NOVEL EXTENSIONS OF LABEL PROPAGATION FOR BIOMARKER DISCOVERY IN GENOMIC DATA

by

Matthew E. Stokes

B.S. Systems and Control Engineering, Case Western Reserve University, 2008

M.S. Intelligent Systems Program / Biomedical Informatics, University of Pittsburgh 2011

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation proposal was presented

by

Matthew E. Stokes

on

July 17, 2014

and approved by

M. Michael Barmada, PhD, Department of Human Genetics

Gregory F. Cooper, MD, PhD, Department of Biomedical Informatics and the Intelligent Systems Program

Milos Hauskrecht, PhD, Department of Computer Science and the Intelligent Systems Program

Dissertation Advisor: Shyam Visweswaran, MD, PhD, Department of Biomedical Informatics and the Intelligent Systems Program

NOVEL EXTENSIONS OF LABEL PORPAGATION FOR BIOMARKER DISCOVERY IN GENOMIC DATA

Matthew E. Stokes, M.S

University of Pittsburgh, 2014

NOVEL EXTENSIONS OF LABEL PROPAGATION FOR BIOMARKER DISCOVERY IN GENOMIC DATA

Matthew E. Stokes, PhD

University of Pittsburgh, 2014

One primary goal of analyzing genomic data is the identification of biomarkers which may be causative of, correlated with, or otherwise biologically relevant to disease phenotypes. In this work, I implement and extend a multivariate feature ranking algorithm called label propagation (LP) for biomarker discovery in genome-wide single-nucleotide polymorphism (SNP) data. This graph-based algorithm utilizes an iterative propagation method to efficiently compute the strength of association between a SNP and a phenotype.

I developed three extensions to the LP algorithm, with the goal of tailoring it to genomic data. The first extension is a modification to the LP score which yields a variable-level score for each SNP, rather than a score for each SNP genotype. The second extension incorporates prior biological knowledge that is encoded as a prior value for each SNP. The third extension enables the combination of rankings produced by LP and another feature ranking algorithm.

The LP algorithm, its extensions, and two control algorithms (chi squared and sparse logistic regression) were applied to 11 genomic datasets, including a synthetic dataset, a semi-synthetic dataset, and nine genome-wide association study (GWAS) datasets covering eight diseases. The quality of each feature ranking algorithm was evaluated by using a subset of top-ranked SNPs to construct a classifier, whose predictive power was evaluated in terms of the area under the Receiver Operating Characteristic curve. Top-ranked SNPs were also evaluated for prior evidence of being associated with disease using evidence from the literature.

The LP algorithm was found to be effective at identifying predictive and biologically meaningful SNPs. The single-score extension performed significantly better than the original algorithm on the GWAS datasets. The prior knowledge extension did not improve on the feature ranking results, and in some cases it reduced the predictive power of top-ranked variants. The ranking combination method was effective for some pairs of algorithms, but not for others. Overall, this work’s main results are the formulation and evaluation of several algorithmic extensions of LP for use in the analysis of genomic data, as well as the identification of several disease-associated SNPs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

Attribute – also called a dimension, feature, or variable, it represents a measured quality of the data

Biomarker – any biochemical signal in the body which can be quantified

Dimensionality reduction – any method which reduces the number of variables in a dataset

Feature selection – any dimensionality reduction method which keeps original dimensions intact

Feature ranking – any feature selection method which orders variables from most to least important

Feature subset selection – any feature selection method which outputs a set of relevant features, while discarding the others

GERP – genomic evolutionary rate profiling, a measure of SNP conservation

***k*NN –** k-nearest neighbor classifier

LP – label propagation

LP1 – single-score LP algorithm extension

LP3 – original allele-scoring LP algorithm

MAF – minor allele frequency

Sample – a single data point composed of many attributes

SLR – sparse logistic regression

SNV – single nucleotide variant

SNP – single nucleotide polymorphism; a SNP is a common SNV (>5% MAF)

SWRF – sigmoid weighted ReliefF

1.0 INTRODUCTION

In today's genomic era, DNA sequencing technology has become so inexpensive that it is readily available to everyday healthcare consumers. Genomic data has the potential to impact the way that healthcare providers prevent, diagnose, treat, and monitor disease. In the future, personalized medicine will enable physicians to generate a precise individualized assessment of risk of developing disease, and to pursue individualized therapy based on variations that are present in the individual's genome. In order to achieve this goal of precision medicine, it is necessary to find useful and predictive biomarkers in high-dimensional genomic data. These data require sophisticated algorithms to analyze, because there are millions of variants measured for only a few thousand individuals. Given the comparative paucity of samples compared to the number of genomic loci that are measured, the challenge is to find the relatively few variants that are associated with disease. In this proposal I investigate the application of a graphical algorithm called label propagation (LP) for feature selection in genomic variant data. I also develop several novel extensions to the algorithm, tailoring it specifically to genomic variant data. In particular, I apply and evaluate the algorithm and its extensions on genome-wide single nucleotide polymorphism (SNP) data.

1.1 DIMENSIONALITY REDUCTION

Dimensionality reduction (DR) techniques comprise a class of algorithms which are particularly applicable to the biomedical domain. Many forms of biomedical data, including genomic sequences, gene expression data, mass spectrometry data, or electronic health records, by their nature consist of a large number of variables. DR techniques either combine features to construct a smaller number of new features (feature construction), or select a subset of features that capture the important patterns in the data (feature selection). In the context of genomic analysis, DR techniques are often used to identify variants that are predictive of the disease or phenotype of interest.

Many DR techniques have been applied to genomic data with the goals of biomarker discovery, identifying features which may be used for prediction. DR techniques are directly applicable for biomarker discovery, identifying disease-associated (and potentially causal) variants that can illuminate the genetic underpinnings of disease. Moreover, the selected variants can be used in developing predictive models, which utilize discriminative variants to predict a disease or phenotype from genomic data. No one DR technique is best across all domains, so it is imperative to find algorithms that are well-suited to the genomic domain.

1.2 OVERVIEW OF LABEL PROPAGATION

Label propagation (LP) is a multivariate, semi-supervised, graphical algorithm that has been applied for classification and ranking of features in a variety of domains. In the context of biomarker discovery, label propagation represents genomic data as a bipartite network where genomic variants are represented by one set of nodes, and individuals with their disease status (case or control) are denoted by a second set of nodes. Links are allowed only between a variant node and a sample node, indicating which individuals exhibit which variants. LP labels the sample nodes with case or control status and propagates this information according to network topology. Ultimately, features are scored according to their association with the case or control group. This scoring is a multivariate optimization of a particular cost criterion which attempts to balance the strength of the initial labeling with the strength of network diffusion. In addition to evaluating LP on genomic data for ranking variants associated with case or control status, I have further extended it for application to genomic data.

1.3 HYPOTHESIS AND SPECIFIC AIMS

I hypothesize that a semi-supervised LP method can be improved for feature ranking (and selection) in genomic data, and that its application to high-dimensional SNP data will yield better results than currently used feature selection methods in terms of both predictive performance and biological function. To test this hypothesis, I propose the following specific aims:

Specific Aim 1. Extend the LP algorithm for SNP data to i) produce a probabilistic, single-SNP score rather than the current SNP-state score, ii) incorporate knowledge about the genome as priors, and iii) combine the single score LP method's output with another feature ranking method's output.

Specific Aim 2. Evaluate the LP extensions in Aim 1 on synthetic, semi-synthetic, and real GWAS datasets and compare its performance to that of chi square (univariate feature selection method), Relief, and Sparse Logistic Regression (multivariate feature selection methods).

The two main aims are to develop several extensions to the LP algorithm, and to evaluate these extensions on a variety of data.

1.4 CONTRIBUTIONS

The goal of this work was to develop an effective computational method for feature selection in high-dimensional genomic data. From a machine learning standpoint, the algorithmic extensions to LP provide improvements to an already effective, efficient, multivariate feature ranking method. The LP extensions that I have developed are applicable to many other types of high-dimensional data, and the combination method can enable the combination of existing effective algorithms.

From a biomedical standpoint, the methods described in this work can be applied to genomic data to discover new variants that are associated with disease. Moreover, since LP is multivariate it can be applied to discover not only variants with main effects but also interacting genomic variants. In this dissertation, LP and its extensions were characterized and extensively evaluated using synthetic, semi-synthetic, and GWAS data

1.5 OVERVIEW OF DISSERTATION

Chapter 2 provides relevant background to set the context for biomarker discovery in genomic data and surveys the related work in machine learning and genomic analysis. Chapter 3 describes the LP algorithm in detail, including previous LP variants and their applications in the literature. The novel extensions to LP are also described in this chapter. Chapter 4 describes the experimental method applied, including the performance metrics used to evaluate each algorithm, and a description of each dataset analyzed. Chapter 5 provides the results of the feature ranking experiments, and Chapter 6 summarizes the discoveries and conclusions drawn from this work. Full tables of results from all experiments are given in Appendix A and Appendix B.

2.0 BACKGROUND

This chapter describes relevant background in the biomedical and machine learning domains. I present the overall problem of biomarker feature selection in high-dimensional genomic data, and review some of the methods that have been applied successfully as described in the literature. Section 2.1 describes how genetic variation can impact disease risk, and Section 2.2 describes several types of variation commonly analyzed. Section 2.3 explains methods of discovering disease-associated SNPs. Section 2.4 covers some of the challenges encountered when performing large-scale genomic analysis. Section 2.5 gives an overview of dimensionality reduction methods, while Section 2.6 describes some particular DR methods in greater detail.

2.1 GENETIC BASIS OF COMPLEX DISEASES

Genetic inheritance has been understood for centuries as traits that are passed down from generation to generation. Genetics influences readily apparent traits such as hair, skin, or eye color, height, and several other physical attributes. Other genetic traits include susceptibility to many diseases, some of which are well-understood, and others whose genetic components are only just being unraveled.

Mendelian diseases are the simplest class of genetic diseases, directly caused by a variation at a single locus or gene. Presence of the genetic variant indicates presence of the disease with high probability. Examples of Mendelian diseases are cystic fibrosis, which results from mutation in the *CFTR* gene, and Tay-Sach's disease, which is caused by mutations in the *HEXA* gene. The genetics of many Mendelian diseases are relatively well understood; however, Mendelian diseases tend to be rare, recessive disorders that affect only a small portion of the population [1].

Common diseases such as late-onset Alzheimer's disease or coronary artery disease stand in contrast to Mendelian diseases in that their prevalence is much higher in the population, and that their genetic component is more complex and less well-characterized. The genetic basis of common diseases does not lie in a single locus or gene, but is instead hypothesized to be the result of many variants working together. Dozens of genetic variants have been shown to be associated with many common diseases, including diabetes, bipolar disorder, and hypertension.

A number of genetic models have been proposed to explain the genetic basis of common diseases. One prevalent hypothesis is the "common disease – common variant" hypothesis, which states that many common variants in combination determine risk of common diseases. This is supported by the observation that in several genome-wide association studies, results hold true across distinct populations with varying allele frequencies. If rare variants were responsible, varying allele frequencies would result in widely varying associations between different populations [2, 3].

In contrast to this hypothesis is the "common disease – rare variant" hypothesis, which states that rare variants rather than common ones that underlie common diseases. In support of this hypothesis, allele rarity has been shown to be proportional to the likelihood of disease association [4]. Intuitively, this is a result of selective pressure selecting against the deleterious alleles, and it has been shown that particularly deleterious alleles are indeed rare in the population.

It is likely that both common variants and rare variants are involved in common diseases and a combination of rare and common variants may be responsible for most common diseases. Thus, feature selection methods that are developed for genomic data should be able to identify both rare and common variants.

2.2 GENETIC VARIATIONS

The human genomic sequence consists of approximately 3 billion nucleotide pairs. Because the vast majority of the genomic sequence is identical across all humans, it is sufficient to focus analysis only on the variant regions of the sequence when analyzing genomic data in the context of disease.

The commonest sequence variation is the single nucleotide variant (SNV) which is a variation that occurs when a single nucleotide (A, T, C, or G) at a specific location in the genome (called a locus) differs between individuals in a population. On a particular chromosome, a SNV can be one of two nucleotide pairs (A-T or C-G). These states are called alleles, the more common of which is called the major allele “A” (not to be confused with the specific nucleotide A) and the less frequent one is called minor allele “a”. Because humans have two versions of each chromosome (one from the mother and one from the father), the genotype is defined as the combination of the SNV alleles on both chromosomes. As such, there are three possible genotype states for each SNV – the major homozygous (AA), the heterozygous (Aa), and the minor homozygous (aa).

SNVs can be broadly grouped into common SNVs and rare SNVs. When the minor allele frequency (MAF) is 5% or higher in the population, a SNV is considered to be a common SNV [5], also called a single nucleotide polymorphism (SNP). Among the approximately 3 billion nucleotides in the human genome, tens of millions are common SNPs. In contrast, when the MAF is less than 5%, the SNV is considered to be a rare SNV. These uncommon SNVs account for very rare population-wide alleles, as well as loci of unexpected variation such as an individual's unique random mutation in a particular locus.

There are a variety of technologies for measuring SNPs. Several companies such as Illumina and Affymetrix offer low-cost SNP arrays which can genotype an individual at millions of SNPs at once. This is only a small subset of all SNPs in the genome, so SNP arrays often oversample non-synonymous SNPs (which alter the amino acid chain produced) and exonic SNPs (which are within protein-coding regions) in order to find biologically interesting variants. SNP arrays target loci of common variation because the genome is queried at particular sites, rather than being read as a string of nucleotides – to be cost-effective, coverage is not focused on loci where variation is not expected or extremely rare.

Exome sequencing is a more recent technology used for measuring the genome. This method reads all nucleotide pairs in the exome, which is the protein coding portion of the genome, consisting of only about 1.5% of the full genome. This method can discover very rare, even unique mutations in the DNA sequence, because it does not just query loci of expected variation. Furthermore, variation within genes may be easier to understand than intronic regions, as there are methods of analyzing the downstream effects in terms of amino acid and protein alterations.

Copy number variation is yet another method of measuring the genome, and involves looking at repeated chunks of DNA throughout the genome. Large portions of DNA are duplicated multiple times, but the exact number of copies can vary from person to person. Gene functionality can be lost if there are fewer DNA copies than normal, and conversely, more copies than normal can lead to hyper-activity of a gene.

2.3 GENOME-WIDE ASSOCIATION STUDIES

In a genome-wide association study (GWAS), high-throughput genotyping technologies are used to assay hundreds of thousands or even millions of SNPs across the genome in a cohort of cases and controls. Since the advent of the GWAS, many common diseases, including late-onset Alzheimer's disease, diabetes mellitus, and heart disease have been studied with the goal of identifying the underlying common genetic variations. GWASs are based on the common disease-common variant hypothesis which posits that many common variants underlie common diseases, each variant increases the risk of disease modestly and an individual manifests disease when he or she has a sufficient number of common variants that cumulatively increase the risk of disease above a threshold level [6].

While GWASs have uncovered several thousand SNPs associated with a range of common diseases, these SNPs explain only a small proportion of the genetic variability. A possible reason for the moderate success of GWASs is the common disease-rare variant hypothesis, which posits that many rare variants underlie common diseases and each variant causes disease in relatively few individuals with high penetrance [7]. However, larger sample sizes and new analytical method designed specifically for rare variants will likely make GWASs useful for detecting rare variants as well [4].

GWAS data has a number of properties that make it particularly attractive to biomedical researchers. First, it is relatively cheap to obtain SNP data. The price of sequencing DNA has fallen dramatically over the past 10 years, to the point where a whole genome can be sequenced for just a few thousand dollars [8]. Eventually, SNP data will be replaced with full sequence data. Second, it is mostly invariant in time and space (ignoring events like tumor mutations). Other types of data, such as gene expression, may change over the course of hours or even minutes, and can be different depending on the type and location of tissue sampled. A person's DNA sequence is fixed from the moment of conception, and is carried by every cell in their body. Finally, the data comes in standardized formats and is available to researches for secondary analyses from large repositories such as dbGaP [9]. Moreover, knowledge about functional aspects of SNPs continues to grow rapidly, and is readily available to researches from knowledge bases such as dbSNP [10]. Hence, genomic variation data is likely to be useful for many clinical applications, ranging from diagnostic aids to guiding treatment decisions.

2.4 ANALYSIS OF GENETIC VARIATION DATA

The analysis of genomic data occurs often in the service of two goals – (1) biomarker discovery, and (2) development of predictive models for assessing risk or predicting the development of disease. Given thousands of variables representing the genomic or other biomolecular state of a person, we would like to find relatively few variables that can explain the data. Generally, biomarkers are sought in the context of a disease or a phenotype, indicating correlation or causation of the disease or phenotype of interest. Biomarkers have a wide array of applications in the clinical domain, aiding in risk assessment, disease diagnosis, patient prognosis, and treatment decisions The task of GWAS SNP analysis raises some practical concerns that are specific to the genomic domain. I discuss some of them here, ranging from genomic population structure to computational issues.

2.4.1 Linkage Disequilibrium

The phenomenon of linkage disequilibrium (LD) appears in localized regions of DNA, and is a measure of statistical correlation between genetic loci. Recombination is the process by which genetic material is exchanged between homologous chromosomes, resulting in a new combination of alleles. SNPs which are physically near one another are more likely to be inherited together, resulting in two loci which carry highly correlated information (strong LD). Just as redundant or correlated variables can complicate traditional machine learning tasks, LD can confound GWAS analysis. For example, a SNP showing significant association with disease may have no causal effect on the phenotype, the culprit instead being a nearby SNP in LD that was never measured. It is also possible that multiple SNPs show a strong signal, when in actuality all belong to a block of SNPs in strong LD. One method of dealing with the problem of LD is by identifying tag SNPs, each of which represents a block of SNPs in strong LD. By analyzing a dataset consisting only of tag SNPs, redundant information is removed.

2.4.2 Population Stratification

Another potential complication in the analysis of GWAS data is population stratification, which can cause spurious associations to appear in a dataset. A stratified dataset exhibits systematic differences in allele frequencies among the population, caused by factors other than association with disease (such as ancestry). Population stratification can be controlled for by filtering out the offending allele differences. By performing a cluster analysis (such as PCA), stratified groups can be identified, and their particular characteristics can be zeroed out in order to work with an unstratified population [11].

2.4.3 Computational Complexity

GWAS datasets are usually high-dimensional and can contain billions of measured genetic loci that can consume several gigabytes of memory. As such, the computational complexity of analysis methods, both in terms of time and memory required, is a significant issue. Algorithms which are of quadratic or combinatorial complexity in terms of the number of features are almost always intractable for analyzing a GWAS on a single machine. The linear complexity of most univariate methods is one reason why they have found widespread use in the genetics literature.

One way of addressing the issue of complexity is to utilize multiple processors in a parallelized fashion. Some algorithms can be readily programmed so that they can take advantage of multiple processors to reduce the runtime. However, other algorithms cannot be easily parallelized. Another possibility for certain algorithms is to shift to sample-space analysis rather than feature-space analysis. Certain matrix-based algorithms utilize pairwise transitions between *N* samples consisting of *d* features. With some matrix rearrangement, one can generate an equivalent expression that uses an N^2 transition matrix, rather than an intractably large d^2 matrix. For a dataset with N=1,000, and d=500,000, a sample-space representation of the transition matrix requires only about 7.6 megabytes of RAM, while a feature-space representation would require over 1.8 terabytes! With problems of this size, the limitations of even a high-powered computing platform can preclude the use of certain methods.

2.4.4 Avoiding Bias

Feature selection in the genomic domain can either be an end in itself or it can be a first step in an analysis pipeline, like one that builds a predictive model to be tested on a validation set. It is important to consider how selected variants will be used when performing the selection process, so as not to introduce unwanted bias. When building a predictive model to describe data, the model is usually learned on a training set, and then applied to a held-out test set to evaluate the performance. If feature selection is performed as a first step before building the model, it must be performed on only the training dataset. Selecting features on the full dataset will produce downstream results that are based in part on the testing data, and will often yield an overly optimistic estimation of the model's generalizability. Full-dataset selection is appropriate, however, if the selected variants are not being applied directly to the original samples, such as the case of mining biological knowledge without building a predictive model.

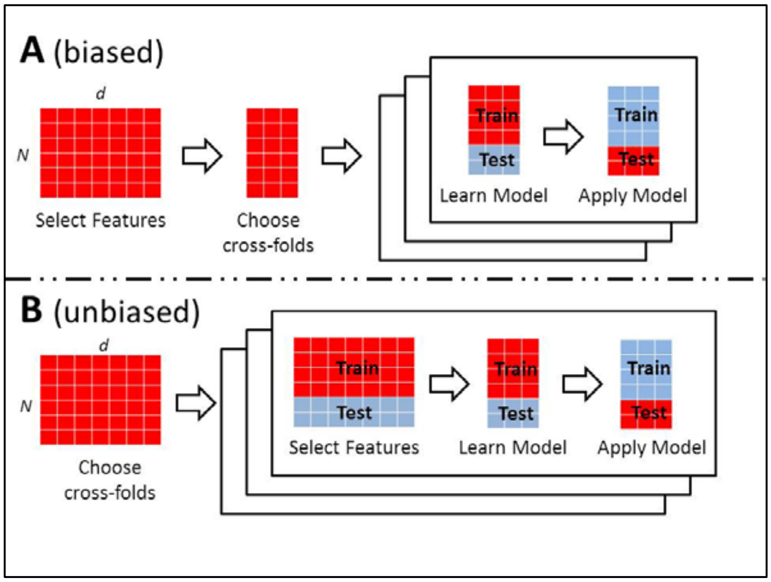

Figure 1 - Two experimental designs incorporating feature selection and cross-fold validation.

The highlighted cells in red are the ones being used in each step's computation, and the boxed sections represent processes that are repeated for each of the cross-folds. Bias is introduced in A, because features are selected in part on the testing data. Design B avoids bias, and is the preferred design.

2.5 DIMENSIONALITY REDUCTION

Because of the high dimensionality of GWAS data, dimensionality reduction methods are of importance in the analyses of such data. Though there are many types of dimensionality reduction methods, they all have the common goal of reducing the number of variables in a dataset. Usually, the goal is to find a parsimonious representation of a high-dimensional dataset using fewer features. This can be achieved in one of two ways – by performing feature ranking and feature subset selection, keeping relevant variables intact, or by mapping the data to lower-dimensional space using combinations of variables as the new dimensions (feature construction).

Dimensionality reduction is important as a preprocessing step in the analysis of high-dimensional data, because many classification and prediction algorithms often perform poorly with variables that are numerous, noisy, meaningless, or redundant. By first reducing a dataset to a smaller number of predictor variables, overfitting can be reduced and statistical performance can be improved. Dimensionality reduction also has more specific domain applications, in that the features remaining after reduction are good candidates to explore for knowledge about the domain.

2.5.1 Types of Dimensionality Reduction

Dimensionality reduction, feature ranking, and feature subset selection are related methods that seek to remove irrelevant dimensions from a dataset [12]. Dimensionality reduction is the most general of the three concepts, and encompasses methods that can remove, combine, scale, or otherwise transform variables in order to produce a low-dimensional representation of the original data. A dimension in the new representation may not correspond directly to any dimension in the original representation, instead being combinations of multiple variables. Such dimensionality reduction methods may lead to new variables that are not readily interpretable like the variables in the original space, which can be important in the biomedical domain. Typically a new variable that is a function of several biomarkers has little meaning in terms of biological knowledge; we would instead prefer to identify individual biomarkers of interest.

Feature selection methods, in contrast, preserve the original dimensions of the dataset. This can be achieved by individually ranking the dimensions, or by directly identifying a subset of the dimensions. In ranking methods, a score is assigned to each dimension, indicating a level of importance. These scores can be directly compared between features, yielding a ranked list of features. This representation is useful for finding features which are individually important. While feature ranking methods put features on a scale from most to least important, they do not establish a significance threshold. The ranking allows for sequential consideration of potentially important biomarkers in a logical order, but does not establish when features become irrelevant.

Feature subset selection returns a subset of variables that are important in some sense from the original set of variables. The subset can be of varying size, but is generally much smaller than the original set. Each selected feature corresponds directly to a dimension in the original data, making interpretation simple. However, feature selection methods often consider the selected subset as a whole, meaning that individual features in the subset may not be important on their own.

Among the methods of dimensionality reduction, feature ranking may be the most appropriate in the biomedical domain. In the context of a biomedical dataset, a feature ranking method will return a list of biomarkers sorted by their importance. These biomarkers are unmodified dimensions which correspond directly to observed variables in the dataset, meaning they may be easily interpreted. Furthermore, the top-ranked biomarkers are each important on their own, making them good candidates for being causal biomarkers, drug targets, or other another kind of individually meaningful variant.

Feature subset selection methods may also be suitable for GWAS analysis, in that the selected features correspond directly to measured biomarkers. However, these biomarkers may not be important on their own, making univariate analysis for drug targets or therapy difficult.

Dimensionality reduction methods which construct new features as some combination of the measured variables are generally unsuitable for biomarker discovery in genomic analysis. While these algorithms may perform well from a statistical or machine learning standpoint, the lack of feature interpretability is a problem. It must be kept in mind that one of the primary goals of GWAS analysis is to find variants of interest, which can then be studied in terms of biological function, practical application, and clinical relevance. A variable which is a combination of many biomarkers may not give a clear indication of how those biomarkers function individually. A constructed variable could also be cumbersome for the task of prediction, because each individual biomarker must still be measured in practice.

2.6 FEATURE SELECTION METHODS

Feature selection methods perform either feature ranking or feature subset selection. In feature ranking a weight or importance value is assigned to each feature and the method returns a list of features are ranked according to the weight. In feature subset selection the method attempts to identify and return an optimal subset of features. A feature ranking method can be converted into a feature subset selection method by choosing a threshold and returning the features with ranks lower than the threshold. Typically, it is not possible to convert a feature subset selection method into a feature ranking method.

There are many types of algorithmic methods for feature selection. Supervised methods leverage information about the target variable to find variables that are useful for prediction. They are usually inductive algorithms which learn a model independent of the unlabeled test data Unsupervised methods do not use information about the target, and instead try to find hidden cluster structure in the data. These methods try to find a parsimonious representation of the data, regardless of target values. In between these two paradigms are semi-supervised algorithms. These methods use limited target information and use the structure of the unlabeled data to find relevant variables. In contrast to supervised inductive algorithms, semi-supervised algorithms are transductive because models are learned using samples that are not in the labelled training set [13].

A range of selection and feature ranking methods have been developed and a recent review of the methods is provided in [14]. There are three major families of feature selection methods, namely, filter methods, wrapper methods, and embedded methods. Filter methods evaluate features directly, independent of how the features will be used subsequently. For example, if features are to be used to develop a classification model, a filter method will select features based on some data-dependent criterion and then pass them to an independent classifier. Wrapper methods evaluate features in the context of the how they will be used. In terms of classification, features are evaluated directly in terms of their ability to improve the performance of the classification model. Embedded methods perform feature selection during the classifier construction, deriving feature subsets as a direct result of the classification itself. In addition to these three main types of methods, there are also combination methods that aggregate several feature rankings into a single ranking.

2.6.1 Filter Methods

Filter methods assess the relevance of features by considering only the intrinsic properties of the data. Univariate filter methods compute the relevance of each feature independently of other features. They are computationally fast and scale to high-dimensional data because the complexity is linear in the number of features and interactions between features are ignored. Typically, such methods compute a statistic or a score for each feature such as chi squared or information gain. Multivariate filter methods model correlations and dependencies among the features; they are computationally somewhat slower and may be less scalable to high-dimensional data. Examples of multivariate methods include ReliefF and Markov blanket feature selection. Filter methods are particularly appropriate for GWAS analysis because they are agnostic to the final task at hand, and the method used to accomplish it. Filter methods can be directly compared by passing their results through the same downstream classifier.

The chi squared statistic is commonly used in SNP analysis is a univariate filter method [15]. This test measures whether outcome distributions are significantly different among SNP states, indicating features that have an impact on disease. The chi squared statistic is very fast to compute and has a simple statistical interpretation. However, it cannot detect higher-order effects such as SNPs that interact to produce an effect on disease.

Bayesian methods comprise an entire class of GWAS analysis techniques, representing data in terms of a well-defined probability distribution. By utilizing prior information and observational data, model probabilities may be estimated. In the simplest Naïve Bayes case, all variables are assumed independent, and probabilities are assigned according to a maximum a posteriori (MAP) estimate. Alternatively, instead of picking a single most likely model, models may be averaged according to their likelihood. Other implementations, such as the BD, BDe, BDeu, and K2 methods, score Bayesian models according to the data likelihood under a particular parameterization, with each method accomplishing smoothing via prior pseudocounts slightly differently. Still other score-based methods include minimum description length (MDL), minimum message length (MML), Bayesian information criterion (BIC), or Akaike information criterion (AIC), each of which seeks to balance model likelihood with the number of parameters required to specify it. For these methods, feature selection can be performed by looking for models which show a higher likelihood under an assumption of association, rather than independence. Variables which are present in high-likelihood models and absent in low-likelihood models have the most explanatory power [16].

The Relief algorithm [17] is a multivariate filter method that has been applied to SNP data to rank SNPs. This method computes the relevance of a SNP by examining patterns in a local neighborhood of training samples. The algorithm examines whether, among reasonably similar samples, a change in SNP state is accompanied by a change in the disease state. The Relief algorithm can detect multivariate interaction effects by means of the neighborhood locality measure, but does so at the cost of increased computation time. The Relief algorithm has been adapted in several ways for application to SNP data. The most recently described adaptations of Relief include Spatially Uniform ReliefF (SURF) [18, 19] and Sigmoid Weighted ReliefF (SWRF) [20] that were developed specifically for application to high-dimensional SNP data.

2.6.2 Wrapper Methods

Wrapper methods contain a feature selection algorithm as well as a way to apply those features to a task (e.g., classification). Features are evaluated based on their contribution to the performance on the final task, and selection is performed for features which improve performance. In this way, the selection process is a “wrapper” around the classification model, iteratively searching the feature space for good classification results. As such, the selection and classification algorithms are closely tied, and cannot be separated. Comparison of different feature selection methods is not completely straightforward with wrapper methods, because the classifier is integrated into the selection process [14].

