistic regression model contained three terms ; a term capturing the additive main effect of each of the two SNPs included in the model, and a multiplicative interaction term. This conditional analysis does not include an intercept term -- each model is assessed within a case/pseudo-control stratum and a stratum specific intercept term (representing the baseline risk of MS within the trio) is generated but *conditioned out* of the analysis.

The initial screening phase revealed 5,463 significant models (5,965 SNPs) from the conditional logistic regression analysis of biologically derived two-SNP models using case/pseudo-control pairs ($p < 0.001$ for model fit and interaction significance using likelihood ratio tests), indicating an effect from simultaneous over-transmission of two alleles to affected offspring. These models were selected for inclusion in the second-stage analysis. Figure 24 (IMSGC) shows these significant results This plot is an interpolated surface with $-\log_{10}$ p-values of two-SNP model fits on the z-axis, plotted with the physical genomic position of SNP 1 on the x-axis and the physical genomic position of SNP 2 on the y-axis. Peaks on this surface are also color coded, with each color shade indicating an additional decreasing order of magnitude of the p-value.

As the MHC region is a strong genetic component of MS, we examined the contribution of the MHC to these significant models. Fifty-eight significant models contained two SNPs from chromosome 6, and of these there were 24 models where both markers were in the MHC region, and 5 models where

only one marker was in the MHC region. 420 models contained one SNP from chromosome 6, and of these, 141 models contained a SNP in the MHC region and 117 models with only one SNP in the MHC region. 236 models have significant model fits and inconsistent directions across studies. In short, a small number of models contained SNPs in the MHC region, indicating that with respect to this analysis, the MHC is a largely independent effect.

To maximize information from the screening phase, we compared the probands from the 931 family trios with control data from 2950 individuals from the Wellcome Trust Case Control Consortium (WTCCC). A drawback of the case/pseudocontrol multi-locus analysis is that two SNPs on the same chromosome may be joinly over-transmitted to affected offspring due to chromosomal linkage rather than the joint effect of those SNPs on the phenotype. As a result, the false positive rate of the interaction statistic from this analysis may be inflated. A comparison between the probands and unrelated controls using logistic regression will reduce false positives due to this over-transmission in the family-based analysis. Using the proband/control analysis, we reduced the 5463 models to 326 models ($p < 0.001$ for model fit and interaction significance) containing 469 SNPs. An interpolated surface of $-\log_{10}$ p-values for the model fit statistics of these models is shown in figure 24 (WTCCC).

Replication analysis

For a separate study, the Affymetrix 6.0 platform was used to genotype 808 MS cases and 1720 controls ascertained at Brigham and Women's Hospital (BWH). From this panel, we selected the 453 QC+ SNPs included in the 326 significant models from the genome-wide analysis. Twenty models were not evaluated due to one or more SNPs failing the QC procedures in the BWH sample.

The results of the replication analysis are shown in figure 24 (BWH) as an interpolated surface of $-\log_{10}$ p-values, and the annotated results of both the screening and replication analyses are shown in table 10. Twenty multi-locus models had significant models fits ($p < 0.05$) in the screen and replication sets. Of these, two are particularly notable, as they have both a significant model fit and a significant interaction component ($p < 0.05$ for both).

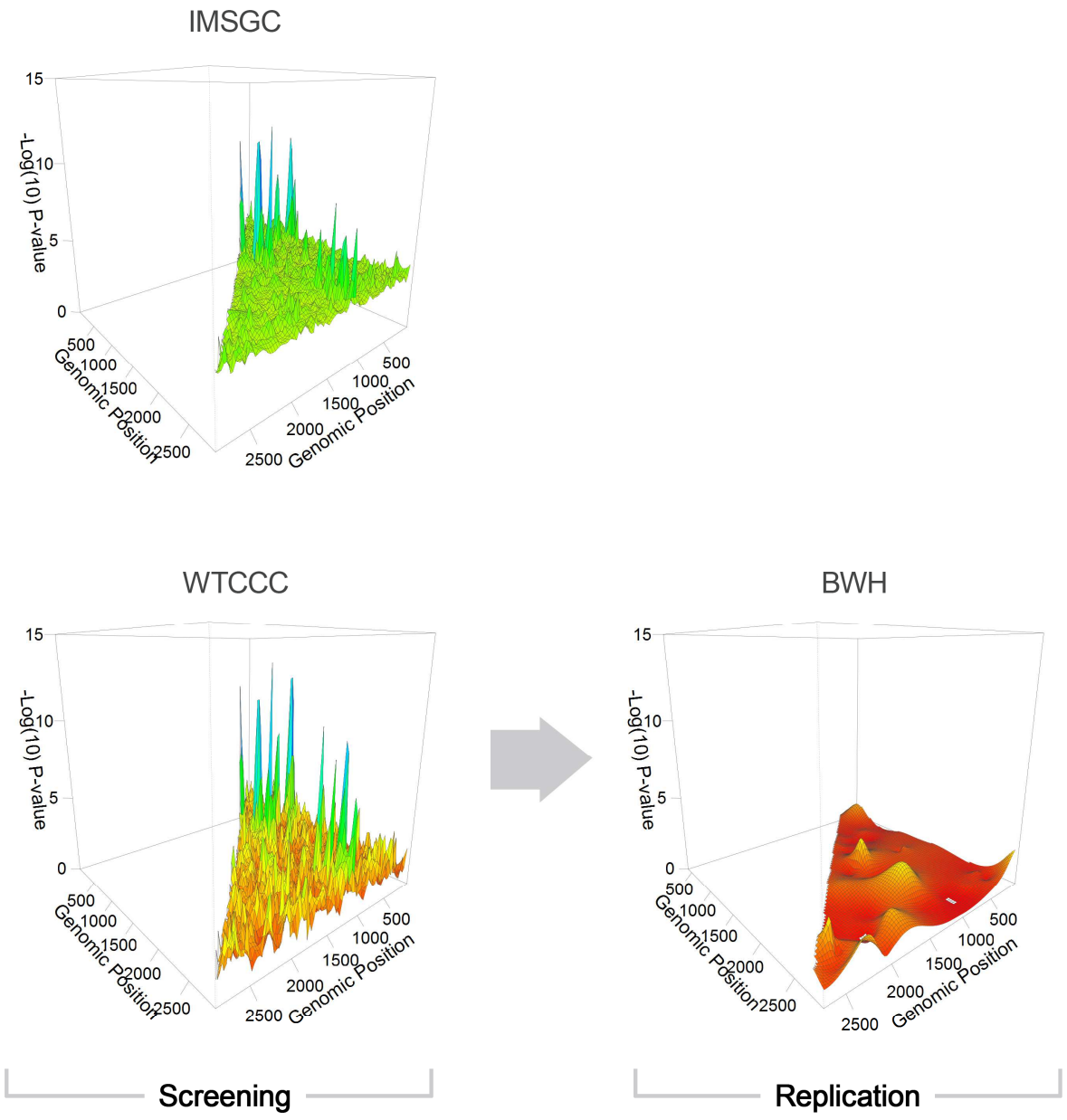

Figure 24. Overview of Knowledge-based Genome-wide Interaction Analysis.
Model fit p-values are shown for screening (conditional logistic regression of 931 MS family trios (IMSGC), logistic regression of 931 MS affecteds and 2950 unaffecteds (WTCCC)), and replication (logistic regression of 808 MS affecteds and 1720 unaffecteds (BWH)) phases. $-\log_{10}$ values for 5463 significant IMSGC models (top panel) and 306 significant WTCCC models (bottom panel) were processed to create genomic interaction surfaces, where each point on the grid represents the interaction of two loci across the genome.

SNP rs528011 in the muscarinic cholinergic receptor 3 (*CHRM3*) gene located on chromosome 1 interacts with SNP rs4677905 in the myosin-light-chain kinase (*MYLK*, EC:2.7.11.8) gene located on chromosome 3 (replication model fit p = 0.0235, replication interaction test p = 0.0026). This model was generated based on prior associations of each gene to an immune disease (via the GAD), and because the genes are related by functioning in the calcium signaling pathway (via KEGG). *CHRM3* has been previously associated to asthma and atopy (Donfack et al., 2003). *MYLK* has also been previously associated to asthma(Flores et al., 2007; Gao et al., 2007). *CHRM3* and *MYLK* both function in the calcium signaling pathway(Berridge, Bootman, & Roderick, 2003) (KEGG pathway: ko04020) and the regulation of actin cytoskeleton pathway (KEGG pathway: ko04810).

SNP rs4816129 in the phospholipase C beta 4 (*PLCβ4*) gene located on chromosome 20 interacts with SNP rs6516415 in the phospholipase C beta 1 (*PLCβ1*) gene located on chromosome 20 (replication model fit p = 0.0475, replication interaction test p = 0.0095). *PLCβ4* and *PLCβ1* function in multiple KEGG pathways, including the calcium signaling pathway shown in figure 25 (KEGG pathway:ko04020), Wnt signaling (KEGG: ko04310), and inositol phosphate metabolism (KEGG:ko00562). Also, *PLCβ4* and *PLCβ1* are both members of the phosphatidylinositol-specific phospolipase C family, and contain the PI-PLC-X and PI-PLC-Y domains, via PFAM (Meldrum, Parker, & Carozzi, 1991).

Seven other models had significant model fit in the replication sample ($p < 0.05$) but a non-significant interaction component by likelihood ratio test. Notably, two of these models contain genes that also function in the actin cytoskeletal regulation, shown in figure 26. *ACTN1* located on chromosome 14 and *MYH9* located on chromosome 22 (replication model fit p=0.0087, replication interaction significance p=0.1066) functions in the formation of actin stress fibers and cytoskeletal contraction (via KEGG). *CYFIP1* located on chromosome 15 and *SCIN* located on chromosome 7 (replication model fit p=0.0041, replication interaction test p = 0.3235) function in lamellepodia formation (via KEGG).

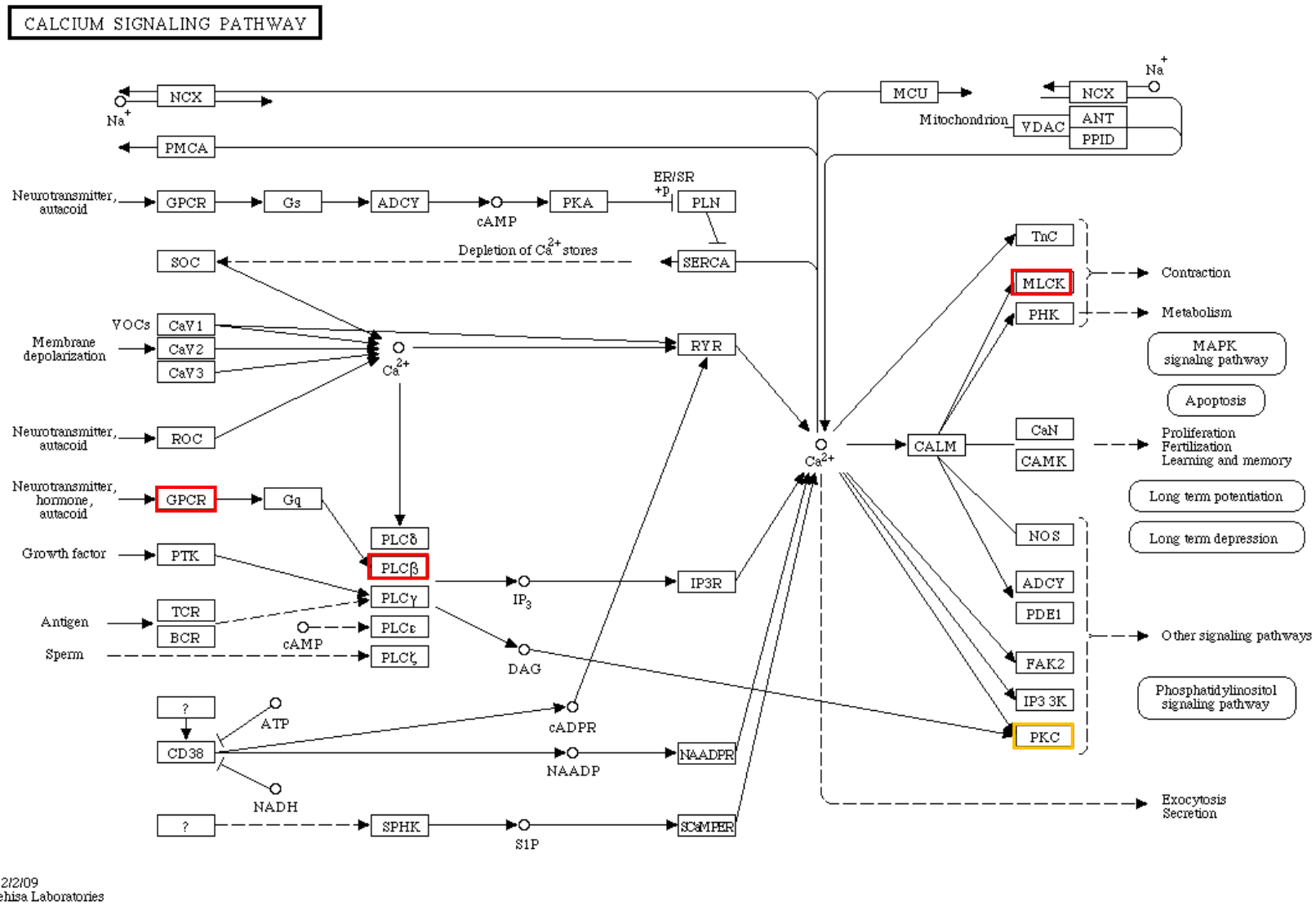

Figure 25. The calcium signaling pathway (adapted from KEGG). Genes represented by two-SNP models with significant model fit and interaction term statistics are outlined in red (*CHRM3* is a GPCR). Previously implicated gene *PRKCA* is outlined in yellow.

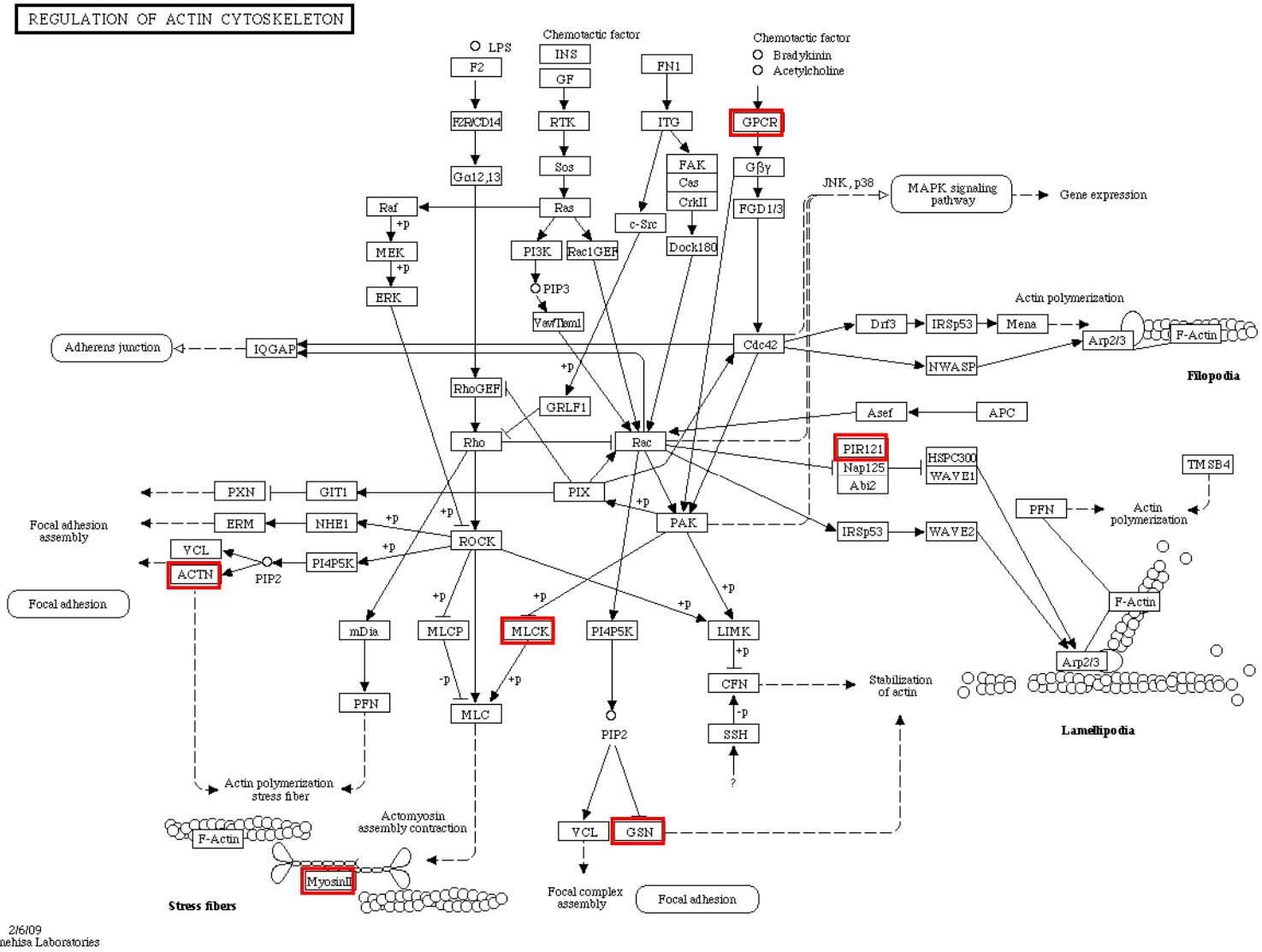

Figure 26. The regulation of actin cytoskeleton pathway (adapted from KEGG). Genes represented by two-SNP models with significant model fit statistics only are outlined in red. (*CYFIP1* is represented as PIR121, and *SCIN* is represented as GSN)

Also of note, the thyroid stimulating hormone receptor (*TSHR*) on chromosome 14 and the andronergic receptor A1A (*ADRA1A*) on chromosome 8 (replication model fit p=0.270, replication interaction test p=0.9820) are both G-protein coupled receptors with interesting functional implications. *ADRA1A* is one of several receptors that can activate the calcium signaling pathway by binding epinephrine. *TSHR* is a hormone receptor that functions in immune signaling pathways, and may have common gene regulatory factors with the major histocompatibility complex(Ohmori et al., 1996). Polymorphisms in *TSHR* have been previously associated to Grave's disease(Hiratani et al., 2005) and autoimmune thyroid disease(Akamizu et al., 2000).