A genetic algorithm (GA) search is one such example of a wrapper method. In this method, subsets of SNPs are randomly selected and used for sample classification. The worst-performing subsets are removed, and the best-performing ones are randomly combined and mutated. Because the algorithm directly evaluates SNP subset on their ability to maximize the final performance metric (classification accuracy), it is considered a wrapper method. GAs have been applied to SNP data to find multilocus effects in GWAS data [21].

2.6.3 Embedded Methods

Embedded methods, like wrapper methods, have an interaction between the feature selection and classification processes. While wrapper methods select features and evaluate these subsets in terms of classification performance, embedded methods derive features subsets from the classification process itself. In a way, the two are even more closely linked.

A decision tree algorithm is one example of an embedded feature selection technique that has been used to find interacting SNPs in GWAS data. Classification and regression tree (CART) algorithms build a tree classifiers where decision splits are SNPs and samples are classified at the leaf nodes [22]. The CART algorithm greedily chooses the next SNP to split on based on some criterion (e.g., Gini impurity). This leads to trees that have highly discriminative SNPs at the top, with predictions being refined further down the tree. Because the tree is built in an iterative greedy fashion, the feature selection process is intimately tied with the classification performance. The feature search process is embedded in the classification method.

The random forest (RF) method is an improvement to CART, using multiple decision trees to aggregate and smooth performance, reducing the variance in estimates. Many decision trees are built on random subsets of the data, and then feature importance overall is computed from the fraction of trees that contain a particular feature. RF has been applied successfully for finding biomarkers in biomedical datasets [23, 24].

Other embedded feature selection techniques include classification algorithms from which a feature scoring can be derived. The support vector machine (SVM), for example, is a classification algorithm that builds a maximum-margin decision boundary anchored to relatively few training points near the boundary. Several methods leveraging the properties of this classification for feature selection have been developed, including a recursive feature elimination method (SVM-RFE) [25] as well as gradient-based leave-one-out gene selection (GLGS), which is based on the least squares SVM classifier [26].

Some feature selection methods have been specifically designed to find epistatic SNPs with complex interactions, and a full review is provided in [27]. Some of these methods, such as SNP Harvester [28], filter strong main effects before searching for smaller interaction effects. Others, like maximum entropy conditional probability modelling (MECPM), utilize a greedy search to build higher-order interactions [29]. Still others, like multifactor dimensionality reduction (MDR), model epistasis by searching over the space of low-order interactions [30].

2.6.4 Combining Feature Rankings

Each feature selection method has differing strengths and weaknesses, so there can be benefit in combining multiple methods. Methods that combine multiple feature selection techniques are called ranking aggregation methods. Ranking aggregation methods have been shown to improve the rankings obtained from several different feature selection methods [31]. Rankings may be aggregated over different algorithms applied to the same data, or even over the same algorithm applied to different datasets.

Two categories of ranking aggregation methods have been described: i) order-based aggregation and ii) score-based aggregation. Ranking aggregation is typically based on order-based aggregation, that is, only the order of the features in the ranking is taken into account. The advantages of order-based aggregation include that order-based aggregation is naturally calibrated and scale insensitive. A simple order-based aggregation method is the Borda method, which takes two or more ranked lists of attributes, and averages the ranks of each attribute over all the lists. Another method is the Condorcet method, which looks at pairwise feature rankings on each input list. The top ranked variable is the one which outranks the other variables on a majority of the input rankings [32].

In score-based aggregation, the scores (weights) of the features from different algorithms are combined. Score-based aggregations have to deal with several challenges. One, weights produced by each algorithm has to be rescaled to the same range (say, between 0 and 1) so that different absolute scales do not influence in the aggregate result. Two, the same weight produced by different algorithms might represent different feature relevance and the weights have to be calibrated (say, by multiplying the scores of each algorithm by some factor). A score-based Borda method exists as well, which averages the scores among all feature lists. In general, rank-based methods can be converted to score-based methods by normalizing the rank value, while the reverse can be achieved by ranking based on scores [33].

Table 1 - Feature selection methods that have been applied to GWAS data.

Method	**Type**	**Variations**	**Software**
Chi Squared test	Univariate		http://pngu.mgh.harvard.edu/~purcell/plink/
Logistic Regression	Univariate	LR + Interaction Terms, Sparse LR	http://www.cns.atr.jp/~oyamashi/SLR_WEB/
Support Vector Machine (SVM)	Multivariate	Greedy Regularized Least Squares	http://svmlight.joachims.org/ http://www.esat.kuleuven.be/sista/lssvmlab/
Naïve Bayes (NB)	Univariate	Model-Averaged Naïve Bayes, BIC, MDL, K2	http://www.dbmi.pitt.edu/content/manb
Decision Tree	Univariate		

Random Forest	Multivariate		http://cran.r-project.org/web/packages/randomForest/
SNPHarvester	Multivariate		http://bioinformatics.ust.hk/SNPHarvester.html
Maximum entropy conditional probability modelling (MECPM)	Multivariate		http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm
Multifactor Dimensionality Reduction (MDR)	Multivariate		http://www.multifactordimensionalityreduction.org/
ReliefF	Multivariate	Sigmoid-Weighted ReliefF, Tuned ReliefF	https://code.google.com/p/ensemble-of-filters/ https://github.com/mattstokes42/MoRF
Label Propagation (LP)	Multivariate	Spectral clustering	

3.0 ALGORITHMIC METHODS

This chapter provides background about the label propagation (LP) algorithm and describes in detail the version of LP that I implemented for application to genomic data, as well as the extensions to LP that I developed. Section 3.1 gives an introduction to LP, with a broad overview of the algorithm and its applications. Section 3.2 includes the mathematical and algorithmic details of the specific applied version of LP. Other forms of the LP algorithm are described in section 3.3, giving a more complete perspective on the general family of propagation algorithms. Section 3.4 details the novel algorithmic extensions to LP that I developed, including a single-score extension, a prior knowledge method, and a ranking combination method.

3.1 ALGORITHMIC DETAILS

One approach in machine learning involves representing a dataset as a graph where each node denotes a sample and the weight of an edge between a pair of nodes is measure of similarity between the two corresponding samples. Many algorithms utilize this representation; however, a family of algorithms leverages additional information in the form of labels that are added to small set of nodes in the graph. This setting gives rise to a semi-supervised learning wherein the additional information is used to label the unlabeled nodes while obeying the constraints imposed by the topology of the graph. This family of methods is called label propagation (LP) algorithms.

LP is a semi-supervised algorithm that can be used for classification and for multivariate feature ranking (e.g., ranking SNPs in a case/control genomic data). The data is represented as a bipartite graph. A bipartite graph contains two sets of nodes (i.e., *sample nodes* that represent individuals and *feature nodes* that represent SNP-states in GWAS data) and edges that link nodes from one set to nodes in the other set. The sample nodes are labeled with case/control status, and the LP algorithm diffuses the labels across graph edges to the feature nodes and back again, until a stable solution is reached. The solution results in a final labeling of all nodes in the graph which balances the diffusion of the labels with consistency with the original labeling. The labeling of the feature nodes can be used to rank the features.

LP scales well for thousands of samples and features. It has complexity $O(kNd)$, where N is the number of samples, d is the number of features and k is the number of iterations required for convergence. Typically, k is much smaller than N or d, which makes LP a relatively efficient method. Moreover, LP can handle missing data, as well as both continuous and discrete data.

Because of its wide applicability, fast running time, and multivariate nature, LP has been applied to several bioinformatics problems. For example, LP has been used in breast cancer gene expression data in order to find functional modules of co-expressed genes [34]. It has been applied to gene function prediction, utilizing known gene functions and interactions to infer the function of other genes [35]. It has applied successfully in classifying patients with Alzheimer’s disease using protein array data [36]. To my knowledge, LP has not been applied to SNP data outside of my previous work. Unique challenges in the SNP domain include a huge feature space (on the order of hundreds of thousands), as well as the discrete, nominal nature of SNP states (in contrast to the continuous nature of expression data).

3.2 DETAILS OF LP ALGORITHM FOR SNP DATA

In LP, SNP data are represented as a bipartite graph $G = (V, U, E)$ which consists of two sets of nodes V and U where nodes in V represent samples (individuals) and nodes in U represent features (SNP-states). Note that if a SNP has three states (major homozygote, heterozygote and minor homozygote) then it will be represented by three nodes in U. In addition to the two sets of nodes, the graph contains a set of edges E where each edge links a node in V with a node in U. An edge $e(v,u)$ that links node v with node u is associated with a link weight $w(v, u) = 1$. These edges connect sample nodes to feature nodes, representing the presence of SNP state u in individual v. Initial labels $y(v)$ and $y(u)$ are applied to sample and feature nodes, and take values {-1, 0, +1}, representing known training information about case/control status (+1 and -1, respectively), or a lack of information (0). Labels on sample nodes represent disease status, and labels on SNP allele nodes represent a level of association with disease status. An example graph initialization is shown in Figure 2.

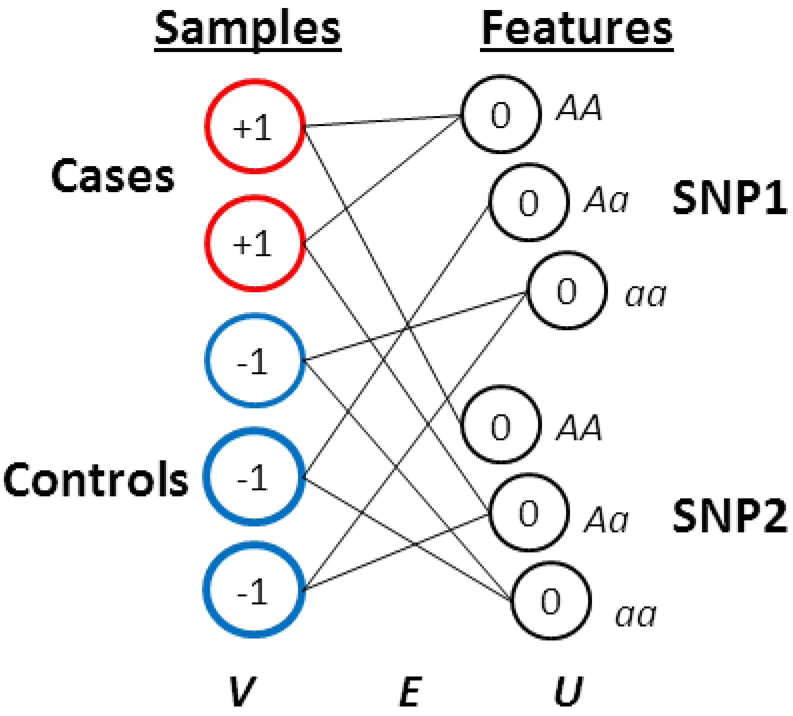

Figure 2 - A small bipartite graph for a hypothetical dataset with five samples and two SNPs. ***The five samples are represented by the nodes at the left (V), and are labeled with case or control status (+1 and -1, respectively). Each SNP is represented by three nodes at the right (U) with one node for each SNP state. Edges represent actual observations in the dataset and connect samples to the SNP states that they exhibit.***

Given the graph initialization, the propagation algorithm finds an optimal assignment of node labels *f*(*v*) and *f*(*u*), which minimizes the objective function

$$Q(f) = \sum_{(v,u)\in E} w(v,u)\left(\frac{f(v)}{\sqrt{d(v)}} - \frac{f(u)}{\sqrt{d(u)}}\right)^2 + \mu\left(\sum_{v\in V}\big(f(v) - y(v)\big)^2 + \sum_{u\in U}(f(u) - y(u))^2\right)$$

where μ is a parameter controlling the relative effect of the two parts of the cost function.

The first part of the equation is a smoothness constraint, ensuring that strongly connected nodes in *V* and *U* get similar labels. Here, *d*(*v*) and *d*(*u*) are the degree of each node in *V* and *U*, such that $d(v) = \sum_{(v,u)\in E} w(v,u)$ and $d(u) = \sum_{(v,u)\in E} w(v,u)$**.** The second part of the equation is a fitting constraint. For labeled nodes, this ensures that nodes labels are consistent with the initial labels. For unlabeled nodes, this term constrains the overall cost. In the discrete-label case where *f*→{-1, 0, +1}, the optimization of this cost function is NP-hard. By relaxing the labels so that *f*→*R*, however, the optimization of this equation becomes straightforward as derived in Zhou [37], and has the solution $f^* = (1-\alpha)(I-\alpha S)^{-1}Y$. Here, I is the identity matrix, Y is the vector of initial labels, and S is the normalized connectivity matrix $S = \begin{bmatrix} 0 & D_V^{-1/2} W D_U^{-1/2} \\ D_U^{-1/2} W^T D_V^{-1/2} & 0 \end{bmatrix}$, where *W* is the |*V*| x |*U*| sized matrix of edge weights and D_V and D_U are the |*V*| x |*V*| and |*U*| x |*U*| diagonal matrices containing node degrees, respectively. The parameter α (range [0, 1]) is analogous to the scaling parameter μ (range [0, +∞)) in the objective function, with $\mu = \frac{\alpha}{1-\alpha}$.

While the solution may be computed directly by algebraic evaluation, it requires the inversion of a *T* x *T* matrix where *T* is the total number of nodes in the network (*T* = |*V*| + |*U*|). This requires between $O(T^2)$ and $O(T^3)$ time, depending on the inversion method used. Instead, an iterative procedure may be used which diffuses node labels from one node set to another. First, the normalized graph Laplacian is computed as $B = D_V^{-1/2} W D_U^{-1/2}$. This is a special encoding of the graph which represents node degrees and adjacency. It has an interpretation as a random walk transition matrix, allowing labels to travel across graph edges. The node labels on *V* and *U* are computed iteratively as

$$f_{t+1}(V) = (1-\alpha)y(V) + \alpha B f_t(U) \quad \text{and} \quad f_{t+1}(U) = (1-\alpha)y(U) + \alpha B f_t(V)$$

where α is a user-specified parameter in the range [0, 1] that controls the balance between the initial labeling *y* and the diffusion of current labels *f*. This procedure ultimately converges to the same optimized node labels as the direct algebraic evaluation. The complexity of the direct algebraic evaluation is at least $O((|V| + |U|)^2)$, while the complexity of the iterative procedure is $O(k|V||U|)$, where k is the number of iterations required for convergence. The exact value of k depends on the properties of the graph as well as the convergence criteria, but in practice is found to be an order of magnitude less than both $|V|$ and $|U|$ even for large graphs (>100,000 nodes) and large alpha (>0.9).

The final labels of the nodes indicate association with the case or control group. Nodes with scores near +1 are associated with the case group, nodes with scores near -1 are associated with the control group, and nodes with scores near 0 are uninformative. For sample nodes, this score can be viewed as a prediction of case/control status based on SNP information. For feature nodes, this score can be interpreted as a case/control association measure that can be used to identify case/control associated biomarkers. The feature node scores may be ordered to obtain a ranking of biomarkers according to their association with case/control status.

3.3 ADDITIONAL LP ALGORITHMS

The LP algorithm has a number of implementations and interpretations. The iterative algorithm described in Section 3.2 can be expressed succinctly as $f^{t+1} = \propto Lf^t + (1-\propto)Y$, where L is the graph Laplacian, f^t contains node labels at iteration t, and Y contains the initial labels. A slightly different mathematical formulation can be expressed as $f^{t+1} = A^{-1}(\mu W f^t + Y)$, with the weight matrix W representing edge weights ($W_{ii} = 0$) and the diagonal matrix A containing entries $A_{ii} = I(i) + \mu D_{ii} + \mu\epsilon$, where D_{ii} represents node degrees and ϵ is a small constant that prevents degenerate, disconnected networks. These two formulations are very similar in their convex quadratic optimization framework, but seek to optimize slightly different cost functions. The first optimizes $Q(Y) = \|f_l - Y_l\|^2 + \|f_u\|^2 + \frac{\propto}{1-\propto}\left(D^{-1/2}f\right)^T L\left(D^{-1/2}f\right)$, while the second optimizes $Q(Y) = \|f_l - Y_l\|^2 + \mu\epsilon\|f\|^2 + \mu f^T L f$, with the subscripts l and u indicating portions of the vector corresponding to labeled and unlabeled nodes, respectively The main difference between the formulations is a somewhat stronger regularization in the former. While both methods seek to fit the given training labels on the labeled nodes, the former formulation more strongly drives the labels on unlabeled nodes to 0 in the absence of evidence.

The graph Laplacian L can be viewed as an operator on functions defined over the graph, and encodes the network geometry in terms of node connectivity and degree. The eigenvectors of L can be used for spectral decomposition of the graph, in that eigenvectors with the smallest eigenvalues correspond to the smoothest functions over the graph. It is possible to smooth any function on the manifold by projecting it onto p eigenvectors ($p < d$) with the smallest eigenvalues. An algorithm similar to LP is derived by smoothing a graph using eigenvector projection, and then fitting labels on the projected graph [38].

The LP cost criterion in Section 3.2 implicitly utilizes the graph Laplacian in the smoothness constraint. For any set of labels y, the smoothness between labels can be measured as $\frac{1}{2}\sum_{i,j}\left(y_i - y_j\right)w_{ij} = \boldsymbol{y}^T L\boldsymbol{y}$ [39]. The smoothness constraint penalizes labelings with rapid changes in y between strongly connected nodes. It is balanced with the fitting constraint using Tikhonov regularization, which minimizes the squared error from the initial labeling [40]. Various versions of the label propagation algorithm are obtained by combining different smoothness measurements with different error regularizations. An excellent review of the different quadratic criteria which can be optimized using iterative label propagation is provided in {Bengio, 2006 #319}.

Many applications of LP use a unipartite graph representation, where all nodes represent the same type of object (e.g., proteins) and edges represent strength of association between objects (e.g., the degree of interaction between a pair of proteins). This is in contrast to the bipartite network representation that was described in the Section 3.2, where nodes represent two different types of objects (e.g., individuals and SNP alleles) and edges represent which objects co-occur (which individuals exhibit which alleles). The unipartite graph has a different geometry than the bipartite graph, and is particularly useful for assigning nominal labels to objects, but is not directly applicable as a feature selection algorithm.

One variant of the LP algorithm involves clamping the training labels. In the implementation described in Section 3.2, each node’s score is updated in each iteration, regardless of whether the node was labeled or unlabeled to begin with. This allows for some flexibility in the initial labeling, and is useful in problems where classes cannot be linearly separated. In the clamped version of LP, all initially labeled nodes have their scores permanently fixed to their initial values. This version of the algorithm is useful in the unipartite representations, especially when initial labels are sparse. Clamping the initial values can prevent weak diffusion, because the signal sources will not attenuate over time. In the bipartite network representation that I use for genomic data, however, the clamped version is unsuitable. All sample nodes in the graph begin with labels, and are only one edge away from unlabeled nodes. Fixing the labels on all the sample nodes would prevent diffusion through the graph, as any information passed back from the feature nodes to the sample nodes would be lost at the clamped nodes. Every diffusion path is in effect blocked by a clamped node.

This clamped version of the algorithm also has a physical interpretation as an electric network. The initial node labels can be considered as positive or negative voltage sources, and the edge weights represent electrical conductance between nodes. In this case, the final labeling on the unlabeled nodes is equivalent to the voltage that would be observed on those nodes in the real-world electric network, obeying physical constraints such as Ohm's Law and Kirchoff's Law. A similar interpretation can be made with a heat diffusion network, where labels represent heat sources or sinks, edges represent thermal conductance, and final labels represent temperature. These interpretations provide an intuitive understanding of the algorithm's behavior as it attempts to achieve smooth gradients while being driven toward the initial labels.

The distance from labeled nodes to unlabeled nodes is an important property of the LP graph, which again differs widely from the unipartite to the bipartite representations. In the bipartite network, unlabeled nodes are only a single edge away from labeled nodes. In the unipartite network, it can take multiple hops to find a labeled node. The average distance to a labeled node has a direct impact on the optimal amount of diffusion in the LP algorithm. Sparsely labeled graphs with many distant unlabeled nodes will require more diffusion, while densely labeled graphs with only nearby unlabeled nodes will require less diffusion. Obviously, the bipartite implementation is in the second category, requiring relatively little diffusion to spread the training labels to every corner of the network.

Network geometry has been examined as a means to choose appropriate diffusion parameters, with a number of methods proposed. One heuristic method is to build a minimum spanning tree over the graph using Kruskal's algorithm, which iteratively adds the shortest possible edge that connects an unconnected component. The neighborhood size is then taken to be some fraction (1/3 is suggested) of the shortest edge that connects differently labeled components. This allows diffusion to take place mostly within the local class neighborhood [13].

The LP method can also be interpreted as a Markov random walk algorithm. If edge weights are cast as transition probabilities, it is possible to calculate the probability of arriving at any particular node given a starting point and random walk length. Using this information, it is straightforward to calculate the probability of arriving at a node with a particular label. The random walk length here is crucial to the final solution. As the random walk length increases, the steady-state probabilities are approached, but without a long enough walk, we might not find labelled examples. It is possible to remove the walk length variable entirely by stipulating that the walk continues until a labeled node is found. In this case, the LP algorithm becomes equivalent to the Markov random walk, in that the final LP label on an unlabeled node is proportional to the random walk probability of arriving at a node with a particular label. Once again, in the bipartite network representation, this interpretation is of limited value, because all unlabeled nodes are just one step away from labeled nodes.

The probabilistic interpretation can be extended to the bipartite LP algorithm. The final label on sample node represents probability of that individual belonging to the case or control class. The label on a feature node represents a random walk probability of reaching a node of a particular class. The walk length is directly related to the diffusion parameter α. For $\alpha = 0$, we have a one-step walk that uses no diffusion, and is a simple proportion of cases and controls. For $\alpha = 1$, we have an infinite-step walk, where the starting node is irrelevant (given a connected graph).

3.4 ALGORITHMIC EXTENSIONS

Section 3.2 gave the details of the main LP algorithm that I have implemented for GWAS SNP analysis. This section provides details of several extensions to the main LP algorithm that I have developed, implemented and evaluated.

3.4.1 LP1 score

The LP feature scoring method has a few drawbacks relating to the representation of SNP features in the bipartite graph. Each three-state SNP is represented by three binary variables, resulting in interdependent features. A score is given to each SNP feature (SNP allele), rather than to each SNP variable, making the interpretation of the final ranked list more complex. Ideally, each SNP should be given a single score to indicate its association with disease.

The feature scores themselves can be improved. The scores given by the LP algorithm exist on an arbitrary (-1, +1) scale which has no meaning outside of the relative ranking for a particular experiment. Furthermore, the scores are not associated with a degree of confidence to provide a measure of uncertainty about the score. Preferably, associated SNPs supported by lots of evidence should rise to the top of the ranking over associated SNPs with little support.

In order to overcome these shortcomings, I leveraged the soft labels discovered by the LP algorithm. In addition to labeling the unlabeled feature nodes, the method produces a new labeling even for the originally labeled sample nodes. I view these soft labels as probabilistic class identities, and use them to perform inference.

Many scoring methods utilize a contingency count table that tabulates observations in a dataset. A contingency table for a particular SNP consists of three rows representing the SNP's three genotypes (*AA*, *Aa*, and *aa*), and two columns representing the phenotypes (case or control). The table is filled in simply by tallying every individual's genotype-phenotype combination, resulting in the total count of all observations in the dataset. Scoring methods which operate directly on these observed counts include the chi squared statistic, the Naïve Bayes model, and the Bayesian Information Criterion.

Instead of using the counts derived from direct observations, I use partial pseudocounts derived from the LP method to fill in the contingency table. By running LP before filling in the contingency table, I allow some information to diffuse around the network, softening the hard labels. The amount of diffusion is controlled by the parameter α. At $\alpha = 0$, the algorithm relies solely on the initial labeling and allows no diffusion. This setting keeps the hard labeling, and produces a count table that is derived only from observations. At $\alpha = 1$, the propagation process dominates, resulting in diffuse, uniform labeling. This leads to an uninformative count table where every column has the same distribution of counts. An intermediate setting for α between 0 and 1 allows for some diffusion, while being sensitive to the initial labeling.

The soft labeling method results in a contingency table based on partial pseudocounts, to which I applied the chi squared test. The result is a likelihood that the phenotype distributions are different across the SNP states. This approach provides a single score for each SNP, as well as a readily interpretable probability value is a measure of a SNP's ability to discriminate between cases and controls.

Figure 3 shows the pseudocode for the LP1 algorithm. First, the LP3 scores on each sample node (in the range -1 to +1) are converted into case/control pseudocounts (in the range 0 to 1). These pseudocounts represent the probability of a sample belonging to class 1 (cases) or class 0 (controls). These pseudocounts are then used to fill in the contingency table, counting samples only for the features states (SNP alleles) that they exhibit, as dictated by the edge weight matrix *w*. Note that *w* has three times as many columns as SNPs, representing each SNP as a set of three grouped alleles. Finally, a simple chi squared statistic is computed using the LP3-derived pseudocount table.

```
//Turn LP3 sample labels (-1, +1) into case/control pseudocounts (0, 1)
for i = 0 to #Samples-1
    Vcount[i, 0] = 1 - (V[i]+1)/2
    Vcount[i, 1] = (V[i]+1)/2
//Now use pseudocounts to construct contingency table
Initialize all arrays (Obs, Row, Col, Exp, LP1) to 0 for all array indices
//Compute a score for each SNP
for j = 0 to #SNPs-1
    //Combine allele-specific LP3 scores
    for a = 0 to 2
        //Collect observed pseudocounts (Obs) in 2x3 contingency table
        for i=1 to #samples
            //Only collect counts for individuals who exhibit the SNP allele
            if w[i, 3j+a] > 0
                Obs[j, 0, a] += Vcount[i, 0]
                Obs[j, 1, a] += Vcount[i, 1]
        //Compute table column and row totals (Col and Row)
        Col[j, a] = Obs[j, 0, a] + Obs[j, 1, a]
        for d = 0, 1
            Row[j, d] += Obs[j, d, a]
    //Compute LP1 statistic over 2x3 contingency table
    for d = 0, 1
        for a = 0 to 2
            //Compute expected table values (Exp) under no association
            Exp[j, d, a,] = Row[j, d] Col[j, a] / #Samples
            //Chi squared statistic using observed pseudocounts
            LP1(j) += $(Obs[j, d, a] - Exp[j, d, a,])^2/ Exp[j, d, a,]$
```

Figure 3 - LP1 pseudocode

Finally, because the method is based on pseudocounts and not just proportions, the method is sensitive to the amount of evidence presented. Two SNPs with identical proportions of cases and controls would get the same score in the original LP algorithm, but now the actual number of cases and controls is important – the SNP with more data has a greater chance of having a significant statistic. I call the single-score LP method "LP1", and refer to the original three-score method described in 3.2 as "LP3".

3.4.2 Incorporation of prior knowledge

Data-driven methods examine just one dataset, tuning parameters and determining a feature ranking based solely on the presented data. These may be complemented by knowledge-driven methods, which utilize external sources of knowledge data to inform the learning process. The corpus of publicly available knowledge about the genome is large and growing, and contains many types of data which can be leveraged in performing feature ranking. I utilized multiple sources of prior knowledge in the LP algorithm.

3.4.2.1 Sources of Knowledge

There are several kinds of knowledge about the genome. We can glean information about genetic polymorphisms based on something as simple as their frequency in the population. It is also possible to leverage knowledge about the functions of gene production, to predict downstream functional effects of variations in genes. By using the genomes of multiple species, we can infer the importance of a specific locus in the genome.

Minor Allele Frequency

One simple type of knowledge about a SNP is based on population-wide allele frequencies. Using a database independent of the study population (e.g., the Hapmap Project

[41]), we can estimate the minor allele frequency (MAF) for each SNP. A SNP's MAF has been shown to be inversely correlated with it being a damaging variant [4]. That is, variants with rare minor alleles are more likely to be damaging. Purifying selection against damaging SNP variants causes them to appear at lower rates in the population than neutral or beneficial variants. Thus, we can have a prior assessment of a SNP's likelihood of disease association based on its MAF.

Substitution Effect

The deleteriousness of exonic SNPs can be predicted based on the structural changes in the amino acid sequence, and ultimately the functional changes in the corresponding protein, caused by the nucleotide replacement. Nucleotides are parsed in sets of three, called codons, each of which codes for one amino acid. Because there are 64 possible codons (3 positions, 4 possible nucleotides = 4^3), but only 20 amino acids (plus a stop codon), there is significant redundancy in the amino acid coding scheme. A nucleotide change which does not change the corresponding amino acid is called a synonymous substitution, and is unlikely to have an effect on the phenotype because the corresponding protein is not altered. Nucleotide substitutions which change the amino acid sequence are called non-synonymous mutations, and can have a wide range of effects on the phenotype. If the amino acid is changed to one that has relatively similar physical properties (such as charge, polarity, hydrophobicity, or volume), there may be little effect on the phenotype. On the other hand, a radical change in amino acid properties can have a significant effect on the phenotype; for example, an amino acid change can cause structural changes in the corresponding protein, rendering it non-functional.

Possibly the most damaging type of substitution is the nonsense mutation. Here, the nucleotide change results in an amino acid codon being changed to a stop codon. This signals the biochemical machinery of the cell to stop transcription of the DNA, resulting in a truncated protein. The further the stop codon is from the normal end of the protein, the more genetic information goes unused, and the more deleterious the change. Less commonly, a normal stop codon may be changed to an amino acid codon, resulting in a run-on protein that typically has diminished functionality.

Several online tools are available for computing the deleteriousness of a SNP allele based on the amino acid change caused. Examples include SIFT (**S**orting **I**ntolerant **F**rom **T**olerant [42]) and PolyPhen (**Poly**morphism **Phen**otyping [43]).

Conservation

Another method of predicting a SNP's importance for biological function stems from the analysis of conservation across species. Orthologs are conserved genomic sequences in DNA that appear relatively unchanged from species to species. The fact that a sequence has been preserved through millions of years of evolution and multiple speciation events suggests that the sequence is functionally important and intolerant to change. Therefore, a SNP which occurs in a highly conserved region is more likely to be deleterious than a SNP in a non-conserved region. Online tools for computing SNP deleterious based on cross-species conservation include PhyloP (**Phylo**genetic **p**-values [44]) and **G**enomic **E**volutionary **R**ate **P**rofiling 2 (GERP++ [45]).

The GERP score is a real-valued number which represents a quantity called rejected substitutions (RS). Using the genetic sequences of multiple species, the background neutral mutation rate is estimated, along with the actual number of mutations at any given locus. The RS value is computed as the number of mutations expected under a neutral mutation rate, minus the number of observed mutations. A positive score indicates fewer mutations than expected, i.e., a region that is conserved. A score of zero represents a neutrally evolving locus. While a negative score seems to imply faster mutation than expected, it is rather a result of variability in the neutral rate estimation, and should be interpreted simply as a lack of evolutionary constraint.

3.4.2.2 Use of Prior Knowledge in LP

I employed two types of knowledge in LP, namely, SNP MAFs and GERP scores. The standard LP algorithm described in 3.2 is a spatially uniform method that treats all features identically. I developed two methods for incorporating prior knowledge into the LP algorithm. The first is through the use of edge weights, and the second is through the use of prior pseudocounts in the contingency table analysis.

The iterative LP diffusion formula propagates labels along all dimensions equally. By using prior knowledge, however, it is possible to weight the features according to their prior likelihood of being associated with a phenotype of interest. This allows for greater diffusion through likely associated nodes, allowing them to have a greater impact on the final scoring. This is achieved this through the use of edge weights.

Previously, each edge $w(v,u)$ in the bipartite graph was given a weight of 1. I instead used a relative weighting, where all edges connected to a SNP with a higher prior likelihood of association are given higher weights. That is, $\forall v \in V, w(v,u) \propto prior(u)$. This allows more label propagation through nodes with higher prior likelihood, possibly making them have a greater impact on the final labeling. SNP nodes that are unlikely to be associated would permit less propagation, having little impact on the overall scoring of other nodes.

As an alternative to changing the edge weights in the network, prior knowledge is also incorporated as prior psuedocount observations. In the chi squared contingency table analysis, prior information is added as virtual counts. Simple Laplace smoothing could be performed by adding a count of 1 to each cell, or prior information could be incorporated as a larger number of skewed counts. I add these counts to the contingency table at the final LP1 scoring.

The distribution of these pseudocounts across the 2x3 contingency table should reflect the strength of the prior belief. Under an assumption of no prior knowledge (or knowledge against association), we would like to reinforce the null distribution of equal disease distributions for each SNP state. Ideally, we should also respect the natural distribution of genotypes, which are generally in Hardy-Weinberg equilibrium in the absence of association. So, in order to reinforce the notion of no association, I distribute the prior counts (*P*) as follows, where *d* represents the prevalence of disease in the dataset and *q* represents the MAF of the SNP.

Table 2 - Null distribution of prior counts.

	AA	***Aa***	***aa***
D^-	$(1-q)^2)(1-d)P$	$2q(1-q)(1-d)P$	$q^2(1-d)P$
D^+	$(1-q)^2dP$	$2q(1-q)dP$	q^2dP

For SNPs that do have prior evidence, the pseudocounts should skew accordingly. The column totals are fixed according to HWE, but row totals are not. For both the MAF and GERP score, we operate under the assumption that the minor allele is the one increasing risk of disease. So, as prior belief in association increases, counts move from D^+ to D^- in the *AA* column, and from D^- to D^+ in the *aa* column.

For the GERP score, we assume that scores less than 0 indicate neutrally evolving sites that should have the null distribution of no association reinforced. We then normalize positive scores from the range [0, 6.4] (where 6.4 is the maximum GERP score in practice) to the range [0, 1], and use this as a factor *m* for moving pseudocounts away from the null distribution. The MAF scoring follows a similar method, normalizing MAFs from the range [0, 0.5] to the range

[0, 1] and setting the factor *m* as 1 minus the normalized MAF. For both GERP and MAF scores, an *m* factor near 0 indicates a prior belief of no association, while a factor of 1 indicates maximum possible prior belief in association. Using the *m* factor, I modify the null count table (Table 2) with prior pseudocounts *a* through *f* as follows.

Table 3 - Null prior count table (left, see Table 2) and prior count table with a prior belief of strength *m* (right).

	AA	***Aa***	***aa***
$\boldsymbol{D^-}$	*a*	*b*	*c*
$\boldsymbol{D^+}$	*d*	*e*	*f*

→

	AA	***Aa***	***aa***
$\boldsymbol{D^-}$	$a + m^2 d$	*b*	$c(1-m^2)$
$\boldsymbol{D^+}$	$d(1-m^2)$	*e*	$f + cm^2$

For an *m* factor of 0 (no association), we just add the null count table to the actual distribution. As *m* increases toward 1, counts move up in the first column and down in the last column, indicating disease association with the minor allele. When $m = 1$ (the highest possible association factor), the *AA* column has 0 pseudocounts added to the D^+ row and the rest added to the D^- row, while the opposite is true for the *aa* column. The exponent on the *m* term in the contingency table significantly upweights SNPs with high priors (near $m = 1$), while having less effect on SNPs with low and moderate priors.

3.4.3 Combining feature rankings

Each feature selection algorithm has its own strengths and weaknesses. In previous work, I found that the univariate method chi squared tends to identify strong single-variable signals while missing more subtle multivariate effects. Multivariate ranking methods including LP, on the other hand, tend to rank the strong univariate signals somewhat lower, instead favoring variables with smaller independent effects. This suggests that combining the rankings from univariate and multivariate methods may produce a ranking that is superior to either the univariate or the multivariate method alone.

To this end, I combined feature scores from multiple algorithms. With the algorithmic extension, the LP scores are given SNP by SNP on a probabilistic scale, as opposed to three scores per SNP on an arbitrary scale. These scores can be used for both order-based and score-based rank aggregation. I combined scores from chi squared (univariate method) and from LP (multivariate method), as well as from SLR and LP using the Borda methods of combining ranks and scores.

Given two variable scorings, the Borda method generates a third variable scoring that is a combination of both. For the score-based, Borda method, the variable scores are simply averaged together arithmetically. This is simple for variable scoring methods that operate on the same scale, e.g., a probabilistic [0, 1] range. The rank-based Borda method is very similar, except that instead of directly averaging variable scores, the cardinal variable ranks are determined from the scores, and it is the ranks that are averaged. The rank-based version has the advantage of being insensitive to the absolute scale of the scoring methods used.

Because LP1 and the chi squared test use the same scale for their scoring metric (chi squared is in fact equivalent to LP for $\alpha = 0$), the feature scores may be combined directly. The SLR method's scores, in contrast, are just feature coefficients in the regression model, and don't map to a probabilistic scale. Because of this, we do not perform score-based Borda combination for LP1 and SLR. Instead, we use only the rank-based Borda method to combine LP1 and SLR. Because the SLR method is implicitly selective, most of the variables get identical scores of 0. When converting to a rank, all of these non-selected variants are assigned the maximum rank, which is just the number of variables in the dataset.

4.0 EXPERIMENTAL METHODS

This chapter describes the experimental methods used to evaluate the feature selection algorithms. Section 4.1 describes the eleven datasets used for analysis, Section 4.2 describes the performance metrics used to quantify each algorithm's performance, and Section 4.3 details the comparison algorithms used.

4.1 DATASETS

For the experiments, I used a synthetic SNP dataset, a semi-synthetic SNP dataset and nine GWAS SNP datasets. The synthetic dataset is low-dimensional, the GWAS datasets are high-dimensional with hundreds of thousands of SNPs, and the semi-synthetic dataset has a moderate number of features. All datasets have one binary target variable that denotes the case/control status of an individual, and many trinary SNP variables that indicate SNP alleles. The datasets are summarized in Table 4, and more details about the datasets are provided in the following subsections.

Table 4 - Summary of datasets.

Dataset	Cases	Controls	SNPs
Synthetic	134	866	1,000
Semi-synthetic (GAW17)	1,100	5,870	24,487
Alzheimer’s Disease (TGen)	861	550	234,665
Alzheimer’s Disease (ADRC)	1,291	958	682,685
Bipolar Disorder (WTCCC)	1,868	2,938	394,290
Crohn’s Disease (WTCCC)	1,748	2,938	393,861
Coronary Artery Disease (WTCCC)	1,926	2,938	394,265
Hypertension (WTCCC)	1,952	2,938	393,549
Rheumatoid Arthritis (WTCCC)	1,960	2,938	393,502
Type 1 Diabetes (WTCCC)	1,963	2,938	394,217
Type 2 Diabetes (WTCCC)	1,924	2,938	394,283

4.1.1 Synthetic dataset

The synthetic dataset contains 1,000 SNVs and a binary phenotype that is a function of 35 “causal” SNVs. Of the 35 causal SNVs, 10 of them were modeled as common SNPs with MAFs that were sampled uniformly from the range 0.0500 to 0.5000 with odds ratios in the range 1.05 to 1.50. The other 25 SNVs were modeled as rare SNVs that were sampled uniformly from the range 0.0001 to 0.0100 and odds ratios in the range 2 to 10. The remaining 965 SNVs (“noise” SNVs) ranged from common to rare, but do not have an effect on the phenotype. Phenotype status was assigned using an additive threshold model, with each causal SNV conferring an independent risk of disease. We created a set of 1,000 individuals and in that set 13.3% of individuals had a positive phenotype. The comparable number of samples and features make this model fairly robust to variations across instantiations of the data, reducing the need for multiple runs to observe “average” statistical performance.

4.1.2 GAW17 semi-synthetic dataset

The GAW17 dataset is a mini-exome semi-synthetic dataset that was constructed for the Genetic Analysis Workshop 17 that was held in 2010 at Boston, Massachusetts. The genomic data was obtained from 697 unrelated individuals whose exomes were sequenced in the 1000 Genomes Project and the genomic data consists of 24,487 autosomal SNVs that map to 3,205 genes [46, 47]. This is a mini-exome dataset since the 3,205 genes comprise a subset of all human genes.

The synthetic portion of the dataset consists of four quantitative risk factors that were simulated as normally distributed phenotypes. The genes associated with each of the risk factors were chosen from the cardiovascular disease (CVD) risk and inflammation pathways. Finally, a binary disease phenotype representing CVD was modeled as a function of the four quantitative risk factors. In the synthetic phenotype data, the values of three of the four risk factors (named Q1, Q2, and Q4) and the binary phenotype were provided for each individual. The values of the risk factor Q3 were not provided to simulate a latent factor. Q1 was modeled as a function of age and 39 SNVs in nine genes and included a genotype-smoking interaction. Q2 was modeled as a function of 72 SNVs in 13 genes and was not influenced by age, sex, or smoking status. Q4 was modeled as a function of age, sex and smoking; while it had a genetic component; it was not influenced by any of the SNVs in this dataset. The latent factor Q3 was influenced by 51 SNPs in 15 genes. A total of 200 replicate datasets of are provided. In each replicate an individual had the same genotypes and the phenotype was simulated.

This dataset simulated a common disease as function of both rare and common SNVs based on the current thinking that both common and rare SNVs contribute to the genetic basis of common diseases. Over 75% of the SNVs in the GAW17 have MAF below 1%, and nearly 40% have MAF below 0.1% (see Figure 4).

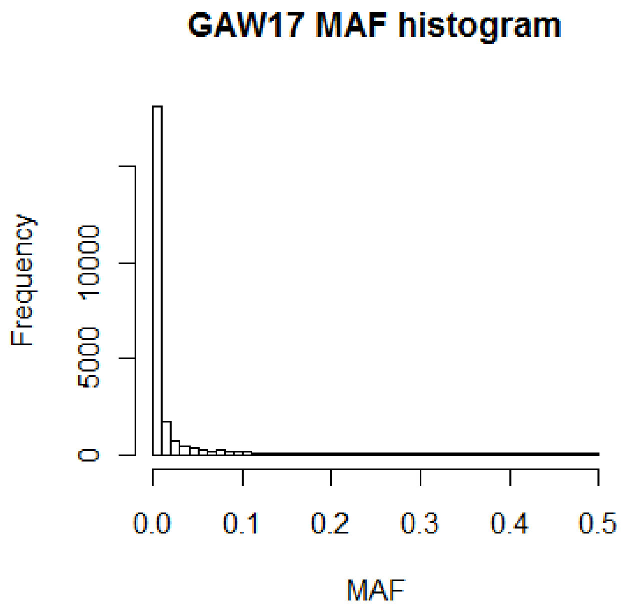

Figure 4 - Distribution of MAFs in the GAW17 data shows many rare variants.

For my experiments, I pooled the data in the first ten GAW17 replicates to create a dataset with 24,487 SNVs and 6,970 individuals. I used the binary Q2 risk factor as the phenotype of interest, because its underlying model uses only genetic variables and does not include any latent features.

4.1.3 Late-onset Alzheimer's disease datasets

Two late-onset Alzheimer's disease (LOAD) GWAS datasets were used in the experiments.

TGen Dataset

The TGen GWAS data comes from the Translational Genomics Research Institute (TGen) located in Phoenix, Arizona. This dataset was originally collected and analyzed by Reiman et al. [48]. The genotype data were collected on 1,411 individuals, of which 861 had LOAD and 550 did not. Of the 1,411 individuals, 644 were APOE 34 carriers (one or more copies of the e4 allele) and 767 were non-carriers. Of the 1,411 individuals, 1,047 are brain donors in whom the status of LOAD or control was neuropathologically determined, and 364 are living individuals in whom the status was clinically determined. The average age of the brain donors at death was 73.5 years for LOAD and 75.8 years for controls. The average age of the living individuals is 78.9 years for LOAD and 81.7 years for controls. The target phenotype variable is the presence or absence of LOAD. In this dataset, 61% (861 of 1,411) had LOAD. In the original study an Affymetrix chip was used with 502,627 SNPs for each individual. After quality control 234,665 autosomal SNPs were retained for analysis.

ADRC Dataset

The ADRC LOAD dataset comes from the University of Pittsburgh Alzheimer’s Disease Research Center (ADRC) [15]. This dataset consists of 2,229 individuals of which 1,291 were diagnosed with LOAD and 938 were healthy age-matched controls. All of the patients met National Institute of Neurological and Communication Disorders and Stroke (NINCDS) and Alzheimer’s Disease and Related Disorders Association (ADRDA) criteria for probable or definite AD. In the original study 1,016,423 SNPS were measured and after quality controls were applied by the original investigators 682,685 SNPs located on autosomal chromosomes were retained for analysis.

4.1.4 WTCCC GWAS datasets

The Wellcome Trust Case-Control Consortium (WTCCC) is a British study covering thousands of individuals and spanning multiple diseases. I used the WTCCC phase 1 datasets, which consists of bipolar disorder (BD), coronary artery disease (CAD), Crohn’s disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). Each disease dataset contains genotypes for approximately 2,000 cases. In addition two sets of healthy controls are provided that contain SNPs for approximately 3,000 individuals. After quality control filters were applied, individuals in each dataset contain approximately 400,000 SNPs.

4.1.5 Quality control for GWAS datasets

Each GWAS dataset was preprocessed by applying a number of quality control criteria. Both individuals and SNPs were filtered for missingness (<1% for SNPs, <5% for individuals) to eliminate poor-quality data points, and the remaining missing genotype values were imputed. SNPs were further filtered according to minor allele frequency (MAF > 0.01). Some SNPs passed full data-set MAF filters, but appeared monogenic in cases or controls, so the MAF filter was applied to cases and control separately. SNPs were further filtered according to Hardy-Weinberg equilibrium *p*-value. Finally, datasets were examined for population stratification. The PLINK software [49] was used to perform the principal component analysis (PCA) for each dataset, and clustering was performed using the first two principal components. Though mixed groups appear in the PCA plots, there is no significant clustering of cases and control separately, indicating that these datasets are suitable for analysis without requiring adjustment for population stratification (see Figure 5).

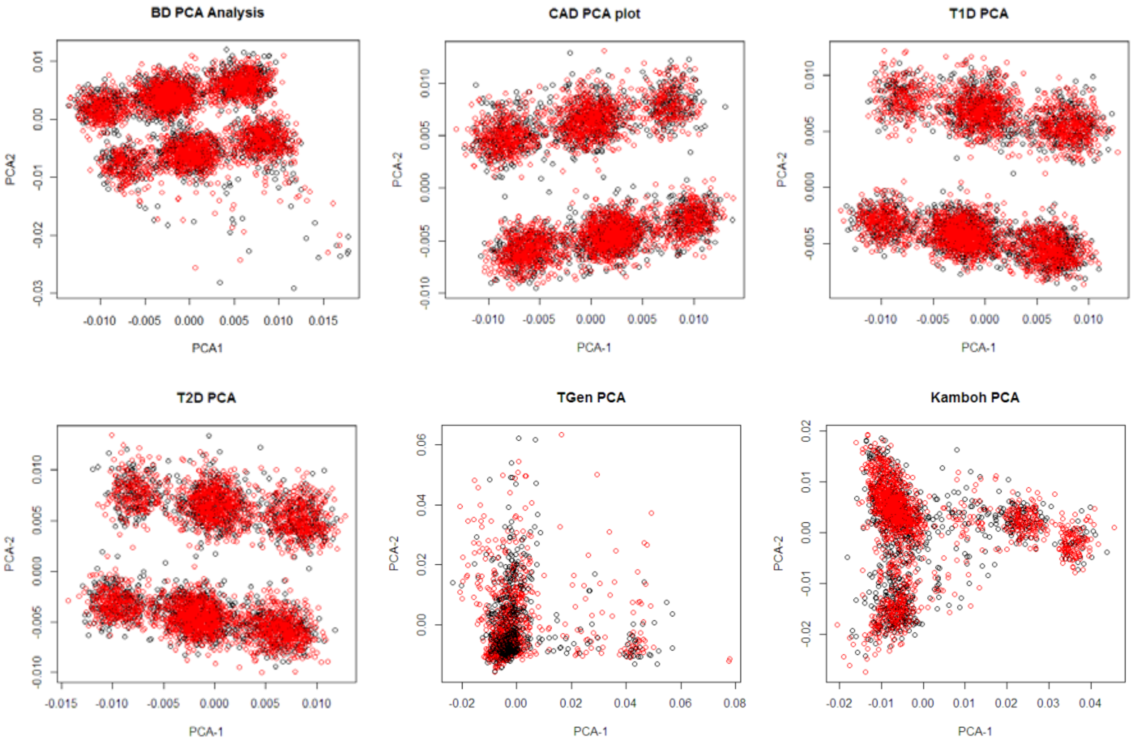

Figure 5 - Principal Component plots for select datasets.

While the datasets show some population structure, individuals are not stratified by case/control status

4.1.6 Prior knowledge sources

SNP minor allele frequencies and GERP scores were obtained using the Genome Variation Server (GVS). This online portal can be used to submit batch queries to the dbSNP server, downloading the relevant information for roughly half a million variants in about 24 hours. Records for nearly all SNPs in each dataset (>99.5%) were retrieved successfully, with both a GERP score and a MAF for each mapped rsID. SNPs that did not have an associated MAF or GERP score in the database were set to the mean score value for that dataset. As expected, GERP scores were centered near a null score of 0, indicating a neutral mutation rate. MAFs were generally distributed uniformly between 5% and 50%, but showed a deficit of rare SNPs with MAF < 5%. No correlation was found between the MAF and the GERP score in any dataset.

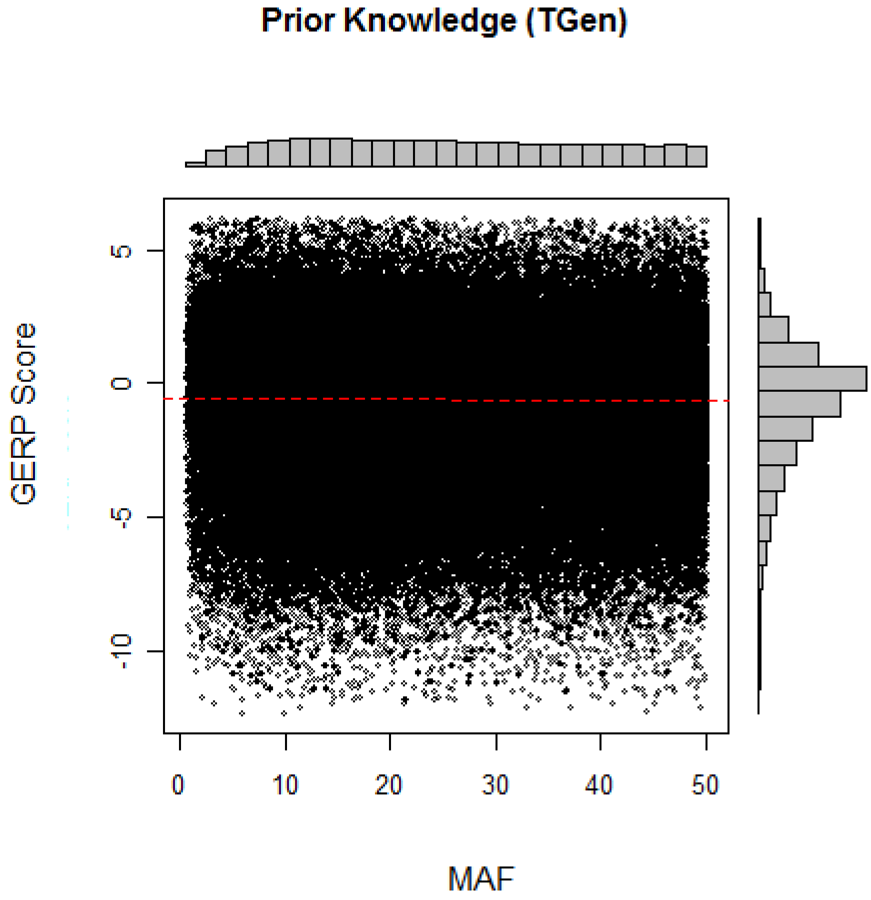

Figure 6 - Plot of GERP score versus MAF for each SNP in the TGen dataset, with distribution histograms. ***There is no correlation between the measures, as indicated by the trend line in red. All other datasets showed similar distributions of scores, with no correlation between MAF and GERP***

4.2 EVALUATION

Several methods are available for evaluating the performance of algorithms that are used for biomarker ranking and biomarker discovery. The evaluation methods that I used included precision-recall curves, evaluation of predictive performance using Receiver Operating Characteristic (ROC) curves, reproducibility of biomarkers across datasets, evidence of biological validity obtained from the literature for top-ranked biomarkers, and computational efficiency.

4.2.1 Precision-recall curves

Biomarker selection can be viewed as information retrieval with binary classification. In this context, precision (also called positive predictive value) is the fraction of selected biomarkers that are true (or causal) biomarkers, and recall (also called sensitivity) is the fraction of true biomarkers that are retrieved. High precision indicates that an algorithm selected substantially more true biomarkers than irrelevant biomarkers, while high recall indicates that an algorithm selected most of the true biomarkers.

In the context of ranked biomarkers, appropriate sets of selected biomarkers are naturally given by the top *k* ranked biomarkers. For each set of *k* biomarkers, precision and recall values can be plotted on the y-axis and x-axis respectively to give a precision-recall (PR) curve. In addition to the PR curve, it is also possible to view the ranking in terms of an ROC curve which utilizes the true positive rate (sensitivity) and false positive rate (fraction of negatives which are incorrect). By modifying the feature selection threshold value, it is possible to compute the TPR and FPR over all possible settings, yielding the ROC curve.

Since computation of precision and recall requires knowledge of true biomarkers in a dataset, this evaluation is performed only for experiments that use synthetic and semi-synthetic datasets where the true SNPs are known.

4.2.2 Evaluation of predictive performance

Meaningful features should be predictive of disease, and classifiers developed from highly predictive SNPs should have good performance in discriminating between cases and controls. I evaluated the predictive performance of the top-ranked SNPs for each feature ranking method and dataset by measuring the performance of a series of classification models that were developed using progressively larger number of top-ranked SNPs. Given a set of top-ranked SNPs obtained from a ranking method applied to a training dataset, I applied a *k*NN classification algorithm to a test dataset containing genotypes for the corresponding SNPs. I evaluated the performance of the classification algorithm using fivefold cross-validation. The dataset was randomly partitioned into five approximately equal sets such that each set had a similar proportion of individuals who developed the disease. I applied the ranking algorithm on four sets taken together as the training data, and evaluated the top-ranked SNPs' predictions on the remaining test data. I repeated this process for each possible test set to obtain a prediction for each individual in the dataset. I used the predictions to compute the area under the Receiver Operating Characteristic curve (AUC) which is a widely used measure of classification performance. The LP algorithm was evaluated using α=0.25, which was found to have the best classification performance among values tested between 0.0 and 0.9. This setting puts more emphasis on matching the case/control training labels while still utilizing some network diffusion, and is suitable for finding discriminative SNPs.

The *k*NN algorithm is a simple non-parametric classification algorithm that utilizes pairwise distances between a query sample and the training samples. For SNP data, the pairwise distance simply counts the number of SNPs which have different values between the query and reference individuals. The classification result for the query sample is then computed as the average target value among the *k* most similar training samples to the query. I utilized a setting of $k = 10$. The *k*NN algorithm is suitable for a small number of features, but suffers from the curse of dimensionality as irrelevant, noisy or redundant variables are added to the dataset. Because of this, I perform feature selection as a first step, using the downstream *k*NN classification performance as a proxy for evaluating the feature selection itself.

4.2.3 Reproducibility of biomarkers

I evaluated the feature ranking methods for reproducibility across the two LOAD datasets. The two datasets were reduced so that they contained only the genotypes for the 64,984 SNPs that were common to both. After running the feature ranking methods separately on each of the reduced datasets, the ranked SNPs were evaluated for reproducibility as follows. Given two ranked list of SNPs obtained by applying a feature ranking method to the two reduced datasets the ranked lists were examined for common SNPs in the top-ranked 10 SNPs, 50 SNPs, 100 SNPs, and so on. Reproducibility was calculated as the number of SNPs in common to both lists divided by the total number of SNPs in a list, yielding a value in the range from 0 (no SNPs in common) to 1 (both lists contain exactly the same SNPs). This measure only checks for presence or absence of SNP in a list, and ignores actual ranks within the list. Since only LOAD had two separate datasets, reproducibility was evaluated only across these two datasets.

4.2.4 Evidence of biological validity

For the GWAS datasets, I examined the top-ranked SNPs for biological significance and evidence of previously documented association with disease. I used several publically available databases and resources including SNPedia [50], GeneCards [51], and dbSNP [10] to identify evidence linking variants and diseases. In addition to SNPs directly named in the literature as having an association with disease, I also considered a wider range of plausible associations. For each SNP, I searched whether it was in strong linkage disequilibrium with disease-related SNPs, whether the SNP was in a disease-related gene, whether the associated gene was part of a strongly conserved, disease-related family, or whether the variant has been associated with a similar condition or a plausible pathway.

The actual protocol for identifying literature evidence required several steps. First, the each SNP's rsID was entered into Google, and the top 20 hits were examined for mention of the disease of interest or similar diseases/symptoms (e.g., neurological conditions for Alzheimer's disease datasets, bowel-related conditions for Crohn's disease dataset). If this search failed to produce evidence, another search was run using both the rsID and disease name as search terms (e.g., "rs1234 bipolar disorder"), and the top 20 hits were examined. The rsID was entered into dbSNP to identify the chromosome, and if applicable, the associated gene of the locus. The rsID was also researched using SNPedia, which returns citations of literature mentioning the SNP. If no literature association was found at the SNP level, I searched for association at the gene level (for exonic SNPs). Similar Google searches were run as above, replacing the rsID with the gene name, and searching for literature citations and gene information on the GeneCards database.

Due to the time required for manual validation of SNPs, I evaluated all algorithms' top-ranked SNPs for the TGen and ADRC LOAD datasets, but only the LP1 extension's results on the rest of the GWAS datasets. The chi squared test was also evaluated on the CD and HT WTCCC datasets for baseline comparison. These datasets were chosen because the performance of the LP1 algorithm was at its highest on the CD dataset, and showed somewhat poorer performance on the HT dataset.

4.3 COMPARISON ALGORITHMS

Three algorithms were used as comparison methods for LP. The SWRF algorithm was only applied the two LOAD datasets; while the chi squared test and SLR were applied to all datasets.

4.3.1 Chi squared test

The chi squared test is a commonly used univariate statistic that has a probabilistic interpretation. The test tabulates observations in a contingency table, which records co-occurrences of variable states and the target variable. The chi squared statistic computes the deviance from the null contingency table, in which all variable states have the same distribution of the target variable. This is done by determining the "expected" value for each cell in the contingency table, which is the row total multiplied by the column total, divided by the total number of samples. The chi squared statistic is then computed as $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$, where r is the number of rows in the contingency table and c is the number of columns, $O_{i,j}$ is the observed value in the ith row and jth column, and $E_{i,j}$ is the expected value in the same cell. The resulting statistic can be compared to the chi squared distribution to determine the probability of association between the SNP's alleles and the target variable.

4.3.2 Sigmoid Weighted ReliefF

The Relief algorithm was first described by Kira and Rendell [17] as a simple, fast, and effective approach to attribute weighting. The output of the Relief algorithm is a weight between -1 and 1 for each attribute, with more positive weights indicating more predictive attributes. The weight of an attribute is updated iteratively as follows. A sample is selected from the data, and the nearest neighboring sample that belongs to the same class (*nearest hit*) and the nearest neighboring sample that belongs to the opposite class (*nearest miss*) are identified. A change in attribute value accompanied by a change in class leads to upweighting of the attribute based on the intuition that the attribute change could be responsible for the class change. On the other hand, a change in attribute value accompanied by no change in class leads to downweighting of the attribute based on the observation that the attribute change had no effect on the class. This procedure of updating the weight of the attribute is performed for each sample in the dataset. The weight updates are then averaged so that the final weight is in the range [-1, 1]. The attribute weight estimated by Relief has a probabilistic interpretation. It is proportional to the difference between two conditional probabilities, namely, the probability of the attribute's value being different conditioned on the given nearest miss and nearest hit respectively.