Another notable model contains estrogen receptor 1 (*ESR1*) on chromosome 6 and the interleukin-2 receptor alpha chain (*ILR2A*) on chromosome 7 (replication model fit p=0.0041, interaction model fit p=0.2148). These two genes have been previously implicated in multiple sclerosis (via GAD). One study indicates *ESR1* has significant associations in the Japanese population(Niino et al., 2000), and a second study finds the effect of *ESR1* is modulated by HLA status(Mattila et al., 2001).

Pathway enrichment

Two-hundred and twenty-two distinct gene pairs were evaluated in the replication set (using 306 two-SNP models). Of these, there were 9 gene pairs associated with cytoskeletal regulation (via KEGG), 11 gene pairs associated to calcium signaling (via KEGG), and 19 gene pairs associated with the union of these pathways. Three of the 9 cytoskeletal regulation gene pairs were statistically significant, 2 of the 11 calcium signaling gene pairs were statistically significant, and 4 of the union pathway gene pairs were statistically significant at $p < 0.05$. As such, the "regulation of actin cytoskeleton" pathway was significantly enriched ($p = 0.005$, exact test), and the union of the calcium signaling and regulation of actin cytoskeleton pathways was significantly enriched ($p = 0.006$). The calcium signaling pathway itself was marginally enriched ($p = 0.081$, exact test).

Table 10. Summary of results from the GWA study and replication studies

	Locus 1				Locus 2					IMSGC Proband/Pseudo-Control		IMSGC Proband/WTCCC Control		BWH Case/BWH Control	
Number	**Chr**	**Gene**	**Source**	**SNP**	**Chr**	**Gene**	**Source**	**SNP**	**Pair source**	**Model Fit**	**Interaction**	**Model Fit**	**Interaction**	**Model Fit**	**Interaction**
1	1	*TNFRSF1B*	E,M,I	rs235219	22	*APOBEC3G*	L	rs8177832	N	1.36E-05	1.23E-04	1.43E-04	3.55E-04	0.0031	0.5273
1	15	*CYFIP1*		rs8025779	7	*SCIN*		rs2240571	G,K	3.75E-04	1.51E-04	1.50E-04	8.13E-05	0.0041	0.3235
2	6	*ESR1*	M,I,C	rs9340817	10	*IL2RA/PFKB3*	I	rs12722489	N	6.81E-05	8.45E-04	3.11E-06	8.87E-04	0.0046	0.2148
2	14	*ACTN1*		rs17106421	22	*MYH9*		rs1009150	K,N	8.93E-04	6.38E-05	9.54E-05	6.90E-06	0.0087	0.1066
2	10	*PTPRE*		rs10829321	19	*PTPRS*		rs7259497	P,G	1.07E-04	6.83E-04	4.85E-04	7.25E-05	0.0189	0.4735
1	**1**	***CHRM3***	**I**	**rs528011**	**3**	***MYLK***	**I**	**rs4677905**	**K**	**5.57E-04**	**3.74E-05**	**4.75E-04**	**1.22E-04**	**0.0235**	**0.0026**
1	14	*TSHR*		rs179250	8	*ADRA1A*		rs4732652	P,K	8.26E-04	4.35E-04	2.04E-04	1.35E-04	0.0270	0.9820
8	19	*CARD8*	I,L	rs2910400	11	*ZFP91-CNTF*	M,I,C	rs11229555		6.42E-06	2.53E-07	1.71E-04	1.13E-05	0.0289	0.8937
2	**20**	***PLCB4***	**L**	**rs4816129**	**20**	***PLCB1***	**L**	**rs6516415**	**P,K**	**9.23E-04**	**8.50E-05**	**7.73E-04**	**8.91E-04**	**0.0475**	**0.0095**
4	19	*SULT2A1*	I,L	rs2932766	11	*ZFP91-CNTF*	M,I,C	rs11229555		8.26E-05	4.19E-06	5.45E-04	4.49E-05	0.0756	0.8741
2	10	*COL17A1*	I	rs805693	11	*ETS1*	I	rs7117768	N	2.10E-05	1.60E-06	5.60E-04	5.27E-05	0.0929	1.0000
6	10	*PRKG1*		rs1903996	3	*PRKAR2A/DALRD3*		rs6446205	P,N	6.87E-04	7.96E-04	6.27E-04	1.70E-04	0.1135	0.0442
9	1	*GALNT2*		rs1967707	2	*GALNT13*		rs11677858	P,K	1.90E-04	2.30E-04	6.54E-05	4.90E-04	0.1465	0.3837
2	18	*MBP*	M,I	rs509620	20	*CD40*	M,I	rs1569723		8.31E-04	2.72E-04	2.93E-06	2.02E-04	0.1583	0.2141
2	5	*C7*	M	rs1551090	5	*PIK3R1*	I	rs2302976	N	6.22E-04	2.58E-04	1.67E-04	2.88E-05	0.2099	0.0398
4	6	*HA25_HUMAN*	M,I	rs9272219	6	*C6orf10/BTNL2*	M,I	rs12528797		1.77E-06	2.11E-05	7.59E-06	1.43E-04	0.2105	0.2499
22	19	*SYNGR4/GRIN2D*	L	rs190672	22	*CSF2RB*	I	rs4821560	N	1.22E-04	7.06E-04	1.73E-05	4.16E-04	0.2143	0.2673
2	12	*FGD4*		rs11052069	21	*TIAM1*		rs845960	P,R	8.83E-04	1.74E-04	1.67E-04	6.10E-04	0.2234	0.9938
1	17	*PRKCA*	M,L,C	rs3803821	20	*PAK7*	L	rs6118717	P,K	7.68E-04	4.02E-04	2.45E-04	4.06E-04	0.2290	0.6483
4	1	*PTPN7*		rs12735966	11	*PAK1*		rs3015993	K,N	3.01E-04	3.98E-05	8.33E-04	3.41E-04	0.2638	0.5845
4	1	*LGR6*		rs12735966	11	*PAK1*		rs3015993	K,N	3.01E-04	3.98E-05	8.33E-04	3.41E-04	0.2638	0.5845
2	2	*PLA2R1*		rs2667012	2	*IL1RN*	E,M,I,C	rs4251961	N	1.48E-04	7.56E-05	6.58E-04	8.37E-05	0.2742	0.0537
2	5	*GRIA1*	L	rs12515563	3	*ITPR1*	L	rs6774037	K	9.29E-04	3.54E-04	3.27E-05	2.81E-04	0.3013	0.3101
6	2	*ALK*		rs6715185	9	*ROR2/SPTLC1*		rs1569141	P,G	1.60E-04	9.87E-06	6.49E-04	4.82E-05	0.3013	0.8468
1	2	*NRXN1*		rs1895132	17	*PECAM1*	M,I,L	rs1122800	K	9.19E-04	1.41E-04	4.44E-04	3.73E-05	0.3147	0.5164
2	6	*ESR1*	M,I,C	rs2207396	5	*HBEGF*	L	rs2237077	N	7.56E-04	5.39E-05	3.42E-04	1.02E-04	0.3299	0.1786
1	19	*PRKD2*		rs314662	2	*ACVR1*		rs7595478	P,N	4.55E-04	3.97E-04	1.31E-04	1.80E-04	0.3615	0.1070
1	22	*IL2RB*	I,L	rs3218295	12	*NOS1*	M,I	rs6490121		5.78E-04	8.22E-04	9.85E-04	4.51E-04	0.3669	0.2129
9	2	*PDE1A*		rs1430154	14	*OSGEP/NP/APEX1*		rs999692	K,N	2.11E-04	1.41E-05	6.99E-04	4.93E-05	0.3758	0.1891
1	3	*RARB*		rs1153589	20	*HNF4A*		rs6031579	P,N	1.42E-04	4.78E-04	3.25E-04	3.91E-04	0.4626	0.5184
2	5	*MCC/TSSK1B*		rs10043783	17	*PRKCA*	M,L,C	rs9896549	P	1.18E-04	9.64E-05	3.10E-05	1.23E-06	0.4790	0.1656

The GWA study was performed in 931 affected trios with multiple sclerosis diagnosed using the McDonald criteria and 2950 control individuals from the Wellcome Trust Case Control Consortium. The replication study was performed in 808 cases with multiple sclerosis and 1720 controls ascertained from Brigham and Women's Hospital. The selected 31 two-locus models had significant model fit and interaction likelihood ratio tests and had consistent direction of effects across both screening and replication phases. They are ranked by replication model fit p-value (only BWH model fit p-value < 0.5 shown). Rows shaded gray indicate significant model fit in both screening phases ($p < 0.001$) and in the replication phase (at $p < 0.05$). Models in boldface have significant model fit and interaction likelihood ratio tests in both screening phases ($p < 0.001$) and in the replication phase (at $p < 0.05$). Number indicates the number of two-SNP models supporting the interaction of these two loci. Gene data sources are coded: E, differential expression ; M, prior MS association ; I, prior immune-related disease association ; L, linkage ; C, selected candidate genes. Gene pairing data sources are coded: P, PFAM ; N, Netpath ; K, KEGG ; R, Reactome ; G, GO.