In contrast to most other feature ranking or feature selection methods that consider attributes univariately, Relief algorithms are able to capture attribute interactions because the global distance measure which defines sample proximity is a multivariate function. However, because the nearest neighbors are identified by a distance measure that incorporates all attributes, the presence of many irrelevant or noisy attributes (as in SNP data) can lead to suboptimal identification of nearest neighbors.

I used a variant of ReliefF that I developed called sigmoid weighted ReliefF (SWRF). It utilizes a soft neighborhood inclusion threshold, and has been shown in synthetic data to have greater power than ReliefF [20] for selecting predictive variables.

4.3.3 Sparse Logistic Regression

The logistic regression (LR) algorithm is an algebraic method that learns a function to map the input variables ($\boldsymbol{X}$) to the target variable ($\boldsymbol{Y}$). Coefficient-weighted variables ($\boldsymbol{wX}$) are summed and passed through a logistic function to perform a mapping from inputs to an estimated target classification ($\widehat{\boldsymbol{Y}}$). The ideal variable coefficients minimize the residuals obtained when mapping inputs to target ($\left|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\right|^{p}$), where p is a regularization factor. In contrast to the linear regression model, the optimal solution to the logistic regression model cannot be expressed in closed form, so an iterative optimization technique such as Newton's method must be used. The logistic regression model space is much larger than the linear regression model space, and can represent many more function.

The cost criterion that the LR method optimizes can be altered to change its performance. By increasing the regularization of the cost equation, the LR algorithm will give lower coefficients to most variables. So-called sparse solutions are regularized such that the weights of many variables are pulled to 0, eliminating them from the model entirely. One sparse logistic regression (SLR) method utilizes automatic relevance detection to inform the iterative optimization procedure. The SLR algorithm's current estimate of variable weights is used to alter the prior distribution over which weights are estimated for the next iteration. Variables with low weights have their prior distribution shrink around 0, eventually eliminating many of the variables entirely. I used a MATLAB implementation of SLR which was developed for analyzing high-dimensional fMRI voxel data [52].

5.0 EXPERIMENTAL RESULTS

This section provides experimental results of applying the algorithms described in Chapter 3.0 using the evaluation metrics and datasets described in Chapter 4.0 Section 5.1 gives results of the original LP3 algorithm and the LP1 extension on the synthetic, semi-synthetic, and GWAS datasets. Section 5.2 provides results from the use of prior knowledge in the LP1 algorithm, and Section 5.3 provides results from the feature ranking combination method.

5.1 EVALUATION OF LP3 AND LP1

This section describes the evaluation of two LP algorithms that I developed for application to genomic data. The LP3 algorithm represents the data as a bipartite graph in which each SNP is represented by three nodes, with each SNP state being score differently. The LP1 algorithm also represents the data as a bipartite graph; however each SNP receives a single score representing association over all SNP states. My goal in developing LP1 was to adapt the LP algorithm to output directly a single rank for each SNP. The two algorithms are described in detail in Chapter 3.0

The performance of LP1 and LP3 algorithms were compared with three control algorithms – chi squared, SWRF and SLR. Each algorithm were evaluated on synthetic data, semi-synthetic GWA17 data, two LOAD GWAS datasets and WTCCC GWAS datasets on seven diseases. From the GWAS datasets, results for the SWRF algorithm are presented only for the LOAD datasets because SWRF had very long runtimes. Given that the WTCCC datasets have about double the number of SNPs and more than double the number of samples compared to the LOAD datasets, the $O(N^2d)$ SWRF algorithm was estimated to take 10 times as long to run, resulting in intolerable month-long computation times for a single fold on a single dataset. I replaced the SWRF algorithm with the SLR algorithm as a multivariate control algorithm since it has been recently shown in the literature to have good performance on GWAS data and has shorter running time than SWRF [52].

5.1.1 Synthetic Data

On the synthetic data, the algorithms were evaluated in their ability to identify the 35 causal SNVs (that include 10 common SNVs and 25 rare SNVs) using precision-recall and ROC curves, but not in terms of predictive performance. The LP algorithms were run with multiple settings for the α parameter.

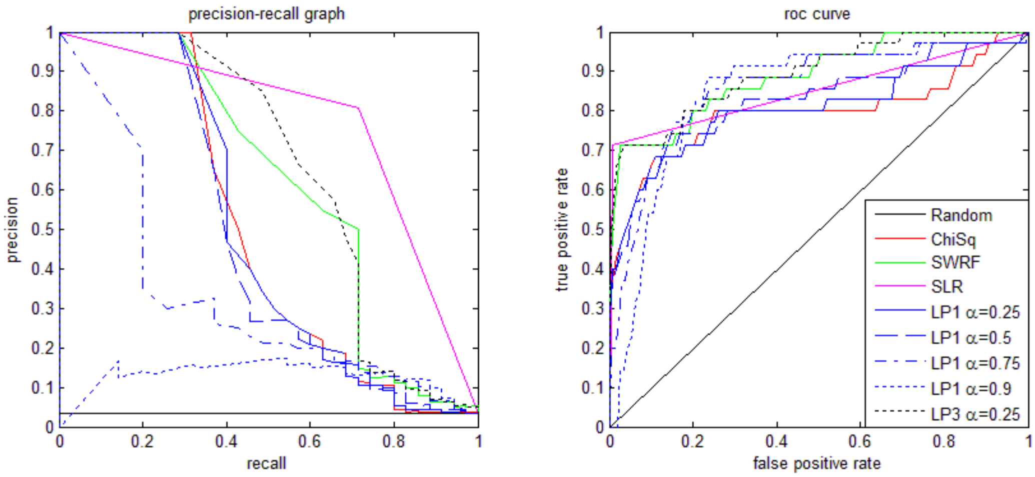

Figure 7 - Precision-recall curves and ROC plots for recovering 35 causal SNVs in synthetic data.

The SLR method performed the best on this dataset, with very good precision and recall. LP3 and SWRF have the next best performances, followed by LP1 and chi squared (see Figure 7). As the LP1 method's α parameter increases beyond 0.5, its performance rapidly decreases. Performances in the range of α = [0, 0.5] were roughly equivalent, with α = 0.25 performing marginally better than other parameterizations. All of the algorithms easily recover the 10 common SNVs with smaller effect sizes, as evidenced by the dropoff point on the precision-recall plot after all common variants have been found, at 10/35 (~0.285) on the x-axis. None of the algorithms ranked the rare SNVs highly, and have rapidly decreasing precision after identifying approximately 70% of the causal SNVs.

5.1.2 GAW17 Data

On the semi-synthetic GAW17 data, the algorithms were evaluated in their ability to identify the 72 causal SNVs in terms of precision-recall curves and predictive performance. Overall, all algorithms had poor precision-recall performance, as shown in Figure 8. The LP1 algorithm performed somewhat better than the other methods, but still had low precision. The LP3 algorithm did no better than random. Many SNP genotypes were not observed in the data, due to the rarity of the alleles. These disconnected nodes in the graph retain their initial scores of 0, and are all indistinguishable in rank.

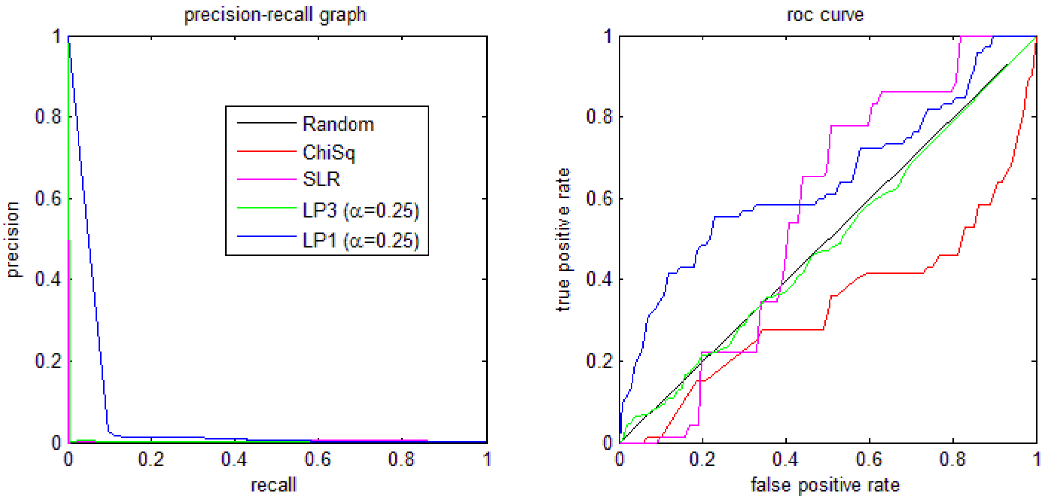

Figure 8 - Precision-recall and ROC results for GAW17 data.

The poor PR performance is due to the observation that the majority of the causal SNVs are very rare in this dataset, and some of them are private mutations in that the minor homozygote occurs only in one individual. In addition, the phenotypic model used in this dataset is multivariate and complex.

Without the ability to recover the causal variants, the predictive performance is also very poor, as is expected. None of the algorithms do better than random classification until at least 50 SNVs are used, and even then it is only a marginal improvement. The LP1 algorithm performs slightly better than chi squared and SLR for 100, 500, and 1000 features, but the difference is not significant. Classification AUCs plateau around 0.6 as the number of SNVs added to the classifier is increased. Table 5 Table 5shows the *k*NN prediction AUCs for the GAW17 data.

Table 5 - Prediction AUCs on GAW17 data.

Method	***Features Used***							
	1	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.4995 ±0.0595	0.5006 ±0.0599	0.5416 ±0.0601	0.5415 ±0.0607	0.5628 ±0.0603	0.5703 ±0.0591	0.5867 ±0.0650	0.5717 ±0.0656
SLR	0.5030 ±0.0597	0.5311 ±0.0601	0.5401 ±0.0601	0.5412 ±0.0607	0.5598 ±0.0607	0.5775 ±0.0591	-	-
LP3 **(*α* = 0.25)**	0.4934 ±0.0604	0.5128 ±0.0605	0.5015 ±0.0602	0.5184 ±0.0605	0.5244 ±0.0620	0.6051 ±0.0622	0.6001 ±0.0622	0.5524 ±0.0607
LP1 **(*α* = 0.25)**	0.5134 ±0.0605	0.5311 ±0.625	0.5311 ±0.0625	0.5311 ±0.0627	0.5650 ±0.0607	0.5806 ±0.0627	0.5963 ±0.0631	0.6058 ±0.0654

5.1.3 GWAS Datasets

The algorithms were evaluated on a total of nine GWAS datasets in terms of predictive performance, in terms of biological significance of the top-ranked SNPs (evidence of documented association with disease), and in terms of computational efficiency. In addition, LP1 was evaluated on the two LOAD datasets for feature reproducibility.

5.1.3.1 Predictive Performance

Each feature ranking algorithm was applied to each GWAS dataset, and the top-ranked SNPs were used to construct a *k*-nearest neighbor classifier. This was done using fivefold cross-validation, repeating the process of selecting features on four training folds and classifying the samples in the remaining test fold. Experiments were conducted where classifiers were constructed with increasing number of top-ranked SNPs; the number of SNPs ranged from 1 top-ranked SNP to 1000 top-ranked SNPs. The AUCs from all experiments, along with confidence intervals, are given in Appendix A. A portion of these results are presented in Table 6, which shows the AUC results for classifiers that were constructed using a small to moderate number of top-ranked SNPs (5, 10, 50 and 100 SNPs).

Table 6 - Prediction AUCs for nine GWAS datasets.

Features	***Algorithm***	***Dataset***								
		TGen	**ADRC**	**BD**	**CAD**	**CD**	**HT**	**RA**	**T1D**	**T2D**
5	**ChiSq**	0.7220	0.7433	0.6056	0.7702	0.6225	0.5519	0.7013	0.7448	0.7780
	SLR	0.7291	0.7100	0.5975	0.6291	0.5859	0.5756	0.6260	0.7346	0.6751
	LP3	0.7088	0.7342	0.5478	0.6612	0.5414	0.4959	0.7013	0.6891	0.6903
	LP1	0.7230	0.7433	0.6270	0.7668	0.6288	0.5601	0.7028	0.7448	0.7787
10	**ChiSq**	0.7394	0.7184	0.5846	0.8201	0.6346	0.5530	0.7273	0.7293	0.7329
	SLR	0.7424	0.7354	0.6166	0.7840	0.6285	0.5885	0.6352	0.7202	0.7424
	LP3	0.7369	0.7315	0.5532	0.7109	0.5548	0.4904	0.7284	0.6900	0.6907
	LP1	0.7118	0.7058	0.6137	0.8372	0.6371	0.5711	0.7282	0.7329	0.7622
50	**ChiSq**	0.7060	0.6438	0.5372	0.5643	0.5796	0.5263	0.6107	0.6562	0.5707
	SLR	0.7264	0.6970	0.5874	0.6568	0.6027	0.5774	0.5852	0.6841	0.7258

	LP3	0.7519	0.7151	0.5219	0.5860	0.5314	0.5060	0.6398	0.6678	0.6071
	LP1	0.6694	0.6264	0.5348	0.6254	0.5791	0.5684	0.6377	0.6704	0.5804
100	**ChiSq**	0.6574	0.6034	0.5300	0.5368	0.5950	0.5362	0.5884	0.5905	0.5611
	SLR	-	0.6874	0.5742	0.5993	0.6188	0.5671	0.5719	0.6639	0.6671
	LP3	0.7286	0.7154	0.5190	0.5420	0.5146	0.4921	0.5946	0.6301	0.5773
	LP1	0.6473	0.5862	0.5328	0.5452	0.5872	0.5391	0.6009	0.6054	0.5614

A moderately predictive genetic signal is found in each dataset, with peak AUCs ranging from 0.62 to 0.83. The LP1 algorithm is comparable with the chi squared and SLR algorithms, outperforming either one in select cases. The LP1 algorithm is an improvement over the LP3 algorithm. LP3 does only slightly worse than LP1 on some datasets, but fails to improve on random classification for the BD, CD, and HT datasets. The SLR method selects 500 or fewer variants in each dataset, showing good classification performance over its operable region. However, SLR is somewhat slower to increase AUC as the number of features increases, sometimes requiring 2-5 features to move beyond random classification, while the other algorithms tend to select predictive variants in the top 2.

To statistically compare the predictive performance of the different algorithms, I computed the paired Wilcoxon signed-rank test over the GWAS AUCs. For each algorithm I used two rows of AUCs from Table 6 (two different feature selection thresholds for all 9 datasets), yielding 18 AUC samples over which to compare algorithms. Two feature selection thresholds were used in order to increase sample size enough to have a meaningful p-value, even though the AUCs are not totally independent from one feature selection threshold to another. I examined feature selection thresholds of 5 and 10 SNPs (Table 7), as well as 50 and 100 SNPs (Table 8). These represent characteristic regions when using few SNVs, and when using a moderate number of SNPs. A large number of SNPs (i.e., 500 and 1000 SNPs) was not examined, because all algorithms tested tend to have diminishing performance in this range.

Below are all pairwise Wilcoxon comparisons, with the better-performing algorithm listed in each case along with the *p*-value of the comparison result.

Table 7- Wilcoxon signed-ranked test p-value comparing AUCs of algorithms over 9 GWAS datasets for feature selection thresholds of 5 and 10.

	ChiSq	**SLR**	**LP3**
SLR	ChiSq (*p*=0.00906)	-	-
LP3	ChiSq (*p*=0.00124)	LP3 (*p*=0.06432)	-
LP1	LP1 (*p*=0.00596)	LP1 (*p*=0.00906)	LP1(*p*=0.00096)

Table 8 - Wilcoxon signed-ranked test p-value comparing AUCs of algorithms over 9 GWAS datasets for feature selection thresholds of 50 and 100.

	ChiSq	**SLR**	**LP3**
SLR	SLR (*p*=0.0035)	-	-
LP3	ChiSq (*p*=0.1096)	SLR (*p*=0.00096)	-
LP1	LP1 (*p*=0.01108)	SLR (*p*=0.01782)	LP1 (*p*=0.00528)

For small feature set sizes, the LP1 algorithm outperforms all other algorithms by a significant margin. The chi squared algorithm has the next-best performance, outperforming the SLR and LP3 methods. The SLR algorithm has the worst performance when using only 5 or 10 features, but is not significantly different from LP3. When using 50 or 100 features, however, the SLR method performs the best by a significant margin. The LP1 algorithm outperforms all other methods besides SLR when using a larger feature set size. However, when correcting for multiple testing, LP1 is not significantly different from the chi squared test or SLR in the 50 to 100 feature range.

In addition to performing feature selection and classification on each dataset individually, I also performed cross-dataset experiments for the LOAD datasets. In these experiments, I filtered both LOAD datasets to the same set of 64,984 SNPs, then selected features on one dataset and used the features to construct a classifier on the other dataset. The results are shown in Table 9.

Table 9 - Prediction AUCs from selecting features on one LOAD dataset and predicting on the other.

Dataset	*# SNPs*	*Method*	*Number of SNPs used in classifier*							
			1	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
TGen (Feature selection from ADRC)	**64,984 (ADRC overlap, chr1-22)**	**Chi Sq**	0.6086 ±0.0294	0.6863 ±0.0280	0.7099 ±0.0270	0.6958 ±0.0253	0.6563 ±0.0286	0.6097 ±0.0296	0.5593 ±0.0310	0.5563 ±0.0308
		SWRF	0.5952 ±0.0296	0.6980 ±0.0274	0.6994 ±0.0272	0.7005 ±0.0274	0.6756 ±0.0284	0.6677 ±0.0284	0.5635 ±0.0306	0.5195 ±0.0310
		SLR	0.6086 ±0.0294	0.6863 ±0.0280	0.7164 ±0.0269	0.7289 ±0.0263	0.6522 ±0.0292	0.6084 ±0.0300	-	-
		LP3	0.5023 ±0.0306	0.6039 ±0.0300	0.7023 ±0.0272	0.7037 ±0.0274	0.6888 ±0.0276	0.6543 ±0.0286	0.6114 ±0.0298	0.5690 ±0.0306
		LP1	0.6086 ±0.0294	0.6863 ±0.0280	0.7069 ±0.0271	0.7058 ±0.0256	0.6643 ±0.0276	0.6497 ±0.0270	0.5893 ±0.0299	0.5443 ±0.0307
ADRC (Feature selection from TGen)	**64,984 (TGen overlap, chr1-22)**	**Chi Sq**	0.6172 ±0.0231	0.6385 ±0.0229	0.7419 ±0.0204	0.7362 ±0.0208	0.6695 ±0.0225	0.6479 ±0.0227	0.5396 ±0.0239	0.5259 ±0.0122
		SWRF	0.5397 ±0.0239	0.5345 ±0.0241	0.5350 ±0.0241	0.5401 ±0.0243	0.5042 ±0.0243	0.5257 ±0.0241	0.5201 ±0.0241	0.5053 ±0.0241
		SLR	0.5397 ±0.0214	0.7006 ±0.0214	0.7003 ±0.0214	0.7048 ±0.0216	0.6048 ±0.0233	0.5854 ±0.0237	-	-
		LP3	0.5397 ±0.0239	0.6021 ±0.0235	0.7283 ±0.0210	0.7366 ±0.0208	0.6853 ±0.0220	0.6598 ±0.0225	0.5678 ±0.0239	0.5306 ±0.0239
		LP1	0.6172 ±0.0231	0.6385 ±0.0229	0.7422 ±0.0203	0.7364 ±0.0206	0.6702 ±0.0218	0.6479 ±0.0227	0.5377 ±0.0236	0.5264 ±0.00241

The results on the cross-dataset experiments are very similar to the results when performing feature selection and classification on the same dataset. This indicates that the selected features have predictive power outside of each individual cohort, suggesting that the top-ranked features are indeed generalizable.

5.1.3.2 Biological Validity

The LP1 algorithm was applied to each GWAS dataset, and the top-ranked features were examined for biological validity. Due to the time required to manually validate each SNP, the control algorithms' results were generally only examined for the Alzheimer's disease datasets. The chi squared algorithm was also evaluated on the CD and HT datasets. Table 10 summarizes the validation results, listing the number of SNPs having evidence of association in the literature for each dataset.

Table 10 - Summary of literature validation results.

Dataset	**Algorithm**	**SNPs validated (% precision)**
TGen	Chi Sq	6 (24%)
	SWRF	5 (20%)
	SLR	7 (28%)
	LP3	14 (56%)
	LP1	11 (44%)
ADRC	Chi Sq	10 (40%)
	SWRF	2 (8%)
	SLR	5 (20%)
	LP3	10 (40%)
	LP1	12 (48%)
BD	LP1	17 (68%)
CAD	LP1	11 (44%)
CD	Chi Sq	20 (80%)
	LP1	19 (76%)
HT	Chi Sq	9 (36%)
	LP1	12 (48%)
RA	LP1	11 (44%)
T1D	LP1	7 (28%)
T2D	LP1	15 (60%)
Total	LP1	108 (48%)

The LP1 algorithm returned many biologically validated SNPs, overall finding evidence for 48% of the 225 SNPs examined over the nine datasets. On the LOAD datasets, the LP1 and LP3 algorithms outperformed the control methods. Applying a *z*-test to compare proportions, LP3 significantly outperforms ChiSq, SWRF and SLR on the TGen dataset, and LP1 significantly outperforms SWRF and SLR on the ADRC dataset. On the CD and HT datasets, the LP1 method is statistically equivalent to the chi squared algorithm, and in fact returns many of the same top-ranked SNPs. Below are the top 25 SNPs as ranked by LP1 for each dataset (Table 11 through Table 19). The control algorithms' results are given in Appendix B.

Table 11 - Top 25 SNPs for TGen using LP1 ($\alpha = 0.25$).

Rank	rsID	Chr	Gene	Notes
1	rs429358	19	APOE	APOE risk allele determined by rs7412 and rs429358 [53]
2	rs4420638	19	APOC	In strong linkage disequilibrium with APOE SNPs [54]
3	rs7412	19	APOE	APOE risk allele determined by rs7412 and rs429358 [53]
4	rs10824310	10	PRKG1	Significant association with LOAD [55]
5	rs7079348	10	C10orf11	-
6	rs3732443	3	GXYLT2	-
7	rs582790	11	-	-
8	rs934745	18	MAPK4	-
9	rs7964760	12	NAV3	NAV3 mRNA levels elevated in AD brains [56]
10	rs16974268	15	SLCO3A1	-
11	rs10499687	7	VWC2	-
12	rs6717497	2	-	-
13	rs17330779	7	NRCAM	Associated with axonal degeneration in LOAD [57]
14	rs3007246	13	-	-
15	rs7077757	10	RBM20	Meta-analysis of multiple studies showed association [58]
16	rs12162084	16	-	Significant association with LOAD [59]
17	rs3905173	1	-	Associated in Bayesian analysis of TGen data [60]
18	rs17048190	2	-	-
19	rs2968848	7	-	-
20	rs12041702	1	-	-
21	rs9934599	16	IL34	-
22	rs950922	1	ALPL	-
23	rs10115381	9	-	-
24	rs1038891	11	LRRC4C	SNP associated with LOAD in genome-wide analysis [61]
25	rs9398855	6	THEMIS	Possible gene association via GAB2 pathway [62]

Table 12 - Top 25 SNPs for ADRC using LP1 ($\alpha = 0.25$).

Rank	rsID	Chr	Gene	Notes
1	rs439401	19	APOE	In strong LD with rs7412 and rs429358 [63]
2	rs5157	19	APOC4	In strong LD with other APOC risk SNPs [64]
3	rs157582	19	TOMM40	Showed LOAD association in African-American cohort [65]
4	rs2075650	19	APOE4	Predictive of longevity of LOAD patients [66, 67]
5	rs445925	19	-	Located between APOE and APOC genes [68]
6	rs8106922	19	TOMM40	Meta-analysis finds significant association with LOAD [69]
7	rs11076978	16	-	-
8	rs405509	19	APOE	APOE promoter varies LOAD risk [70]
9	rs157580	19	TOMM40	Associated with LOAD in Chinese population [71]
10	rs3738269	1	IGFN1	-
11	rs832156	1	IGFN1	-
12	rs12507679	4	STAP1	-
13	rs26845	16	ECI1	-
14	rs13132585	4	STAP1	-
15	rs17428956	1	-	-
16	rs10994553	10	-	-
17	rs9909412	17	COX10	Gene differentially expressed in LOAD patients [72]
18	rs206081	13	BRCA2	-
19	rs4558873	4	SORCS2	Gene differentially expressed in LOAD patients [73]
20	rs151716	1	-	-

21	rs12140610	1	-	-
22	rs9487940	6	-	-
23	rs279877	9	DMRT3	CNV in related DMRT1 gene associated with LOAD [74]
24	rs537761	1	-	-
25	rs8082842	18	RAB31	Gene involved in potential treatment [75]

Table 13 - Top 25 SNPs for BD using LP1 ($\alpha = 0.25$).

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs1909936	8	-	Significant in RF analysis of WTCCC [76]
2	rs7653441	3	FNDC3B	Significant association in independent cohort [77]
3	rs6577370	1	-	-
4	rs17116117	11	HTR3B	Gene mutations disrupts serotonin regulation [78]
5	rs12355606	10	CACNB2	Associated in Han Chinese population [79]
6	rs1442650	18	LINC00907	-
7	rs12938916	17	-	Discovered in nonparametric analysis of WTCCC data [80]
8	rs7260296	19	-	-
9	rs11059460	12	-	-
10	rs420259	16	PALB2	SNP associated with BD in Scandinavian cohort [81]
11	rs2837588	21	DSCAM	Gene associated with BD [82]
12	rs858719	11	ZBTB44	Significant in WTCCC study [83]
13	rs914715	11	ZBTB44	Significant in WTCCC study [83]
14	rs2953146	2	RNPEPL1	Significant in WTCCC study [83]
15	rs16857512	1	CACNA1E	Gene implicated in treatment efficacy [84]
16	rs514636	3	LAMP3	Discovered in secondary analysis of WTCCC [80]
17	rs2683780	3	-	-
18	rs6458307	6	-	BD association discovered in Finnish population [85]
19	rs6414500	3	LAMP3	Discovered in secondary analysis of WTCCC [80]
20	rs6414498	3	LAMP3	Discovered in secondary analysis of WTCCC [80]
21	rs12472797	2	-	-
22	rs12980129	19	-	Discovered as part of epistatic interactions in WTCCC data [86]
23	rs7152966	14	-	-
24	rs9318400	13	-	-
25	rs682970	10	CELF2	Gene associated with major depression [87]

Table 14 - Top 25 SNPs for CAD using LP1 ($\alpha = 0.25$).

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs4799934	16	PALB2	-
2	rs7906587	10	PNLIPRP3	Gene associated with mean arterial pressure [88]
3	rs11671119	19	MEF2BNB-MEF2B	Gene associated with cardiac development [89]
4	rs17042882	3	PLCL2	Gene expression modified by cardiovascular disease risk reduction therapy [90]
5	rs159171	5	-	-
6	rs16955238	16	-	-
7	rs16891338	8	SAMD12-AS1	-
8	rs16908145	8	FLJ45872	-
9	rs6989092	8	-	-
10	rs16883114	8	-	-
11	rs7653441	3	FNDC3B	Locus associated with heart rate and rhythm disorders [91]
12	rs4970605	1		Interaction found in WTCCC study [92]
13	rs17022496	4	BMPR1B	-
14	rs12724674	1	-	-
15	rs1333049	9	-	Replicated in German, Japanese and Korean populations [93, 94]
16	rs9884478	4	NPFFR2	-
17	rs17146094	7	EIF4H	Discovered in exhaustive epistatic analysis of WTCCC data [95]
18	rs906766	3	MED12L	-

19	rs7002837	8	-	-
20	rs17672135	1	FMN2	SNP discovered in independent experiments [96, 97]
21	rs4846770	1	MIA3	Meta analysis implicated gene [98]
22	rs326296	3	-	-
23	rs6490506	13	ZMYM2	-
24	rs523096	9	CDKN2B-AS1	WTCCC gene replicated in independent German cohort [93]
25	rs518394	9	CDKN2B-AS1	WTCCC gene replicated in independent German cohort [93]

Table 15 - Top 25 SNPs for CD using LP1 (α = 0.25).

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs17116117	11	HTR3B	-
2	rs10483456	14	RALGAPA1	Associated with CD in independent study of Chinese individuals [99]
3	rs11209026	1	IL23R	Association found in multiple studies [100, 101]
4	rs2076756	16	NOD2	Associated in independent study [102]
5	rs10210302	2	ATG16L1	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103]
6	rs6752107	2	ATG16L1	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103]
7	rs6431654	2	ATG16L1	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103]
8	rs3828309	2	ATG16L1	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103, 104]
9	rs17234657	5	-	WTCCC finding replicated in independent cohorts [105, 106]
10	rs2066843	16	NOD2	Associated with CD in independent study [102]
11	rs3792106	2	ATG16L1	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103, 104]
12	rs11805303	1	IL23R	IL23 implicated in CD [107]
13	rs11957215	5	-	-
14	rs9292777	5	-	WTCCC SNP replicated in independent study [108]
15	rs17221417	16	NOD2	NOD2 implicated in CD [102]
16	rs10489629	1	IL23R	Replicated in multiple studies [102, 109, 110]
17	rs2201841	1	IL23R	SNP implicated in distinct populations [111]
18	rs4957295	5	-	-
19	rs10213846	5	-	-
20	rs6871834	5	-	-
21	rs4957297	5	-	Replicated independent of WTCCC [112]
22	rs4957300	5	-	-
23	rs16869934	5	-	Discovered in BIC analysis of WTCCC [113]
24	rs12119179	1	-	Associated with a disease with similar genetic profile [114]
25	rs11209033	1	-	Cited in patent for testing for autoimmune-associated polymorphisms [115]

Table 16 - Top 25 SNPs for HT using LP1 (α = 0.25).

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs4765066	12	-	-
2	rs488101	9		Associated with arterial plaque [116]
3	rs4867173	5	-	-
4	rs11782342	8	KCNB2	Discovered as part of epistatic interactions in WTCCC data [86]
5	rs11024327	11	OTOG	Found in combined analysis of WTCCC data [117]
6	rs16857512	1	CACNA1E	Calcium gate channels implicated in BP regulation [118]
7	rs2820037	1	-	SNP associated with BP regulation [119]
8	rs2790622	1	-	-
9	rs2820038	1	-	-
10	rs6574988	14	-	-
11	rs2820046	1	-	-
12	rs16945811	17	YWHAE	Gene implicated in HT [120]
13	rs9428826	1	-	-
14	rs2398162	15	NR2F2-AS1	Population-specific association has been replicated [121]
15	rs2820026	1	-	-
16	rs921535	15	-	-

17	rs17018584	4	CCSER1	SNP associated with heart disease in rats and humans[122]
18	rs10889923	1	NEGR1	-
19	rs1022684	20	SEC23B	-
20	rs2191003	4	-	-
21	rs41515647	1	ST6GALNAC5	Gene associated with heart disease [123]
22	rs300916	4	GAB1	Discovered in combined WTCCC + Australian cohort [117]
23	rs1935683	6	-	-
24	rs13119672	4	PPARGC1A	Gene associated with HT [124]
25	rs17201619	17	-	Discovered in combined WTCCC + Australian cohort [117]

Table 17 - Top 25 SNPs for RA using LP1 (α = 0.25)

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs4718582	7	-	-
2	rs12670243	7	-	-
3	rs9271850	6	-	SNP associated in Swedish study [125]
4	rs17104722	14	-	-
5	rs6679677	1	PHTF1	Approached significance in low-power, independent study [126]
6	rs1733717	10	MBL2	SNP associated with risk of RA [127]
7	rs1369036	1	-	-
8	rs3129768	6	-	-
9	rs2282859	6	FGFR1OP	-
10	rs1230666		-	-
11	rs9272346	6	HLA-DQA1	Gene implicated in Taiwanese population [128]
12	rs1711029	15	-	-
13	rs2943570	8	-	-
14	rs16874205	8	-	Replicated in Spanish population [129]
15	rs2488457	1	PTPN22	Replicated in Caucasian but not Korean population[130]
16	rs11776005	8	-	-
17	rs1028850	13	LINC00598	-
18	rs10834744	11	ART1	-
19	rs9272723	6	HLA-DQA1	Gene implicated in Taiwanese population [128]
20	rs1217396	1	RSBN1	Marginal association found in independent cohort [131]
21	rs1230649	1	PHTF1	-
22	rs1230658	1	MAGI3	Gene associated with RA [132]
23	rs1217200	1	MAGI3	Gene associated with RA [132]
24	rs1562694	14	-	-
25	rs2837960	21	-	Validated in independent study [133]

Table 18 - Top 25 SNPs for T1D using LP1 (α = 0.25)

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs9272346	6	HLA-DQA1	Associated in WTCCC data [83]
2	rs3129768	6	-	-
3	rs6679677	1	PHTF1	Associated with T1D, may be in strong LD with causal variant [134]
4	rs17116117	11	HTR3B	-
5	rs9989228	14	MIPOL1	-
6	rs17696736	12	NAA25	In LD with possible causal SNPs, validated in independent cohort [135]
7	rs11171739	12	-	In LD with possible causal SNPs, validated in independent cohort [135]
8	rs10483456	14	RALGAPA1	-
9	rs6894569	5	-	-
10	rs765534	11	-	-
11	rs1977	6	BTN3A2	Multilocus analysis implicated gene [136]
12	rs9358932	6	-	-
13	rs7745603	6	-	-
14	rs10494787	1	FLJ43585	-

15	rs9393713	6	BTN3A2	Multilocus analysis implicated gene [136]
16	rs2237236	6	BTN3A3	-
17	rs1873914	12	RAB5B	-
18	rs1343125	1	MAGI3	-
19	rs9393848	6	-	-
20	rs9468203	6	-	-
21	rs3734536	6	BTN3A2	-
22	rs7776351	6	-	-
23	rs4711165	6	ZKSCAN8	-
24	rs12708716	16	CLEC16A	Gene associated with T1D, SNP discovered in WTCCC [137]
25	rs17711344	6	-	-

Table 19 - Top 25 SNPs for T2D using LP1 (α = 0.25)

Rank	**rsID**	**Chr**	**Gene**	**Notes**
1	rs3777582	6	CLIC5	-
2	rs11042656	11	SBF2	-
3	rs1477523	7	AC009264.1	Linked to T2D through HDL regulation [138]
4	rs17116117	11	HTR3B	-
5	rs17117531	15	-	-
6	rs10492267	12	-	Found in Bayesian analysis of WTCCC data[139]
7	rs4506565	10	TCF7L2	Validated in multiple distinct populations [140-142]
8	rs10483456	14	RALGAPA1	Linked to T2D through HDL regulation [138]
9	rs7193144	16	FTO	Validated in Indian population [143, 144]
10	rs9405484	6	LOC102723944	-
11	rs9939609	16	FTO	Replicated in Norwegian population [145]
12	rs7917983	10	TCF7L2	Validated in Indian population [146, 147]
13	rs13373826	1	SLC44A5	SNP associated with T2D [148]
14	rs1025450	18	-	-
15	rs7901275	10	TCF7L2	Discovered in independent study [149]
16	rs9926289	16	FTO	Associated with obesity Polish population [150]
17	rs9465871	6	CDKAL1	Validated in Chinese population [151]
18	rs9939973	16	FTO	Implicated in obesity [152]
19	rs9940128	16	FTO	Linked to T2D in Indian population [153]
20	rs9367532	6	-	-
21	rs1121980	16	FTO	Validated in Swedish population [154]
22	rs1957779	14	RHOJ	-
23	rs9930506	16	FTO	Implicated in obesity [152]
24	rs358806	3	-	-
25	rs903228	2	-	-

5.1.3.3 Computational Efficiency

The LP1 algorithm is no more computationally complex than the original LP3 algorithm, requiring only a single extra iteration of propagation to collect the soft labels in a contingency table, and is $O(kNd)$. The SLR algorithm's computational complexity is at least $O(N^2d)$, stemming from the matrix inversions necessary to compute the regression coefficients. The chi squared algorithm is simply $O(Nd)$. All algorithms were benchmarked on one fold of the BD

WTCCC data, running on a 2.33 GHz processor with 8GB of RAM available and excluding time required to read in the data files. The chi squared test ran the fastest, in about 11 minutes. The SLR method was slowest, taking nearly 3 days to complete. The LP1 method ran fairly quickly for a multivariate algorithm, requiring 42 minutes. The SWRF method was not applied to the large-scale GWAS datasets, in light of the fact that the increased sample size would have increased the computation time to at least several weeks for a single fold.

5.1.3.4 Feature Reproducibility on LOAD data

For the ranking reproducibility experiments, I filtered the two LOAD datasets so they contain only the intersection of the features. There are 64,984 SNPs in common between the two datasets which were ranked by each algorithm, and the intersection of these rankings is compared.

Figure 9 shows the reproducibility results on the LOAD datasets. Chi squared identifies the first few SNPs reproducibly; these are SNPs that are located in genes apolipoprotein-E (APOE) and apolipoprotein-C (APOC) and are known to have large effects sizes. Beyond the first few SNPs, however, the reproducibility of chi squared drops rapidly to a level which is effectively random. The SWRF algorithm produces results that are no better than random for the genome-wide datasets. The implicitly selective SLR method is not shown on this graph because only two features overlap in the selected subset, yielding virtually no reproducibility.

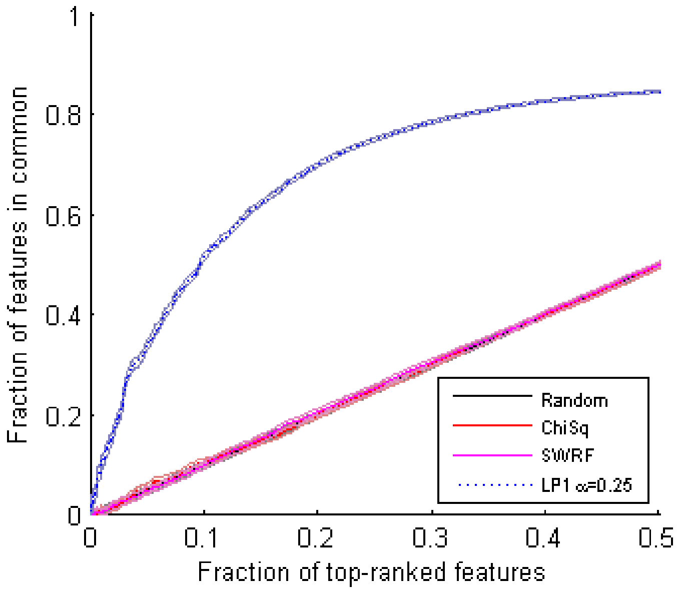

Figure 9 - Reproducibility curves of top-ranked features with 95% confidence envelope.

The x-axis shows the fraction of top-ranked features being considered, and the y-axis shows the fraction of features in common to rankings obtained from each of the two datasets independently (TGen and ADRC). For this plot, the chi squared and SWRF methods are virtually indistinguishable from the random performance curve along the diagonal.

LP1, in contrast to these two methods, shows good reproducibility for many of the top-ranked SNPs, and does so even in the high-dimensional datasets. The algorithm has low reproducibility for the first few SNPs but quickly surpasses chi squared and SWRF.

5.2 EVALUATION OF LP1 WITH PRIOR KNOWLEDGE

This section presents the results of the experiments on the second algorithmic extension described in 3.4.2.2. Two methods of incorporating knowledge were initially proposed: an edge weighting method and a prior pseudocount method. After numerous experiments with the edge weighting method, it was ultimately found to be ineffective. Even drastic changes in the network edge weights had little impact on the final LP1 score. SNP rankings remained virtually unchanged, with only miniscule changes to SNP *p*-value scores. It appears the edge weights do not have enough impact on the labeling of the sample nodes, from which the LP1 score is directly computed. By changing the weight of all edges connecting to a particular SNP identically, the overall proportion of signal coming from cases or controls remains unchanged. One solution might be to have differential weighting of cases versus controls for SNPs with a prior likelihood of association. This would require knowledge (or assumption) of the genetic model underlying the association.

The prior pseudocount method, in contrast, resulted in promising changes to the rankings. Prior knowledge experiments were performed on the synthetic, semi-synthetic, and GWASdatasets. For the GWAS datasets, I used GERP and MAF prior knowledge scores corresponding to the SNPs in the data. For the synthetic dataset, the variables have no biological meaning, so I instead used a “gold standard” prior knowledge. This prior gives the maximum prior to the 35 true causal variants, and a null prior to all other variants. The semi-synthetic dataset contains real SNPs, but the phenotypic model does not take into account evolutionary conservation. Because of this, I use a gold standard prior which upweights the 72 true causal SNVs, as well as the MAF prior. The use of gold standard priors is useful for testing the algorithmic validity of the method, because the prior values are known to correspond to meaningful variants. For the GWAS datasets, a lack of good performance could either be a result of an invalid method, or an invalid prior. By ensuring valid priors, we can test the appropriateness of the pseudocount method.

Figure 10 shows the results of the prior knowledge experiments on the synthetic dataset, for increasing prior equivalent sample size (PESS). When PESS=0, the LP1 algorithm does not use any prior information and is equivalent to the original LP1 algorithm. As the PESS increases, the algorithm's ability to recovery the causal variants increases as well. The PESS of 1000 at first glance appears worse than the PESS of 100, but in fact has better performance in the low-precision tail region, recovering more of the rare variants as evidenced by the better ROC performance.

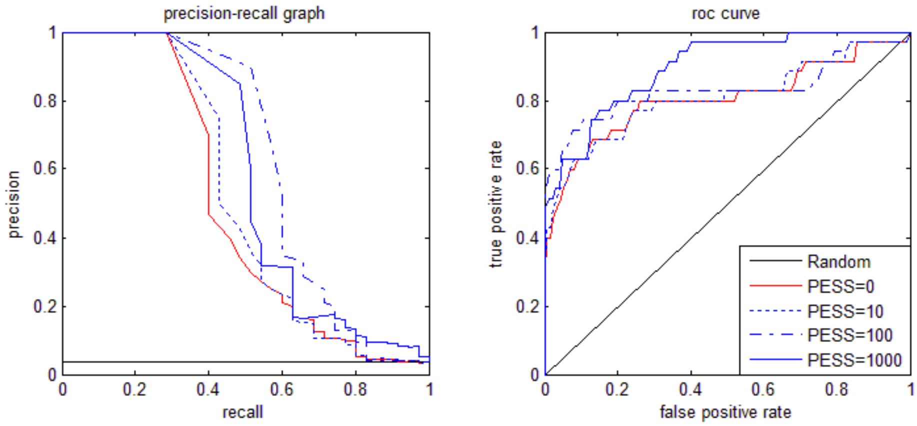

Figure 10 - Precision-recall and ROC curves for prior knowledge experiments on synthetic data.

Similar results are found on the semi-synthetic GAW17 dataset. The LP1 algorithm with PESS = 0 does especially poorly on this dataset, showing difficulty in discovering the rare causal SNPs. A small to moderate PESS of up to 1000 yields marginal improvement in the algorithm's performance. The sample size for the GAW17 is larger than the synthetic dataset, and requires a larger PESS values for improved performance.

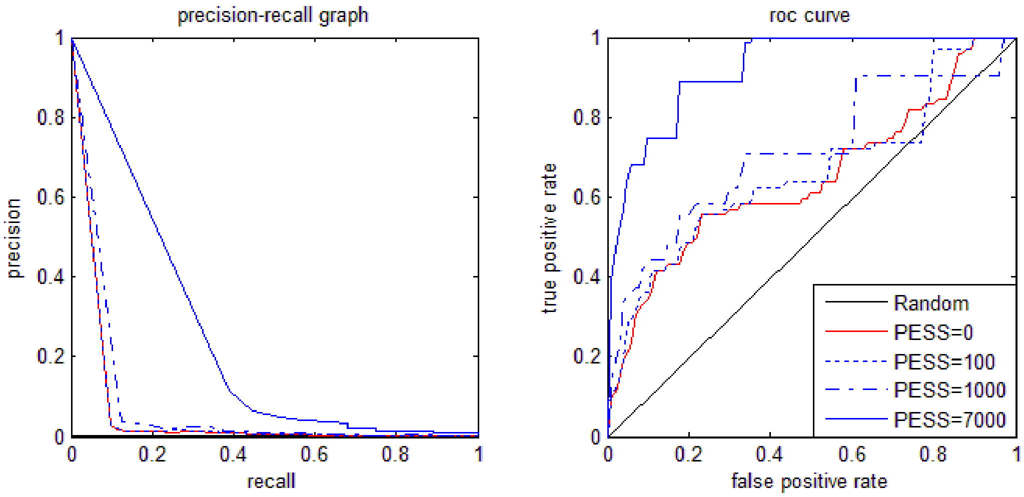

Figure 11 - Precision-recall and ROC curves for prior knowledge experiments on semi-synthetic data using gold standard prior knowledge.

The GAW17 data was also analyzed using the MAF prior, with the knowledge that many of the causal SNVs in the model are very rare. The MAF prior does not significantly improve LP1's performance on the GAW17 data, and in fact shows decrease performance for large PESS. The MAF prior's failure could be explained by the fact that not all SNVs in the causal model are the rarest variants, and also that rare SNVs which appear in only one individual (about 40% of the variants) are all given the exact same upweighting, making it difficult to distinguish among them.

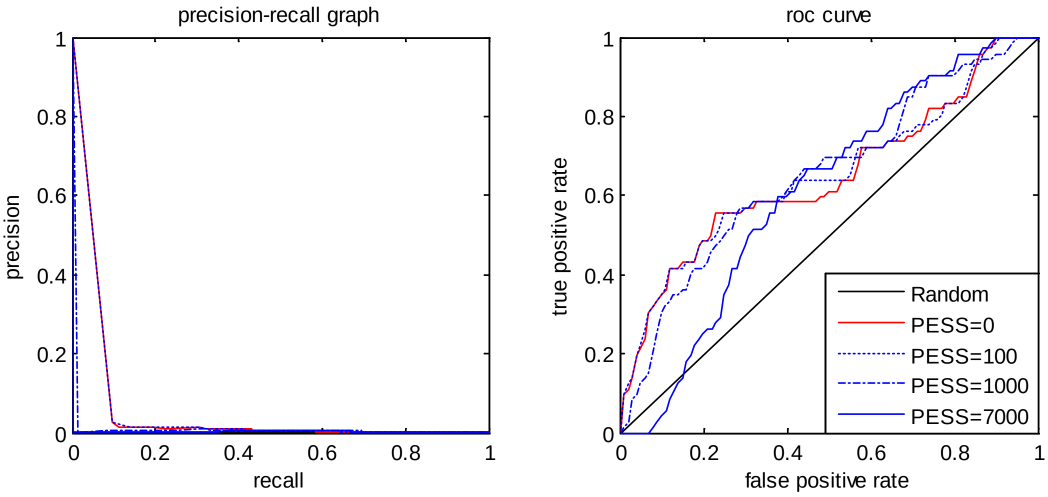

Figure 12 - Precision-recall and ROC curves for prior knowledged experiments on semi-synthetic data using MAF prior.

The prior knowledge method was applied to the GWAS datasets using both GERP and MAF priors in place of the gold standard priors used in the synthetic and semi-synthetic experiments. I tested a PESS of 50 and 500, and the classification AUC results are given in Table 20 for feature set sizes between 5 and 100. The full set of results with 95% confidence intervals is given in Appendix A.

Table 20 - Prior knowledge AUCs for nine GWAS datasets.

Features	*Algorithm*	*Dataset*								
		TGen	**ADRC**	**BD**	**CAD**	**CD**	**HT**	**RA**	**T1D**	**T2D**
5	**LP1**	0.7230	0.7433	0.6270	0.7668	0.6288	0.5601	0.7013	0.7448	0.7787
	LP1+ GERP50	0.7228	0.7433	0.6270	0.7668	0.6236	0.5625	0.7028	0.7445	0.7739
	LP1+ GERP500	0.5112	0.5043	0.5108	0.7489	0.5085	0.4946	0.6005	0.7093	0.7320
	LP1+ MAF50	0.7014	0.7345	0.6160	0.7502	0.6212	0.5522	0.6969	0.7445	0.7669
	LP1+ MAF500	0.5102	0.5068	0.5111	0.7023	0.5065	0.5002	0.5985	0.6875	0.6255
10	**LP1**	0.7118	0.7058	0.6137	0.8372	0.6371	0.5711	0.7284	0.7329	0.7622
	LP1+ GERP50	0.7145	0.7102	0.6131	0.8372	0.6358	0.5715	0.7282	0.7306	0.7551
	LP1+ GERP500	0.5108	0.5138	0.5118	0.8063	0.5465	0.4742	0.6654	0.7040	0.7573
	LP1+ MAF50	0.6983	0.7001	0.6034	0.8361	0.6316	0.5604	0.7118	0.7296	0.7485
	LP1+ MAF500	0.5129	0.5264	0.5118	0.7356	0.5115	0.4841	0.6254	0.6755	0.6548

50	**LP1**	0.6694	0.6264	0.5348	0.6254	0.5791	0.5684	0.6398	0.6704	0.5804
	LP1+ GERP50	0.6587	0.6133	0.5376	0.5950	0.5870	0.5430	0.6377	0.6624	0.5703
	LP1+ GERP500	0.5145	0.5057	0.5037	0.5903	0.5046	0.4987	0.5914	0.6383	0.5715
	LP1+ MAF50	0.6455	0.6023	0.5485	0.6055	0.5901	0.5497	0.6291	0.6544	0.5693
	LP1+ MAF500	0.5123	0.5148	0.5024	0.5842	0.5048	0.4964	0.5724	0.6212	0.5685
100	**LP1**	0.6478	0.5862	0.5328	0.5452	0.5872	0.5391	0.5946	0.6054	0.5614
	LP1+ GERP50	0.6426	0.5789	0.5278	0.5514	0.5805	0.5364	0.6009	0.5983	0.5640
	LP1+ GERP500	0.5088	0.5102	0.5095	0.5307	0.5073	0.4905	0.5698	0.6077	0.5455
	LP1+ MAF50	0.6326	0.5643	0.5153	0.5489	0.5640	0.5456	0.6015	0.5838	0.5542
	LP1+ MAF500	0.5101	0.5089	0.5084	0.5267	0.5013	0.4921	0.5448	0.6014	0.5326

For PESS = 50, the downstream classification performance is diminished, but not significantly. The PESS of 500, however, is a severe detriment to the classification performance. On some datasets, the prior masks the true genomic signal, leading to poor feature selection and essentially random classification. For some datasets, the PESS of 500 results in a moderate drop in AUC, but does not completely mask the genomic signal.

To statistically compare the performance of the prior method on GWAS data, I computed the paired Wilcoxon signed-rank test over the GWAS AUCs. Using each of the nine GWAS datasets, I used two rows of the AUCs in Table 20 (corresponding to two different feature selection thresholds), yielding 18 paired AUC samples over which to compare algorithms. Below are the pairwise Wilcoxon comparisons with the better-performing algorithm listed in each case along with the *p*-value of the comparison result.

Table 21 - Wilcoxon signed-ranked test p-value comparing AUCs of algorithms over 9 GWAS datasets for feature selection thresholds of 5 and 10.

	LP1	**LP1+GERP50**	**LP1+GERP500**	**LP1+MAF50**
LP1+GERP50	LP1 (p = 0.2801)	-	-	-
LP1+GERP500	LP1 (p = 0.0002)	LP1+GERP50 (p = 0.0002)	-	-

LP1+MAF50	LP1 (*p* = 0.0002)	LP1+GERP50 (*p* = 0.003)	LP1+MAF50 (*p* = 0.0002)	-
LP1+MAF500	LP1 (*p* = 0.0002)	LP1+GERP50 (*p* = 0.0002)	LP1+GERP500 (*p* = 0.0548)	LP1+MAF50 (*p* = 0.0002)

Table 22 - Wilcoxon signed-ranked test p-value comparing AUCs of algorithms over 9 GWAS datasets for feature selection thresholds of 50 and 100.

	LP1	**LP1+GERP50**	**LP1+GERP500**	**LP1+MAF50**
LP1+GERP50	LP1 (*p* = 0.0155)	-	-	-
LP1+GERP500	LP1 (*p* = 0.0002)	LP1+GERP50 (*p* = 0.0005)	-	-
LP1+MAF50	LP1 (*p* = 0.0278)	LP1+GERP50 (*p* = 0.2501)	LP1+MAF50 (*p* = 0.001)	-
LP1+MAF500	LP1 (*p* = 0.0002)	LP1+GERP50 (*p* = 0.0002)	LP1+GERP500 (*p* = 0.0069)	LP1+MAF50 (*p* = 0.0003)

The prior knowledge method performs significantly worse than the original LP1 algorithm with no prior knowledge. A larger PESS value gives significantly worse performance than the smaller PESS value. This indicates either that a useful prior is not being used, or that it is being incorporated incorrectly. A discussion of possible causes and potential solutions can be found in Section 6.1.2. In addition, the GERP score generally outperforms the MAF when using the same PESS.

5.3 EVALUATION OF LP1 WITH RANKING COMBINATION

The ranking combination experiments combined the LP1 feature scores and ranks with the chi squared test's scores and ranks, as well as SLR's method's ranks. The combination method was tested on the synthetic dataset and the GWAS datasets. The method was not tested on the

GAW17 semi-synthetic dataset, due to overall poor performance of each algorithm on this dataset. Figure 13Figure 13 shows the precision-recall and ROC results of the combination methods on the synthetic data.

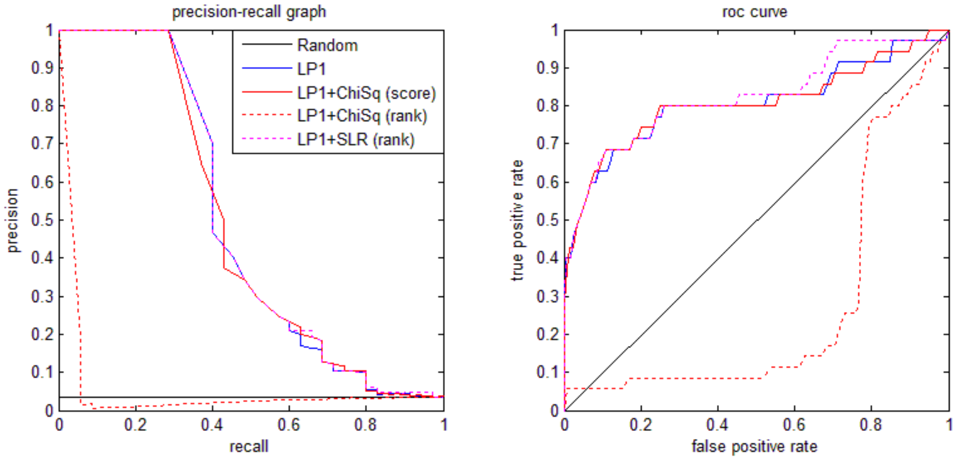

Figure 13 - Combination method results on synthetic data.

The original LP1 method's performance is almost indistinguishable the score-based combination with the chi squared test. The rank-based LP1+ChiSq method, in contrast, performs quite poorly. The LP1+SLR rank-based method also performs very similarly to the LP1 method.

The results of the combination methods on the GWAS datasets are shown in Table 23 for 5 through 100 features. The LP1+ChiSq score-based method once again has fairly similar results to the LP1 method alone. The rank based method does not change much for the first few feature set sizes (i.e. 1, 2, and 5), but leads to slightly diminished performance for larger feature sets. Compared to LP1, the LP1+SLR rank-based method had some diminished performance for small feature set sizes (where SLR performs poorly), but improved for larger feature sets (where SLR has the best performance).

Table 23 - Prediction AUCs for combination methods on GWAS data.

Features	***Algorithm***	***Dataset***								
		TGen	**ADRC**	**BD**	**CAD**	**CD**	**HT**	**RA**	**T1D**	**T2D**
5	**LP1+ChiSq (score)**	0.7230	0.7433	0.6150	0.7668	0.6259	0.5546	0.7013	0.7448	0.7783

	LP1+ChiSq (rank)	0.7230	0.7233	0.6053	0.7501	0.6284	0.5501	0.7013	0.7448	0.7546
	LP1+SLR (rank)	0.7214	0.7433	0.6053	0.7504	0.6118	0.5661	0.7007	0.7418	0.7603
10	**LP1+ChiSq (score)**	0.7234	0.7111	0.6011	0.8118	0.6349	0.5670	0.7280	0.7301	0.7514
	LP1+ChiSq (rank)	0.6873	0.7027	0.6022	0.8007	0.6224	0.5529	0.7225	0.7316	0.7465
	LP1+SLR (rank)	0.7336	0.7244	0.6022	0.8016	0.6371	0.5802	0.7318	0.7311	0.7575
50	**LP1+ChiSq (score)**	0.6905	0.6361	0.5392	0.5718	0.5618	0.5407	0.6218	0.6700	0.5794
	LP1+ChiSq (rank)	0.6635	0.6254	0.5217	0.5525	0.5652	0.5548	0.6010	0.6702	0.5771
	LP1+SLR (rank)	0.7251	0.6518	0.5656	0.6548	0.6012	0.5715	0.6623	0.6819	0.6853
100	**LP1+ChiSq (score)**	0.6458	0.5915	0.5324	0.5540	0.5873	0.5391	0.5881	0.6045	0.5610
	LP1+ChiSq (rank)	0.6217	0.5619	0.5246	0.5314	0.5676	0.5307	0.5891	0.5984	0.5596
	LP1+SLR (rank)	0.6378	0.6152	0.5612	0.5228	0.6033	0.5642	0.6583	0.6547	0.6608

To statistically compare the performance of the prior method on GWAS data, I computed the paired Wilcoxon signed-rank test over the GWAS AUCs. Using each of the nine GWAS datasets, I used two rows of AUCs from Table 23 (two different feature selection thresholds), yielding 18 paired AUC samples over which to compare algorithms. Below are the relevant AUC values across all datasets, followed by the pairwise Wilcoxon comparisons with the better-performing algorithm listed in each case along with the *p*-value of the comparison result.

Table 24 - Wilcoxon signed-ranked test p-value comparing AUCs of algorithms over 9 GWAS datasets for feature selection thresholds of 5 and 10.

	LP1	**LP1+ChiSq (score)**	**LP1+ChiSq (rank)**
LP1+ChiSq (score)	LP1 (*p* = 0.0466)	-	-
LP1+ChiSq (rank)	LP1 (*p* = 0.0006)	LP1+ChiSq (score) (*p* = 0.0021)	--
LP1+SLR (rank)	LP1 (*p* = 0.3030)	LP1+SLR (rank) (*p* = 0.9282)	LP1+SLR (rank) (*p* = 0.0384)

Table 25 - Wilcoxon signed-ranked test p-value comparing AUCs of algorithms over 9 GWAS datasets for feature selection thresholds of 50 and 100.

	LP1	**LP1+ChiSq (score)**	**LP1+ChiSq (rank)**
LP1+ChiSq (score)	LP1 (*p* = 0.0687)	-	-
LP1+ChiSq (rank)	LP1 (*p* = 0.0002)	LP1+ChiSq (score) (*p* = 0.0032)	-
LP1+SLR (rank)	LP1+SLR (rank) (*p* = 0.0037)	LP1+SLR (rank) (*p* = 0.0002)	LP1+SLR (rank) (*p* = 0.0012)

The Wilcoxon tests show that the LP1+SLR rank-based method performed better for larger feature set sizes of 50 and 100, but does not outperform the LP1 method alone for smaller feature set sizes. For the LP1+ChiSq methods, the score-based method performed better than the rank-based method.