Discussion

We report on a knowledge-based two-locus analysis of a genome-wide association study of multiple sclerosis that examined biologically plausible combinations of variants in genes across the genome. Using this analysis strategy, we identified several consistent and replicating two-locus models that influence MS susceptibility. Many of these models are functionally related to actin cytoskeletal regulation and the calcium signaling of multiple downstream events, both of which are mediated by inositol-based signaling molecules IP2 and IP3.

The muscarinic acetylcholine receptor M3 (*CHRM3*) is a G-protein coupled receptor which binds its cognate ligand acetylcholine to activate phospholipase C (PLC), generating inositol 1,45-tripohsphate (IP_3). IP_3 binds to the IP_3-receptor to release Ca^{2+} ions from intracellular stores (Furuichi & Mikoshiba, 1995). Calcium signaling triggers a wide variety of downstream events, but notably intracellular calcium levels have been experimentally shown to induce IL-2 production (Mills et al., 1985). *CHRM3* statistically interacts with myosin light-chain kinase (*MYLK*), which is activated downstream of intra-cellular calcium release. *MYLK* mediates myosin II motor activity responsible for actin cytoskelelon contraction, which plays an important role in many cellular functions, including cell spreading, motility, cell division, and focal adhesion. Intracellular calcium levels may contribute to traumatically induced axonal injury by altering cytoskeletal structure and alignment, an subsequently axonal transport(Fitzpatrick, Maxwell, & Graham, 1998). It is plausible, therefore, that the coinheritance of variations in *CHRM3* and *MYLK* alter an acetylcholine-based response to neuronal injury, reducing the capacity to correct structural changes to the cell via cyctoskeleton regulation.

PLCβ1 and *PLCβ4* are two isozymes in the larger phospholipase-C family (Suh et al., 2008). PLC-beta hydrolyses phosphatidylinositol 4,5-bisphosphate (PIP_2) to produce IP_3 and diacylglycerol (DAG). Notably, DAG activates various protein kinase C (PRC) isoforms, and specifically, alleles of the *PRKCA* gene were found to confer increased MS risk in Finnish and Canadian populations by fine-mapping after linkage analysis (Saarela et al., 2006). This locus

may explain the recurrent linkage to 17q20-24 by multiple studies, and corroborates a potential role of this signaling pathway in MS etiology (GAMES & Transatlantic Multiple Sclerosis Genetics Cooperative, 2003; Sawcer et al., 2005). *PLCβ* isoforms show tissue specific expression, and both *PLCβ1* and *PLCβ4* are expressed in the central nervous system, with *PLCβ1* highly expressed in the cerebral cortex and hippocampus (Homma et al., 1989), and *PLCβ4* expressed in the cerebellum and retina(Adamski, Timms, & Shieh, 1999). Model systems also illustrate a role for both isoforms in proper conduction of nerve signals. *PLCβ1* null mice have recurrent seizure attacks that ultimately lead to death near three weeks of growth, indicating an essential role of *PLCβ1* in inhibitory neuronal circuitry (Kim et al., 1997). *PLCβ4* null mice showed smaller cerebellum growth and motor defects likely due to impaired PLC-linked signal transduction (Kim et al., 1997). It is certainly plausible that altered expression or function of these genes could increase the impact of MS lesions on nervous signal conduction or weaken essential repair mechanisms for cells of the central nervous system.

Four genes in the replicating models directly impact PIP2 concentrations. *PLCβ1* and *PLCβ4* phosphorylate PIP2 to produce IP3. *ACTN1* and *SCIN* are both regulated by PIP2. This could indicate a role of intracellular PIP2 or IP3 concentrations in neuronal damage repair, signal transduction, or some other mechanism of MS disease etiology.

Conclusions

Our findings illustrate the utility of applying a knowledge-based approach to GWAS analysis and highlight the potential role of calcium signaled cytoskeletal response in the etiology of multiple sclerosis. While these new findings have small effect sizes and likely explain little of the total genetic-based MS risk, there may be multiple distinct genetic architectures exhibiting both epistasis and genetic heterogeneity that give rise to the same disease phenotype or delineate clinical subtype and disease course. Given this potential complexity, further examination of these

pathways in other clinical populations may reveal multiple combinations of alleles in these functionally related genes that modulate MS risk.

Acknowledgements

This work was conducted as a secondary analysis for the International Multiple Sclerosis Genetics Consortium, and was supported by NS049477 from the National Institute of Neurological Disorders and Stroke.

CHAPTER V

CONCLUSION

The last thirty years of human genetics research has shaped our view of the generic etiology of common disease from a single mechanism driven by a highly influential sequence variant to a complex web of interacting genetic variants and environmental exposures, each contributing a small degree of disease risk. As such, there has also been a renewed interest in epistasis or gene-gene interaction, coupled with a technological push towards study designs that capture nearly all common variation within the human genome. Simultaneously through this period, legions of scientists began to develop and curate databases of genomic and pathway data. The work described in this dissertation attempts to take advantage of the multiple advances in knowledge and technology to both illustrate the principle that incorporating structured biological knowledge into an analysis can be fruitful and to elucidate new mechanisms involved in the development and progression of multiple sclerosis.

Numerous examples of pathway-based mechanisms have been discovered in model organisms and for Mendelian diseases (see Chapter III). In many cases, these mechanisms were discovered by traditional linkage mapping, or by observing segregation patterns when crossing strains. Once the gene or genes responsible for the phenotype were found, the biological details explaining what the gene product is and how it influences the larger biological system to cause the phenotype were uncovered through experimentation, generally with model organisms or cell lines. A knowledge-based approach would not have discovered Mendelian effects, generally because *genetic analysis of a Mendelian trait formed the foundation for discovering the biological system*. In contract, complex disease is likely caused by the combined effect of numerous genetic variants with small independent or epistatic effects. Some of these genetic variants certainly occur in biological systems that are already well characterized experimentally. As such, it is logical to use

the biological systems that we do understand *as a framework for discovering the combined effects of common human variation*. In this manner, human genetic studies can both provide a springboard for experimental discovery of biological systems, and can also use biological systems to aid genetic discovery in human populations.

The genome-wide association studies described in Chapter I have established new risk factors for several complex disorders and phenotypes. In many cases, however, the small genetic effects found by these studies only scratch the surface of the true genetic picture. The methods outlined and described in Chapters II and III attempt to synthesize multiple knowledge sources to intelligently explore the multitude of potential gene-gene interactions that may be influencing common disease.

We have illustrated through a simulation study that patterns of linkage disequilibrium from modest samples can be used to reliably mark the broader genomic region a particular polymorphism represents using LD-Spline, the method described in Chapter II. The LD-Spline method is also a technological advance, in that these regions can be quickly and seamlessly identified within a contained database system. When applied to two popular large-scale genotyping platforms, the LD-Spline approach dramatically increased the number of genes mapped versus more simplistic approaches. In future studies, LD-Spline could be modified, processing HapMap-based statistics to provide a SNP-centric version of the Gabriel et al. approach, which may produce results with a more sound population genetics interpretation.

We have also illustrated the benefit of incorporating gene-gene relationship information into data analysis by real-world application (see Chapter III). Two-SNP models were generated using the Biofilter method, and were overlaid to calculate an implication index, denoting the number of knowledge-bases that support a given model. By applying the Biofilter to a genome-wide association study of multiple sclerosis, we illustrate that models supported by more knowledge sources are more likely to have significant model fit and interaction term statistics. In future work, we hope to apply the Biofilter approach to other phenotypes and determine if the

relationship between the implication index and significant model statistics is a generalizable phenomenon. We also hope to include gene regulatory networks and evolutionary conservation as data sources for the Biofilter method. Also, one of the great potential uses of the Biofilter system is to pose specific hypotheses to test using genome-wide association data. The Biofilter could easily allow quick evaluations of overlapping gene groups (such as genes of the same pathway under significant linkage peaks), which could prove to be an incredibly powerful tool.