6.0 CONCLUSIONS AND FUTURE WORK

In this dissertation, I developed and applied several extensions to label propagation (LP), a multivariate feature selection algorithm, with the goal of performing feature selection for biomarker discovery in GWAS SNP data. I implemented the LP algorithm as well as several extensions to tailor the method to genomic data. I applied these methods to synthetic, semi-synthetic, and GWAS datasets and evaluated their performance in terms of precision-recall, predictive power, reproducibility, and biological validity. The LP1 extension was found to improve upon the original LP3 method under several conditions, namely, the ability to identify variants with population-wide predictive power. The prior knowledge incorporation methods did not significantly improve performance over not using prior knowledge, and the ranking combination method had limited success. A summary of the findings is presented in the next section followed by some directions for future work in the last section.

6.1 CONTRIBUTIONS AND FINDINGS

This section summarizes the main contributions and the findings of the research presented in this dissertation.

6.1.1 LP1 Extension

On the synthetic data, the LP1 extension performed worse than the original LP3 method. The LP1 method was unable to identify rare variants, because the signal from just a few samples is being lost in the course of the population-wide contingency table analysis. In contrast, the LP3 score for a SNP state is more or less independent of sample size, so even a variant with just a few occurrences in a population can get a highly ranked score. The synthetic experiments also provided support for using a value between 0 and 0.5 for the parameter α, limiting the amount of diffusion in the propagation graph.

All algorithms performed poorly on the GAW17 data, stymied by the rarity of the causal variants and the complex phenotypic model. These experiments indicate that LP as formulated in this dissertation may not be well-suited to the discovery of rare SNVs. However, with the increasing sample sizes of genomic datasets, it is possible that rare variants will occur in large enough numbers for LP to be useful.

Results of the LP1 algorithm on GWAS data showed that LP1 is an improvement over the original LP3 algorithm. By using the soft sample labels as pseudocounts, LP1 computes a population-wide score that is not specific to a particular SNP allele. This score ranks a SNP highly only if there is a significant amount of data to support the association, and is not susceptible to highly-ranked rare variants that are not predictive on a population scale. While the LP1 method is effective at finding common predictive variants, it showed reduced power in finding the rare causal variants in the synthetic dataset.

The LP1 algorithm had good performance in selecting SNPs that could be validated in the literature, finding evidence for nearly half of the top-ranked 25 SNPs across all GWAS datasets.

The other top-ranked SNPs that were not validated are good candidates for further study to uncover possible mechanisms of association with disease.

6.1.2 Prior Knowledge

The edge weighting method of incorporating prior knowledge was found to be ineffective. This could be a result of the LP1 score utilizing the sample node scores directly in the contingency table analysis. Because the sample nodes begin with labels, they are relatively inflexible when the diffusion parameter α is low. Referring back to the iterative update equations, it can be seen that the current label on a node is a linear combination of the initial label and the currently labels on from the other side of the network. With α less than 0.5, the initial labeling dominates, and in fact, cannot be overcome by contributions from the other side of the network. That is, a node with an initial label of +1 will ultimately have a positive score no matter what the network geometry and labeling, and similarly a node with an initial label of -1 will always finish with a negative score.

In the course of experimentation, I found a low α of 0.25 to be very effective. Under this parameterization, a labeled node will not change class from +1 to -1, despite being given an updated soft labeling. Unlabeled nodes still have flexibility to drift positive or negative, within a limited range around 0. It seems that the edge weight prior knowledge method simply does not have enough effect on the relabeling of the initially labeled sample nodes to impact the final LP1 score.

In contrast, the prior pseudocount method had a marked effect on the feature ranking. The utility of the method was shown on the synthetic and semi-synthetic datasets. The gold standard prior increased LP1's precision and recall, especially when using a large PESS. The prior method results on the GWAS datasets were less promising, having no significant impact on the downstream classification accuracy for small values of PESS, and sometimes decreasing the accuracy for large values of PESS.

The poor performance of the prior knowledge method could stem from a few sources. It is possible that the MAF and GERP scores being used are simply not well-correlated with the predictiveness of common variants. The lack of correlation between the MAF and GERP scores suggests that either one or both of these scores is uncorrelated with the predictive variants. Either one is well-correlated with predictiveness and the other is not, or both are only somewhat correlated with predictiveness. If both the MAF and GERP scores were indeed correlated with predictiveness, there would be some level of correlation between them.

It is also possible that an incorrect genetic model is being used. The prior pseudocount method assumes that the minor allele is the disease-causing allele, skewing the prior count table such that the disease prevalence is higher for the *aa* genotype, lower for the *AA* genotype, and unchanged for the *Aa* genotype. In effect, this assumes an additive model of disease association. If this model is not true for a SNP, the counts could skew in the wrong direction, potentially diluting genomic signal in the data rather than enhancing it.

Finally, it is possible that the mapping from MAF or GERP score to the prior count table is being performed incorrectly. My experiments utilized an exponential mapping from MAF or GERP to the actual prior knowledge factor. While this does put more emphasis on the extreme MAF and GERP scores, it is possible that this weighting is not enough. An analysis by Gorlov et al. [4] suggests that it is indeed the rarest mutations which have the strongest associations with disease, but this may not hold true for SNPs with moderate to large MAF. It is possible that GERP and MAF are useful as priors only for rare to very rare SNVs.

6.1.3 Combination Method

The combination score-based combination method was found to have relatively little impact when combining LP1 and the chi squared test. It is possible that the chi squared test is dominating this combination, because the chi squared test tends to result in smaller *p*-values than the LP1 metric. As the LP1's α parameter increases, the resulting contingency table morphs from the original chi squared table with hard counts to a contingency table that represent diffusion using soft labels. These counts are by their nature less extreme, leading to less extreme distribution differences, and ultimately larger (less significant) *p*-values. The rank-based chi-squared combination method had somewhat better results than the score-based method.

The LP1+SLR rank-based method was generally worse than LP1 for small feature set sizes (where SLR performs worse), but better for larger feature set sizes (where SLR performs better). Applying ranks to the SLR-scored variables is somewhat different than the chi squared test, because most of the features under the SLR model get an identical score of 0. This means that all features not used in the SLR model get the maximum possible rank, while usually only about 500 features have true ranks. When averaging these ranks with LP1, this is almost performing an implicit feature selection step. Most variables' combination ranks will be the LP1 rank averaged with the maximum possible rank, putting them far down the ranking no matter what the LP1 rank is. Only features selected by SLR have a chance of being ranked highly. If SLR does select meaningful features, however, the ranking can be fine-tuned by the addition of the LP1 ranking.

6.2 FUTURE WORK

The experimental work presented in this research explored the application of LP for biomarker discovery in genomic data. Several extensions and directions for future work are possible.

While the prior knowledge method did not significantly improve the performance of LP1, there are a number of ways to modify it that might lead to improved performance. Fitting the proper genetic model to each SNP could improve the method, eliminating the case where the prior counts dilute the signal in the data rather than enhance it. It would be possible to utilize the LP3 scores directly to glean information about the genetic model of each SNP. In the LP3 model, each SNP allele is scored according to its association with the case or control group, so it is very simple to identify the risk and protective alleles simply by finding the allele with the largest positive and negative scores, respectively. This prior method is not totally independent of the data, but only utilizes the data to determine the direction of association and not the strength of the prior.

Another way to determine the proper genetic model would be to examine each model exhaustively, and select the best-fitting model. A fully genotypic model could be analyzed by computing a statistic such as the chi squared criterion or the BIC for the 2x3 contingency table. Other models would be represented by a collapsed, 2x2 version of the contingency table. The recessive model, for instance, would be a result of combining the *AA* and *Aa* columns, while the dominant model would combine the *Aa* and *aa* columns.

Yet another method of including prior knowledge could be to include pseudocount bias nodes in the network propagation. In this method the graph would be initialized with two extra sample nodes having labels +1 and -1 which are connected to every feature node in the graph. These nodes’ labels are fixed, and do not update at any step of the iterative propagation process.

The prior knowledge is encapsulated as edge weights which connects the bias nodes to the feature nodes. For a SNP with no prior belief of association, the weights connecting to the bias nodes are 0, meaning they have no effect on the propagation equations for that SNP. On the other hand, SNPs with prior evidence of association are have increased weights connecting to the bias nodes, indicating a larger pseudocount of cases or controls for a particular SNP allele. This allows the prior knowledge to actually have an effect on the propagation itself, rather than just being added as pseudocounts at the end (as is described in 3.4.2.2).

The ranking combination method proved to be effective for the LP1+SLR rank-based method using a moderate number of features. This could possibly be a result of the implicit feature selection step that SLR undertakes. Instead of directly averaging ranks, it might also be effective to perform a two-step ranking by using an implicitly selective algorithm like SLR followed by ranking only those selected variants. This could potentially overcome SLR's problem of low predictive performance for small feature set sizes.

Finally, the LP algorithm is a semi-supervised algorithm, but I did not fully explore how the algorithm might utilize unlabeled data. Unlabeled sample nodes would represent individuals for whom we have the genotype, but no phenotype. Additional unlabeled samples could add extra diffusion paths to the graph, reinforcing functional modules and affecting the network propagation. This might ultimately allow the LP algorithm to leverage disparate data sources where the phenotype is unlabeled or uncertain, potentially improving the algorithm's ability to select meaningful variants.

APPENDIX A

AUC RESULTS FOR GWAS DATASTES

This Appendix contains the AUC results obtained from all classification experiments. Each table (Table 26 though Table 34) gives results for one GWAS dataset and gives the dataset name, the feature ranking algorithms that were used, and the number of features used in the *k*NN classifier. ChiSq refers to the chi squared test, SLR is the sparse logistic regression method, LP3 is the original LP algorithm with α = 0.25, and LP1 is the single-score extension with α = 0.25. LP1+GERP50 refers to the LP1 score using the GERP prior with PESS set to 50, while LP1+GERP500 refers to the GERP prior with PESS set to 500. The experiments using the MAF prior are named in an analogous fashion. The last set of experiments is the combination methods, which combine LP1 with chi squared or SLR, using both rank and score-based methods.

Table 26 - Classification AUC results for TGen data.

TGen	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.6992	0.6945	0.7220	0.7394	0.7060	0.6574	0.6013	0.5953
95% CI	±0.0139	0.0141	±0.0136	±0.0132	±0.0140	±0.0148	±0.0155	±0.0154
SLR	0.6783	0.6876	0.7291	0.7424	0.7264	-	-	-
95% CI	±0.0144	±0.0137	±0.0134	±0.0131	±0.0131	-	-	-
LP3	0.6733	0.6904	0.7088	0.7369	0.7519	0.7286	0.6138	0.5735
95% CI	±0.0144	±0.0137	±0.0139	±0.0131	±0.0135	±0.0137	±0.0131	±0.0154
LP1	0.6992	0.6945	0.7230	0.7118	0.6694	0.6473	0.5958	0.5820
95% CI	±0.0139	±0.0141	±0.0136	±0.0138	±0.0145	±0.0149	±0.0154	±0.0156
LP1+GERP50	0.6992	0.6945	0.7228	0.7145	0.6587	0.6426	0.5725	0.5620
95% CI	±0.0139	±0.0141	±0.0136	±0.0138	±0.0145	±0.0149	±0.0154	±0.0156
LP1+GERP500	0.5165	0.5076	0.5112	0.5108	0.5145	0.5088	0.4923	0.4986
95% CI	±0.0164	±0.0164	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166
LP1+MAF50	0.6992	0.6945	0.7014	0.6983	0.6455	0.6326	0.5488	0.5313

95% CI	±0.0139	±0.0141	±0.0136	±0.0138	±0.0145	±0.0149	±0.0154	±0.0156
LP1+MAF500	0.5025	0.5110	0.5102	0.5129	0.5123	0.5101	0.4988	0.4976
95% CI	±0.0164	±0.0164	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166
LP1+ChiSq_Score	0.6992	0.6945	0.7230	0.7234	0.6905	0.6458	0.5852	0.5871
95% CI	±0.0139	±0.0141	±0.0136	±0.0138	±0.0139	±0.0149	±0.0154	±0.0156
LP1+ChiSq_Rank	0.6992	0.6945	0.7230	0.6873	0.6635	0.6217	0.5746	0.5321
95% CI	±0.0139	±0.0141	±0.0136	±0.0140	±0.0145	±0.0149	±0.0154	±0.0156
LP1+SLR_Rank	0.6992	0.6945	0.7214	0.7336	0.7251	0.6378	0.5755	0.5713
95% CI	±0.0139	±0.0141	±0.0134	±0.0131	±0.0132	±0.0149	±0.0154	±0.0156

Table 27 - Classification AUC results for ADRC data.

ADRC	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.6834	0.7369	0.7433	0.7184	0.6438	0.6034	0.5445	0.5349
95% CI	±0.0112	±0.0105	±0.0104	±0.0108	±0.0116	±0.0120	±0.0122	±0.0122
SLR	0.6834	0.6911	0.7100	0.7354	0.6970	0.6874	-	-
95% CI	±0.0112	±0.0111	±0.0109	±0.0105	±0.0111	±0.0112	-	-
LP3	0.6325	0.6756	0.7342	0.7315	0.7151	0.7154	0.6096	0.5435
95% CI	±0.0117	±0.0105	±0.0110	±0.0110	±0.0107	±0.0109	±0.0104	±0.0122
LP1	0.6834	0.7369	0.7433	0.7058	0.6264	0.5862	0.5383	0.5228
95% CI	±0.0112	±0.0105	±0.0104	±0.0110	±0.0118	±0.0121	±0.0123	±0.0123
LP1+GERP50	0.6834	0.7369	0.7433	0.7102	0.6133	0.5789	0.5464	0.5173
95% CI	±0.0112	±0.0105	±0.0104	±0.0110	±0.0118	±0.0121	±0.0123	±0.0123
LP1+GERP500	0.4975	0.5033	0.5043	0.5138	0.5057	0.5102	0.5013	0.5082
95% CI	±0.0164	±0.0164	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166
LP1+MAF50	0.6834	0.7369	0.7345	0.7001	0.6023	0.5643	0.5523	0.5069
95% CI	±0.0112	±0.0105	±0.0104	±0.0110	±0.0118	±0.0121	±0.0123	±0.0123
LP1+MAF500	0.5025	0.5054	0.5068	0.5264	0.5148	0.5089	0.5166	0.5009
95% CI	±0.0164	±0.0164	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166	±0.0166
LP1+ChiSq_Score	0.6834	0.7369	0.7433	0.7111	0.6361	0.5915	0.5441	0.5101
95% CI	±0.0112	±0.0105	±0.0104	±0.0109	±0.0117	±0.0121	±0.0123	±0.0129
LP1+ChiSq_Rank	0.6834	0.7369	0.7233	0.7027	0.6254	0.5619	0.5303	0.5216
95% CI	±0.0112	±0.0105	±0.0104	±0.0111	±0.0119	±0.0122	±0.0125	±0.0126
LP1+SLR_Rank	0.6834	0.7369	0.7433	0.7244	0.6518	0.6152	0.5401	0.5233
95% CI	±0.0112	±0.0105	±0.0104	±0.0110	±0.0116	±0.0119	±0.0122	±0.0123

Table 28 - Classification AUC results for BD WTCCC data.

BD	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.5358	0.5724	0.6056	0.5846	0.5372	0.5300	0.5301	0.5220
95% CI	±0.0164	±0.0166	±0.0164	±0.0166	±0.0166	±0.0168	±0.0166	±0.0166
SLR	0.5353	0.5461	0.5975	0.6166	0.5874	0.5742	-	-
95% CI	±0.0166	±0.0164	±0.0164	±0.0164	±0.0165	±0.0165	-	-
LP3	0.5211	0.5478	0.5478	0.5532	0.5219	0.5190	0.5089	0.4911
95% CI	±0.0167	±0.0169	±0.0169	±0.0169	±0.0169	±0.0169	±0.0167	±0.0167
LP1	0.5366	0.5653	0.6270	0.6137	0.5348	0.5328	0.5318	0.5365
95% CI	±0.0165	±0.0163	±0.0161	±0.0163	±0.0167	±0.0167	±0.0167	±0.0167
LP1+GERP50	0.5349	0.5653	0.6270	0.6131	0.5376	0.5278	0.5331	0.5379
95% CI	±0.0165	±0.0163	±0.0161	±0.0163	±0.0167	±0.0167	±0.0167	±0.0167
LP1+GERP500	0.5075	0.5067	0.5108	0.5118	0.5037	0.5095	0.4933	0.5042
95% CI	±0.0165	±0.0165	±0.0167	±0.0167	±0.0167	±0.0167	±0.0167	±0.0167
LP1+MAF50	0.5349	0.5653	0.6160	0.6034	0.5485	0.5153	0.5355	0.5324
95% CI	±0.0165	±0.0163	±0.0161	±0.0163	±0.0167	±0.0167	±0.0167	±0.0167
LP1+MAF500	0.5066	0.5061	0.5111	0.5118	0.5024	0.5084	0.4926	0.4958
95% CI	±0.0165	±0.0165	±0.0167	±0.0167	±0.0167	±0.0167	±0.0167	±0.0167
LP1+ChiSq_Score	0.5366	0.5701	0.6150	0.6011	0.5392	0.5324	0.5366	0.5214

95% CI	±0.0165	±0.0163	±0.0162	±0.0164	±0.0167	±0.0167	±0.0167	±0.0168
LP1+ChiSq_Rank	0.5262	0.5514	0.6053	0.6022	0.5217	0.5246	0.5291	0.5107
95% CI	±0.0166	±0.0164	±0.0164	±0.0164	±0.0168	±0.0168	±0.0168	±0.0169
LP1+SLR_Rank	0.5302	0.5514	0.6053	0.6022	0.5656	0.5612	0.5291	0.5107
95% CI	±0.0166	±0.0164	±0.0164	±0.0164	±0.0168	±0.0168	±0.0168	±0.0169

Table 29 - Classification AUC results for CAD WTCCC data.

CAD	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.5991	0.6804	0.7702	0.8201	0.5643	0.5368	0.5295	0.5224
95% CI	±0.0159	±0.0149	±0.0129	±0.0116	±0.0165	±0.0167	±0.0165	±0.0165
SLR	0.4949	0.5408	0.6291	0.7840	0.6568	0.5993	-	-
95% CI	±0.0163	±0.0163	±0.0157	±0.0127	±0.0149	±0.0157	-	-
LP3	0.5503	0.5999	0.6612	0.7109	0.5860	0.5420	0.5218	0.5120
95% CI	±0.0171	±0.0167	±0.0163	±0.0159	±0.0165	±0.0167	±0.0167	±0.0167
LP1	±0.6804	0.6804	0.7668	0.8372	0.6254	0.5452	0.5224	0.5239
95% CI	±0.0149	±0.0149	±0.0129	±0.0108	±0.0161	±0.0165	±0.0167	±0.0165
LP1+GERP50	0.6804	0.6804	0.7668	0.8372	0.5950	0.5514	0.5102	0.5114
95% CI	±0.0149	±0.0149	±0.0129	±0.0108	±0.0163	±0.0165	±0.0165	±0.0165
LP1+GERP500	0.6021	0.6804	0.7489	0.8063	0.5903	0.5307	0.5152	0.5036
95% CI	±0.0157	±0.0149	±0.0135	±0.0120	±0.0163	±0.0165	±0.0165	±0.0165
LP1+MAF50	0.6804	0.6804	0.7502	0.8361	0.6055	0.5489	0.5202	0.5164
95% CI	±0.0149	±0.0149	±0.0129	±0.0108	±0.0163	±0.0165	±0.0165	±0.0165
LP1+MAF500	0.5568	0.6523	0.7023	0.7356	0.5842	0.5267	0.5173	0.4943
95% CI	±0.0157	±0.0149	±0.0135	±0.0120	±0.0163	±0.0165	±0.0165	±0.0165
LP1+ChiSq_Score	±0.6719	0.6804	0.7668	0.8118	0.5718	0.5515	0.5316	0.5249
95% CI	±0.0149	±0.0149	±0.0129	±0.0120	±0.0165	±0.0165	±0.0167	±0.0165
LP1+ChiSq_Rank	±0.6619	0.6744	0.7501	0.8007	0.5525	0.5489	0.5304	0.5275
95% CI	±0.0149	±0.0149	±0.0129	±0.0120	±0.0165	±0.0165	±0.0167	±0.0165
LP1+SLR_Rank	±0.6628	0.6689	0.7504	0.8016	0.6548	0.5540	0.5314	0.5228
95% CI	±0.0149	±0.0149	±0.0129	±0.0120	±0.0149	±0.0165	±0.0167	±0.0165

Table 30 - Classification AUC results for CD WTCCC data.

CD	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.5509	0.5766	0.6225	0.6346	0.5796	0.5950	0.5656	0.5605
95% CI	±0.0171	±0.0169	±0.0165	±0.0163	±0.0171	±0.0167	±0.0171	±0.0169
SLR	0.5509	0.5486	0.5859	0.6285	0.6027	0.6188	-	-
95% CI	±0.0171	±0.0172	±0.0167	±0.0163	±0.0165	±0.0163	-	-
LP3	0.5506	0.5540	0.5414	0.5548	0.5314	0.5146	0.5073	0.5039
95% CI	±0.0171	±0.0171	±0.0172	±0.0172	±0.0171	±0.0172	±0.0171	±0.0171
LP1	0.5509	0.5766	0.6288	0.6371	0.5791	0.5872	0.5664	0.5365
95% CI	±0.0171	±0.0169	±0.0165	±0.0163	±0.0169	±0.0167	±0.0169	±0.0171
LP1+GERP50	0.5509	0.5822	0.6236	0.6358	0.5870	0.5805	0.5750	0.5488
95% CI	±0.0171	±0.0169	±0.0165	±0.0163	±0.0169	±0.0169	±0.0169	±0.0171
LP1+GERP500	0.4986	0.4977	0.5085	0.5465	0.5046	0.5073	0.5530	0.5550
95% CI	±0.0167	±0.0167	±0.0171	±0.0172	±0.0171	±0.0171	±0.0171	±0.0169
LP1+MAF50	0.5509	0.5822	0.6212	0.6316	0.5901	0.5640	0.5547	0.5364
95% CI	±0.0171	±0.0169	±0.0165	±0.0163	±0.0169	±0.0169	±0.0169	±0.0171
LP1+MAF500	0.4986	0.4977	0.5065	0.5115	0.5048	0.5013	0.5530	0.5317
95% CI	±0.0167	±0.0167	±0.0171	±0.0172	±0.0171	±0.0171	±0.0171	±0.0169
LP1+ChiSq_Score	0.5509	0.5766	0.6259	0.6349	0.5618	0.5873	0.5648	0.5422
95% CI	±0.0171	±0.0169	±0.0165	±0.0163	±0.0169	±0.0167	±0.0169	±0.0171
LP1+ChiSq_Rank	0.5509	0.5766	0.6284	0.6224	0.5652	0.5676	0.5548	0.5434
95% CI	±0.0171	±0.0169	±0.0165	±0.0163	±0.0169	±0.0169	±0.0169	±0.0171

LP1+SLR_Rank	0.5509	0.5512	0.6118	0.6371	0.6012	0.6033	0.5562	0.5431
95% CI	±0.0171	±0.0171	±0.0165	±0.0163	±0.0165	±0.0165	±0.0169	±0.0171

Table 31 - Classification AUC results for HT WTCCC data.

HT	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.5344	0.5359	0.5519	0.5530	0.5263	0.5362	0.5355	0.5400
95% CI	±0.0165	±0.0165	±0.0167	±0.0165	±0.0167	±0.0165	±0.0165	±0.0163
SLR	0.5354	0.5463	0.5756	0.5885	0.5774	0.5671	-	-
95% CI	±0.0165	±0.0167	±0.0165	±0.0163	±0.0163	±0.0165	-	-
LP3	0.5003	0.5066	0.4959	0.4904	0.5060	0.4921	0.4934	0.5021
95% CI	±0.0163	±0.0163	±0.0165	±0.0165	±0.0165	±0.0165	±0.0165	±0.0167
LP1	0.5322	0.5481	0.5601	0.5711	0.5684	0.5391	0.5280	0.5378
95% CI	±0.0165	±0.0165	±0.0167	±0.0165	±0.0165	±0.0165	±0.0165	±0.0165
LP1+GERP50	0.5322	0.5518	0.5625	0.5715	0.5430	0.5364	0.5368	0.5354
95% CI	±0.0165	±0.0165	±0.0167	±0.0165	±0.0167	±0.0167	±0.0167	±0.0167
LP1+GERP500	0.5036	0.5092	0.4946	0.4742	0.4987	0.4905	0.5019	0.5118
95% CI	±0.0163	±0.0163	±0.0165	±0.0165	±0.0165	±0.0165	±0.0165	±0.0167
LP1+MAF50	0.5322	0.5518	0.5522	0.5604	0.5497	0.5456	0.5325	0.5276
95% CI	±0.0165	±0.0165	±0.0167	±0.0165	±0.0165	±0.0165	±0.0165	±0.067
LP1+MAF500	0.5036	0.5092	0.5002	0.4841	0.4964	0.4921	0.5009	0.5017
95% CI	±0.0163	±0.0163	±0.0165	±0.0165	±0.0165	±0.0165	±0.0165	±0.0167
LP1+ChiSq_Score	0.5301	0.5368	0.5546	0.5670	0.5407	0.5391	0.5280	0.5378
95% CI	±0.0165	±0.0165	±0.0167	±0.0165	±0.0167	±0.0167	±0.0167	±0.0167
LP1+ChiSq_Rank	0.5322	0.5481	0.5501	0.5529	0.5548	0.5307	0.5344	0.5299
95% CI	±0.0165	±0.0165	±0.0167	±0.0165	±0.0165	±0.0165	±0.0165	±0.0165
LP1+SLR_Rank	0.5352	0.5462	0.5661	0.5802	0.5715	0.5642	0.5318	0.5312
95% CI	±0.0165	±0.0165	±0.0164	±0.0163	±0.0163	±0.0165	±0.0165	±0.0165

Table 32 - Classification AUC results for RA WTCCC data.

RA	**1**	**2**	**5**	**10**	**50**	**100**	**500**	**1000**
ChiSq	0.5891	0.6313	0.7013	0.7273	0.6107	0.5884	0.5467	0.5434
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0163	±0.0165	±0.0167	±0.0167
SLR	0.5891	0.6159	0.6902	0.7382	0.7055	0.6711	-	-
95% CI	±0.0169	±0.0169	±0.0155	±0.0145	±0.0151	±0.0157	-	-
LP3	0.5849	0.6109	0.6260	0.6352	0.5852	0.5719	0.5247	0.5133
95% CI	±0.0169	±0.0169	±0.0169	±0.0167	±0.0169	±0.0169	±0.0169	±0.0167
LP1	0.5891	0.6313	0.7013	0.7284	0.6398	0.5946	0.5565	0.5531
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0163	±0.0165	±0.0167	±0.0167
LP1+GERP50	0.5891	0.6313	0.7028	0.7282	0.6377	0.6009	0.5547	0.5410
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0163	±0.0165	±0.0165	±0.0165
LP1+GERP500	0.5891	0.5950	0.6005	0.6654	0.5914	0.5698	0.5496	0.5588
95% CI	±0.0169	±0.0167	±0.0169	±0.0159	±0.0167	±0.0167	±0.0165	±0.0167
LP1+MAF50	0.5891	0.6313	0.6969	0.7118	0.6291	0.6015	0.5436	0.5398
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0163	±0.0165	±0.0165	±0.0165
LP1+MAF500	0.5502	0.5680	0.5985	0.6254	0.5724	0.5448	0.5305	0.5418
95% CI	±0.0169	±0.0167	±0.0169	±0.0159	±0.0167	±0.0167	±0.0165	±0.0167
LP1+ChiSq_Score	0.5891	0.6313	0.7013	0.7280	0.6218	0.5881	0.5538	0.5502
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0163	±0.0165	±0.0167	±0.0167
LP1+ChiSq_Rank	0.5891	0.6313	0.7013	0.7225	0.6010	0.5891	0.5322	0.5294
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0163	±0.0165	±0.0167	±0.0167
LP1+SLR_Rank	0.5891	0.6313	0.7007	0.7318	0.6623	0.6583	0.5646	0.5514
95% CI	±0.0169	±0.0165	±0.0153	±0.0149	±0.0153	±0.0152	±0.0167	±0.0167

Table 33 - Classification AUC results for T1D WTCCC data.

T1D	1	2	5	10	50	100	500	1000
ChiSq	0.6777	0.7047	0.7448	0.7293	0.6562	0.5905	0.5733	0.5480
95% CI	±0.0149	±0.0143	±0.0135	±0.0139	±0.0153	±0.0161	±0.0163	±0.0163
SLR	0.5007	0.5490	0.7346	0.7202	0.6841	0.6639	-	-
95% CI	±0.0163	±0.0165	±0.0137	±0.0131	±0.0139	±0.0147	-	-
LP3	0.6093	0.6617	0.6891	0.6900	0.6678	0.6301	0.5461	0.5384
95% CI	±0.0157	±0.0155	±0.0151	±0.0151	±0.0153	±0.0159	±0.0165	±0.0165
LP1	0.6863	0.7047	0.7448	0.7329	0.6704	0.6054	0.5614	0.5602
95% CI	±0.0147	±0.0143	±0.0135	±0.0137	±0.0151	±0.0159	±0.0163	±0.0163
LP1+GERP50	0.6863	0.7047	0.7445	0.7306	0.6624	0.5983	0.5609	0.5585
95% CI	±0.0147	±0.0143	±0.0135	±0.0139	±0.0153	±0.0161	±0.0163	±0.0163
LP1+GERP500	0.6777	0.7047	0.7093	0.7040	0.6383	0.6077	0.5553	0.5499
95% CI	±0.0149	±0.0143	±0.0141	±0.0143	±0.0155	±0.0161	±0.0165	±0.0165
LP1+MAF50	0.6863	0.7047	0.7445	0.7296	0.6544	0.5838	0.5567	0.5445
95% CI	±0.0147	±0.0143	±0.0135	±0.0139	±0.0153	±0.0161	±0.0163	±0.0163
LP1+MAF500	0.6213	0.6534	0.6875	0.6755	0.6212	0.6014	0.5448	0.5237
95% CI	±0.0147	±0.0139	±0.0139	±0.0141	±0.0153	±0.0161	±0.0165	±0.0165
LP1+ChiSq_Score	0.6779	0.7047	0.7448	0.7301	0.6700	0.6045	0.5718	0.5542
95% CI	±0.0149	±0.0143	±0.0135	±0.0137	±0.0151	±0.0159	±0.0163	±0.0163
LP1+ChiSq_Rank	0.6863	0.7047	0.7448	0.7316	0.6702	0.5984	0.5673	0.5519
95% CI	±0.0147	±0.0143	±0.0135	±0.0137	±0.0151	±0.0159	±0.0163	±0.0163
LP1+SLR_Rank	0.6434	0.6627	0.7418	0.7311	0.6819	0.6547	0.6012	0.5631
95% CI	±0.0147	±0.0143	±0.0135	±0.0137	±0.0151	±0.0147	±0.0161	±0.0163

Table 34 - Classification AUC results for T2D WTCCC data.

T2D	1	2	5	10	50	100	500	1000
ChiSq	0.5599	0.6498	0.7780	0.7329	0.5707	0.5611	0.5358	0.5201
95% CI	±0.0167	±0.0157	±0.0129	±0.0141	±0.0165	±0.0163	±0.0167	0.0167
SLR	0.4916	0.6015	0.6751	0.7424	0.7258	0.6671	-	-
95% CI	±0.0165	±0.0159	±0.0151	±0.0139	±0.0143	±0.0155	-	-
LP3	0.5521	0.6091	0.6903	0.6907	0.6071	0.5773	0.5293	0.5232
95% CI	±0.0167	±0.0167	±0.0161	±0.0161	±0.0165	±0.0165	±0.0165	0.0165
LP1	0.6862	0.6862	0.7787	0.7622	0.5804	0.5614	0.5432	0.5506
95% CI	±0.0149	±0.0149	±0.0129	±0.0135	±0.0165	±0.0165	±0.0167	±0.0165
LP1+GERP50	0.6862	0.6862	0.7739	0.7551	0.5703	0.5640	0.5492	0.5544
95% CI	±0.0149	±0.0149	±0.0131	±0.0135	±0.0165	±0.0165	±0.0165	±0.0165
LP1+GERP500	0.6009	0.6862	0.7320	0.7573	0.5715	0.5455	0.5160	0.5162
95% CI	±0.0159	±0.0149	±0.0139	±0.0135	±0.0165	±0.0165	±0.0167	±0.0165
LP1+MAF50	0.6862	0.6862	0.7669	0.7485	0.5693	0.5542	0.5313	0.5315
95% CI	±0.0149	±0.0149	±0.0131	±0.0135	±0.0165	±0.0167	±0.0167	±0.0167
LP1+MAF500	0.5595	0.5786	0.6255	0.6548	0.5685	0.5326	0.5061	0.5081
95% CI	±0.0163	±0.0157	±0.0153	±0.0141	±0.0165	0±.0167	±0.0167	±0.0165
LP1+ChiSq_Score	0.6862	0.6582	0.7783	0.7514	0.5794	0.5610	0.5412	0.5386
95% CI	±0.0149	±0.0153	±0.0129	±0.0135	±0.0165	±0.0165	±0.0167	±0.0165
LP1+ChiSq_Rank	0.5832	0.6245	0.7546	0.7465	0.5771	0.5596	0.5483	0.5372
95% CI	±0.0164	±0.0159	±0.0129	±0.0139	±0.0165	±0.0165	±0.0167	±0.0165
LP1+SLR_Rank	0.6641	0.6714	0.7603	0.7575	0.6853	0.6608	0.6035	0.5616
95% CI	±0.0152	±0.0151	±0.0129	±0.0136	±0.0150	±0.0165	±0.0163	±0.0165

APPENDIX B

BIOLOGICAL VALIDITY RESULTS FOR GWAS DATASETS

This Appendix contains the results of the biological validation experiment for each of the control algorithms (ChiSq, SWRF, SLR, and LP3) on the two LOAD datasets (TGen and ADRC). Each table (see Tables 35 to 42) gives the top 25 SNPs as ranked by each control algorithm, with the associated chromosome, gene, and literature reference. The LP1 biological validation results may be found in Section 5.1.3.2. SNPs with literature validation results are highlighted in grey.

Table 35 - Top 25 SNPs as ranked by LP3 ($\alpha = 0.25$) for TGen data.

Rank	rsID	Gene	Chr	Comment
1	rs7412	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
2	rs4420638	APOC	19	In strong linkage disequilibrium with APOE SNPs [156]
3	rs429358	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
4	rs10824310	PRKG1	10	Significant association with LOAD [55]
5	rs12162084	-	16	Significant association with LOAD [157]
6	rs17330779	NRCAM	7	Associated with axonal degeneration in LOAD [158]
7	rs7077757	RBM20	10	Meta-analysis of multiple studies showed association [159]
8	rs10115381	-	9	-
9	rs4356530	-	17	Association found in another analysis of TGen data [59]
10	rs6717497	-	2	-
11	rs2913719	-	5	Association in systematic meta-analysis of AD [69]
12	rs12476792	-	2	-
13	rs17169622	BMPER	7	-
14	rs1038891	LRRC4C	11	SNP associated with LOAD in genome-wide analysis [61]
15	rs10499687	VWC2	7	-
16	rs7335085	-	13	-
17	rs16974268	SLCO3A1	15	-
18	rs10996618	-	10	SNP selected in logistic regression analysis [62]
19	rs950922	ALPL	1	-
20	rs17151710	-	5	Found in meta-analysis of 3 studies [160]
21	rs9934599	IL34	16	-
22	rs473367	-	9	SNP may interact with APOE to affect LOAD [161]
23	rs4862146	-	4	Glycoprotein buildup affects nerve cells in the brain [162]
24	rs6013406	ZFP64	20	-
25	rs1712417	TMEM87A	15	-

Table 36 - Top 25 SNPs as ranked by LP3 (α = 0.25) for ADRC data.

Rank	rsID	Gene	Chr	Comment
1	rs439401	APOE	19	In strong LD with rs7412 and rs429358 [63]
2	rs5157	APOC4	19	In strong LD with other APOC risk SNPs [64]
3	rs2075650	TOMM40	19	Predictive of longevity of LOAD patients [163, 164]
4	rs445925	-	19	Showed LOAD association in African-American cohort [65]
5	rs157182	ZNF433	19	Other zinc finger proteins linked to LOAD [165]
6	rs283129	PIN1	5	PIN1 linked to neural apoptosis in LOAD [166, 167]
7	rs17428956	-	1	-
8	rs11076978	-	16	-
9	rs5749272	NDRG1	22	NDRG family linked to neuron development, LOAD [168, 169]
10	rs6754487	-	2	-
11	rs439401	-	19	Near APOE, associated with LOAD [170]
12	rs3738269	IGFN1	1	-
13	rs10106829	LOC157273	8	-
14	rs12520115	-	5	-
15	rs17018886	-	2	-
16	rs17821171	-	15	-
17	rs523079	-	3	-
18	rs2314221	-	2	-
19	rs13059988	-	3	-
20	rs356611	-	5	-
21	rs10976056	KDM4C	9	-
22	rs10489926	PRG5	1	Brain-specific protein linked to axonal health [171]
23	rs10489924	PRG5	1	Brain-specific protein linked to axonal health [171]
24	rs2712599	-	12	-
25	rs10459209	-	12	-

Table 37 - Top 25 SNPs as ranked by SLR for TGen data.

Rank	rsID	Gene	Chr	Comment
1	rs7412	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
2	rs429358	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
3	rs7662187	PDGFC	4	-
4	rs10778921	TMTC2	12	-
5	rs7335085	-	13	-
6	rs12162084	-	16	Association found in another analysis of TGen data [157]
7	rs16923249	-	9	-
8	rs6508182	DCC	18	Implicated in axonal development [172]
9	rs4902299	-	14	-
10	rs10510990	-	3	-
11	rs16916338	GABBR2	9	Gene involved in neurotransmitters [10]
12	rs10894424	NTM	11	Gene implicated in LOAD [173]
13	rs1728390	-	16	-
14	rs4351927	GPC5	13	Involved in neuronal development [174]
15	rs16907781	ZBTB10	8	-
16	rs10871528	-	18	-
17	rs7848622	-	9	-
18	rs11846241	EML5	14	-
19	rs17044664	-	3	-
20	rs6540253	-	16	-
21	rs10176594	-	2	-
22	rs7243005	-	18	-
23	rs10740667	-	10	-
24	rs10966006	-	9	-
25	rs6824979	MMRN1	4	-

Table 38 - Top 25 SNPs as ranked by SLR for ADRC data.

Rank	rsID	Gene	Chr	Comment
1	rs429358	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
2	rs4420638	APOC	19	In strong linkage disequilibrium with APOE SNPs [156]
3	rs7412	APOE	19	APOE risk allele determined by rs7412 and rs429358 [53]
4	rs8083752	LOC643542	18	-
5	rs1978326	MAGI2	7	Gene associated with hippocampal volume reduction in AD [170]
6	rs6015314	APCDD1L-AS1	20	-
7	rs17767748	BTRC	10	-
8	rs11695991	NEU2	2	-
9	rs7210298	-	17	-
10	rs12190755	ZNF318	6	Gene expression level linked to AD [175]
11	rs7606208	SLC9A2	2	-
12	rs9932776	-	16	-
13	rs11680648	DIRC3	2	-
14	rs12100042	-	13	-
15	rs12257119	MYO3A	10	-
16	rs4147209	-	1	-
17	rs17099379	SYT16	14	-
18	rs7009155	-	-	-
19	rs801289	-	2	-
20	rs10862184	MYF5	12	-
21	rs16846388	SPATA16	3	-
22	rs9299784	KIAA1217	10	-
23	rs1759320	-	10	-
24	rs10507341	-	13	-
25	rs2276754	CCDC174	3	-

Table 39 - Top 25 SNPs as ranked by SWRF for TGen data.

Rank	rsID	Gene	Chr	Comment
1	rs7412	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
2	rs250857	FSTL4	5	-
3	rs9328529	-	9	-
4	rs934745	MAPK4	18	-
5	rs11077058	RBFOX1	16	RBFOX1 linked to brain volume in older adults [176]
6	s17124810	CBFA2T2	20	-
7	rs1251059	-	12	-
8	rs9908065	-	17	-
9	rs13213247	-	6	Significant association in meta-analysis of LOAD [177]
10	rs8112622	-	19	-
11	rs2779556	GABBR2	9	Gene involved in neurological pathways [178]
12	rs188429	RCL1	9	-
13	rs8108780	-	19	-
14	rs2796460	TLE1	9	-
15	rs16910463	-	9	-
16	rs16915130	GRM5	11	Gene is a coreceptor for LOAD-related protein [179]
17	rs16967491	-	15	-
18	rs200556	-	9	-
19	rs250855	FSTL4	5	-
20	rs4394475	-	9	-
21	rs10454604	-	13	-
22	rs8006542	FOXN3	14	-
23	rs865505	-	12	-
24	rs2712271	-	1	-
25	rs17141368	-	7	-

Table 40 - Top 25 SNPs as ranked by SWRF for ADRC data.

Rank	rsID	Gene	Chr	Comment
1	rs439401	-	19	Significant LOAD association [63]
2	rs445925	-	19	Located between APOE and APOC genes [68]
3	rs6434513	-	2	-
4	rs17245472	-	16	-
5	rs182662	RAB23	6	-
6	rs4494677	-	2	-
7	rs16906827	-	10	-
8	rs11108379	LTA4H	12	-
9	rs12683673	KDM4C	9	-
10	rs2712599	-	12	-
11	rs9297095	-	6	-
12	rs4270681	-	5	-
13	rs12592188	-	15	-
14	rs2442968	-	18	-
15	rs2442966	-	18	-
16	rs1834804	-	14	-
17	rs16963657	-	13	-
18	rs11820815	-	11	-
19	rs1892786	-	11	-
20	rs7004779	KCNK9	8	-
21	rs6669982	NFIA	1	-
22	rs9493552	-	6	-
23	rs11004700	-	10	-
24	rs6678065	NFIA	1	-
25	rs10095543	-	8	-

Table 41 - Top 25 SNPs as ranked by chi squared for TGen data.

Rank	rsID	Gene	Chr	Comment
1	rs4420638	APOC	19	In strong linkage disequilibrium with APOE SNPs [156]
2	rs7412	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
3	rs934745	MAPK4	18	-
4	rs429358	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
5	rs7079348	C10orf11	10	-
6	rs188429	RCL1	9	-
7	rs10824310	PRKG1	10	Significant association with LOAD [55]
8	rs6717497	-	2	-
9	rs16938663	STAU2	8	-
10	rs3732443	GXYLT2	3	-
11	rs6453333	-	5	-
12	rs12041702	-	1	-
13	rs17048190	-	2	-
14	rs2968848	-	7	-
15	rs16909497	-	10	-
16	rs10499687	VWC2	7	-
17	rs17169622	BMPER	7	-
18	rs41479848	MBIP	14	Involved in LOAD-associated MAPK pathway [180]
19	rs3007246	-	13	-
20	rs6429224	RGS7	1	Gene involved in brain signaling [181]
21	rs10845804	-	12	-
22	rs12109727	-	5	-
23	rs12476792	-	2	-
24	rs6455005	-	6	-
25	rs11804140	FBXO28	1	-

Table 42 - Top 25 SNPs as ranked by chi squared for ADRC data.

Rank	rsID	Gene	Chr	Comment
1	rs429358	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
2	rs4420638	APOC	19	In strong linkage disequilibrium with APOE SNPs [156]
3	rs157582	TOMM40	19	Showed LOAD association in African-American cohort [65]
4	rs2075650	APOE4	19	Predictive of longevity of LOAD patients [163, 164]
5	rs7412	APOE	19	APOE risk allele determined by rs7412 and rs429358 [155]
6	rs405509	APOE	19	APOE promoter varies LOAD risk [70]
7	rs8106922	TOMM40	19	Meta-analysis finds significant association with LOAD [69]
8	rs26845	ECI1	16	-
9	rs12507679	STAP1	4	-
10	rs13132585	STAP1	4	-
11	rs157580	TOMM40	19	Associated with LOAD in Chinese population [71]
12	rs4496012	-	13	-
13	rs8082842	RAB31	18	Gene involved in potential treatment [75]
14	rs9487940	-	6	-
15	rs34276	ACACB	12	-
16	rs4865859	-	5	-
17	rs7985095	-	13	-
18	rs16976268	-	18	-
19	rs4796922	-	18	-
20	rs16841336	PYHIN1	1	-
21	rs832156	IGFN1	1	-
22	rs9438881	-	1	-
23	rs283129	PIN1	5	PIN1 linked to neural apoptosis in LOAD [166, 167]
24	rs4480661		13	-
25	rs11985315	TRAPPC9	8	-

Table 43 - Top 25 SNPs as ranked by chi squared for CD data.

Rank	rsID	Gene	Chr	Comment
1	rs2076756	NOD2	16	Associated in independent study [102]
2	rs10210302	ATG16L1	2	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103]
3	rs6752107	ATG16L1	2	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103]
4	rs6431654	ATG16L1	2	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103]
5	rs3828309	ATG16L1	2	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103, 104]
6	rs17234657	-	5	WTCCC finding replicated in independent cohorts [105, 106]
7	rs2066843	NOD2	16	Associated with CD in independent study [102]
8	rs3792106	ATG16L1	2	Defects in gene cause susceptibility to IBD, sex-related risk for CD [103, 104]
9	rs11805303	IL23R	1	IL23 implicated in CD [107]
10	rs11957215	-	5	-
11	rs9292777	-	5	WTCCC SNP replicated in independent study [108]
12	rs10489629	IL23R	1	Replicated in multiple studies [102, 109, 110]
13	rs17221417	NOD2	16	NOD2 implicated in CD [102]
14	rs2201841	IL23R	1	SNP implicated in distinct populations [111]
15	rs4957295	-	5	-
16	rs10213846	-	5	-
17	rs6871834	-	5	-
18	rs4957297	-	5	Replicated independent of WTCCC [112]
19	rs4957300	-	5	-
20	rs16869934	-	5	Discovered in BIC analysis of WTCCC [113]
21	rs12119179	-	1	Associated with a disease with similar genetic profile [114]
22	rs11209033	-	1	Cited in patent for testing for autoimmune-associated polymorphisms [115]
23	rs10512734	-	5	SNP validated in independent study [182]
24	rs7546245	-	1	In moderate LD with rs11805303 [107]
25	rs41396545	IL23R	1	Discovered in BIC analysis of WTCCC [113]

Table 44 - Top 25 SNPs as ranked by chi squared for HT data.

Rank	**rsID**	**Gene**	**Chr**	**Comment**
1	rs4765066	-	12	-
2	rs488101	-	9	Associated with arterial plaque [116]
3	rs4867173	-	5	-
4	rs11782342	KCNB2	8	Discovered as part of epistatic interactions in WTCCC data [86]
5	rs11024327	OTOG	11	Found in combined analysis of WTCCC data [117]
6	rs2820037	1	-	SNP associated with BP regulation [119]
7	rs2790622	-	1	-
8	rs2820038	-	1	-
9	rs6574988	-	14	-
10	rs2820046	-	1	-
11	rs16945811	YWHAE	17	Gene implicated in HT [120]
12	rs9428826	-	1	-
13	rs2398162	NR2F2-AS1	15	Population-specific association has been replicated [121]
14	rs2820026	-	1	-
15	rs921535	-	15	-
16	rs10889923	NEGR1	1	-
17	rs2191003	-	4	-
18	rs300916	GAB1	4	Discovered in combined WTCCC + Australian cohort [117]
19	rs1935683	-	6	-
20	rs13119672	PPARGC1A	4	Gene associated with HT [124]
21	rs11110912	MYBPC1	12	WTCCC SNP replicated in independent study [119]
22	rs2840584	SLAMF9	1	-
23	rs633568	-	11	-
24	rs1036392	-	2	-
25	rs973009	ACTN4	4	-

BIBLIOGRAPHY

1. Korosok MR, Wei WH, Farrell PM: **The incidence of cystic fibrosis.** *Statistics in medicine* 1996, **15:**449-462.
2. Shriner D, Adeyemo A, Gerry NP, Herbert A, Chen G, Doumatey A, Huang H, Zhou J, Christman MF, Rotimi CN: **Transferability and fine-mapping of genome-wide associated loci for adult height across human populations.** *PLoS ONE* 2009, **4:**e8398.
3. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, Monroe KR, Kolonel LN, Altshuler D, Henderson BE: **Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups.** *PLoS genetics* 2010, **6:**e1001078.
4. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI: **Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms.** *Am J Hum Genet* 2008, **82:**100-112.
5. Wray NR, Purcell SM, Visscher PM: **Synthetic associations created by rare variants do not explain most GWAS results.** *PLoS Biol* 2011, **9:**e1000579.
6. Helfand BT, Fought AJ, Loeb S, Meeks JJ, Kan D, Catalona WJ: **Genetic prostate cancer risk assessment: common variants in 9 genomic regions are associated with cumulative risk.** *J Urol* 2010, **184:**501-505.
7. Schork NJ, Murray SS, Frazer KA, Topol EJ: **Common vs. rare allele hypotheses for complex diseases.** *Curr Opin Genet Dev* 2009, **19:**212-219.
8. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)** [www.genome.gov/sequencingcosts]
9. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat Genet* 2007, **39:**1181-1186.
10. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.
11. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association.** *Lancet* 2003, **361:**598-604.
12. Masaeli M, Dy JG, Fung GM: **From transformation-based dimensionality reduction to feature selection.** In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010: 751-758.
13. Zhu X, Ghahramani Z: **Learning from labeled and unlabeled data with label propagation.** In *Book Learning from labeled and unlabeled data with label propagation* (Editor ed.^eds.). City: Technical Report CMU-CALD-02-107, Carnegie Mellon University; 2002.

14. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23:**2507-2517.

15. Kamboh MI, Demirci FY, Wang X, Minster RL, Carrasquillo MM, Pankratz VS, Younkin SG, Saykin AJ, Jun G, Baldwin C, et al: **Genome-wide association study of Alzheimer's disease.** *Transl Psychiatry* 2012, **2:**e117.

16. Wei W, Visweswaran S, Cooper GF: **The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.** *J Am Med Inform Assoc* 2011, **18:**370-375.

17. Kira K, Rendell L: **A practical approach to feature selection.** In *ML92: Proceedings of the ninth international workshop on Machine learning*. Morgan Kaufmann Publishers Inc.; 1992: 249-256.

18. Greene CS, Penrod NM, Kiralis J, Moore JH: **Spatially uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions.** *BioData Mining* 2009, **2:**5.

19. Greene C, Himmelstein D, Kiralis J, Moore J: **The Informative Extremes: Using Both Nearest and Farthest Individuals Can Improve Relief Algorithms in the Domain of Human Genetics.** In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Volume* 6023. Edited by Pizzuti C, Ritchie M, Giacobini M: Springer Berlin / Heidelberg; 2010: 182-193: *Lecture Notes in Computer Science*].

20. Stokes M, Visweswaran S: **Application of a spatially-weighted Relief algorithm for ranking genetic predictors of disease.** *BioData Mining* 2012, **5:**20.

21. Mooney M, Wilmot B, Study TBG, McWeeney S: **The GA and the GWAS: Using Genetic Algorithms to Search for Multilocus Associations.** *IEEE/ACM Trans Comput Biol Bioinformatics* 2012, **9:**899-910.

22. Nolan VG, Sebastiani P, Baldwin C, Wyszynski DF, Farrer LA, Steinberg MH: **Modeling genetic polymorhphisms and sickle cell associated vasoocclusive events using classification and regression trees (CART).** *Annals of epidemiology* 2005, **15:**644.

23. Winham S, Colby C, Freimuth R, Wang X, de Andrade M, Huebner M, Biernacka J: **SNP interaction detection with Random Forests in high-dimensional genetic data.** *BMC Bioinformatics* 2012, **13:**1-13.

24. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF: **An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings.**

25. Duan KB, Rajapakse JC, Wang H, Azuaje F: **Multiple SVM-RFE for gene selection in cancer classification with expression data.** *IEEE Trans Nanobioscience* 2005, **4:**228-234.

26. Tang EK, Suganthan P, Yao X: **Gene selection algorithms for microarray data based on least squares support vector machine.** *BMC Bioinformatics* 2006, **7:**95.

27. Chen L, Yu G, Langefeld CD, Miller DJ, Guy RT, Raghuram J, Yuan X, Herrington DM, Wang Y: **Comparative analysis of methods for detecting interacting loci.** *BMC Genomics* 2011, **12:**344.

28. Yang C, He Z, Wan X, Yang Q, Xue H, Yu W: **SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies.** *Bioinformatics* 2009, **25:**504-511.

29. Miller DJ, Zhang Y, Yu G, Liu Y, Chen L, Langefeld CD, Herrington D, Wang Y: **An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions.** *Bioinformatics* 2009, **25:**2478-2485.
30. Hahn LW, Ritchie MD, Moore JH: **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.** *Bioinformatics* 2003, **19:**376-382.
31. Slavkov Ip, Zenko Bp, Dzeroski Sp: *Evaluation Method for Feature Rankings and their Aggregations for Biomarker Discovery.\par*. 2010.
32. Prati RC: **Combining feature ranking algorithms through rank aggregation.**
33. Renda ME, Straccia U: **Web metasearch: rank vs. score based rank aggregation methods.** In *Book Web metasearch: rank vs. score based rank aggregation methods* (Editor ed.^eds.). pp. 841-846. City: ACM; 2003:841-846.
34. Hwang T, Sicotte H, Tian Z, Wu B, Kocher JP, Wigle DA, Kumar V, Kuang R: **Robust and efficient identification of biomarkers by classifying features on graphs.** *Bioinformatics (Oxford, England)* 2008, **24:**2023-2029.
35. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q: **GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.** *Genome Biology* 2008, **9:**S4.
36. Teramoto R: **Prediction of Alzheimer's diagnosis using semi-supervised distance metric learning with label propagation.** *Comput Biol Chem* 2008, **32:**438-441.
37. Zhou D, Bousquet O, Lal T, Weston J, Scholkopf B: **Learning with local and global consistency.** In *Advances in Neural Information Processing Systems 16*. 2004
38. Belkin M, Niyogi P: **Using manifold stucture for partially labeled classification.** In *Advances in neural information processing systems*. 2002: 929-936.
39. Belkin M, Niyogi P: **Laplacian eigenmaps for dimensionality reduction and data representation.** *Neural computation* 2003, **15:**1373-1396.
40. Belkin M, Matveeva I, Niyogi P: **Regularization and Semi-supervised Learning on Large Graphs.** In *Learning Theory. Volume* 3120. Edited by Shawe-Taylor J, Singer Y: Springer Berlin Heidelberg; 2004: 624-638: *Lecture Notes in Computer Science*].
41. **The International HapMap Project.** *Nature* 2003, **426:**789-796.
42. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073-1081.
43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7:**248-249.
44. Hubisz MJ, Pollard KS, Siepel A: **PHAST and RPHAST: phylogenetic analysis with space/time models.** *Briefings in Bioinformatics* 2011, **12:**41-51.
45. Davydov Ev Fau - Goode DL, Goode Dl Fau - Sirota M, Sirota M Fau - Cooper GM, Cooper Gm Fau - Sidow A, Sidow A Fau - Batzoglou S, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++.**
46. Ghosh S, Bickeböller H, Bailey J, Bailey-Wilson JE, Cantor R, Culverhouse R, Daw W, DeStefano AL, Engelman CD, Hinrichs A: **Identifying rare variants from exome scans: the GAW17 experience.** In *BMC proceedings*. BioMed Central Ltd; 2011: S1.

47. Almasy L, Dyer TD, Peralta JM, Kent JW, Charlesworth JC, Curran JE, Blangero J: **Genetic Analysis Workshop 17 mini-exome simulation.** In *BMC proceedings*. BioMed Central Ltd; 2011: S2.

48. Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshipura KD, Pearson JV, Hu-Lince D, Huentelman Matthew J, et al: **GAB2 Alleles Modify Alzheimer's Risk in APOE e4 Carriers.** 2007, **54:**713-720.

49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *The American Journal of Human Genetics* 2007, **81:**559-575.

50. Cariaso M, Lennon G: **SNPedia: a wiki supporting personal genome annotation, interpretation and analysis.** *Nucleic Acids Res* 2012, **40:**D1308-1312.

51. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: integrating information about genes, proteins and diseases.** *Trends Genet* 1997, **13:**163.

52. Yamashita O, Sato MA, Yoshioka T, Tong F, Kamitani Y: **Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns.** *Neuroimage* 2008, **42:**1414-1429.

53. Izaks Gj Fau - Gansevoort RT, Gansevoort Rt Fau - van der Knaap AM, van der Knaap Am Fau - Navis G, Navis G Fau - Dullaart RPF, Dullaart Rp Fau - Slaets JPJ, Slaets JP: **The association of APOE genotype with cognitive function in persons aged 35 years or older.**

54. Bertram L Fau - Lange C, Lange C Fau - Mullin K, Mullin K Fau - Parkinson M, Parkinson M Fau - Hsiao M, Hsiao M Fau - Hogan MF, Hogan Mf Fau - Schjeide BMM, Schjeide Bm Fau - Hooli B, Hooli B Fau - Divito J, Divito J Fau - Ionita I, Ionita I Fau - Jiang H, et al: **Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE.**

55. Fallin M, Szymanski M, Wang R, Gherman A, Bassett S, Avramopoulos D: **Fine mapping of the chromosome 10q11-q21 linkage region in Alzheimer's disease cases and controls.** *neurogenetics* 2010, **11:**335-348.

56. Shioya M, Obayashi S, Tabunoki H, Arima K, Saito Y, Ishida T, Satoh J: **Aberrant microRNA expression in the brains of neurodegenerative diseases: miR-29a decreased in Alzheimer disease brains targets neurone navigator 3.** *Neuropathol Appl Neurobiol* 2010, **36:**320-330.

57. Hu W, Chen-Plotkin A, Arnold S, Grossman M, Clark C, Shaw L, Pickering E, Kuhn M, Chen Y, McCluskey L, et al: **Novel CSF biomarkers for Alzheimer’s disease and mild cognitive impairment.** *Acta Neuropathol* 2010, **119:**669-678.

58. Shi H Fau - Medway C, Medway C Fau - Bullock J, Bullock J Fau - Brown K, Brown K Fau - Kalsheker N, Kalsheker N Fau - Morgan K, Morgan K: **Analysis of Genome-Wide Association Study (GWAS) data looking for replicating signals in Alzheimer's disease (AD).**

59. Jiang X, Barmada MM, Becich MJ: **Evaluating De Novo Locus-Disease Discoveries in GWAS Using the Signal-to-Noise Ratio.** *AMIA Annu Symp Proc* 2011, **2011:**617-624.

60. Wei W, Visweswaran S, Cooper G: **The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.** *Journal of the American Medical Informatics Association* 2011, **18:**370-375.

61. Liu W, Ding J, Gibbs JR, Wang SJ, Hardy J, Singleton A: **A simple and efficient algorithm for genome-wide homozygosity analysis in disease.** *Mol Syst Biol* 2009, **5:**304.

62. Briones N, Dinu V: **Data mining of high density genomic variant data for prediction of Alzheimer's disease risk.** *BMC Med Genet* 2012, **13:**7.

63. Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'Donovan MC, et al: **A genome-wide association study for late-onset Alzheimer's disease using DNA pooling.** *BMC Med Genomics* 2008, **1:**44.

64. Cervantes S, Samaranch L, Vidal-Taboada JM, Lamet I, Bullido MJ, Frank-García A, Coria F, Lleó A, Clarimón J, Lorenzo E, et al: **Genetic variation in APOE cluster region and Alzheimer's disease risk.** *Neurobiology of Aging* 2011, **32:**2107.e2107-2107.e2117.