Finally, a thorough knowledge-driven analysis of the multiple sclerosis data was conducted (see Chapter IV). In this analysis, we discover two consistent epistasis models that replicate in an independent dataset. A collection of sixteen other models had significant model fit statistics in the replication set, but lacked a significant interaction term. When viewed as a whole, many of these results fall into a collection of three inter-related pathways, and several genes involved in these pathways are regulated by or help metabolize inositol-based signaling molecules. This could suggest a potential new mechanism for multiple sclerosis etiology, and nicely illustrates the benefit of knowledge-based analysis. In the near future, we hope to acquire additional replication sets to further evaluate these interacting models.

The work outlined here has established a process for discovering what patterns and processes might cause epistasis, produced a collection of tools that transform a popular agnostic genetic study design into a more precise, knowledge-based study, and illustrated the utility of that analysis for a complex disease. Much more, however, could be done to systematically integrate established biological knowledge into genome-wide association analysis.

While thoughtfully executed, each of the bioinformatic steps in the Biofilter procedure has distinct challenges in interpretation and definition. For example, the Ensembl database defines a gene as the 3′ to 5′ region that is transcribed and spliced into an mRNA. Gene regulatory regions (promoters, transcription factors, etc.) likely contain variants that alter gene expression, potentially with functional consequences, and those regions are not represented in the Ensembl gene definition. Databases that identify known regulatory elements or points of

multi-species conservation could provide additional information about sequence regions surrounding the Ensembl gene definition that should be represented in our analysis pipeline. Expanding the definition of a "gene" will likely enhance the ability to detect genetic effects in complex disease, since influential variants probably have minor effects on the function of a gene product, such as altering functional quantity or an altered splice pattern.

In addition to improving informatics, new genetic data types are on the research horizon. Whole-genome sequence will likely be cost-effective for large sample sizes in the near future, which will collect data on an unprecedented scale. Storage and analysis of sequence data will *require* large online collaborative database systems. Also, whole genome sequencing is generally conducted with an alternative hypothetical disease model: common diseases are caused (in part) by multiple rare variants in the population. Under this disease model, convincing evidence of a statistical association will require either enormous sample sizes, or some type of procedure to "bin" multiple different sequence variants into a common category that can be associated to disease. The most logical approach to binning these rare variants is to assess functional relevance using biological knowledge. Also, these rare variants may be dispersed across multiple genes of a common pathway that all ultimately share the same functional consequence. In this case, pathway-based analyses, followed by careful molecular studies, will be critical to elucidating the true nature of rare variant disease effects.

In addition to sequencing, methods for assaying structural variation and methylation patterns in the human genome are becoming more high-throughput. As these genetic mechanisms may be less stable than simple sequence variations in the population (i.e. de novo copy number variations, fetal environment-based methylation, etc), building convincing evidence of an association requires methods for addressing heterogeneity of effect, similar to the problems with rare variants mentioned above.

Similar technological advances are occurring in structural biology, pharmacology, and even environmental data collection for epidemiological studies, among many other fields and

disciplines, providing enormous rich datasets for the study of human disease. Because high-throughput data is inherently more structured that single experiments, database systems that can incorporate and cross-reference multiple data types will exponentially expand the set of experimental questions an investigator can ask with a collection of data.

High-throughput technologies like these temporarily shift the paradigm of science away from explicit hypothesis testing and toward establishing and refining *models of biological systems*. If complex disease processes are truly driven by multiple interacting genetic and environmental components, then synthesizing masses of collected data into a model of the system will be the best way to fully utilize all of the available information. System models will provide a context for interpreting and analyzing new data, and will provide investigators with a renewed framework to test explicit hypotheses. Despite changes in funding mechanisms, technology, and a shift toward large collaborative projects, human genetics will remain firmly rooted in the same rigors of basic science and the scientific method. Large-scale high-throughput science is simply accelerating the pace at which we provide a systemic *context* to experimental questions.

REFERENCES

Adamski, F. M., Timms, K. M., and Shieh, B. H. (1999) A unique isoform of phospholipase Cbeta4 highly expressed in the cerebellum and eye. *Biochim Biophys Acta*, **1444**, 55-60.

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55-55.

Akamizu, T., Sale, M. M., Rich, S. S., Hiratani, H., Noh, J. Y., Kanamoto, N., Saijo, M., Miyamoto, Y., Saito, Y., Nakao, K., and Bowden, D. W. (2000) Association of autoimmune thyroid disease with microsatellite markers for the thyrotropin receptor gene and CTLA-4 in Japanese patients. *Thyroid*, **10**, 851-858.

Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H., and Wjst, M. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet*, **69**, 936-950.

Amisten, S., Braun, O. O., Bengtsson, A., and Erlinge, D. (2008) Gene expression profiling for the identification of G-protein coupled receptors in human platelets. *Thromb Res*, **122**, 47-57.

Amos, C. I. (2007) Successful design and conduct of genome-wide association studies. *Hum Mol Genet*, **16 Spec No. 2**, R220-R225.

Anderson, D. W., Ellenberg, J. H., Leventhal, C. M., Reingold, S. C., Rodriguez, M., and Silberberg, D. H. (1992) Revised estimate of the prevalence of multiple sclerosis in the United States. *Ann Neurol*, **31**, 333-336.

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., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.

Aubert, J., Bar-Hen, A., Daudin, J. J., and Robin, S. (2004) Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, **5**, 125-125.

Barrett, J. C. and Cardon, L. R. (2006) Evaluating coverage of genome-wide association studies. *Nat Genet*, **38**, 659-662.

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

Bastone, L., Reilly, M., Rader, D. J., and Foulkes, A. S. (2004) MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. *Hum Hered*, **58**, 82-92.

Bateson, W. (1909) Mendel's Principles of Heredity.

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

Beltrao, P., Kiel, C., and Serrano, L. (2007) Structures in systems biology. *Curr Opin Struct Biol*, **17**, 378-384.

Berridge, M. J., Bootman, M. D., and Roderick, H. L. (2003) Calcium signalling: dynamics, homeostasis and remodelling. *Nat Rev Mol Cell Biol*, **4**, 517-529.

Bomprezzi, R., Ringner, M., Kim, S., Bittner, M. L., Khan, J., Chen, Y., Elkahloun, A., Yu, A., Bielekova, B., Meltzer, P. S., Martin, R., McFarland, H. F., and Trent, J. M. (2003) Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Hum Mol Genet*, **12**, 2191-2199.

Borecki, I. B. and Province, M. A. (2008) Linkage and association: basic concepts. *Adv Genet*, **60**, 51-74.

Brassat, D., Motsinger, A. A., Caillier, S. J., Erlich, H. A., Walker, K., Steiner, L. L., Cree, B. A., Barcellos, L. F., Pericak-Vance, M. A., Schmidt, S., Gregory, S., Hauser, S. L., Haines, J. L., Oksenberg, J. R., and Ritchie, M. D. (2006) Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. *Genes Immun*, **7**, 310-315.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) Classification and Regression Trees.

Bridges, C. B. (1919) Specific modifiers of eosin eye color in Drosophila melanogaster. *J Experimental Zoology*, **28**, 337-384.

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2006) Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics*, **22**, 2173-2174.

Bush, W. S., Thornton-Wells, T. A., and Ritchie, M. D. (2007) Association Rule Discovery Has the Ability to Model Complex Genetic Effects. *Computational Intelligence and Data Mining, 2007 CIDM 2007 IEEE Symposium on*, 624-629.

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

Carrasquillo, M. M., McCallion, A. S., Puffenberger, E. G., Kashuk, C. S., Nouri, N., and Chakravarti, A. (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet*, **32**, 237-244.

Chanda, P., Sucheston, L., Zhang, A., Brazeau, D., Freudenheim, J. L., Ambrosone, C., and Ramanathan, M. (2008) AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics*, **180**, 1191-1210.

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., 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., and Collins, F. S. (2007) Replicating genotype-phenotype associations. *Nature*, **447**, 655-660.

Chataway, J., Feakes, R., Coraddu, F., Gray, J., Deans, J., Fraser, M., Robertson, N., Broadley, S., Jones, H., Clayton, D., Goodfellow, P., Sawcer, S., and Compston, A. (1998) The genetics of multiple sclerosis: principles, background and updated results of the United Kingdom systematic genome screen. *Brain*, **121 (Pt 10)**, 1869-1887.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2002) Bayesian treed models. *Machine Learning*, **48**, 299-320.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998) Bayesian CART model search. *Journal of the American Statistical Association*, **93**, 935-948.

Cohen, J. (1968) Weighted Kappa - Nominal Scale Agreement with Provision for Scaled Disagreement Or Partial Credit. *Psychological Bulletin*, **70**, 213-213.

Collaborative Linkage Study of Autism (2001) Incorporating language phenotypes strengthens evidence of linkage to autism. *Am J Med Genet*, **105**, 539-547.

Collaco, J. M. and Cutting, G. R. (2008) Update on gene modifiers in cystic fibrosis. *Curr Opin Pulm Med*, **14**, 559-566.

Comabella, M. and Martin, R. (2007) Genomics in multiple sclerosis--current state and future directions. *J Neuroimmunol*, **187**, 1-8.

Cook, N. R., Zee, R. Y., and Ridker, P. M. (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med*, **23**, 1439-1453.

Cordell, H. J., Barratt, B. J., and Clayton, D. G. (2004) Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol*, **26**, 167-185.

Curtis, D., Vine, A. E., and Knight, J. (2007) A pragmatic suggestion for dealing with results for candidate genes obtained from genome wide association studies. *BMC Genet*, **8**, 20-20.

Deeb, S. S., Fajas, L., Nemoto, M., Pihlajamaki, J., Mykkanen, L., Kuusisto, J., Laakso, M., Fujimoto, W., and Auwerx, J. (1998) A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat Genet*, **20**, 284-287.

Devlin, B. and Risch, N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311-322.

Donfack, J., Kogut, P., Forsythe, S., Solway, J., and Ober, C. (2003) Sequence variation in the promoter region of the cholinergic receptor muscarinic 3 gene and asthma and atopy. *J Allergy Clin Immunol*, **111**, 527-532.

Eaton, W. A. and Hofrichter, J. (1987) Hemoglobin S gelation and sickle cell disease. *Blood*, **70**, 1245-1266.

Eberle, M. A., Ng, P. C., Kuhn, K., Zhou, L., Peiffer, D. A., Galver, L., Viaud-Martinez, K. A., Lawley, C. T., Gunderson, K. L., Shen, R., and Murray, S. S. (2007) Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet*, **3**, 1827-1837.

Eberle, M. A., Rieder, M. J., Kruglyak, L., and Nickerson, D. A. (2006) Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. *PLoS Genet*, **2**, e142-e142.

Ebers, G. C. (1996) Genetic epidemiology of multiple sclerosis. *Curr Opin Neurol*, **9**, 155-158.

Edwards, A. O., Ritter, R., Abel, K. J., Manning, A., Panhuysen, C., and 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., and Ritchie, M. D. (2008) Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA. 24-35.

Fallin, D. and Schork, N. J. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet*, **67**, 947-959.

Falush, D., Stephens, M., and Pritchard, J. K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567-1587.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*, **34**, D247-D251.

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., and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res*, **36**, D281-D288.

Fisher, R. A. (1918) The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399-433.

Fitzpatrick, M. O., Maxwell, W. L., and Graham, D. I. (1998) The role of the axolemma in the initiation of traumatically induced axonal injury. *J Neurol Neurosurg Psychiatry*, **64**, 285-287.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson,

D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. (2008) Ensembl 2008. *Nucleic Acids Res*, **36**, D707-D714.

Flores, C., Ma, S. F., Maresso, K., Ober, C., and Garcia, J. G. (2007) A variant of the myosin light chain kinase gene is associated with severe asthma in African Americans. *Genet Epidemiol*, **31**, 296-305.

Franke, L., van, B. H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, **78**, 1011-1025.

Frankel, W. N. and Schork, N. J. (1996) Who's afraid of epistasis? *Nat Genet*, **14**, 371-373.

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

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A. M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M. R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N., Doney, A. S., Morris, A. D., Smith, G. D., Hattersley, A. T., and McCarthy, M. I. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889-894.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R.,

Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-861.

Frohman, E. M., Filippi, M., Stuve, O., Waxman, S. G., Corboy, J., Phillips, J. T., Lucchinetti, C., Wilken, J., Karandikar, N., Hemmer, B., Monson, N., De, K. J., Hartung, H., Steinman, L., Oksenberg, J. R., Cree, B. A., Hauser, S., and Racke, M. K. (2005) Characterizing the mechanisms of progression in multiple sclerosis: evidence and new hypotheses for future directions. *Arch Neurol*, **62**, 1345-1356.

Furuichi, T. and Mikoshiba, K. (1995) Inositol 1, 4, 5-trisphosphate receptor-mediated Ca2+ signaling in the brain. *J Neurochem*, **64**, 953-960.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., Defelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225-2229.

GAMES and Transatlantic Multiple Sclerosis Genetics Cooperative (2003) A meta-analysis of whole genome linkage screens in multiple sclerosis. *J Neuroimmunol*, **143**, 39-46.

Gao, L., Grant, A. V., Rafaels, N., Stockton-Porter, M., Watkins, T., Gao, P., Chi, P., Munoz, M., Watson, H., Dunston, G., Togias, A., Hansel, N., Sevransky, J., Maloney, J. P., Moss, M., Shanholtz, C., Brower, R., Garcia, J. G., Grigoryev, D. N., Cheadle, C., Beaty, T. H., Mathias, R. A., and Barnes, K. C. (2007) Polymorphisms in the myosin light chain

kinase gene that confer risk of severe sepsis are associated with a lower risk of asthma. *J Allergy Clin Immunol*, **119**, 1111-1118.

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

Gregory, S. G., Schmidt, S., Seth, P., Oksenberg, J. R., Hart, J., Prokop, A., Caillier, S. J., Ban, M., Goris, A., Barcellos, L. F., Lincoln, R., McCauley, J. L., Sawcer, S. J., Compston, D. A., Dubois, B., Hauser, S. L., Garcia-Blanco, M. A., Pericak-Vance, M. A., and Haines, J. L. (2007) Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat Genet*, **39**, 1083-1091.

Hageman, G. S., Anderson, D. H., Johnson, L. V., Hancox, L. S., Taiber, A. J., Hardisty, L. I., Hageman, J. L., Stockman, H. A., Borchardt, J. D., Gehrs, K. M., Smith, R. J., Silvestri, G., Russell, S. R., Klaver, C. C., Barbazetto, I., Chang, S., Yannuzzi, L. A., Barile, G. R., Merriam, J. C., Smith, R. T., Olsh, A. K., Bergeron, J., Zernant, J., Merriam, J. E., Gold, B., Dean, M., and Allikmets, R. (2005) A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc Natl Acad Sci U S A*, **102**, 7227-7232.

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., and Pericak-Vance, M. A. (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science*, **308**, 419-421.

Haines, J. L., Terwedow, H. A., Burgess, K., Pericak-Vance, M. A., Rimmler, J. B., Martin, E. R., Oksenberg, J. R., Lincoln, R., Zhang, D. Y., Banatao, D. R., Gatto, N., Goodkin, D. E., and Hauser, S. L. (1998) Linkage of the MHC to familial multiple sclerosis suggests genetic heterogeneity. The Multiple Sclerosis Genetics Group. *Hum Mol Genet*, **7**, 1229-1234.

Hao, K., Di, X., and Cawley, S. (2007) LdCompare: rapid computation of single- and multiple-marker r2 and genetic coverage. *Bioinformatics*, **23**, 252-254.

Hindorff, L. A., Junkins, H. A., and Manolio, T. A. (2008) A Catalog of Published Genome-Wide Association Studies. *NHGRI*,

Hiratani, H., Bowden, D. W., Ikegami, S., Shirasawa, S., Shimizu, A., Iwatani, Y., and Akamizu, T. (2005) Multiple SNPs in intron 7 of thyrotropin receptor are associated with Graves' disease. *J Clin Endocrinol Metab*, **90**, 2898-2903.

Hirschhorn, J. N. and Daly, M. J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, **6**, 95-108.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002) A comprehensive review of genetic association studies. *Genet Med*, **4**, 45-61.

Homma, Y., Takenawa, T., Emori, Y., Sorimachi, H., and Suzuki, K. (1989) Tissue- and cell type-specific expression of mRNAs for four types of inositol phospholipid-specific phospholipase C. *Biochem Biophys Res Commun*, **164**, 406-412.

Hosmer, D. W. and Lemeshow, S. (2000) Applied Logistic Regression. **2**,

Huang, J. and Manning, B. D. (2008) The TSC1-TSC2 complex: a molecular switchboard controlling cell growth. *Biochem J*, **412**, 179-190.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007) Ensembl 2007. *Nucleic Acids Res*, **35**, D610-D617.

Ingram, V. M. (1957) Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature*, **180**, 326-328.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299-1320.

Jansen, R. C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet*, **4**, 145-151.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de, B. B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, **33**, D428-D432.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, **36**, D480-D484.

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

Kanehisa, M., Goto, S., Hattori, M., oki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**, D354-D357.

Kenealy, S. J., Pericak-Vance, M. A., and Haines, J. L. (2003) The genetic epidemiology of multiple sclerosis. *J Neuroimmunol*, **143**, 7-12.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L. C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, 1073-1080.

Kim, D., Jun, K. S., Lee, S. B., Kang, N. G., Min, D. S., Kim, Y. H., Ryu, S. H., Suh, P. G., and Shin, H. S. (1997) Phospholipase C isozymes selectively couple to specific neurotransmitter receptors. *Nature*, **389**, 290-293.

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., and Hoh, J. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385-389.

Knowler, W. C., Bennett, P. H., Hamman, R. F., and Miller, M. (1978) Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *Am J Epidemiol*, **108**, 497-505.

Kooperberg, C., Ruczinski, I., LeBlanc, M. L., and Hsu, L. (2001) Sequence analysis using logic regression. *Genet Epidemiol*, **21 Suppl 1**, S626-S631.