65. Logue Mw SMVBN, et al.: **A comprehensive genetic association study of Alzheimer disease in African Americans.** *Archives of Neurology* 2011, **68:**1569-1579.

66. Shi H Fau - Belbin O, Belbin O Fau - Medway C, Medway C Fau - Brown K, Brown K Fau - Kalsheker N, Kalsheker N Fau - Carrasquillo M, Carrasquillo M Fau - Proitsi P, Proitsi P Fau - Powell J, Powell J Fau - Lovestone S, Lovestone S Fau - Goate A, Goate A Fau - Younkin S, et al: **Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS).**

67. Deelen J Fau - Beekman M, Beekman M Fau - Uh H-W, Uh Hw Fau - Helmer Q, Helmer Q Fau - Kuningas M, Kuningas M Fau - Christiansen L, Christiansen L Fau - Kremer D, Kremer D Fau - van der Breggen R, van der Breggen R Fau - Suchiman HED, Suchiman He Fau - Lakenberg N, Lakenberg N Fau - van den Akker EB, et al: **Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited.**

68. Jun G, Vardarajan BN, Buros J, Yu C, Hawk MV, Dombroski BA, Crane PK, Larson EB, Mayeux R, Haines JL, et al: **Comprehensive Search for Alzheimer Disease Susceptibility Loci in the APOE Region.** *Arch Neurol* 2012**:**1-10.

69. **The AlzGene Database** [http://www.alzgene.org]

70. Bizzarro A, Seripa, D., Acciarri, A. Matera, M.G., Pilotto, A., Tiziano, F.D., Brache, C., et al.: **The complex interaction between APOE promoter and AD: an Italian case study.** *European Journal of Human Genetics* 2009, **17:**7.

71. Ma XY, Yu JT, Wang W, Wang HF, Liu QY, Zhang W, Tan L: **Association of TOMM40 polymorphisms with late-onset Alzheimer's disease in a Northern Han Chinese population.** *Neuromolecular Med* 2013, **15:**279-287.

72. Vitali M, Venturelli E, Galimberti D, Benerini Gatta L, Scarpini E, Finazzi D: **Analysis of the genes coding for subunit 10 and 15 of cytochrome c oxidase in Alzheimer's disease.** *J Neural Transm* 2009, **116:**1635-1641.

73. Reitz C, Tosto G, Vardarajan B, Rogaeva E, Ghani M, Rogers RS, Conrad C, Haines JL, Pericak-Vance MA, Fallin MD, et al: **Independent and epistatic effects of variants in VPS10-d receptors on Alzheimer disease risk and processing of the amyloid precursor protein (APP).** *Transl Psychiatry* 2013, **3:**e256.

74. Hooli BV, Kovacs-Vajna ZM, Mullin K, Blumenthal MA, Mattheisen M, Zhang C, Lange C, Mohapatra G, Bertram L, Tanzi RE: **Rare autosomal copy number variations in early-onset familial Alzheimer/'s disease.** *Mol Psychiatry* 2014, **19:**676-681.

75. Polhner JH, DE): **Diagnostic and therapeutic use of a rab family gtp-binding protein for neurodegenerative diseases.** In *Book Diagnostic and therapeutic use of a rab family gtp-binding protein for neurodegenerative diseases* (Editor ed.^eds.). City: Evotec Neurosciences GmbH (Schnackenburgallee 114, Hamburg, DE); 2006.

76. Botta V: **A walk into random forests: adaptation and application to Genome-Wide Association Studies.** 2013.

77. Pirooznia M, Seifuddin F, Judy J, Goes FS, Potash JB, Zandi PP: **Metamoodics: meta-analysis and bioinformatics resource for mood disorders.** *Mol Psychiatry* 2014, **19:**748-749.

78. Frank B, Niesler B, Nöthen MM, Neidt H, Propping P, Bondy B, Rietschel M, Maier W, Albus M, Rappold G: **Investigation of the human serotonin receptor gene HTR3B in bipolar affective and schizophrenic patients.** *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2004, **131:**1-5.

79. Lee M, Chen C, Lee C, Chen C, Chong M, Ouyang W, Chiu N, Chuo L, Chen C, Tan H: **Genome-wide association study of bipolar I disorder in the Han Chinese population.** *Molecular psychiatry* 2010, **16:**548-556.

80. Jiang Y, Zhang H: **Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder.** *Genet Epidemiol* 2011, **35:**125-132.

81. Tesli M, Athanasiu L, Mattingsdal M, Kähler AK, Gustafsson O, Andreassen BK, Werge T, Hansen T, Mors O, Mellerup E: **Association analysis of PALB2 and BRCA2 in bipolar disorder and schizophrenia in a Scandinavian case–control sample.** *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2010, **153:**1276-1282.

82. Amano K, Yamada K, Iwayama Y, Detera-Wadleigh SD, Hattori E, Toyota T, Tokunaga K, Yoshikawa T, Yamakawa K: **Association study between the Down syndrome cell adhesion molecule (DSCAM) gene and bipolar disorder.** *Psychiatr Genet* 2008, **18:**1-10.

83. **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.

84. Drago A, Giegling I, Schafer M, Hartmann AM, Friedl M, Konte B, Moller HJ, De Ronchi D, Stassen HH, Serretti A, Rujescu D: **AKAP13, CACNA1, GRIK4 and GRIA1 genetic variations may be associated with haloperidol efficacy during acute treatment.** *Eur Neuropsychopharmacol* 2013, **23:**887-894.

85. Ollila HM, Soronen P, Silander K, Palo OM, Kieseppa T, Kaunisto MA, Lonnqvist J, Peltonen L, Partonen T, Paunio T: **Findings from bipolar disorder genome-wide association studies replicate in a Finnish bipolar family-cohort.** *Mol Psychiatry* 2009, **14:**351-353.

86. Goudey B, Rawlinson D, Wang Q, Shi F, Ferra H, Campbell RM, Stern L, Inouye MT, Ong CS, Kowalczyk A: **GWIS--model-free, fast and exhaustive search for epistatic interactions in case-control GWAS.** *BMC Genomics* 2013, **14 Suppl 3:**S10.

87. Belzeaux R, Bergon A, Jeanjean V, Loriod B, Formisano-Treziny C, Verrier L, Loundou A, Baumstarck-Barrau K, Boyer L, Gall V, et al: **Responder and nonresponder patients exhibit different peripheral transcriptional signatures during major depressive episode.** *Transl Psychiatry* 2012, **2:**e185.

88. de las Fuentes L, Yang W, Dávila-Román VG, Gu CC: **Pathway-based genome-wide association analysis of coronary heart disease identifies biologically important gene sets.** *European Journal of Human Genetics* 2012, **20:**1168-1173.

89. Iida K, Hidaka K, Takeuchi M, Nakayama M, Yutani C, Mukai T, Morisaki T: **Expression of MEF2 genes during human cardiac development.** *Tohoku J Exp Med* 1999, **187:**15-23.

90. Ellsworth DL, Croft DT, Weyandt J, Sturtz LA, Blackburn HL, Burke A, Haberkorn MJ, McDyer FA, Jellema GL, van Laar R: **Intensive Cardiovascular Risk Reduction Induces Sustainable Changes in Expression of Genes and Pathways Important to Vascular Function.** *Circulation: Cardiovascular Genetics* 2014**:**CIRCGENETICS. 113.000121.

91. den Hoed M, Eijgelsheim M, Esko T, Brundel BJ, Peal DS, Evans DM, Nolte IM, Segrè AV, Holm H, Handsaker RE: **Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders.** *Nature Genetics* 2013, **45:**621-631.

92. Brinza D, Schultz M, Tesler G, Bafna V: **RAPID detection of gene-gene interactions in genome-wide association studies.** *BioInformatics* 2010, **26:**2856-2862.

93. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, et al: **Genomewide association analysis of coronary artery disease.** *N Engl J Med* 2007, **357:**443-453.

94. Hinohara K, Nakajima T, Takahashi M, Hohda S, Sasaoka T, Nakahara K, Chida K, Sawabe M, Arimura T, Sato A, et al: **Replication of the association between a chromosome 9p21 polymorphism and coronary artery disease in Japanese and Korean populations.** *J Hum Genet* 2008, **53:**357-359.

95. Lippert C, Listgarten J, Davidson RI, Baxter J, Poon H, Kadie CM, Heckerman D: **An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data.** *Sci Rep* 2013, **3**.

96. Baudhuin LM: **Genetics of coronary artery disease: focus on genome-wide association studies.** *Am J Transl Res* 2009, **1:**221-234.

97. Yan Y, Hu Y, North KE, Franceschini N, Lin D: **Evaluation of population impact of candidate polymorphisms for coronary heart disease in the Framingham Heart Study Offspring Cohort.** *BMC Proc* 2009, **3 Suppl 7:**S118.

98. Li X, Huang Y, Yin D, Wang D, Xu C, Wang F, Yang Q, Wang X, Li S, Chen S, et al: **Meta-analysis identifies robust association between SNP rs17465637 in MIA3 on chromosome 1q41 and coronary artery disease.** *Atherosclerosis* 2013, **231:**136-140.

99. Li MX, Gui HS, Kwan JS, Sham PC: **GATES: a rapid and powerful gene-based association test using extended Simes procedure.** *Am J Hum Genet* 2011, **88:**283-293.

100. Gallagher G, Brazaitis J, Yu R: **The Crohn's Disease protective SNP rs11209026 mediates alternative splicing in human IL23R.** *The Journal of Immunology* 2010, **184:**51.14.

101. Lacher M, Schroepf S, Helmbrecht J, von Schweinitz D, Ballauff A, Koch I, Lohse P, Osterrieder S, Kappler R, Koletzko S: **Association of the interleukin-23 receptor gene variant rs11209026 with Crohn's disease in German children.** *Acta Paediatr* 2010, **99:**727-733.

102. Glas J, Seiderer J, Wetzke M, Konrad A, Torok HP, Schmechel S, Tonenchi L, Grassl C, Dambacher J, Pfennig S, et al: **rs1004819 is the main disease-associated IL23R variant**

in German Crohn's disease patients: combined analysis of IL23R, CARD15, and OCTN1/2 variants. *PLoS ONE* 2007, **2:**e819.

103. Liu LY, Schaub MA, Sirota M, Butte AJ: **Transmission distortion in Crohn's disease risk gene ATG16L1 leads to sex difference in disease association.** *Inflamm Bowel Dis* 2012, **18:**312-322.

104. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, et al: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40:**955-962.

105. Weersma RK, Stokkers PC, Cleynen I, Wolfkamp SC, Henckaerts L, Schreiber S, Dijkstra G, Franke A, Nolte IM, Rutgeerts P, et al: **Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch-Belgian cohort.** *Am J Gastroenterol* 2009, **104:**630-638.

106. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D, et al: **Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility.** *Nat Genet* 2007, **39:**830-832.

107. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, et al: **Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease.** *Am J Hum Genet* 2009, **84:**399-405.

108. Fisher SA, Tremelling M, Anderson CA, Gwilliam R, Bumpstead S, Prescott NJ, Nimmo ER, Massey D, Berzuini C, Johnson C, et al: **Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease.** *Nat Genet* 2008, **40:**710-712.

109. Amre DK, Mack D, Israel D, Morgan K, Lambrette P, Law L, Grimard G, Deslandres C, Krupoves A, Bucionis V, et al: **Association between genetic variants in the IL-23R gene and early-onset Crohn's disease: results from a case-control and family-based study among Canadian children.** *Am J Gastroenterol* 2008, **103:**615-620.

110. Kanaan Z, Ahmad S, Bilchuk N, Vahrenhold C, Pan J, Galandiuk S: **Perianal Crohn's disease: predictive factors and genotype-phenotype correlations.** *Dig Surg* 2012, **29:**107-114.

111. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, et al: **A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.** *Science* 2006, **314:**1461-1463.

112. Hoffman GE, Logsdon BA, Mezey JG: **PUMA: a unified framework for penalized multiple regression analysis of GWAS data.** *PLoS computational biology* 2013, **9:**e1003101.

113. Dolejsi E, Bodenstorfer B, Frommlet F: **Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian information criterion.** *arXiv preprint arXiv:14036623* 2014.

114. Mizuki N, Meguro A, Ota M, Ohno S, Shiota T, Kawagoe T, Ito N, Kera J, Okada E, Yatsu K, et al: **Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behcet's disease susceptibility loci.** *Nat Genet* 2010, **42:**703-706.

115. Schrodi SJ, Li Y: **Genetic polymorphisms associated with autoinflammatory diseases, methods of detection and uses thereof.** In *Book Genetic polymorphisms associated with*

autoinflammatory diseases, methods of detection and uses thereof (Editor ed.^eds.). City: Google Patents; 2009.

116. Gardener H, Beecham A, Cabral D, Yanuck D, Slifer S, Wang L, Blanton SH, Sacco RL, Juo SH, Rundek T: **Carotid plaque and candidate genes related to inflammation and endothelial function in Hispanics from northern Manhattan.** *Stroke* 2011, **42:**889-896.

117. Fowdar JY: **Identification of Hypertension Genes Following a Genome-Wide Association Scan.** Griffith University, School of Medical Science; 2012.

118. Adeyemo A, Gerry N, Chen G, Herbert A, Doumatey A, Huang H, Zhou J, Lashley K, Chen Y, Christman M, Rotimi C: **A genome-wide association study of hypertension and blood pressure in African Americans.** *PLoS Genet* 2009, **5:**e1000564.

119. Kostis WJ, Cabrera J, Hooper WC, Whelton PK, Espeland MA, Cosgrove NM, Cheng JQ, Deng Y, De Staerck C, Pyle M, et al: **Relationships between selected gene polymorphisms and blood pressure sensitivity to weight loss in elderly persons with hypertension.** *Hypertension* 2013, **61:**857-863.

120. Schiff M, Delahaye A, Andrieux J, Sanlaville D, Vincent-Delorme C, Aboura A, Benzacken B, Bouquillon S, Elmaleh-Berges M, Labalme A: **Further delineation of the 17p13. 3 microdeletion involving YWHAE but distal to PAFAH1B1: Four additional patients.** *European journal of medical genetics* 2010, **53:**303-308.

121. Ehret GB, Morrison AC, O'Connor AA, Grove ML, Baird L, Schwander K, Weder A, Cooper RS, Rao DC, Hunt SC, et al: **Replication of the Wellcome Trust genome-wide association study of essential hypertension: the Family Blood Pressure Program.** *Eur J Hum Genet* 2008, **16:**1507-1511.

122. Langley SR, Bottolo L, Kunes J, Zicha J, Zidek V, Hubner N, Cook SA, Pravenec M, Aitman TJ, Petretto E: **Systems-level approaches reveal conservation of trans-regulated genes in the rat and genetic determinants of blood pressure in humans.** *Cardiovasc Res* 2013, **97:**653-665.

123. Inanloo Rahatloo K, Parsa AFZ, Huse K, Rasooli P, Davaran S, Platzer M, Kramer M, Fan J-B, Turk C, Amini S: **Mutation in ST6GALNAC5 identified in family with coronary artery disease.** *Scientific reports* 2014, **4**.

124. Rojek A, Cielecka-Prynda M, Przewlocka-Kosmala M, Laczmanski L, Mysiak A, Kosmala W: **Impact of the PPARGC1A Gly482Ser polymorphism on left ventricular structural and functional abnormalities in patients with hypertension.** *J Hum Hypertens* 2014.

125. Rantapaa-Dahlqvist SM: **Single Nucleotide Polymorphisms within the HLA-DRB1 Gene in Relation to Antibodies Against Citrullinated Peptides in Individuals Prior to the Development of Rheumatoid Arthritis.** In.; 2012: 412.

126. El-Gabalawy HS, Robinson DB, Daha NA, Oen KG, Smolik I, Elias B, Hart D, Bernstein CN, Sun Y, Lu Y, et al: **Non-HLA genes modulate the risk of rheumatoid arthritis associated with HLA-DRB1 in a susceptible North American Native population.** *Genes Immun* 2011, **12:**568-574.

127. Plant D, Deborah P. Symmons, Jane Worthington, David Strachan, Anne Barton: **Genome-Wide Association Study of Rheumatoid Arthritis, Stratified by Smoking Status.** In *American College of Rheumatology/Association of Rheumatology Health Professionals Annual Scientific Meeting; Chicago, Illinois*. 2011: 164.

128. Yen JH, Chen CJ, Tsai WC, Ou TT, Lin CH, Lin SC, Liu HW: **HLA-DQA1 genotyping in patients with rheumatoid arthritis in Taiwan.** *Kaohsiung J Med Sci* 2001, **17:**183-189.

129. Julia A, Ballina J, Canete JD, Balsa A, Tornero-Molina J, Naranjo A, Alperi-Lopez M, Erra A, Pascual-Salcedo D, Barcelo P, et al: **Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility.** *Arthritis Rheum* 2008, **58:**2275-2286.

130. Thompson SD, Sudman M, Ramos PS, Marion MC, Ryan M, Tsoras M, Weiler T, Wagner M, Keddache M, Haas JP, et al: **The susceptibility loci juvenile idiopathic arthritis shares with other autoimmune diseases extend to PTPN2, COG6, and ANGPT1.** *Arthritis Rheum* 2010, **62:**3265-3276.

131. Prahalad S, Hansen S, Whiting A, Guthery SL, Clifford B, McNally B, Zeft AS, Bohnsack JF, Jorde LB: **Variants in TNFAIP3, STAT4, and C12orf30 loci associated with multiple autoimmune diseases are also associated with juvenile idiopathic arthritis.** *Arthritis Rheum* 2009, **60:**2124-2130.

132. Steer S, Abkevich V, Gutin A, Cordell HJ, Gendall KL, Merriman ME, Rodger RA, Rowley KA, Chapman P, Gow P, et al: **Genomic DNA pooling for whole-genome association scans in complex disease: empirical demonstration of efficacy in rheumatoid arthritis.** *Genes Immun* 2007, **8:**57-68.

133. Barton A, Thomson W, Ke X, Eyre S, Hinks A, Bowes J, Plant D, Gibbons LJ, Wilson AG, Bax DE: **Identification of novel RA susceptibility loci at chromosomes 10p15, 12q13 and 22q13.** *Nature Genetics* 2008, **40:**1156.

134. Smyth DJ, Cooper JD, Howson JM, Walker NM, Plagnol V, Stevens H, Clayton DG, Todd JA: **PTPN22 Trp620 explains the association of chromosome 1p13 with type 1 diabetes and shows a statistical interaction with HLA class II genotypes.** *Diabetes* 2008, **57:**1730-1737.

135. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, et al: **Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes.** *Nat Genet* 2007, **39:**857-864.

136. Viken MK, Blomhoff A, Olsson M, Akselsen HE, Pociot F, Nerup J, Kockum I, Cambon-Thomsen A, Thorsby E, Undlien DE, Lie BA: **Reproducible association with type 1 diabetes in the extended class I region of the major histocompatibility complex.** *Genes Immun* 2009, **10:**323-333.

137. Butty V, Campbell C, Mathis D, Benoist C: **Impact of diabetes susceptibility loci on progression from pre-diabetes to diabetes in at-risk individuals of the diabetes prevention trial-type 1 (DPT-1).** *Diabetes* 2008, **57:**2348-2359.

138. Salonen J: **Novel genes and markers associated with high-density lipoprotein-cholesterol (HDL-C).** In *Book Novel genes and markers associated with high-density lipoprotein-cholesterol (HDL-C)* (Editor ed.^eds.). City: Google Patents; 2006.

139. Abellan JJ, Abellan C, Gonzalez JR: **A Bayesian shared component model for genetic association studies.** 2010.

140. An P, Feitosa M, Ketkar S, Adelman A, Lin S, Borecki I, Province M: **Epistatic interactions of CDKN2B-TCF7L2 for risk of type 2 diabetes and of CDKN2B-JAZF1 for triglyceride/high-density lipoprotein ratio longitudinal change: evidence from the Framingham Heart Study.** *BMC Proc* 2009, **3 Suppl 7:**S71.

141. Gupta V, Khadgawat R, Ng HK, Kumar S, Aggarwal A, Rao VR, Sachdeva MP: **A validation study of type 2 diabetes-related variants of the TCF7L2, HHEX, KCNJ11, and ADIPOQ genes in one endogamous ethnic group of north India.** *Ann Hum Genet* 2010, **74:**361-368.

142. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, Hitman GA, Walker M, Wiltshire S, Hattersley AT, McCarthy MI: **Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk.** *Diabetes* 2006, **55:**2640-2644.

143. Rong R, Hanson RL, Ortiz D, Wiedrich C, Kobes S, Knowler WC, Bogardus C, Baier LJ: **Association analysis of variation in/near FTO, CDKAL1, SLC30A8, HHEX, EXT2, IGF2BP2, LOC387761, and CDKN2B with type 2 diabetes and related quantitative traits in Pima Indians.** *Diabetes* 2009, **58:**478-488.

144. Yajnik CS, Janipalli CS, Bhaskar S, Kulkarni SR, Freathy RM, Prakash S, Mani KR, Weedon MN, Kale SD, Deshpande J, et al: **FTO gene variants are strongly associated with type 2 diabetes in South Asian Indians.** *Diabetologia* 2009, **52:**247-252.

145. Hertel JK, Johansson S, Raeder H, Midthjell K, Lyssenko V, Groop L, Molven A, Njolstad PR: **Genetic analysis of recently identified type 2 diabetes loci in 1,638 unselected patients with type 2 diabetes and 1,858 control participants from a Norwegian population-based cohort (the HUNT study).** *Diabetologia* 2008, **51:**971-977.

146. Chandak GR, Janipalli CS, Bhaskar S, Kulkarni SR, Mohankrishna P, Hattersley AT, Frayling TM, Yajnik CS: **Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population.** *Diabetologia* 2007, **50:**63-67.

147. Pang DX, Smith AJ, Humphries SE: **Functional analysis of TCF7L2 genetic variants associated with type 2 diabetes.** *Nutr Metab Cardiovasc Dis* 2013, **23:**550-556.

148. Chan KHK: **Assessing the genetic architecture of metabolic diseases using candidate gene and genome-wide approach.** In *Book Assessing the genetic architecture of metabolic diseases using candidate gene and genome-wide approach* (Editor ed.^eds.). City; 2012.

149. Gul H, Y. C. Acikel, Y. Aydin Son: **Discovering missing heritability and early risk prediction for type 2 diabetes; a new perspective for genome-wide association study analysis with the Nurses Health Study and the Health Professionals Follow-Up Study.** *Turkish Journal of Medical Sciences* 2014.

150. Sitek A, Rosset I, Strapagiel D, Majewska M, Ostrowska-Nawarycz L, Żądzińska E: **Association of FTO gene with obesity in Polish schoolchildren.** *AnthropologicAl review* 2014, **77:**33-44.

151. Wu Y, Li H, Loos RJ, Yu Z, Ye X, Chen L, Pan A, Hu FB, Lin X: **Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8, and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population.** *Diabetes* 2008, **57:**2834-2842.

152. Chang YC, Liu PH, Lee WJ, Chang TJ, Jiang YD, Li HY, Kuo SS, Lee KC, Chuang LM: **Common variation in the fat mass and obesity-associated (FTO) gene confers risk of obesity and modulates BMI in the Chinese population.** *Diabetes* 2008, **57:**2245-2252.

153. Ramya K, Radha V, Ghosh S, Majumder PP, Mohan V: **Genetic variations in the FTO gene are associated with type 2 diabetes and obesity in south Indians (CURES-79).** *Diabetes Technol Ther* 2011, **13:**33-42.

154. Renstrom F, Payne F, Nordstrom A, Brito EC, Rolandsson O, Hallmans G, Barroso I, Nordstrom P, Franks PW: **Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden.** *Hum Mol Genet* 2009, **18:**1489-1496.

155. Izaks GJ, Gansevoort RT, van der Knaap AM, Navis G, Dullaart RP, Slaets JP: **The association of APOE genotype with cognitive function in persons aged 35 years or older.** *PLoS ONE* 2011, **6:**e27415.

156. Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, Schjeide BM, Hooli B, Divito J, Ionita I, et al: **Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE.** *Am J Hum Genet* 2008, **83:**623-632.

157. Jiang X, Barmada MM, Cooper GF, Becich MJ: **A Bayesian Method for Evaluating and Discovering Disease Loci Associations.** *PLoS ONE* 2011, **6:**e22075.

158. Hu WT, Chen-Plotkin A, Arnold SE, Grossman M, Clark CM, Shaw LM, Pickering E, Kuhn M, Chen Y, McCluskey L, et al: **Novel CSF biomarkers for Alzheimer's disease and mild cognitive impairment.** *Acta Neuropathologica* 2010, **119:**669-678.

159. Shi H, Medway C, Bullock J, Brown K, Kalsheker N, Morgan K: **Analysis of Genome-Wide Association Study (GWAS) data looking for replicating signals in Alzheimer's disease (AD).** *Int J Mol Epidemiol Genet* 2010, **1:**53-66.

160. Shi H: *Complementary Approaches to Analyse Genetic Data in Late Onset Alzheimer's Disease (LOAD).* University of Nottingham; 2012.

161. Jiang X, Neapolitan RE, Barmada MM, Visweswaran S, Cooper GF: **A fast algorithm for learning epistatic genomic relationships.** *AMIA Annu Symp Proc* 2010, **2010:**341-345.

162. Saarela J, von Schantz C, Peltonen L, Jalanko A: **A novel aspartylglucosaminuria mutation affects translocation of aspartylglucosaminidase.** *Hum Mutat* 2004, **24:**350-351.

163. Shi H, Belbin O, Medway C, Brown K, Kalsheker N, Carrasquillo M, Proitsi P, Powell J, Lovestone S, Goate A, et al: **Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS).** *Neurobiol Aging* 2012, **33:**1849 e1845-1818.

164. Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, Kremer D, van der Breggen R, Suchiman HE, Lakenberg N, et al: **Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited.** *Aging Cell* 2011, **10:**686-698.

165. Li G, Jiang H, Chang M, Xie H, Hu L: **HDAC6 α-tubulin deacetylase: A potential therapeutic target in neurodegenerative diseases.** *Journal of the neurological sciences* 2011, **304:**1-8.

166. Butterfield DA, Abdul HM, Opii W, Newman SF, Joshi G, Ansari MA, Sultana R: **Pin1 in Alzheimer's disease.** *J Neurochem* 2006, **98:**1697-1706.

167. Driver JA, Lu KP: **Pin1: a new genetic link between Alzheimer's disease, cancer and aging.** *Curr Aging Sci* 2010, **3:**158-165.

168. Okuda T, Higashi Y, Kokame K, Tanaka C, Kondoh H, Miyata T: **Ndrg1-deficient mice exhibit a progressive demyelinating disorder of peripheral nerves.** *Mol Cell Biol* 2004, **24:**3949-3956.

169. Kalaydjieva L, Gresham D, Gooding R, Heather L, Baas F, de Jonge R, Blechschmidt K, Angelicheva D, Chandler D, Worsley P, et al: **N-myc downstream-regulated gene 1 is mutated in hereditary motor and sensory neuropathy-Lom.** *Am J Hum Genet* 2000, **67:**47-58.

170. Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH, Saykin AJ, Orro A, Lupoli S, Salvi E, et al: **Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease.** *PLoS ONE* 2009, **4:**e6501.

171. Broggini T, Nitsch R, Savaskan NE: **Plasticity-related gene 5 (PRG5) induces filopodia and neurite growth and impedes lysophosphatidic acid- and nogo-A-mediated axonal retraction.** *Mol Biol Cell* 2010, **21:**521-537.

172. Harter PN, Bunz B, Dietz K, Hoffmann K, Meyermann R, Mittelbronn M: **Spatio-temporal deleted in colorectal cancer (DCC) and netrin-1 expression in human foetal brain development.** *Neuropathol Appl Neurobiol* 2010, **36:**623-635.

173. Pan Y, Wang KS, Aragam N: **NTM and NR3C2 polymorphisms influencing intelligence: family-based association studies.** *Prog Neuropsychopharmacol Biol Psychiatry* 2011, **35:**154-160.

174. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, Barkhof F, Radue EW, Lindberg RL, Uitdehaag BM, Johnson MR, et al: **Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis.** *Hum Mol Genet* 2009, **18:**767-778.

175. Bossers K, Wirz KT, Meerhoff GF, Essing AH, van Dongen JW, Houba P, Kruse CG, Verhaagen J, Swaab DF: **Concerted changes in transcripts in the prefrontal cortex precede neuropathology in Alzheimer's disease.** *Brain* 2010, **133:**3699-3723.

176. Kohannim O, Hibar DP, Stein JL, Jahanshad N, Hua X, Rajagopalan P, Toga AW, Jack CR, Jr., Weiner MW, de Zubicaray GI, et al: **Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression.** *Front Neurosci* 2012, **6:**115.

177. Shi H, Medway C, Brown K, Kalsheker N, Morgan K: **Using Fisher's method with PLINK 'LD clumped' output to compare SNP effects across Genome-wide Association Study (GWAS) datasets.** *Int J Mol Epidemiol Genet* 2011, **2:**30-35.

178. Williams C, Mehrian Shai R, Wu Y, Hsu YH, Sitzer T, Spann B, McCleary C, Mo Y, Miller CA: **Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease.** *PLoS ONE* 2009, **4:**e4936.

179. Um JW, Kaufman AC, Kostylev M, Heiss JK, Stagi M, Takahashi H, Kerrisk ME, Vortmeyer A, Wisniewski T, Koleske AJ, et al: **Metabotropic glutamate receptor 5 is a coreceptor for Alzheimer abeta oligomer bound to cellular prion protein.** *Neuron* 2013, **79:**887-902.

180. Munoz L, Ammit AJ: **Targeting p38 MAPK pathway for the treatment of Alzheimer's disease.** *Neuropharmacology* 2010, **58:**561-568.

181. Zhang JH, Barr VA, Mo Y, Rojkova AM, Liu S, Simonds WF: **Nuclear localization of G protein beta 5 and regulator of G protein signaling 7 in neurons and brain.** *J Biol Chem* 2001, **276:**10284-10289.

182. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, et al: **Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4.** *PLoS Genet* 2007, **3:**e58.