Krailo, M. D. and Pike, M. C. (1984) Algorithm AS 196: Conditional multivariate logistic analysis of stratified case-control studies. *Applied Statistics*, **33**, 95-103.

Lancefield, D. E. (1918) An autosomal bristle modifier affecting a sex-linked character. *American Naturalist*, **52**, 462-464.

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., and Thomas, D. C. (2007) Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol*, **31**, 871-882.

Lewontin, R. C. and Kojima, K. i. (2001) The Evolutionary Dynamics of Complex Polymorphisms. *Evolution*, **14**, 458-472.

Li, C., Li, M., Lange, E. M., and Watanabe, R. M. (2008) Prioritized subset analysis: improving power in genome-wide association studies. *Hum Hered*, **65**, 129-141.

Lloyd, V., Ramaswami, M., and Kramer, H. (1998) Not just pretty eyes: Drosophila eye-colour mutations and lysosomal delivery. *Trends Cell Biol*, **8**, 257-259.

Lock, C., Hermans, G., Pedotti, R., Brendolan, A., Schadt, E., Garren, H., Langer-Gould, A., Strober, S., Cannella, B., Allard, J., Klonowski, P., Austin, A., Lad, N., Kaminski, N., Galli, S. J., Oksenberg, J. R., Raine, C. S., Heller, R., and Steinman, L. (2002) Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nat Med*, **8**, 500-508.

Lou, X. Y., Chen, G. B., Yan, L., Ma, J. Z., Mangold, J. E., Zhu, J., Elston, R. C., and Li, M. D. (2008) A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am J Hum Genet*, **83**, 457-467.

Lowis, G. W. (1990) The social epidemiology of multiple sclerosis. *Sci Total Environ*, **90**, 163-190.

Lundmark, F., Duvefelt, K., Iacobaeus, E., Kockum, I., Wallstrom, E., Khademi, M., Oturai, A., Ryder, L. P., Saarela, J., Harbo, H. F., Celius, E. G., Salter, H., Olsson, T., and Hillert, J. (2007) Variation in interleukin 7 receptor alpha chain (IL7R) influences risk of multiple sclerosis. *Nat Genet*, **39**, 1108-1113.

Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008) A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*, **118**, 1590-1605.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, **37**, 413-417.

Martin, E. R., Ritchie, M. D., Hahn, L., Kang, S., and Moore, J. H. (2006) A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genet Epidemiol*, **30**, 111-123.

Mattila, K. M., Luomala, M., Lehtimaki, T., Laippala, P., Koivula, T., and Elovaara, I. (2001) Interaction between ESR1 and HLA-DR2 may contribute to the development of MS in women. *Neurology*, **56**, 1246-1247.

McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H. P., Lublin, F. D., McFarland, H. F., Paty, D. W., Polman, C. H., Reingold, S. C., Sandberg-Wollheim, M., Sibley, W., Thompson, A., van den, N. S., Weinshenker, B. Y., and Wolinsky, J. S. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol*, **50**, 121-127.

McKinney, B. A., Reif, D. M., White, B. C., Crowe, J. E., and Moore, J. H. (2007) Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics*, **23**, 2113-2120.

Meldrum, E., Parker, P. J., and Carozzi, A. (1991) The PtdIns-PLC superfamily and signal transduction. *Biochim Biophys Acta*, **1092**, 49-71.

Merlo, C. A. and Boyle, M. P. (2003) Modifier genes in cystic fibrosis lung disease. *J Lab Clin Med*, **141**, 237-241.

Mills, G. B., Cheung, R. K., Grinstein, S., and Gelfand, E. W. (1985) Increase in cytosolic free calcium concentration is an intracellular messenger for the production of interleukin 2 but not for expression of the interleukin 2 receptor. *J Immunol*, **134**, 1640-1643.

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

Moore, J. H., Barney, N., Tsai, C. T., Chiang, F. T., Gui, J., and White, B. C. (2007) Symbolic modeling of epistasis. *Hum Hered*, **63**, 120-133.

Moore, J. H., Hahn, L. W., Ritchie, M. D., Thornton, T. A., and White, B. C. (2004) Routine discovery of complex genetic models using genetic algorithms. *Applied Soft Computing*, **4**, 79-86.

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

Moore, J. H. and Ritchie, M. D. (2004) STUDENTJAMA. The challenges of whole-genome approaches to common diseases. *JAMA*, **291**, 1642-1643.

Moore, J. H. and White, B. C. (2007) Tuning Relief-F for Genome-Wide Genetic Analysis. 166-175.

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

Morton, N. E. (2008) Into the post-HapMap era. *Adv Genet*, **60**, 727-742.

Motsinger, A. A., Brassat, D., Caillier, S. J., Erlich, H. A., Walker, K., Steiner, L. L., Barcellos, L. F., Pericak-Vance, M. A., Schmidt, S., Gregory, S., Hauser, S. L., Haines, J. L., Oksenberg, J. R., and Ritchie, M. D. (2007) Complex gene-gene interactions in multiple sclerosis: a multifactorial approach reveals associations with inflammatory genes. *Neurogenetics*, **8**, 11-20.

Motsinger, A. A., Lee, S. L., Mellick, G., and 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**, 39-39.

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

Motsinger-Reif, A. A. and Ritchie, M. D. (2008) Neural networks for genetic epidemiology: past, present, and future. *BioData Min*, **1**, 3-3.

Mumford, C. J., Wood, N. W., Kellar-Wood, H., Thorpe, J. W., Miller, D. H., and Compston, D. A. (1994) The British Isles survey of multiple sclerosis in twins. *Neurology*, **44**, 11-15.

Myers, J. K., Beihoffer, L. A., and Sanders, C. R. (2005) Phenotology of disease-linked proteins. *Hum Mutat*, **25**, 90-97.

Nagel, R. L. (2005) Epistasis and the genetics of human diseases. *C R Biol*, **328**, 606-615.

Nagel, R. L. (2001) Pleiotropic and epistatic effects in sickle cell anemia. *Curr Opin Hematol*, **8**, 105-110.

National Center for Health and Statistics (2008) Leading Causes of Death. *Center for Disease Control*,

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res*, **15**, 1566-1575.

Niino, M., Kikuchi, S., Fukazawa, T., Yabe, I., and Tashiro, K. (2000) Estrogen receptor gene polymorphism in Japanese patients with multiple sclerosis. *J Neurol Sci*, **179**, 70-75.

Ohmori, M., Ohta, M., Shimura, H., Shimurat, Y., Suzuki, K., and Kohn, L. D. (1996) Cloning of the single strand DNA-binding protein important for maximal expression and thyrotropin (TSH)-induced negative regulation of the TSH receptor. *Mol Endocrinol*, **10**, 1407-1424.

Oksenberg, J. R., Baranzini, S. E., Barcellos, L. F., and Hauser, S. L. (2001) Multiple sclerosis: genomic rewards. *J Neuroimmunol*, **113**, 171-184.

Oksenberg, J. R. and Barcellos, L. F. (2005) Multiple sclerosis genetics: leaving no stone unturned. *Genes Immun*, **6**, 375-387.

Oturai, A., Larsen, F., Ryder, L. P., Madsen, H. O., Hillert, J., Fredrikson, S., Sandberg-Wollheim, M., Laaksonen, M., Koch-Henriksen, N., Sawcer, S., Fugger, L., Sorensen, P. S., and Svejgaard, A. (1999) Linkage and association analysis of susceptibility regions on chromosomes 5 and 6 in 106 Scandinavian sibling pair families with multiple sclerosis. *Ann Neurol*, **46**, 612-616.

Pan, W. (2008) Network-based model weighting to detect multiple loci influencing complex diseases. *Hum Genet*, **124**, 225-234.

Pandey, L. and Institute of Bioinformatics (2008) NetPath. *Pandey Lab*,

Park, M. Y. and Hastie, T. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**, 30-50.

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

Pauling, L. and Itano, H. A. (1949) Sickle cell anemia a molecular disease. *Science*, **110**, 543-548.

Pearson, T. A. and Manolio, T. A. (2008) How to interpret a genome-wide association study. *JAMA*, **299**, 1335-1344.

Pericak-Vance, M. A., Rimmler, J. B., Martin, E. R., Haines, J. L., Garcia, M. E., Oksenberg, J. R., Barcellos, L. F., Lincoln, R., Goodkin, D. E., and Hauser, S. L. (2001) Linkage and

association analysis of chromosome 19q13 in multiple sclerosis. *Neurogenetics*, **3**, 195-201.

Polman, C. H., Reingold, S. C., Edan, G., Filippi, M., Hartung, H. P., Kappos, L., Lublin, F. D., Metz, L. M., McFarland, H. F., O'Connor, P. W., Sandberg-Wollheim, M., Thompson, A. J., Weinshenker, B. G., and Wolinsky, J. S. (2005) Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Ann Neurol*, **58**, 840-846.

Povey, S., Burley, M. W., Attwood, J., Benham, F., Hunt, D., Jeremiah, S. J., Franklin, D., Gillett, G., Malas, S., Robson, E. B., and . (1994) Two loci for tuberous sclerosis: one on 9q34 and one on 16p13. *Ann Hum Genet*, **58**, 107-127.

Powell, I. J. (1998) Prostate cancer in the African American: is this a different disease? *Semin Urol Oncol*, **16**, 221-226.

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

Province, M. A. and Borecki, I. B. (2008) Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans. *Pac Symp Biocomput*, 190-200.

Purcell, S. PLINK 1.01.

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., and Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559-575.

Reich, D. E. and Lander, E. S. (2001) On the allelic spectrum of human disease. *Trends Genet*, **17**, 502-510.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and 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., Bush, W. S., Coffey, C. S., and Moore, J. H. (2007) Genetic programming neural networks: A powerful bioinformatics tool for human genetics. *Applied Soft Computing*, **7**, 471-479.

Robertson, N. P., Fraser, M., Deans, J., Clayton, D., Walker, N., and Compston, D. A. (1996) Age-adjusted recurrence risks for relatives of patients with multiple sclerosis. *Brain*, **119 (Pt 2)**, 449-455.

Rouhani-Kalleh, O. (2007) Algorithms for Fast Large Scale Data Mining Using Logistic Regression. *Computational Intelligence and Data Mining, 2007 CIDM 2007 IEEE Symposium on*, 155-162.

Saarela, J., Kallio, S. P., Chen, D., Montpetit, A., Jokiaho, A., Choi, E., Asselta, R., Bronnikov, D., Lincoln, M. R., Sadovnick, A. D., Tienari, P. J., Koivisto, K., Palotie, A., Ebers, G. C., Hudson, T. J., and Peltonen, L. (2006) PRKCA and multiple sclerosis: association in two independent populations. *PLoS Genet*, **2**, e42-e42.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L.,

Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., and Peterson, J. L. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913-918.

Sadovnick, A. D. and Ebers, G. C. (1995) Genetics of multiple sclerosis. *Neurol Clin*, **13**, 99-118.

Sarkijarvi, S., Kuusisto, H., Paalavuo, R., Levula, M., Airla, N., Lehtimaki, T., Kaprio, J., Koskenvuo, M., and Elovaara, I. (2006) Gene expression profiles in Finnish twins with multiple sclerosis. *BMC Med Genet*, **7**, 11-11.

Sawcer, S., Ban, M., Maranian, M., Yeo, T. W., Compston, A., Kirby, A., Daly, M. J., De Jager, P. L., Walsh, E., Lander, E. S., Rioux, J. D., Hafler, D. A., Ivinson, A., Rimmler, J., Gregory, S. G., Schmidt, S., Pericak-Vance, M. A., Akesson, E., Hillert, J., Datta, P., Oturai, A., Ryder, L. P., Harbo, H. F., Spurkland, A., Myhr, K. M., Laaksonen, M., Booth, D., Heard, R., Stewart, G., Lincoln, R., Barcellos, L. F., Hauser, S. L., Oksenberg, J. R., Kenealy, S. J., and Haines, J. L. (2005) A high-density screen for linkage in multiple sclerosis. *Am J Hum Genet*, **77**, 454-467.

Schmegner, C., Hoegel, J., Vogel, W., and Assum, G. (2005) Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the European population. *Hum Genet*, **118**, 276-286.

Segal, R. B. (1998) Machine Learning as Massive Search.

Shao, H. (2008) Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis.

Shao, Y., Raiford, K. L., Wolpert, C. M., Cope, H. A., Ravan, S. A., shley-Koch, A. A., Abramson, R. K., Wright, H. H., DeLong, R. G., Gilbert, J. R., Cuccaro, M. L., and Pericak-Vance, M. A. (2002) Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. *Am J Hum Genet*, **70**, 1058-1061.

Siegmund, K. D., Langholz, B., Kraft, P., and Thomas, D. C. (2000) Testing linkage disequilibrium in sibships. *Am J Hum Genet*, **67**, 244-248.

Sinclair, C., Kirk, J., Herron, B., Fitzgerald, U., and McQuaid, S. (2007) Absence of aquaporin-4 expression in lesions of neuromyelitis optica but increased expression in multiple sclerosis lesions and normal-appearing white matter. *Acta Neuropathol*, **113**, 187-194.

Slatkin, M. (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, **9**, 477-485.

Spanakis, E., Milord, E., and Gragnoli, C. (2008) AVPR2 variants and mutations in nephrogenic diabetes insipidus: review and missense mutation significance. *J Cell Physiol*, **217**, 605-617.

Steyn, N. P., Mann, J., Bennett, P. H., Temple, N., Zimmet, P., Tuomilehto, J., Lindstrom, J., and Louheranta, A. (2004) Diet, nutrition and the prevention of type 2 diabetes. *Public Health Nutr*, **7**, 147-165.

Stuart, M. J. and Nagel, R. L. (2004) Sickle-cell disease. *Lancet*, **364**, 1343-1360.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, **23**, 3251-3253.

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

Suh, P. G., Park, J. I., Manzoli, L., Cocco, L., Peak, J. C., Katan, M., Fukami, K., Kataoka, T., Yun, S., and Ryu, S. H. (2008) Multiple roles of phosphoinositide-specific phospholipase C isozymes. *BMB Rep*, **41**, 415-434.

Suter, B., Kittanakom, S., and Stagljar, I. (2008) Two-hybrid technologies in proteomics research. *Curr Opin Biotechnol*, **19**, 316-323.

Swanberg, M., Lidman, O., Padyukov, L., Eriksson, P., Akesson, E., Jagodic, M., Lobell, A., Khademi, M., Borjesson, O., Lindgren, C. M., Lundman, P., Brookes, A. J., Kere, J., Luthman, H., Alfredsson, L., Hillert, J., Klareskog, L., Hamsten, A., Piehl, F., and Olsson, T. (2005) MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. *Nat Genet*, **37**, 486-494.

Tahri-Daizadeh, N., Tregouet, D. A., Nicaud, V., Manuel, N., Cambien, F., and Tiret, L. (2003) Automated detection of informative combined effects in genetic association studies of complex traits. *Genome Res*, **13**, 1952-1960.

Thakkinstian, A., Han, P., McEvoy, M., Smith, W., Hoh, J., Magnusson, K., Zhang, K., and Attia, J. (2006) Systematic review and meta-analysis of the association between complement

factor H Y402H polymorphisms and age-related macular degeneration. *Hum Mol Genet*, **15**, 2784-2790.

The International Multiple Sclerosis Genetics Consortium (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med*, **357**, 851-862.

Thomas, P. D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A*, **101**, 15398-15403.

Thornton-Wells, T. A., Moore, J. H., and Haines, J. L. (2006) Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data. *BMC Bioinformatics*, **7**, 204-204.

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

Tian, C., Breyer, R. M., Kim, H. J., Karra, M. D., Friedman, D. B., Karpay, A., and Sanders, C. R. (2005) Solution NMR spectroscopy of the human vasopressin V2 receptor, a G protein-coupled receptor. *J Am Chem Soc*, **127**, 8010-8011.

Torkamani, A., Topol, E. J., and Schork, N. J. (2008) Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, **92**, 265-272.

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., and Stein, L. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*, **8**, R39-R39.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol*, **4**, e72-e72.

Wang, K., Li, M., and Bucan, M. (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet*, **81**,

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet*, **71**, 1227-1234.

Weir, B. S. (1979) Inferences about linkage disequilibrium. *Biometrics*, **35**, 235-254.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G. (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res*, **15**, 1468-1476.

Wickens, T. D. (1989) Multiway contingency tables analysis for the social sciences.

Wise, A., Gearing, K., and Rees, S. (2002) Target validation of G-protein coupled receptors. *Drug Discov Today*, **7**, 235-246.

Wooster, R. and Weber, B. L. (2003) Breast and ovarian cancer. *N Engl J Med*, **348**, 2339-2347.

Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M., and Eisenberg, D. (2001) DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res*, **29**, 239-241.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res*, **28**, 289-291.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and 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.

Young, J. and Povey, S. (1998) The genetic basis of tuberous sclerosis. *Mol Med Today*, **4**, 313-319.

Zareparsi, S., Branham, K. E., Li, M., Shah, S., Klein, R. J., Ott, J., Hoh, J., Abecasis, G. R., and Swaroop, A. (2005) Strong association of the Y402H variant in complement factor H at 1q32 with susceptibility to age-related macular degeneration. *Am J Hum Genet*, **77**, 149-153.

Zaykin, D. V. and Zhivotovsky, L. A. (2005) Ranks of genuine associations in whole-genome scans. *Genetics*, **171**, 813-823.

Zhang, W., Collins, A., Gibson, J., Tapper, W. J., Hunt, S., Deloukas, P., Bentley, D. R., and Morton, N. E. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc Natl Acad Sci U S A*, **101**, 18075-18080.

Zielenski, J. (2000) Genotype and phenotype in cystic fibrosis. *Respiration*, **67**, 117-133.